UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 26

Can Autonomous AI Agents Discover the Fundamental Value of Real Estate?

Evidence from Quebec Transactions and Multi-Agent Deliberation

Simon-Pierre Boucher

Département des sciences administratives
Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché
283, boulevard Alexandre-Taché
Gatineau, Québec, Canada J9A 1L8

October 2026

Version 1.0

Abstract

Committees of role-specialised language-model agents that debate before they answer are increasingly proposed as valuation tools. We ask whether the debate adds information. In a preregistered experiment on 631,866 arm's-length residential sales registered in Quebec between 2021 and 2026, evaluated under a leakage-audited as-of information set, we compare a hedonic regression, a hierarchical repeat-sales model, gradient boosting and a conformal automated valuation model with a single language model, twelve independent role-specialised agents, structured debate of one to four rounds, debate with a blinded skeptic and deterministic verification, a placebo committee and a cost-matched single agent, all on identical evidence packets. A model of valuation as decentralised information aggregation shows that opinion exchange contracts dispersion without reducing error, that evidence exchange helps only outside the common view, and that no linear deliberation beats the best round-0 pool. The data agree: on the test subsample the debate reaches a mean absolute log error of 0.158 against 0.159 for the silent committee, 0.157 for the placebo and 0.163 for the cost-matched agent; every preregistered paired contrast includes zero, and rounds halve the agents' disagreement without moving accuracy. Committees match the automated valuation model (0.171 on the full test partition) but state intervals thirty percent too narrow, and in the 2026 holdout every agentic arm overvalues by ten log points where the AVM is unbiased. A two-round debate costs about 1.5 US dollars per property. Forecasting accuracy is kept distinct from fundamental value, which is triangulated with 92,089 resales, rents and user cost.

Keywords: automated valuation models, large language models, multi-agent deliberation, information aggregation, house prices, calibration, appraisal, Quebec

JEL Classification: R31, C53, C45, D83, G21

1. Introduction

Few numbers matter as much to a household, a lender or a municipality as the value of a dwelling, and few are as hard to observe. A transaction price is realised once, under the particular circumstances of one buyer and one seller; between transactions the value must be estimated, and the estimate governs how much a bank will lend, how much tax the owner pays, and at what price a listing is pitched. For most of the twentieth century the estimate was produced by an appraiser applying the sales-comparison method to a handful of recent sales. Over the last two decades automated valuation models, descended from the hedonic regression of Rosen (1974) and now dominated by gradient-boosted trees, have taken over the volume end of this work, with accuracy that rivals or exceeds manual appraisal in liquid markets and deteriorates, as appraisal does, in thin ones (Kok et al., 2017; Bogin and Shui, 2020; Mullainathan and Spiess, 2017). The appraisal they replace is itself known to be a biased instrument: appraisers shrink toward comparables and anchor on the contract price, so that the recorded price partly reflects the valuation that financed it (Quan and Quigley, 1991; Eriksen et al., 2019; Calem et al., 2021).

Large language models have entered this landscape with a different proposition. A model that can read a dossier of attributes, comparable sales and market indicators, reason about them in words and return a number is, in form, an appraiser, and a committee of such models given distinct roles, a comparables specialist, a market economist, an underwriter, a skeptic, is, in form, a valuation panel. The multi-agent literature in machine learning has argued that such committees improve on a single model when their members debate: independent answers are exchanged, criticised and revised, and the revised answers are aggregated (Du et al., 2023; Liang et al., 2024). If that argument carries over to valuation, the lender of the near future will consult a deliberating panel of artificial analysts rather than a regression. The question this paper asks is whether the deliberation adds anything. Put precisely: does communication among heterogeneous autonomous valuation agents add measurable information about the value of a property, beyond what the same agents know before they communicate, beyond what the same inference budget buys from one agent, and beyond what a strong tabular model extracts from the same data?

The question is open because two literatures give opposite answers to it. The theory of forecast combination says that the value of a crowd is bounded by the correlation of its errors (Bates and Granger, 1969; Clemen and Winkler, 1985), that diversity rather than individual ability is what averaging rewards (Hong and Page, 2004), and that social influence on a network is wise only when no member dominates (DeGroot, 1974; Golub and Jackson, 2010); experiments with human crowds find that even mild exposure to others' estimates narrows the range of opinion without improving its centre (Lorenz et al., 2011), although decentralised exchange can help under the right structure (Becker et al., 2017; Navajas et al., 2018). The language-model literature is divided along the

same line. Against the early debate results stand findings that majority voting over independent samples captures most of the gain (Wang et al., 2022), that debate rarely beats voting at matched cost (Smit et al., 2024; Wu et al., 2025), that similar models debating converge on the majority view including its errors (Estornell and Liu, 2024), and that homogeneous agents with uniform updating cannot improve expected correctness by debating at all (Zhu et al., 2026). Since the gains attributed to deliberation are routinely obtained by spending the same tokens on repeated sampling (Brown et al., 2024; Snell et al., 2024), any test of deliberation must separate communication from compute. In real estate, the handful of recent papers that apply language models to valuation use one model at a time: they report competitive point estimates and overconfident intervals from in-context examples (Geerts et al., 2025), improve a boosted AVM with text features extracted by a model (Faxvaag et al., 2025), fine-tune a model for quantile prediction (Vedula et al., 2025), or ask whether a model has internalised user-cost logic (Li, 2026). None has tested a deliberating committee against information-matched and cost-matched controls on registered transactions.

We answer the question with a preregistered experiment on 631,866 arm's-length residential sales registered in Quebec between January 2021 and October 2026, drawn from the public land register and joined to the municipal assessment roll (422,542 houses, 122,976 condominium units and 86,348 small multi-unit buildings). The sample is split chronologically into a training window to June 2024, a validation window to March 2025, a test window from April to December 2025 and an untouched 2026 holdout, with a spatial holdout of three entire regions and a prospective set of 1,182 sales registered after the protocol was frozen. Every arm of the experiment receives the same information set, defined by a cutoff on the day before the sale and enforced by an audit of nine leakage checks, and the same opportunity to select comparables. Four non-agentic baselines, a hedonic regression, a hierarchical model with a repeat-sales adjustment, a tuned gradient-boosted model and an automated valuation model with as-of comparable features and conformal intervals, are fitted on the training window and tuned on validation only. Seven agentic arms share one evidence packet per property: a single generalist model; twelve role-specialised agents with least-privilege views of the packet, aggregated without communication; the same twelve agents in a structured debate of one, two or four rounds of critique and revision; the debate with a blinded skeptic and deterministic verification of citations and adjustments; a placebo committee with generic prompts on the same views; a full-view variant that removes the evidence partition; and a cost-matched single agent sampled twelve times and self-refined. The agents are drawn from two providers at identical list prices and assigned to roles by a seeded balanced permutation, so that a provider contrast is not a budget contrast. Every number in the paper traces to a stored transcript, and the four primary contrasts, each paired on the same properties and tested by a block bootstrap over municipality-by-month cells with Holm's correction (Cameron et al., 2008; Holm, 1979), were written down before any test property was valued.

The experiment is built on a model of valuation as decentralised information aggregation that tells us what to expect and what would count as evidence. Under correlated signals the equal-weight pool of K agents has a mean-squared error floor of $\rho\sigma^2$ that no committee size removes; a revision rule in which agents move toward the committee statistic contracts the dispersion of their estimates by $(1-\lambda)^r$ after r rounds without changing the pooled error, so that convergence is a measure of the anchoring weight, not of accuracy; evidence exchange lowers the pooled error only when it redistributes weight toward evidence outside the common view, and can raise it when it concentrates attention on evidence that is already popular; permuting role prompts while holding the evidence views fixed leaves the error distribution unchanged under additive separability; and no linear deliberation can beat the best pool of the round-0 estimates. Each proposition is mapped to a preregistered contrast, and the mapping is what allows a null result to be read as a finding rather than as an absence of one.

The result is a precise null on deliberation and a more nuanced picture of the committee. On the test subsample at the mid-tier models, the single model reaches a mean absolute log error of 0.166, the twelve independent roles 0.159, the two-round debate 0.158, the debate with skeptic and verification 0.158, the placebo committee 0.157 and the cost-matched single agent 0.163; the automated valuation model scores 0.171 on the full test partition and 0.163 on the same eighty properties. Every preregistered paired contrast includes zero: the debate improves on the independent committee by 0.001 log points with a 95 percent interval of $[-0.006, 0.005]$, on the placebo committee by nothing at all, $[-0.005, 0.006]$, on the cost-matched single agent by 0.006, $[-0.017, 0.004]$, and on the AVM by 0.005, $[-0.033, 0.026]$, with Holm-adjusted p-values of one. The intervals rule out relative improvements from deliberation larger than about four percent against the independent committee and about seven percent against the compute-matched agent, which is not quite the two percent threshold that the preregistration set as the falsification criterion; we return to this in Section 13. Rounds of debate behave exactly as the theory predicts: the interquartile range of the agents' log estimates falls by roughly half after the first round and keeps falling, the anchoring index of the revisions is 0.7 to 0.8, and accuracy does not move. In the 2026 holdout and in the prospective set the debate is marginally worse than the independent committee, by 0.003 and 0.009 log points. The committees are not inaccurate: they are statistically indistinguishable from the AVM on the test subsample, and their median absolute percentage error is lower. But their intervals are about thirty percent narrower than the AVM's conformal intervals and cover 86 to 89 percent of prices at a nominal 90 percent, 78 to 81 percent at the small-model tier, and in the 2026 holdout the three agentic arms overvalue by ten log points where the AVM is unbiased, a drift that no amount of deliberation corrects because every agent in the committee shares it. The committee's explanations are faithful in the narrow sense the audit can verify, every cited evidence identifier exists in the packet, 96 to 99 percent of stated adjustments lie within forty percent of the

raw comparable price and about 99 percent of points lie inside the hull of the adjusted comparables, but faithful explanations do not translate into calibrated intervals. A two-round debate costs about 1.52 US dollars and 5.6 minutes per property, against 0.23 dollars for the independent committee and about two cents for a single model; the AVM costs nothing at the margin. Finally, on the 92,089 sales with a later resale, the gradient-boosted baselines lie measurably closer to the market-adjusted resale value than the realised first price does; the same test for the agentic arms is specified but deferred, because too few of the evaluated properties have yet resold.

The paper makes four contributions. The first is a design that isolates communication from its confounds. The independent committee supplies the round-0 estimates of the debate, so that their difference is communication and nothing else; the cost-matched single agent supplies the same tokens without roles or communication; the placebo committee supplies the same views and mechanics without mandates; the full-view variant supplies the same mandates without the partition; and the information set, the comparable pool, the providers and the budget are held fixed across all of them. To our knowledge no evaluation of language-model deliberation in any domain has run all of these controls on the same properties under a frozen protocol. The second is the theory, which is elementary but, we think, clarifying: it shows that the convergence statistics on which debate systems are often judged are uninformative about accuracy, identifies the single channel through which communication can add information, and shows that the placebo contrast identifies prompt-evidence complementarity rather than the value of role diversity. The third is the Quebec evidence itself, the first on language-model valuation in the province and, to our knowledge, the first to use registered sales from a public land register rather than listing or portal data, with a leakage-defended as-of pipeline, chronological, spatial and prospective holdouts, calibration at four levels, a faithfulness audit, measured run-to-run variation and full cost accounting, and with the distinction between transaction-price prediction and fundamental value kept explicit throughout. The fourth is the set of open artefacts: the code, the configuration, the role prompts with their hashes, the freeze record, the ledgers and the evaluation outputs are released, together with a labelled synthetic dataset and a mock provider that allow the entire pipeline to run without keys or proprietary data, so that the null can be challenged by extending the same runs rather than by repeating them.

Three bodies of prior work are closest. The hedonic and AVM literature, from Rosen (1974) to the tabular benchmarks of Kok et al. (2017) and Bogin and Shui (2020), provides the baselines and the standard that any new method must meet; the Quebec hedonic studies reviewed in Section 2 provide the institutional antecedents. The forecast-combination and social-learning literature, from Bates and Granger (1969) and Clemen and Winkler (1985) to Golub and Jackson (2010) and Lorenz et al. (2011), provides the propositions, which we restate for a committee with an explicit evidence partition and a provider structure. The multi-agent debate literature, from Du et al. (2023) to the

sceptical results of Smit et al. (2024), Estornell and Liu (2024) and Zhu et al. (2026), provides the hypothesis under test and the warning that compute must be matched; our contribution relative to it is a domain with a continuous, noisy, selected target and strong non-neural baselines, and a design in which the channels those papers conflate are separately identified. We do not claim priority for applying language models to property valuation, which Geerts et al. (2025), Faxvaag et al. (2025), Vedula et al. (2025) and others have done; our claim is about what deliberation among such models does and does not add. The policy relevance follows from the finding rather than preceding it: if committees of agents built on two or three foundation models are no more accurate than a tabular model, systematically narrower in their stated uncertainty, and jointly biased when the market moves, then the concern about algorithmic monoculture raised by Kleinberg and Raghavan (2021) and Bommasani et al. (2022) applies to collateral valuation with particular force, and the welfare model of Section 12 shows why the variance of system losses, not their mean, is where that concern bites.

The remainder of the paper is organised as follows. Section 2 reviews the five literatures on which the study draws and states what it adds. Section 3 sets out the model of valuation as information aggregation, defines the objects that the empirical work can and cannot identify, and derives the propositions. Section 4 describes the Quebec records and the institutional facts that shape the identification strategy, and Section 5 the sample, the partitions, the information-set rules and the leakage audit. Section 6 describes the evidence packet, the roles, the output contract, the providers, the protocols and the audit trail, and Section 7 the arms, the preregistered hypotheses, the metrics, the inference and the freeze. Section 8 reports the main results, Section 9 the ablations and robustness checks, including the rounds grid, the provider committees, the spatial holdout and the prospective set, and Section 10 the calibration and faithfulness audits. Section 11 triangulates fundamental value with repeat sales, rents and user cost. Section 12 works out what the accuracy results do and do not imply for buyers, sellers and lenders. Section 13 records the limitations, and Section 14 concludes. The appendices contain the reproducibility record, the role prompts, the output contract, additional tables and the proofs.

2. Related Literature

This paper draws on five literatures, reviewed briefly below before we state what the study adds and what it does not claim.

2.1. Hedonic Models and Automated Valuation

The hedonic approach treats a dwelling as a bundle of characteristics whose implicit prices are revealed in equilibrium (Rosen, 1974); applied reviews document wide variation in functional form,

attributes and spatial treatment (Malpezzi, 2003; Sirmans et al., 2005), and Hill (2013) surveys the resulting index methods. Repeat-sales indexes (Bailey et al., 1963; Case and Shiller, 1987) abstract from attribute measurement but require repeated transfers. Spatial dependence is first-order: Pace et al. (1998a) and Pace et al. (1998b) introduce spatial and spatiotemporal autoregressive structures for real estate, systematised by LeSage and Pace (2009). For Quebec, the closest hedonic antecedents are the Quebec City and Montreal studies of Des Rosiers, Thériault and Dubé on access and neighbourhood factors (Des Rosiers et al., 2000), transit investments (Dubé et al., 2014), the joint determination of price and time on the market (Dubé and Legros, 2016) and residential reconversions (Dubé et al., 2023). On the automated-valuation side, Mullainathan and Spiess (2017) frame prediction as a distinct econometric task; Kok et al. (2017), Mayer et al. (2019) and Pérez-Rave et al. (2019) compare manual appraisal, hedonic updating and big-data regression; gradient-boosted trees (Chen and Guestrin, 2016; Prokhorenkova et al., 2018) remain the strongest tabular baselines (Grinsztajn et al., 2022); Bogin and Shui (2020) show that both AVM and appraisal accuracy deteriorate in thin rural markets; and listing text adds information at the margin (Nowak and Smith, 2017; Shen and Ross, 2021; Zhang et al., 2024).

2.2. Appraisal Accuracy and Bias

Quan and Quigley (1991) model appraisal as signal extraction with shrinkage toward comparables, and Northcraft and Neale (1987) show that professionals anchor on listing prices. Cho and Megbolugbe (1996) link appraisal to redlining and Nakamura (2010) to house-price uncertainty. Purchase-mortgage data reveal confirmation bias toward contract prices, with a sharply asymmetric appraisal-to-contract ratio (Calem et al., 2021; Eriksen et al., 2019, 2020), and low appraisals pull contract prices down (Fout et al., 2022). Agarwal et al. (2015) measure the collateral-valuation response of constrained borrowers, Agarwal et al. (2020) ask whether regulation de-biases appraisers, and Shui and Murthy (2018) find no quality difference between appraisals routed through management companies and others. Racial valuation gaps are documented by Howell and Korver-Glenn (2018), Howell and Korver-Glenn (2021), Perry et al. (2018) and Freddie Mac (2021), and Fuster et al. (2022) show that algorithmic credit decisions reshape such disparities. These findings motivate our treatment of prices as noisy, possibly biased proxies.

2.3. Information Aggregation and Deliberation

The case for averaging independent judgements (Galton, 1907; Surowiecki, 2004) is formalised in the forecast-combination literature (Bates and Granger, 1969; Clemen, 1989; Timmermann, 2006), where trimmed, extremised or logit-pooled aggregates dominate simple means when forecasts are miscalibrated (Ranjan and Gneiting, 2010; Satopää et al., 2014; Jose et al., 2014). The value of

a crowd is bounded by the correlation of its errors (Clemen and Winkler, 1985), and diversity often matters more than individual ability (Hong and Page, 2004; Lamberson and Page, 2012; Davis-Stober et al., 2014). Social influence is double-edged: averaging on a network is wise only if no agent is too influential (DeGroot, 1974; Golub and Jackson, 2010), Bayesian learning on networks has its own limits (Acemoglu et al., 2011), sequential observation produces cascades (Bikhchandani et al., 1992), and public signals crowd out private information (Morris and Shin, 2002). Experimentally, Lorenz et al. (2011) find that mild social influence narrows the diversity of estimates and undermines crowd wisdom, Becker et al. (2017) find that decentralised communication can improve accuracy, Navajas et al. (2018) report that small deliberating groups outperform large non-interacting crowds, and Prelec et al. (2017) recover the truth from meta-knowledge when the majority is wrong.

The literature on large language models (LLMs) revisits these questions with artificial agents. Multi-agent debate was proposed to improve factuality and reasoning (Du et al., 2023; Liang et al., 2024; Chan et al., 2023), building on the debate paradigm of Irving et al. (2018) and on evidence that persuasive debaters help weaker judges (Khan et al., 2024; Michael et al., 2023); frameworks such as CAMEL, AutoGen, MetaGPT and generative agents (Li et al., 2023; Wu et al., 2023; Hong et al., 2023; Park et al., 2023) made role-based agent societies easy to build, and Guo et al. (2024) survey the design space. Whether debate adds information beyond ensembling is contested. Wang et al. (2022) show that majority voting over independent samples captures much of the gain; Smit et al. (2024), Wang et al. (2024), Kaesberg et al. (2025) and Wu et al. (2025) find that debate often fails to beat voting at matched cost and that majority pressure suppresses independent correction; Estornell and Liu (2024) show that debate among similar models converges to the majority view, including shared misconceptions; and Zhu et al. (2026) show that with homogeneous agents and uniform updating debate cannot improve expected correctness. The compute-scaling results of Chen et al. (2024), Li et al. (2024), Brown et al. (2024) and Snell et al. (2024) are decisive for our design: any benefit of deliberation must be separated from more samples or more inference compute. Finally, LLMs are calibrated only under some elicitation schemes and are frequently overconfident (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024), and their stated reasoning is not always faithful to the computation behind the answer (Turpin et al., 2023; Lanham et al., 2023); we therefore evaluate explanations by the faithfulness standard of Jacovi and Goldberg (2020) rather than by plausibility.

2.4. LLMs as Forecasters and Valuers

Halawi et al. (2024) approach human crowd accuracy on forecasting tournaments, and Schoenegger et al. (2024) find that an ensemble of twelve LLMs rivals a human crowd whereas single models lag it (Schoenegger and Park, 2023). In real estate, prior work is recent and sparse. Geerts et al.

(2025) evaluate single LLMs on international housing datasets with in-context examples and report competitive point estimates but overconfident intervals and weak spatial reasoning; Faxvaag et al. (2025) use LLM-extracted listing-text features to improve a gradient-boosting AVM; Vedula et al. (2025) fine-tune LLMs for quantile price prediction; Chen and Si (2024) compare ChatGPT with machine learning for Shanghai rents; Bae et al. (2025) propose a single-model valuation and reasoning pipeline; and Li (2026) ask whether one LLM internalises user-cost logic when valuing dwellings for surveyed UK households. None implements deliberation among specialised agents, cost-matched comparisons with ensembles and strong non-LLM baselines on arm's-length transactions, or a joint evaluation of calibration and explanation faithfulness.

2.5. Uncertainty, Evaluation and Model Comparison

Proper scoring rules (Matheson and Winkler, 1976; Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014) and quantile loss (Koenker and Bassett, 1978) reward calibrated and sharp predictive distributions, and conformal prediction (Vovk et al., 2005; Lei et al., 2018; Romano et al., 2019; Angelopoulos and Bates, 2023) provides finite-sample coverage against which LLM-elicited intervals are benchmarked. Accuracy comparisons use Diebold and Mariano (1995) with the correction of Harvey et al. (1997), the conditional framework of Giacomini and White (2006) and the model confidence set of Hansen et al. (2011); dependence across properties, neighbourhoods and time is handled by block and stationary bootstraps (Künsch, 1989; Politis and Romano, 1994) and clustered bootstrap inference (Cameron et al., 2008), and multiplicity by Holm (1979).

2.6. Fundamental Value and Transaction Prices

A transaction price is the outcome of search and bargaining, not an observation of fundamental value. The user-cost framework (Poterba, 1984; Himmelberg et al., 2005) and variance decompositions of the rent-price ratio (Campbell et al., 2009) define fundamental value through discounted housing services, financing costs, taxes, risk and expected appreciation, and evidence of predictable returns and extrapolative expectations (Case and Shiller, 1989; Glaeser and Nathanson, 2017) implies persistent deviations from fundamentals. Search frictions (Wheaton, 1990; Han and Strange, 2015), seller loss aversion (Genesove and Mayer, 2001) and collateral constraints (Stein, 1995; Kiyotaki and Moore, 1997; Iacoviello, 2005; Mian and Sufi, 2011) add transaction-specific noise and liquidity premia; official Canadian and Quebec data products (Statistics Canada, 2024; Ministère des Affaires municipales et de l'Habitation du Québec, 2024) frame the institutional setting. Finally, Kleinberg and Raghavan (2021) show that algorithmic monoculture can lower welfare even when each algorithm is individually more accurate, and Bommasani et al. (2022) document outcome homogenisation when decision-makers share a model; because our agents rest on a few foundation

models, correlated errors are a market-design concern as well as a statistical one.

2.7. Position of This Paper

The paper makes three contributions. First, it brings the theory of aggregation under correlated signals to bear on LLM deliberation, deriving testable conditions on correlation, diversity and communication structure under which deliberation reduces or increases mean-squared valuation error. Second, it provides, to our knowledge, the first evaluation of specialised multi-agent deliberation for residential valuation on arm's-length transactions that is matched on information, compute and provider against hedonic, repeat-sales, gradient-boosting, AVM and single-LLM baselines, reporting calibration, tail errors and explanation faithfulness alongside point accuracy. Third, it separates transaction-price prediction from fundamental value, triangulated with repeat sales, rents and user-cost diagnostics, in a Quebec setting with hedonic but no LLM-based evidence. We do not claim to be the first to apply LLMs to property valuation, to study multi-agent debate, or to model Quebec housing; the novelty lies in the joint design that isolates whether communication among heterogeneous agents adds information, and in reporting a negative answer should the data support it.

3. Theory: Valuation as Decentralised Information Aggregation

This section formalises the object that every arm of the experiment is asked to estimate, and derives what a committee of partially informed, possibly correlated agents can and cannot achieve by deliberating. The model is deliberately simple: linear signals, Gaussian errors, and a fixed-weight revision rule. Its purpose is not realism but discipline. It delivers four results that the empirical design is built to test: the pooled error of a committee has a floor that no number of agents removes when their errors share a common component; opinion-only deliberation manufactures consensus without information; evidence exchange lowers the pooled error only when it redistributes weight towards evidence that the committee under-uses; and a role permutation that leaves the evidence partition untouched does not change the expected error. Proofs are collected in Appendix E. Table 1 fixes the notation.

3.1. Prices, values and the estimation target

Definition 1 (Transaction price)**.** The transaction price P_{it} is the consideration at which property i transfers between unrelated parties at date t. All empirical metrics in this paper are computed on $p_{it} = \log P_{it}$ for arm's-length residential sales.

Definition 2 (Market-consistent value and the selected target)**.** Given the information set $\mathscr{I}_t$, the

Table 1. Notation.

$P,\ p = \log P$	transaction price of the subject property and its logarithm
$\mathscr{I}_t$	information available at the valuation date t (cutoff enforced in every arm)
D	sale indicator; $D = 1$ if the property transacts in the evaluation window
$v^* = \mathbb{E}[p \mid \mathscr{I}_t]$	latent market-consistent log value
$\tilde{v} = \mathbb{E}[p \mid \mathscr{I}_t, D = 1]$	selected market value, the operational estimation target; $\kappa = \tilde{v} - v^*$
$\varepsilon = p - \tilde{v}$	transaction noise (bargaining, urgency, liquidity)
$V^F,\ R_t,\ u_t$	structural fundamental value, net rent (housing-service flow), user cost of capital
K	number of agents; $\mathbf{1}$ is the K-vector of ones
$s_k = v + b_k + \eta_k$	signal (round-0 estimate) of agent k; bias b_k, error η_k
$\Sigma,\ \sigma^2,\ \rho$	error covariance; common variance and pairwise correlation in the exchangeable case
$w,\ \hat{y}(w) = w's$	linear pool weights ($w'\mathbf{1} = 1$) and pooled estimate
$x^{(r)},\ \lambda,\ T$	vector of estimates after round r; anchoring weight; committee statistic (mean or median)
$W,\ \pi$	influence matrix of a linear deliberation and its consensus weights
$\zeta,\ \Gamma$	revision term carrying new evidence or persuasion; its covariance
$z_j = v + c + u_j$	evidence item $j = 1,\ldots,J$; common evidence error c (variance σ_c^2), item error u_j (variance σ_u^2)
$A_k,\ a_k,\ \omega$	evidence view of agent k, its within-view weights, implied committee weights $\omega = K^{-1}\sum_k a_k$
$g_m,\ e_k$	provider-common error ($m = 1,\ldots,M$; variance σ_g^2) and idiosyncratic reasoning error (variance τ_k^2)

latent market-consistent log value is $v^*_{it} = \mathbb{E}[p_{it} \mid \mathscr{I}_t]$, the conditional expectation of the price the property would fetch at t. Because prices are observed only for properties that sell, the quantity any price-trained method actually targets is the selected market value $\tilde{v}_{it} = \mathbb{E}[p_{it} \mid \mathscr{I}_t, D_{it} = 1]$, and we write $\kappa_{it} = \tilde{v}_{it} - v^*_{it}$ for the selection term.

Definition 3 (Appraisal estimate and investment value)**.** An appraisal estimate is any $\mathscr{I}_t$-measurable function $\hat{y}_{it} = \hat{y}(\mathscr{I}_t)$, whether produced by a hedonic regression, a gradient-boosted model, a single language model or a committee of agents. An investment value is the present value of the property to a particular holder under that holder's discount rate, tax position and horizon; it is neither the target nor the benchmark of any arm.

Definition 4 (Structural fundamental value)**.** Let R_t denote the net flow of housing services (the net rent the property commands or would command), i_t the nominal cost of mortgage funds, τ_p the property-tax rate, δ maintenance and depreciation, γ_t a risk and liquidity premium and $\pi^e_t = \mathbb{E}_t[P_{t+1}/P_t] - 1$ the expected one-period capital gain. The one-period user-cost condition of Poterba (1984), in the form used by Himmelberg et al. (2005), states that a marginal household is indifferent between owning and renting when

$$R_t = P_t\, u_t, \qquad u_t = i_t + \tau_p + \delta + \gamma_t - \pi^e_t, \tag{1}$$

where the Canadian tax treatment (no deductibility of mortgage interest or property taxes for owner-occupiers, no tax on imputed rent and an exemption of principal-residence gains) removes the marginal-tax-rate terms of the original United States formulation. Writing $k_t = i_t + \tau_p + \delta + \gamma_t$ and iterating (1) forward,

$$P_t = \sum_{j=0}^{H-1} \mathbb{E}_t \left[\frac{R_{t+j}}{\prod_{h=0}^{j}(1+k_{t+h})} \right] + \mathbb{E}_t \left[\frac{P_{t+H}}{\prod_{h=0}^{H-1}(1+k_{t+h})} \right]. \tag{2}$$

The structural fundamental value V_t^F is the first term of (2) as $H \to \infty$ under the transversality condition that the second term vanishes; the gap $P_t - V_t^F$ is a rational bubble or a persistent expectational component.

Three objects therefore coexist: the price P, the market-consistent value $V^* = \exp(v^*)$ that an unbiased appraisal targets, and the structural value V^F anchored in the service flow. They differ by economically meaningful terms. The price differs from $\tilde{v}$ by transaction noise ε, which collects bargaining outcomes, seller urgency, time on market and the idiosyncratic match between buyer and dwelling; $\tilde{v}$ differs from v^* by the selection term κ, which is non-zero whenever the decision to sell is correlated with the unobserved component of the price, as the loss-aversion evidence of Genesove and Mayer (2001) implies; and v^* differs from $\log V^F$ by expectational and liquidity components that Stein (1995) shows can be large when down-payment constraints bind. The following lemma states what price-based evaluation can and cannot rank.

Lemma 1 (Price-based evaluation ranks estimators against the selected value)**.** *For any $\mathscr{I}_t$-measurable estimate $\hat{y}$,*

$$\mathbb{E}\big[(\hat{y}-p)^2 \mid D=1\big] = \mathbb{E}\big[(\hat{y}-\tilde{v})^2 \mid D=1\big] + \mathrm{Var}(p \mid \mathscr{I}_t, D=1). \tag{3}$$

The second term is common to every estimator. Hence the ranking of estimators by mean-squared error against realised prices equals their ranking by distance to $\tilde{v}$, and nothing in (3) identifies the distance to v^* or to $\log V^F$.

Proof. Expand $(\hat{y}-p)^2 = (\hat{y}-\tilde{v})^2 + (p-\tilde{v})^2 + 2(\hat{y}-\tilde{v})(p-\tilde{v})$. Conditioning the cross term on $(\mathscr{I}_t, D=1)$, where $\hat{y}-\tilde{v}$ is fixed, gives $\mathbb{E}[p-\tilde{v} \mid \mathscr{I}_t, D=1] = 0$ by the definition of $\tilde{v}$. Taking expectations conditional on $D=1$ yields (3). □

Lemma 1 licenses the empirical comparisons of Sections 8 and 9: differences in price-based error between arms are differences in distance to a well-defined target, the selected market value. It also delimits them. A method can rank first against $\tilde{v}$ while being uninformative about V^F, and the triangulation of Section 3.6 is required to say anything about fundamentals. In what follows we write v for the target of the agents and treat it as fixed, so that every error is measured relative to $\tilde{v}$.

3.2. One-shot aggregation of correlated opinions

Assumption 1 (Linear-Gaussian signals)**.** Each agent $k = 1, \dots, K$ reports $s_k = v + b_k + \eta_k$, where $b = (b_1, \dots, b_K)'$ is a vector of biases and $\eta = (\eta_1, \dots, \eta_K)' \sim N(0, \Sigma)$ with Σ positive definite. K is finite. Where optimal weights are discussed, Σ is treated as known.

A linear pool with weights w, $w'\mathbf{1} = 1$, produces $\hat{y}(w) = w's$ with error $w'b + w'\eta$ and

$$\mathrm{MSE}(w) = (w'b)^2 + w'\Sigma w. \tag{4}$$

The two pools of empirical interest are the equal-weight mean, $w = \mathbf{1}/K$, and the minimum-variance pool of Bates and Granger (1969). The exchangeable case, in which every agent is equally accurate and every pair of agents is equally correlated, is the natural model of a committee of interchangeable language models that read the same dossier.

Proposition 1 (Correlated opinions)**.** *Let $b = 0$ and $\Sigma = \sigma^2[(1-\rho)I + \rho\mathbf{1}\mathbf{1}']$ with $-1/(K-1) \le \rho < 1$. Then the equal-weight mean has*

$$\mathrm{MSE}_{\mathrm{mean}}(K, \rho) = \sigma^2\Big[\frac{1-\rho}{K} + \rho\Big], \tag{5}$$

equal weights are the optimal linear pool, and for $\rho > 0$ the pooled error is bounded below by $\rho\sigma^2$ for every K. Adding agents removes only the idiosyncratic component $(1-\rho)\sigma^2/K$.

Figure 1 plots (5) against its Monte Carlo counterpart. The floor $\rho\sigma^2$ is the central quantity of the paper: it is the part of the committee's error that reflects shared evidence, a shared model of the market, or a shared provider, and it is invisible in the cross-sectional dispersion of the estimates, which measures only $(1-\rho)\sigma^2$. With $\rho = 0.6$, the pooled error falls from $0.73\sigma^2$ with three agents to $0.63\sigma^2$ with twelve, and no committee size brings it below $0.60\sigma^2$.

Proposition 2 (General covariance, bias and optimal weights)**.** *Under Assumption 1 with $b = 0$, the minimum-variance pool is $w^* = \Sigma^{-1}\mathbf{1}/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$ with*

$$\mathrm{MSE}^* = \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \;\le\; \mathrm{MSE}_{\mathrm{eq}} = \frac{\mathbf{1}'\Sigma\mathbf{1}}{K^2}, \tag{6}$$

with equality if and only if every row of Σ has the same sum, that is $\Sigma\mathbf{1} \propto \mathbf{1}$. If $b \neq 0$ the minimum of (4) is attained at $w \propto (\Sigma + bb')^{-1}\mathbf{1}$; when the bias is common to all agents, $b = \beta\mathbf{1}$, it enters every pool as β^2 and no weighting removes it.

The equality condition explains why inverse-variance weights estimated on a validation sample, as in Section 7, can be expected to help only when agents differ in accuracy or in correlation

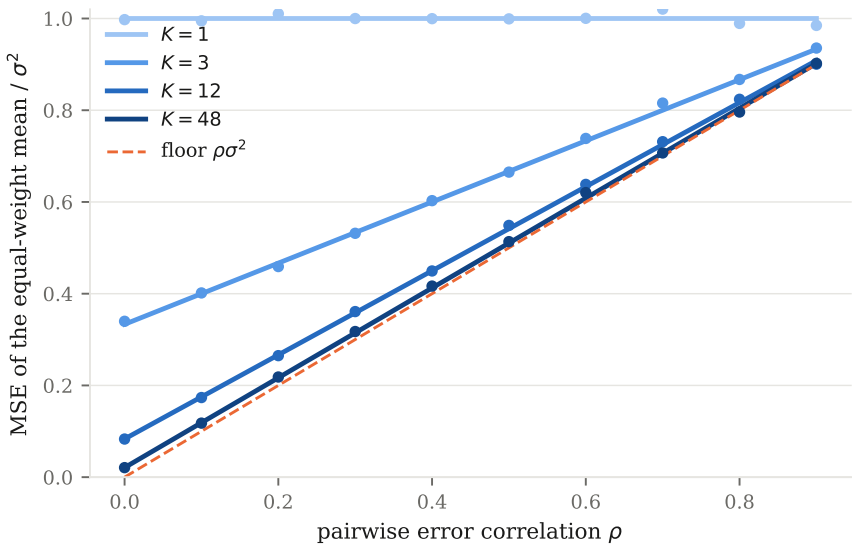


Figure 1. Mean-squared error of the equal-weight mean of K exchangeable signals as a function of the pairwise error correlation ρ (lines: equation (5); dots: Monte Carlo, 20,000 draws; $\sigma^2 = 1$). The dashed line is the floor $\rho\sigma^2$, which no committee size removes. Source: `scripts/41_theory_sim.py`.

structure; with interchangeable agents they add estimation noise without any gain, the forecast-combination puzzle documented by Clemen (1989) and Timmermann (2006). Trimmed means and medians (Jose et al., 2014) trade a small loss under Gaussian errors for protection against the heavy-tailed hallucinations that language-model committees produce; the paper reports them but the theory below is stated for the mean.

Remark 1 (Anchoring on a public signal). Suppose every agent shrinks towards a public value $a = v + \xi$ with weight β, so that $s_k = (1-\beta)(v+\eta_k) + \beta a$, with ξ independent of η and of variance σ_ξ^2. The equal-weight pool has error $(1-\beta)\bar{\eta} + \beta\xi$ and mean-squared error $f(\beta) = (1-\beta)^2 \mathrm{MSE}_{\mathrm{eq}} + \beta^2\sigma_\xi^2$, minimised at $\beta^* = \mathrm{MSE}_{\mathrm{eq}}/(\mathrm{MSE}_{\mathrm{eq}} + \sigma_\xi^2)$ and worse than no anchoring if and only if $\beta > 2\beta^*$. Anchoring on the assessment-roll value is therefore not a pathology in itself; it is harmful when the weight placed on a public signal exceeds twice its optimal weight, the over-reaction to public information analysed by Morris and Shin (2002). The same logic applies to the committee median, which becomes a public signal in every deliberation round.

The exchangeable case hides where the correlation comes from. A nested structure makes the sources explicit and shows which of them the committee's composition can address.

Proposition 3 (Common-model dependence). *Let* $\eta_k = c + g_{m(k)} + e_k$, *where* $c \sim N(0, \sigma_c^2)$ *is common to all agents,* $g_m \sim N(0, \sigma_g^2)$ *is common to the* n_m *agents served by provider* $m \in \{1, \ldots, M\}$,

$e_k \sim N(0, \sigma_e^2)$ *is idiosyncratic, and all components are independent. Then*

$$\mathrm{MSE}_{\mathrm{eq}} = \sigma_c^2 + \sigma_g^2 \sum_{m=1}^{M} \left(\frac{n_m}{K}\right)^2 + \frac{\sigma_e^2}{K}, \tag{7}$$

which for a balanced committee ($n_m = K/M$) equals $\sigma_c^2 + \sigma_g^2/M + \sigma_e^2/K$. At fixed K the provider term is minimised by the balanced assignment, and σ_c^2 is unaffected by both K and M.

Equation (7) is the formal content of the concern about algorithmic monoculture raised by Kleinberg and Raghavan (2021) and Bommasani et al. (2022): a committee drawn from one provider cannot average away that provider's shared error, and a market in which every lender uses the same valuation model inherits $\sigma_c^2 + \sigma_g^2$ as a common shock. It also motivates the provider-balanced committees of the design, which hold K fixed and vary M.

3.3. Deliberation

The protocol of Section 7 has two ingredients. In every round each agent sees the anonymised table of the other estimates and the committee median, and may revise towards them; this is opinion exchange. Each agent also sees the identifiers of evidence cited by others that it had not been shown, within its permitted view; this is evidence exchange. The two have opposite implications for accuracy and identical implications for dispersion, which is why dispersion is not a measure of success. The machine-learning literature on multi-agent debate reports accuracy gains from rounds of mutual critique (Du et al., 2023); the model specifies the conditions under which such gains can be genuine rather than an artefact of convergence.

Definition 5 (Opinion-only revision)**.** Let $x^{(0)} = s$ and, for $r \geq 0$,

$$x^{(r+1)} = (1-\lambda)\, x^{(r)} + \lambda\, T\big(x^{(r)}\big)\mathbf{1}, \qquad \lambda \in (0,1], \tag{8}$$

where the committee statistic $T : \mathbb{R}^K \to \mathbb{R}$ is translation- and positively scale-equivariant, $T(ax + b\mathbf{1}) = aT(x) + b$ for $a > 0$. The mean, the median and every trimmed mean satisfy this condition.

Proposition 4 (Consensus without information)**.** *Under Assumption 1 with $b = 0$ and the revision (8):*

(i) The committee statistic is invariant, $T(x^{(r)}) = T(x^{(0)})$ for all r, and

$$x^{(r)} = (1-\lambda)^r x^{(0)} + \left[1 - (1-\lambda)^r\right] T(x^{(0)})\, \mathbf{1}. \tag{9}$$

Consequently $x_k^{(r)} - x_l^{(r)} = (1-\lambda)^r (x_k^{(0)} - x_l^{(0)})$ for every pair, and every dispersion statistic that is translation invariant and homogeneous of degree one (standard deviation, interquartile

range, range) equals $(1-\lambda)^r$ times its round-0 value, whatever Σ is.

(ii) If T is the mean, $\bar{x}^{(r)} = \bar{x}^{(0)}$ for all r: the equal-weight pool and its mean-squared error are unchanged by any number of rounds.

(iii) If T is the median and Σ is exchangeable with $0 \le \rho < 1$, then $\mathbb{E}[\bar{x}^{(r)}] = v$, the median-aggregated estimate is invariant, $\mathrm{med}(x^{(r)}) = \mathrm{med}(x^{(0)})$, and

$$\mathrm{MSE}\left(\bar{x}^{(r)}\right) = \mathrm{MSE}\left(\bar{x}^{(0)}\right) + \left[1-(1-\lambda)^r\right]^2 \Delta_K, \qquad \Delta_K = \mathrm{Var}\left(\mathrm{med}(e)\right) - \mathrm{Var}(\bar{e}) \ge 0, \quad (10)$$

where e is a vector of K independent $N(0,(1-\rho)\sigma^2)$ variables. The pooled error is non-decreasing in r and converges to that of the round-0 median.

Part (i) is the mechanism identified experimentally by Lorenz et al. (2011): social influence narrows the range of estimates at a geometric rate while leaving the location of the group estimate where it was. Part (ii) says that when agents anchor on the mean, the pooled estimate is literally the round-0 mean. Part (iii) says that when they anchor on the median, as in our protocol, the pooled mean drifts towards the round-0 median and, under Gaussian errors, loses accuracy by at most Δ_K, which is modest (for $K = 12$, $\Delta_K \approx 0.41\,(1-\rho)\sigma^2/K$, the excess variance of the median over the mean) but of the wrong sign. In no case does opinion exchange reduce the pooled error. The dispersion-based stopping rule of the protocol and any confidence measure derived from the agreement of the agents are therefore uninformative about accuracy; they measure λ and r.

The revision (8) with the mean is a special case of the linear model of DeGroot (1974), $x^{(r)} = W^r x^{(0)}$ with W row-stochastic. Allowing heterogeneous weights and incomplete attention changes the limit but not the conclusion.

Proposition 5 (Linear deliberation conserves at most the round-0 information)**.** *Let $x^{(r)} = W^r x^{(0)}$ with W row-stochastic and $b = 0$.*

(i) For every r and every pool w with $w'\mathbf{1} = 1$ there is a round-0 pool $u = (W^r)'w$ with $u'\mathbf{1} = 1$ and $w'x^{(r)} = u'x^{(0)}$. Hence $\min_w \mathrm{MSE}(w'x^{(r)}) \ge \mathrm{MSE}^*$ for all r, with equality whenever W^r is non-singular.

(ii) If W is primitive, $x^{(r)} \to (\pi' x^{(0)})\mathbf{1}$ where π is the left Perron vector of W, and the limiting mean-squared error of every pool is $\pi'\Sigma\pi \ge \mathrm{MSE}^*$, with equality if and only if $\pi = w^*$.

(iii) For the anonymised-table protocol with a common weight λ and the mean as anchor, $W = (1-\lambda)I + \lambda\mathbf{1}\mathbf{1}'/K$, $\pi = \mathbf{1}/K$, and the consensus is the round-0 equal-weight mean.

Part (ii) is the wisdom-of-crowds condition of Golub and Jackson (2010): consensus is accurate only if influence is spread in proportion to precision, a knife-edge that an anonymised table does not implement. The experimental evidence of Becker et al. (2017) that decentralised networks can improve group estimates is consistent with the proposition: improvement there comes from a π that happens to weight accurate individuals more, not from communication as such.

Evidence exchange is different in kind because it changes the error of each agent rather than mixing errors across agents. We model it as an additive revision and then specialise to a model in which the revision is generated by evidence.

Lemma 2 (Additive revisions)**.** *Let the estimates x with error η be revised to $x+\zeta$, where ζ has mean zero and covariance Γ. For any pool w,*

$$\mathrm{MSE}\left(w'(x+\zeta)\right) - \mathrm{MSE}(w'x) = w'\Gamma w + 2\,w'\,\mathrm{Cov}(\zeta,\eta)\,w. \tag{11}$$

If ζ is uncorrelated with η, every pool is weakly worse after the revision, strictly if Γ is positive definite. A revision lowers the pooled error only if it is negatively correlated with the existing errors, which is what evidence does and what persuasion by narrative does not.

Assumption 2 (Evidence model)**.** There are J evidence items (comparable sales, roll data, market indices), $z_j = v + c + u_j$, with $c \sim N(0,\sigma_c^2)$ common to all items and $u_j \sim N(0,\sigma_u^2)$ independent across items. Agent k is shown the view $A_k \subseteq \{1,\dots,J\}$ and reports

$$x_k^{(0)} = v + a_k'(c\mathbf{1}+u) + g_{m(k)} + e_k, \qquad a_{kj} = \frac{\mathbf{1}\{j\in A_k\}}{|A_k|}, \tag{12}$$

where the uniform within-view weights are the minimum-variance weights given the view, g_m and e_k are the provider-common and idiosyncratic reasoning errors of Proposition 3 with $\mathrm{Var}(e_k)=\tau_k^2$, and all components are independent. An evidence exchange replaces the views by $A_k' \supseteq A_k$ and the weights accordingly, leaving c, u, g and e unchanged. Let $\omega = K^{-1}\sum_k a_k$ and ω' denote the implied evidence weights before and after the exchange.

Proposition 6 (Evidence exchange)**.** *Under Assumption 2 the equal-weight pool has*

$$\mathrm{MSE}_{\mathrm{eq}} = \sigma_c^2 + \sigma_u^2\,\|\omega\|^2 + \sigma_g^2 \sum_m \left(\frac{n_m}{K}\right)^2 + \frac{1}{K^2}\sum_k \tau_k^2, \tag{13}$$

and an evidence exchange changes it by exactly $\sigma_u^2(\|\omega'\|^2 - \|\omega\|^2)$. In particular:

- *(i) exchange of items in the common view $\bigcap_k A_k$, and more generally any exchange that leaves ω unchanged, has no effect; the common evidence error σ_c^2 and the reasoning terms are never affected;*
- *(ii) $\|\omega\|^2 \geq 1/|\bigcup_k A_k|$ with equality if and only if ω is uniform on the union of the views; the full exchange $A_k' = \bigcup_l A_l$ attains this minimum, so that the committee's evidence error equals that of a single agent shown the union;*
- *(iii) a partial exchange can raise the pooled error: with $J=2$, $K=3$, $A_1=\{1\}$, $A_2=A_3=\{2\}$, giving item 2 to agent 1 moves $\sigma_u^{-2}\mathrm{MSE}_{\mathrm{eq}}$ from $5/9$ to $13/18$.*

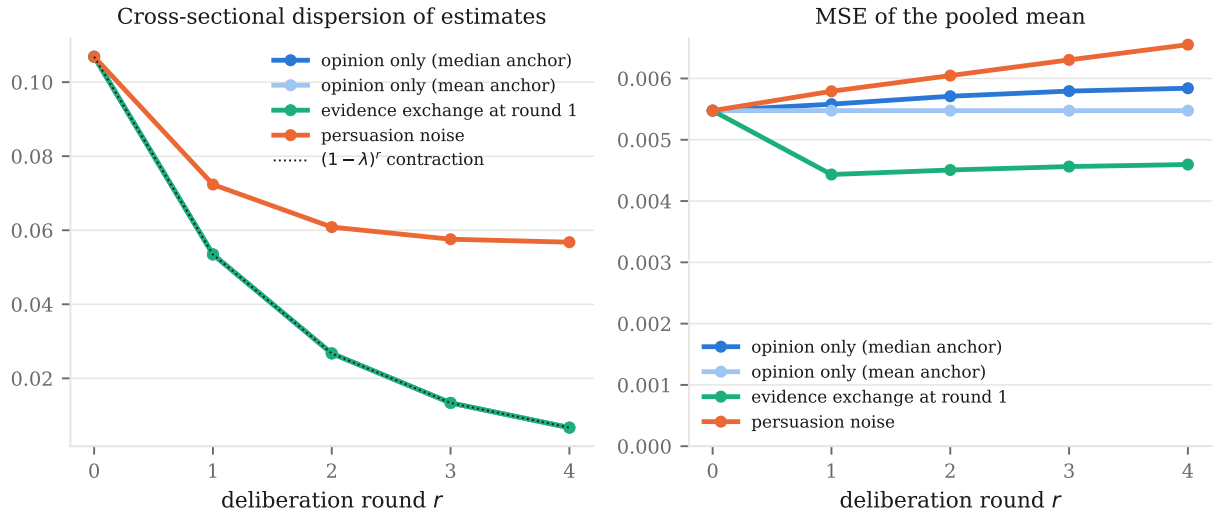


Figure 2. Four rounds of deliberation with $\lambda = 0.5$ for $K = 12$ agents in the evidence model (four comparables shared by all agents and two private ones per agent, $J = 28$, $\sigma_c = 0.05$, $\sigma_u = 0.12$, $\sigma_g = 0.04$, $\tau = 0.10$; 40,000 draws). Left: cross-sectional standard deviation of the estimates, which contracts by $(1-\lambda)$ per round under every mechanism (dotted line). Right: mean-squared error of the pooled mean; opinion-only revision leaves it unchanged (mean anchor) or raises it slightly (median anchor), evidence exchange at round 1 lowers it, and a zero-mean persuasion term raises it. Source: `scripts/41_theory_sim.py`.

The proposition identifies precisely when deliberation can add information. Exchange helps when it moves weight towards evidence that the committee under-uses, which happens when views are specialised and their union is rich; it is neutral when the cited items are already in everyone's dossier, which is the case when all roles work from the same comparable pool; and it hurts when it concentrates attention on items that are already popular, the herding on shared comparables that an information cascade produces (Bikhchandani et al., 1992). Part (ii) also bounds what communication can achieve: after a full exchange the committee's evidence error equals that of one agent with the union view, so any residual advantage over the evidence-matched single agent of arm E must come from the averaging of reasoning errors, $\sigma_g^2 \sum_m (n_m/K)^2 + K^{-2} \sum_k \tau_k^2$ against $\sigma_g^2 + \tau^2$, which arm F obtains without any communication. Figure 2 illustrates Propositions 4 and 6: every variant contracts dispersion at the same rate, only the exchange variant lowers the pooled error, and a persuasion term that is independent of the errors raises it, as Lemma 2 requires.

Deliberation also damages the aggregator's ability to exploit heterogeneity among agents, because mixing makes the agents' errors more alike and hides the precision differences that inverse-variance weights are meant to exploit.

Proposition 7 (Over-shrinkage of the variance-weighted pool)**.** *Let round-0 errors be independent with variances $\sigma_1^2, \ldots, \sigma_K^2$ not all equal, and let agents revise towards the mean with a common $\lambda \in (0,1)$. Let the aggregator apply inverse-variance weights $w_k \propto 1/\operatorname{Var}(x_k^{(r)})$ to the round-r*

estimates. At round 0 this pool is optimal, with $\mathrm{MSE}^* = (\sum_k \sigma_k^{-2})^{-1}$. *For every* $r \geq 1$ *it is strictly dominated by the round-0 pool, and as* $r \to \infty$ *its error converges to* $\mathrm{MSE}_{\mathrm{eq}} = K^{-2} \sum_k \sigma_k^2 > \mathrm{MSE}^*$.

The practical implication is that aggregation weights estimated on validation data should be applied to round-0 estimates, or at least that the comparison between round-0 and round-r pools must hold the weighting rule fixed; otherwise part of what looks like a deliberation effect is a change in the effective weights.

3.4. Placebo roles and what the role contrast identifies

Arm I keeps the mechanics of arm G and permutes the role prompts across agents. What that contrast identifies depends on whether the evidence view travels with the prompt.

Proposition 8 (Placebo roles)**.** *Let the committee be as in Assumption 2 and let the role prompt* π_k *affect the signal of slot* k *only through the reasoning-error variance,* $\tau_k^2 = \tau^2(\pi_k)$ *(additive separability of prompt and view). Let* σ *be a permutation of* $\{1, \ldots, K\}$.

(i) If prompts and views are permuted together, so that slot k *receives* $(\pi_{\sigma(k)}, A_{\sigma(k)})$*, the vector of signals is a relabelling of the original and every permutation-invariant aggregator (mean, median, trimmed mean) has exactly the same error distribution as in arm G.*

(ii) If prompts alone are permuted and views stay with the slots, the equal-weight pool has the same mean-squared error as in arm G for every σ*, because* $\sum_k \tau^2(\pi_{\sigma(k)}) = \sum_k \tau^2(\pi_k)$ *in (13).*

(iii) If the reasoning-error variance depends on the match between prompt and view, $\tau_k^2 = \tau^2(\pi_k, A_k)$*, the expected difference between arms I and G is* $K^{-2} \sum_k [\tau^2(\pi_{\sigma(k)}, A_k) - \tau^2(\pi_k, A_k)]$*, which is non-zero only through this complementarity.*

Under (i) the placebo is vacuous: arm I is arm G with relabelled agents and the contrast can only measure run-to-run noise, which is useful as a check but tests no hypothesis. Under (ii) and (iii) the contrast does not measure the value of role labels, nor the value of the evidence partition, which is held fixed; it measures the complementarity between a prompt and the evidence it is paired with. The value of the partition itself is identified by comparing committees that differ in their views with prompts held fixed, and the proposition shows that under separability a committee of role-prompted agents and a committee of generically prompted agents differ only through the average reasoning variance $K^{-1} \sum_k \tau^2(\pi_k)$, which is a prompt-quality effect and not a diversity effect in the sense of Hong and Page (2004). The preregistration therefore specifies arm I as case (ii), views fixed to the slot and prompts permuted, and the paper reports the view-homogeneous committee as a secondary contrast.

Table 2. Mapping between the theory and the preregistered hypotheses.

Hypothesis	Contrast	Theoretical content
H1	G2 vs D	No prediction on the level; Propositions 1 and 3 imply that the committee cannot beat an AVM whose error lies below the floor $\sigma_c^2 + \sigma_g^2/M$.
H2	G2 vs F	Propositions 4–6: a gain requires evidence exchange with $\|\omega'\|^2 < \|\omega\|^2$; opinion exchange alone predicts no gain and a small loss with median anchoring.
H3	G2 vs J	Equation (13): the committee beats the compute-matched single agent only through provider diversification $\sigma_g^2(1 - \sum_m (n_m/K)^2)$ and the exchange term, net of the evidence-concentration loss $\sigma_u^2(\|\omega\|^2 - 1/J)$.
H4	G2 vs I	Proposition 8: zero under separability; a non-zero contrast identifies prompt–evidence complementarity.
H5	coverage	Proposition 4(i): agreement-based intervals contract by $(1-\lambda)^r$ while the pooled error does not; coverage falls with rounds unless agents' own quantiles are insensitive to the table; linear pools of quantiles require recalibration in any case (Ranjan and Gneiting, 2010).
H6	faithfulness	Lemma 2: cited items that are not in the packet are persuasion, not evidence, and raise the pooled error; H6 is the precondition for the mechanism behind H2.

3.5. Testable implications and the preregistered hypotheses

The propositions translate into predictions that are matched one-to-one with the preregistered hypotheses of Section 7; Table 2 summarises the mapping.

Under either mechanism the dispersion of the estimates falls with rounds at the rate $(1-\lambda)$, so a finding that the committee converges, that its interquartile range narrows, or that the agents report higher confidence after deliberation is not evidence for H2; this is the content of Proposition 4(i) and the reason the protocol records but never selects on convergence. The pooled error falls with rounds only through evidence exchange, and Proposition 6 predicts that the gain is largest when roles have specialised views with a rich union and smallest when every role works from the same comparable pool; the paper tests this by measuring, for each property, the share of cited identifiers that lie outside the common view, and by interacting it with the G2–F contrast. The compute channel is isolated by arm J: a single generalist sampled repeatedly draws its e_k independently but shares one provider and one view, so that (13) assigns it $\sigma_c^2 + \sigma_u^2/J + \sigma_g^2 + \tau^2/K$, and the G2–J contrast is predicted to be positive if and only if provider diversification and exchange outweigh the partition's concentration of evidence weight. The communication channel is isolated by arm F, whose estimates are the round-0 vector of arm G; by Proposition 5(i), no pool of round-r estimates can beat the

best pool of round-0 estimates unless evidence is exchanged, so a G2–F gain is attributable to the exchange term and to nothing else in the model. Finally, Proposition 7 prescribes that the stacking and inverse-variance weights be estimated and applied at the same round, and that the primary comparison use a weighting rule that is invariant to deliberation, the mean or the median.

3.6. Identifying fundamental value from repeat sales, rents and user cost

Lemma 1 shows that price-based accuracy is silent about V^F. Three proxies allow partial triangulation, each with its own selection problem. First, a subsequent resale of the same property at $t_2 > t_1$ identifies the change in value net of quality, $p_{t_2} - p_{t_1} = (v^*_{t_2} - v^*_{t_1}) + (\varepsilon_{t_2} - \varepsilon_{t_1})$, once the intervening market movement is removed by a repeat-sales index in the manner of Case and Shiller (1989); an appraisal at t_1 that is closer to the market-adjusted resale than the t_1 price was is evidence that it captured a component of v^* that the first transaction's noise obscured. The selection problem is that properties that resell quickly are not a random sample: flips, forced sales, renovations and the loss-averse listing behaviour documented by Genesove and Mayer (2001) all correlate the resale decision with ε_{t_1}, so that the resale-based error is informative only after conditioning on holding period and on observable changes to the property. Second, where a market rent R_t exists for a comparable unit, (1) delivers $V^F_t = R_t / u_t$ up to the user cost. The selection problems are that the rental stock differs from the owner-occupied stock in quality and location, that Quebec's rent regulation detaches contract rents from market rents for sitting tenants, and above all that u_t contains the unobservable expected appreciation π^e_t, so that the price-to-rent diagnostic of Himmelberg et al. (2005) identifies V^F only under a maintained assumption about expectations; we report it as a bound under a range of π^e_t rather than as a point value. Third, independent professional appraisals, where legally accessible, are $\mathscr{I}_t$-measurable estimates like any other arm and are subject to the same lemma; their known tendency to confirm the contract price when one exists makes them a proxy for $\tilde{v}$ at best. A method that forecasts transaction prices well may therefore fail every one of these checks, and the converse is also possible; the paper keeps the two conclusions separate.

4. Institutional setting: valuation and transaction records in Quebec

Quebec offers an unusually clean environment for studying valuation because three public records describe the same dwelling from three different angles, each produced by a distinct institution with its own timing. The first is the land register (*Registre foncier du Québec*), in which every transfer of ownership is published by a notary. A registered deed records the parties, the consideration paid, the cadastral lot and the date of the act; it does not record the dwelling's attributes, the marketing history, or whether the transfer was at arm's length. The second is the municipal assessment roll, prepared under the *Loi sur la fiscalité municipale* by municipal or regional assessors and

supervised by the Ministère des Affaires municipales et de l'Habitation (MAMH). The roll is triennial: each unit of assessment carries a land value and a building value that are meant to reflect market value at a reference date set eighteen months before the roll comes into force, together with the physical attributes used by the assessor (building type, construction year, floor area, lot area, number of storeys and dwellings, and the land-use code CUBF). The third is the brokerage record (the MLS/Centris system operated under the authority of the OACIQ), which contains listing prices, descriptions, photographs and marketing dates, but only for the roughly three quarters of transactions that go through a licensed broker; it is proprietary and is not used in the arms of this study.

Valuation itself is a regulated profession: chartered appraisers (évaluateurs agréés) are members of the Ordre des évaluateurs agréés du Québec and produce appraisal reports for lenders, courts and expropriations following the sales-comparison, cost and income approaches. The agents studied here are simulated analytic roles modelled on this division of labour; none of them is a chartered appraiser and nothing in the paper should be read as an appraisal within the meaning of Quebec law.

Two institutional facts shape the identification strategy. First, the assessment roll is public, stale by construction (its reference date precedes the sale by eighteen months to four and a half years), and known to every market participant: it is a natural, information-set-consistent anchor that any valuation method, human or artificial, legitimately observes. We therefore give it to every arm, together with its reference date, and we treat the price-to-assessment ratio of recent nearby sales as the canonical "local market state" signal. Second, registration in the land register lags the signature of the deed by a few days, and publication in the data platform we use lags registration by two to eight weeks (measured in a companion paper, Boucher, 2026); a valuation dated at the deed therefore cannot use sales registered after it, and the comparable pool of every arm is truncated at the valuation date on the *sale* date, not the publication date, which is the conservative choice.

The sample period, January 2021 to October 2026, spans a complete cycle: the pandemic boom of 2021–2022, the correction that followed the rise of policy rates in 2022–2023, and the recovery of 2024–2026. This matters for the question at hand because a valuation method that merely extrapolates recent comparables performs differently at turning points than one that reasons about the state of the market, and because distribution shift between the training window and the test window is a feature of the design rather than a nuisance.

5. Data

This section describes the transaction records, the construction of the analytic sample, the chronological and spatial partitions, the subsamples on which the language-model arms are evaluated, and the rules that define the information set $\mathscr{I}_t$ of Section 3 for every arm. The last two subsections

report the leakage audit that enforces those rules and the privacy safeguards applied before any record leaves the research environment. Every count below is written by a numbered script from the versioned data layer (`results/data/*.json`, `results/tables/*.csv`); none is transcribed by hand.

5.1. Sources

The primary source is the registered-sales table of Vrai Prix, a platform operated by the author that assembles residential transfers from the public Quebec land register and links each transfer to its unit of assessment on the municipal roll (Section 4). A registered deed contributes the consideration paid, the registration date and the cadastral lot; the roll contributes the physical attributes used by the assessor (floor area, land area, construction year, storeys, number of dwellings, land-use code), the land and building values in force at the sale and the roll's market reference date; the link is made on the roll unit identifier and verified with coordinates. We work from an immutable snapshot of the production database taken on 2026-10-21 and processed by the cleaning pipeline of a companion paper on municipal price indices, which geocodes the records, attaches the administrative geography of the Ministère des Ressources naturelles et des Forêts and the municipality repertory of the Ministère des Affaires municipales et de l'Habitation, deduplicates repeated registrations of the same transfer, and assigns an exclusion reason to records that fail its headline-sample criteria. The snapshot contains 769,995 transfers registered between 2021-01-04 and 2026-10-07; the publication lag of the register, measured in the companion paper at two to eight weeks, explains why the last registered sale precedes the snapshot date. The checksums of the raw files are pinned in the repository and the record-level data are not redistributed.

A second platform of the author, Immo-Ka, aggregates brokerage listings with asking prices, marketing dates and free-text descriptions. We audited it as a candidate source and excluded it from every arm, for three reasons that are documented in the data-access checklist of the repository. Its coverage begins on 2026-08-12, after the end of the test window, so that no listing could be matched to a test or validation sale; asking prices and days on market would be observable for some subjects and not for others and would therefore create an information asymmetry across properties and across arms that the design is meant to prevent; and its records contain broker contact details and third-party remarks that cannot be transmitted to external model providers. The audit found 479,265 listings first seen up to 2026-10-27, of which 195,503 are linked to a roll unit; they are reserved for a descriptive comparison of the prospective sales, registered after the protocol freeze, with their listing history, which is reported separately and plays no role in the evaluation.

Three external geospatial layers complete the evidence. The provincial flood maps of the Quebec government provide 24,997 flood-zone polygons and 1,597 polygons delimiting the area that the maps actually cover; a point outside the covered area has unknown exposure, which we

keep as a missing value rather than coding it as safe. A public-transit layer contributes 37,475 stops (bus, metro, light rail and commuter rail). The 2021 Census of Population contributes 13,805 dissemination areas with their population, density, median age, median household income and share of renter households (Statistics Canada, 2024). Each layer is a single snapshot and is time invariant by construction; it carries no transaction information and is given identically to every arm. Of the analytic sales, 39.9 percent fall inside a mapped flood-coverage area, 1.2 percent of those lie in a mapped flood zone, and 97.0 percent are matched to a dissemination area.

5.2. The analytic sample

The analytic sample is the subset of the snapshot that satisfies the companion paper's headline-sample criteria and the additional restrictions of this study. Of the 769,995 raw records, 137,989 are removed by the headline-sample flag, which excludes non-residential and vacant-land transfers, transfers that could not be matched to a roll unit, duplicate registrations, sales whose ratio of price to assessment lies outside the range consistent with an arm's-length transfer, dwellings built after the sale date, transfers of shared or multi-parcel units, and units whose land-use classification falls outside the three residential classes retained here, namely single-family houses, condominium units and small multi-unit buildings (plexes). The remaining 632,006 records all have coordinates, a roll identifier and one of the three classes. A further 140 records with prices below 50,000 or above 5,000,000 CAD are dropped, leaving 631,866 sales on 541,244 distinct roll units. The price band removes nominal transfers and the extreme upper tail where comparables are too sparse to support any method; its sensitivity is examined in the appendix.

Table 3 describes the sample by property class. Houses account for two thirds of the sales, condominiums for a fifth and plexes for the remainder. The median price is 390,000 CAD overall, from 350,000 for condominiums to 475,000 for plexes; the median roll value in force at the sale is 297,900 CAD, well below the median price, which reflects the lag between the roll's reference date and the transaction and the strong appreciation of the period. Floor and land areas, construction years and the share of sales preceded by an earlier registered sale of the same unit differ markedly across classes; condominiums are younger, smaller and resold more often, plexes older and larger. The share with a prior sale is 14.3 percent overall, bounded by the fact that only sales from 2021 onward can serve as the earlier transaction. Across years the sample contains between 83,489 (the partial year 2026) and 131,800 (2021) sales, and across the four region groups used for stratification it contains 274,294 sales in the Montréal census metropolitan area, 56,163 in the Québec metropolitan area, 36,280 in Outaouais and 265,129 elsewhere in the province.

Table 3. Descriptive statistics of the analytic sample by property class.

Class	n	Median price (CAD)	Floor area (m^2) mean	Floor area (m^2) median	Land area median (m^2)	Year built	Roll value median (CAD)	Prior sale (%)
House	422,542	397,000	130.9	113.7	652	1983	293,200	11.6
Condominium	122,976	350,000	90.1	89.6	104	2006	276,800	25.0
Plex	86,348	475,000	205.0	194.5	507	1959	383,300	12.3
All classes	631,866	390,000	133.0	111.5	544	1986	297,900	14.3

Notes. Registered sales of 2021-01-04 to 2026-10-07 in the analytic sample ($n = 631{,}866$). Floor area, land area and year built are the attributes of the assessment-roll unit. Roll value is the total assessment in force at the sale date. Prior sale is the share of sales preceded by a strictly earlier registered sale of the same roll unit within the sample window. Medians are unweighted. Source: `scripts/15_descriptives.py`, `results/tables/descriptives.csv`.

5.3. Variables and their as-of status

Table 4 lists the variables of the analytic table with the status that governs their use. The status answers one question: at the valuation date, which is the sale date, could an analyst have observed the value without knowledge of the sale? Roll attributes and roll values are public before the sale; static geography is time invariant; derived quantities are computed from strictly earlier sales; the price, its logarithm and the price tercile used for stratification are post-sale and are never shown to any arm. Two variables deserve comment. The total roll value is the assessment recorded by the provider at the time of the sale, that is, the roll in force, and not the value on the roll current at the snapshot date; because the snapshot does not retain the land and building split of earlier rolls, that split is taken from the current roll and rescaled to the assessment in force, which we label an as-of approximation. The roll's market reference date is inferred by stepping the current roll's reference date back by whole triennial cycles until the reference date plus eighteen months precedes the sale; the inference agrees with the recorded assessment in 98.6 percent of the rows, and the reference date is disclosed to the agents so that they treat the roll as a lagged rather than a current value, as the assessment manual prescribes (Ministère des Affaires municipales et de l'Habitation du Québec, 2024). Missingness is low: floor area is absent for 1.2 percent of sales and construction year for 1.0 percent; the borough is defined only in the six cities that have boroughs; the prior-sale fields are structurally missing for the 85.7 percent of sales without an earlier transaction in the window.

5.4. Chronological and spatial partitions

The partitions are closed date windows fixed in the configuration file before any model was fitted. Sales registered from 2021-01-01 to 2024-06-30 form the training partition (379,310 sales); 2024-07-01 to 2025-03-31 the validation partition (72,132), on which hyper-parameters, aggregation weights and every design choice are tuned; 2025-04-01 to 2025-12-31 the test partition (96,935), evaluated once after the protocol freeze; and 2026-01-01 to 2026-10-07 the holdout partition (83,489), an

Table 4. Variables of the analytic table and their as-of status.

Variable	Definition	As-of status	Missing (%)
`tx_id`, `parcel_id`	Transaction identifier; roll-unit identifier grouping repeat sales	key	0.0
`date`	Registration date of the deed; valuation date; cut-off is the previous day	anchor	0.0
`price`, `log_price`	Transaction price in CAD and its logarithm	target, never shown	0.0
`lat`, `lng`, `x_m`, `y_m`	Coordinates of the roll unit (WGS84; NAD83 Quebec Lambert metres)	static	0.0
`region_name`, `mun_code`, `borough_name`	Administrative region, municipality code, borough where defined	static	0.0 / 0.0 / 70.6
`region_group`, `urban_class`	Montréal CMA, Québec CMA, Outaouais, other; metro, suburban, rural by municipal population	static (derived)	0.0
`property_class`, `subtype`	House, condo, plex; detached, semi, row, integrated, unknown	static	0.0
`floor_area_m2`, `land_area_m2`	Living floor area and lot area from the roll	roll, public	1.2 / 0.1
`year_built`, `storeys`, `n_units`	Construction year, storeys, dwellings from the roll	roll, public	1.0 / 0.5 / 0.0
`new_construction`	Built in the sale year or the year before	roll, public	1.0
`roll_total_value`	Assessment in force at the sale, as recorded at the time of the sale	roll, public	0.0
`roll_land_value`, `roll_building_value`	Land and building split of the current roll rescaled to the assessment in force	as-of approximation	1.7 / 0.0
`roll_reference_date`	Market reference date of the roll in force, inferred from the triennial cycle	as-of inference	0.0
`prior_sale_price`, `prior_sale_date`	Most recent strictly earlier sale of the same roll unit in the window	strictly earlier sale	85.7
`grid1k`, `grid5k`	1 km and 5 km grid cells in the Lambert plane	static (derived)	0.0
`price_tercile`	Tercile of price within class; stratification only	post-sale, never a feature	0.0

Notes. Missingness measured on the 631,866 analytic rows. "Roll, public" denotes attributes of the municipal assessment roll, published before the sale. The external geospatial variables (flood exposure, transit stops within 1 km, dissemination-area statistics, distance to the nearest central business district) are static layers joined on coordinates; flood exposure is missing for the 60.1 percent of sales outside the mapped coverage. Source: `docs/data_dictionary.md`, `results/data/analytic_summary.json`.

Table 5. Sales by partition and region group.

Partition	Sale dates	Montréal CMA	Québec CMA	Outaouais	Other regions	Total
Train	2021-01-01 to 2024-06-30	162,988	34,348	22,439	159,535	379,310
Validation	2024-07-01 to 2025-03-31	32,062	5,889	4,334	29,847	72,132
Test	2025-04-01 to 2025-12-31	42,778	8,577	5,326	40,254	96,935
Holdout	2026-01-01 to 2026-10-07	36,466	7,349	4,181	35,493	83,489
Total	2021-01-04 to 2026-10-07	274,294	56,163	36,280	265,129	631,866

Notes. Partition windows are those of `configs/data.yaml`; sale dates are registration dates. Region groups: Montréal CMA (regions Montréal, Laval, Montérégie, Laurentides, Lanaudière within 45 km of the city centre), Québec CMA (Capitale-Nationale within 35 km of the city centre), Outaouais, and all other sales. The spatial-generalisation experiment withholds the whole of Outaouais, Abitibi-Témiscamingue and Bas-Saint-Laurent (59,770 sales) from training. Source: `scripts/15_descriptives.py`, `results/tables/splits_by_region.csv`.

untouched later year evaluated once at the very end. The windows are contiguous, every sale belongs to exactly one partition, and Table 5 gives the counts by partition and region group. A chronological split does not by itself prevent a roll unit from appearing on both sides of a boundary; the leakage audit therefore verifies that for each of the 45,747 units with a sale in the training partition and a later sale in a later partition, the training sale is the earlier one, so that no future transfer of a test property is ever in the training data. The validation window straddles the trough and the recovery of the 2024 market and the test and holdout windows cover the subsequent expansion, which exposes every arm to a regime it did not see in training.

The spatial experiment withholds three entire administrative regions from training, Outaouais, Abitibi-Témiscamingue and Bas-Saint-Laurent, 59,770 sales in total, and evaluates the arms on their test-window sales with models and comparables drawn from the rest of the province. Outaouais is of particular interest because it is a border market whose prices track Ottawa rather than Montréal; the two other regions are thin, rural markets where comparables are sparse and where Proposition 6 predicts the largest benefit from evidence exchange. A prospective partition is reserved for sales registered in the live database after the protocol freeze, whose identifiers are by construction absent from the snapshot; it is fetched by a dedicated script and evaluated last.

5.5. Subsamples for the language-model arms

The statistical and machine-learning baselines are evaluated on the full partitions. The language-model arms, whose cost per property runs from a few cents to more than a dollar, are evaluated on stratified random subsamples of 240 validation, 600 test and 300 holdout sales, drawn once with the seed recorded in the configuration. Strata are the cells of region group by property class by within-class price tercile, 36 cells of which 35 are non-empty in the test partition; the allocation is proportional to the stratum's share of the partition by the largest-remainder method, capped at the stratum size, and the draw within each stratum uses a generator seeded from the hash of the

seed, the partition and the stratum name, so that the file can be regenerated exactly and the audit confirms that it is. Because the subsample is ordered deterministically and the runner takes the first n properties, any run of size n is nested in every run of larger size; Section 7 uses this property to extend the evaluated samples without redrawing. The baselines are also scored on these subsamples so that paired comparisons are available on identical properties.

5.6. Information-set rules

The information set $\mathscr{I}_t$ of a subject sold on date t is defined operationally by six rules that hold for every arm, statistical or agentic. The information cutoff is $t-1$: the sale itself is not yet observed. Comparable sales are drawn from sales registered strictly before the cutoff, within 540 days, of the same property class and of other roll units; the retrieval rule is described in Section 6 and is identical for all arms, so that the comparable-selection opportunity is equalised. The subject's own earlier sale, when one exists strictly before the cutoff, is disclosed to every arm with its date as repeat-sales information. The roll values are those in force at the sale and are accompanied by the reference date of that roll. The market context, consisting of municipal and regional price levels, a monthly municipal index, liquidity and the local ratio of price to assessment, is computed from sales dated on or before the cutoff only, with the cutoff month flagged as partial; the platform's own published price index and the automated estimates it displays are computed from the full history and are excluded from every arm. The baselines receive the same information through the feature matrix: every neighbourhood and municipal feature is a function of sales dated strictly before the sale date, computed with left-closed rolling windows and monthly spatial indices, and no model is fitted outside the training partition. External layers are static and shared. Nothing in the pipeline distinguishes the arms by what they may know, only by how they process it.

5.7. Leakage audit

An audit script re-derives the partitions and features and runs nine checks, each of which must pass for the pipeline to proceed; all nine pass on the real data and are also exercised on a labelled synthetic dataset in the test suite. The first two verify that the partition windows are contiguous and non-overlapping and that every sale is assigned to the window containing its date, with no duplicate identifiers. The third is the group-ordering check described above. The fourth recomputes the prior-sale fields by brute force on 4,255 roll units and verifies that every disclosed prior sale is strictly earlier than the subject's sale and matches the stored value. The fifth draws 300 subjects from the subsamples, retrieves their comparables and verifies that none is dated on or after the cutoff or beyond the look-back window, that none belongs to the subject's own roll unit or to another class, that identifiers follow score order and that the count never exceeds twelve; on these subjects the

mean number of comparables is 11.2 and the minimum is three. The sixth verifies that the market context of forty subjects is unchanged when every sale after the cutoff is deleted. The seventh is the strongest test of the feature matrix: it multiplies the price of every sale dated on or after 2025-04-01 by seven, then deletes those sales altogether, rebuilds the matrix and checks that no feature of any earlier sale has changed in either case, which it does on more than fifteen thousand compared rows with zero changes. The eighth verifies that no identifier, date or price column is among the model inputs. The ninth regenerates the stratified subsamples and confirms their sizes, their strata and their exact reproducibility.

Two leaks were caught by this suite during development and are worth recording because they illustrate how as-of violations arise in otherwise careful code. The first concerned the neighbourhood features. An early implementation obtained the K nearest same-class sales of each subject from a single spatial index built on the whole table and only then filtered them by date. In dense urban cells the K nearest sales were crowded with transactions registered after the subject's sale; after the date filter, the number of surviving neighbours, and hence the count and the medians, depended on how many later sales had occurred nearby, which is post-sale information about local turnover. The perturbation check detected it because deleting the later sales changed the features of earlier ones. The current implementation builds one spatial index per class and calendar month of candidate sales and queries, for a subject sold in month m, only the indices of months $m-12$ to m, applying exact date masks afterwards, and is verified against a brute-force computation. The second concerned the municipal rolling medians. A rolling join on the sale date with the default right-closed window included the subject's own transaction and all sales registered the same day in the window from which its municipal median was computed, so that the target entered its own feature through the median. The fix is a left-closed window, under which same-day and later sales are excluded by construction; the brute-force test now asserts the exclusion. Neither leak would have been visible in a conventional chronological evaluation, because both features were legitimately computed from past data for most sales and were contaminated only at the margin; both would have flattered the baselines and biased the central contrast of the paper against the agentic arms.

5.8. Privacy

The transaction records contain civic addresses and the roll-unit identifier, and the ancillary platform data contain names and remarks of third parties. No address, identifier, name or free text is ever transmitted to a model provider. The evidence packet of Section 6 carries only the attributes listed in Table 4, the municipality and borough names, and coordinates rounded to three decimals, roughly one hundred metres, which does not identify a dwelling in any built-up area; the packet builder enforces a deny list of forbidden keys and an assertion that the subject's price is absent, and the audit hashes every packet so that the transcripts can be matched to what was actually sent. The

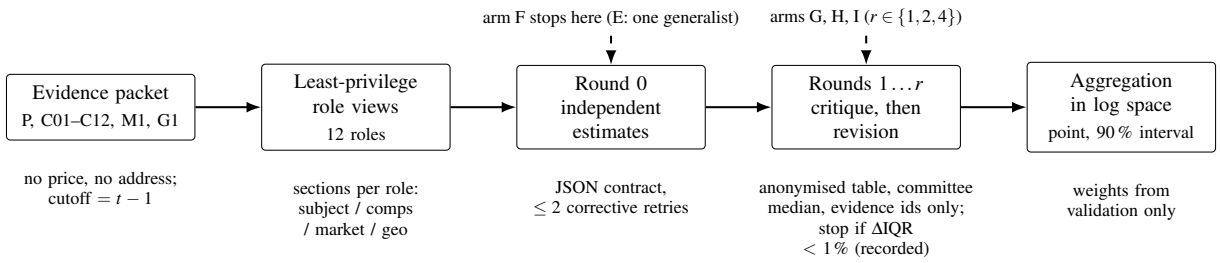


Figure 3. Flow of one property through the agentic arms.
Notes. The evidence packet is built once per property under the information-set rules of Section 5.6 and is identical for all arms. Each role receives only the sections of its view. Arm F stops after round 0; arms G, H and I continue through the pre-specified number of critique-and-revision rounds. Aggregation applies the rules of Section 6.6 to the final estimates of the roles. Source: `figures/fig_architecture.tex`.

transaction identifier that labels a packet is an opaque platform key with no external meaning. Model providers receive no information that would allow them to recover the sale from the land register other than what is public there already; the research use of the platform data is authorised by its owner, no human subjects are involved, and record-level data are not redistributed with the paper. The repository ships a clearly labelled synthetic dataset with the same schema so that the entire pipeline can be run without the proprietary records.

6. Multi-agent architecture

The architecture is designed to make the theory of Section 3 testable rather than to produce a convincing valuation dialogue. Every design choice serves one of three purposes: to fix the information set so that the arms differ only in how they process evidence, to partition that evidence across roles in a way that can be removed or permuted, and to record every message so that the audit of Section 7 can trace each number in the paper to a stored artefact. Figure 3 gives the overall flow; the subsections describe the evidence packet, the roles and their views, the output contract, the provider layer, the deliberation protocols, the aggregation rules and the audit trail.

6.1. The evidence packet

For each subject property the pipeline assembles one evidence packet containing four sections, each identified by an evidence code that agents must cite. The subject section, coded P, carries the valuation date, the information cutoff, the property class and subtype, the municipality and borough, the rounded coordinates, the roll attributes (floor area, land area, year built, storeys, number of dwellings, new-construction flag), the land, building and total assessment in force at the sale with the roll's reference date, the prior sale of the same unit when one exists strictly before the cutoff, and an explicit list of the fields that are missing. The subject's price never enters the packet; the

builder asserts its absence on every packet.

The comparables section, coded C01 to C12, is produced by a deterministic retrieval rule that is the same for every arm and is therefore not a source of difference between them. Candidates are sales of the same property class on other roll units, registered strictly before the cutoff and within 540 days of it. The search radius expands from one to three to ten kilometres and stops at the first radius that yields at least three candidates; if ten kilometres still yields fewer than three, the search is extended to 25 kilometres and the packet is annotated accordingly, which occurred for one of the 1,140 subsample subjects. Candidates are ranked by a similarity score, lower being better, that adds the distance divided by the retained radius, one and a half times the absolute difference in log floor area, half the absolute difference in construction year divided by fifty, and half a recency term proportional to the age of the sale within the look-back window; the twelve best receive the codes in score order. Each comparable carries its sale date, price, attributes, distance to the subject, and its ratio of price to assessment. The retrieval is intentionally simple and transparent; a role that disagrees with it can discard comparables and must say why.

The market section, coded M1, summarises the local market from sales dated on or before the cutoff: the municipal median price, median price per square metre and sales count for the same class over the preceding twelve months; a monthly municipal index of median price per square metre, with base 100 at the first month with at least five sales and with the cutoff month flagged as partial, falling back to the regional series when the municipality has fewer than twenty sales in twelve months; the regional median and its year-over-year change; the number of sales of all classes within three kilometres as a liquidity measure; and the median ratio of price to assessment within three kilometres. A mortgage-rate field exists in the schema and is null in this snapshot, which the agents are told. The geographic section, coded G1, carries the flood-zone status where mapped, the number of transit stops within one kilometre, the population density and median household income of the dissemination area, and the distance to the nearest of six central business districts, with notes stating when flood exposure is unmapped. The packet closes with notes about the retrieval radius, the number of candidates and any extension.

6.2. Roles and least-privilege views

The default committee comprises twelve roles listed in Table 6. Each role is a simulated analytic role defined by a system prompt and a view, the subset of packet sections it may see. The agents are not licensed Quebec appraisers, lenders, brokers or inspectors, and their prompts say so; nothing they produce is a professional opinion of value. The views implement the evidence partition A_k of Assumption 2: the building-condition specialist sees the subject alone and is instructed to state that no inspection has been performed and that condition is unobserved; the market economist sees the subject and the market context but no comparables; the comparables specialist sees the

Table 6. The twelve default roles, their evidence views and mandates.

Role identifier	View	Mandate (one line)
`appraisal_method`	P, C, M	Reconcile the sales-comparison, cost and income approaches and derive the point from the reconciliation
`comparables`	P, C	Rank, discard and adjust comparables; one traceable adjustment per retained comparable
`hedonic`	P, C, M	Treat the comparables as a local sample and reason from implicit prices of attributes and time
`market_economist`	P, M	Convert the roll value to the valuation date from price levels, index path, liquidity and assessment ratio
`building_condition`	P	State that condition is unobserved; reason only from vintage, size and the land–building split
`geospatial_risk`	P, G, M	Quantify flood exposure, transit access, density, income and distance to the centre against local levels
`underwriting`	P, C, M, G	Prudent collateral value; downside stress in the lower quantile and in the counterarguments
`buyer_wtp`	P, C, M, G	Price at which the marginal household is indifferent between the subject and its best substitute
`seller_strategy`	P, C, M	Expected closing price after normal marketing time, with the assumed negotiation gap
`rental_usercost`	P, C, M	Income-multiplier or user-cost reasoning with yield assumptions stated as assumptions
`data_auditor`	P, C, M, G	Report inconsistencies and apparent post-sale information; estimate from the reliable evidence only
`skeptic`	P, C, M, G	Find the strongest reasons the committee could be wrong; propose a value only when justified

Notes. P, C, M and G denote the subject, comparables, market and geographic sections of the evidence packet. Identifiers and views are those of `src/agentic_valuation/agents/roles.py`; the mandates paraphrase the role-specific prompts, whose SHA-256 hashes are part of the protocol freeze. All roles are simulated analytic roles, not licensed professionals.

subject and the comparables but no market context; the geospatial specialist sees the subject, the geographic section and the market context; four roles, among them the data auditor and the skeptic, see the whole packet. The union of the views is the whole packet, and the overlap between views is substantial, as it is in any committee of practitioners; the partition is what the full-view variant of arm G removes and what the placebo of arm I preserves.

Eight optional roles extend the committee to twenty for the robustness analysis: a construction-cost estimator, an urban-planning analyst, a macro and mortgage-rate analyst, an energy-efficiency analyst, a climate-risk analyst, a regional market expert, an uncertainty auditor and a fairness auditor, each with its own view. A generalist role with the full packet serves arms E and J. The placebo committee of arm I consists of twelve members whose prompt contains no specialised mandate, only an instruction to apply general valuation judgement to the evidence visible, and whose views are, member for member, those of the twelve default roles; an alternative placebo permutes the twelve default prompts across the twelve default views with a seeded permutation. The system prompt of

every role is the concatenation of a shared preamble, which states the task, the information cutoff, the evidence discipline and the answer format, the role identifier, the role-specific text, and the output rules; it is constant across properties, and its hash is recorded.

6.3. The output contract

Every estimate turn must return one JSON object validating against a typed schema. The object contains a point estimate in CAD as of the valuation date; five quantiles at the 5th, 25th, 50th, 75th and 95th percentiles, required to be non-decreasing and to bracket the point; the list of evidence codes actually relied upon; a list of adjustments, each naming a comparable, the adjusted price after the stated adjustments and the reason; key assumptions; missing data; a confidence rationale; counterarguments; and a reasoning summary of at most 120 words. Critique turns return a set of critiques, each naming a target role, an agreement flag, the evidence-based points, the codes cited and an optional suggested value. The runtime validates the object against the schema and the ordering constraints and, on failure, returns the validation error to the model for at most two corrective retries; a turn that still fails is recorded with its error and a null estimate. Failed turns are excluded from the committee state and from aggregation and are never imputed, so that a committee of twelve may on occasion aggregate eleven estimates, and the transcript says so. The contract is what makes the faithfulness audit possible: cited codes can be checked against the packet, adjusted prices against raw comparable prices, and the point against the hull of the adjusted comparables.

6.4. Provider layer

Two providers are supported through their official software development kits, the Anthropic Messages API and the OpenAI Responses API, each with the vendor's structured-output facility used to constrain the response to the schema. Reasoning is left to the vendor's adaptive mode at a medium effort setting; no temperature, nucleus or other sampling parameter is sent, both because the current Anthropic models reject them and because the design relies on the vendor's default sampling to generate the variation between repeated calls that arm J needs. The two models of the main experiments are cost-matched: the mid tier pairs `claude-sonnet-5-5` with `gpt-6.1-sol`, both listed at two dollars per million input tokens and ten dollars per million output tokens on the vendors' pricing pages on 2026-10-27; the small tier pairs `claude-haiku-5-5` with `gpt-6-luna`, both at ten cents and fifty cents. Identical list prices mean that a committee with one provider and a committee with the other face the same token budget, so that a provider contrast is not a budget contrast. Three committee compositions are run: all-Anthropic, all-OpenAI, and a mixed committee in which the twelve roles are assigned to the two models by a balanced permutation seeded by the run seed and the property identifier, recorded in the transcript, so that no role is systematically tied

to a provider. A mock provider that parses the packet and returns schema-valid estimates allows the whole pipeline and its tests to run without keys.

The prompts are laid out for the vendors' prefix caching. The system prompt of a role is identical for every property of a run and carries the cache breakpoint; everything that varies, the packet view, the committee state, the verification flags and the task line, is in the user message. The twelve role prompts are therefore cached after the first property, and critique and revision turns of the same role within a property share the cached prefix. The ledger records, for every call, the model identifier, the prompt hash, the input, cached and output token counts, the estimated cost at list price, the latency, the request identifier and the number of validation retries; costs in the paper are estimates from this ledger, not invoices.

6.5. Deliberation protocols

Arm F runs the twelve roles concurrently on their views and stops; its estimates are the round-0 vector $x^{(0)}$ of the theory. Arm E is the same with a single generalist on the full packet. Arm G is the structured debate. Round 0 is identical to arm F. In each subsequent round every role first receives the committee state and returns a critique set, then receives the committee state together with the critiques addressed to it and returns a revised estimate. The committee state is an anonymised table of the current point estimates, 5th and 95th percentiles and reasoning summaries of the other members, identified by role and never by model or provider, together with the committee median of the current points and the list of evidence codes that other members cited and that lie outside the recipient's own view. Evidence exchange is thus by identifier only: a role learns that another member relied on the geographic section, never its content, and never anything about the price. The revision instruction states that an estimate may be kept unchanged and that moving toward the committee without an evidential reason is an error, which is the operational form of the distinction between the revision term ζ and the anchoring weight λ of Definition 5. The number of rounds is fixed in advance from the grid $\{1,2,4\}$. A stopping rule ends the debate early when the interquartile range of the log points changes by less than one percent between consecutive rounds; the event is recorded in the transcript with the round at which it occurred, and the rounds actually run are reported, but the rule is never used to select the best round, in keeping with Proposition 4, under which convergence is not evidence.

Arm H adds two elements to a two-round debate. The skeptic is removed from the rounds and blinded: it sees the packet and the committee's final estimates only, never the discussion, and its verdict is appended as a separate estimate. A deterministic verification tool runs on every estimate after each round and flags cited codes that do not exist in the packet, cited comparables that do not exist, adjusted prices more than forty percent away from the raw comparable price, and points outside the range of the adjusted comparables widened by a quarter; the flags are injected into

the next prompt of the flagged member, which must address them. Arm I keeps the mechanics of arm G, two rounds, the same views, the same committee state, and replaces the role prompts by the placebo prompts, so that the only difference between G and I is the specialised mandate. The full-view variant of arm G keeps the twelve default prompts and gives every role the whole packet, so that the only difference between G and G-fullview is the evidence partition. Arm J is the cost-matched single agent: the generalist is sampled twelve times independently on the same prompt, which matches the number of round-0 turns and, in the pilot token counts, the token budget of arm F, followed by a two-step chain in which the agent critiques its own first sample and then refines it in the manner of self-consistency and self-refinement (Wang et al., 2022); its outputs are the sample mean, the sample median and the refined estimate. Matching arm J to the token budget of the two-round debate rather than to arm F would require roughly seventy samples, a configuration the pilot did not run; the contrast reported in the paper is therefore the conservative one, G against a single agent given the tokens of F, and the secondary contrast F against J isolates ensembling of one model from ensembling of twelve roles at equal tokens.

6.6. Aggregation

All aggregation is done on the logarithms of the roles' final estimates and converted back. Six point rules are computed for every property: the arithmetic mean; the median; the twenty-percent trimmed mean (Jose et al., 2014); the inverse-variance pool, with weights proportional to the inverse of each role's variance of log errors; stacking, a non-negative least-squares regression of the log price on the roles' log points with weights normalised to sum to one (Bates and Granger, 1969; Timmermann, 2006); and a robust pool that places a Student-t likelihood with four degrees of freedom on each role's estimate and locates the committee by iteratively reweighted least squares, with the scale taken from validation. The weights and scales of the last three rules are estimated once on the validation subsample, written to a file and applied unchanged to the test and holdout subsamples; when a run has no fitted weights, the inverse-variance and stacking rules fall back to equal weights and the robust rule to a scale from the committee's own dispersion, and the run manifest records the source of the weights so that the fallback is never silent. Following Proposition 7, the primary comparisons use the median, which is invariant to the deliberation, and the fitted rules are reported as secondary. The interval of a committee is the median, across roles, of each of the five quantiles, which yields a 90 percent interval $[q_{.05}, q_{.95}]$; an equal-weight mixture of the roles' distributions, with quantile functions interpolated in log space, is computed as an alternative and is reported in the calibration analysis. Every prediction file contains one row per aggregation rule, so that the choice of rule is a column rather than a run.

6.7. Audit trail

Each run writes, for every property, the packet that was built, the full transcript of every turn with its prompt hash, model identifier, tokens, cost, latency, validation retries and error, the committee assignment, the stopping and verification events, the final estimates and quantiles of each role, and the aggregated values; a manifest records the arm, committee, tier, split, seed, model identifiers, package version, weights source and spend-cap status, and a ledger appends one line per API call. Runs are resumable: an interrupted run skips properties whose transcript exists and rebuilds its prediction file from all transcripts. Before any test-set evaluation, a freeze file records the SHA-256 hash of every role prompt, of the four configuration files, of the schema module, and of the agents, debate and provider code, together with the model identifiers of the cost-matched pairs; a check command lists every deviation from the frozen state, and any deviation after the freeze is reported as a protocol deviation. Hosted models are not deterministic, and the paper does not claim bitwise reproducibility of the agentic arms; it claims instead that every number traces to a stored transcript and that run-to-run variation is measured by repeating a tenth of the properties.

7. Experimental design and preregistration

The design has one purpose: to attribute any difference between a deliberating committee and its alternatives to a named channel, communication, role diversity, evidence partition, inference compute, verification or provider, under a fixed information set and a fixed budget. This section states the estimands, the arms and the contrast each pair identifies, the preregistered hypotheses, the metrics and inference procedures, the sample sizes and budget actually used in this version of the paper, the freeze, and the arms that are pending.

7.1. Estimands

For a property i sold at date t_i with log price p_i, every arm produces an $\mathscr{I}_{t_i}$-measurable log estimate $\hat{y}_i$ and, for the arms with intervals, quantiles $q_{i,\tau}$ at $\tau \in \{0.05, 0.25, 0.50, 0.75, 0.95\}$. The operational target is the selected market value $\tilde{v}_i = \mathbb{E}[p_i \mid \mathscr{I}_{t_i}, D_i = 1]$ defined in Section 3.1; by Lemma 1 the expected squared distance to the price decomposes into the distance to $\tilde{v}_i$ and the irreducible transaction noise $\mathrm{Var}(p \mid \mathscr{I}_t, D = 1)$, so that a ranking of arms by price-prediction error is a ranking by distance to $\tilde{v}$ and is silent about the structural fundamental value V^F. The per-record error is $e_i = \hat{y}_i - p_i$, the log of the ratio of the estimate to the price. The primary estimand is the population mean of $|e_i|$ over arm's-length residential sales in the evaluation window, and the primary contrast between two arms A and B is the mean of the paired difference $|e_i^B| - |e_i^A|$ over the same properties; a negative value favours B. Calibration estimands are the coverage probabilities $\Pr(q_{i,.05} \le P_i \le q_{i,.95})$

and $\Pr(q_{i,.25} \leq P_i \leq q_{i,.75})$, whose nominal values are 0.90 and 0.50. All estimands are defined on the test partition and, separately, on the 2026 holdout partition; the validation partition serves only to choose.

7.2. Arms and what each contrast identifies

Table 7 lists the eleven arms. Four are non-agentic baselines fitted on the training partition and tuned on validation only: a hedonic regression of log price with municipality-by-quarter and grid fixed effects (A); a hierarchical hedonic model with municipality and grid random intercepts and a repeat-sales adjustment when a prior sale exists (B); a gradient-boosted model on the feature matrix of Section 5.6 (C); and an automated valuation model that adds comparable-based and spatial-temporal lag features to the boosted model and produces conformal intervals (D). Seven are agentic and share the packet of Section 6: the single generalist (E), the independent committee (F), the structured debate with one, two or four rounds (G1, G2, G4), the debate with a blinded skeptic and verification (H), the placebo committee (I), the full-view debate (G-fullview) and the cost-matched single agent (J). Two rows of the table describe the same protocol at different rounds because the rounds grid is itself a treatment.

The contrasts are read as follows. F against E isolates what twelve differently instructed and differently informed agents add over one agent given everything, without any communication; under Proposition 1 the gain is bounded by the common component of their errors. G against F isolates communication, because the round-0 estimates of G are, in distribution, those of F, and Proposition 5 states that no pool of later rounds can beat the best pool of round 0 unless evidence is exchanged. G against J separates communication from compute: J spends a comparable token budget on repeated reasoning of one model with one view, so that by equation (13) the contrast is positive only if provider diversification and evidence exchange outweigh the concentration of evidence weight that a partition induces; the paper does not treat test-time compute as a nuisance but as a rival explanation that must be matched (Snell et al., 2024). G against I holds the views, the mechanics and the budget fixed and changes only the mandates, so that by Proposition 8 a non-zero contrast identifies complementarity between the prompt and the evidence, not the value of the partition; G against G-fullview holds the mandates fixed and removes the partition, so the two secondary contrasts together separate the two components of role diversity. H against G isolates the effect of flagging fabricated citations and out-of-band adjustments and of a reviewer who has not read the discussion. The rounds grid identifies the shape of the returns to deliberation and allows the convergence, anchoring and interquartile-range statistics of the transcripts to be plotted against the error. The three committee compositions, run for E, F and G2 on validation, identify provider dependence: the term $\sigma_g^2(1-\sum_m(n_m/K)^2)$ of the theory predicts that the mixed committee has the lowest pooled error if provider-common errors are material, and no difference if they are not.

Table 7. Treatment arms and the contrast each pair identifies.

Arm	Description	Contrast and channel identified
A	Hedonic ridge regression, fixed effects	Reference level of a transparent statistical model
B	Hierarchical hedonic with repeat-sales adjustment	Value of the prior sale and of spatial pooling
C	Gradient boosting (LightGBM), tuned on validation	Reference level of tabular machine learning (Grinsztajn et al., 2022)
D	AVM: boosting plus comparable and lag features, conformal intervals	H1 benchmark; the strongest non-agentic arm
E	Single generalist, full packet	Reference level of one model on the same evidence
F	Twelve roles, independent, aggregated	F vs E: ensembling and role diversity without communication
G1, G2, G4	Structured debate, 1, 2 or 4 rounds	G vs F: communication at fixed roles and views; rounds 0/1/2/4: diminishing returns, convergence, anchoring
H	Debate with blinded skeptic and deterministic verification	H vs G2: verification and a blinded reviewer
I	Placebo committee: generic prompts, same views, same mechanics	G2 vs I: prompt–evidence complementarity at a fixed partition
G-fullview	Default prompts, full packet for every role	G2 vs G-fullview: value of the evidence partition
J	Generalist sampled twelve times plus self-refinement	G2 vs J: communication versus compute at matched tokens; F vs J: twelve roles versus one model resampled

Notes. Arm letters follow the preregistration and are used throughout the tables. All agentic arms receive the same evidence packet under the same information cutoff; the mid-tier model pair is `claude-sonnet-5-5` and `gpt-6.1-sol`, assigned to roles by a seeded balanced permutation in the mixed committee. Each contrast is paired on the same properties. Source: `configs/treatments.yaml`.

7.3. Preregistered hypotheses

Six hypotheses were written before any test-set evaluation and are reproduced from the preregistration document. H1 states that G2, the two-round debate of twelve roles in the mixed committee, has a lower mean absolute log error than the AVM (D) on the test subsample. H2 states that G2 has a lower mean absolute log error than F on the same properties. H3 states that G2 has a lower mean absolute log error than J. H4 states that G2 has a lower mean absolute log error and a lower 95th percentile of absolute log error than I. H5 states that the 90 percent interval of G2 under median-of-quantiles aggregation has a coverage closer to 0.90 than that of E and of F. H6 is descriptive: at least ninety percent of the evidence codes cited by G2 agents exist in their packets, and the adjusted-comparable consistency check passes for at least eighty percent of turns. The primary family comprises the four mean-absolute-log-error contrasts of H1 to H4; the family-wise error rate is controlled at five percent by the step-down procedure of Holm (1979), and all tests are two-sided. The P95 component of H4, H5, the rounds grid, H against G2, G2 against G-fullview, the provider committees, the stratified errors, the spatial holdout, the cost frontier and the repeat-sales triangulation are secondary and are reported with confidence intervals and without claims of significance. The preregistration also states what would falsify the claim that deliberation adds information: H2 and H3 both not rejected, with confidence intervals excluding a two percent relative improvement in mean absolute log error. The mapping between each hypothesis and the propositions of Section 3 is given in Table 2.

7.4. Metrics

The evaluation protocol of the repository fixes the definitions; we summarise them. The primary metric is the mean absolute log error $\frac{1}{n}\sum_i |e_i|$. Secondary accuracy metrics are the mean absolute error in CAD, the median absolute percentage error, the root mean squared log error, the 90th and 95th empirical percentiles of $|e_i|$, the mean log error as a bias measure, and the severe-mispricing rate, the share of properties with $|e_i| > \log 1.25$, reported separately as overvaluation and undervaluation in percent terms as a lender would read them. Interval metrics are the raw coverage of the 50 and 90 percent intervals that the five quantiles bracket directly; the coverage of 80 and 95 percent intervals, which the quantiles do not bracket and which are obtained by fitting a normal distribution in log space to the five quantiles on the probit grid, and are labelled as interpolated wherever they appear; the mean log width of the 90 and 50 percent intervals; the pinball loss averaged over the five quantiles (Koenker and Bassett, 1978); and a lower-biased continuous ranked probability score on the five-point grid (Gneiting and Raftery, 2007). A probability-integral-transform histogram and a reliability diagram per quantile complete the calibration assessment, and a split-conformal rescaling of the intervals, with factors estimated on validation only, is reported as a remedy rather than as a

result (Romano et al., 2019). Each agentic arm also reports its mean estimated cost in dollars per property, its mean wall-clock latency and its mean token count from the ledger. Stratified results are reported by region group, urban class, property class, price tercile, decade of construction and local transaction density, and any cell with fewer than thirty properties is suppressed.

7.5. Inference

All contrasts are paired on the same properties. The sampling unit for inference is the block defined by municipality and calendar month of the sale, which absorbs the dependence between sales in the same local market at the same time; a paired block bootstrap resamples blocks with replacement, 2,000 times with a fixed seed, applying the same block draw to both arms, and reports the percentile 95 percent interval of the contrast and a bootstrap p-value from the recentred distribution (Cameron et al., 2008). Holm's adjustment is applied within the primary family. As a complement, a Diebold–Mariano statistic on the loss differential ordered by sale date, with a Newey–West long-run variance and the small-sample correction of Harvey et al. (1997), is reported for the same pairs (Diebold and Mariano, 1995; Newey and West, 1987); its stationarity assumption is not guaranteed in a cross-section ordered by date, and the block bootstrap remains the primary inference tool. Effect sizes are reported as relative changes in the primary metric, so that a contrast can be practically negligible and statistically precise, or the reverse, and the text says which.

7.6. Budget, sample sizes and the nested design

The preregistered subsamples are 240 validation, 600 test and 300 holdout properties. The experiments of this version of the paper are run under a self-imposed envelope of roughly one thousand US dollars at list prices and use nested prefixes of those subsamples. A pilot of twelve validation properties with E, F and G2 in the two single-provider committees, capped at sixty dollars, served to catch prompt errors and to calibrate the cost of a turn; the live sanity run measured a single-agent turn at the mid tier at about 5,500 input and 1,000 output tokens, which prices arm E at roughly two cents per property, arm F at about a quarter and the two-round debate at between 1.2 and 1.4 dollars. The validation tier uses forty properties for E, F and G2 in all three committees, for G1, H, I, G-fullview and J in the mixed committee, and twenty properties for G4, capped at five hundred dollars; it is the only tier on which anything is chosen, and it supplies the aggregation weights and the conformal factors. The test tier uses the first eighty test properties at the mid tier for the six main arms E, F, G2, H, I and J in the mixed committee, capped at five hundred dollars, and the first three hundred test properties at the small tier for E, F and G2, capped at one hundred and twenty dollars, which provides power for the central contrasts and a second point on the cost frontier. The holdout tier uses the first sixty 2026 properties for E, F and G2, capped at one hundred and twenty

dollars. These sizes are those of the budget configuration of the repository; with eighty properties the primary contrasts are powered to detect only large differences, and the paper reports the width of the intervals rather than claiming precision it does not have.

Two properties of the runner make the design extensible without compromising it. Every run is resumable, so that a run of size n can be continued to $n' > n$ by valuing the additional properties only; and because the subsample order is fixed by the seed and the runner takes the first n properties, the eighty-property test run is a prefix of the six-hundred-property preregistered sample, not a different sample. Extending the test tier to two hundred or to the full six hundred properties at the mid tier for the six main arms is estimated at roughly 1,100 and 3,300 dollars respectively at list prices before caching. A spend cap set in the environment stops any run when the ledger exceeds it, writes what exists and records the event in the manifest, so that a truncated run is never mistaken for a complete one.

7.7. Freeze procedure

Before the first test-set run, the freeze command records in a versioned file the hash of every role prompt, including the placebo prompts, of the four configuration files, of the schema module and of the agents, debate and provider code, together with the model identifiers of the cost-matched pairs, the package version, the Python version and the timestamp. The preregistration document is frozen at the same moment. The check command compares the working tree to the frozen hashes and lists every deviation; the paper reports any such deviation, with its reason, in the limitations section. No prompt, role, round count, model identifier, effort setting, packet construction rule or aggregation weight was changed after the freeze on the basis of a test-set number; the validation tier is where every such choice was made, and the test and holdout tiers were each evaluated once.

7.8. Pending arms and the prospective set

A human-appraiser arm, in which licensed or trained appraisers would value the same packets under the same cutoff with blinded prices, randomised case order and recorded confidence intervals, is part of the research program but has not been performed: it requires a research-ethics certificate from the Université du Québec en Outaouais, which has not yet been obtained, and no human benchmark is reported or simulated in this paper. The arm is marked pending wherever a table could have contained it. A prospective evaluation set is defined by sales registered in the live database after the freeze timestamp, whose identifiers are absent from the snapshot and which therefore could not have influenced any choice; it is fetched by a dedicated script, valued under the frozen protocol, and reported as the final robustness check, together with the descriptive comparison of these sales with their listing history where one exists.

8. Main results

This section reports the preregistered evaluation on the test partition, the reading of hypotheses H1 to H4, the effect sizes and what the confidence intervals exclude, the cheaper model tier, the evaluations on the untouched 2026 holdout and on the prospective set fetched after the freeze, the run-to-run variation of the hosted models, and the cost and latency of every arm. All numbers are those of the versioned files written by the evaluation script; the preregistered contrasts were computed once, after the freeze, and nothing in this section was used to choose anything. Throughout, $|e|$ denotes the absolute log error of Section 7.1, the agentic arms use the mid-tier model pair in the mixed committee with the median aggregation unless stated otherwise, and n is the number of properties.

8.1. Accuracy on the test partition

Table 8 gives the accuracy of every arm. The non-agentic baselines are scored on the full test partition of 96,935 sales; the agentic arms on the first eighty properties of the preregistered subsample at the mid tier, and on the first three hundred at the small tier. Among the baselines the ordering is the expected one: the hedonic regression (A) has a mean absolute log error of 0.191, the hierarchical model with the repeat-sales adjustment (B) 0.182, the three boosted models (C) between 0.177 and 0.180, and the AVM (D), which adds comparable-based and spatial-temporal lag features to the boosted model, 0.171, with a median absolute percentage error of 10.5 percent and a severe-mispricing rate, the share of sales valued more than 25 percent away from the price, of 21.8 percent. The AVM is the strongest non-agentic arm on every accuracy metric and is the H1 benchmark.

The six agentic arms at the mid tier are close to one another and close to the AVM. The single generalist (E) has a mean absolute log error of 0.166, the independent committee (F) 0.159, the two-round debate (G2) 0.158, the debate with the blinded skeptic (H) 0.158, the placebo committee (I) 0.157 and the cost-matched single agent (J) 0.163. The committees have a smaller median error than the generalist, 8.7 to 9.3 percent against 10.7 percent, and the lowest 95th percentile of $|e|$ among all arms, 0.493 to 0.499 against 0.508 for E; their mean log error is within 0.004 of zero. Because the eighty properties are not the full partition, the relevant AVM figure is its paired value on the same eighty, 0.163, which appears in the contrast tables and is lower than its full-partition value, so that the subsample is somewhat easier than the partition as a whole. Figure 4 shows the distribution of the log error by arm: the agentic arms have a slightly tighter interquartile range than the baselines and whiskers that end earlier on the right, but the picture is one of methods that make errors of the same order on the same properties.

Table 8. Accuracy by arm on the test partition.

Arm	Mean $\lvert e\rvert$	MAE (CAD)	MdAPE	RMSE (log)	P90 $\lvert e\rvert$	P95 $\lvert e\rvert$	Bias $\bar{e}$	$\lvert e\rvert > \log 1.25$	N
A (hedonic)	0.191	88962	0.126	0.286	0.412	0.641	0.014	0.270	96935
A (hedonic spatial)	0.208	76181	0.154	0.279	0.417	0.552	−0.078	0.359	8988
B (hierarchical)	0.182	84925	0.115	0.281	0.402	0.645	0.026	0.242	96935
C (catboost)	0.177	82153	0.111	0.278	0.375	0.640	0.021	0.226	96935
C (gbm)	0.179	83227	0.114	0.278	0.388	0.630	0.023	0.238	96935
C (gbm spatial)	0.176	65412	0.128	0.258	0.358	0.537	0.050	0.241	8988
C (xgboost)	0.180	85232	0.112	0.286	0.392	0.668	0.055	0.231	96935
D (avm)	0.171	80059	0.105	0.270	0.373	0.622	0.010	0.218	96935
D (avm spatial)	0.161	57681	0.105	0.248	0.345	0.527	0.045	0.204	8988
E (single llm)	0.166	76060	0.107	0.250	0.354	0.508	0.013	0.237	80
E (single llm small)	0.182	90125	0.104	0.287	0.406	0.634	0.038	0.273	300
F (independent)	0.159	76764	0.091	0.240	0.314	0.499	0.003	0.212	80
F (independent small)	0.176	86596	0.100	0.281	0.417	0.572	0.035	0.257	300
G (debate r2)	0.158	73765	0.087	0.243	0.307	0.493	−0.004	0.225	80
G (debate r2 small)	0.180	88861	0.106	0.283	0.413	0.602	0.028	0.280	300
H (debate skeptic)	0.158	73923	0.092	0.242	0.322	0.493	−0.002	0.225	80
I (placebo)	0.157	75095	0.093	0.239	0.321	0.497	0.004	0.212	80
J (cost matched)	0.163	75760	0.118	0.249	0.347	0.497	0.009	0.250	80

Notes. $e = \log \hat{P} - \log P$. Mean $\lvert e\rvert$ is the primary metric; MAE is the mean absolute error in Canadian dollars; MdAPE the median absolute percentage error; RMSE the root mean squared log error; P90 and P95 the empirical percentiles of $\lvert e\rvert$; Bias the mean of e; the last-but-one column the share of sales with $\lvert e\rvert > \log 1.25$. Baselines A–D are scored on the full test partition (sales registered 2025-04-01 to 2025-12-31); rows marked *spatial* are the versions trained without Outaouais, Abitibi-Témiscamingue and Bas-Saint-Laurent, scored on the 8,988 test sales of those regions. Agentic arms E–J use the mid-tier pair (`claude-sonnet-5-5`, `gpt-6.1-sol`) in the mixed committee with the median aggregation on the first 80 properties of the preregistered test subsample; rows marked *small* use the small-tier pair (`claude-haiku-5-5`, `gpt-6-luna`) on the first 300. Source: `results/evaluation/summary_main.csv`, `scripts/20_evaluate.py`.

8.2. Preregistered contrasts and the reading of H1 to H4

Table 9 reports the paired contrasts. The first four rows are the primary family: G2 against D, F, J and I on the mean absolute log error, with the block bootstrap of Section 7.5 over 65 municipality-by-month blocks and the step-down adjustment of Holm (1979); the fifth row is the tail component of H4; the remaining rows are secondary. Every primary contrast is small and every interval contains zero. Against the AVM, the debate lowers the mean absolute log error by 0.005, with a 95 percent interval from −0.033 to 0.026; against the independent committee by 0.001, interval −0.006 to 0.005; against the cost-matched single agent by 0.006, interval −0.017 to 0.004; against the placebo committee it raises it by 0.000, interval −0.005 to 0.006. The bootstrap p-values are 0.75, 0.70, 0.29 and 0.95, the Holm-adjusted values are all equal to one, and the Diebold–Mariano statistics on the date-ordered loss differentials give the same reading, with p-values between 0.25 and 0.95. The P95 component of H4 is −0.004 with interval −0.022 to 0.065. None of H1 to H4 is rejected.

Three readings follow from the pattern of the estimates rather than from any one test. First, a committee of twelve deliberating agents is, on these properties, about as accurate as the AVM: a difference of 0.005 in mean absolute log error, three percent in relative terms, in favour of the

committee, estimated with an interval that extends from a 20 percent advantage to a 16 percent disadvantage. Second, communication adds nothing measurable to independent aggregation: the G2–F contrast is the most precisely estimated of the family because the two arms are the same twelve roles on the same views and differ only in the two rounds of exchange, and its interval, -0.006 to 0.005, is narrow relative to the others. Third, the compute-matched single agent is close to the committee, 0.163 against 0.158, and the placebo committee is indistinguishable from it, which is the prediction of Proposition 8(ii) under additive separability of prompt and view. The secondary contrasts complete the picture: the committee beats the single generalist by 0.008, interval -0.023 to 0.004, and by 0.020 in median absolute percentage error, interval -0.060 to 0.007; the independent committee beats the generalist by 0.007; neither difference is distinguishable from zero at these sample sizes, but both are in the direction that the averaging of reasoning errors in Proposition 6 predicts, and both are larger than the G2–F difference.

Table 10 states what the intervals rule out. In relative terms the G2–D contrast is -3.1 percent with an interval from -20.0 to $+16.2$ percent; G2–F is -0.7 percent, interval -4.0 to $+3.2$ percent; G2–J is -3.5 percent, interval -10.6 to $+2.5$ percent; G2–I is $+0.1$ percent, interval -2.9 to $+3.5$ percent. The minimum effect that the design could have detected with 80 percent power, given the bootstrap standard errors, is 0.043 log points or 26 percent of the AVM's error for H1, 0.008 or 5.1 percent for H2, 0.015 or 9.5 percent for H3, and 0.007 or 4.6 percent for H4. The design was therefore powered to detect a moderate gain from communication and from roles, and only a large gain over the AVM. The preregistration stated that the claim that deliberation adds information would be falsified if H2 and H3 were both not rejected with intervals excluding a two percent relative improvement. Both are not rejected, and the H2 interval excludes improvements larger than 4.0 percent while the H3 interval excludes improvements larger than 10.6 percent; neither excludes a two percent improvement. The falsification criterion is therefore not met in its strict form: the evidence is that any information added by two rounds of deliberation is worth less than four percent of the committee's error, not that it is zero, and the eighty-property test tier cannot separate the two. What the test tier does establish is the ordering of the channels: whatever twelve agents add over one, they add it before the first round of communication.

8.3. The small tier and the cost frontier

The small-tier runs provide the second point on the cost frontier and a larger sample for the central contrast. With `claude-haiku-5-5` and `gpt-6-luna` on three hundred properties, the generalist has a mean absolute log error of 0.182, the independent committee 0.176 and the two-round debate 0.180, between 0.017 and 0.022 above the mid-tier values of the same arms, with a positive mean log error of 0.028 to 0.038 and severe-mispricing rates of 25.7 to 28.0 percent against 21.2 to 23.7 percent at the mid tier. The cheaper models are also worse calibrated: their raw 90 percent coverage

is 0.777, 0.787 and 0.807 against 0.887 for the three mid-tier arms, their 50 percent coverage 0.417 to 0.427 against 0.463 to 0.512, with intervals that are narrower, not wider, than those of the mid tier. Section 10 returns to this. Within the small tier the independent committee is again the best of the three and the debate does not improve on it; the ordering E, G2, F reproduces the mid-tier finding that the gain from twelve roles over one is obtained without communication.

Figure 5 places every arm in the cost–accuracy plane. The frontier runs from the AVM, which costs nothing per property and has a mean absolute log error of 0.161 on its spatial variant and 0.171 on the full partition, through the independent committee at 23 cents and 0.159, to the debate and skeptic arms at 1.35 to 1.52 dollars and 0.158. The small-tier arms lie above the frontier at every cost: the small-tier debate costs 12.5 cents per property, six times the mid-tier generalist, and is worse than it by 0.014. The frontier is almost flat: moving from the AVM to the most expensive arm buys a reduction of about 0.005 in mean absolute log error for a cost of 1.5 dollars and five to six minutes of wall-clock time per property. These are the numbers that the lender of Section 12 would weigh.

8.4. Temporal holdout and prospective evaluation

The 2026 holdout partition, 83,489 sales registered from 2026-01-01 to 2026-10-07, was evaluated once, after the test results were written, for the baselines on the full partition and for E, F and G2 on the first sixty properties of the holdout subsample. The prospective set comprises sales present in the live database but absent from the 2026-10-21 snapshot, registered between 2026-10-06 and 2026-10-13, whose identifiers could not have influenced any decision; the baselines are scored on all 1,182 of them and the three agentic arms on forty. Table 11 collects the accuracy and calibration of the AVM and the agentic arms on the three evaluation sets; the full tables for the holdout and prospective sets are in the repository.

The holdout reverses the test-tier ordering between the AVM and the committees and reveals a drift that the test partition did not show. On the full 2026 partition the AVM's mean absolute log error is 0.180 with a mean log error of 0.002; on the sixty paired properties it is 0.189 with a mean log error of 0.055. The agentic arms have mean absolute log errors of 0.195 (E), 0.186 (F) and 0.189 (G2), their 95th percentiles of $|e|$ rise from about 0.50 on the test subsample to 0.81–0.85, and all three carry a mean log error of $+0.10$ to $+0.11$: in 2026 they overvalue, and 18.3 percent of their valuations are more than 25 percent above the price against 5 percent below. The paired contrasts on the holdout are G2 against D -0.000, interval -0.035 to 0.037, and G2 against F $+0.003$, interval -0.001 to 0.007, with a Holm-adjusted p-value of 0.26; the generalist is worse than the independent committee by 0.009, interval -0.021 to 0.002. The median error of the agentic arms remains low, 8.6 to 9.0 percent against 11.2 percent for the AVM on the full partition, so that the drift is in the tails and in the level, not in the typical case. One reading, which the data cannot confirm, is that the

agents extrapolate the 2025 expansion into 2026 from the market section of their packets, while the AVM's lag features are re-computed as of each valuation date; the fundamental-value triangulation of Section 11 is the only place where a bias of this sign could be interpreted as something other than an error.

The prospective set shows a bias of the same sign in every arm, including the baselines, which suggests that the market itself moved between the training window and October 2026. On the 1,182 prospective sales the AVM has a mean absolute log error of 0.202 and a mean log error of $+0.039$, the hedonic regression 0.219, the boosted models 0.215 to 0.222 with mean log errors of $+0.050$ to $+0.095$. On the forty paired properties the AVM is at 0.197 with a mean log error of $+0.059$, and the agentic arms at 0.187 (E), 0.179 (F) and 0.187 (G2) with mean log errors of $+0.058$ to $+0.061$. The contrasts are G2 against D -0.010, interval -0.057 to 0.043, and G2 against F $+0.009$, interval -0.000 to 0.021, Holm-adjusted p-value 0.19. Across the three evaluation sets the independent committee is never worse than the debate and is better than it on the two later sets by 0.003 and 0.009, with intervals whose lower bounds touch zero; the direction is that of Proposition 4(iii), under which opinion exchange with median anchoring can only lose accuracy, but the magnitudes are at the edge of what forty to sixty properties can resolve.

8.5. Run-to-run variation

Hosted models are not deterministic, and the paper does not claim bitwise reproducibility of the agentic arms. Eight of the eighty test properties, one tenth, were re-run through the two-round debate with the same packets, the same protocol, the same committee assignment and the same model identifiers, so that only the providers' sampling differed. The mean absolute difference between the two runs' log point estimates is 0.0069, the median 0.0044, the maximum 0.0186 and the standard deviation of the difference 0.0092; the correlation between the two runs' absolute errors is 0.998, the mean absolute log error on the eight properties is 0.195 in the original run and 0.198 in the repeat, the mean 90 percent log width 0.504 and 0.518, and every repeated point lies inside the original run's 90 percent interval. Run-to-run noise of the committee output is therefore of the order of 0.005 to 0.01 log points, an order of magnitude below the differences between arms and the baselines, and about the size of the G2–F and G2–I contrasts, which is one more reason to read those two contrasts as zero. The repeat run cost 12.39 dollars.

8.6. Cost and latency

Table 12 reports the ledger cost, wall-clock latency and token count per property. At the mid tier the generalist costs 2.2 cents, 20 seconds and 6,600 tokens per property; the independent committee 23 cents, 79 seconds and 74,000 tokens; the cost-matched single agent 26 cents, 94 seconds and

92,000 tokens, which confirms that J is matched to F in tokens, not to G2, as Section 6.5 explains; the two-round debate 1.52 dollars, 336 seconds and 521,000 tokens, seven times the tokens of the independent committee for the same accuracy; the skeptic and placebo variants 1.35 and 1.40 dollars. The small tier is about twelve times cheaper per property at the same latency, since latency is dominated by the number of sequential turns rather than by the model. The six mid-tier test runs together cost 382 dollars, the three small-tier runs 44 dollars, and the ledgers record 1,078.82 US dollars of live API spend over every run of the paper, pilots, validation ablations, test, holdout, prospective and the repeat run included, against the envelope of roughly one thousand dollars stated in Section 7.6. At list prices and before caching, valuing the full test partition of 96,935 sales with the two-round debate would cost about 147,000 dollars and the AVM nothing, which is the practical content of a flat cost frontier.

9. Ablations and robustness

The ablations of this section were run on the validation subsample, forty properties for every arm and twenty for the four-round debate, before the freeze, and they are the runs on which the aggregation weights and the conformal factors were estimated. They are exploratory in the sense of the preregistration: they are reported with block-bootstrap intervals and without claims of significance, and nothing in them was re-estimated after the test results were known. Two further ablations use the test partition itself, because they involve no choice: the comparison of aggregation rules, which is a column of the prediction files, and the spatial holdout of the baselines. The stratified errors are reported for both partitions. With forty properties the minimum detectable effect for a within-protocol contrast such as G2 against G1 is of the order of 0.01 log points and for a cross-protocol contrast such as G2 against D of the order of 0.05, so that the section can rule out large effects and nothing smaller.

9.1. Rounds of deliberation

Figure 6 and Table 13 follow the committee through its rounds. For every property and every round the transcripts give the twelve current log estimates; the table reports the mean absolute error of their median and of their mean, with block-bootstrap intervals, together with the mean interquartile range of the twelve estimates. Round 0 is the independent vector of arm F. When the stopping rule of Section 6.5 ends a debate early, because the interquartile range changed by less than one percent between consecutive rounds, the committee's last estimates are carried forward to the later rounds, so that every round averages the same properties and the curve is not contaminated by the selection of the committees that kept debating. In the two-round debate of the mixed committee the mean absolute error of the median goes from 0.180 at round 0 to 0.178 after one round and

0.176 after two, a change of 0.004 inside an interval of width 0.11, while the interquartile range of the estimates falls from 0.070 to 0.044 and then 0.034, a contraction to about half in two rounds. The one-round debate reproduces the first step, 0.180 to 0.178 and 0.069 to 0.043. The four-round debate on twenty properties starts lower, 0.162, because the twenty properties are a different sample, and stays at 0.163 through every later round while its interquartile range falls from 0.057 to 0.032, 0.026, 0.025 and 0.024: after the second round the committee has stopped moving and the accuracy has not changed. The all-Anthropic and all-OpenAI committees give the same reading at opposite extremes of convergence, 0.181, 0.179 and 0.178 with the interquartile range falling from 0.056 to 0.018 for the former, 0.177, 0.178 and 0.178 with the range barely moving from 0.042 to 0.038 for the latter. The full-view committee goes from 0.173 to 0.174 and 0.175. The paired contrasts on the final outputs are G2 against G1, -0.002 with interval -0.008 to 0.002, and G4 against G2 on the same twenty properties, -0.001 with interval -0.005 to 0.002.

The dispersion collapses and the accuracy does not move. This is the pattern of Proposition 4: opinion exchange contracts every dispersion statistic at a geometric rate whatever the error structure, while the pooled error is unchanged, so that convergence is not evidence. The transcript statistics of Table 14 say how the contraction happens. The anchoring index, the correlation across roles between the revision $x_{k,r} - x_{k,0}$ and the distance to the initial committee median $m_0 - x_{k,0}$, is 0.67 on validation and 0.71 on the test runs of G2, 0.68 at the small tier, 0.70 on the holdout and 0.76 on the prospective set; between 43 and 52 percent of the roles end closer to the initial median than they started. Revisions are thus mostly moves toward the table that every role is shown, and the revision instruction, which states that moving toward the committee without an evidential reason is an error, does not prevent them. Under the remark on public signals in Section 3.2, anchoring of this kind is harmless as long as the weight placed on the median does not exceed twice its optimal weight, and the flat accuracy curve is consistent with a weight in that range; what it is not consistent with is the evidence-exchange gain of Proposition 6, which would have moved the pooled error. Cited evidence identifiers that lie outside a role's view are passed by identifier only, and the roles' comparable pools are drawn from the same retrieval, so that the exchange term $\sigma_u^2(\|\omega'\|^2 - \|\omega\|^2)$ has little to act on.

9.2. Blinded skeptic and deterministic verification

Arm H removes the skeptic from the rounds, blinds it to the discussion, and injects deterministic verification flags into the next prompt of any role whose estimate cites a non-existent code, adjusts a comparable by more than forty percent, or lies outside the range of the adjusted comparables widened by a quarter. On the test subsample H and G2 are identical to three decimals on the primary metric, 0.158 against 0.158, with a paired difference of -0.000 and interval -0.003 to 0.003; the median error is 9.2 against 8.7 percent, interval for the difference -0.017 to 0.007;

the 95th percentile, the severe-mispricing rate and the pinball loss do not differ; the 90 percent coverage is 0.875 against 0.887. On validation the picture is the same, 0.175 against 0.176. The verification tool has little to correct because the faithfulness audit of Section 10.6 finds that every cited identifier exists and that 98 to 99 percent of adjustments already lie within the forty percent band in the unverified arms; the flags that remain concern individual adjustments and do not move the committee median. H is, however, cheaper than G2, 1.35 against 1.52 dollars, because the skeptic's turns are removed from the rounds, so that at equal accuracy it dominates the full debate on the cost frontier.

9.3. Placebo roles

Arm I keeps the two rounds, the twelve evidence views and the committee state of G2 and replaces the twelve specialised mandates by generic prompts. On the test subsample the placebo committee has a mean absolute log error of 0.157 against 0.158, a difference of 0.000 with interval -0.005 to 0.006 and a Holm-adjusted p-value of one; its 95th percentile of $|e|$ is 0.497 against 0.493, interval for the difference -0.022 to 0.065. On validation the same contrast had pointed the other way, -0.008 with interval -0.023 to 0.003 on the mean and -0.044 with interval -0.156 to 0.011 on the 95th percentile, which is why H4 was retained in the primary family; the test tier did not reproduce it. By Proposition 8, a contrast between I and G2 with views fixed to the slots identifies the complementarity between a prompt and the evidence it is paired with, and nothing else; its absence means that, on this packet, a role told to reason as a mortgage underwriter about the underwriting section does no better than a role told to value the property from the same section. The placebo committee also deliberates like the role committee: its anchoring index is 0.75 against 0.71, its interquartile range contracts from 0.087 to 0.041 against 0.088 to 0.038, and its summaries are shorter, 72 words against 78. Role labels change the prose and not the numbers.

9.4. Full-view committee: the value of the evidence partition

The second component of role diversity is the partition of the packet into least-privilege views. The full-view variant of G2 keeps the twelve default prompts and gives every role the whole packet; it was run on validation only. Table 15 reports the paired contrasts. Against G2 the full-view committee has a mean absolute log error of 0.175 against 0.176, difference -0.001 with interval -0.006 to 0.004; the median error is unchanged, the 95th percentile differs by 0.005, and the 90 percent coverage is 0.850 against 0.825. Against the placebo committee it is better by 0.009, interval -0.027 to 0.004, and against the independent committee by 0.005, interval -0.017 to 0.004. The partition therefore carries no measurable value on this packet, which is what Proposition 6(ii) predicts when the union of the views is what a single agent would be shown anyway: after the

exchange the committee's evidence error is at best that of one agent with the union view, and giving the union to every role from the start reaches the same point without communication. Two secondary observations follow. First, with the full packet the share of turns in which every stated adjustment lies within forty percent of the raw comparable price rises from 0.71 to 0.88, close to the 0.90 to 0.98 of the single-agent arms, which is consistent with part of the flagged adjustments in the partitioned committee coming from roles that reason about comparables they were not shown; the paper does not test this. Second, the full-view committee anchors slightly less, 0.65 against 0.67, with a larger share of roles moving toward the median, 0.52 against 0.44, which is the expected behaviour of agents whose views no longer differ.

9.5. Provider committees

Proposition 3 predicts that the mixed committee has the lowest pooled error if provider-common errors are material, and no difference if they are not. Table 16 reports E, F and G2 in the all-Anthropic, all-OpenAI and mixed committees on validation, with paired contrasts of each single-provider committee against the mixed one. For the two-round debate the three committees have mean absolute log errors of 0.178, 0.178 and 0.176, differences of 0.002 from the mixed committee with intervals -0.014 to 0.016 and -0.011 to 0.019; for the independent committee 0.182, 0.178 and 0.179; for the single generalist 0.194, 0.177 and 0.182, the largest spread, with the all-Anthropic generalist worse than the mixed one by 0.013, interval -0.025 to 0.048. The provider-common term is thus not detectable at forty properties, where the minimum detectable effect for the G2 contrasts is about 0.02 log points, or twelve percent. What differs between providers is the shape of the error and the dynamics of the debate rather than its level. The Anthropic models carry a smaller upward bias, 0.043 to 0.062 against 0.073 to 0.080 for the OpenAI models, and wider intervals, 0.58 against 0.49 to 0.52 in log width; the mixed committee sits between. In deliberation the two providers behave very differently: the all-Anthropic committee has an anchoring index of 0.97 and contracts its interquartile range from 0.056 to 0.018, almost complete convergence on the initial median, while the all-OpenAI committee has an anchoring index of 0.38, moves only 7.5 percent of its roles toward the median, and keeps its interquartile range at 0.038 against 0.042 initially; the mixed committee is at 0.67 and 0.034 against 0.070. Neither behaviour changes the accuracy of the median, which is the point of Proposition 4: a committee that converges completely and a committee that barely revises reach the same pooled error, because in neither case is the revision generated by evidence. The three committees do not differ in faithfulness either, with every cited identifier valid in all of them, but the Anthropic roles write shorter summaries, 63 to 69 words against 83 to 87.

9.6. Aggregation rules

Every prediction file carries one row per aggregation rule, so that the choice of rule can be examined on the test partition without any new run. Table 17 gives the mean absolute log error of the committee arms under the median, the arithmetic mean, the twenty-percent trimmed mean, the inverse-variance pool and the stacking pool with weights estimated on the validation run of G2, the robust Student-t pool, and the equal-weight mixture of the roles' distributions. The weights themselves are nearly uniform for the inverse-variance rule, between 0.082 and 0.088 for eleven roles and 0.060 for the building-condition role, whose log-error standard deviation on validation is 0.296 against 0.244 to 0.252 for the others; the stacking weights are concentrated, 0.39 on the hedonic role, 0.24 on the geospatial role, 0.16 on the market economist, 0.13 on the data auditor and 0.08 on the underwriter, with zero on the seven others. At the mid tier the rule changes the mean absolute log error of G2 between 0.158 and 0.162 and that of F between 0.155 and 0.162, a range of 0.004 to 0.007 that is below the run-to-run noise of Section 8.5 and far inside the intervals of the primary contrasts; the mean is the best rule for the independent committee and the median for the debate, and stacking is the worst for both. At the small tier the ranges are 0.006 and 0.009 and stacking is the best rule for F. This is the forecast-combination result of Clemen (1989) and Timmermann (2006) in the form of Proposition 2: with roles of nearly equal accuracy, weights estimated on forty transcripts add estimation noise and cannot be expected to help, and a rule that concentrates on five roles discards seven nearly equivalent signals. For the cost-matched single agent the mean and the median of the twelve samples give 0.164 and 0.163 and the self-refined estimate 0.161, so that a critique-and-refine step on one model is worth about as much as two rounds among twelve. The primary comparisons use the median, which Proposition 7 identifies as the rule that is invariant to the deliberation; the table shows that nothing in the conclusions depends on this choice.

9.7. Spatial holdout of the baselines

The spatial experiment withholds Outaouais, Abitibi-Témiscamingue and Bas-Saint-Laurent from training and scores the withheld-region models on the 8,988 test-window sales of those regions; Table 18 pairs each withheld model with its in-region counterpart on the same sales. The hedonic regression suffers most: its mean absolute log error rises from 0.183 to 0.208, difference 0.025 with interval 0.019 to 0.030, and its mean log error falls from $+0.021$ to -0.078, a systematic undervaluation that follows from municipality fixed effects it cannot estimate for municipalities it has not seen; its 90 percent coverage falls from 0.921 to 0.886. The boosted model does not suffer at all: trained without the regions it is more accurate on them, 0.176 against 0.187, difference -0.011 with interval -0.013 to -0.009, a precisely estimated difference for which the paper offers

no mechanism beyond the observation that the three regions are a small share of the training data and that the in-region fit may be dominated by the rest of the province. The AVM is unchanged in accuracy, 0.161 against 0.163, interval for the difference -0.006 to 0.001, and in coverage, 0.907 against 0.910, with a larger upward bias, $+0.045$ against $+0.022$, interval for the difference 0.019 to 0.027. The methods that travel across regions are those whose information comes from comparables and lag features computed as of the valuation date rather than from region-specific parameters. The agentic arms were not run on the spatial holdout in this version of the paper: their packets contain comparables retrieved from the whole province and no fitted region effect, so that the experiment would withhold nothing from them, and the Outaouais stratum of the eighty-property test subsample contains two properties.

9.8. Stratified errors

Figure 7 and Table 19 report the mean absolute log error by region group, urban class, property class, price tercile and decade of construction, under the rule that a cell with fewer than thirty properties is suppressed. For the AVM the cells are those of the full partition and are all populated; for the mid-tier agentic arms only five of the twenty-four cells reach thirty properties, the Montréal CMA and the residual region group, the metropolitan class, houses and the top price tercile; for the small-tier debate seventeen cells do. The comparison between the AVM's full-partition cell and an agentic arm's subsample cell is not paired and should be read as indicative. Two patterns are nonetheless clear. The agentic arms do markedly better than the AVM in the Montréal CMA, 0.112 to 0.118 against 0.149 at the mid tier and 0.137 against 0.149 at the small tier, and no better outside it, 0.189 to 0.206 against 0.199 in the residual region group and 0.223 at the small tier; the Montréal cell is where comparables are densest and where the market section of the packet is most informative. The small-tier cells, which cover more of the distribution, show the agentic arms better than the AVM for condominiums, 0.088 against 0.115, for the middle price tercile, 0.125 against 0.131, and for houses built in the 2000s, 0.117 against 0.141, and worse for rural sales, 0.267 against 0.219, for houses, 0.200 against 0.179, for the cheapest tercile, 0.337 against 0.319, and above all for buildings of the 1970s, 0.274 against 0.185, and of the 2010s, 0.160 against 0.137. The cheapest tercile is the hardest for every method, with errors above 0.3, as the hedonic literature on thin and heterogeneous low-price segments would predict (Dubé et al., 2023). The stratification by local transaction density, thin, mid and liquid terciles of the number of sales within three kilometres in the preceding twelve months, is NOT YET ESTIMATED: the analytic table of this version does not carry the density variable, and the cell is absent from the strata file.

9.9. What was not run

Three robustness checks that the research plan lists were not run in this version and are stated as such rather than approximated. The sensitivity of the results to the exclusion flags of WP 22, non-residential, land, duplicate, assessment-ratio outlier, built-after-sale and multi-parcel transfers, has not been estimated: all partitions apply the flags and no run without them exists. The agentic arms were not run on the spatial holdout, for the reason given above. The rounds grid beyond two rounds was run at the mid tier in the mixed committee only, on twenty properties; the interaction of the exchange term with the share of cited identifiers outside the common view, which Section 3.5 proposes as the direct test of Proposition 6, is NOT YET ESTIMATED.

10. Calibration, interval quality and faithfulness

A valuation that is used as collateral or as an appraisal is a distribution, not a point, and the welfare model of Section 12 prices its width and its honesty. This section reports the coverage of the intervals at the four nominal levels, the probability-integral-transform histograms and the reliability of each quantile, the trade-off between sharpness and coverage and the conformal remedy, the conditional coverage by subgroup, the behaviour of the intervals under the drift of Section 8.4, the severe-mispricing rates that a lender reads, and the faithfulness audit of the transcripts that H6 concerns. The 50 and 90 percent coverages are raw, since the five quantiles bracket those intervals directly; the 80 and 95 percent coverages are interpolated from a normal fit in log space and are labelled as such.

10.1. Coverage at the four nominal levels

Table 20 and Figure 8 report the coverage, width and proper scores of every arm with intervals on the test partition. The hedonic and hierarchical baselines, whose intervals come from a Gaussian residual model, are close to nominal at every level, 0.489 and 0.499 at 50 percent, 0.909 and 0.913 at 90 percent, with a mean 90 percent log width of 1.00, that is, intervals spanning a factor of e between their bounds. The AVM's conformal intervals, with factors estimated on validation, cover 0.896 at 90 percent and 0.936 at 95 percent with a width of 0.83, but only 0.446 at 50 percent. The mid-tier agentic arms cover 0.887 at 90 percent for E, F, G2 and I, 0.875 for H and 0.863 for J, with widths of 0.56 to 0.60; at 50 percent they cover 0.463 (E), 0.512 (F), 0.475 (G2), 0.525 (H), 0.487 (I) and 0.475 (J); at the interpolated 95 percent level they cover 0.887 to 0.912, which is the first sign that their tails are too light. H5 stated that the 90 percent interval of G2 would have a coverage closer to 0.90 than that of E and of F. The three coverages are identical, 0.8875, 71 of 80 properties in each case, and the paired contrast of G2 against E is 0.000 with interval -0.050 to 0.051; H5

is not supported, and the reason is structural rather than statistical. The committee interval is the median across roles of each quantile, and Table 14 shows that the roles' own quantiles do not contract with the rounds, the 90 percent width being 0.590 for F and 0.602 for G2; Proposition 4(i) predicted that agreement-based intervals would contract while the pooled error did not, and the protocol's intervals are not agreement-based, which is why the two rounds leave coverage where it was. A linear pool of quantiles is not calibrated even when its components are (Ranjan and Gneiting, 2010); here the components are not calibrated either.

The small-tier arms are worse calibrated on every count: 0.777, 0.787 and 0.807 at 90 percent for E, F and G2, 0.417 to 0.427 at 50 percent, 0.813 to 0.860 at the interpolated 95 percent level, with intervals that are narrower than the mid tier's, 0.45 to 0.49 against 0.56 to 0.60, while their errors are larger. The cheaper models are therefore overconfident in the ordinary sense, and the committee of twelve of them, whose interval is the median of twelve overconfident intervals, inherits the overconfidence; this is the common-component term of Proposition 1 applied to the quantiles rather than to the points. The pattern is consistent with the general finding that language models' verbal confidence is poorly calibrated and that calibration improves with model scale (Kadavath et al., 2022; Xiong et al., 2024).

10.2. Probability integral transform and quantile reliability

Figure 9 shows the PIT histograms, the position of the realised log price in each arm's predictive distribution, interpolated in probit space between the five quantiles. A calibrated arm has a flat histogram (Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014). The AVM has a spike in the 0.2 to 0.3 bin, 19.5 percent of the mass against 10, and too little mass above 0.5: its distribution is shifted upward relative to the prices, which the quantile reliability confirms, with 0.578 of prices below the AVM's median quantile and 0.339 below its lower quartile, against 0.50 and 0.25 nominal. The hedonic baselines show a smaller spike in the same bin. The mid-tier agentic arms are closer to flat at eighty properties, with bins between 5 and 16 percent, and their reliability is good at the extremes and poor in the middle: for G2 the shares of prices below the 5th, 25th, 50th, 75th and 95th quantiles are 0.062, 0.225, 0.475, 0.700 and 0.950, so that the lower tail and the median are right, the upper quartile sits too low, and the interquartile band is narrow on its upper side. The small-tier histograms are U-shaped, with 15 percent of the mass in each of the two extreme bins for G2 and 18 and 15 percent for E, and reliabilities of 0.133 at the 5th and 0.910 at the 95th quantile for E: too many prices fall outside the stated 90 percent range on both sides, which is the signature of intervals that are too narrow rather than mislocated.

10.3. Sharpness, coverage and the conformal remedy

The agentic intervals are sharper than the AVM's at almost the same 90 percent coverage: 0.56 to 0.60 in log width against 0.83 on the full partition and 0.846 on the eighty paired properties, 27 to 34 percent narrower, for coverages of 0.887 against 0.896 and 0.875. Sharpness subject to calibration is the goal of probabilistic forecasting (Gneiting and Raftery, 2007), and on that criterion the committees' intervals are the more useful ones on the test partition. The proper scores are less favourable than the widths suggest: the pinball loss of G2 is 0.053 against 0.058 for the AVM on the full partition but 0.052 on the paired eighty, and the F–D contrast on the pinball loss is $+0.001$ with interval -0.008 to 0.011; the narrower intervals are paid for by the mass that falls outside them, and on the same properties the two families' predictive distributions score the same.

Table 21 applies the split-conformal rescaling of the evaluation protocol, with the factors estimated on the forty validation properties of each arm and committee and frozen before the test runs (Vovk et al., 2005; Lei et al., 2018; Romano et al., 2019; Angelopoulos and Bates, 2023). The factors are 1.55 to 1.63 for the 90 percent interval and 1.24 to 1.26 for the 50 percent interval, which says that the models' own 90 percent intervals needed to be stretched by more than half to cover 90 percent of validation prices. On the test partition the rescaled intervals cover 0.925 (E), 0.950 (F), 0.938 (G2) and 0.938 (J) at 90 percent and 0.525 to 0.637 at 50 percent, with 90 percent widths of 0.86 to 0.93: the conformal guarantee holds, and the sharpness advantage over the AVM disappears, since the rescaled widths match or exceed the AVM's 0.846 on the same properties. On the prospective set the rescaled coverages are 0.900 to 0.925. On the 2026 holdout they are 0.883 to 0.900 with widths of 0.80 to 0.82, below the guarantee for E and G2, because split conformal prediction assumes exchangeability between the calibration and the evaluation sets and the holdout is a different year with a different level of prices. The remedy restores marginal coverage where the distribution has not moved and does not where it has; the honest interval for 2026 would have had to be calibrated on 2026 data.

10.4. Conditional coverage

Marginal coverage can hide subgroups in which the intervals fail, and lenders care about the subgroups. Table 22 gives the raw 90 and 50 percent coverages by region group, property class and price tercile, under the thirty-property rule, for the AVM on the full partition cells and for the agentic arms on the subsample cells. At the mid tier the committees cover 0.944 of Montréal CMA prices and 0.861 of prices elsewhere at the 90 percent level, 0.860 to 0.895 for houses and 0.906 in the top tercile, against 0.911, 0.885, 0.896 and 0.928 for the AVM; the cost-matched single agent is lower in three of the four cells, 0.917, 0.861, 0.860 and 0.844. The small-tier debate, whose seventeen populated cells reach further into the distribution, covers 0.900 of condominiums and 0.853 of the

top tercile but only 0.789 of houses, 0.756 of plexes and 0.667 of the cheapest tercile, where the AVM itself covers only 0.770. The cheapest tercile is where every method's intervals fail: these are the sales for which comparables are heterogeneous and the price is most affected by condition that the packet does not observe. At the 50 percent level the mid-tier cells lie between 0.417 and 0.562 and the small-tier debate covers 0.222 of cheapest-tercile prices with its interquartile band. Conditional coverage therefore degrades in the same cells where accuracy degrades in Section 9.8, and the degradation is larger for the cheaper models.

10.5. Calibration under drift and severe mispricing

Table 11 already showed what happens to the intervals on the later evaluation sets. On the 2026 holdout the agentic arms' raw 90 percent coverage falls to 0.800, 0.833 and 0.833 and their widths to 0.50 to 0.53, narrower than on the test partition although the errors are larger, while the AVM covers 0.878 on the full partition and 0.900 on the sixty paired properties with a width of 0.82. On the prospective set the agentic coverages are 0.750, 0.800 and 0.825 against 0.866 and 0.900 for the AVM. The agents' intervals do not know that the market has moved: their width is a function of the packet, which is as of the valuation date, and of the model's habits, and neither widens when the model's bias changes sign. The AVM's conformal factor does not know either, but its interval is wide enough to absorb a bias of 0.05.

The severe-mispricing rates are the quantity a lender reads. On the test partition the share of valuations more than 25 percent away from the price is 0.218 for the AVM on the full partition and 0.212 on the paired properties, 0.237 for the generalist, 0.212 for the independent committee, 0.225 for the debate; over- and undervaluation are balanced for the committees, 0.100 and 0.088 for G2, whereas the AVM overvalues more often than it undervalues on the full partition, 0.110 against 0.059. On the holdout the rates are 0.233 to 0.267 for the agentic arms and 0.250 for the paired AVM, but their composition changes: 0.183 of the agentic valuations are more than 25 percent above the price and 0.050 below, and the paired AVM shows the same 0.183 above. On the prospective set the agentic rates are 0.275 to 0.300 with 0.175 above, and the AVM's 0.281 on all 1,182 sales with 0.157 above. A collateral haircut calibrated on the test partition's symmetric errors would have been too small in 2026 for every method, and the committees' narrower intervals would have made the shortfall larger for them than for the AVM. Proposition 10 gives the mapping from these rates to the lender's expected loss.

10.6. Faithfulness audit

H6 is descriptive: at least ninety percent of the evidence codes cited by G2 agents should exist in their packets, and the adjusted-comparable consistency check should pass for at least eighty percent

of turns. Table 23 reports the audit for every run, computed on the stored transcripts and packets. The first criterion is met with room to spare: in every run of every arm, on test, holdout, prospective and validation, 100 percent of the cited evidence identifiers exist in the role's packet, the packet hash recorded in each transcript matches the packet rebuilt from the snapshot, and the mean number of schema-validation retries per turn is at most 0.033. No role cited a comparable or a section it had not been shown, and in the terms of Lemma 2 no revision was persuasion dressed as evidence. The second criterion is met for the single-agent arms and not for the committees. Between 95.4 and 99.9 percent of individual adjustments lie within forty percent of the raw comparable price, but the share of turns in which every adjustment does is 0.900 for the generalist and 0.917 for the cost-matched agent against 0.709 for the independent committee, 0.721 for the debate, 0.717 for the skeptic variant and 0.711 for the placebo on the test partition, and 0.70 to 0.75 in every other committee run. The committees fail H6's second threshold, and the full-view ablation of Section 9.4, where the share rises to 0.883 when every role sees the whole packet, locates the failure in the partition: roles whose view does not contain the comparable section reason about comparables at second hand and state adjustments that the deterministic check cannot reconcile with a price they were not shown. The verification flags of arm H, which are injected into the next prompt, do not raise the share, 0.717, because a flagged role that cannot see the comparable cannot repair the adjustment. The remaining checks pass almost uniformly: the point estimate lies within the hull of the adjusted comparables widened by a quarter in 97.5 to 100 percent of turns, the five quantiles are positive, ordered, bracket the point and span a ratio between 1.01 and 4 in 98.5 to 100 percent, and the reasoning summaries of the test, holdout and prospective runs average 71 to 80 words with none above the 120-word limit.

These results bear on H6 and on nothing larger. The audit establishes that the explanations are consistent with the evidence the agents were given: the identifiers exist, the arithmetic is within bounds, the point lies among the comparables, and the stated uncertainty has the right shape. It does not establish that the stated reasons are the causes of the estimate. An explanation can cite valid evidence and be numerically consistent while the number was produced by a different process, the one that the placebo committee of Section 9.3 exposes when generic prompts reach the same estimates as specialised mandates with different prose; the literature on language-model explanations documents that stated reasoning can be unfaithful to the computation that produced the answer (Jacovi and Goldberg, 2020; Turpin et al., 2023; Lanham et al., 2023). The hull check is also weak by construction, since a band from three quarters of the lowest adjusted comparable to five quarters of the highest contains most defensible values. The paper therefore reports faithfulness as a precondition for the mechanism of H2, which is satisfied, and not as a property of the valuations, which the accuracy and calibration results of this and the two preceding sections measure directly. The counterfactual-stability and human-rated usefulness assessments listed in the research plan are

NOT YET ESTIMATED.

11. Identifying Fundamental Value: Repeat Sales, Rents and User Cost

The accuracy results of Sections 8 and 9 measure the distance between an arm's estimate and the realised transaction price. Lemma 1 states exactly what such a distance identifies: the ranking of estimators by price-based error is their ranking by distance to the selected market value $\tilde{v}$, and nothing in that ranking speaks to the latent market-consistent value v^* or to the structural fundamental value V^F of Section 3.1. A realised price differs from $\tilde{v}$ by transaction noise, which collects the bargaining outcome, the seller's urgency, the time the dwelling spent on the market and the idiosyncratic match between buyer and property; $\tilde{v}$ differs from v^* by a selection term that is non-zero whenever the decision to sell is correlated with the unobserved component of the price (Genesove and Mayer, 2001); and v^* differs from $\log V^F$ by expectational and liquidity components that can be large when down-payment constraints bind (Stein, 1995) or when expectations extrapolate recent growth (Glaeser and Nathanson, 2017). A method that forecasts prices well may therefore track a bubble as faithfully as it tracks fundamentals, and a method that is closer to fundamentals may be penalised by the price-based metric when the market departs from them. This section builds the three proxies announced in Section 3.6, documents the selection problem of each, reports the comparison of the statistical baselines against the sharpest proxy, the market-adjusted resale, and specifies how the same comparison will be read for the agentic arms once the evaluation partitions have had time to resell. Every number below is produced by `scripts/50_repeat_sales.py`, `51_eval_vs_resale.py` and `52_rents_usercost.py` from the snapshot of Section 5, and nothing in it is shown to any arm.

11.1. Three proxies and what selects into each

The first proxy is the subsequent resale of the same property. If parcel i sells at t_1 for P_{t_1} and again at $t_2 > t_1$ for P_{t_2}, and if the movement of the market between the two dates is common to the cell in which the property trades, then the repeat-sales logic of Bailey et al. (1963) and Case and Shiller (1987, 1989) brings the later price back to the first date,

$$\log V^R_{t_1} = \log P_{t_2} - \big[I(m,c,\tau_2) - I(m,c,\tau_1)\big], \tag{14}$$

where $I(m,c,\tau)$ is a log-price index for municipality m, property class c and month τ. Under unchanged quality, $\log V^R_{t_1} - \log P_{t_1} = \varepsilon_{t_2} - \varepsilon_{t_1}$ is the difference of two transaction noises, so that an estimate formed at t_1 which lies closer to $V^R_{t_1}$ than P_{t_1} does has captured a component of $v^*_{t_1}$ that the first transaction's noise obscured. Its selection problem is that properties which resell within a

six-year window are not a random draw: flips, relocations, forced sales and renovated dwellings are over-represented, and the loss-averse listing behaviour documented by Genesove and Mayer (2001) ties the decision to resell to the sign of ε_{t_1}. Renovations are unobserved, the roll being a single snapshot shared by both legs, and the holding period is right-censored by the snapshot date. We flag pairs held less than twelve months, report the proxy by holding period, and never correct for selection.

The second proxy is the rent. Where a comparable unit commands a market rent R_t, the user-cost condition (1) of Poterba (1984), in the form used by Himmelberg et al. (2005), delivers $V_t^F = R_t/u_t$ with $u_t = i_t + \tau_p + \delta + \gamma_t - \pi_t^e$. Three selection problems stand between this identity and the data. The rental stock differs from the owner-occupied stock in size, quality and location, so that the rent of an average registered apartment is not the rent of the condominium or plex that sold. Quebec rent regulation detaches contract rents from market rents for sitting tenants and compresses the rents of new leases toward those of the previous occupant. And u_t contains the expected appreciation π_t^e, which is unobservable, so that the diagnostic of Himmelberg et al. (2005) identifies V^F only under a maintained assumption about expectations; following Campbell et al. (2009), who show that movements in rent–price ratios are dominated by the expected-return and expected-growth components rather than by rents, we report the implied value under a range of π_t^e and read its time profile and dispersion, not its level.

The third proxy, an independent professional appraisal, is an $\mathcal{I}_t$-measurable estimate like any other arm and is subject to Lemma 1; appraisers update toward observed prices (Quan and Quigley, 1991) and confirm the contract price in the large majority of mortgage transactions (Eriksen et al., 2019; Calem et al., 2021), so that even where accessible they would proxy $\tilde{v}$ at best. None is accessible for this study and none is used.

11.2. The repeat-sales deflator and the resale sample

The deflator is deliberately transparent. For every municipality, class and calendar month we take the median of the log price per square metre of floor area over the sales of the cell; cells with fewer than five sales fall back to the region-group by class cell; each series is smoothed by a centred three-month rolling median computed on valid months only; and the index is chained, the log index being the cumulative sum of month-to-month changes of the smoothed series, each change taken from the municipality series when both months are valid there and from the region series otherwise, so that a switch of source never produces a jump in level. On the 2021–2026 snapshot this yields 2,283 municipality-by-class series over 70 months, of which 16,003 monthly changes come from the municipality and 141,524 from the region: at the monthly frequency most Quebec municipalities are thin, and the deflator is in practice a regional index with municipal corrections in the large cities.

Pairs are formed from consecutive sales of the same roll unit in the full snapshot of 769,995

deeds rather than in the headline sample alone, after removing deeds that cannot serve as a leg (duplicates, shared and multi-parcel deeds, land, non-residential and built-after-sale rows) and same-date pairs; the later leg keeps its data-quality flags. The resulting sample, described in Table 24, contains 92,089 headline sales with a later resale, that is 14.6 per cent of the analytic table: 30,900 condominiums, 50,178 houses and 11,011 plexes. The median holding period is 23.5 months and 28.8 per cent of the pairs are held less than twelve months; among those held less than a month, one in nine resells at an identical price, which marks re-registrations rather than resales. The median annualised log return is 8.0 per cent against a median index change of 0.124 over the holding period. The gap $\log V^R - \log P$ has a mean of $+0.015$, so the deflator removes the market movement without a systematic residual, but a standard deviation of 0.438 (0.30 for condominiums, 0.49 for houses, 0.50 for plexes) and a median absolute value of 0.128. Figure 10 shows that the dispersion is essentially flat in the holding period: it is the noise floor of the proxy, the sum of two transaction noises, unobserved changes to the dwelling and index error, and it is of the same order as the price-based errors of the arms. A resale-based contrast is therefore informative only with samples of hundreds of pairs.

The censoring of holding periods by the snapshot date has a direct consequence for the evaluation design. First sales in the test partition (April to December 2025) have at most eighteen months in which to resell before the snapshot of October 2026, and 80 per cent of their pairs are short holds; the 2026 holdout has none that is not. After excluding short holds and flagged legs, 729 of the 96,935 test-partition sales carry a usable V^R, which suffices for the baselines evaluated on the full partition, but only 6 of the 600 test subsample properties and none of the 300 holdout properties on which the language-model arms are evaluated do. The resale check of the agentic arms is thus a deferred test: it is specified now, its code is frozen with the protocol, and it will be run on the same predictions once a later snapshot has been fetched.

11.3. Price-to-rent and user-cost diagnostics

The rent data are a sample of 57,433 residential leases with monthly rent, number of rooms, lease year and coordinates, obtained through the Immo-Ka platform and described as leases registered with the Tribunal administratif du logement. Their limits bound what the diagnostic can say. The sample is self-selected and concentrated in 2022–2025; it records no floor area, so that no bridge from rooms to the square metres of the sales table exists; the rooms field is an undocumented integer that we read as bedrooms; the lease start date is missing for four in ten rows, so leases are placed in calendar years; and the raw table contains addresses outside Quebec. After removing implausible rents and years and exact duplicates, 56,305 leases remain, of which 49,082 are matched to a municipality by the nearest analytic sale within two kilometres (32,232 in the Montréal area, 5,359 in the Québec area, 1,127 in the Outaouais and 10,364 elsewhere) and 48,188 to a 2021

dissemination area. Median rents are then computed by municipality, year and rooms bucket.

Because no area bridge exists, the price-to-rent comparison is deliberately crude: for each condominium and plex sale we take the median monthly rent of all matched leases in the same municipality and year, annualise it, multiply it by the number of dwellings for plexes, and divide by the price. This gross rental yield matches 162,981 sales (103,553 condominiums and 59,428 plexes) to a municipality-year cell with at least ten leases. Its median is 3.6 per cent for condominiums and 5.6 per cent for plexes; in 2025 it ranges from 3.5 and 4.8 per cent in the Montréal area to 5.0 and 6.2 per cent in the Outaouais. Figure 11 plots the distributions by region group and year. Two features are robust to the crudeness of the measure. Yields are ordered across regions in the way that prices are ordered in reverse, which is what a common discount rate applied to heterogeneous rent levels predicts, and the plex yields compress from 2021 to 2024 while condominium yields are nearly flat, which reflects the faster appreciation of small income properties over the period. The level of the condominium yield, by contrast, is not interpretable, since the rent of an average registered apartment is compared with the price of a sold condominium.

The user-cost diagnostic divides the gross yield by u_t to obtain the rent-implied value relative to the price, $(R_t/u_t)/P_t$. Table 25 lists the inputs. The mortgage rate is the only measured one: the Bank of Canada's weekly series for the five-year conventional mortgage rate, averaged by calendar year, rises from 4.8 per cent in 2021 to 6.7 per cent in 2023–2024 and eases to 6.1 per cent in 2026. Property tax, depreciation and the risk premium are assumptions in the range used by Himmelberg et al. (2005), and expected appreciation is bracketed by three scenarios. Figure 12 is accordingly labelled illustrative. Under the central scenario the median rent-implied value is 0.34 to 0.46 of the price for condominiums and 0.54 to 0.75 for plexes; under four per cent expected appreciation the plex ratio reaches 0.99 in 2021 and 0.70 in 2023–2024. We do not read these levels as evidence of over-valuation, because the quality mismatch and the regulation of contract rents both push the measured yield below the yield of the sold unit. What the figure does show is the mechanism that separates price from fundamental value in this period: the rise of the mortgage rate between 2021 and 2023 raised the user cost by about two percentage points while gross yields barely moved, so that the price-to-rent ratio could remain where it was only if expected appreciation absorbed the change. An arm that forecasts 2025 prices accurately is implicitly forecasting those expectations; whether it has identified the fundamental is a different question, and one that the resale proxy, not the price error, is designed to answer.

11.4. How the arm comparison against resale-implied values will be read

For every evaluated property with a usable resale, an arm's estimate $\hat{y}$ has two errors, $e^P = \log\hat{y} - \log P_{t_1}$ against the realised price and $e^R = \log\hat{y} - \log V^R_{t_1}$ against the resale-implied value, and the realised price itself has $e^R_P = \log P_{t_1} - \log V^R_{t_1}$. The quantity of interest is the paired contrast

$\Delta = \overline{|e^R|}(\text{arm}) - \overline{|e^R|}(\text{price})$, estimated by the block bootstrap of Section 7 with municipality-by-month blocks and Holm adjustment within the family, together with the pre-registered arm pairs evaluated against V^R rather than against P. A negative Δ means that the arm is closer to the market-adjusted resale than the first transaction price was, which under the maintained assumptions is evidence that the arm captured a component of v^* obscured by transaction noise; a Δ near zero means that the arm's error is of the same nature as the price's own noise; a positive Δ means that the arm is further from the resale than the transaction it was asked to predict, which is the expected outcome for an estimator that fits the price well, since by Lemma 1 price-based training rewards proximity to $\tilde{v}$, not to V^R. The arm-versus-arm contrasts are read in the same way as in Section 8, with two caveats that the sample size imposes: the noise floor of the proxy is a log standard deviation of about 0.44, and the selection of resellers is not corrected, so that a contrast is reported with its holding-period composition. Table 26 reports the result for the statistical baselines of Section 8, which are evaluated on the full test partition and therefore on the 729 test sales with a usable resale (399 municipality-by-month blocks; the spatial-holdout variants on 63). The language-model arms are not yet in the table: no live run exists on the test or holdout partitions at the time of writing, and when the runs exist their resale subsample will, for the reasons given above, remain too small until a later snapshot is fetched.

Three features of the baseline result deserve notice, each read with the selection caveat in mind. First, the realised price is itself 0.073 log points below the resale-implied value on average in this subsample, with a mean absolute distance of 0.243: the properties that resold within twelve to eighteen months were, on average, bought below what their resale implies, which is the signature of the flip and renovation selection discussed in Section 11.1 and not a property of the market. Second, the gradient-boosting arms are measurably closer to the resale-implied value than the first transaction price was, with Δ between -0.028 and -0.038 and Holm-adjusted p-values below 0.05, while their error against the realised price on the same properties is 0.18; the hedonic baseline is no closer than the price ($\Delta = 0.006$, $p = 1$) and the hierarchical baseline is further from it ($\Delta = 0.030$). The ordering of the baselines against V^R is therefore the same as their ordering against P, which is what Lemma 1 predicts when the arms differ mainly in their distance to $\tilde{v}$, and the gain of the boosting arms over the price is of the order of the price's own selection bias, so that it is as consistent with the arms being unbiased for the resellers as with their having identified a component of value. Third, the bias against V^R is close to zero for the boosting arms and positive for the hedonic and hierarchical ones, so that on this subsample the proxy separates estimators by their bias rather than by their dispersion. None of this bears on the agentic arms, whose test will be run on the same footing when the data allow it.

The conclusion of this section is methodological. Forecasting accuracy and the identification of fundamentals are distinct properties of a valuation method, measured against distinct targets with

distinct noise, and the paper keeps them separate. An arm that wins the price-based comparison of Section 8 has demonstrated proximity to the selected market value; whether it has also demonstrated proximity to the economic fundamental is decided by the resale contrast of Table 26, available for the statistical baselines and deferred for the agentic arms, and by nothing else in the paper. The rent-based diagnostics bound the gap between the two targets over the sample period but do not close it, and we do not claim that any arm discovers fundamental value until a test that could have refuted the claim has been run.

12. Economic Welfare and Market Design

The accuracy metrics of the preceding sections measure the distance between an estimate and the selected market value. Whether a more accurate valuation is worth having depends on how valuations enter decisions, and the mapping is not monotone: a valuation that is unbiased on average can still ration credit, shift bargaining surplus, and, when it is shared by every lender, concentrate losses in the same states of the world. This section sets out a deliberately stylised model of one transaction with a buyer, a seller and a lender, derives three comparative statics that hold in general, and illustrates them numerically. Only the loan-to-value cap and the form of the haircut rule are transparent calibrations; every other parameter is illustrative and the numbers that follow should be read as orders of magnitude under stated assumptions, not as estimates.

12.1. The model

Assumption 3 (Transaction stage)**.** All quantities are expressed as fractions of the market-consistent value V of the dwelling, normalised to one. The seller's reservation value is $S = 1 + u_S$ and the buyer's willingness to pay is $B = 1 + u_B$, with (u_S, u_B) drawn from a continuous joint distribution; a match is mutually beneficial when $B \geq S$. The lender observes an appraisal $\hat{A} = 1 + b + sZ$, $Z \sim N(0,1)$ independent of (u_S, u_B), with bias b and standard deviation s, and lends $L = \ell(1-h)\hat{A}$, where ℓ is the loan-to-value cap and $h = h_0 + h_1 s$ is a haircut that may rise with valuation uncertainty. The buyer holds cash $W = (1-\ell)\kappa$ with $\kappa \geq 1$. The transaction completes if and only if $B \geq S$ and $S \leq W + L$, at the price $P = \min\{S + \theta(B-S), W + L\}$, where $\theta \in (0,1)$ is the seller's bargaining share.

The structure follows the down-payment mechanism of Stein (1995): the buyer's capacity to pay is the sum of own cash and what the lender will advance against the collateral, and the lender's advance is tied to a valuation rather than to the negotiated price. The financing cap $W + L$ truncates the price from above; when it binds, the match either fails or completes at a price below the bargaining outcome, which transfers surplus from the seller to the buyer. The collateral channel through which valuations propagate to credit is the one formalised by Kiyotaki and Moore (1997).

Assumption 4 (Default stage)**.** The loan defaults with exogenous probability p_d. At default the lender recovers $(1-c_\ell)V_1$, where c_ℓ is a liquidation cost and $V_1 = 1+g$ with $g \sim N(\mu_g, \sigma_g^2)$ independent of Z; the loss is $\max\{L-(1-c_\ell)V_1, 0\}$ and the expected loss at origination is $\mathrm{EL}(b,s) = p_d\,\mathbb{E}[\max\{L-(1-c_\ell)V_1, 0\}]$. There are N lenders whose appraisal errors are exchangeable with correlation ρ, $Z_n = \sqrt{\rho}\,C + \sqrt{1-\rho}\,E_n$, where $C, E_1, \ldots, E_N$ are independent standard normal; ρ measures common-model dependence across lenders' valuation models.

The independence of V_1 from Z is restrictive: it rules out the case in which the valuation error is itself informative about the subsequent price path, as when a model overvalues precisely the dwellings that later underperform. The results below are therefore lower bounds on the cost of correlated mispricing in that sense.

12.2. Three comparative statics

Proposition 9 (Credit rationing)**.** *Conditional on the seller's draw, let $a(u_S) = (1+u_S) - (1-\ell)\kappa - \ell(1-h)(1+b)$ denote the financing gap at the error-free appraisal, and suppose financing is needed, $(1+u_S) > (1-\ell)\kappa$. The probability that a mutually beneficial match fails for lack of financing is*

$$F(u_S; b, h, s) = \Phi\left(\frac{a(u_S)}{\ell(1-h)s}\right). \tag{15}$$

(i) F is strictly decreasing in b and strictly increasing in h. (ii) If $a(u_S) < 0$, the match is financeable at the error-free appraisal and F is strictly increasing in s; if $a(u_S) > 0$, F is strictly decreasing in s. (iii) The failure rate among beneficial matches is increasing in s whenever $a(u_S) < 0$ almost surely on $\{B \geq S\}$; in general it is the sum of a rationing effect on financeable matches and a lucky-financing effect on non-financeable ones. (iv) A haircut rule with $h_1 > 0$ reinforces the rationing effect.

The proposition separates two things that discussions of valuation accuracy often conflate. For the matches that an accurate valuation would finance, which are the bulk of beneficial matches under any reasonable distribution of reservation values, valuation noise is a pure cost: each unit of s raises the chance that the appraisal falls below what the buyer needs. For matches that an accurate valuation would not finance, noise occasionally lets the transaction through; whether that is a welfare gain depends on whether the lender's rule was right to exclude them, which the model does not adjudicate. The direction of bias matters in the opposite way to its effect on the lender: undervaluation rations credit, overvaluation expands it.

Proposition 10 (Lender's expected loss)**.** *Hold the haircut h fixed. (i)* EL *is strictly increasing in the bias b. (ii)* EL *is strictly increasing in the standard deviation s: a mean-preserving spread of the appraisal raises the expected loss because the loss is a strictly convex function of the loan amount.*

(iii) Under the haircut rule $h = h_0 + h_1 s$, the loan's mean falls with s and the sign of $\partial \mathrm{EL}/\partial s$ is not determined; for every s there is a haircut that holds EL *at its error-free level, and by Proposition 9(i) it raises the failure probability of every financeable match.*

Part (iii) is the sense in which valuation uncertainty has a price that someone must pay. A lender that targets a constant expected loss responds to a noisier valuation technology by widening the haircut; the cost of the noise is then borne by borrowers in the form of rationed credit rather than by the lender in the form of losses. The empirical interval widths of Section 8 map directly into s for a lender that uses the committee's interval as its measure of uncertainty, and the coverage results determine whether that mapping is honest.

Proposition 11 (Common-model dependence across lenders)**.** *Let $\ell_n = g(Z_n)$ be lender n's expected loss conditional on its appraisal, $g(z) = p_d\,\mathbb{E}[\max\{\ell(1-h)(1+b+sz) - (1-c_\ell)V_1, 0\}]$. (i) The expected system loss $\mathbb{E}[\sum_n \ell_n] = N\mathbb{E}[g(Z)]$ does not depend on ρ. (ii) The variance of the system loss, $N\,\mathrm{Var}(g(Z)) + N(N-1)\,\mathrm{Cov}(g(Z_1), g(Z_2))$, is strictly increasing in ρ on $[0,1]$, rising from $N\,\mathrm{Var}(g(Z))$ at $\rho = 0$ to $N^2\,\mathrm{Var}(g(Z))$ at $\rho = 1$. (iii) Replacing N distinct models with loss variance v_0 and loss correlation $\tilde{\rho}_0$ by a single common model with loss variance $v_1 < v_0$ raises the variance of the system loss if and only if $v_1/v_0 > [1 + (N-1)\tilde{\rho}_0]/N$.*

Part (i) is the reason that the welfare case for accurate valuation cannot be made at the level of expected losses alone: correlation does not change what lenders lose on average, it changes when they lose it. Part (ii) is the collateral-market counterpart of Proposition 3; it formalises the monoculture risk discussed by Kleinberg and Raghavan (2021) and Bommasani et al. (2022) for a market in which lenders converge on the same valuation model. Part (iii) is the trade-off that a market with a single dominant provider faces: a common model that is more accurate for each lender can nevertheless raise the dispersion of aggregate losses, and the condition shows that the threshold is easy to cross when N is large and the existing models are only moderately correlated. Within the paper's design, the provider-balanced committees are the within-committee analogue of this proposition; the market-level implication is that the diversity that helps a committee is also what a lending system loses when every participant adopts the best available committee.

12.3. Illustrative simulation

Figure 13 and Figure 14 illustrate the three propositions with $\ell = 0.80$, the conventional cap for an uninsured mortgage in Canada, two haircut rules ($h_1 = 0$ and $h_1 = 0.5$), and illustrative values for everything else: $\kappa = 1.25$, $\theta = 0.5$, $u_S \sim N(-0.05, 0.05^2)$, $u_B \sim N(0.05, 0.08^2)$, $p_d = 0.02$, $c_\ell = 0.15$, $g \sim N(-0.10, 0.10^2)$ and $N = 50$ lenders. These values are not estimated from Quebec data; they are chosen so that most beneficial matches are financeable at the error-free

appraisal, which is the regime in which Proposition 9(iii) applies. All quantities are produced by `scripts/40_welfare_sim.py` and stored in `results/theory/welfare_sim.csv`.

In the left panel of Figure 13, with no bias and no uncertainty haircut, the failure rate rises from about 1 percent at $s = 0$ to 4 percent at $s = 0.05$, 13 percent at $s = 0.10$ and 26 percent at $s = 0.20$; with the haircut rule the same path runs through 8, 23 and 43 percent. A negative bias of three percent shifts the curve up and a positive bias shifts it down, as Proposition 9(i) states. Among completed transactions the financing cap also moves the split of surplus: the seller's average surplus falls and the buyer's rises as s grows, because a binding cap truncates the price from above. The lender's expected loss, with the haircut fixed, rises from 0.109 to 0.182 percent of V between $s = 0$ and $s = 0.20$ (Proposition 10(ii)); with $h = 0.5s$ it falls below its error-free value, which is the transfer of the cost of uncertainty from the lender to rationed borrowers described in part (iii). In Figure 14 the mean system loss is flat in ρ while its standard deviation at $s = 0.10$ rises sevenfold between $\rho = 0$ and $\rho = 1$ and the expected shortfall rises from 0.082 to 0.202. The concentration case of Proposition 11(iii) is computed in the same file: moving from fifty lenders with $s = 0.15$ and $\rho = 0.2$ to a single shared model with $s = 0.10$ and $\rho = 1$ lowers the expected system loss by 15 percent and raises its standard deviation by 53 percent and its expected shortfall by 26 percent.

12.4. What accuracy does and does not establish

The empirical sections ask whether a committee of agents reduces the error s and the bias b of a valuation relative to the alternatives. This section shows that the welfare consequences of such a reduction are mediated by three objects that the accuracy metrics do not contain: the distribution of reservation values and cash, which determines how many beneficial matches sit on the financing margin; the lender's haircut rule, which determines who bears the cost of uncertainty; and the correlation of valuation errors across lenders, which determines the dispersion of aggregate losses without affecting their mean. A reduction in s that is achieved by every lender adopting the same committee lowers the failure rate and the per-lender expected loss while raising ρ, and Proposition 11(iii) shows that the net effect on the dispersion of system losses can have either sign. Predictive accuracy is therefore necessary for a welfare gain in this model but not sufficient for one, and the paper does not claim a welfare gain on the strength of the accuracy results. What it can claim, if the empirical results support it, is a measured reduction in s and b under a stated information set, together with the mapping above from which a reader with a view on the distribution of matches, the haircut rule and the market structure can draw the welfare implication.

13. Limitations

The limitations of the study are of three kinds: those of the sample on which the agentic arms were run, those of the evidence that any arm could see, and those of the agents and the protocol themselves. We take them in that order, state what each implies for the reading of the results, and record what has been done or could be done about it. The register from which this section is drawn is a living document in the repository.

The first limitation is statistical power. The preregistered test subsample has six hundred properties; the version of the experiment reported here values the first eighty of them at the mid tier for the six main arms, the first three hundred at the small tier for three arms, the first sixty of the holdout subsample and forty prospective sales, under a self-imposed envelope of roughly one thousand US dollars at list prices. The paired design and the low variance of the within-property differences make the contrasts between agentic arms far more precise than their sample size suggests, but the precision has limits that the intervals state exactly. Against the independent committee the two-round debate's 95 percent interval of $[-0.006, 0.005]$ log points excludes relative improvements larger than about four percent of the mean absolute log error; against the placebo committee the interval of $[-0.005, 0.006]$ excludes about three percent; against the cost-matched single agent the interval of $[-0.017, 0.004]$ excludes about seven percent; and against the automated valuation model, whose errors are not correlated with the committee's in the way the agentic arms' errors are with one another, the interval of $[-0.033, 0.026]$ excludes only improvements of about twenty percent. The preregistration specified that the claim of deliberation adding information would be falsified if the contrasts against the independent committee and the compute-matched agent both failed to reject with intervals excluding a two percent relative improvement; the first interval reaches four percent and the second seven, so that the strict criterion is not met. What the data establish is a point estimate at zero with that resolution, repeated in the validation, holdout and prospective partitions with the same sign or the opposite one. Because the subsamples are nested, the first eighty properties of the six hundred in a fixed stratified order, the runs can be resumed to the full preregistered size without redrawing; at six hundred properties the interval against the independent committee would shrink by a factor of about 2.7, to roughly plus or minus 0.002 log points, which is the resolution the falsification criterion requires. The stratified results by region, property class and price tercile are reported only for cells with thirty or more properties and, for the agentic arms, exist for few of them; nothing in the paper should be read as a claim about thin markets or any other stratum for those arms.

The second limitation is the target. All accuracy metrics measure the distance between an estimate and a realised transaction price, which Lemma 1 shows to be the distance to the selected market value plus a common noise term. A price is not a value: it embeds the bargaining outcome,

the seller's urgency, the marketing history and the appraisal that financed the mortgage, and it is observed only for properties that sold. The repeat-sales proxy of Section 11 is the only check on the selected target that the data allow, and it is available for the statistical baselines but not yet for the agentic arms, because the evaluated properties of 2025 and 2026 have had too little time to resell; six of the six hundred test properties and none of the holdout properties carry a usable resale, and the test is deferred rather than reported on that basis. The rent and user-cost diagnostics bound the gap between price and fundamental over the sample period but, for want of an area bridge between leases and sales, cannot close it. The title's question is therefore answered only in part, and the part that concerns fundamental value is answered for the tabular baselines alone.

The third limitation is the evidence. No arm observes the condition of the dwelling: there is no inspection report, no photograph, no description of finishes or renovations, and the building-condition role is instructed to say so and to keep its interval wide. Every arm is therefore blind to the component of value that an appraiser visiting the property would see, and the comparison between arms is a comparison of reasoning on identical, incomplete evidence, not of valuation as practised. Listing data were audited and excluded by design, because their coverage begins after the test window and because asking prices would be observable for some subjects and not for others; the arms consequently see no asking price, no days on market and no listing text, information that the literature shows to be valuable and that a commercial deployment would have. The mortgage-rate field of the market context is null in this snapshot and the agents are told so. The roll's land and building split is an as-of approximation rescaled from the current roll, and the roll's reference date is inferred from the triennial cycle rather than recorded; both are disclosed to the agents with their status. More generally, every role's view is a subset of the same four sections of one packet, so the union of the views is small and the scope that Proposition 6 leaves for evidence exchange to help is correspondingly small. The null on deliberation is a null for this information set; a richer or more specialised evidence base is the most plausible route to a different result.

The fourth limitation concerns the providers. The two models of each tier are matched on their published list prices per token, not on the compute they consume, and identical prices are a vendor decision rather than a measurement; the token counts of every call are logged and the cost frontier is reported, but the compute matching of arm J to arm F, and of the debate to either, is approximate. Reasoning was left to each vendor's adaptive mode at a medium effort setting and no sampling parameters were sent, because the current models of one vendor reject them; the variation between repeated calls on which arm J relies is therefore the vendor's default, not a controlled temperature. Hosted models are not deterministic and the paper does not claim bitwise reproducibility of any agentic result: the repeat of eight test properties under the frozen two-round protocol gave a mean absolute difference of 0.007 log points in the committee median, a maximum of 0.019, a correlation of 0.998 between the two runs' absolute errors and every repeated point inside the original 90

percent interval, which is small relative to the contrasts but is measured on eight properties. The models themselves are fixed by identifier at the freeze and will be deprecated by their vendors; the archived transcripts, not the live endpoints, are the reproducible object.

The fifth limitation is the protocol. The role prompts, the placebo prompt, the packet layout, the anonymised committee table, the median anchor and the revision instruction were each written once, exercised on a pilot of twelve validation properties for the three main arms, and frozen; no sensitivity analysis to prompt wording, to the order of sections in the packet, to the number or identity of comparables shown, or to the choice of the median rather than the mean as the committee statistic was performed, and the paper cannot say how much of any arm's accuracy is attributable to these choices. The twenty-role extension and the shuffled-prompt placebo exist in the code but were not run live. The rounds grid stops at four, and the four-round arm was run on twenty validation properties only. The stopping rule on the interquartile range, the verification thresholds of arm H and the trimming and robust-pool parameters are reasonable defaults rather than tuned values, and were in any case chosen on validation. Because the single design shares one information set across arms, the paper cannot separate what the agents would do with privileged evidence from what they do with the common packet; that is a feature of the identification strategy and a limit on its external validity.

The sixth limitation is the human arm. No licensed or trained appraiser valued any packet; the arm requires a research-ethics certificate from the Université du Québec en Outaouais that has not yet been obtained, and no human benchmark is reported, approximated or simulated. The comparison of the committee with professional judgement on identical information, and the evaluation of a human-plus-committee workflow, remain to be done. Two further points belong in the record. The 2026 holdout and the prospective set are each evaluated once, as the protocol requires, but the prospective set covers one week of registrations and forty agentic subjects and is a consistency check rather than an endpoint; and the upward bias of every arm on that set, the tabular baselines included, may reflect the selection of deeds that register quickly rather than a property of the valuations. Finally, the welfare model of Section 12 is illustrative and calibrated only in its loan-to-value cap; it yields comparative statics, not estimates, and the paper makes no causal welfare claim on the strength of the accuracy results.

14. Conclusion

This paper asked whether communication among heterogeneous autonomous valuation agents adds measurable information about the value of a dwelling, and it answered with a preregistered experiment on registered Quebec transactions in which every rival explanation for a deliberation effect was given its own control arm. The answer, on the evidence of this version of the study, is no.

Twelve role-specialised agents that read least-privilege views of the same timestamped evidence packet and never communicate reach the same accuracy as the same agents after two rounds of critique and revision, as a committee with the mandates removed, and, within a wider interval, as one agent given the same tokens; the four preregistered contrasts have point estimates within 0.006 log points of zero and Holm-adjusted p-values of one, and in the 2026 holdout and the prospective set the debated committee is marginally worse than the independent one. What deliberation does, reliably and in every run, is manufacture agreement: the dispersion of the agents' estimates halves in the first round and keeps shrinking thereafter while the error of their median does not move, which is what the theory of Section 3 says must happen when agents revise toward a committee statistic rather than toward new evidence. The committees themselves are competent price forecasters, indistinguishable from a tuned automated valuation model on the test subsample and better on the median error, but they state their uncertainty too narrowly, their small-model versions markedly so, and when the market moved in 2026 all of them moved with it in the same direction, overvaluing by ten log points where the tabular model was unbiased. None of this is evidence that the agents have discovered fundamental value. By Lemma 1 the price-based metrics rank estimators by their distance to the selected market value, and the only proxy for value that is independent of the first transaction, the market-adjusted resale, is not yet available for enough of the evaluated properties to say anything about the agentic arms.

For lenders, appraisers and the providers of valuation models the findings carry three practical messages. The first concerns common-model risk. A committee whose members are drawn from two foundation models inherits the errors those models share, and Proposition 3 shows that no committee size removes them; the 2026 drift is a direct illustration, since every agent, whatever its role or provider, overvalued the same properties in the same direction. A lending system in which every participant consults the same committee therefore gains little in expected accuracy over a good tabular model and loses the diversification that independent valuation errors provide, and Proposition 11 shows that the cost appears in the dispersion of aggregate losses rather than in their mean. The second concerns verification. The deterministic checks of arm H, which flag citations to evidence not in the packet, adjustments far from the raw comparable price and points outside the hull of the comparables, cost almost nothing, catch the failures that would be most embarrassing in a lending file, and have no measurable effect on accuracy; they are a governance tool, not a performance tool, and should be adopted as such. The third concerns calibration. The intervals that agents state are narrower than the evidence warrants and the agreement of a committee after deliberation is not a measure of confidence; a lender that used either as its haircut rule would be systematically under-protected, and a split-conformal rescaling estimated on held-out data, which the paper reports as a remedy, is the minimum that responsible use requires.

Several changes to the design could overturn the conclusion, and we state them so that the null

is read as a statement about this experiment rather than about deliberation in general. The sample is small: eighty properties at the mid tier rule out relative improvements larger than about four percent against the independent committee and about seven percent against the compute-matched agent, and the preregistered six hundred, which the nested design allows to be reached by resuming the same runs, would bring the resolution to the two percent threshold that the preregistration set as the falsification criterion. The evidence is thin: the packet contains the roll attributes, twelve comparables, a market summary and a geographic context, and every role sees a subset of the same four sections, so that the union of the views is small and, by Proposition 6, the scope for evidence exchange to help is correspondingly small. Richer and more specialised evidence, listing text and photographs of the kind that Nowak and Smith (2017), Zhang et al. (2024) and Faxvaag et al. (2025) show to be informative, inspection reports, permit histories, or tools that let an agent query the register rather than read a fixed excerpt, would enlarge the union of the views and could create the complementarity that the placebo contrast found absent. The agents are general-purpose models prompted once; fine-tuning on Quebec transactions in the manner of Vedula et al. (2025), or a calibration layer trained on the committee's own validation errors, would address the interval problem directly and might also change the round-0 accuracy on which everything else is built. And the deliberation protocol is one of many: the anonymised table and median anchor of arm G are the natural first design, but a protocol that forced agents to exchange evidence rather than opinions, or that weighted members by demonstrated precision as Golub and Jackson (2010) requires, could in principle move the pooled error, although Proposition 5 bounds what any linear scheme can achieve.

The human arm remains to be run. A panel of chartered appraisers valuing the same packets under the same cutoff, with prices blinded and confidence intervals recorded, is the comparison that practitioners will ask for first and the one that would settle whether the committee's accuracy is competitive with professional judgement on identical information; it requires a research-ethics certificate that has not yet been obtained, and the paper reports it as pending rather than simulating it. Until that arm exists, and until the evaluated properties have resold, the paper's claims are deliberately limited: a deliberating committee of language-model agents predicts Quebec transaction prices about as well as an automated valuation model, no better than the same agents kept silent, and with intervals that should not be trusted at face value. That is a smaller result than the question in the title invites, but it is the one the data support, and the artefacts released with the paper are intended to let it be revised rather than repeated.

References

Acemoglu, Daron, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar, "Bayesian Learning in Social Networks," *The Review of Economic Studies*, 2011, *78* (4), 1201–1236.

Agarwal, Sumit, Brent W. Ambrose, and Vincent W. Yao, "Can regulation de-bias appraisers?," *Journal of Financial Intermediation*, 2020, *44*, 100827.

_ , Itzhak Ben-David, and Vincent Yao, "Collateral Valuation and Borrower Financial Constraints: Evidence from the Residential Real Estate Market," *Management Science*, 2015, *61* (9), 2220–2240.

Angelopoulos, Anastasios N. and Stephen Bates, "Conformal Prediction: A Gentle Introduction," *Foundations and Trends in Machine Learning*, 2023, *16* (4), 494–591.

Bae, Seongeun, Leehyun Jung, Sukyung Nam, Sihyun An, and Kwangwon Ahn, "Housing Price Estimation and Reasoning Based on a Large Language Model," in "Blockchain Technologies," Springer Nature Singapore, 2025, pp. 27–42.

Bailey, Martin J., Richard F. Muth, and Hugh O. Nourse, "A Regression Method for Real Estate Price Index Construction," *Journal of the American Statistical Association*, 1963, *58* (304), 933–942.

Bates, J. M. and C. W. J. Granger, "The Combination of Forecasts," *Journal of the Operational Research Society*, 1969, *20* (4), 451–468.

Becker, Joshua, Devon Brackbill, and Damon Centola, "Network dynamics of social influence in the wisdom of crowds," *Proceedings of the National Academy of Sciences*, 2017, *114* (26), E5070–E5076.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch, "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 1992, *100* (5), 992–1026.

Bogin, Alexander N. and Jessica Shui, "Appraisal Accuracy and Automated Valuation Models in Rural Areas," *The Journal of Real Estate Finance and Economics*, 2020, *60* (1-2), 40–52.

Bommasani, Rishi, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang, "Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?," 2022. NeurIPS 2022. arXiv:2211.13972.

Boucher, Simon-Pierre, "SHARP-HPI: A Spatially Hierarchical, Adaptive and Robust Framework for Quality-Adjusted Housing Price Indices. Monthly Municipal Indices for Quebec from 770,000 Registered Sales, 2021–2026," UQO Working Paper 22, Université du Québec en Outaouais, Département des sciences administratives October 2026.

Brown, Bradley, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini, "Large Language Monkeys: Scaling Inference Compute with Repeated Sampling," 2024. arXiv:2407.21787.

Calem, Paul, Jeanna Kenney, Lauren Lambie-Hanson, and Leonard Nakamura, "Appraising home purchase appraisals," *Real Estate Economics*, 2021, *49* (S1), 134–168.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Campbell, Sean D., Morris A. Davis, Joshua Gallin, and Robert F. Martin, "What moves housing markets: A variance decomposition of the rent–price ratio," *Journal of Urban Economics*, 2009, *66* (2), 90–102.

Case, Karl E. and Robert J. Shiller, "Prices of Single-Family Homes since 1970: New Indexes for Four Cities," *New England Economic Review*, September/October 1987, pp. 45–56. Accessed 2026-10-27.

_ and _ , "The Efficiency of the Market for Single-Family Homes," *American Economic Review*, 1989, *79* (1), 125–137. Accessed 2026-10-27.

Chan, Chi-Min, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu, "ChatEval: Towards Better LLM-Based Evaluators through Multi-Agent Debate," 2023. ICLR 2024. arXiv:2308.07201.

Chen, Lingjiao, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou, "Are More LLM Calls All You Need? Towards Scaling Laws of Compound Inference Systems," 2024. arXiv:2403.02419.

Chen, Tianqi and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," in "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" ACM 2016, pp. 785–794.

Chen, Tingting and Shijing Si, "Predicting Rental Price of Lane Houses in Shanghai with Machine Learning Methods and Large Language Models," 2024. arXiv:2405.17505.

Cho, Man and Isaac F. Megbolugbe, "An empirical analysis of property appraisal and mortgage redlining," *The Journal of Real Estate Finance and Economics*, 1996, *13* (1), 45–55.

Clemen, Robert T., "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 1989, *5* (4), 559–583.

_ **and Robert L. Winkler**, "Limits for the Precision and Value of Information from Dependent Sources," *Operations Research*, 1985, *33* (2), 427–442.

Davis-Stober, Clintin P., David V. Budescu, Jason Dana, and Stephen B. Broomell, "When is a crowd wise?," *Decision*, 2014, *1* (2), 79–101.

DeGroot, Morris H., "Reaching a Consensus," *Journal of the American Statistical Association*, 1974, *69* (345), 118–121.

Diebold, Francis X. and Roberto S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 1995, *13* (3), 253–263.

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch, "Improving Factuality and Reasoning in Language Models through Multiagent Debate," 2023. ICML 2024. arXiv:2305.14325.

Dubé, Jean and Diègo Legros, "A Spatiotemporal Solution for the Simultaneous Sale Price and Time-on-the-Market Problem," *Real Estate Economics*, 2016, *44* (4), 846–877.

_ **,** _ **, Marius Thériault, and François Des Rosiers**, "A spatial Difference-in-Differences estimator to evaluate the effect of change in public mass transit systems on house prices," *Transportation Research Part B: Methodological*, 2014, *64*, 24–40.

_ **, Maha AbdelHalim, François Des Rosiers, and Nicolas Devaux**, "Do residential reconversions affect residential property values? An investigation based on Québec city (Canada)," *Journal of Housing and the Built Environment*, 2023, *38* (4), 2373–2397.

Eriksen, Michael D., Hamilton B. Fout, Mark Palim, and Eric Rosenblatt, "The influence of contract prices and relationships on appraisal bias," *Journal of Urban Economics*, 2019, *111*, 132–143.

_ **,** _ **,** _ **, and** _ , "Contract Price Confirmation Bias: Evidence from Repeat Appraisals," *The Journal of Real Estate Finance and Economics*, 2020, *60* (1-2), 77–98.

Estornell, Andrew and Yang Liu, "Multi-LLM Debate: Framework, Principals, and Interventions," in "Advances in Neural Information Processing Systems 37 (NeurIPS 2024)" 2024. Accessed 2026-10-27.

Faxvaag, Håvard, Ulf Jakob Flø Aarsnes, and Are Oust, "Incorporating Large Language Models in Automated Real Estate Valuation," *Journal of Real Estate Research*, 2025, *48* (3), 454–479.

Fout, Hamilton, Nuno Mota, and Eric Rosenblatt, "When Appraisers Go Low, Contracts Go Lower: The Impact of Expert Opinions on Transaction Prices," *The Journal of Real Estate Finance and Economics*, 2022, *65* (3), 451–491.

Freddie Mac, "Racial and Ethnic Valuation Gaps in Home Purchase Appraisals," Research Note, Freddie Mac Economic & Housing Research 2021. Accessed 2026-10-27.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance*, 2022, *77* (1), 5–47.

Galton, Francis, "Vox Populi," *Nature*, 1907, *75* (1949), 450–451.

Geerts, Margot, Manon Reusens, Bart Baesens, Seppe vanden Broucke, and Jochen De Weerdt, "On the Performance of LLMs for Real Estate Appraisal," 2025. ECML-PKDD 2025. arXiv:2506.11812.

Genesove, David and Christopher Mayer, "Loss Aversion and Seller Behavior: Evidence from the Housing Market," *The Quarterly Journal of Economics*, 2001, *116* (4), 1233–1260.

Giacomini, Raffaella and Halbert White, "Tests of Conditional Predictive Ability," *Econometrica*, 2006, *74* (6), 1545–1578.

Glaeser, Edward L. and Charles G. Nathanson, "An extrapolative model of house price dynamics," *Journal of Financial Economics*, 2017, *126* (1), 147–170.

Gneiting, Tilmann and Adrian E Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 2007, *102* (477), 359–378.

_ **and Matthias Katzfuss**, "Probabilistic Forecasting," *Annual Review of Statistics and Its Application*, 2014, *1* (1), 125–151.

Golub, Benjamin and Matthew O Jackson, "Naïve Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal: Microeconomics*, 2010, *2* (1), 112–149.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," 2022. NeurIPS 2022 Datasets and Benchmarks. arXiv:2207.08815.

Guo, Taicheng, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang, "Large Language Model Based Multi-agents: A Survey of Progress and Challenges," in "Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)" International Joint Conferences on Artificial Intelligence Organization 2024, pp. 8048–8057.

Halawi, Danny, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt, "Approaching Human-Level Forecasting with Language Models," 2024. arXiv:2402.18563.

Han, Lu and William C. Strange, "The Microstructure of Housing Markets," in Gilles Duranton, J. Vernon Henderson, and William C. Strange, eds., *Handbook of Regional and Urban Economics, Volume 5*, Amsterdam: Elsevier, 2015, pp. 813–886.

Hansen, Peter R., Asger Lunde, and James M. Nason, "The Model Confidence Set," *Econometrica*, 2011, *79* (2), 453–497.

Harvey, David, Stephen Leybourne, and Paul Newbold, "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 1997, *13* (2), 281–291.

Hill, Robert J., "Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy," *Journal of Economic Surveys*, 2013, *27* (5), 879–914.

Himmelberg, Charles, Christopher Mayer, and Todd Sinai, "Assessing High House Prices: Bubbles, Fundamentals and Misperceptions," *Journal of Economic Perspectives*, 2005, *19* (4), 67–92.

Holm, Sture, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 1979, *6* (2), 65–70. Accessed 2026-10-27.

Hong, Lu and Scott E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *Proceedings of the National Academy of Sciences*, 2004, *101* (46), 16385–16389.

Hong, Sirui, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber, "MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework," 2023. ICLR 2024. arXiv:2308.00352.

Howell, Junia and Elizabeth Korver-Glenn, "Neighborhoods, Race, and the Twenty-first-century Housing Appraisal Industry," *Sociology of Race and Ethnicity*, 2018, *4* (4), 473–490.

_ **and** _ , "The Increasing Effect of Neighborhood Racial Composition on Housing Values, 1980–2015," *Social Problems*, 2021, *68* (4), 1051–1071.

Iacoviello, Matteo, "House Prices, Borrowing Constraints, and Monetary Policy in the Business Cycle," *American Economic Review*, 2005, *95* (3), 739–764.

Irving, Geoffrey, Paul Christiano, and Dario Amodei, "AI safety via debate," 2018. arXiv:1805.00899.

Jacovi, Alon and Yoav Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?," in "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)" Association for Computational Linguistics 2020, pp. 4198–4205.

Jose, Victor Richmond R., Yael Grushka-Cockayne, and Kenneth C. Lichtendahl, "Trimmed Opinion Pools and the Crowd's Calibration Problem," *Management Science*, 2014, *60* (2), 463–475.

Kadavath, Saurav, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan, "Language Models (Mostly) Know What They Know," 2022. arXiv:2207.05221.

Kaesberg, Lars Benedikt, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp, "Voting or Consensus? Decision-Making in Multi-Agent Debate," 2025. arXiv:2502.19130.

Khan, Akbir, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez, "Debating with More Persuasive LLMs Leads to More Truthful Answers," 2024. ICML 2024. arXiv:2402.06782.

Kiyotaki, Nobuhiro and John Moore, "Credit Cycles," *Journal of Political Economy*, 1997, *105* (2), 211–248.

Kleinberg, Jon and Manish Raghavan, "Algorithmic monoculture and social welfare," *Proceedings of the National Academy of Sciences*, 2021, *118* (22), e2018340118.

Koenker, Roger and Gilbert Bassett, "Regression Quantiles," *Econometrica*, 1978, *46* (1), 33–50.

Kok, Nils, Eija-Leena Koponen, and Carmen Adriana Martínez-Barbosa, "Big Data in Real Estate? From Manual Appraisal to Automated Valuation," *The Journal of Portfolio Management*, 2017, *43* (6), 202–211.

Künsch, Hans R., "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 1989, *17* (3), 1217–1241.

Lamberson, P. J. and Scott E. Page, "Optimal Forecasting Groups," *Management Science*, 2012, *58* (4), 805–810.

Lanham, Tamera, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez, "Measuring Faithfulness in Chain-of-Thought Reasoning," 2023. arXiv:2307.13702.

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman, "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 2018, *113* (523), 1094–1111.

LeSage, James and Robert Kelley Pace, *Introduction to Spatial Econometrics*, Boca Raton, FL: Chapman and Hall/CRC, 2009.

Li, Guohao, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem, "CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society," 2023. NeurIPS 2023. arXiv:2303.17760.

Li, Junyou, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye, "More Agents Is All You Need," 2024. arXiv:2402.05120.

Li, Longtu, "Large Language Models as Structural Housing Valuation Agents: Evidence from User-Cost Theory and UK Household Panel Data," in "32nd Annual European Real Estate Society Conference (ERES 2026)" European Real Estate Society 2026. Conference paper.

Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu, "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate," in "Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)" Association for Computational Linguistics 2024, pp. 17889–17904.

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing, "How social influence can undermine the wisdom of crowd effect," *Proceedings of the National Academy of Sciences*, 2011, *108* (22), 9020–9025.

Malpezzi, Stephen, "Hedonic Pricing Models: A Selective and Applied Review," in Tony O'Sullivan and Kenneth Gibb, eds., *Housing Economics and Public Policy*, Oxford: Blackwell, 2003, pp. 67–89.

Matheson, James E. and Robert L. Winkler, "Scoring Rules for Continuous Probability Distributions," *Management Science*, 1976, *22* (10), 1087–1096.

Mayer, Michael, Steven C. Bourassa, Martin Hoesli, and Donato Scognamiglio, "Estimation and updating methods for hedonic valuation," *Journal of European Real Estate Research*, 2019, *12* (1), 134–150.

Mian, Atif and Amir Sufi, "House Prices, Home Equity-Based Borrowing, and the US Household Leverage Crisis," *American Economic Review*, 2011, *101* (5), 2132–2156.

Michael, Julian, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman, "Debate Helps Supervise Unreliable Experts," 2023. arXiv:2311.08702.

Ministère des Affaires municipales et de l'Habitation du Québec, "Manuel d'évaluation foncière du Québec," https://www.mamh.gouv.qc.ca/evaluation-fonciere/manuel-devaluation-fonciere-du-quebec/ 2024. Accessed 2026-10-27.

Morris, Stephen and Hyun Song Shin, "Social Value of Public Information," *American Economic Review*, 2002, *92* (5), 1521–1534.

Mullainathan, Sendhil and Jann Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Nakamura, Leonard I., "How Much Is That Home Really Worth? Appraisal Bias and House-Price Uncertainty," *Federal Reserve Bank of Philadelphia Business Review*, 2010, *Q1*, 11–22. Accessed 2026-10-27.

Navajas, Joaquin, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman, "Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds," *Nature Human Behaviour*, 2018, *2* (2), 126–132.

Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 1987, *55* (3), 703–708.

Northcraft, Gregory B and Margaret A Neale, "Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions," *Organizational Behavior and Human Decision Processes*, 1987, *39* (1), 84–97.

Nowak, Adam and Patrick Smith, "Textual Analysis in Real Estate," *Journal of Applied Econometrics*, 2017, *32* (4), 896–918.

Pace, R. Kelley, Ronald Barry, and C. F. Sirmans, "Spatial Statistics and Real Estate," *The Journal of Real Estate Finance and Economics*, 1998, *17* (1), 5–13.

_ , _ , John M. Clapp, and Mauricio Rodriquez, "Spatiotemporal Autoregressive Models of Neighborhood Effects," *The Journal of Real Estate Finance and Economics*, 1998, *17* (1), 15–33.

Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in "Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)" ACM 2023, pp. 1–22.

Pérez-Rave, Jorge Iván, Juan Carlos Correa-Morales, and Favián González-Echavarría, "A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes," *Journal of Property Research*, 2019, *36* (1), 59–96.

Perry, Andre M., Jonathan Rothwell, and David Harshbarger, "The Devaluation of Assets in Black Neighborhoods: The Case of Residential Property," Technical Report, Brookings Institution, Metropolitan Policy Program 2018. Accessed 2026-10-27.

Politis, Dimitris N. and Joseph P. Romano, "The Stationary Bootstrap," *Journal of the American Statistical Association*, 1994, *89* (428), 1303–1313.

Poterba, James M., "Tax Subsidies to Owner-Occupied Housing: An Asset-Market Approach," *The Quarterly Journal of Economics*, 1984, *99* (4), 729–752.

Prelec, Dražen, H. Sebastian Seung, and John McCoy, "A solution to the single-question crowd wisdom problem," *Nature*, 2017, *541* (7638), 532–535.

Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, "CatBoost: Unbiased Boosting with Categorical Features," 2018. NeurIPS 2018. arXiv:1706.09516.

Quan, Daniel C. and John M. Quigley, "Price formation and the appraisal function in real estate markets," *The Journal of Real Estate Finance and Economics*, 1991, *4* (2), 127–146.

Ranjan, Roopesh and Tilmann Gneiting, "Combining Probability Forecasts," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2010, *72* (1), 71–91.

Romano, Yaniv, Evan Patterson, and Emmanuel J. Candès, "Conformalized Quantile Regression," 2019. NeurIPS 2019. arXiv:1905.03222.

Rosen, Sherwin, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 1974, *82* (1), 34–55.

Rosiers, François Des, Marius Thériault, and Paul-Y. Villeneuve, "Sorting out access and neighbourhood factors in hedonic price modelling," *Journal of Property Investment & Finance*, 2000, *18* (3), 291–315.

Satopää, Ville A., Jonathan Baron, Dean P. Foster, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar, "Combining multiple probability predictions using a simple logit model," *International Journal of Forecasting*, 2014, *30* (2), 344–356.

Schoenegger, Philipp and Peter S. Park, "Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament," 2023. arXiv:2310.13014.

_ , Indre Tuminauskaite, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock, "Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy," *Science Advances*, 2024, *10* (45), eadp1528.

Shen, Lily and Stephen Ross, "Information value of property description: A Machine learning approach," *Journal of Urban Economics*, 2021, *121*, 103299.

Shui, Jessica and Shriya Murthy, "Are Appraisal Management Companies Value-Adding? Stylized Facts from AMC and Non-AMC Appraisals," Staff Working Paper 18-01, Federal Housing Finance Agency 2018. Accessed 2026-10-27.

Sirmans, Stacy, David Macpherson, and Emily Zietz, "The Composition of Hedonic Pricing Models," *Journal of Real Estate Literature*, 2005, *13* (1), 1–44.

Smit, Andries, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius, "Should We Be Going MAD? A Look at Multi-Agent Debate Strategies for LLMs," 2024. ICML 2024. arXiv:2311.17371.

Snell, Charlie, Jaehoon Lee, Kelvin Xu, and Aviral Kumar, "Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters," 2024. arXiv:2408.03314.

Statistics Canada, "Canadian Housing Statistics Program (CHSP): Housing Statistics Portal," https://www.statcan.gc.ca/en/subjects-start/housing 2024. Accessed 2026-10-27.

Stein, Jeremy C., "Prices and Trading Volume in the Housing Market: A Model with Down-Payment Effects," *The Quarterly Journal of Economics*, 1995, *110* (2), 379–406.

Surowiecki, James, *The Wisdom of Crowds*, New York: Doubleday, 2004. Accessed 2026-10-27.

Tian, Katherine, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning, "Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback," in "Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)" Association for Computational Linguistics 2023, pp. 5433–5442.

Timmermann, Allan, "Forecast Combinations," in Graham Elliott, Clive W. J. Granger, and Allan Timmermann, eds., *Handbook of Economic Forecasting, Volume 1*, Amsterdam: Elsevier, 2006, pp. 135–196.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman, "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting," 2023. NeurIPS 2023. arXiv:2305.04388.

Vedula, Nikhita, Dushyanta Dhyani, Laleh Jalali, Boris Oreshkin, Mohsen Bayati, and Shervin Malmasi, "Quantile Regression with Large Language Models for Price Prediction," 2025. Findings of ACL 2025. arXiv:2506.06657.

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer, *Algorithmic Learning in a Random World*, New York: Springer, 2005.

Wang, Qineng, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song, "Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?," 2024. NAACL 2024. arXiv:2402.18272.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," 2022. ICLR 2023. arXiv:2203.11171.

Wheaton, William C., "Vacancy, Search, and Prices in a Housing Market Matching Model," *Journal of Political Economy*, 1990, *98* (6), 1270–1292.

Wu, Haolun, Zhenkun Li, and Lingyao Li, "Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning," 2025. arXiv:2511.07784.

Wu, Qingyun, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang, "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," 2023. arXiv:2308.08155.

Xiong, Miao, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi, "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs," 2024. ICLR 2024. arXiv:2306.13063.

Zhang, Hanxiang, Yansong Li, and Paula Branco, "Describe the house and I will tell you the price: House price prediction with textual description data," *Natural Language Engineering*, 2024, *30* (4), 661–695.

Zhu, Xiaochen, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos, "Demystifying Multi-Agent Debate: The Role of Confidence and Diversity," 2026. arXiv:2601.19921.

Appendices

A. Reproducibility

A.1. Repository and one-command reproduction

The research code is a Python package, `agentic_valuation`, installed in editable mode from the repository root, with a command-line entry point `av` and a numbered pipeline of scripts. The layout is as follows: `configs/` holds the four configuration files that fix the partition windows and seeds (`data.yaml`), the model identifiers and list prices (`models.yaml`), the arms, aggregation rules and rounds grid (`treatments.yaml`) and the budget tiers (`budgets.yaml`); `src/agentic_valuation/` contains the sub-packages `data` (analytic schema, build, splits, leakage audit, synthetic generator), `features` (spatial index, as-of comparables, market context, geographic layers, feature matrix), `baselines` (hedonic ridge, hierarchical and repeat-sales model, LightGBM, CatBoost, XGBoost, AVM with conformal intervals), `agents` (role prompts, evidence packets and views, turn runner), `debate` (protocols, committees, aggregation, stopping rule), `experiments` (resumable arm runner, freeze, provider check), `evaluation` (metrics, block bootstrap, Holm and Diebold–Mariano tests, calibration and conformal rescaling, strata, faithfulness audit, table generation), `fundamental` (repeat-sales deflator, resale-implied values, rents and user cost), `theory` (aggregation and welfare simulations) and `providers` (Anthropic, OpenAI and mock adapters with the cost ledger); `scripts/` contains the pipeline in execution order; `tests/` contains 143 tests that run on the synthetic dataset with the mock provider; `docs/` contains the research plan, the data-access audit, the preregistration, the architecture, the data dictionary, the evaluation protocol, the literature map, the fundamental-value note and the prospective-set note; `reports/` contains the phase reports and the limitations register; `results/` contains every run, ledger, evaluation output and table; and `paper/` contains this manuscript.

Everything that does not require a paid API call is reproduced by one command, `make all`, which runs in sequence the test suite, the synthetic data layer, the real data layer and leakage audit, the geographic layers and feature matrix, the baselines, the theory and welfare simulations, the fundamental-value scripts, a mock-provider smoke run of arms F and G2 on five synthetic properties, the evaluation, the figures and tables, and the LaTeX build. Table A1 lists the command of each phase. Live runs are explicit: a tier is run with `scripts/33_run_tier.py <tier> -live` and a single arm with `scripts/30_run_arm.py`, whose options `-arm`, `-split`, `-n`, `-live`, `-committee` and `-run-id` select the arm, the partition, the number of properties, the live providers, the committee and the run identifier; the environment variable `AV_MAX_SPEND_USD` caps the spend of a run and `AV_PROVIDER_MODE=mock` routes every call to the mock provider. The prospective set is fetched by

scripts/60_fetch_prospective.py, built by scripts/61_build_prospective.py, scored by scripts/62_prospective_baselines.py and evaluated by scripts/20_evaluate.py with the option -split prospective.

A.2. Environment

The runs reported in the paper were executed on an Apple-silicon workstation under Python 3.14.4 with polars 1.40, pandas 3.0, numpy 2.4, scikit-learn 1.6, lightgbm 4.6, xgboost 3.2, catboost 1.2.10, statsmodels 0.14, pydantic 2.13, anthropic 2.3.0 and openai 3.40.0, and the manuscript was built with TeX Live through latexmk; the package version recorded in the freeze is 0.1.0 and the minimum supported Python version declared in pyproject.toml is 3.12. The test suite and the mock provider do not depend on hardware. Heavy baseline tuning may be delegated to a compute node without changing the results, which are fixed by the seed 20261027 recorded in configs/data.yaml and reused by the evaluation bootstrap.

A.3. Data snapshot, freeze and archive

The raw data layer consists of three Parquet files produced by the cleaning pipeline of the companion index paper (Boucher, 2026) from the Vrai Prix snapshot of 2026-10-21, pinned by SHA-256 in data/raw/SHA256SUMS and not redistributed:

```
wp22_clean_transactions.parquet   d6e63ae0b9b480b66572c16a1fcd9a1ea18f0734b28163584532eac42c587fc9
wp22_raw_market_index.parquet     6c8c70d1a6d86e7704480c4527c90ff84026d51f38a2cf80ab5bd686e01d599a
wp22_raw_units_stock.parquet      6d0ba7c384111c41c09d34f46b37f3202f1fc91ed34471f90f9f643fd70314a7
```

The external layers (flood maps, transit stops, census dissemination areas, rental-board leases) and the prospective fetch of 2026-10-27 are documented with their provenance in data/raw/README.md and docs/prospective_set.md. The protocol freeze was recorded on 2026-10-28 at 01:05:40 UTC in results/freeze.json, which stores the SHA-256 hash of each of the thirty role prompts (twelve default, eight optional, the generalist and nine placebo members), of the four configuration files, of the schema module and of the sixteen source files of the agents, debate and providers sub-packages, together with the model identifiers of the cost-matched pairs, the rounds grid and the aggregation rules; the digest of the whole record is

```
f0e9d6a94b0387f92645346af2c8c63200c8b9f83c9a85095f507cffa2e85e1c
```

and scripts/31_freeze.py -check lists every deviation of the working tree from it. No deviation was recorded between the freeze and the last evaluation. Every live run writes to results/runs/<run_id>/ the manifest, one transcript per property with every turn, prompt hash, model identifier, token counts, cost, latency, validation retries and error, the evidence packet as sent with its hash, and the prediction file with one row per aggregation rule; the ledger of the run,

results/ledger/<run_id>.jsonl, appends one line per API call. The archive of transcripts and packets is about 0.5 GB and is kept locally; the prediction files, manifests, ledgers, aggregation weights, conformal scales and every evaluation output are in the repository. Runs whose identifiers begin with an underscore are mock, synthetic, pilot or superseded runs and are excluded from every evaluation by the manifest. Summing the ledgers of the live runs that enter the paper gives a total spend of about 1,080 US dollars at list prices (about 1,100 dollars including sanity and mock-priced calls).

A.4. What is and is not deterministic

The data layer, the partitions, the stratified subsamples, the feature matrix, the comparable retrieval, the evidence packets, the baselines, the aggregation rules, the evaluation and the simulations are deterministic given the seed and the configuration, up to the floating-point differences that gradient-boosting libraries produce across hardware and thread counts; the leakage audit verifies that the subsamples regenerate exactly. The agentic arms are not deterministic: hosted models return different completions to identical requests, no sampling parameters are sent, and the vendors' adaptive reasoning adds variation of its own. The paper therefore does not claim bitwise reproducibility of any agentic number; it claims that every such number traces to a stored transcript, and it measures the variation by repeating the two-round debate on eight test properties under the frozen protocol (scripts/36_repeat_variation.py, results/evaluation/repeat_variation.json): the committee median differed from the original run by 0.007 log points on average and 0.019 at most, with a standard deviation of 0.009, the absolute errors of the two runs had a correlation of 0.998, and every repeated point lay inside the original 90 percent interval. Re-running an agentic arm reproduces the design, not the transcript, and extends the archived runs rather than replacing them, because the runner skips properties whose transcript exists.

B. Role Prompts

The system prompt of every role is the concatenation of a shared preamble, the line ROLE_ID: <id>, the role-specific text and a block of output rules; it is constant across properties, which allows the vendors to cache it as a prefix, and its SHA-256 hash is part of the freeze. The texts below are reproduced verbatim from src/agentic_valuation/agents/roles.py. Table 6 in the main text paraphrases the twelve default mandates; two of them and the placebo text are given in full here.

B.1. Shared preamble

You are one simulated analytic role inside a research committee that estimates the market value of a residential property in Quebec, Canada. This is an academic experiment on information aggregation. You are NOT a licensed Quebec appraiser (évaluateur agréé), lender, broker or inspector, and nothing you produce is a professional opinion of value, a lending decision or an offer.

TASK
Estimate the MARKET VALUE of the subject property in Canadian dollars (CAD) AS OF THE VALUATION DATE given in the packet. Market value here means the most probable price the property would fetch in an arm's-length sale on the open market at the valuation date, with buyer and seller typically motivated and reasonably informed, after normal marketing time. It is a prediction of a transaction price, not a statement about the property's unobservable fundamental value and not an investment value for a particular buyer.

INFORMATION SET AND CUTOFF
– The packet states a valuation date and an information cutoff (the day before the sale). You may use ONLY evidence dated on or before the cutoff. The subject's own sale price is never in the packet and must not be guessed from the packet identifiers.
– Comparable sales (ids C01, C02, ...) all closed before the cutoff. Market context (M1) is computed from sales up to the cutoff. Geospatial context (G1), when present, is static. The subject (P) carries attributes and the municipal assessment roll in force at the sale date; the roll's market reference date is stated and usually precedes the valuation date by 18 months or more, so the roll is NOT a current value.
– A prior sale of the same parcel, if any, is disclosed in the subject section; account for the market movement between the two dates.
– You only see the sections your role is entitled to. If a section or field is absent, it is absent for you: do not infer it, do not assume typical values for it, and list what is missing in missing_data.

EVIDENCE DISCIPLINE
– Cite the evidence ids you actually relied on (P, C01.., M1, G1) in evidence_ids_used. Never cite an id that is not in your packet.
– No inspection has been performed. Do not invent condition, renovations, finishes, defects, views, noise or neighbourhood quality that are not in the evidence. Unobserved means unobserved.
– Adjusted comparable prices in adjustments must be numerically traceable: adjusted_price is the comparable's raw price after your stated adjustments, and must remain within a plausible band of the raw price.

– Keep missing data missing: if the floor area, land area, year built or roll values are null, say so and widen your interval instead of filling the gap.
– Quantities: CAD for money, square metres for areas, kilometres for distances, ISO dates.

ANSWER FORMAT
Return ONLY one JSON object that validates against the schema requested in the user message. No prose outside the JSON, no markdown fences, no comments. Numbers are plain numbers (no currency symbols, no thousands separators). Quantiles must be non-decreasing (q05 $\leq$ q25 $\leq$ q50 $\leq$ q75 $\leq$ q95) and the point estimate must lie within [q05, q95]; the 90 % interval [q05, q95] should reflect your genuine uncertainty given the evidence you can see. reasoning_summary is at most 120 words and must mention the evidence ids that drove the estimate. Be concise, specific and quantitative; do not pad.

B.2. Two role-specific texts

Role `comparables` *(view: subject, comparables).*

You are the comparable-sales selection and adjustment specialist. Rank the comparables by similarity to the subject (class, subtype, floor area, land area, year built, distance, months before valuation), discard the ones that are not credible substitutes and say why, then adjust the retained ones for time, size, age and location. Report one adjustment entry per retained comparable with the adjusted price. Your point estimate is a weighted reconciliation of the adjusted prices; your interval reflects their dispersion and the quality of the match.

Role `building_condition` *(view: subject only).*

You are the building-condition specialist. You see only the subject's attributes and roll values, and NO inspection has been performed. State plainly that condition, renovations and defects are unobserved. Reason only about what age, storeys, floor area, construction year, new-construction flag and the land/building split of the roll imply for typical depreciation and functional obsolescence. Your estimate starts from the roll values and the reference date stated; your interval must be wide and your missing_data must list the inspection items you would need.

B.3. Placebo and generalist texts

Placebo member k (arm I; the view is that of default role k).

You are committee member number {k}. You have no specialised role: apply general real-estate valuation judgement to the evidence available to you, cite the ids you rely on, state your assumptions and return your best estimate with a calibrated 90 % interval.

Generalist (arms E and J; full packet).

> You are a single generalist valuation analyst working alone. You see the full packet. Use comparable sales, the market context, the roll values and the geospatial context as you see fit, state your assumptions, and return your best estimate and a calibrated 90 % interval.

C. The JSON Output Contract

Every estimate turn must return one object validating against the `AgentEstimate` schema and every critique turn one object validating against `CritiqueSet`, both defined as typed Pydantic models in `src/agentic_valuation/schemas.py` and enforced through the vendors' structured-output facilities and a server-side validation with at most two corrective retries. Table A2 lists the fields. The validator of `AgentEstimate` rejects any object in which a quantile or the point estimate is not strictly positive, in which the five quantiles are not non-decreasing, or in which the point estimate lies outside $[q_{05}, q_{95}]$; a turn that fails after the retries is stored with its error and a null estimate and is excluded from aggregation without imputation. The schema also defines the evidence packet (`SubjectProperty`, `Comparable`, `MarketContext`, `GeoContext`), whose evidence identifiers are the universe against which `evidence_ids_used` and `evidence_ids_cited` are checked by the faithfulness audit, and the transcript objects (`AgentTurn`, `DebateTranscript`) that record each turn's role, round, kind, model, provider, tokens, cost, latency, prompt hash, request identifier, validation retries and error, together with the committee assignment and the protocol events of each property.

D. Additional Tables

This appendix reports, for the validation partition on which every design choice was made, for the untouched 2026 holdout and for the prospective set of sales registered after the freeze, the same tables that the main text reports for the test partition: accuracy by arm, the paired contrasts, interval coverage and sharpness, errors by stratum and cost. Every fragment is written by `scripts/20_evaluate.py` from the versioned prediction files and is not edited by hand. Baselines are evaluated on the full partition; the agentic arms on the nested subsamples described in Section 7.6 (forty validation properties per committee, twenty for the four-round debate, sixty holdout properties and forty prospective sales, all in the mixed committee at the mid tier unless the committee column says otherwise), with the median of the roles' final estimates as the aggregation rule. Strata with fewer than thirty properties are suppressed, which for the agentic arms leaves few cells. The faithfulness audit is reported once, in the main text, because its table covers every run of every partition.

D.1. Validation partition

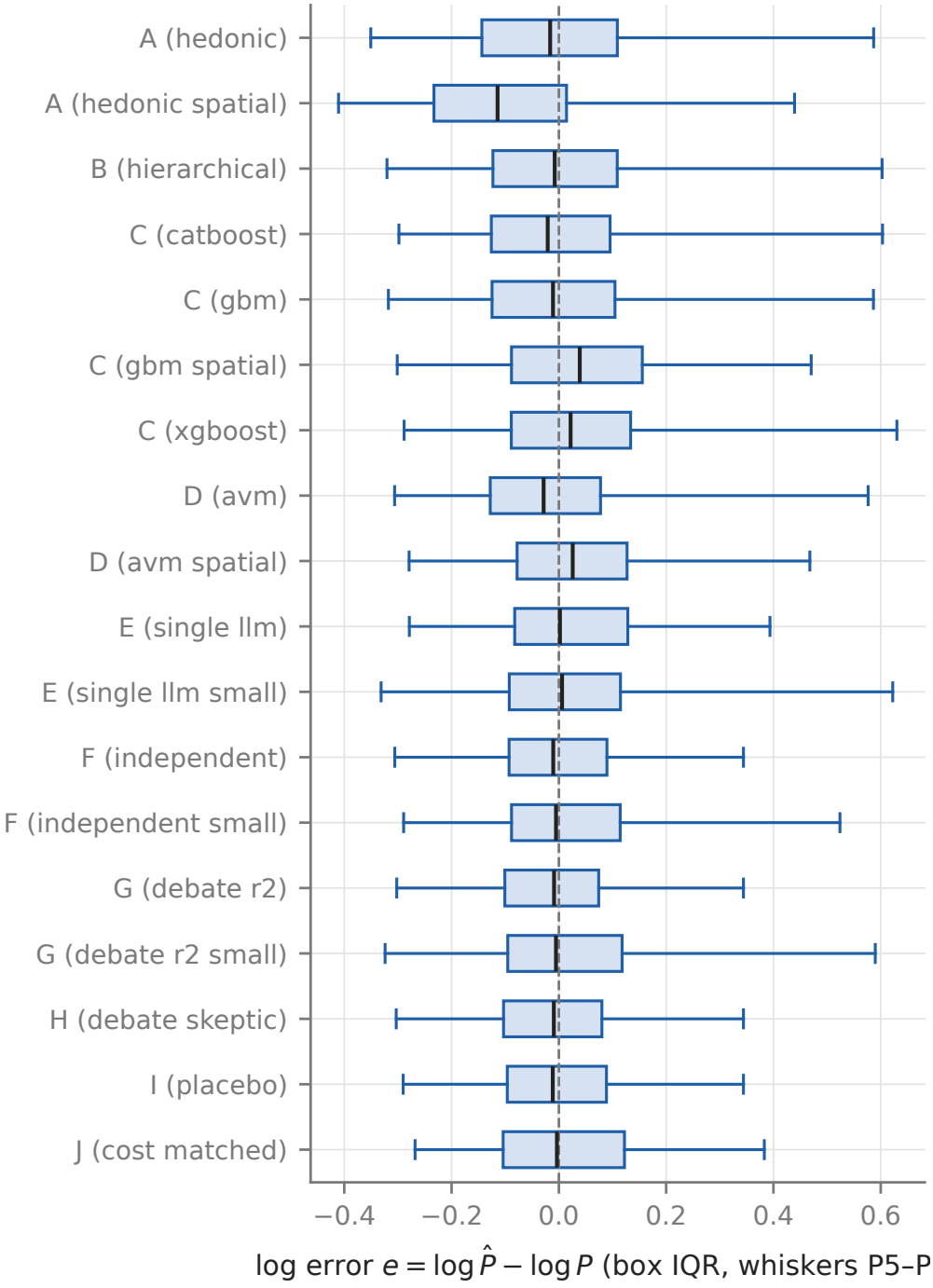


Figure 4. Distribution of the log error $e = \log \hat{P} - \log P$ by arm on the test partition. Boxes are the interquartile range with the median marked; whiskers end at the 5th and 95th percentiles. Baselines are scored on the full partition (spatial rows on the 8,988 withheld-region sales), mid-tier agentic arms on 80 properties and small-tier arms on 300. Source: `results/evaluation/records.parquet`, `scripts/21_figures.py`.

Table 9. Paired contrasts on the test subsample: primary family and secondary pairs.

Contrast (B vs A)	Metric	N	A	B	Δ	95% CI	p (boot)	p (Holm)	p (DM)
G (debate r2) vs D (avm)	Mean $\lvert e\rvert$	80	0.163	0.158	−0.005	[-0.033, 0.026]	0.750	1.000	0.746
G (debate r2) vs F (independent)	Mean $\lvert e\rvert$	80	0.159	0.158	−0.001	[-0.006, 0.005]	0.702	1.000	0.655
G (debate r2) vs J (cost matched)	Mean $\lvert e\rvert$	80	0.163	0.158	−0.006	[-0.017, 0.004]	0.291	1.000	0.246
G (debate r2) vs I (placebo)	Mean $\lvert e\rvert$	80	0.157	0.158	0.000	[-0.005, 0.006]	0.946	1.000	0.948
G (debate r2) vs I (placebo)	P95 $\lvert e\rvert$	80	0.497	0.493	−0.004	[-0.022, 0.065]	0.731	–	–
G (debate r2) vs E (single llm)	Mean $\lvert e\rvert$	80	0.166	0.158	−0.008	[-0.023, 0.004]	0.236	–	0.201
G (debate r2) vs E (single llm)	MdAPE	80	0.107	0.087	−0.020	[-0.060, 0.007]	0.263	–	–
G (debate r2) vs E (single llm)	P95 $\lvert e\rvert$	80	0.508	0.493	−0.014	[-0.151, 0.070]	0.690	–	–
G (debate r2) vs E (single llm)	$\lvert e\rvert > \log 1.25$	80	0.237	0.225	−0.012	[-0.069, 0.040]	0.747	–	–
G (debate r2) vs E (single llm)	Cov. 90%	80	0.887	0.887	0.000	[-0.050, 0.051]	1.000	–	–
G (debate r2) vs E (single llm)	Pinball (log)	80	0.055	0.053	−0.002	[-0.006, 0.001]	0.272	–	0.243
G (debate r2) vs H (debate skeptic)	Mean $\lvert e\rvert$	80	0.158	0.158	−0.000	[-0.003, 0.003]	0.973	–	0.968
G (debate r2) vs H (debate skeptic)	MdAPE	80	0.092	0.087	−0.006	[-0.017, 0.007]	0.240	–	–
G (debate r2) vs H (debate skeptic)	P95 $\lvert e\rvert$	80	0.493	0.493	0.000	[-0.025, 0.024]	0.979	–	–
G (debate r2) vs H (debate skeptic)	$\lvert e\rvert > \log 1.25$	80	0.225	0.225	0.000	[-0.037, 0.038]	1.000	–	–
G (debate r2) vs H (debate skeptic)	Cov. 90%	80	0.875	0.887	0.012	[0.000, 0.041]	0.558	–	–
G (debate r2) vs H (debate skeptic)	Pinball (log)	80	0.053	0.053	0.000	[-0.001, 0.001]	0.787	–	0.805
G (debate r2) vs A (hedonic)	Mean $\lvert e\rvert$	80	0.163	0.158	−0.006	[-0.035, 0.027]	0.715	–	0.668
G (debate r2) vs A (hedonic)	MdAPE	80	0.109	0.087	−0.023	[-0.058, 0.027]	0.276	–	–
G (debate r2) vs A (hedonic)	P95 $\lvert e\rvert$	80	0.430	0.493	0.063	[-0.111, 0.429]	0.562	–	–
G (debate r2) vs A (hedonic)	$\lvert e\rvert > \log 1.25$	80	0.263	0.225	−0.038	[-0.126, 0.056]	0.477	–	–
G (debate r2) vs A (hedonic)	Cov. 90%	80	0.925	0.887	−0.038	[-0.110, 0.024]	0.306	–	–
G (debate r2) vs A (hedonic)	Pinball (log)	80	0.054	0.053	−0.001	[-0.011, 0.010]	0.790	–	0.760
G (debate r2) vs C (gbm)	Mean $\lvert e\rvert$	80	0.168	0.158	−0.011	[-0.041, 0.025]	0.531	–	0.517
G (debate r2) vs C (gbm)	MdAPE	80	0.124	0.087	−0.037	[-0.061, 0.014]	0.084	–	–
G (debate r2) vs C (gbm)	P95 $\lvert e\rvert$	80	0.530	0.493	−0.037	[-0.220, 0.372]	0.653	–	–
G (debate r2) vs C (gbm)	$\lvert e\rvert > \log 1.25$	80	0.225	0.225	0.000	[-0.063, 0.060]	1.000	–	–
G (debate r2) vs C (gbm)	Cov. 90%	80	–	0.887	–	–	–	–	–
G (debate r2) vs C (gbm)	Pinball (log)	80	–	0.053	–	–	–	–	–
F (independent) vs D (avm)	Mean $\lvert e\rvert$	80	0.163	0.159	−0.004	[-0.029, 0.024]	0.787	–	0.791
F (independent) vs D (avm)	MdAPE	80	0.122	0.091	−0.031	[-0.060, 0.014]	0.142	–	–
F (independent) vs D (avm)	P95 $\lvert e\rvert$	80	0.506	0.499	−0.007	[-0.192, 0.284]	0.916	–	–
F (independent) vs D (avm)	$\lvert e\rvert > \log 1.25$	80	0.212	0.212	0.000	[-0.056, 0.064]	1.000	–	–
F (independent) vs D (avm)	Cov. 90%	80	0.875	0.887	0.012	[-0.056, 0.071]	0.766	–	–
F (independent) vs D (avm)	Pinball (log)	80	0.052	0.053	0.001	[-0.008, 0.011]	0.833	–	0.827
F (independent) vs E (single llm)	Mean $\lvert e\rvert$	80	0.166	0.159	−0.007	[-0.023, 0.006]	0.340	–	0.308
F (independent) vs E (single llm)	MdAPE	80	0.107	0.091	−0.016	[-0.049, 0.013]	0.361	–	–
F (independent) vs E (single llm)	P95 $\lvert e\rvert$	80	0.508	0.499	−0.009	[-0.154, 0.082]	0.835	–	–
F (independent) vs E (single llm)	$\lvert e\rvert > \log 1.25$	80	0.237	0.212	−0.025	[-0.085, 0.031]	0.495	–	–
F (independent) vs E (single llm)	Cov. 90%	80	0.887	0.887	0.000	[-0.050, 0.051]	1.000	–	–
F (independent) vs E (single llm)	Pinball (log)	80	0.055	0.053	−0.002	[-0.007, 0.002]	0.319	–	0.311

Notes. Each row compares arm B to arm A on the same N properties; Δ is the metric of B minus the metric of A, so that a negative value favours B for every loss. The 95% CI is the percentile interval of a paired block bootstrap (blocks = municipality $\times$ calendar month, 65 blocks, 2,000 resamples, seed 20261027); p (boot) is the recentred bootstrap p-value; p (Holm) is the step-down adjusted value within the primary family of the four mean-$\lvert e\rvert$ contrasts (rows 1–4); p (DM) is the Diebold–Mariano test on the date-ordered loss differential with Newey–West variance and the Harvey–Leybourne–Newbold correction, reported for mean $\lvert e\rvert$ and pinball loss only. The baselines' values are their paired values on the same properties, not their full-partition values. Cov. 90% is the raw coverage of the 90 percent interval; arm C has no intervals. Source: `results/evaluation/contrasts.csv`.

Table 10. Effect sizes, relative confidence intervals and minimum detectable effects for the primary family.

Contrast	Metric	Δ	95% CI	SE	Δ (%)	95% CI (%)	MDE	MDE (%)	p (Holm)
G2 vs D	Mean $\lvert e\rvert$	−0.005	[-0.033, 0.026]	0.0152	−3.1	[-20.0, 16.2]	0.043	26.2	1.000
G2 vs F	Mean $\lvert e\rvert$	−0.001	[-0.006, 0.005]	0.0029	−0.7	[-4.0, 3.2]	0.008	5.1	1.000
G2 vs J	Mean $\lvert e\rvert$	−0.006	[-0.017, 0.004]	0.0055	−3.5	[-10.6, 2.5]	0.015	9.5	1.000
G2 vs I	Mean $\lvert e\rvert$	0.000	[-0.005, 0.006]	0.0026	0.1	[-2.9, 3.5]	0.007	4.6	1.000
G2 vs I	P95 $\lvert e\rvert$	−0.004	[-0.022, 0.065]	0.0177	−0.9	[-4.4, 13.1]	0.050	10.0	–

Notes. Δ and its CI are those of Table 9; SE is the bootstrap standard error of Δ; Δ (%) and its CI are Δ and the CI divided by the value of the reference arm A (the first-named arm in the design, e.g. D in "G2 vs D"). MDE is the minimum detectable effect at 80 percent power for a two-sided test at the 5 percent level implied by the observed standard error, $(1.96 + 0.84)$ SE, in absolute log points and as a percentage of the reference value. p (Holm) is reported for the four mean-$\lvert e\rvert$ contrasts. Source: `results/evaluation/contrasts.csv`, `scripts/22_extra_tables.py`.

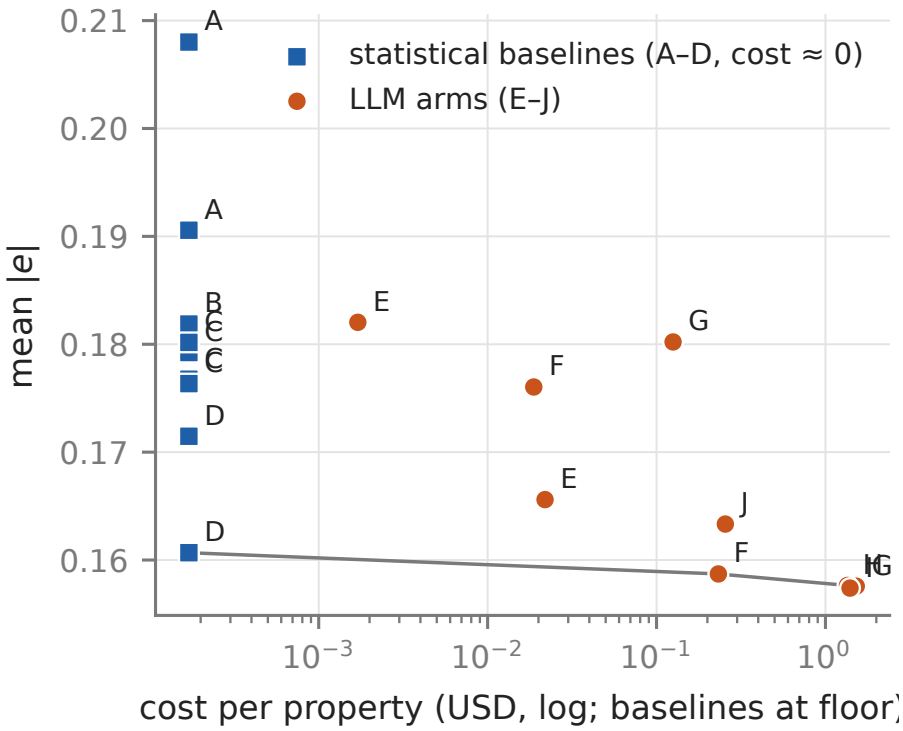


Figure 5. Cost per property and mean absolute log error on the test partition. Baselines (squares) have no API cost and are plotted at a floor; agentic arms (circles) at their ledger cost per property, mid tier on 80 properties and small tier on 300. Letters are arm codes; the two points per letter for E, F and G are the small and mid tiers. The grey line joins the non-dominated arms. Source: `results/evaluation/summary_main.csv`, `scripts/21_figures.py`.

Table 11. Accuracy and calibration across the test, holdout and prospective evaluation sets.

Split	Arm	n	Mean $\lvert e\rvert$	MdAPE	P95 $\lvert e\rvert$	Bias $\bar{e}$	$\lvert e\rvert > \log 1.25$	Over +25%	Cov. 50%	Cov. 90%	Width 90%
test	D (full partition)	96935	0.171	0.105	0.622	0.010	0.218	0.110	0.446	0.896	0.830
	D (paired subsample)	80	0.163	0.122	0.506	−0.031	0.212	0.075	0.425	0.875	0.846
	E	80	0.166	0.107	0.508	0.013	0.237	0.125	0.463	0.887	0.562
	F	80	0.159	0.091	0.499	0.003	0.212	0.087	0.512	0.887	0.590
	G2	80	0.158	0.087	0.493	−0.004	0.225	0.100	0.475	0.887	0.602
holdout 2026	D (full partition)	83489	0.180	0.112	0.643	0.002	0.239	0.115	0.434	0.878	0.827
	D (paired subsample)	60	0.189	0.113	0.689	0.055	0.250	0.183	0.367	0.900	0.815
	E	60	0.195	0.086	0.846	0.110	0.267	0.183	0.467	0.800	0.497
	F	60	0.186	0.090	0.805	0.107	0.233	0.183	0.517	0.833	0.508
	G2	60	0.189	0.089	0.828	0.102	0.250	0.183	0.533	0.833	0.528
prospective	D (full partition)	1182	0.202	0.129	0.740	0.039	0.281	0.157	0.410	0.866	0.889
	D (paired subsample)	40	0.197	0.096	0.698	0.059	0.275	0.150	0.350	0.900	0.902
	E	40	0.187	0.085	0.645	0.059	0.300	0.175	0.425	0.750	0.558
	F	40	0.179	0.088	0.660	0.061	0.275	0.175	0.500	0.800	0.574
	G2	40	0.187	0.096	0.662	0.058	0.300	0.175	0.450	0.825	0.571

Notes. Columns as in Table 8; Over +25% is the share of sales valued more than 25 percent above the price; Cov. 50% and 90% are the raw coverages of the 50 and 90 percent intervals and Width 90% their mean log width. "D (full partition)" is the AVM on all sales of the set; "D (paired subsample)" the same AVM on the properties valued by the agentic arms, so that it is directly comparable with E, F and G2. Test: sales of 2025-04-01 to 2025-12-31; holdout: 2026-01-01 to 2026-10-07; prospective: registered 2026-10-06 to 2026-10-13 and absent from the snapshot. Source: `results/evaluation{,_holdout,_prospective}/summary_main.csv` and `records.parquet`, `scripts/22_extra_tables.py`.

Table 12. Cost, latency and tokens per property, test partition.

Arm	Cost (USD)	Latency (s)	Tokens	Total cost (USD)	Mean $\lvert e\rvert$	N
A (hedonic)	0.0000	0.0	0	0.00	0.191	96935
A (hedonic spatial)	0.0000	0.0	0	0.00	0.208	8988
B (hierarchical)	0.0000	0.0	0	0.00	0.182	96935
C (catboost)	0.0000	0.0	0	0.00	0.177	96935
C (gbm)	0.0000	0.0	0	0.00	0.179	96935
C (gbm spatial)	0.0000	0.0	0	0.00	0.176	8988
C (xgboost)	0.0000	0.0	0	0.00	0.180	96935
D (avm)	0.0000	0.0	0	0.00	0.171	96935
D (avm spatial)	0.0000	0.0	0	0.00	0.161	8988
E (single llm)	0.0219	19.7	6596	1.75	0.166	80
E (single llm small)	0.0017	19.1	8133	0.51	0.182	300
F (independent)	0.2323	79.1	73947	18.58	0.159	80
F (independent small)	0.0188	78.5	89504	5.63	0.176	300
G (debate r2)	1.5217	336.2	521127	121.73	0.158	80
G (debate r2 small)	0.1253	358.4	667795	37.58	0.180	300
H (debate skeptic)	1.3501	304.8	459057	108.01	0.158	80
I (placebo)	1.3973	268.1	491998	111.78	0.157	80
J (cost matched)	0.2554	93.8	91937	20.43	0.163	80

Notes. Cost is the mean estimated API cost per property in US dollars at list prices as recorded in the ledger; Latency the mean wall-clock time per property in seconds, including validation retries; Tokens the mean input plus output tokens per property; Total cost the sum over the evaluated properties. Baselines record no API cost; their fitting time is not shown. Mean $\lvert e\rvert$ and N repeat Table 8. Source: `results/tables/cost.csv`, `results/ledger/`.

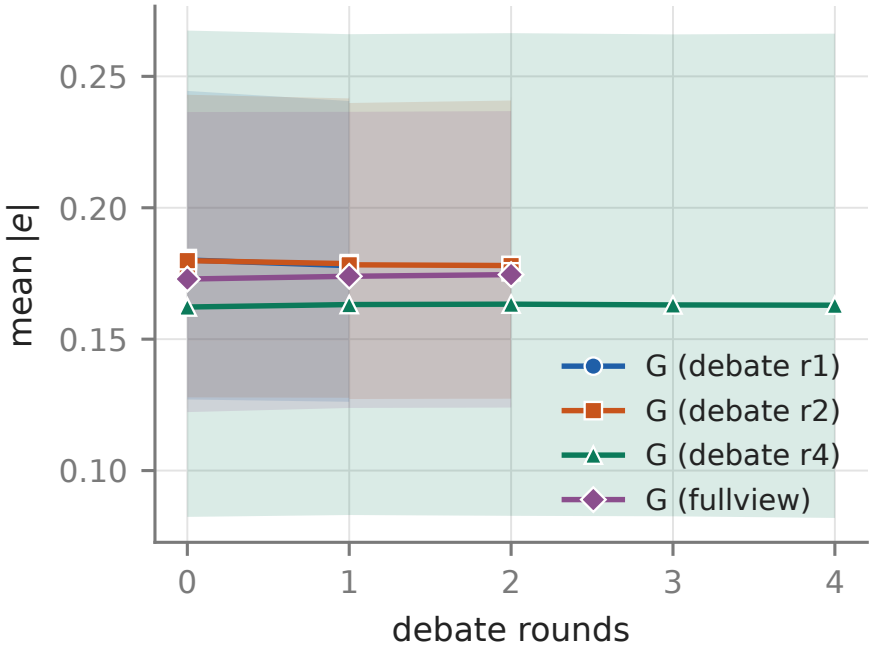


Figure 6. Mean absolute log error of the committee median by deliberation round on the validation subsample, mixed committee (G1, G2, G4, G-fullview). Shaded bands are 95 percent block-bootstrap intervals. Committees stopped early by the interquartile-range rule keep their last estimates in the later rounds, so that every round averages the same 40 properties (20 for G4). Source: `results/evaluation/rounds.csv`, `scripts/35_rounds_curve.py`.

Table 13. Accuracy and dispersion of the committee by round, validation subsample.

Arm	Committee	Round	n	Mean $\lvert e\rvert$ (median)	95% CI	Mean $\lvert e\rvert$ (mean)	IQR	95% CI
G1	mixed	0	40	0.180	[0.127, 0.244]	0.177	0.069	[0.048, 0.095]
		1	40	0.178	[0.126, 0.241]	0.175	0.043	[0.031, 0.058]
G2	mixed	0	40	0.180	[0.128, 0.243]	0.178	0.070	[0.049, 0.098]
		1	40	0.178	[0.127, 0.240]	0.174	0.044	[0.031, 0.061]
		2	40	0.176	[0.126, 0.238]	0.173	0.034	[0.027, 0.044]
G2	all-Anthropic	0	40	0.181	[0.129, 0.245]	0.182	0.056	[0.037, 0.080]
		1	40	0.179	[0.128, 0.242]	0.176	0.026	[0.019, 0.035]
		2	40	0.178	[0.127, 0.241]	0.177	0.018	[0.013, 0.023]
G2	all-OpenAI	0	40	0.177	[0.124, 0.242]	0.173	0.042	[0.033, 0.052]
		1	40	0.178	[0.124, 0.242]	0.174	0.039	[0.031, 0.048]
		2	40	0.178	[0.124, 0.242]	0.174	0.038	[0.031, 0.048]
G4	mixed	0	20	0.162	[0.082, 0.267]	0.168	0.057	[0.040, 0.076]
		1	20	0.163	[0.083, 0.266]	0.161	0.032	[0.024, 0.042]
		2	20	0.163	[0.083, 0.266]	0.161	0.026	[0.019, 0.035]
		3	20	0.163	[0.083, 0.266]	0.161	0.025	[0.018, 0.034]
		4	20	0.163	[0.082, 0.266]	0.161	0.024	[0.017, 0.033]
G-fullview	mixed	0	40	0.173	[0.122, 0.236]	0.177	0.072	[0.046, 0.108]
		1	40	0.174	[0.124, 0.236]	0.174	0.036	[0.024, 0.051]
		2	40	0.175	[0.124, 0.237]	0.174	0.028	[0.020, 0.037]

Notes. For each property and round r, the twelve roles' current log estimates are pooled by their median and by their mean; the columns give the mean over properties of the absolute error of each pool and the mean interquartile range (IQR, in log points) of the twelve estimates. n is the number of properties; committees stopped early by the interquartile-range rule carry their last estimates forward, so that every round is computed on the same properties. 95% CIs are percentile block-bootstrap intervals (municipality $\times$ month blocks, 2,000 resamples). Source: `results/evaluation/rounds_validation.csv`, `scripts/35_rounds_curve.py`.

Table 14. Deliberation statistics by run: dispersion, convergence and anchoring.

Split	Arm	Committee	n	IQR_0	IQR_R	Convergence	Anchoring	To median
validation	F	all-Anthropic	40	0.058	0.058	1.000	–	–
validation	F	all-OpenAI	40	0.041	0.041	1.000	–	–
validation	F	mixed	40	0.064	0.064	1.000	–	–
validation	G1	mixed	40	0.069	0.043	0.664	0.672	0.427
validation	G2	all-Anthropic	40	0.056	0.018	0.457	0.970	0.525
validation	G2	all-OpenAI	40	0.042	0.038	0.944	0.384	0.075
validation	G2	mixed	40	0.070	0.034	0.560	0.673	0.442
validation	G4	mixed	20	0.057	0.024	0.469	0.781	0.429
validation	G-fullview	mixed	40	0.072	0.028	0.520	0.651	0.517
validation	H	mixed	40	0.074	–	–	–	–
validation	I	mixed	40	0.073	0.035	0.577	0.776	0.435
test (mid)	F	mixed	80	0.086	0.086	1.000	–	–
test (mid)	G2	mixed	80	0.088	0.038	0.549	0.713	0.451
test (mid)	H	mixed	80	0.092	–	–	–	–
test (mid)	I	mixed	80	0.087	0.041	0.612	0.751	0.429
test (small)	F	mixed	300	0.088	0.088	1.000	–	–
test (small)	G2	mixed	300	0.087	0.034	0.475	0.677	0.521
holdout 2026	F	mixed	60	0.075	0.075	1.000	–	–
holdout 2026	G2	mixed	60	0.073	0.037	0.528	0.700	0.465
prospective	F	mixed	40	0.077	0.077	1.000	–	–
prospective	G2	mixed	40	0.082	0.041	0.616	0.757	0.452

Notes. IQR_0 and IQR_R are the mean interquartile ranges of the twelve roles' log estimates at round 0 and at the final round; Convergence is their ratio; Anchoring is the mean over properties of the correlation across roles between the revision $x_{k,R} - x_{k,0}$ and the distance $m_0 - x_{k,0}$ to the initial committee median (1 = every revision is a move toward the initial median); To median is the share of roles whose final estimate is closer to m_0 than their initial one. Arm F has no rounds (convergence 1 by construction); arm H's skeptic sits outside the rounds and its final-round statistics are not defined. Source: `results/evaluation/faithfulness.csv`, `scripts/22_extra_tables.py`.

Table 15. Exploratory paired contrasts on validation: the full-view committee.

Contrast (B vs A)	Metric	n	A	B	Δ	95% CI	p (boot)
G-fullview vs G2 (evidence partition)	Mean $\|e\|$	40	0.176	0.175	−0.001	[-0.006, 0.004]	0.657
	MdAPE	40	0.113	0.113	−0.000	[-0.005, 0.009]	0.967
	P95 $\|e\|$	40	0.420	0.425	0.005	[-0.024, 0.019]	0.492
	Cov. 90%	40	0.825	0.850	0.025	[0.000, 0.077]	0.567
G-fullview vs I (prompts at full view)	Mean $\|e\|$	40	0.184	0.175	−0.009	[-0.027, 0.004]	0.259
	MdAPE	40	0.106	0.113	0.007	[-0.016, 0.039]	0.348
	P95 $\|e\|$	40	0.463	0.425	−0.038	[-0.178, 0.000]	0.354
	Cov. 90%	40	0.825	0.850	0.025	[0.000, 0.077]	0.546
G-fullview vs F	Mean $\|e\|$	40	0.179	0.175	−0.005	[-0.017, 0.004]	0.441
	MdAPE	40	0.118	0.113	−0.005	[-0.011, 0.037]	0.498
	P95 $\|e\|$	40	0.443	0.425	−0.018	[-0.102, 0.018]	0.710
	Cov. 90%	40	0.825	0.850	0.025	[0.000, 0.077]	0.567

Notes. Same conventions as Table 9: Δ = B − A on the same 40 validation properties, mixed committee, median aggregation; percentile block-bootstrap CI (municipality × month blocks, 2,000 resamples, seed 20261027) and recentred bootstrap p-value. These pairs were not preregistered and are not adjusted for multiplicity. Source: `results/evaluation_validation/records.parquet`, `scripts/22_extra_tables.py`.

Table 16. Provider committees on validation: accuracy, calibration, deliberation and contrasts against the mixed committee.

Arm	Committee	n	Mean $\|e\|$	P95 $\|e\|$	Bias	Cov. 90%	Width 90%	IQR_R	Anchoring	Δ vs mixed	95% CI	p
E	all-Anthropic	40	0.194	0.502	0.062	0.825	0.589	–	–	0.013	[-0.025, 0.048]	0.482
E	all-OpenAI	40	0.177	0.458	0.080	0.800	0.486	–	–	−0.005	[-0.016, 0.006]	0.398
E	mixed	40	0.182	0.457	0.065	0.800	0.493	–	–	–	–	–
F	all-Anthropic	40	0.182	0.454	0.049	0.825	0.581	–	–	0.003	[-0.023, 0.023]	0.776
F	all-OpenAI	40	0.178	0.463	0.073	0.825	0.507	–	–	−0.001	[-0.010, 0.007]	0.733
F	mixed	40	0.179	0.443	0.065	0.825	0.541	–	–	–	–	–
G2	all-Anthropic	40	0.178	0.433	0.043	0.825	0.582	0.018	0.970	0.002	[-0.014, 0.016]	0.741
G2	all-OpenAI	40	0.178	0.460	0.074	0.825	0.515	0.038	0.384	0.002	[-0.011, 0.019]	0.774
G2	mixed	40	0.176	0.420	0.056	0.825	0.553	0.034	0.673	–	–	–

Notes. All-Anthropic: every role served by `claude-sonnet-5-5`; all-OpenAI: by `gpt-6.1-sol`; mixed: seeded balanced assignment of the two. Median aggregation, 40 validation properties. Bias is the mean log error; Cov. 90% and Width 90% the raw coverage and mean log width of the 90 percent interval; IQR_R and Anchoring as in Table 14 (G2 only). Δ vs mixed is the mean $|e|$ of the single-provider committee minus that of the mixed committee of the same arm on the same properties, with its percentile block-bootstrap CI and recentred p-value (2,000 resamples). Source: `results/evaluation_validation/summary.csv`, `results/runs/validation_*/predictions.parquet`, `scripts/22_extra_tables.py`.

Table 17. Mean absolute log error by aggregation rule, test partition.

Aggregation rule	F	G2	H	I	J	F (small)	G2 (small)
Median (primary)	0.159	0.158	0.158	0.157	0.163	0.176	0.180
Arithmetic mean	0.155	0.160	0.159	0.159	0.163	0.170	0.174
Trimmed mean (20%)	0.156	0.158	0.158	0.158	0.163	0.173	0.179
Inverse variance (validation weights)	0.155	0.160	0.159	0.159	0.163	0.171	0.175
Stacking (validation weights)	0.162	0.162	0.162	0.159	0.163	0.167	0.174
Robust Student-t pool	0.157	0.160	0.159	0.159	0.163	0.173	0.177
Mixture of role distributions	0.155	0.159	0.158	0.158	0.163	0.172	0.177
J: mean of 12 samples	–	–	–	–	0.164	–	–
J: median of 12 samples	–	–	–	–	0.163	–	–
J: self-refined estimate	–	–	–	–	0.161	–	–

Notes. Each cell is the mean $|e|$ of the same run under a different pooling of the twelve roles' final log estimates (80 properties at the mid tier, 300 at the small tier). Inverse-variance and stacking weights, and the scale of the Student-t pool, were estimated once on the validation run of G2 (`results/aggregation_weights.json`) and applied unchanged; the mixture pools the roles' quantile functions with equal weights. For arm J the twelve rows are the independent samples of the generalist; "self-refined" is the estimate after the critique-and-refine step. Source: `results/evaluation/summary.csv`, `scripts/22_extra_tables.py`.

Table 18. Spatial holdout: in-region and withheld-region baselines paired on the withheld sales.

Arm	Metric	n	In-region model	Withheld model	Δ	95% CI	p (boot)
A	Mean $\lvert e\rvert$	8988	0.183	0.208	0.025	[0.019, 0.030]	<0.001
	MdAPE	8988	0.132	0.154	0.023	[0.016, 0.030]	<0.001
	P95 $\lvert e\rvert$	8988	0.542	0.552	0.010	[-0.009, 0.024]	0.251
	Bias $\bar{e}$	8988	0.021	−0.078	−0.099	[-0.120, -0.069]	<0.001
	Cov. 90%	8988	0.921	0.886	−0.035	[-0.045, -0.024]	<0.001
C (gbm)	Mean $\lvert e\rvert$	8988	0.187	0.176	−0.011	[-0.013, -0.009]	<0.001
	MdAPE	8988	0.141	0.128	−0.013	[-0.017, -0.010]	<0.001
	P95 $\lvert e\rvert$	8988	0.546	0.537	−0.008	[-0.019, 0.004]	0.185
	Bias $\bar{e}$	8988	0.048	0.050	0.001	[-0.005, 0.009]	0.780
	Cov. 90%	8988	–	–	–	–	–
D	Mean $\lvert e\rvert$	8988	0.163	0.161	−0.002	[-0.006, 0.001]	0.254
	MdAPE	8988	0.106	0.105	−0.001	[-0.007, 0.002]	0.654
	P95 $\lvert e\rvert$	8988	0.525	0.527	0.001	[-0.012, 0.013]	0.845
	Bias $\bar{e}$	8988	0.022	0.045	0.023	[0.019, 0.027]	<0.001
	Cov. 90%	8988	0.910	0.907	−0.004	[-0.009, 0.002]	0.175

Notes. n = 8,988 test-window sales in Outaouais, Abitibi-Témiscamingue and Bas-Saint-Laurent. "In-region model" is the baseline trained on the full training partition; "Withheld model" the same specification trained without the three regions. Δ = withheld − in-region, with percentile block-bootstrap CI and recentred p-value (municipality × month blocks, 2,000 resamples, seed 20261027); p values below 0.001 are floored. Arm C has no intervals. Source: `results/evaluation/records.parquet`, `scripts/22_extra_tables.py`.

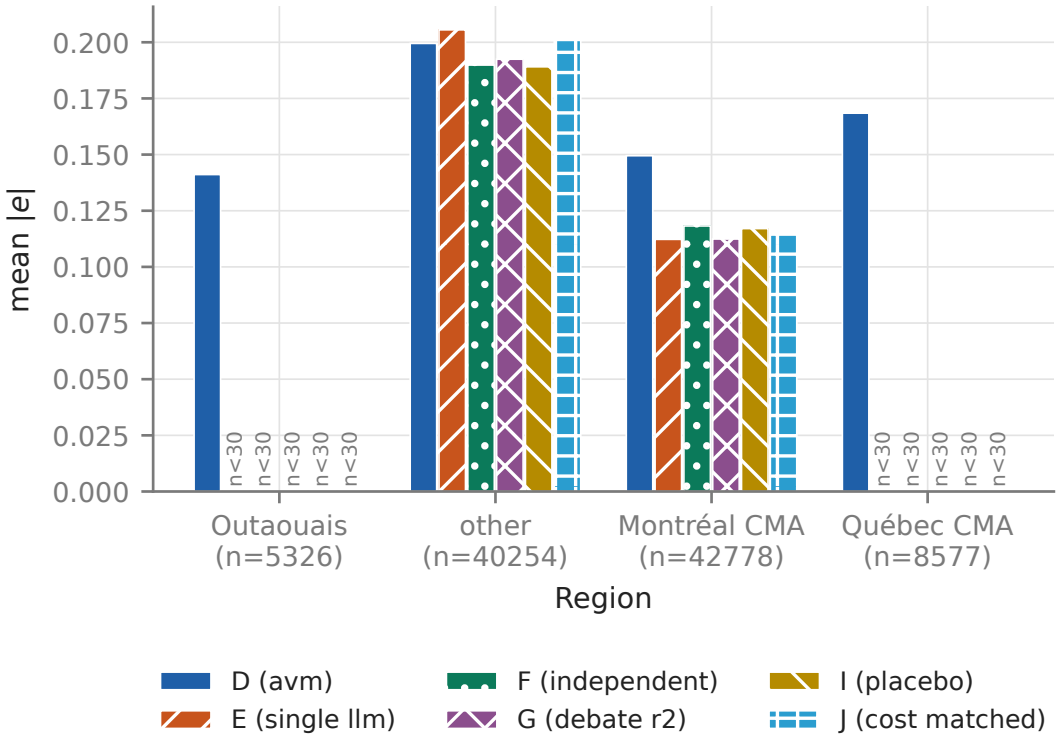


Figure 7. Mean absolute log error by region group and arm on the test partition. The AVM (D) is scored on the full partition cell (*n* in the axis labels); the mid-tier agentic arms on the cells of the 80-property subsample, of which only "other" and "Montréal CMA" reach the thirty-property threshold; suppressed cells are marked $n < 30$. Source: `results/evaluation/strata.csv`, `scripts/21_figures.py`.

Table 19. Mean absolute log error by stratum, test partition.

Stratum	*n* full	*n* mid	*n* small	D	E	F	G2	I	J	G2 (small)
Region: Outaouais	5326	2	14	0.141	–	–	–	–	–	–
Region: other	40254	36	128	0.199	0.206	0.190	0.192	0.189	0.201	0.223
Region: Montréal CMA	42778	36	134	0.149	0.112	0.118	0.112	0.117	0.114	0.137
Region: Québec CMA	8577	6	24	0.168	–	–	–	–	–	–
Urban class: metro	40948	37	127	0.154	0.157	0.162	0.158	0.160	0.156	0.151
Urban class: suburban	33826	23	93	0.162	–	–	–	–	–	0.146
Urban class: rural	22161	20	80	0.219	–	–	–	–	–	0.267
Property class: house	65306	57	199	0.179	0.165	0.156	0.155	0.155	0.162	0.200
Property class: condo	18763	15	60	0.115	–	–	–	–	–	0.088
Property class: plex	12866	8	41	0.215	–	–	–	–	–	0.220
Price tercile: 1	21029	23	72	0.319	–	–	–	–	–	0.337
Price tercile: 2	33482	25	99	0.131	–	–	–	–	–	0.125
Price tercile: 3	42424	32	129	0.130	0.113	0.122	0.116	0.118	0.113	0.135
Construction decade: 1950s	7566	9	21	0.205	–	–	–	–	–	–
Construction decade: 1960s	9253	4	18	0.196	–	–	–	–	–	–
Construction decade: 1970s	12570	12	40	0.185	–	–	–	–	–	0.274
Construction decade: 1980s	11595	15	43	0.162	–	–	–	–	–	0.151
Construction decade: 1990s	9925	6	32	0.150	–	–	–	–	–	0.132
Construction decade: 2000s	13104	10	45	0.141	–	–	–	–	–	0.117
Construction decade: 2010+	20101	12	63	0.137	–	–	–	–	–	0.160
Construction decade: <1950	11834	10	34	0.229	–	–	–	–	–	0.248
Construction decade: unknown	987	2	4	0.260	–	–	–	–	–	–

Notes. *n* full, *n* mid and *n* small are the cell sizes in the full test partition (on which D is scored), in the 80-property mid-tier subsample (E, F, G2, I, J) and in the 300-property small-tier subsample (G2 small). Cells with fewer than 30 properties are suppressed (–). Price terciles are terciles of the sale price over the test partition. The full stratified table with every arm and metric is `results/tables/strata.tex`. Source: `results/evaluation/strata.csv`, `scripts/22_extra_tables.py`.

Table 20. Interval coverage, width and proper scores by arm, test partition.

Arm	Cov. 50%	Cov. 80%	Cov. 90%	Cov. 95%	Width 90% (log)	Pinball (log)	CRPS (log)	n_with_interval
A (hedonic)	0.489	0.869	0.909	0.939	1.001	0.065	0.133	96935
A (hedonic spatial)	0.423	0.834	0.886	0.952	1.012	0.065	0.134	8988
B (hierarchical)	0.499	0.874	0.913	0.939	0.998	0.063	0.129	96935
C (catboost)	–	–	–	–	–	–	–	0
C (gbm)	–	–	–	–	–	–	–	0
C (gbm spatial)	–	–	–	–	–	–	–	0
C (xgboost)	–	–	–	–	–	–	–	0
D (avm)	0.446	0.868	0.896	0.936	0.830	0.058	0.119	96935
D (avm spatial)	0.425	0.856	0.907	0.938	0.741	0.058	0.120	8988
E (single llm)	0.463	0.762	0.887	0.912	0.562	0.055	0.114	80
E (single llm small)	0.417	0.670	0.777	0.813	0.464	0.064	0.129	300
F (independent)	0.512	0.775	0.887	0.900	0.590	0.053	0.109	80
F (independent small)	0.420	0.680	0.787	0.840	0.451	0.061	0.124	300
G (debate r2)	0.475	0.787	0.887	0.900	0.602	0.053	0.109	80
G (debate r2 small)	0.427	0.703	0.807	0.860	0.489	0.062	0.126	300
H (debate skeptic)	0.525	0.800	0.875	0.912	0.604	0.053	0.109	80
I (placebo)	0.487	0.812	0.887	0.887	0.590	0.052	0.108	80
J (cost matched)	0.475	0.787	0.863	0.900	0.561	0.054	0.112	80

Notes. Cov. 50% and 90% are the raw shares of sale prices inside $[q_{.25}, q_{.75}]$ and $[q_{.05}, q_{.95}]$; Cov. 80% and 95% are interpolated from a normal distribution fitted in log space to the five quantiles on the probit grid and are model-based. Width 90% is the mean of $\log q_{.95} - \log q_{.05}$. Pinball is the average over the five quantiles of the pinball loss in log units; CRPS the trapezoidal approximation of the continuous ranked probability score on the five-point grid, lower-biased and comparable across arms. Arm C produces no intervals. Samples as in Table 8. Source: `results/evaluation/summary_main.csv`.

Table 21. Split-conformal rescaling of the agentic intervals with validation factors.

Split	Arm	n	Cov. 50% raw	Cov. 90% raw	Width 90% raw	k_{50}	k_{90}	Cov. 50% conf.	Cov. 90% conf.	Width 90% conf.
test (mid)	E	80	0.463	0.887	0.562	1.264	1.629	0.525	0.925	0.916
	F	80	0.512	0.887	0.590	1.261	1.584	0.613	0.950	0.934
	G2	80	0.475	0.887	0.602	1.264	1.548	0.637	0.938	0.932
	J	80	0.475	0.863	0.561	1.243	1.532	0.525	0.938	0.859
holdout 2026	E	60	0.467	0.800	0.497	1.264	1.629	0.633	0.883	0.810
	F	60	0.517	0.833	0.508	1.261	1.584	0.583	0.900	0.804
	G2	60	0.533	0.833	0.528	1.264	1.548	0.583	0.883	0.817
prospective	E	40	0.425	0.750	0.558	1.264	1.629	0.600	0.900	0.909
	F	40	0.500	0.800	0.574	1.261	1.584	0.575	0.925	0.910
	G2	40	0.450	0.825	0.571	1.264	1.548	0.575	0.925	0.884

Notes. Raw columns are the stated intervals. k_{50} and k_{90} are the split-conformal factors of the asymmetric scaled-residual score, estimated on the 40 validation properties of the same arm and mixed committee (`results/evaluation/llm_conformal_scales.json`) and applied unchanged: the rescaled 90 percent interval is $[\exp(c - k_{90}\ell), \exp(c + k_{90}u)]$ with $c = \log q_{.50}$, $\ell = c - \log q_{.05}$, $u = \log q_{.95} - c$, and likewise at 50 percent with the quartiles. Under exchangeability the rescaled coverage is at least nominal. Arms H and I have no validation factors and are omitted. Source: `results/evaluation{,_holdout,_prospective}/records.parquet`, `scripts/22_extra_tables.py`.

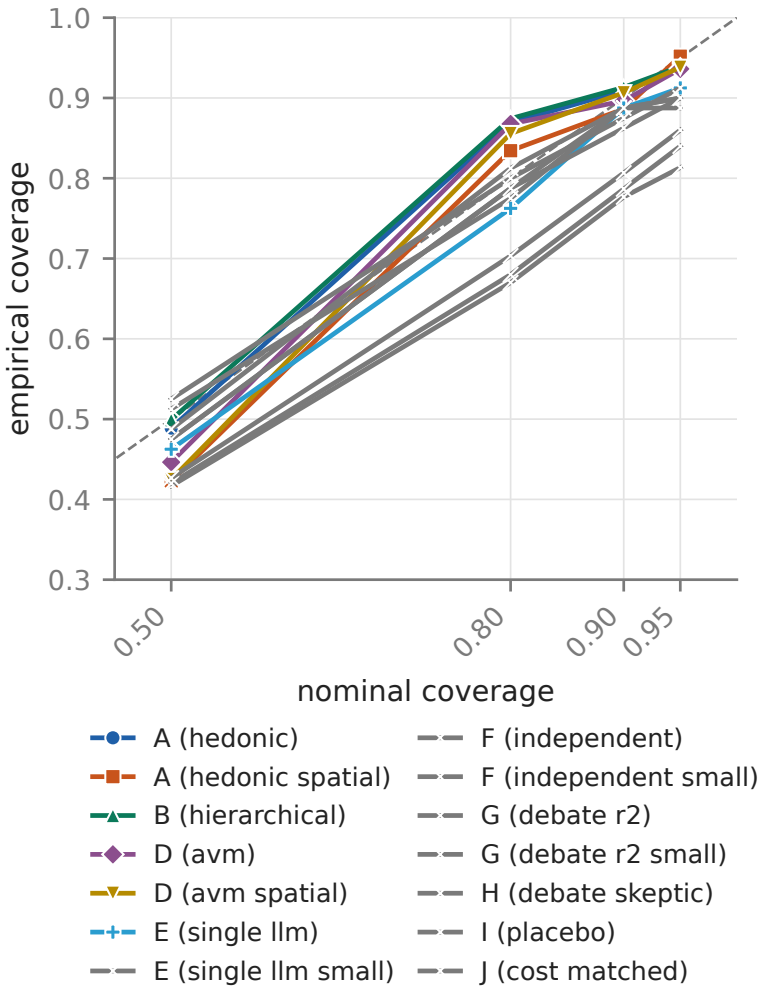


Figure 8. Empirical against nominal interval coverage by arm on the test partition at the 50, 80, 90 and 95 percent levels (80 and 95 interpolated). The dashed diagonal is perfect calibration; arms beyond the sixth are drawn in grey. Source: `results/evaluation/coverage.csv`, `scripts/21_figures.py`.

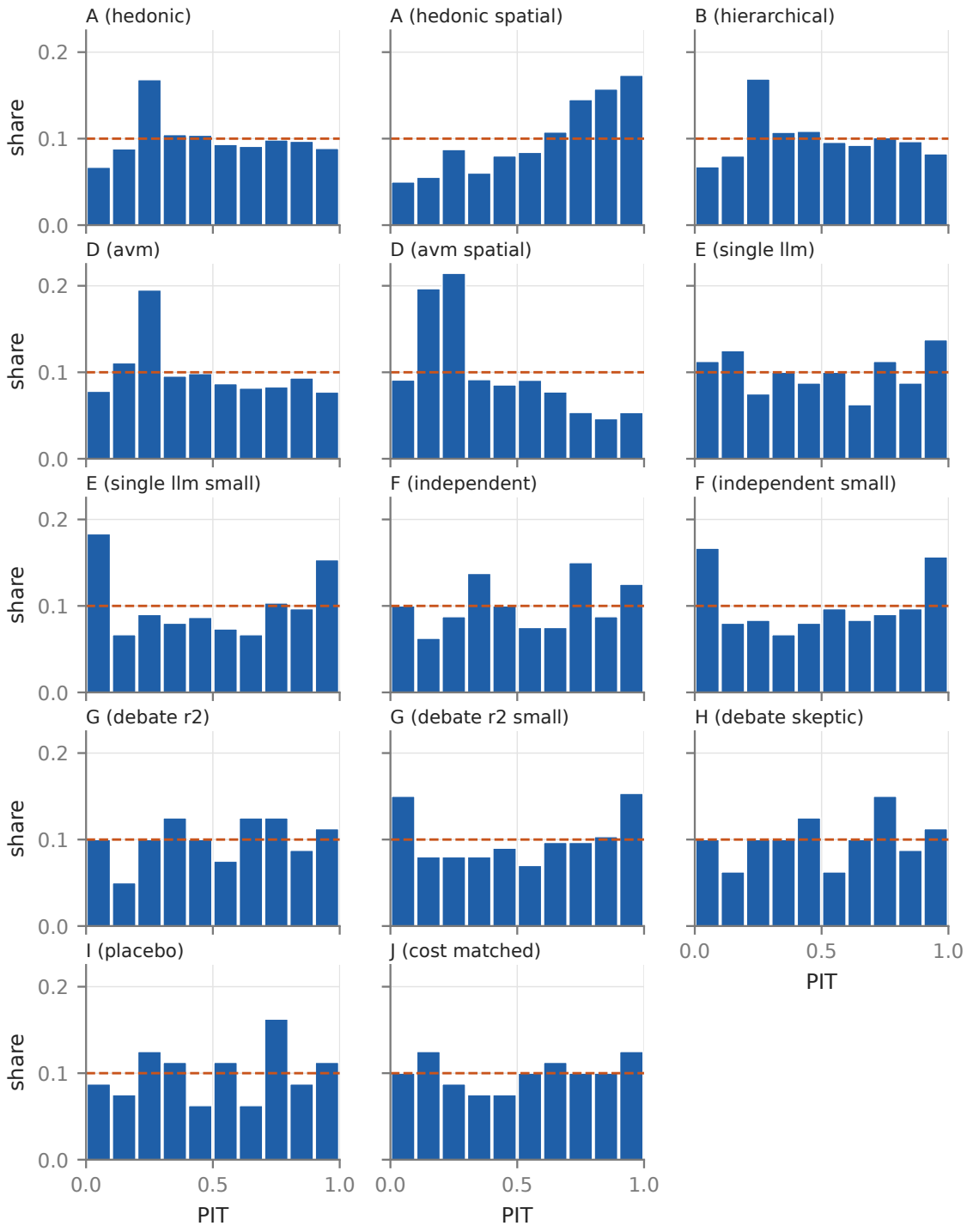


Figure 9. Probability integral transform histograms by arm on the test partition, ten bins; the dashed line is the uniform reference. The PIT is $F_i(\log P_i)$ with F_i interpolated in probit space between the five stated quantiles and normal tails beyond them. Source: `results/evaluation/pit.csv`, `scripts/21_figures.py`.

Table 22. Conditional coverage of the 90 percent (upper panel) and 50 percent (lower panel) intervals by subgroup, test partition.

Subgroup	*n* mid	*n* small	D	E	F	G2	H	I	J	G2 (small)
Region: Montréal CMA	36	134	0.911	0.944	0.944	0.944	0.944	0.944	0.917	0.843
Region: Outaouais	2	14	0.932	–	–	–	–	–	–	–
Region: Québec CMA	6	24	0.855	–	–	–	–	–	–	–
Region: other	36	128	0.885	0.861	0.861	0.861	0.833	0.861	0.861	0.797
Property class: condo	15	60	0.910	–	–	–	–	–	–	0.900
Property class: house	57	199	0.896	0.895	0.877	0.877	0.860	0.877	0.860	0.789
Property class: plex	8	41	0.877	–	–	–	–	–	–	0.756
Price tercile: 1	23	72	0.770	–	–	–	–	–	–	0.667
Price tercile: 2	25	99	0.936	–	–	–	–	–	–	0.848
Price tercile: 3	32	129	0.928	0.906	0.906	0.906	0.906	0.906	0.844	0.853
Region: Montréal CMA	36	134	0.447	0.528	0.556	0.528	0.556	0.500	0.528	0.500
Region: Outaouais	2	14	0.385	–	–	–	–	–	–	–
Region: Québec CMA	6	24	0.386	–	–	–	–	–	–	–
Region: other	36	128	0.466	0.417	0.500	0.444	0.528	0.500	0.444	0.359
Property class: condo	15	60	0.424	–	–	–	–	–	–	0.617
Property class: house	57	199	0.447	0.439	0.491	0.474	0.509	0.474	0.439	0.377
Property class: plex	8	41	0.475	–	–	–	–	–	–	0.390
Price tercile: 1	23	72	0.347	–	–	–	–	–	–	0.222
Price tercile: 2	25	99	0.478	–	–	–	–	–	–	0.545
Price tercile: 3	32	129	0.470	0.531	0.562	0.500	0.562	0.500	0.562	0.450

Notes. Raw coverages within each subgroup. D is scored on the full test partition cell; the mid-tier arms E–J on the cells of the 80-property subsample (*n* mid) and the small-tier debate on the cells of the 300-property subsample (*n* small); cells with fewer than 30 properties are suppressed (–). Source: `results/evaluation/coverage.csv`, `scripts/22_extra_tables.py`.

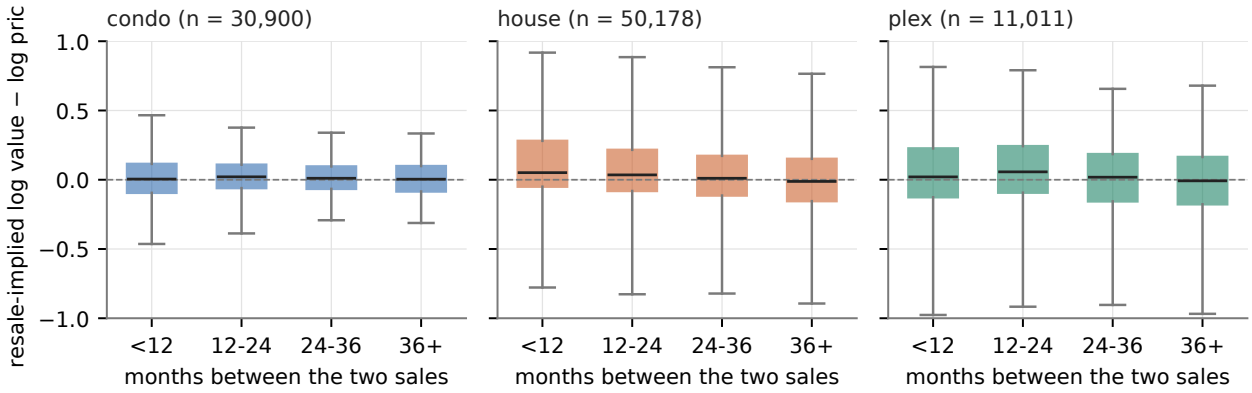


Figure 10. Resale-implied log value minus realised log price, by property class and holding period (boxes: interquartile range; whiskers: 5th–95th percentiles). A value of zero means that the later price, deflated by the index, reproduces the first price exactly. Source: `scripts/50_repeat_sales.py`.

Table 23. Faithfulness audit of the transcripts by run.

Split	Arm	Committee	n	Turns	Valid ids	Adj. ±40%	Turns ok	In hull	Quantiles ok	Words	≤120	Retries
test (mid)	E	mixed	80	80	1.000	0.954	0.900	1.000	1.000	77	1.000	0.000
test (mid)	F	mixed	80	960	1.000	0.983	0.709	0.994	0.991	78	1.000	0.000
test (mid)	G2	mixed	80	2856	1.000	0.988	0.721	0.998	0.988	78	1.000	0.001
test (mid)	H	mixed	80	2654	1.000	0.992	0.717	0.998	0.985	78	1.000	0.001
test (mid)	I	mixed	80	2832	1.000	0.981	0.711	0.995	0.988	72	1.000	0.001
test (mid)	J	mixed	80	1040	1.000	0.965	0.917	0.997	0.996	76	1.000	0.001
test (small)	E	mixed	300	300	1.000	0.992	0.977	0.993	0.997	77	1.000	0.017
test (small)	F	mixed	300	3600	1.000	0.985	0.716	0.996	0.998	78	1.000	0.025
test (small)	G2	mixed	300	10740	1.000	0.986	0.702	0.996	0.995	80	1.000	0.025
holdout 2026	E	mixed	60	60	1.000	0.997	0.983	1.000	1.000	74	1.000	0.033
holdout 2026	F	mixed	60	720	1.000	0.995	0.732	0.998	0.994	78	1.000	0.010
holdout 2026	G2	mixed	60	2147	1.000	0.992	0.728	1.000	0.992	77	1.000	0.001
prospective	E	mixed	40	40	1.000	0.995	0.975	0.975	1.000	71	1.000	0.000
prospective	F	mixed	40	479	1.000	0.981	0.704	0.992	1.000	77	1.000	0.002
prospective	G2	mixed	40	1404	1.000	0.983	0.707	0.988	0.999	77	1.000	0.002
validation	E	all-Anthropic	40	40	1.000	1.000	1.000	1.000	1.000	63	1.000	0.000
validation	E	all-OpenAI	40	40	1.000	0.985	0.950	0.975	1.000	83	1.000	0.000
validation	E	mixed	40	40	1.000	0.956	0.900	1.000	1.000	76	1.000	0.000
validation	F	all-Anthropic	40	480	1.000	0.995	0.738	1.000	0.992	68	1.000	0.004
validation	F	all-OpenAI	40	480	1.000	0.979	0.710	0.994	0.998	86	1.000	0.000
validation	F	mixed	40	480	1.000	0.987	0.731	0.994	1.000	77	1.000	0.002
validation	G1	mixed	40	960	1.000	0.988	0.727	0.999	0.996	78	1.000	0.001
validation	G2	all-Anthropic	40	1428	1.000	0.999	0.748	1.000	0.993	69	1.000	0.001
validation	G2	all-OpenAI	40	1104	1.000	0.986	0.719	0.991	0.998	87	1.000	0.000
validation	G2	mixed	40	1440	1.000	0.978	0.713	0.999	1.000	78	1.000	0.001
validation	G4	mixed	20	1020	1.000	0.992	0.726	0.999	1.000	78	1.000	0.000
validation	G-fullview	mixed	40	1416	1.000	0.984	0.883	0.997	0.993	78	1.000	0.003
validation	H	mixed	40	1360	1.000	0.994	0.717	0.999	0.998	78	1.000	0.001
validation	I	mixed	40	1404	1.000	0.997	0.734	0.999	0.997	73	1.000	0.003
validation	J	mixed	40	520	1.000	0.978	0.935	0.979	1.000	76	1.000	0.002

Notes. n is the number of properties and Turns the number of estimate or revision turns audited. Valid ids: share of cited evidence identifiers present in the role's packet (H6 threshold 0.90). Adj. ±40%: share of stated adjustments with $|\text{adjusted}/\text{raw} - 1| \leq 0.40$ where raw is the comparable's price in the packet; Turns ok: share of turns in which every adjustment passes (H6 threshold 0.80). In hull: share of turns whose point lies in $[\min \text{adjusted} \times 0.75, \max \text{adjusted} \times 1.25]$. Quantiles ok: positive, non-decreasing, point inside $[q_{.05}, q_{.95}]$ and $q_{.95}/q_{.05} \in [1.01, 4]$. Words: mean length of the reasoning summary; ≤120: share within the limit; Retries: mean number of schema-validation retries per turn. Every transcript's packet hash matches the rebuilt packet. The deliberation statistics of the same runs are in Table 14. Source: `results/evaluation/faithfulness.csv`, `scripts/22_extra_tables.py`.

Table 24. The resale sample: analytic sales with a later resale of the same parcel in the snapshot. Hold <12 m is the share of pairs held less than twelve months; Months is the median holding period; Ann. return is the median annualised log return; gap is $\log V^R - \log P$, the resale-implied log value of equation (14) minus the realised log price. Source: `scripts/50_repeat_sales.py`, `results/fundamental/resale_sample.csv`.

Sample	N	Hold <12 m	Months	Ann. return	Mean gap	SD gap	Med. |gap|
All pairs	92089	0.288	23.5	0.080	0.015	0.438	0.128
Condo	30900	0.505	11.7	0.059	0.004	0.301	0.094
House	50178	0.174	27.2	0.088	0.028	0.491	0.146
Plex	11011	0.199	27.5	0.091	−0.010	0.500	0.174
Montréal CMA	39256	0.350	21.6	0.058	0.011	0.423	0.112
Québec CMA	8641	0.255	24.5	0.097	−0.001	0.402	0.117
Outaouais	5731	0.332	22.1	0.068	0.049	0.354	0.100
other	38461	0.226	24.9	0.099	0.018	0.471	0.154
Held <12 months	26539	1.000	3.7	0.133	0.033	0.401	0.121
Held 12-24 months	20501	0.000	17.9	0.092	0.042	0.451	0.125
Held 24-36 months	18890	0.000	29.7	0.074	0.012	0.435	0.124
Held 36+ months	26159	0.000	46.5	0.072	−0.021	0.462	0.138
First sale in train	80559	0.232	26.6	0.080	0.014	0.437	0.130
First sale in validation	5251	0.441	13.1	0.095	0.032	0.474	0.116
First sale in test	3778	0.804	6.1	0.076	0.025	0.426	0.100
First sale in holdout	2501	1.000	1.0	0.002	−0.003	0.388	0.127

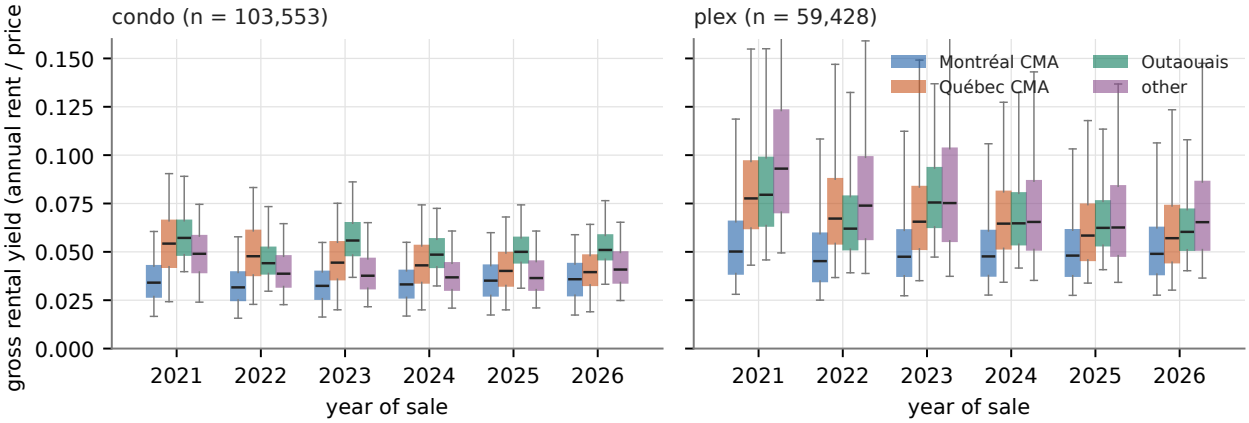


Figure 11. Crude gross rental yields by region group and year of sale (boxes: interquartile range; whiskers: 5th–95th percentiles; cells with fewer than 30 sales omitted). The yield is twelve times the municipality-year median rent of the registered leases, times the number of dwellings for plexes, divided by the sale price; no adjustment for the size or quality of the sold unit is possible. Source: `scripts/52_rents_usercost.py`.

Table 25. Inputs of the user-cost diagnostic $u = i + \tau_p + \delta + \gamma - \pi^e$. The mortgage rate is measured (Bank of Canada Valet series V80691335, weekly, annual means); the other inputs are assumptions and scenarios. Source: `scripts/52_rents_usercost.py`, `results/fundamental/usercost_parameters.csv`.

Parameter	Value	Status
i: 5-year conventional mortgage rate (annual mean of weekly observations)	2021: 4.79%; 2022: 5.65%; 2023: 6.68%; 2024: 6.73%; 2025: 6.23%; 2026: 6.10%	measured
τ_p: property tax rate (municipal and school, share of market value)	1.0%	assumption
δ: maintenance and depreciation	2.5%	assumption
γ: risk and liquidity premium	2.0%	assumption
expected nominal appreciation π^e (three scenarios)	0%, 2%, 4%	scenario

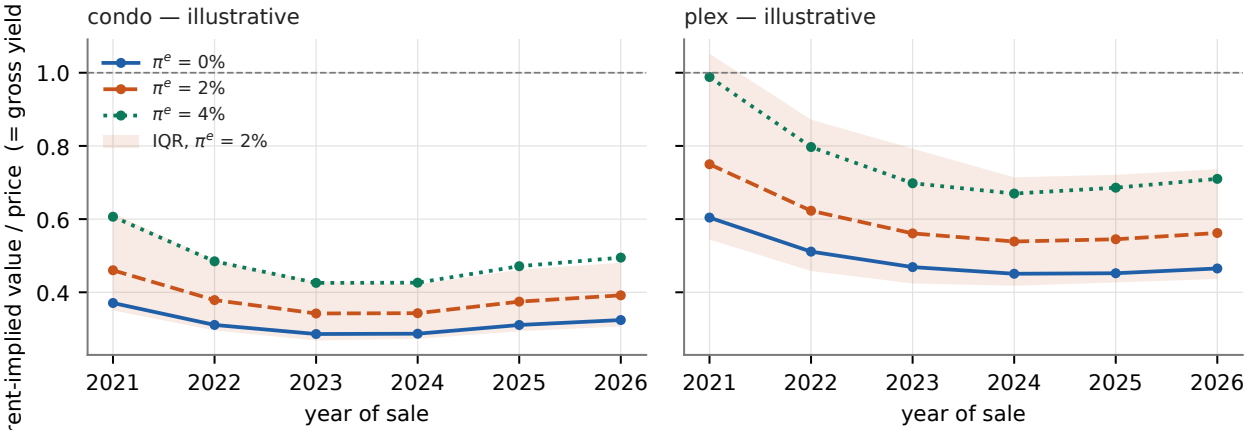


Figure 12. Illustrative: median rent-implied value relative to the price, $(R/u)/P$, by year of sale under three expected-appreciation scenarios, with the interquartile range for the central scenario. The gross yield is the crude measure of Figure 11; u uses the inputs of Table 25. Levels are not interpretable; the time profile and the dispersion are. Source: `scripts/52_rents_usercost.py`.

Table 26. Arm errors against the realised price and against the resale-implied value V^R on the evaluated properties with a usable later resale (holding period of at least twelve months, no data-quality flag on the later leg). e^P and e^R are the log errors against the realised price and against V^R; Δ is the paired difference between the arm's mean $|e^R|$ and that of the realised price used as an estimate, with the block-bootstrap 95 per cent interval and Holm-adjusted p-value; the reason for a placeholder is recorded in `results/fundamental/eval_vs_resale.json`. Source: `scripts/51_eval_vs_resale.py`, `results/fundamental/eval_vs_resale.csv`.

Arm	N	Mean $\lvert e^P \rvert$	Mean $\lvert e^R \rvert$	Δ	95% CI	p (Holm)
A (hedonic)	729	0.215	0.239	−0.004	[-0.024, 0.014]	1.000
A (hedonic spatial)	63	0.248	0.274	0.040	[-0.020, 0.095]	1.000
B (hierarchical)	729	0.209	0.230	−0.014	[-0.032, 0.004]	1.000
C (catboost)	729	0.195	0.222	−0.021	[-0.041, -0.003]	0.257
C (gbm)	729	0.196	0.221	−0.022	[-0.042, -0.003]	0.225
C (gbm spatial)	63	0.206	0.245	0.011	[-0.046, 0.067]	1.000
C (xgboost)	729	0.196	0.213	−0.031	[-0.050, -0.012]	0.030
D (avm)	729	0.184	0.216	−0.027	[-0.046, -0.010]	0.038
D (avm spatial)	63	0.186	0.231	−0.003	[-0.060, 0.050]	1.000
Realised price P_{t_1}	729	–	0.243	–	–	–

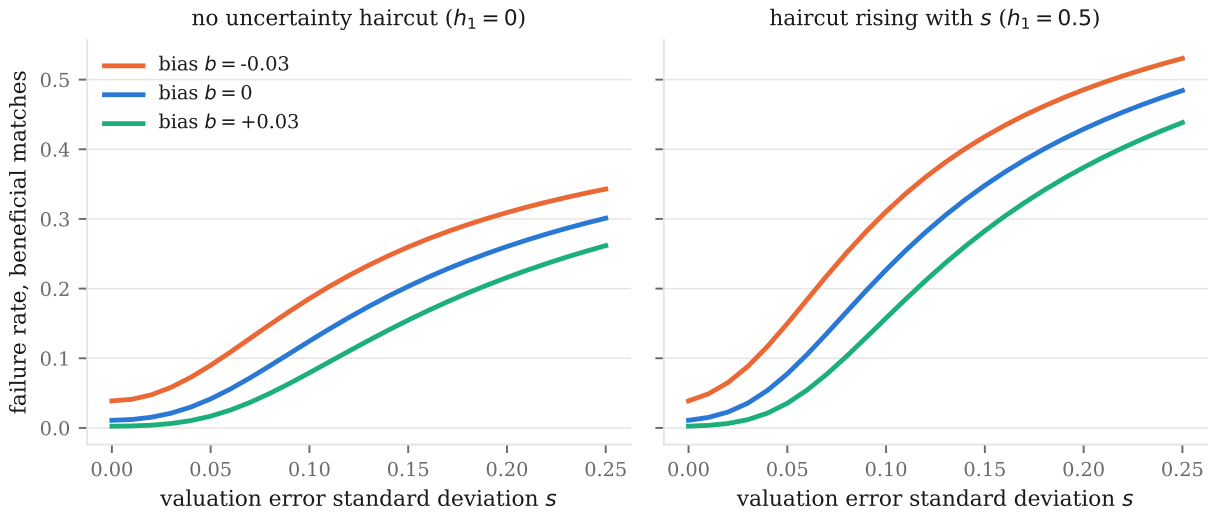


Figure 13. Share of mutually beneficial matches that fail for lack of financing as a function of the valuation error standard deviation s, for three values of the bias b, without an uncertainty haircut (left) and with $h = 0.5s$ (right). Equation (15) integrated over the seller draws, 200,000 draws. Illustrative parameters; see text. Source: `scripts/40_welfare_sim.py`.

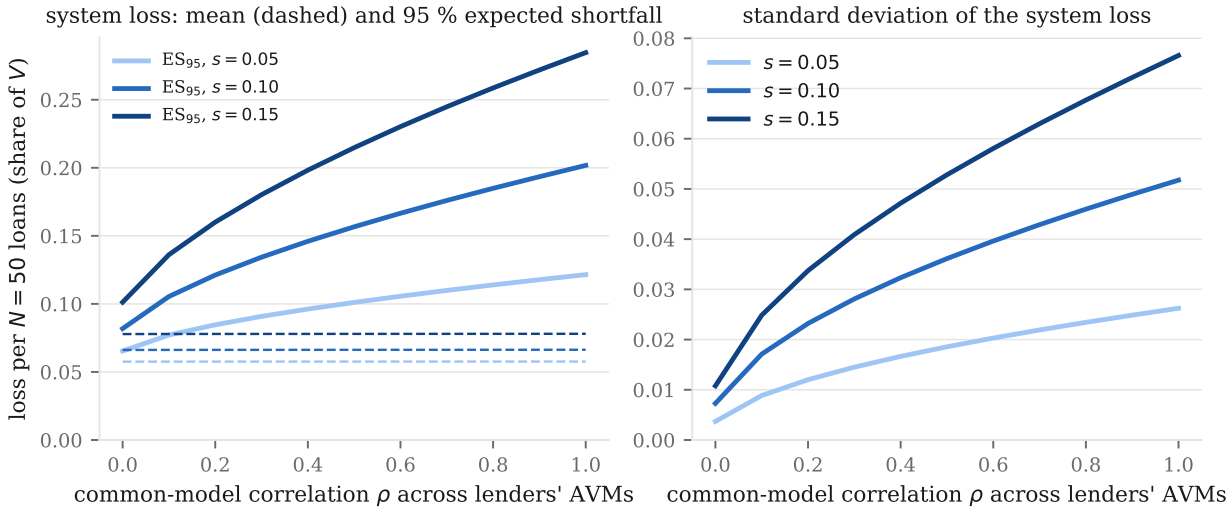


Figure 14. System loss of $N = 50$ lenders whose appraisal errors have common-model correlation ρ, for three values of the per-lender valuation error s. Left: mean (dashed, flat in ρ by Proposition 11(i)) and 95 percent expected shortfall. Right: standard deviation. Losses are expected losses conditional on the appraisal, as fractions of V, summed over lenders; 20,000 draws with common random numbers across ρ. Illustrative parameters; see text. Source: `scripts/40_welfare_sim.py`.

Table A1. Phases of the pipeline and the command that reproduces each.

Phase	Command (from the repository root)
Install and test	`python3 -m pip install -e ".[dev]"; python3 -m pytest tests -q`
Synthetic layer (no data needed)	`make synth`
Analytic table, partitions, leakage audit	`make data` (`scripts/01_build_analytic.py`, `02_splits.py`, `05_leakage_audit.py`)
Geographic layers, feature matrix	`make features` (`scripts/04_geo.py`, `03_features.py`)
Baselines A–D	`make baselines` (`scripts/10_train_baselines.py`)
Theory and welfare simulations	`make theory` (`scripts/41_theory_sim.py`, `40_welfare_sim.py`)
Repeat sales, rents, user cost	`make fundamental` (`scripts/50_repeat_sales.py`, `52_rents_usercost.py`, `51_eval_vs_resale.py`)
Mock smoke run of the agentic arms	`AV_PROVIDER_MODE=mock make smoke`
Live tiers (keys in `.env`)	`make pilot`; `make validation`; `make weights` (aggregation weights, conformal scales, freeze); `make final`; holdout and prospective runs as documented in `configs/budgets.yaml` and `docs/prospective_set.md`
Run-to-run variation	`scripts/36_repeat_variation.py`
Evaluation, rounds curve, figures	`make evaluate` (`scripts/15_descriptives.py`, `20_evaluate.py`, `35_rounds_curve.py`, `21_figures.py`)
Tables and figures into the paper	`make tables`
Manuscript	`make paper` (`latexmk -pdf main.tex` in `paper/`)

Table A2. Fields of the two output schemas.

Field	Type	Meaning and constraint
AgentEstimate (estimate and revision turns)		
`point_estimate`	number	most likely market value in CAD as of the valuation date; positive; within $[q_{05}, q_{95}]$
`q05`, `q25`, `q50`, `q75`, `q95`	numbers	quantiles of the predictive distribution in CAD; positive and non-decreasing
`evidence_ids_used`	list of strings	packet identifiers actually relied upon (P, C01–C12, M1, G1)
`adjustments`	list of objects	each with `comparable_id`, `adjusted_price` (CAD after the stated adjustments) and `reason`
`key_assumptions`	list of strings	assumptions on which the estimate rests
`missing_data`	list of strings	fields and observations that were absent and would have mattered
`confidence_rationale`	string	why the interval has the width it has
`counterarguments`	list of strings	the strongest reasons the estimate could be wrong
`reasoning_summary`	string	at most 120 words; must mention the evidence identifiers that drove the estimate
CritiqueSet (critique turns): a list of Critique objects		
`target_role`	string	the committee member addressed
`agree`	boolean	whether the critic agrees with the target's estimate
`points`	list of strings	evidence-based objections or confirmations
`evidence_ids_cited`	list of strings	packet identifiers supporting the points
`suggested_value`	number or null	an alternative value, only when the critic can justify one

Table A3. Accuracy by arm, validation partition (2024-07 to 2025-03). Mean $|e|$ is the mean absolute log error; MAE in CAD; MdAPE the median absolute percentage error; P90 and P95 the empirical percentiles of $|e|$; bias the mean log error; the last column but one the share of properties with $|e| > \log 1.25$. Source: `results/evaluation_validation/summary.csv`.

Arm	Mean $\lvert e\rvert$	MAE (CAD)	MdAPE	RMSE (log)	P90 $\lvert e\rvert$	P95 $\lvert e\rvert$	Bias $\bar{e}$	$\lvert e\rvert > \log 1.25$	N
A (hedonic)	0.197	84618	0.125	0.296	0.451	0.680	0.008	0.280	72132
B (hierarchical)	0.190	81929	0.116	0.293	0.444	0.687	0.021	0.259	72132
C (catboost)	0.173	73856	0.099	0.284	0.396	0.691	0.039	0.212	72132
C (gbm)	0.183	78141	0.112	0.283	0.412	0.659	0.008	0.246	72132
C (xgboost)	0.181	78659	0.105	0.293	0.424	0.714	0.054	0.228	72132
D (avm)	0.180	77392	0.109	0.280	0.407	0.659	0.004	0.238	72132
E (single llm)	0.182	64424	0.118	0.259	0.419	0.457	0.065	0.275	40
F (independent)	0.179	63380	0.118	0.255	0.385	0.443	0.065	0.300	40
G (debate r1)	0.178	63019	0.115	0.254	0.377	0.420	0.059	0.300	40
G (debate r2)	0.176	62715	0.113	0.250	0.370	0.420	0.056	0.300	40
G (debate r4)	0.163	65063	0.100	0.270	0.349	0.422	0.042	0.250	20
G (fullview)	0.175	62351	0.113	0.250	0.346	0.425	0.059	0.300	40
H (debate skeptic)	0.175	62649	0.119	0.250	0.350	0.430	0.055	0.275	40
I (placebo)	0.184	64127	0.106	0.264	0.423	0.463	0.064	0.300	40
J (cost matched)	0.179	63449	0.118	0.257	0.385	0.476	0.070	0.275	40

Table A4. Paired contrasts, validation partition. Δ is the mean of $|e_i^B| - |e_i^A|$ over the same properties; the interval and the bootstrap p-value come from the paired block bootstrap over municipality-by-month cells (2,000 draws); Holm adjustment within the primary family; DM is the Diebold–Mariano p-value with the Harvey–Leybourne–Newbold correction. Source: `results/evaluation_validation/contrasts.csv`.

Contrast (B vs A)	Metric	N	A	B	Δ	95% CI	p (boot)	p (Holm)	p (DM)
G (debate r2) vs D (avm)	Mean $\lvert e\rvert$	40	0.185	0.176	−0.009	[-0.048, 0.028]	0.608	1.000	0.692
G (debate r2) vs F (independent)	Mean $\lvert e\rvert$	40	0.179	0.176	−0.003	[-0.013, 0.004]	0.453	1.000	0.417
G (debate r2) vs J (cost matched)	Mean $\lvert e\rvert$	40	0.179	0.176	−0.003	[-0.024, 0.014]	0.746	1.000	0.810
G (debate r2) vs I (placebo)	Mean $\lvert e\rvert$	40	0.184	0.176	−0.008	[-0.023, 0.003]	0.233	0.930	0.309
G (debate r2) vs I (placebo)	P95 $\lvert e\rvert$	40	0.463	0.420	−0.044	[-0.156, 0.011]	0.249	–	–
G (debate r2) vs E (single llm)	Mean $\lvert e\rvert$	40	0.182	0.176	−0.006	[-0.026, 0.012]	0.526	–	0.658
G (debate r2) vs E (single llm)	MdAPE	40	0.118	0.113	−0.005	[-0.061, 0.032]	0.707	–	–
G (debate r2) vs E (single llm)	P95 $\lvert e\rvert$	40	0.457	0.420	−0.037	[-0.185, 0.065]	0.609	–	–
G (debate r2) vs E (single llm)	$\lvert e\rvert > \log 1.25$	40	0.275	0.300	0.025	[0.000, 0.077]	0.581	–	–
G (debate r2) vs E (single llm)	Cov. 90%	40	0.800	0.825	0.025	[-0.051, 0.103]	0.701	–	–
G (debate r2) vs E (single llm)	Pinball (log)	40	0.062	0.058	−0.003	[-0.011, 0.003]	0.365	–	0.470
G (debate r2) vs H (debate skeptic)	Mean $\lvert e\rvert$	40	0.175	0.176	0.001	[-0.003, 0.006]	0.814	–	0.856
G (debate r2) vs H (debate skeptic)	MdAPE	40	0.119	0.113	−0.006	[-0.010, 0.007]	0.226	–	–
G (debate r2) vs H (debate skeptic)	P95 $\lvert e\rvert$	40	0.430	0.420	−0.010	[-0.025, 0.023]	0.362	–	–
G (debate r2) vs H (debate skeptic)	$\lvert e\rvert > \log 1.25$	40	0.275	0.300	0.025	[0.000, 0.077]	0.578	–	–
G (debate r2) vs H (debate skeptic)	Cov. 90%	40	0.825	0.825	0.000	[0.000, 0.000]	1.000	–	–
G (debate r2) vs H (debate skeptic)	Pinball (log)	40	0.058	0.058	0.000	[-0.001, 0.002]	0.708	–	0.756
G (debate r2) vs A (hedonic)	Mean $\lvert e\rvert$	40	0.211	0.176	−0.035	[-0.076, 0.007]	0.099	–	0.143
G (debate r2) vs A (hedonic)	MdAPE	40	0.157	0.113	−0.044	[-0.091, 0.036]	0.223	–	–
G (debate r2) vs A (hedonic)	P95 $\lvert e\rvert$	40	0.567	0.420	−0.147	[-0.234, 0.070]	0.070	–	–
G (debate r2) vs A (hedonic)	$\lvert e\rvert > \log 1.25$	40	0.275	0.300	0.025	[-0.125, 0.179]	0.826	–	–
G (debate r2) vs A (hedonic)	Cov. 90%	40	0.925	0.825	−0.100	[-0.244, 0.025]	0.175	–	–
G (debate r2) vs A (hedonic)	Pinball (log)	40	0.070	0.058	−0.011	[-0.024, 0.002]	0.098	–	0.110
G (debate r2) vs C (gbm)	Mean $\lvert e\rvert$	40	0.199	0.176	−0.024	[-0.064, 0.015]	0.236	–	0.317
G (debate r2) vs C (gbm)	MdAPE	40	0.129	0.113	−0.016	[-0.087, 0.054]	0.591	–	–
G (debate r2) vs C (gbm)	P95 $\lvert e\rvert$	40	0.588	0.420	−0.169	[-0.268, 0.094]	0.089	–	–
G (debate r2) vs C (gbm)	$\lvert e\rvert > \log 1.25$	40	0.325	0.300	−0.025	[-0.129, 0.077]	0.705	–	–
G (debate r2) vs C (gbm)	Cov. 90%	40	–	0.825	–	–	–	–	–
G (debate r2) vs C (gbm)	Pinball (log)	40	–	0.058	–	–	–	–	–
F (independent) vs D (avm)	Mean $\lvert e\rvert$	40	0.185	0.179	−0.006	[-0.043, 0.029]	0.731	–	0.789
F (independent) vs D (avm)	MdAPE	40	0.147	0.118	−0.028	[-0.070, 0.052]	0.405	–	–
F (independent) vs D (avm)	P95 $\lvert e\rvert$	40	0.529	0.443	−0.086	[-0.218, 0.093]	0.266	–	–
F (independent) vs D (avm)	$\lvert e\rvert > \log 1.25$	40	0.275	0.300	0.025	[-0.051, 0.122]	0.716	–	–
F (independent) vs D (avm)	Cov. 90%	40	0.850	0.825	−0.025	[-0.154, 0.100]	0.793	–	–
F (independent) vs D (avm)	Pinball (log)	40	0.061	0.060	−0.001	[-0.012, 0.009]	0.772	–	0.800
F (independent) vs E (single llm)	Mean $\lvert e\rvert$	40	0.182	0.179	−0.002	[-0.017, 0.012]	0.730	–	0.810
F (independent) vs E (single llm)	MdAPE	40	0.118	0.118	0.000	[-0.061, 0.016]	0.987	–	–
F (independent) vs E (single llm)	P95 $\lvert e\rvert$	40	0.457	0.443	−0.014	[-0.109, 0.037]	0.726	–	–
F (independent) vs E (single llm)	$\lvert e\rvert > \log 1.25$	40	0.275	0.300	0.025	[0.000, 0.077]	0.581	–	–
F (independent) vs E (single llm)	Cov. 90%	40	0.800	0.825	0.025	[-0.051, 0.103]	0.701	–	–
F (independent) vs E (single llm)	Pinball (log)	40	0.062	0.060	−0.002	[-0.007, 0.003]	0.509	–	0.622
G (debate r2) vs G (debate r1)	Mean $\lvert e\rvert$	40	0.178	0.176	−0.002	[-0.008, 0.002]	0.414	–	0.395
G (debate r2) vs G (debate r1)	MdAPE	40	0.115	0.113	−0.002	[-0.010, 0.009]	0.677	–	–
G (debate r2) vs G (debate r1)	P95 $\lvert e\rvert$	40	0.420	0.420	0.000	[-0.045, 0.010]	0.953	–	–
G (debate r2) vs G (debate r1)	$\lvert e\rvert > \log 1.25$	40	0.300	0.300	0.000	[0.000, 0.000]	1.000	–	–
G (debate r2) vs G (debate r1)	Cov. 90%	40	0.825	0.825	0.000	[0.000, 0.000]	1.000	–	–
G (debate r2) vs G (debate r1)	Pinball (log)	40	0.060	0.058	−0.001	[-0.003, 0.000]	0.187	–	0.191
G (debate r4) vs G (debate r2)	Mean $\lvert e\rvert$	20	0.164	0.163	−0.001	[-0.005, 0.002]	0.390	–	0.371
G (debate r4) vs G (debate r2)	MdAPE	20	0.101	0.100	−0.002	[-0.012, 0.014]	0.580	–	–
G (debate r4) vs G (debate r2)	P95 $\lvert e\rvert$	20	0.440	0.422	−0.018	[-0.019, 0.008]	0.185	–	–
G (debate r4) vs G (debate r2)	$\lvert e\rvert > \log 1.25$	20	0.250	0.250	0.000	[0.000, 0.000]	1.000	–	–
G (debate r4) vs G (debate r2)	Cov. 90%	20	0.800	0.800	0.000	[0.000, 0.000]	1.000	–	–
G (debate r4) vs G (debate r2)	Pinball (log)	20	0.058	0.058	0.000	[-0.001, 0.001]	0.931	–	0.930

Table A5. Interval coverage and sharpness, validation partition. Coverage at 50 and 90 percent is read directly from the five quantiles; coverage at 80 and 95 percent is interpolated from a normal fit in log space; width is the mean log width of the 90 percent interval; pinball loss is averaged over the five quantiles; CRPS is the lower-biased score on the five-point grid. Source: `results/evaluation_validation/coverage.csv`.

Arm	Cov. 50%	Cov. 80%	Cov. 90%	Cov. 95%	Width 90% (log)	Pinball (log)	CRPS (log)	n_with_interval
A (hedonic)	0.500	0.855	0.900	0.930	1.001	0.067	0.137	72132
B (hierarchical)	0.500	0.857	0.900	0.930	0.998	0.065	0.134	72132
C (catboost)	–	–	–	–	–	–	–	0
C (gbm)	–	–	–	–	–	–	–	0
C (xgboost)	–	–	–	–	–	–	–	0
D (avm)	0.491	0.873	0.900	0.937	0.877	0.059	0.121	72132
E (single llm)	0.350	0.675	0.800	0.825	0.493	0.062	0.126	40
F (independent)	0.375	0.775	0.825	0.850	0.541	0.060	0.123	40
G (debate r1)	0.375	0.775	0.825	0.850	0.553	0.060	0.122	40
G (debate r2)	0.400	0.775	0.825	0.850	0.553	0.058	0.120	40
G (debate r4)	0.500	0.800	0.800	0.850	0.521	0.058	0.118	20
G (fullview)	0.450	0.750	0.850	0.850	0.560	0.059	0.120	40
H (debate skeptic)	0.425	0.775	0.825	0.850	0.557	0.058	0.119	40
I (placebo)	0.400	0.725	0.825	0.850	0.541	0.061	0.126	40
J (cost matched)	0.350	0.775	0.825	0.825	0.497	0.060	0.123	40

Table A6. Cost, latency and tokens per property by arm, validation partition. Costs are estimates from the ledger at list prices, not invoices. Source: `results/evaluation_validation/cost.csv`.

Arm	Cost (USD)	Latency (s)	Tokens	Total cost (USD)	Mean $\lvert e \rvert$	N
A (hedonic)	0.0000	0.0	0	0.00	0.197	72132
B (hierarchical)	0.0000	0.0	0	0.00	0.190	72132
C (catboost)	0.0000	0.0	0	0.00	0.173	72132
C (gbm)	0.0000	0.0	0	0.00	0.183	72132
C (xgboost)	0.0000	0.0	0	0.00	0.181	72132
D (avm)	0.0000	0.0	0	0.00	0.180	72132
E (single llm)	0.0206	19.6	6381	0.83	0.182	40
F (independent)	0.2269	84.7	71977	9.08	0.179	40
G (debate r1)	0.8773	252.2	294497	35.09	0.178	40
G (debate r2)	1.5012	402.3	514936	60.05	0.176	40
G (debate r4)	2.2712	575.3	776036	45.42	0.163	20
G (fullview)	1.5952	411.5	545943	63.81	0.175	40
H (debate skeptic)	1.3740	383.2	464506	54.96	0.175	40
I (placebo)	1.3519	317.8	476454	54.08	0.184	40
J (cost matched)	0.2523	107.9	89422	10.09	0.179	40

Table A7. Mean absolute log error by stratum and arm, validation partition. Cells with fewer than thirty properties are suppressed. Source: `results/evaluation_validation/strata.csv`.

Stratum	n	A (hedonic)	B (hierarchical)	C (catboost)	C (gbm)	C (xgboost)	D (avm)	E (single llm)	F (independent)	G (debate r1)	G (debate r2)	G (debate r4)	G (fullview)	H (debate skeptic)	I (placebo)	J (cost matched)
Region: Outaouais	4334	0.166	0.158	0.136	0.143	0.158	0.138	–	–	–	–	–	–	–	–	–
Region: other	29847	0.237	0.228	0.205	0.219	0.215	0.215	–	–	–	–	–	–	–	–	–
Region: Montréal CMA	32062	0.163	0.161	0.148	0.154	0.155	0.154	–	–	–	–	–	–	–	–	–
Region: Québec CMA	5889	0.196	0.183	0.171	0.185	0.169	0.174	–	–	–	–	–	–	–	–	–
Urban class: metro	30588	0.173	0.167	0.151	0.160	0.160	0.157	–	–	–	–	–	–	–	–	–
Urban class: suburban	25227	0.185	0.181	0.166	0.175	0.172	0.173	–	–	–	–	–	–	–	–	–
Urban class: rural	16317	0.257	0.248	0.224	0.236	0.235	0.234	–	–	–	–	–	–	–	–	–
Property class: house	47910	0.211	0.205	0.187	0.198	0.195	0.195	0.197	0.199	0.198	0.195	–	0.195	0.195	0.204	0.196
Property class: condo	14528	0.126	0.120	0.108	0.112	0.117	0.109	–	–	–	–	–	–	–	–	–
Property class: plex	9694	0.231	0.223	0.201	0.213	0.207	0.212	–	–	–	–	–	–	–	–	–
Price tercile: 1	19965	0.311	0.312	0.308	0.301	0.325	0.297	–	–	–	–	–	–	–	–	–
Price tercile: 2	24907	0.159	0.149	0.123	0.140	0.128	0.134	–	–	–	–	–	–	–	–	–
Price tercile: 3	27260	0.146	0.138	0.120	0.135	0.125	0.137	–	–	–	–	–	–	–	–	–
Construction decade: 1950s	5701	0.218	0.214	0.204	0.211	0.211	0.210	–	–	–	–	–	–	–	–	–
Construction decade: 1960s	7201	0.213	0.207	0.196	0.204	0.201	0.202	–	–	–	–	–	–	–	–	–
Construction decade: 1970s	9186	0.209	0.203	0.192	0.202	0.200	0.199	–	–	–	–	–	–	–	–	–
Construction decade: 1980s	8544	0.179	0.173	0.162	0.172	0.170	0.170	–	–	–	–	–	–	–	–	–
Construction decade: 1990s	7191	0.174	0.167	0.155	0.164	0.163	0.162	–	–	–	–	–	–	–	–	–
Construction decade: 2000s	9532	0.164	0.156	0.143	0.154	0.150	0.151	–	–	–	–	–	–	–	–	–
Construction decade: 2010+	15162	0.177	0.169	0.138	0.151	0.149	0.147	–	–	–	–	–	–	–	–	–
Construction decade: <1950	8912	0.250	0.246	0.223	0.229	0.232	0.229	–	–	–	–	–	–	–	–	–
Construction decade: unknown	703	0.318	0.311	0.266	0.277	0.284	0.278	–	–	–	–	–	–	–	–	–

D.2. Holdout partition (2026)

Table A8. Accuracy by arm, holdout partition (2026-01 to 2026-10), evaluated once after the test results were written. Columns as in Table A3. Source: `results/evaluation_holdout/summary.csv`.

Arm	Mean $\lvert e\rvert$	MAE (CAD)	MdAPE	RMSE (log)	P90 $\lvert e\rvert$	P95 $\lvert e\rvert$	Bias $\bar{e}$	$\lvert e\rvert > \log 1.25$	N
A (hedonic)	0.198	95868	0.132	0.291	0.436	0.649	−0.004	0.291	83489
B (hierarchical)	0.185	90072	0.117	0.284	0.415	0.655	0.016	0.252	83489
C (catboost)	0.191	94518	0.124	0.289	0.413	0.659	0.012	0.265	83489
C (gbm)	0.193	94553	0.127	0.288	0.418	0.651	0.008	0.274	83489
C (xgboost)	0.192	95645	0.121	0.295	0.425	0.689	0.046	0.261	83489
D (avm)	0.180	88512	0.112	0.278	0.399	0.643	0.002	0.239	83489
E (single llm)	0.195	96344	0.086	0.320	0.605	0.846	0.110	0.267	60
F (independent)	0.186	92553	0.090	0.310	0.584	0.805	0.107	0.233	60
G (debate r2)	0.189	93535	0.089	0.314	0.587	0.828	0.102	0.250	60

Table A9. Paired contrasts, holdout partition. Columns as in Table A4. Source: `results/evaluation_holdout/contrasts.csv`.

Contrast (B vs A)	Metric	N	A	B	Δ	95% CI	p (boot)	p (Holm)	p (DM)
G (debate r2) vs D (avm)	Mean $\|e\|$	60	0.189	0.189	−0.000	[-0.035, 0.037]	0.998	0.998	0.998
G (debate r2) vs F (independent)	Mean $\|e\|$	60	0.186	0.189	0.003	[-0.001, 0.007]	0.131	0.261	0.152
G (debate r2) vs E (single llm)	Mean $\|e\|$	60	0.195	0.189	−0.006	[-0.017, 0.005]	0.240	–	0.181
G (debate r2) vs E (single llm)	MdAPE	60	0.086	0.089	0.003	[-0.025, 0.029]	0.843	–	–
G (debate r2) vs E (single llm)	P95 $\|e\|$	60	0.846	0.828	−0.019	[-0.117, 0.036]	0.228	–	–
G (debate r2) vs E (single llm)	$\|e\| > \log 1.25$	60	0.267	0.250	−0.017	[-0.052, 0.000]	0.517	–	–
G (debate r2) vs E (single llm)	Cov. 90%	60	0.800	0.833	0.033	[-0.034, 0.098]	0.369	–	–
G (debate r2) vs E (single llm)	Pinball (log)	60	0.072	0.069	−0.002	[-0.006, 0.001]	0.229	–	0.256
G (debate r2) vs A (hedonic)	Mean $\|e\|$	60	0.212	0.189	−0.023	[-0.053, 0.007]	0.150	–	0.186
G (debate r2) vs A (hedonic)	MdAPE	60	0.132	0.089	−0.043	[-0.082, -0.003]	0.030	–	–
G (debate r2) vs A (hedonic)	P95 $\|e\|$	60	0.650	0.828	0.178	[-0.106, 0.249]	0.105	–	–
G (debate r2) vs A (hedonic)	$\|e\| > \log 1.25$	60	0.267	0.250	−0.017	[-0.098, 0.052]	0.759	–	–
G (debate r2) vs A (hedonic)	Cov. 90%	60	0.917	0.833	−0.083	[-0.179, 0.000]	0.105	–	–
G (debate r2) vs A (hedonic)	Pinball (log)	60	0.070	0.069	−0.000	[-0.011, 0.012]	0.954	–	0.961
G (debate r2) vs C (gbm)	Mean $\|e\|$	60	0.212	0.189	−0.023	[-0.064, 0.017]	0.277	–	0.358
G (debate r2) vs C (gbm)	MdAPE	60	0.144	0.089	−0.055	[-0.096, -0.020]	0.011	–	–
G (debate r2) vs C (gbm)	P95 $\|e\|$	60	0.695	0.828	0.132	[-0.125, 0.293]	0.290	–	–
G (debate r2) vs C (gbm)	$\|e\| > \log 1.25$	60	0.250	0.250	0.000	[-0.107, 0.094]	1.000	–	–
G (debate r2) vs C (gbm)	Cov. 90%	60	–	0.833	–	–	–	–	–
G (debate r2) vs C (gbm)	Pinball (log)	60	–	0.069	–	–	–	–	–
F (independent) vs D (avm)	Mean $\|e\|$	60	0.189	0.186	−0.003	[-0.036, 0.033]	0.869	–	0.886
F (independent) vs D (avm)	MdAPE	60	0.113	0.090	−0.023	[-0.065, 0.005]	0.215	–	–
F (independent) vs D (avm)	P95 $\|e\|$	60	0.689	0.805	0.116	[-0.059, 0.370]	0.236	–	–
F (independent) vs D (avm)	$\|e\| > \log 1.25$	60	0.250	0.233	−0.017	[-0.103, 0.068]	0.818	–	–
F (independent) vs D (avm)	Cov. 90%	60	0.900	0.833	−0.067	[-0.145, 0.016]	0.118	–	–
F (independent) vs D (avm)	Pinball (log)	60	0.065	0.069	0.003	[-0.008, 0.015]	0.582	–	0.650
F (independent) vs E (single llm)	Mean $\|e\|$	60	0.195	0.186	−0.009	[-0.021, 0.002]	0.102	–	0.070
F (independent) vs E (single llm)	MdAPE	60	0.086	0.090	0.004	[-0.032, 0.025]	0.789	–	–
F (independent) vs E (single llm)	P95 $\|e\|$	60	0.846	0.805	−0.041	[-0.108, 0.029]	0.179	–	–
F (independent) vs E (single llm)	$\|e\| > \log 1.25$	60	0.267	0.233	−0.033	[-0.085, 0.000]	0.227	–	–
F (independent) vs E (single llm)	Cov. 90%	60	0.800	0.833	0.033	[-0.034, 0.098]	0.372	–	–
F (independent) vs E (single llm)	Pinball (log)	60	0.072	0.069	−0.003	[-0.007, 0.001]	0.108	–	0.144

Table A10. Interval coverage and sharpness, holdout partition. Columns as in Table A5. Source: `results/evaluation_holdout/coverage.csv`.

Arm	Cov. 50%	Cov. 80%	Cov. 90%	Cov. 95%	Width 90% (log)	Pinball (log)	CRPS (log)	n_with_interval
A (hedonic)	0.480	0.853	0.898	0.935	1.001	0.066	0.137	83489
B (hierarchical)	0.495	0.865	0.907	0.936	0.998	0.064	0.131	83489
C (catboost)	–	–	–	–	–	–	–	0
C (gbm)	–	–	–	–	–	–	–	0
C (xgboost)	–	–	–	–	–	–	–	0
D (avm)	0.434	0.846	0.878	0.926	0.827	0.061	0.125	83489
E (single llm)	0.467	0.717	0.800	0.833	0.497	0.072	0.143	60
F (independent)	0.517	0.750	0.833	0.833	0.508	0.069	0.138	60
G (debate r2)	0.533	0.733	0.833	0.850	0.528	0.069	0.140	60

Table A11. Cost, latency and tokens per property by arm, holdout partition. Source: `results/evaluation_holdout/cost.csv`.

Arm	Cost (USD)	Latency (s)	Tokens	Total cost (USD)	Mean $\lvert e \rvert$	N
A (hedonic)	0.0000	0.0	0	0.00	0.198	83489
B (hierarchical)	0.0000	0.0	0	0.00	0.185	83489
C (catboost)	0.0000	0.0	0	0.00	0.191	83489
C (gbm)	0.0000	0.0	0	0.00	0.193	83489
C (xgboost)	0.0000	0.0	0	0.00	0.192	83489
D (avm)	0.0000	0.0	0	0.00	0.180	83489
E (single llm)	0.0228	17.6	7343	1.37	0.195	60
F (independent)	0.2419	81.2	77650	14.51	0.186	60
G (debate r2)	1.5783	341.6	537705	94.70	0.189	60

Table A12. Mean absolute log error by stratum and arm, holdout partition. Cells with fewer than thirty properties are suppressed. Source: `results/evaluation_holdout/strata.csv`.

Stratum	*n*	A (hedonic)	B (hierarchical)	C (catboost)	C (gbm)	C (xgboost)	D (avm)	E (single llm)	F (independent)	G (debate r2)
Region: Outaouais	4181	0.176	0.169	0.226	0.208	0.244	0.154	–	–	–
Region: other	35493	0.224	0.208	0.205	0.208	0.206	0.198	–	–	–
Region: Montréal CMA	36466	0.170	0.164	0.169	0.171	0.171	0.165	0.160	0.146	0.148
Region: Québec CMA	7349	0.225	0.194	0.211	0.214	0.195	0.183	–	–	–
Urban class: metro	34870	0.189	0.175	0.187	0.188	0.184	0.170	–	–	–
Urban class: suburban	29374	0.178	0.170	0.173	0.174	0.174	0.166	–	–	–
Urban class: rural	19245	0.244	0.229	0.226	0.230	0.231	0.219	–	–	–
Property class: house	56917	0.199	0.187	0.194	0.193	0.194	0.182	0.218	0.213	0.214
Property class: condo	15381	0.163	0.151	0.154	0.156	0.158	0.139	–	–	–
Property class: plex	11191	0.241	0.227	0.230	0.240	0.227	0.228	–	–	–
Price tercile: 1	16568	0.347	0.348	0.354	0.352	0.376	0.339	–	–	–
Price tercile: 2	27874	0.172	0.154	0.151	0.156	0.150	0.136	–	–	–
Price tercile: 3	39047	0.153	0.139	0.151	0.151	0.143	0.144	0.098	0.093	0.095
Construction decade: 1950s	6613	0.220	0.210	0.219	0.223	0.219	0.212	–	–	–
Construction decade: 1960s	8001	0.210	0.198	0.208	0.210	0.206	0.199	–	–	–
Construction decade: 1970s	11340	0.206	0.193	0.200	0.202	0.200	0.191	–	–	–
Construction decade: 1980s	10108	0.181	0.169	0.182	0.182	0.181	0.168	–	–	–
Construction decade: 1990s	8653	0.171	0.158	0.171	0.169	0.169	0.154	–	–	–
Construction decade: 2000s	11068	0.169	0.156	0.166	0.166	0.168	0.149	–	–	–
Construction decade: 2010+	16807	0.181	0.166	0.165	0.166	0.166	0.154	–	–	–
Construction decade: <1950	10088	0.255	0.246	0.241	0.245	0.243	0.234	–	–	–
Construction decade: unknown	811	0.325	0.318	0.281	0.281	0.297	0.280	–	–	–

D.3. Prospective set

Table A13. Accuracy by arm, prospective set: 1,182 sales registered between the snapshot of 2026-10-21 and the fetch of 2026-10-27, absent from every partition and valued under the frozen protocol; the agentic arms on a stratified draw of forty. Columns as in Table A3. Source: `results/evaluation_prospective/summary.csv`.

Arm	Mean $\lvert e\rvert$	MAE (CAD)	MdAPE	RMSE (log)	P90 $\lvert e\rvert$	P95 $\lvert e\rvert$	Bias $\bar{e}$	$\lvert e\rvert > \log 1.25$	N
A (hedonic)	0.219	101233	0.142	0.316	0.536	0.738	0.006	0.331	1182
B (hierarchical)	0.207	95377	0.123	0.312	0.522	0.757	0.035	0.291	1182
C (catboost)	0.215	103040	0.138	0.327	0.497	0.759	0.077	0.297	1182
C (gbm)	0.218	103314	0.142	0.321	0.479	0.763	0.050	0.324	1182
C (xgboost)	0.222	105344	0.144	0.333	0.504	0.817	0.095	0.305	1182
D (avm)	0.202	95300	0.129	0.309	0.460	0.740	0.039	0.281	1182
E (single llm)	0.187	83022	0.085	0.275	0.491	0.645	0.059	0.300	40
F (independent)	0.179	77780	0.088	0.268	0.431	0.660	0.061	0.275	40
G (debate r2)	0.187	81111	0.096	0.278	0.487	0.662	0.058	0.300	40

Table A14. Paired contrasts, prospective set. Columns as in Table A4. Source: `results/evaluation_prospective/contrasts.csv`.

Contrast (B vs A)	Metric	N	A	B	Δ	95% CI	p (boot)	p (Holm)	p (DM)
G (debate r2) vs D (avm)	Mean $\|e\|$	40	0.197	0.187	−0.010	[-0.057, 0.043]	0.700	0.700	0.601
G (debate r2) vs F (independent)	Mean $\|e\|$	40	0.179	0.187	0.009	[-0.000, 0.021]	0.097	0.193	0.094
G (debate r2) vs E (single llm)	Mean $\|e\|$	40	0.187	0.187	0.001	[-0.020, 0.020]	0.925	–	0.898
G (debate r2) vs E (single llm)	MdAPE	40	0.085	0.096	0.011	[-0.044, 0.030]	0.332	–	–
G (debate r2) vs E (single llm)	P95 $\|e\|$	40	0.645	0.662	0.017	[-0.058, 0.087]	0.631	–	–
G (debate r2) vs E (single llm)	$\|e\| > \log 1.25$	40	0.300	0.300	0.000	[-0.071, 0.071]	1.000	–	–
G (debate r2) vs E (single llm)	Cov. 90%	40	0.750	0.825	0.075	[0.000, 0.167]	0.087	–	–
G (debate r2) vs E (single llm)	Pinball (log)	40	0.065	0.064	−0.000	[-0.006, 0.005]	0.897	–	0.860
G (debate r2) vs A (hedonic)	Mean $\|e\|$	40	0.206	0.187	−0.018	[-0.054, 0.022]	0.347	–	0.425
G (debate r2) vs A (hedonic)	MdAPE	40	0.123	0.096	−0.027	[-0.073, 0.022]	0.183	–	–
G (debate r2) vs A (hedonic)	P95 $\|e\|$	40	0.643	0.662	0.019	[-0.293, 0.149]	0.829	–	–
G (debate r2) vs A (hedonic)	$\|e\| > \log 1.25$	40	0.325	0.300	−0.025	[-0.123, 0.086]	0.723	–	–
G (debate r2) vs A (hedonic)	Cov. 90%	40	0.925	0.825	−0.100	[-0.212, -0.023]	0.045	–	–
G (debate r2) vs A (hedonic)	Pinball (log)	40	0.068	0.064	−0.004	[-0.018, 0.012]	0.569	–	0.642
G (debate r2) vs C (gbm)	Mean $\|e\|$	40	0.231	0.187	−0.043	[-0.090, 0.010]	0.091	–	0.085
G (debate r2) vs C (gbm)	MdAPE	40	0.151	0.096	−0.055	[-0.120, 0.034]	0.086	–	–
G (debate r2) vs C (gbm)	P95 $\|e\|$	40	0.808	0.662	−0.146	[-0.350, 0.077]	0.126	–	–
G (debate r2) vs C (gbm)	$\|e\| > \log 1.25$	40	0.300	0.300	0.000	[-0.143, 0.167]	1.000	–	–
G (debate r2) vs C (gbm)	Cov. 90%	40	–	0.825	–	–	–	–	–
G (debate r2) vs C (gbm)	Pinball (log)	40	–	0.064	–	–	–	–	–
F (independent) vs D (avm)	Mean $\|e\|$	40	0.197	0.179	−0.019	[-0.062, 0.029]	0.446	–	0.319
F (independent) vs D (avm)	MdAPE	40	0.096	0.088	−0.008	[-0.096, 0.045]	0.619	–	–
F (independent) vs D (avm)	P95 $\|e\|$	40	0.698	0.660	−0.038	[-0.322, 0.084]	0.260	–	–
F (independent) vs D (avm)	$\|e\| > \log 1.25$	40	0.275	0.275	0.000	[-0.128, 0.147]	1.000	–	–
F (independent) vs D (avm)	Cov. 90%	40	0.900	0.800	−0.100	[-0.265, 0.024]	0.231	–	–
F (independent) vs D (avm)	Pinball (log)	40	0.072	0.062	−0.011	[-0.024, 0.004]	0.159	–	0.137
F (independent) vs E (single llm)	Mean $\|e\|$	40	0.187	0.179	−0.008	[-0.025, 0.005]	0.314	–	0.212
F (independent) vs E (single llm)	MdAPE	40	0.085	0.088	0.003	[-0.053, 0.022]	0.751	–	–
F (independent) vs E (single llm)	P95 $\|e\|$	40	0.645	0.660	0.015	[-0.069, 0.037]	0.533	–	–
F (independent) vs E (single llm)	$\|e\| > \log 1.25$	40	0.300	0.275	−0.025	[-0.086, 0.000]	0.551	–	–
F (independent) vs E (single llm)	Cov. 90%	40	0.750	0.800	0.050	[0.000, 0.128]	0.219	–	–
F (independent) vs E (single llm)	Pinball (log)	40	0.065	0.062	−0.003	[-0.008, 0.001]	0.116	–	0.040

Table A15. Interval coverage and sharpness, prospective set. Columns as in Table A5. Source: `results/evaluation_prospective/coverage.csv`.

Arm	Cov. 50%	Cov. 80%	Cov. 90%	Cov. 95%	Width 90% (log)	Pinball (log)	CRPS (log)	n_with_interval
A (hedonic)	0.457	0.812	0.873	0.910	1.001	0.073	0.151	1182
B (hierarchical)	0.478	0.837	0.884	0.909	0.998	0.071	0.146	1182
C (catboost)	–	–	–	–	–	–	–	0
C (gbm)	–	–	–	–	–	–	–	0
C (xgboost)	–	–	–	–	–	–	–	0
D (avm)	0.410	0.827	0.866	0.899	0.889	0.071	0.148	1182
E (single llm)	0.425	0.725	0.750	0.850	0.558	0.065	0.131	40
F (independent)	0.500	0.775	0.800	0.875	0.574	0.062	0.126	40
G (debate r2)	0.450	0.725	0.825	0.900	0.571	0.064	0.132	40

Table A16. Cost, latency and tokens per property by arm, prospective set. Source: `results/evaluation_prospective/cost.csv`.

Arm	Cost (USD)	Latency (s)	Tokens	Total cost (USD)	Mean $\lvert e \rvert$	N
A (hedonic)	0.0000	0.0	0	0.00	0.219	1182
B (hierarchical)	0.0000	0.0	0	0.00	0.207	1182
C (catboost)	0.0000	0.0	0	0.00	0.215	1182
C (gbm)	0.0000	0.0	0	0.00	0.218	1182
C (xgboost)	0.0000	0.0	0	0.00	0.222	1182
D (avm)	0.0000	0.0	0	0.00	0.202	1182
E (single llm)	0.0205	16.1	6944	0.82	0.187	40
F (independent)	0.2341	54.7	75320	9.36	0.179	40
G (debate r2)	1.4996	244.9	514032	59.98	0.187	40

Table A17. Mean absolute log error by stratum and arm, prospective set. Cells with fewer than thirty properties are suppressed. Source: `results/evaluation_prospective/strata.csv`.

Stratum	*n*	A (hedonic)	B (hierarchical)	C (catboost)	C (gbm)	C (xgboost)	D (avm)	E (single llm)	F (independent)	G (debate r2)
Region: Outaouais	57	0.189	0.192	0.279	0.258	0.304	0.191	–	–	–
Region: other	567	0.236	0.221	0.230	0.227	0.229	0.213	–	–	–
Region: Montréal CMA	459	0.199	0.194	0.191	0.203	0.208	0.195	–	–	–
Region: Québec CMA	99	0.228	0.193	0.202	0.212	0.193	0.180	–	–	–
Urban class: metro	427	0.217	0.204	0.214	0.222	0.224	0.200	–	–	–
Urban class: suburban	430	0.193	0.184	0.193	0.195	0.206	0.183	–	–	–
Urban class: rural	325	0.254	0.241	0.244	0.243	0.240	0.231	–	–	–
Property class: house	856	0.217	0.205	0.214	0.214	0.221	0.199	0.201	0.189	0.195
Property class: condo	161	0.232	0.221	0.231	0.239	0.246	0.218	–	–	–
Property class: plex	165	0.215	0.204	0.202	0.217	0.203	0.206	–	–	–
Price tercile: 1	270	0.356	0.365	0.388	0.379	0.409	0.365	–	–	–
Price tercile: 2	411	0.186	0.167	0.170	0.177	0.178	0.154	–	–	–
Price tercile: 3	501	0.171	0.155	0.158	0.165	0.157	0.155	–	–	–
Construction decade: 1950s	91	0.197	0.184	0.206	0.211	0.219	0.204	–	–	–
Construction decade: 1960s	130	0.232	0.221	0.223	0.229	0.239	0.217	–	–	–
Construction decade: 1970s	184	0.247	0.234	0.244	0.256	0.254	0.231	–	–	–
Construction decade: 1980s	144	0.211	0.202	0.214	0.216	0.216	0.208	–	–	–
Construction decade: 1990s	112	0.190	0.173	0.182	0.191	0.197	0.164	–	–	–
Construction decade: 2000s	144	0.209	0.193	0.211	0.209	0.222	0.192	–	–	–
Construction decade: 2010+	217	0.188	0.174	0.182	0.179	0.178	0.161	–	–	–
Construction decade: <1950	147	0.253	0.254	0.238	0.239	0.245	0.234	–	–	–

E. Proofs

Throughout, v is treated as fixed, so that the error of a pool $w's$ with $w'\mathbf{1} = 1$ is $w'b + w'\eta$ and (4) holds.

Proof of Proposition 1. Write $\Sigma = \sigma^2[(1-\rho)I + \rho\mathbf{1}\mathbf{1}']$. Its eigenvalues are $\sigma^2(1-\rho)$, with multiplicity $K-1$ on the subspace orthogonal to $\mathbf{1}$, and $\sigma^2(1-\rho+\rho K)$ on $\mathbf{1}$; both are positive exactly when $-1/(K-1) < \rho < 1$, and the lower endpoint gives a singular but positive semi-definite matrix. Since $\Sigma\mathbf{1} = \sigma^2(1-\rho+\rho K)\mathbf{1}$,

$$\mathrm{MSE}_{\mathrm{eq}} = \frac{\mathbf{1}'\Sigma\mathbf{1}}{K^2} = \frac{\sigma^2(1-\rho+\rho K)}{K} = \sigma^2\Big[\frac{1-\rho}{K} + \rho\Big],$$

which is (5). For $\rho > -1/(K-1)$, $\Sigma^{-1}\mathbf{1} = \mathbf{1}/[\sigma^2(1-\rho+\rho K)]$, so the Bates–Granger weights $\Sigma^{-1}\mathbf{1}/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$ equal $\mathbf{1}/K$ and the equal-weight pool is the minimum-variance pool (Proposition 2 shows that this pool is optimal among all linear pools). Finally $(1-\rho)/K \geq 0$ gives $\mathrm{MSE}_{\mathrm{eq}} \geq \rho\sigma^2$ for every K, and $(1-\rho)\sigma^2/K$ is the only term that depends on K. □

Proof of Proposition 2. Let $b = 0$. Minimising $w'\Sigma w$ subject to $w'\mathbf{1} = 1$ with multiplier μ gives $2\Sigma w = \mu\mathbf{1}$, hence $w^* = \Sigma^{-1}\mathbf{1}/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$ and $\mathrm{MSE}^* = w^{*\prime}\Sigma w^* = 1/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$; the minimiser is unique because Σ is positive definite. For the inequality, apply the Cauchy–Schwarz inequality to $\Sigma^{1/2}\mathbf{1}$ and $\Sigma^{-1/2}\mathbf{1}$:

$$K^2 = (\mathbf{1}'\mathbf{1})^2 = \big((\Sigma^{1/2}\mathbf{1})'(\Sigma^{-1/2}\mathbf{1})\big)^2 \leq (\mathbf{1}'\Sigma\mathbf{1})(\mathbf{1}'\Sigma^{-1}\mathbf{1}),$$

which rearranges to $1/(\mathbf{1}'\Sigma^{-1}\mathbf{1}) \leq \mathbf{1}'\Sigma\mathbf{1}/K^2$. Equality holds if and only if $\Sigma^{1/2}\mathbf{1}$ and $\Sigma^{-1/2}\mathbf{1}$ are proportional, that is $\Sigma\mathbf{1} \propto \mathbf{1}$. If $b \neq 0$, $\mathrm{MSE}(w) = w'(\Sigma + bb')w$ and the same Lagrangian argument gives $w \propto (\Sigma + bb')^{-1}\mathbf{1}$. If $b = \beta\mathbf{1}$ then $w'b = \beta$ for every w with $w'\mathbf{1} = 1$, so the bias contributes β^2 to every pool. □

Proof of Proposition 3. The error of the equal-weight pool is $K^{-1}\sum_k \eta_k = c + \sum_m (n_m/K)g_m + K^{-1}\sum_k e_k$. The three groups of terms are independent, so the variance is $\sigma_c^2 + \sigma_g^2\sum_m (n_m/K)^2 + \sigma_e^2/K$, which is (7). With $\sum_m n_m = K$, the Cauchy–Schwarz inequality $(\sum_m n_m)^2 \leq M\sum_m n_m^2$ gives $\sum_m (n_m/K)^2 \geq 1/M$ with equality if and only if all n_m are equal. Neither σ_c^2 nor the minimised provider term σ_g^2/M depends on K, and σ_c^2 does not depend on M. □

Proof of Proposition 4. (i) Suppose $T(x^{(r)}) = T_0 := T(x^{(0)})$. Using translation and scale equivariance with $a = 1-\lambda > 0$ when $\lambda < 1$,

$$T(x^{(r+1)}) = T\big((1-\lambda)x^{(r)} + \lambda T_0\mathbf{1}\big) = (1-\lambda)T(x^{(r)}) + \lambda T_0 = T_0;$$

when $\lambda = 1$, $x^{(r+1)} = T_0\mathbf{1}$ and $T(T_0\mathbf{1}) = T_0$ by translation equivariance applied to $x = 0$ together with $T(0) = 0$, which follows from scale equivariance. By induction $T(x^{(r)}) = T_0$ for all r. Substituting into (8), $x^{(r+1)} - T_0\mathbf{1} = (1-\lambda)(x^{(r)} - T_0\mathbf{1})$, hence $x^{(r)} - T_0\mathbf{1} = (1-\lambda)^r(x^{(0)} - T_0\mathbf{1})$, which is (9). Differences $x_k^{(r)} - x_l^{(r)}$ follow by subtraction. For a dispersion statistic D with $D(x+b\mathbf{1}) = D(x)$ and $D(ax) = aD(x)$ for $a > 0$, (9) gives $D(x^{(r)}) = D((1-\lambda)^r x^{(0)}) = (1-\lambda)^r D(x^{(0)})$ when $\lambda < 1$, and $D(x^{(r)}) = D(T_0\mathbf{1}) = D(0) = 0$ when $\lambda = 1$, consistent with the formula. Nothing in the argument involves Σ.

(ii) Averaging (9) over k with $T_0 = \bar{x}^{(0)}$ gives $\bar{x}^{(r)} = (1-\lambda)^r\bar{x}^{(0)} + [1-(1-\lambda)^r]\bar{x}^{(0)} = \bar{x}^{(0)}$.

(iii) For $0 \le \rho < 1$ the exchangeable vector can be written $\eta = c\mathbf{1} + e$ with $c \sim N(0, \rho\sigma^2)$ independent of e, whose components are independent $N(0, (1-\rho)\sigma^2)$. Then $\bar{x}^{(0)} = v + c + \bar{e}$ and $\mathrm{med}(x^{(0)}) = v + c + \mathrm{med}(e)$. Averaging (9) with $T = \mathrm{med}$ and writing $\alpha = (1-\lambda)^r$,

$$\bar{x}^{(r)} = \alpha\bar{x}^{(0)} + (1-\alpha)\,\mathrm{med}(x^{(0)}) = v + c + \bar{e} + (1-\alpha)U, \qquad U := \mathrm{med}(e) - \bar{e}.$$

Since e is symmetric about zero, $\mathbb{E}[\mathrm{med}(e)] = 0$, so $\mathbb{E}[\bar{x}^{(r)}] = v$. The invariance $\mathrm{med}(x^{(r)}) = \mathrm{med}(x^{(0)})$ is part (i). Consider the Gaussian location model in which $e_k = \mu + \varepsilon_k$ with ε_k independent $N(0, (1-\rho)\sigma^2)$ and the variance known; $\bar{e}$ is a complete sufficient statistic for μ and is therefore the uniformly minimum-variance unbiased estimator by the Lehmann–Scheffé theorem, while $\mathrm{med}(e)$ is unbiased for μ by symmetry. Hence U is an unbiased estimator of zero with finite variance, and for every real a the estimator $\bar{e} + aU$ is unbiased, so that

$$\mathrm{Var}(\bar{e} + aU) = \mathrm{Var}(\bar{e}) + 2a\,\mathrm{Cov}(\bar{e}, U) + a^2\,\mathrm{Var}(U) \ge \mathrm{Var}(\bar{e}) \quad \text{for all } a,$$

which is possible only if $\mathrm{Cov}(\bar{e}, U) = 0$. The identity holds at every μ, in particular at $\mu = 0$. Therefore, using the independence of c and e,

$$\mathrm{MSE}(\bar{x}^{(r)}) = \rho\sigma^2 + \mathrm{Var}(\bar{e}) + (1-\alpha)^2\,\mathrm{Var}(U) = \mathrm{MSE}(\bar{x}^{(0)}) + (1-\alpha)^2\,\mathrm{Var}(U),$$

and, because $\mathrm{Cov}(\bar{e}, \mathrm{med}(e)) = \mathrm{Var}(\bar{e}) + \mathrm{Cov}(\bar{e}, U) = \mathrm{Var}(\bar{e})$,

$$\mathrm{Var}(U) = \mathrm{Var}(\mathrm{med}(e)) - 2\,\mathrm{Cov}(\bar{e}, \mathrm{med}(e)) + \mathrm{Var}(\bar{e}) = \mathrm{Var}(\mathrm{med}(e)) - \mathrm{Var}(\bar{e}).$$

This is (10) with $\Delta_K = \mathrm{Var}(U) \ge 0$. The factor $(1-\alpha)^2 = [1-(1-\lambda)^r]^2$ is non-decreasing in r and tends to one, so that $\mathrm{MSE}(\bar{x}^{(r)})$ converges to $\rho\sigma^2 + \mathrm{Var}(\mathrm{med}(e))$, the mean-squared error of $\mathrm{med}(x^{(0)})$. □

Proof of Proposition 5. (i) Let $u = (W^r)'w$. Then $u'x^{(0)} = w'W^r x^{(0)} = w'x^{(r)}$ and $u'\mathbf{1} = w'W^r\mathbf{1} = w'\mathbf{1} = 1$ because $W\mathbf{1} = \mathbf{1}$ implies $W^r\mathbf{1} = \mathbf{1}$. Hence every round-$r$ pool is a round-0 pool and its

mean-squared error is at least MSE^*. If W^r is non-singular, then $(W^r)^{-1}\mathbf{1} = \mathbf{1}$, and for any u with $u'\mathbf{1} = 1$ the vector $w = ((W^r)')^{-1}u$ satisfies $w'\mathbf{1} = u'(W^r)^{-1}\mathbf{1} = u'\mathbf{1} = 1$ and $(W^r)'w = u$; taking $u = w^*$ shows that the optimum is attained at round r.

(ii) A primitive row-stochastic matrix has the simple eigenvalue 1 with right eigenvector $\mathbf{1}$ and a strictly positive left eigenvector π normalised by $\pi'\mathbf{1} = 1$, all other eigenvalues having modulus strictly below one, so that $W^r \to \mathbf{1}\pi'$ (Perron–Frobenius). Hence $x^{(r)} \to \mathbf{1}\pi' x^{(0)}$ and, for any w with $w'\mathbf{1} = 1$, $w'x^{(r)} \to \pi' x^{(0)}$. Since $x^{(0)}$ is Gaussian and W^r converges, second moments converge and $\mathrm{MSE}(w'x^{(r)}) \to \pi'\Sigma\pi$. As $\pi'\mathbf{1} = 1$, π is a round-0 pool, so $\pi'\Sigma\pi \geq \mathrm{MSE}^*$ with equality if and only if $\pi = w^*$, the unique minimiser of Proposition 2.

(iii) $W = (1-\lambda)I + \lambda\mathbf{1}\mathbf{1}'/K$ has every entry strictly positive for $\lambda \in (0,1]$ and is therefore primitive; $W\mathbf{1} = \mathbf{1}$; and $(\mathbf{1}/K)'W = (1-\lambda)\mathbf{1}'/K + \lambda\mathbf{1}'/K = (\mathbf{1}/K)'$, so $\pi = \mathbf{1}/K$ and the consensus is $\bar{x}^{(0)}$. □

Proof of Lemma 2. The pooled error after the revision is $w'\eta + w'\zeta$ (biases are unchanged since $\mathbb{E}[\zeta] = 0$). Expanding the square and taking expectations, $\mathbb{E}[(w'\eta + w'\zeta)^2] = w'\Sigma w + w'\Gamma w + 2w'\mathbb{E}[\zeta\eta']w$, which is (11). If $\mathrm{Cov}(\zeta,\eta) = 0$ the change equals $w'\Gamma w \geq 0$, strictly positive for every $w \neq 0$ when Γ is positive definite. A negative change requires $2w'\,\mathrm{Cov}(\zeta,\eta)w < -w'\Gamma w < 0$. □

Proof of Proposition 6. Because $a_k'\mathbf{1} = 1$, the signal (12) equals $v + c + a_k'u + g_{m(k)} + e_k$, and the equal-weight pool has error $c + \omega'u + \sum_m (n_m/K)g_m + K^{-1}\sum_k e_k$ with $\omega = K^{-1}\sum_k a_k$. The four groups are independent and $\mathrm{Var}(\omega'u) = \sigma_u^2\|\omega\|^2$, which gives (13). An exchange changes only the a_k, hence only ω, so the change in the pooled error is $\sigma_u^2(\|\omega'\|^2 - \|\omega\|^2)$.

(i) An item in $\bigcap_k A_k$ already belongs to every view, so adding it changes no A_k and no a_k. Any exchange with $\omega' = \omega$ leaves the pooled error unchanged by the formula just derived, and σ_c^2, σ_g^2 and the τ_k^2 never enter the exchange term.

(ii) Let $U = \bigcup_k A_k$ and $J_U = |U|$. Each a_k is supported on U and sums to one, so ω is supported on U with $\sum_{j\in U}\omega_j = 1$. By the Cauchy–Schwarz inequality, $1 = (\sum_{j\in U}\omega_j)^2 \leq J_U\sum_{j\in U}\omega_j^2 = J_U\|\omega\|^2$, with equality if and only if $\omega_j = 1/J_U$ for all $j \in U$. The full exchange sets every a_k equal to the uniform vector on U, hence ω' is uniform on U and the bound is attained. A single agent shown U has evidence error $c + J_U^{-1}\sum_{j\in U}u_j$ with variance $\sigma_c^2 + \sigma_u^2/J_U$, the same quantity.

(iii) Before the exchange $a_1 = (1,0)$, $a_2 = a_3 = (0,1)$, so $\omega = (1/3, 2/3)$ and $\|\omega\|^2 = 1/9 + 4/9 = 5/9$. After giving item 2 to agent 1, $a_1 = (1/2, 1/2)$, $\omega = (1/6, 5/6)$ and $\|\omega\|^2 = 1/36 + 25/36 = 13/18 > 5/9$. □

Proof of Proposition 7. Let $S = \sum_j \sigma_j^2$ and $\alpha = (1-\lambda)^r \in (0,1)$ for $r \geq 1$. By Proposition 4(ii) with the mean as anchor, $x^{(r)} = \alpha x^{(0)} + (1-\alpha)\bar{x}^{(0)}\mathbf{1}$, so the error of agent k at round r is $\alpha\eta_k + (1-\alpha)\bar{\eta}$.

With independent round-0 errors, $\mathrm{Cov}(\eta_k, \bar{\eta}) = \sigma_k^2/K$ and $\mathrm{Var}(\bar{\eta}) = S/K^2$, hence

$$\mathrm{Var}(x_k^{(r)}) = \beta\sigma_k^2 + \gamma, \qquad \beta = \alpha^2 + \frac{2\alpha(1-\alpha)}{K} > 0, \qquad \gamma = \frac{(1-\alpha)^2 S}{K^2} > 0.$$

The inverse-variance weights are $w_k = (\beta\sigma_k^2 + \gamma)^{-1}/C$ with $C = \sum_j (\beta\sigma_j^2 + \gamma)^{-1}$, and the pooled estimate is

$$\sum_k w_k x_k^{(r)} = v + \alpha \sum_k w_k \eta_k + (1-\alpha)\bar{\eta} = v + \sum_k u_k \eta_k, \qquad u_k = \alpha w_k + \frac{1-\alpha}{K},$$

with $\sum_k u_k = 1$. Its mean-squared error is $\sum_k u_k^2 \sigma_k^2$, which by Proposition 2 is at least $\mathrm{MSE}^* = (\sum_k \sigma_k^{-2})^{-1}$, with equality if and only if $u = w^*$, $w_k^* = \tau\sigma_k^{-2}$, $\tau = (\sum_j \sigma_j^{-2})^{-1}$. At round 0, $\alpha = 1$, $u = w$ and $w_k \propto \sigma_k^{-2}$, so the pool is optimal. Suppose that $u = w^*$ for some $r \geq 1$. Then for every k, $\alpha w_k = \tau\sigma_k^{-2} - (1-\alpha)/K$, and multiplying by $\beta\sigma_k^2 + \gamma$,

$$\frac{\alpha}{C} = \Big(\frac{\tau}{\sigma_k^2} - \frac{1-\alpha}{K}\Big)(\beta\sigma_k^2 + \gamma) = \tau\beta - \frac{(1-\alpha)\gamma}{K} + \frac{\tau\gamma}{\sigma_k^2} - \frac{(1-\alpha)\beta}{K}\,\sigma_k^2.$$

The right-hand side is $\phi(\sigma_k^2)$ for the function $\phi(y) = \mathrm{const} + \tau\gamma/y - (1-\alpha)\beta y/K$, which is strictly decreasing on $(0,\infty)$ because $\tau\gamma > 0$ and $(1-\alpha)\beta/K > 0$. A strictly monotone function takes the same value at all σ_k^2 only if they are all equal, which contradicts the hypothesis. Hence the inverse-variance pool at any round $r \geq 1$ has mean-squared error strictly above MSE^*, the error of the round-0 inverse-variance pool. As $r \to \infty$, $\alpha \to 0$, so $\beta \to 0$, $\gamma \to S/K^2$, $w_k \to 1/K$ and $u_k \to 1/K$; the error converges to $S/K^2 = \mathrm{MSE}_{\mathrm{eq}}$, which exceeds MSE^* by Proposition 2 because a diagonal Σ with unequal entries does not satisfy $\Sigma\mathbf{1} \propto \mathbf{1}$. □

Proof of Proposition 8. (i) Let P_σ be the permutation matrix with $(P_\sigma x)_k = x_{\sigma(k)}$. With provider assignment randomised independently of the slot, as in the design, the signal vector of arm I has the distribution of $P_\sigma x^G$, where x^G is the signal vector of arm G. An aggregator $\mathscr{A}$ that is invariant to permutations satisfies $\mathscr{A}(P_\sigma x) = \mathscr{A}(x)$ for every x, so $\mathscr{A}(x^I)$ and $\mathscr{A}(x^G)$ have the same distribution, and in particular the same error distribution.

(ii) With views fixed to the slots, ω and the provider term in (13) are unchanged, and the reasoning term is $K^{-2}\sum_k \tau^2(\pi_{\sigma(k)}) = K^{-2}\sum_k \tau^2(\pi_k)$ because a permutation does not change a sum. Hence $\mathrm{MSE}_{\mathrm{eq}}$ is the same in both arms.

(iii) Substituting $\tau_k^2 = \tau^2(\pi_{\sigma(k)}, A_k)$ and $\tau_k^2 = \tau^2(\pi_k, A_k)$ into (13) and subtracting gives the stated difference; the other three terms cancel because they do not involve the prompts. □

Proof of Proposition 9. Conditional on the seller's draw, the match fails for lack of financing when $W + L < S$, that is when $\ell(1-h)(1+b+sZ) < (1+u_S) - (1-\ell)\kappa$, or $Z < a(u_S)/(\ell(1-h)s)$, which

has probability $F = \Phi\big(a(u_S)/(\ell(1-h)s)\big)$ with $a(u_S) = (1+u_S) - (1-\ell)\kappa - \ell(1-h)(1+b)$. Write the argument as

$$q(b,h,s) = \frac{(1+u_S)-(1-\ell)\kappa}{\ell(1-h)s} - \frac{1+b}{s}.$$

Since Φ is strictly increasing, the monotonicity of F is that of q. (i) $\partial q/\partial b = -1/s < 0$. When financing is needed, $(1+u_S) > (1-\ell)\kappa$, the first term is positive and strictly increasing in h, so $\partial q/\partial h > 0$. (ii) $\partial q/\partial s = -a(u_S)/(\ell(1-h)s^2)$, which is positive when $a(u_S) < 0$ and negative when $a(u_S) > 0$. (iii) The failure rate among beneficial matches is $\mathbb{E}[F \mid B \geq S]$; if $a(u_S) < 0$ almost surely on $\{B \geq S\}$, each term is increasing in s by (ii) and so is the expectation. In general the expectation splits into the integral over $\{a < 0\}$, increasing in s, and the integral over $\{a > 0\}$, decreasing in s. (iv) Under $h = h_0 + h_1 s$ with $h_1 > 0$, $dq/ds = \partial q/\partial s + h_1\, \partial q/\partial h$, and the second term is positive by (i), so on $\{a < 0\}$ the derivative is larger than without the rule. □

Proof of Proposition 10. Let $Y = (1-c_\ell)V_1$, a Gaussian variable with positive variance, and $\varphi(L) = \mathbb{E}[\max(L-Y,0)]$. Then $\varphi'(L) = \Pr(Y < L) \in (0,1)$ and $\varphi''(L) = f_Y(L) > 0$ for every L, so φ is strictly increasing and strictly convex on $\mathbb{R}$. With $L = \ell(1-h)(1+b+sZ)$ and h fixed, $\mathrm{EL}(b,s) = p_d\mathbb{E}[\varphi(\ell(1-h)(1+b+sZ))]$. (i) L is strictly increasing in b for every Z, and so is $\varphi(L)$; hence EL is strictly increasing in b. (ii) Let $s' > s$ and $\delta = (s'^2 - s^2)^{1/2}$. If Z' is standard normal and independent of Z, then $sZ + \delta Z'$ has the distribution of $s'Z$. Writing $m = \ell(1-h)$,

$$\mathbb{E}\big[\varphi(m(1+b+s'Z))\big] = \mathbb{E}\Big[\mathbb{E}\big[\varphi(m(1+b+sZ)+m\delta Z') \mid Z\big]\Big] > \mathbb{E}\big[\varphi(m(1+b+sZ))\big],$$

where the inequality is Jensen's inequality applied conditionally on Z to the strictly convex φ and the non-degenerate variable $m\delta Z'$. (iii) Under $h = h_0 + h_1 s$ the loan is $L = \ell(1-h_0-h_1 s)(1+b+sZ)$, whose mean $\ell(1-h_0-h_1 s)(1+b)$ is strictly decreasing in s; the effect of s on EL combines the increasing spread effect of (ii) with a decreasing location effect, and its sign is not determined without further restrictions. For any s there is a haircut $h(s)$ that holds EL at its value at $s = 0$, because EL is continuous and strictly increasing in L and tends to zero as $h \to 1$; by Proposition 9(i) this haircut raises the failure probability of every financeable match. □

Proof of Proposition 11. Let $g(z) = p_d\,\varphi(\ell(1-h)(1+b+sz))$ with φ as above; g is continuous, non-decreasing and grows at most linearly, hence square-integrable against the standard normal density. (i) By linearity of expectation, $\mathbb{E}[\sum_n \ell_n] = N\mathbb{E}[g(Z_1)]$, and Z_1 is standard normal for every ρ. (ii) By exchangeability, $\mathrm{Var}(\sum_n \ell_n) = N\,\mathrm{Var}(g(Z_1)) + N(N-1)\,\mathrm{Cov}(g(Z_1), g(Z_2))$, and (Z_1, Z_2) is standard bivariate normal with correlation ρ. Let $\{h_n\}_{n\geq 0}$ be the orthonormal Hermite polynomials with respect to the standard normal density and $g = \sum_{n\geq 0} a_n h_n$ in L^2. Mehler's formula gives

$\mathbb{E}[h_m(Z_1)h_n(Z_2)] = \rho^n \mathbf{1}\{m = n\}$, hence

$$\operatorname{Cov}\big(g(Z_1), g(Z_2)\big) = \sum_{n \geq 1} a_n^2 \rho^n,$$

a power series with non-negative coefficients, which is non-decreasing in ρ on $[0,1]$ and strictly increasing because g is not almost surely constant, so that some a_n with $n \geq 1$ is non-zero. At $\rho = 0$ the covariance is zero and at $\rho = 1$ it equals $\sum_{n \geq 1} a_n^2 = \operatorname{Var}(g(Z_1))$, so the system variance rises from $N \operatorname{Var}(g)$ to $N^2 \operatorname{Var}(g)$. (iii) With N lenders using distinct models, $\operatorname{Var}(\sum_n \ell_n) = N v_0[1 + (N-1)\tilde{\rho}_0]$ where $v_0 = \operatorname{Var}(g_0(Z))$ and $\tilde{\rho}_0 = \operatorname{Cov}(g_0(Z_1), g_0(Z_2))/v_0 \in [0,1]$; with a single common model of loss variance v_1 it equals $N^2 v_1$. The latter exceeds the former if and only if $v_1/v_0 > [1 + (N-1)\tilde{\rho}_0]/N$. □