



UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS
DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 20

From Listing Images to Property Valuation: A Multimodal AI Framework for Automated Building Characterization and the Cost Approach

Evidence from Québec single-family listings matched to the assessment roll

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

The cost approach to value requires a specification of the improvements—materials, finish grade, condition, renovations—that assessment rolls do not record and that appraisers obtain by inspection. We ask whether that specification can be extracted from the photographs and text of a for-sale listing by general-purpose vision–language models, and what a transparent cost indication built on it is worth. We formalise an extraction-to-valuation framework with five evidence classes (administrative, visual, proxy, text-claimed, latent), deterministic cross-image aggregation, text–image–record reconciliation, a fully sourced cost schedule and uncertainty propagation, and we apply it to 1,200 Québec house listings matched within 15 m to their roll unit (14,245 photographs, three extractor tiers, 104 verified references). Visible exterior attributes are extracted with high agreement against broker fields (roof covering 95.8%, garage type 94.1%, cladding 84.5%) but storeys agree with the roll only loosely ($\kappa = 0.51$), and the most cost-relevant interior evidence is simply not shown: within the first ten photographs, 3.1% of listings show a basement and 19.6% a bathroom, a missing-not-at-random structure that no extractor can overcome. Relative to a roll-only baseline, image and text attributes raise the median cost indication and move its ratio to the asking price from 0.75 to 0.95 (median absolute deviation 30.1% to 27.9%), mainly through condition-adjusted depreciation; the gain is real but modest, and the level of the indication depends more on the published unit-cost anchor than on anything seen in the images. A calibration against assessor building values on five metropolitan areas transfers to a geographic hold-out and implies a hard cost near \$300 per square foot, consistent with 2026 Québec industry guides and well above a Montréal hard-cost band. We release the ontology, prompts, schedule and code, and we are explicit about what the data cannot establish: no expert labels exist, latent defects are unobservable, and the cost indication is a screening tool, not an appraisal.

Keywords: cost approach, automated valuation, vision–language models, listing photographs, depreciation, assessment roll, Québec

JEL Classification: R31, R32, C45, C55, D83

1. Introduction

Of the three approaches to value, the cost approach is the one that depends least on the market and most on the building. Its logic is simple: a buyer will not pay more for an existing house than it would cost to acquire a comparable lot and build an equivalent house, less an allowance for the age and condition of the existing one. Its practice is demanding: it requires a specification of the improvements—their size, type, materials, finish grade, condition and renovation history—that only an inspection provides. Mass-appraisal systems, including the Québec assessment roll studied here, carry the size, the type and the year built but not the materials, the grade or the condition; appraisers obtain those on site. The question this paper asks is whether the photographs and text that sellers themselves publish when they list a house can supply that specification, how well, and what a transparent cost indication built on it is worth.

The question has become tractable for two reasons. Vision–language models now answer structured questions about an image—what room is this, what is the counter made of, how worn is the exterior—at a cost of a fraction of a cent per photograph, with an explicit option to say that the feature is not visible. And data infrastructures such as Vrai Prix, the author’s research platform, link the millions of records of the provincial assessment roll to for-sale listings with their media, so that extracted attributes can be set against verified administrative fields, against the fields brokers enter, and against the values professional assessors have produced for the same buildings under a cost-based manual. The combination allows a study that is neither a demonstration on a few photographs nor a black-box price predictor, but an audit of an extraction-to-valuation chain at every link.

We make four contributions. First, we formalise the chain. Section 4 states the cost identity with explicit scope for each term, introduces an attribute ontology with five evidence classes—administrative, visual, proxy, text-claimed and latent—so that what is recorded, seen, judged, claimed and unknowable are never conflated, defines deterministic cross-image aggregation and text–image–record reconciliation rules, and maps the attributes into a cost schedule and a condition-adjusted age–life depreciation. Two propositions characterise the value of an image observation and the bias that misreported condition induces in depreciation. Second, we implement the chain on 1,200 Québec house listings drawn by stratified sampling from a frame of 7,745 listings matched within 15 m to their roll unit, with 14,245 photographs analysed by three extractor tiers, every parameter of the cost schedule sourced or declared, and no parameter fitted on prices. Third, we evaluate each link: agreement of visual attributes with broker fields and the roll, information content of the image and text channels, cross-extractor reliability of judgment attributes, confidence calibration, the missing-not-at-random structure of what sellers photograph, and the alignment of the cost indication with assessor values, asking prices and recorded sales, under a geographic

hold-out. Fourth, we are explicit about what cannot be established with these data—no expert labels, no view of hidden defects, no cross-city cost levels—and we release the ontology, prompts, schedule, code and a pre-registered plan for the expert annotation study that should follow.

Three findings organise the results. Visible exterior facts are extracted well: against broker fields, the mid-size extractor agrees on roof covering in 95.8% of informative cases, on garage type in 94.1% and on the primary cladding in 84.5%, with coverage above 85% for each; storeys, by contrast, agree with the roll only loosely ($\kappa = 0.51$) because roll and image count half-storeys and walk-outs differently. The binding constraint is not the extractor but the seller: within the first ten photographs, only 3.1% of listings show a basement and 19.6% a bathroom, while 67.8% show a kitchen and 95.0% an exterior; basement finish, one of the largest add-ons in a cost schedule, is therefore essentially a text-claimed attribute in this setting, and interior condition is judged from living rooms and kitchens. And the cost indication responds to what is extracted mostly through depreciation: moving from a roll-only baseline to the multimodal baseline lowers the median depreciation from 0.62 to 0.38, raises the median indication from \$407,545 to \$514,919, and moves its ratio to the asking price from 0.75 to 0.95 with a median absolute deviation falling from 30.1% to 27.9%. The improvement is real, replicated across extractor tiers and robust to the depreciation assumptions we vary, but it is modest next to the effect of the published unit-cost anchor: switching from a Québec 2026 industry band to a Montréal hard-cost band moves the median indication by -29.9%. A calibration of the schedule against assessor building values on five metropolitan areas transfers to the Québec City hold-out and, under the Montréal anchor, implies a standard hard cost of about \$306 per square foot—inside the Québec industry band and well above the Montréal band.

The paper is a cost-approach paper, not an image-enhanced hedonic paper, and the distinction matters for how the results should be read. A hedonic model would learn from prices how much a photographed granite counter is worth; it would be more accurate and less interpretable, and it would tell an appraiser nothing about replacement cost or depreciation. Here no price enters the construction of the indication; the photographs serve only to specify the building, and the gap between the indication and the price is a finding, interpretable in the light of the literature on prices and construction costs, not a residual to be minimised. Price prediction, cost estimation and appraisal judgment are three different things, and the paper keeps them apart.

The rest of the paper proceeds as follows. Section 2 describes the Québec assessment framework and the Vrai Prix infrastructure, and states precisely which parts of the chain exist in production and which are new. Section 3 reviews six literatures and locates the gap. Section 4 develops the framework. Section 5 describes the data, the ontology, the rights questions and the sample. Section 6 describes the implementation, the five baselines and the validation protocol. Sections 7 and 8 report the extraction and the valuation results; Section 9 the robustness, uncertainty, missingness and failure analyses; Section 10 the implications for practice; Section 11 the limitations, ethical questions and

the research agenda; Section 12 concludes.

2. Institutional setting

2.1. Residential assessment in Québec and the place of the cost approach

Municipal assessment in Québec is governed by the *Loi sur la fiscalité municipale* (RLRQ, c. F-2.1). Section 14 requires the assessor to prepare a roll every three years for three consecutive fiscal years; section 43 defines the value to be entered as the *valeur réelle*, « sa valeur d'échange sur un marché libre et ouvert à la concurrence, soit le prix le plus probable qui peut être payé lors d'une vente de gré à gré » between informed, unconstrained parties; and section 46 fixes the reference date for both the state of the unit and market conditions at 1 July of the second fiscal year preceding the first year of the roll. The roll therefore reports, for every unit, a total value and its land and building components as of a dated market-conditions reference; in our data this reference is 2022, 2023 or 2024 depending on the municipality's cycle.

The methods are prescribed by the *Manuel d'évaluation foncière du Québec* (MEFQ), to which the Act's regulations refer. Part 3 of the manual, « Processus d'évaluation », has separate chapters for the measurement of market variation (3A), roll equilibration (3B), the comparison method (3C), the income method (3D), the cost method (3E) and the reconciliation of value indications (3F). Chapter 3E provides unit-cost schedules from which a *coût neuf* is built from a base cost, together with annual *facteurs de rajustement* published by the ministry with a 1 July reference date. The full text of Chapter 3E was not available to us; the description above rests on the manual's published structure and on the ministry's public page on adjustment factors, and the paper makes no claim about the detailed content of the schedules or of the depreciation tables. What matters for the design is that the roll's building value is produced within a cost-based mass-appraisal framework and then reconciled with market evidence, so that it is a professionally produced, dated indication rather than either a construction cost or a transaction price.

Professional appraisal practice in Québec is regulated by the Ordre des évaluateurs agréés du Québec, whose standards of practice we could not consult in their current edition; the Canadian Uniform Standards of Professional Appraisal Practice (CUSPAP 2024) and the International Valuation Standards effective 31 January 2025 are the national and international references. All three recognise the cost approach as one of three approaches and require that its inputs—land value, cost new, depreciation—be supported and that the approach's applicability be assessed; none of them contemplates the use of listing photographs as an input, which is the gap this paper documents rather than fills.

2.2. The Vrai Prix data infrastructure

Vrai Prix is a research platform developed by the author that assembles three bodies of data. The first is the provincial assessment roll, published by the Ministère des Affaires municipales et de l'Habitation on Données Québec under a CC-BY 4.0 licence, with owner names, cadastre and mailing addresses redacted at source: 3.75 million units with address, coordinates, land-use code, year built, floor area, lot area, storeys, number of dwellings, and assessed land, building and total values with their market-conditions date. The second is the register of residential transfers since January 2021 (769 333 transactions in the copy used here), linked to the roll unit. The third is a daily feed of for-sale listings from 67 broker and platform connectors operated by a companion platform, each listing carrying its photographs, free-text description, feature list and broker-entered structured fields, and matched to a roll unit by geocoding with a recorded match distance.

The Vrai Prix valuation engine, whose source we inspected in a verbatim copy dated 17 October 2026, combines a gradient-boosted hedonic model trained on the register with an adjusted-comparables estimate and reconciles the two with fixed weights; it also stores, for every evaluated listing, a production estimate labelled as a cost estimate whose generating code was not available to us during this study. Vrai Prix therefore provides the *data substrate* of this paper—verified administrative fields, listing media and text, dated assessor values, and production estimates usable as external benchmarks—but **it does not contain the image-based cost approach studied here**. The extraction schema, the aggregation rules, the cost schedule, the depreciation model and the evaluation design were all developed for this paper and are released with it. Readers should not read the results as a validation of a deployed Vrai Prix component.

3. Related literature and research gap

The paper sits at the intersection of six literatures. We organise the review around the questions the framework must answer rather than around disciplines, and we close with the gap the study addresses. All references were verified against primary bibliographic records; the review characterises each work by its contribution and does not attribute to it figures we did not re-derive.

3.1. Cost approach, depreciation and the land–structure split

The cost approach rests on the principle of substitution and on the decomposition of a property into land and depreciated improvements. The economics of depreciation has been studied largely through prices rather than costs: [Hulten and Wykoff \(1981\)](#) estimate economic depreciation from vintage asset prices, and a sequence of housing studies estimates net depreciation rates of the order of one to two percent a year, with large variation across markets, tenures and maintenance

levels (Malpezzi et al., 1987; Shilling et al., 1991; Knight and Sirmans, 1996; Harding et al., 2007; Wilhelmsson, 2008). Two lessons from this work shape our depreciation model. First, chronological age is a poor proxy for the loss of value because it confounds deterioration, vintage and expected redevelopment (Clapp and Giaccotto, 1998; Coulson and McMillen, 2008; Chau et al., 2005); the age–life method must therefore use an *effective* age. Second, observed maintenance and condition offset age substantially (Knight and Sirmans, 1996; Harding et al., 2007; Wilhelmsson, 2008), which is exactly the information a photograph can carry and a roll cannot. The decomposition of value into land and structure has been pursued with replacement-cost residuals at the aggregate level (Davis and Heathcote, 2007; Davis and Palumbo, 2008) and with hierarchical repeat-sales models at the property level (Francke and van de Minne, 2017); the “builder’s model” of Diewert and Shimizu (2017) is the econometric mirror of the cost identity we implement. These studies also document that house prices can exceed replacement cost plus land for long periods where supply is constrained (Glaeser et al., 2005; Gyourko and Saiz, 2006), which is why we treat the ratio of a cost indication to a price as a market-alignment statistic, not an error. Published construction cost indices are known to be imperfect measures of cost (Somerville, 1999); we use the Statistics Canada index only for temporal updating and treat levels as uncertain. Depreciation has also been characterised for commercial and multifamily property (Bokhari and Geltner, 2018) and across regions (Yoshida, 2020); the regional heterogeneity it reveals motivates our metropolitan-area diagnostics.

3.2. Automated valuation, assessment and valuation standards

Automated valuation models have progressed from hedonic regressions (Rosen, 1974; Malpezzi, 2002; Hill, 2013) through spatial and machine-learning estimators (Case et al., 2004; Bourassa et al., 2010; Zurada et al., 2011; Mayer et al., 2019; Pace and Hayunga, 2020) to large-scale commercial systems whose accuracy depends above all on the breadth of the characteristics they observe (Kok et al., 2017). The information in listing text has been shown to add to structured fields (Nowak and Smith, 2017; Shen and Ross, 2021; Baur et al., 2023), as has condition information from inspection reports (Oust et al., 2025). Unobserved renovations bias both indices and valuations (Bogin and Doerner, 2019). On the appraisal side, valuation accuracy and variation have long been recognised as inherent (Crosby, 2000), and assessed values exhibit systematic error that ratio studies measure (PAGLIN and FOGARTY, 1972; McMillen and Singh, 2020). Interval reporting for AVMs is an active topic, with model-based intervals outperforming error-based ones (Krause et al., 2020). Québec-specific hedonic work documents the capitalisation of landscaping, schools, transit and location-specific attribute values (Rosiers et al., 2001; Des Rosiers et al., 2002; Thériault et al., 2003; Dubé et al., 2014), and the author’s companion working papers describe the Vrai Prix data and the empirical performance of comparable-based methods on them. None of this work implements a cost approach whose inputs are extracted from listing media.

3.3. Computer vision for real estate

A decade of work has used images to predict house prices or to measure what buyers see. Early contributions combined visual and textual features (H. Ahmed and Moustafa, 2016), graded the luxury of interiors (Poursaeed et al., 2018), and exploited satellite and street-view imagery (Bency et al., 2017; Law et al., 2019; You et al., 2017; Kang et al., 2021; Kucklick and Müller, 2023). Exterior appearance has been shown to carry price premia, with the important caveat that looks are partly strategic (Glaeser et al., 2018; Johnson et al., 2020; Lindenthal and Johnson, 2025; Nordstrom et al., 2024), and photographs in listings affect price and marketing time (Benefield et al., 2011; Seiler et al., 2012). Closer to our object, building condition has been estimated from façade patches (Koch et al., 2018), construction epoch and energy demand from photographs (Zeppelzauer et al., 2018; Despotovic et al., 2019), condition classes have been fed into pricing models (Koch et al., 2026), and, most recently, interior condition from listing images has been shown to improve automated valuation for Oslo condominiums (Almaas et al., 2026). Multi-view approaches combine interior, exterior and satellite images (Nouriani and Lemke, 2022). Across this literature the target is the *price*, the model is a supervised predictor trained on prices, and the visual signal is a learned embedding or a condition class whose link to any cost quantity is implicit. Perception-scaling methods such as Streetscore (Naik et al., 2014) show how human judgments can be scaled by vision models but also that the judgments scaled are perceptual, not technical.

3.4. Vision–language models, structured extraction and calibration

Contrastive pre-training (Radford et al., 2021), few-shot multimodal models (Alayrac et al., 2022) and visual instruction tuning (Liu et al., 2023) have produced general-purpose vision–language models that answer schema-constrained questions about an image, a setting that descends from visual question answering (Antol et al., 2015). Schema-constrained extraction with large language models has been validated in scientific text (Dagdelen et al., 2024) and applied to real-estate documents (Zhao et al., 2024; Shehbaz and Moreno-García, 2026), and language models have been tested as appraisers, where they prove sensitive to prompting and to anchors (Geerts et al., 2026). Two risks are documented. Models hallucinate—in language (Ji et al., 2023) and, for vision–language models, by reporting frequent objects that are not in the image (Li et al., 2023)—which is why an abstention value is a design requirement rather than an option. And stated confidence is only partially calibrated (Guo et al., 2017; Kadavath et al., 2022); ensembles and Bayesian approximations provide alternative uncertainty signals (Lakshminarayanan et al., 2016; Gal and Ghahramani, 2015), and conformal methods offer distribution-free intervals once a calibration set exists (Angelopoulos and Bates, 2023). Human-in-the-loop designs organise the interaction between such models and expert review (Mosqueira-Rey et al., 2023).

3.5. Measurement error, missingness, leakage, interpretability and fairness

Attribute extraction errors enter the cost model as measurement error in categorical and continuous inputs, with the attenuation and misclassification consequences surveyed by [Hausman \(2001\)](#), [Bound et al. \(2001\)](#) and [Schennach \(2016\)](#). Which rooms a seller photographs is a missing-data mechanism in the sense of [RUBIN \(1976\)](#), and there is every reason to expect it to be missing not at random; the selection problem is the one [Heckman \(1979\)](#) formalised, here at the level of images within a listing. Leakage—the model seeing the target through a side channel—is a documented cause of over-optimistic machine-learning results ([Kapoor and Narayanan, 2023](#)), and in listings the asking price is literally written in some descriptions, which motivates our masking. Because the cost approach is used for high-stakes decisions, interpretability is a requirement, in the sense of [Lipton \(2018\)](#) and the argument of [Rudin \(2019\)](#) for inherently interpretable models; and because valuation inputs can encode neighbourhood composition, fairness across housing segments must be examined ([Mehrabi et al., 2022](#); [Howell and Korver-Glenn, 2021](#); [Kermani and Wong, 2021](#)). Reliability of human coding, which our proposed expert protocol would require, has standard measures and conventions ([Cohen, 1960, 1968](#); [Landis and Koch, 1977](#); [Hayes and Krippendorff, 2007](#); [Artstein and Poesio, 2008](#); [Hallgren, 2012](#); [McHugh, 2012](#); [Koo and Li, 2016](#)), and non-expert annotation can approach expert quality when aggregated ([Snow et al., 2008](#)).

3.6. Construction cost estimation

Parametric and learning-based estimation of construction cost has a long history in construction management, with regression, neural-network and case-based models reaching single-digit percentage errors at the early design stage when the quantity and the grade are known ([Kim et al., 2004](#); [Murat Günaydin and Zeynep Doğan, 2004](#); [Elfaki et al., 2014](#)), and with dynamic models for the indices that update them ([Hwang, 2009](#)). This literature takes the building’s specification as given; the question of *how to obtain the specification of an existing house from the evidence a listing provides* is not asked.

3.7. The gap

Putting the strands together, the state of knowledge is this. Images carry price-relevant information, and condition extracted from images improves price prediction. Depreciation depends on condition and maintenance, which rolls do not record. Vision–language models can answer structured questions about images with known failure modes. Cost estimation is accurate once the specification is known. What is missing is a study that (i) extracts an *appraisal ontology*—materials, grades, condition, renovation evidence—rather than a price embedding; (ii) separates what is seen from what is claimed and from what is recorded; (iii) maps the extracted attributes into a transparent

cost identity with every parameter sourced; (iv) measures the missing-not-at-random structure of listing photographs that governs what can be extracted at all; and (v) evaluates the result against administrative records and against professionally produced assessor values rather than against prices alone. The study closest to ours, [Almaas et al. \(2026\)](#), does (part of) the condition extraction for condominiums and feeds it to a price model; ours is, to our knowledge, the first to work on the cost approach, on detached houses with a verified administrative link, and with an explicit evidence-class ontology. We state this as a positioning claim, not a priority claim: the field is moving quickly and we searched publisher databases, Crossref and preprint servers but cannot exclude concurrent work.

4. Cost-approach theory and a formal extraction-to-valuation framework

4.1. The cost identity and what each term may legitimately contain

The cost approach produces an *indication of value* for an improved property by adding the value of the land, assumed vacant and available for its highest and best use, to the cost of producing the improvements new, less the loss in value that the existing improvements have suffered relative to a new equivalent. For property i we write

$$\widehat{V}_i^{\text{cost}} = \widehat{L}_i + (\widehat{RCN}_i + \widehat{S}_i)(1 - \widehat{D}_i), \quad (1)$$

where $\widehat{L}_i$ is the land value, $\widehat{RCN}_i$ the replacement cost new of the principal improvement, $\widehat{S}_i$ the cost new of site improvements that are valued separately from the building (driveway, landscaping, pool), and $\widehat{D}_i \in [0, 1]$ the accrued depreciation rate applied to the improvements. Equation (1) differs from the familiar additive display $L + RCN - D + S$ of the author’s specification only in that depreciation is written as a rate applied to all improvements, which is how it is implemented and which makes the no-double-counting requirement explicit: physical deterioration, functional obsolescence and external obsolescence are combined multiplicatively, $1 - D = (1 - D^{\text{phys}})(1 - D^{\text{func}})(1 - D^{\text{ext}})$, so that no component is charged twice. Three scope statements govern the identity throughout the paper.

First, RCN is *replacement* cost, the cost of a building of equivalent utility built with current materials and methods, not *reproduction* cost, the cost of an exact replica. Replacement cost is the appropriate base for a mass-appraisal setting in which the subject’s original materials are unknown, and it is the concept that the published per-area schedules we use describe. The paper never attempts to reproduce historical construction details from photographs.

Second, market price is not construction cost. The identity is an indication of value under the economic principle of substitution; the relation between $\widehat{V}_i^{\text{cost}}$ and an observed asking or transaction price P_i is an empirical question that depends on supply conditions, land scarcity and the depreciation

actually perceived by buyers (Glaeser et al., 2005; Davis and Heathcote, 2007). We therefore call comparisons with prices *market alignment* and never accuracy.

Third, the assessed values of the Québec roll are themselves indications produced under a cost-based mass-appraisal manual (Section 2); the roll’s building value B_i^{roll} is used as an *assessor benchmark* after temporal indexing, and its land value L_i^{roll} as the independent evidence for $\hat{L}_i$. Both are judgments, with their own errors (McMillen and Singh, 2020), not ground truth.

4.2. Observations, evidence classes and the attribute vector

A listing i consists of a set of photographs $\{x_{ik}\}_{k=1}^{K_i}$ in the order chosen by the seller, a free-text description t_i , a set of broker-entered fields b_i , and, through a spatial match, the administrative record a_i of the assessment-roll unit. The appraisal-relevant characteristics of the property are collected in an attribute vector $\theta_i = (\theta_{i1}, \dots, \theta_{iJ})$ whose components belong to one of five *evidence classes* (Table 1):

- ADMIN: recorded by the assessor (year built, floor area, lot area, storeys, assessed land and building values);
- VISUAL: directly observable in a photograph (cladding material, roof covering, garage, kitchen counters, basement finish);
- PROXY: visual constructs that stand in for a latent cost driver (finish grade, apparent condition, renovation evidence);
- TEXT: claimed by the seller or broker (renovations and their years, condition statements, materials listed in structured fields);
- LATENT: not inferable from listing media (foundation integrity, hidden moisture, structural soundness, permits, tenure, exact area).

The pipeline never estimates a LATENT attribute; it treats it as unknown and, where it matters for value, as a scenario. The distinction between VISUAL and PROXY is the distinction between a fact that two careful observers would agree on from the same image and a graded judgment on which they could reasonably differ; the evaluation design (Section 6) tests the former against reference fields and the latter against cross-extractor agreement.

4.3. Per-image observation and cross-image aggregation

For each image x_{ik} the extractor m returns a scene label $s_{ik}^m \in \mathcal{S}$, a vector of attribute observations $\hat{\theta}_{ik}^m = (\hat{\theta}_{ikj}^m)_j$ whose components may take the value `not_visible` (the attribute is relevant to the scene but cannot be seen) or `not_applicable` (the attribute does not pertain to the scene), and a stated confidence $c_{ik}^m \in [0, 1]$. Let $\mathcal{K}_{ij}$ denote the set of images whose scene is relevant to attribute j (exterior scenes for cladding, kitchen scenes for counters, and so on) and $\mathcal{J}_{ij} \subseteq \mathcal{K}_{ij}$ the subset in

which the observation is informative. The property-level estimate is a deterministic function of the per-image observations: for a nominal attribute, the confidence-weighted plurality

$$w_{ij}^m(v) = \sum_{k \in \mathcal{I}_{ij}} \max(c_{ik}^m, \underline{c}) \mathbf{1}\{\hat{\theta}_{ikj}^m = v\}, \quad \hat{\theta}_{ij}^m = \arg \max_v w_{ij}^m(v), \quad (2)$$

$$\pi_{ij}^m = \frac{w_{ij}^m(\hat{\theta}_{ij}^m)}{\sum_v w_{ij}^m(v)}, \quad (3)$$

with $\underline{c} = 0.05$ a floor that keeps zero-confidence votes from vanishing and π_{ij}^m the vote share, which we use as an internal consistency measure; for an ordinal grade, the weighted median of the ordinal codes; for a condition score, the weighted mean over informative images. When $\mathcal{I}_{ij} = \emptyset$ the attribute is `not_visible` at the property level and the cost model falls back to a declared default. Truncating the image set to the first K photographs in seller order, $\mathcal{K}_{ij}(K)$, and recomputing (2) gives the image-count sensitivity of Section 7 without any additional model call.

The composite condition $\bar{c}_i^m$ is the mean of the exterior and interior condition scores when both exist, and the composite finish grade $g_i^m \in \{\text{basic}, \text{standard}, \text{upgraded}, \text{luxury}\}$ is the rounded weighted average of the kitchen grade (weight 0.5), the bathroom grade (0.3) and the interior finish-quality score rescaled to the grade range (0.2), each term entering only when observed. These weights are declared, not fitted; Section 9 varies them.

4.4. Text–image–record reconciliation

Let $\tilde{\theta}_{ij}$ be the text-claimed value of attribute j (from the broker fields and description, with prices and assessments masked) and a_{ij} the administrative value when the attribute is in the roll. The reconciled value and its status are

$$\theta_{ij}^* = \begin{cases} a_{ij} & j \in \text{ADMIN}, \\ \hat{\theta}_{ij}^m & \hat{\theta}_{ij}^m \text{ informative}, \\ \tilde{\theta}_{ij} & \text{otherwise, if informative}, \\ \text{default}_j & \text{otherwise,} \end{cases} \quad (4)$$

with a status $\sigma_{ij} \in \{\text{agree}, \text{contradict}, \text{visual only}, \text{text only}, \text{unknown}\}$ that compares the visual and text channels when both are informative. The ordering in (4) gives precedence to the record, then to what is seen, then to what is claimed; it is a design choice motivated by the incentive structure of listings (Levitt and Syverson, 2008; Nowak and Smith, 2017) and we report how often it is exercised. Contradictions are not resolved silently: they widen the uncertainty interval of Section 4.6.

4.5. Mapping attributes to the cost schedule

The replacement cost new is the product of a quantity, which is administrative, and a unit cost, which is a function of attributes:

$$\widehat{RCN}_i = A_i \kappa(g_i) \mu_{\text{st}}(n_i) \mu_{\text{att}}(\tau_i) \phi_{r(i)} \psi_t + \beta_{\text{bsmt}}(f_i) \frac{A_i}{n_i} + \gamma_{\text{gar}}(h_i), \quad (5)$$

where A_i is the roll floor area (sum of above-grade floor areas), $\kappa(g)$ the per-area hard construction cost for finish grade g taken from a published schedule, $\mu_{\text{st}}(n)$ a storey multiplier that spreads foundation and roof costs over the floor area, $\mu_{\text{att}}(\tau)$ a party-wall adjustment for attached houses, ϕ_r a regional level factor, ψ_t the temporal update of the schedule to the valuation quarter by the Statistics Canada building construction price index, $\beta_{\text{bsmt}}(f)$ the basement-finish add-on per unit of footprint and $\gamma_{\text{gar}}(h)$ the garage add-on. Indirect costs and entrepreneurial incentive are applied as proportional mark-ups, $RCN_i = \widehat{RCN}_i^{\text{hard}} (1+s)(1+e)$. Every parameter is listed with its source or its status as an assumption in Table 6; the two published anchors for $\kappa(\cdot)$ are compared in Section 8.

Physical depreciation follows the age–life method with a condition-adjusted effective age,

$$D_i^{\text{phys}} = \min \left\{ \frac{\max \{ \text{age}_i \lambda(\bar{c}_i) - \rho(\text{ren}_i), 1 \}}{\text{EL}}, \bar{D} \right\}, \quad (6)$$

where $\lambda(\cdot)$ is a decreasing function of the composite condition equal to one at the average score, $\rho(\cdot)$ removes a fixed number of years when renovation evidence is partial or extensive, EL is the economic life and $\bar{D}$ a cap. In the baseline the functional and external components are set to zero, because nothing in a listing supports their measurement; an external-obsolescence scenario is reported.

4.6. Uncertainty: what propagates and what does not

The cost indication inherits three kinds of uncertainty, and we treat them differently. *Attribute uncertainty* arises because $\hat{\theta}_{ij}^m$ may be wrong or missing; we represent it by a distribution over the finish grade and the condition whose spread increases with $1 - c_i^m$, with the absence of an informative observation, and with a recorded contradiction. *Parameter uncertainty* arises because unit costs, economic life, mark-ups, the land index and the regional factor are known only within bands; we represent it by uniform draws over the published or stated bands. *Latent uncertainty*—the possibility that a sound-looking house hides a defective foundation—cannot be propagated because it has no measurable distribution in our data; it is excluded from the intervals, which are therefore explicitly *conditional on the absence of hidden defects*, and it is discussed as the main reason why the cost approach from listing media is a screening tool rather than an appraisal.

Two elementary results frame the empirical analysis. Both follow directly from the identity and

from standard measurement-error arguments; proofs are in Appendix A.

Proposition 1 (Value of an image observation for the cost indication). *Let the finish grade g_i be a cost driver with unit-cost schedule $\kappa(g)$, and let the appraiser’s prior over g_i in the absence of images be the frame distribution p_g . Under a quadratic loss on the improvements component, the expected reduction in loss from observing a visual grade $\hat{g}_i$ with confusion matrix $Q = [q_{g\hat{g}}]$ equals $A_i^2 \mu_i^2 [\text{Var}_p(\kappa(g)) - \mathbb{E}_{\hat{g}} \text{Var}(\kappa(g) \mid \hat{g})]$, where μ_i collects the multiplicative terms of (5); it is zero when Q is uninformative (rows equal) and increases with the diagonal dominance of Q and with the spread of the schedule. In particular, an observation that is correct with probability π and otherwise uniform over the other grades reduces the loss by a factor that is increasing in π and in the between-grade dispersion of κ , and is scaled by the depreciation factor $(1 - D_i)^2$: the value of a visual grade observation is smaller for older houses.*

Proposition 2 (Misclassification of condition and bias of depreciation). *Suppose the true composite condition is c_i and the extractor reports $\hat{c}_i = c_i + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i \mid c_i] = \delta(c_i)$. If λ is affine on the relevant range, the age–life depreciation computed from $\hat{c}_i$ has bias $\text{age}_i \lambda' \delta(c_i) / \text{EL}$ before the cap; a uniform upward bias in reported condition (“everything looks well kept”) therefore understates depreciation by an amount that grows linearly with age and shrinks with the assumed economic life, and a condition-dependent bias ($\delta' < 0$, extremes pulled toward the middle) compresses the dispersion of depreciation across properties without changing its mean.*

Proposition 1 says what the experiments on grade agreement and on the spread of the cost schedule jointly measure; Proposition 2 says why the calibration of condition scores against an external reference, which this study could not perform, matters more for old houses than for new ones. Neither proposition asserts that images improve valuation; that is the empirical question.

5. Data, variable ontology, rights and sample construction

5.1. Sources and provenance

Three sources enter the analysis, all through read-only copies of the Vrai Prix databases taken on 17 October 2026 (Section 2.2). The assessment roll provides the ADMIN fields; its licence (CC-BY 4.0, Données Québec) permits reuse with attribution, owner identifiers are redacted at source and no attempt at re-identification is made. The listing database provides photographs (as URLs), descriptions, feature lists and broker-entered structured fields for houses offered for sale between 12 August and 17 October 2026, together with the Vrai Prix spatial match to a roll unit and its match distance. The transaction register is used only in aggregate, for the 39,278-listing frame’s units that sold after being listed. Two further public sources enter the cost model: Statistics

Canada’s building construction price indexes by census metropolitan area (table 18-10-0289-01, 2023=100, quarterly, retrieved through the Web Data Service) and published per-area construction cost figures for Québec and Montréal in 2026 (Section 5.5). Appendix E lists every institutional source with its access date and the exact wording relied upon.

5.2. Rights and ethics

Photographs and descriptions are copyrighted marketing material of brokers and owners. The study does not download, store or redistribute any photograph: images are passed to the model provider by URL for analysis and are not retained, no photograph is reproduced in this paper, and all reported quantities are aggregates. Hosts whose robots.txt disallows all automated retrieval were excluded from the analysis rather than circumvented; the provider’s fetcher independently refused those URLs. This rule removed 5,287 listings from the frame, essentially the listings of one large franchise network whose media host carries a blanket disallow directive. The exclusion is not neutral: the removed listings have a higher median asking price (\$589,900 against \$563,000) and a higher median roll land value (\$163,900 against \$132,600), so the retained frame under-represents the more urban, higher-land-value segment served by that network. Prices, taxes and municipal assessment amounts are masked in every text passed to a model, because the asking price is literally printed in some descriptions and would otherwise leak into any downstream comparison. The extraction schema contains no field about occupants or personal belongings. Terms of use of the individual broker sites were not reviewed one by one; this residual legal uncertainty is recorded in the project’s data-rights document and the analysis is a non-commercial academic study of aggregate extraction behaviour.

5.3. Variable ontology

Table 1 summarises the attribute ontology; the machine-readable version with allowed values, extraction method, expected effect in the cost model, validation source and known error modes is released with the code. Each attribute belongs to one of the five evidence classes of Section 4.2. The VISUAL attributes were chosen because they are (a) observable in a single photograph by a careful person, (b) cost-relevant in published schedules, and (c) at least partly verifiable against a broker field or a roll field in our data. Interior finish grades and condition scores are PROXY attributes and are anchored with written descriptors in the prompt; they are evaluated by cross-extractor agreement and by their effect on the cost indication, not by accuracy, because no external reference exists. The LATENT attributes are listed so that their exclusion is explicit.

Table 1. Attribute ontology by evidence class (summary)

Class	Attributes	Extraction	Validation in this study
ADMIN	year built, floor area, lot area, storeys, land value, building value, market-conditions date	roll lookup	internal (year built vs listing, 95% within ± 1 year)
VISUAL	scene, building type, storeys, cladding (up to 3), roof form and covering, garage, kitchen counters and cabinets, flooring, basement finish, visible defects, photo quality, staging	per-image VLM, strict JSON schema	broker fields (roofing, siding, basement, garage, type), roll storeys
PROXY	exterior condition, interior condition, interior finish quality, kitchen and bathroom grade, renovation evidence	per-image VLM, anchored scales	cross-extractor agreement; effect on indication
TEXT	building type, storeys, materials, garage, basement, grades, condition and renovation claims, year built, areas, heating, pool, disclosed defects	per-listing LLM on masked text	roll (year built, floor area), broker fields, images
LATENT	foundation, hidden moisture/mould, structure, permits/code, tenure, exact area	never inferred	scenario only

5.4. Sampling frame and sample

Table 2 logs the construction of the frame. Starting from 39,278 house listings evaluated by Vrai Prix, we keep listings whose matched roll unit is a single-family dwelling, whose geocode lies within 15 m of the unit, whose roll record is complete (year built, floor area, land and building values), whose listing year built, when present, is within five years of the roll’s, with at least four photographs, an asking price between \$100,000 and \$3,000,000, from sources that carry broker structured fields (two classified-advertisement platforms are dropped), whose image host permits automated retrieval, and one listing per roll unit when the same house is advertised by several brokers. The frame contains 7,745 listings in 221+ municipalities.

From the frame we draw a stratified sample of 1,200 listings (135 non-empty strata of metropolitan area $\times$ building-age band $\times$ asking-price quartile, proportional allocation with a floor of 20 per stratum, fixed seed) and keep for each the first twelve photographs in the seller’s order, 14,245 image URLs in all. The floor over-represents the smaller metropolitan areas by design, so that subgroup diagnostics have minimum support; design weights (frame stratum size over sample stratum size) are stored and used wherever a frame-level statement is made. Table 3 compares the composition of frame and sample and Table 4 gives descriptive statistics: the median listing asks \$547,850, has 121 m² of above-grade floor area in the roll and 36 photographs. Broker structured fields are

unevenly available: a roofing material is recorded for 38% of the sample, a basement description for 39%, a siding material for 16%, a garage field for 21%, and a municipal building assessment for 26%; these coverages bound the reference sets of the agreement analysis.

Table 2. Construction of the sampling frame

Rule	Description	Removed	Remaining
R0	House listings (property_type in house types) evaluated by Vrai Prix	0	39,278
R1	Vrai Prix classifies the matched roll unit as single-family (unifamilial)	6,041	33,237
R2	Listing geocode within 15 m of the roll unit	12,636	20,601
R3	Roll unit is a single dwelling	1,550	19,051
R4	Complete roll fields: year built, floor area, land and building values	404	18,647
R5	Listing year built (if any) within ± 5 years of the roll year built (match sanity)	331	18,316
R6	At least 4 image URLs	598	17,718
R7	Asking price within [100,000, 3,000,000]	209	17,509
R8	Sources with no structured building fields and few photos excluded: ['kijiji', 'fb_marketplace']	3,633	13,876
R10	Image host permits automated retrieval (robots.txt): hosts with 'Disallow: /' excluded	5,287	8,589
R9	One listing per roll unit (duplicates across sources: keep the listing with most images)	844	7,745

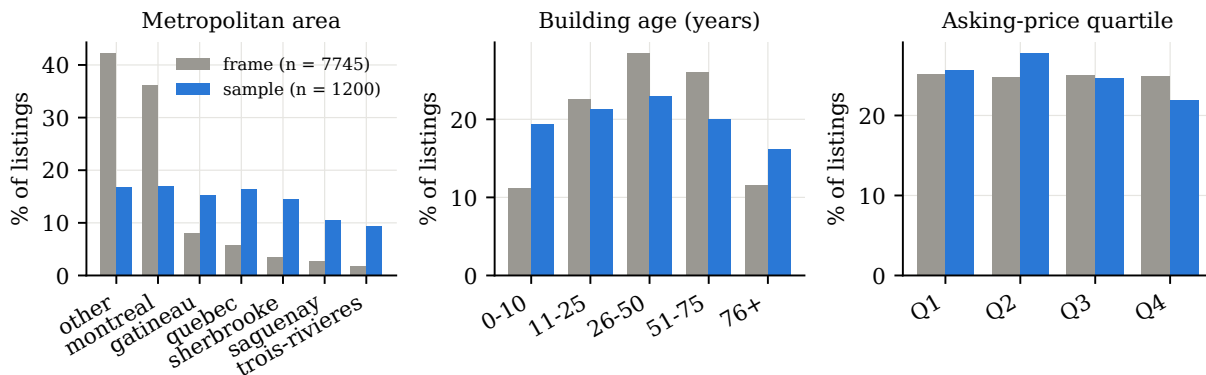
Table 3. Frame and sample composition by stratum

Stratum	Frame n	Frame %	Sample n	Sample %
Metropolitan area: other	3,267	42.2	201	16.8
Metropolitan area: montreal	2,800	36.2	203	16.9
Metropolitan area: gatineau	628	8.1	184	15.3
Metropolitan area: quebec	443	5.7	198	16.5
Metropolitan area: sherbrooke	268	3.5	175	14.6
Metropolitan area: saguenay	207	2.7	127	10.6
Metropolitan area: trois-rivieres	132	1.7	112	9.3
Building age (years): 0-10	872	11.3	233	19.4
Building age (years): 11-25	1,753	22.6	256	21.3
Building age (years): 26-50	2,202	28.4	276	23.0
Building age (years): 51-75	2,016	26.0	241	20.1
Building age (years): 76+	902	11.6	194	16.2
Asking-price quartile: Q1	1,953	25.2	308	25.7
Asking-price quartile: Q2	1,920	24.8	333	27.8
Asking-price quartile: Q3	1,938	25.0	296	24.7
Asking-price quartile: Q4	1,934	25.0	263	21.9

Table 4. Descriptive statistics of the sample

Variable	Sample P10	Sample median	Sample P90	Frame median
Asking price (\$)	339,810	547,850	974,820	563,000
Floor area, roll (m ²)	81	121	228	121
Lot area, roll (m ²)	256	627	3391	630
Age in 2026 (years)	5	37	85	40
Roll land value (\$)	44,460	118,700	267,420	132,600
Roll building value (\$)	156,970	275,600	530,130	291,100
Images available	18	36	58	35
Description length (chars)	317	808	2385	808

Note: floor area, lot area, age and values are those of the matched assessment-roll unit; the roll's market-conditions date is 1 July 2022, 2023 or 2024 depending on the municipality's cycle.

**Figure 1.** Composition of the frame and of the stratified sample. The per-stratum floor over-represents the smaller metropolitan areas; design weights restore frame proportions where needed.

A supplementary set of 146 frame listings, not in the main sample, whose roll unit recorded a sale after the listing first appeared, was processed with the two cheaper extractors for the market-alignment test against transaction prices (Section 8.5).

5.5. Cost-schedule data

Two published anchors describe the per-area cost of a new single-family house in 2026. Anchor A is the Montréal band of hard construction costs for a single-family house with an unfinished basement reported from Altus Group's *2026 Canadian Cost Guide* by a secondary source (\$155–\$230 per square foot after our grade placement; the guide itself is distributed on request and was not consulted directly). Anchor B is the range reported by Québec construction-industry guides updated in August and September 2026—a standard finish at \$280–325 per square foot for a two-storey house, a

cottage with unfinished basement at \$230–285, high-end finishes from \$350, and a bungalow premium of roughly \$100 per square foot—which exclude land, servicing, soil tests, professional fees, landscaping, basement finishing and garages. We place the four grades at \$257, \$302, \$340 and \$400 per square foot of a two-storey reference and derive a one-storey multiplier of 1.25 from the bungalow premium. Anchor B is Québec- and date-specific and is the headline; anchor A is reported throughout as a lower-bound alternative. Both are builder-facing guide figures rather than a cost manual, and the paper treats them as bands, not points (Table 5). All other parameters of the schedule—storey and party-wall multipliers, basement and garage add-ons, indirect-cost and entrepreneurial-incentive mark-ups, site-improvement share, economic life, condition factor, renovation allowance—are author assumptions with declared ranges (Table 6). The Statistics Canada index updates anchor A from early 2026 to the valuation quarter and indexes the assessor’s building value from the roll’s market-conditions date to the same quarter; it does not provide cross-city cost levels, which is why the regional level factor is a sensitivity parameter.

Table 5. Per-area hard construction cost by finish grade: the two published anchors

Anchor	Grade	\$/sq ft (point)	Band	Reference storeys
B	basic	257	230–285	2
B	standard	302	280–325	2
B	upgraded	340	325–375	2
B	luxury	400	350–450	2
A	basic	155	140–170	1
A	standard	175	160–195	1
A	upgraded	195	180–215	1
A	luxury	230	205–280	1

Note: CAD per square foot of above-grade floor area; the band is the sensitivity range used in Sections 9 and 9.3. Anchor A: Altus Group 2026 Canadian Cost Guide, Montréal, via NerdWallet Canada (accessed 2026-10-20). Anchor B: Soumission Rénovation (updated 2026-08-31, attributing its ranges to Société d’habitation du Québec 2026 data) and Dessins Drummond (updated 2026-09-24).

Table 6. Cost-schedule parameters, sources and status

Parameter	Value	Source / status
Storey multiplier, anchor B	1: 1.25, 1.5: 1.12, 2: 1.0, 2.5: 0.97, 3: 0.97	industry bungalow premium (B); assumption (A)
Storey multiplier, anchor A	1: 1.0, 1.5: 0.97, 2: 0.93, 2.5: 0.92, 3: 0.92	industry bungalow premium (B); assumption (A)
Attached-house multiplier	detached: 1.0, semi_detached: 0.97, row_or_townhouse: 0.94	assumption
Basement finish (\$/sq ft of footprint)	finished: 45, partially_finished: 22, unfinished: 0	assumption, $\pm 30\%$
Garage (\$)	attached: 35,000, detached: 45,000, carport: 8,000	assumption, $\pm 30\%$ (Dessins Drummond: single 25–50k)
Soft costs / entrepreneurial incentive	10% / 8%	assumption; ranges 5–15% / 0–15%
Site improvements	2% of RCN + pool	assumption
Economic life / max physical depreciation	60 y / 85%	assumption; 50–75 y
Condition factor on effective age	1: 1.4, 2: 1.2, 3: 1.0, 4: 0.75, 5: 0.55	assumption; strength 0.5–1.5
Renovation (years removed)	extensive 10, partial 4	assumption
Land	roll land value $\times$ market index (roll month $\rightarrow$ 2026-07)	MAMH roll; Vrai Prix index; share 0–1
Temporal / regional index	StatCan 18-10-0289 single-detached, by CMA	official statistics

6. Pipeline implementation, baselines and validation protocol

6.1. Pipeline

Figure 2 shows the pipeline as implemented. Listings are matched to roll units and filtered (Section 5.4); the first twelve photographs are sent one at a time, by URL, to a vision–language model with a frozen system prompt and a strict JSON schema of 22 fields (Appendix B); the masked text is sent once per listing to a language model with a second schema of 22 fields; the per-image outputs are aggregated by the deterministic rules of Section 4.3; visual, text and administrative attributes are reconciled by (4); the cost schedule and depreciation model of Section 4.5 produce the indication; and a Monte Carlo layer propagates attribute and parameter uncertainty. Every raw model response, prompt version, schema version and model identifier is persisted; no step after the model calls involves any fitted parameter.

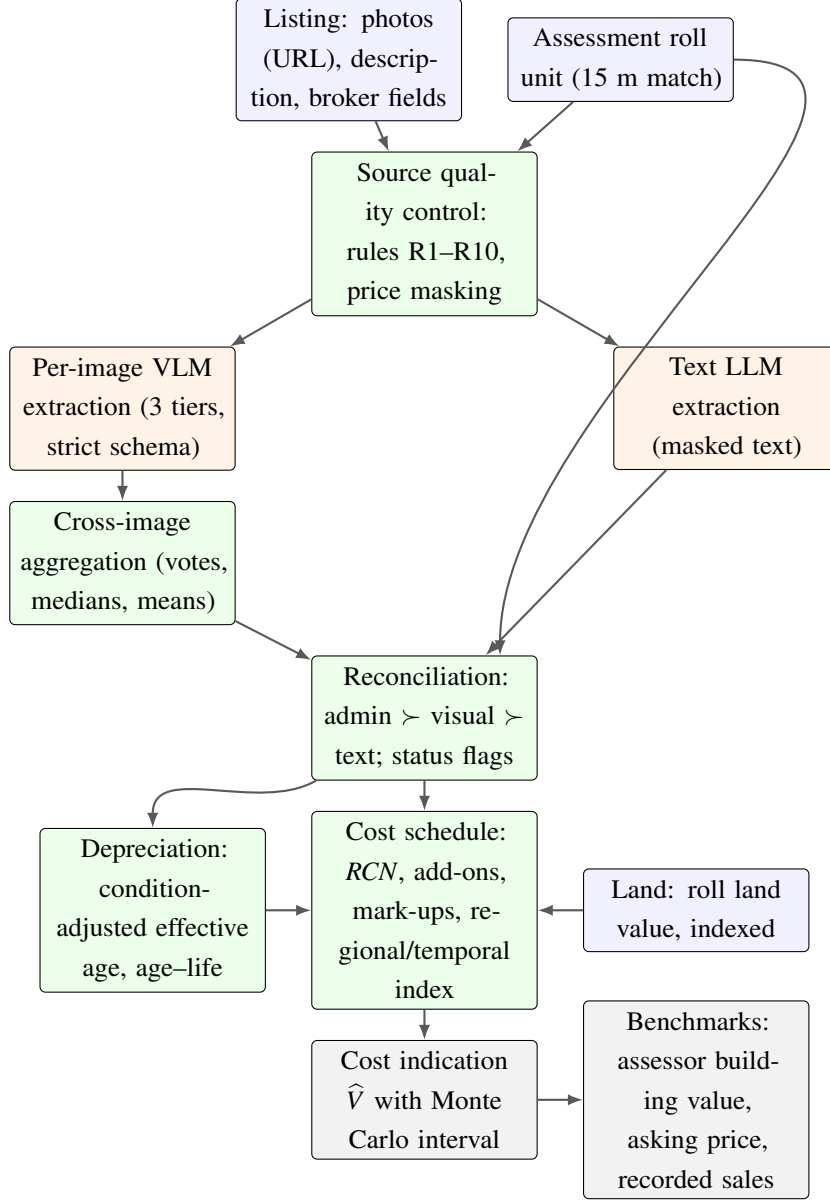


Figure 2. The extraction-to-valuation pipeline. Blue: data sources; orange: model calls; green: deterministic, parameter-free or sourced steps; grey: outputs and benchmarks.

6.2. Extractors

Three model tiers of one provider family were used so that extractor capability is itself a treatment: a small model (Claude Haiku 5.5) on all 1,200 listings and up to twelve photographs each (14,245 image calls), a mid-size model (Claude Sonnet 5.5) on all listings and the first ten photographs (11,916 calls), and the largest model (Claude Opus 5.5) on a nested random subsample of 200 listings and ten photographs. The largest tier’s text batch was returned and is used in the text-channel statistics; its image batch (1,994 requests) had not been returned by the provider when this version

was compiled, so the cross-extractor tables cover the two full-sample tiers and the image-based Opus results are deferred to the next version. All calls went through the provider’s asynchronous batch interface with the same prompt and schema; sampling parameters were left at their defaults and reasoning effort at its lowest setting, since the task is per-image classification. The schema forces every field to be filled, with explicit `not_visible` and `not_applicable` values, and asks for a confidence in $[0, 1]$ and a short statement of the visual cues used; the system prompt instructs the model never to infer hidden conditions, exact year or area. Because comparisons between tiers must hold the evidence fixed, all cross-tier statistics use the common first ten photographs; image-count sensitivity up to twelve uses the small model only. Image calls succeeded for 96.4% (mid-size) and 97.1% (small) of requests; failures are dead URLs and a few residual `robots.txt` refusals on secondary hosts, and are treated as missing images. Table 7 summarises the runs; total provider cost was about US\$99 at batch prices. Text calls used the same three tiers on the masked text.

Table 7. Extraction runs

Extractor	Listings	Image calls	Succeeded	URL failures	robots.txt refusals	Text calls	Mean tokens in / out	Cost (US\$, batch)
Claude Haiku 5.5	1,200	14,245	97.1%	174	33	1,200	5,176 / 323	6
Claude Sonnet 5.5	1,200	11,916	96.4%	185	29	1,200	5,175 / 335	94

Note: “URL failures” are images the provider could not download; “robots.txt refusals” are hosts whose directives disallow retrieval of the specific path. Cost is the provider’s batch price for the tokens actually billed.

6.3. Baselines

Five nested information sets define the baselines (Table 8). B0 uses only the roll and the typical assumptions an appraiser would make without an inspection: standard grade, average condition, unfinished basement, no garage, no pool, storeys and building type from the record. B1 adds the text channel—broker fields and the masked description, through the language model—for grade, condition claims, renovations, basement, garage, pool and, where stated, storeys and type. B2 adds the image channel instead of the text channel and is restricted to visually supported attributes; where no interior photograph exists it falls back to the B0 default and flags the listing. B3 reconciles images and text by (4). B4 adds the administrative override (storeys from the roll) and the uncertainty flags used by the Monte Carlo layer. No baseline sees prices or assessed building values; all share the roll floor area and the roll land value, so differences between them isolate the specification of the improvements. The comparison is fair in the sense that each baseline is given exactly the information its label says and nothing else; it is unfavourable to B2 to the extent that sellers do not photograph what matters.

Table 8. Information sets of the baselines

Level	Channels	Attributes beyond the roll
B0	roll	none: standard grade, condition 3/5, unfinished basement, no garage, no pool
B1	roll + text	grade from kitchen/bathroom claims, condition claim, renovation claim, basement, garage, pool, storeys/type if stated
B2	roll + images	composite grade and condition, renovation evidence, basement finish, garage, storeys and type as seen
B3	roll + images + text	B2 with text filling what images do not show; contradictions flagged
B4	B3 + administrative reconciliation	storeys from the roll; uncertainty flags for missing grade/condition

6.4. Validation protocol and metrics

The protocol was written before any result was examined and is released with the code. For VISUAL attributes with a reference field we report *coverage* (the share of listings for which the extractor gives an informative value when a reference exists), *agreement* among informative cases, Cohen’s κ and macro-F1, with confusion matrices for the key attributes; broker fields are claims, so we speak of agreement, not accuracy, except for the roll’s storeys and the listing’s property type. For PROXY attributes we report cross-extractor agreement (pairwise κ , Krippendorff’s nominal α , correlations of condition scores) and the effect of the attribute on the cost indication. Confidence calibration is assessed by binning per-image stated confidence against agreement with the broker field and by selective-prediction curves. The missing-not-at-random structure of photographs is measured directly: the scene of every photograph by position in seller order, the share of listings showing each room within the first K photographs, and the association of showing a room with price, age, condition and the broker’s own description of that room.

For the cost indication we report, for every baseline: the median indication and its decomposition; the ratio of depreciated improvements to the assessor’s building value indexed from the roll’s market-conditions date to the valuation quarter with the metropolitan-area construction price index (*assessor alignment*: median ratio, median absolute percentage deviation, coefficient of dispersion); and the ratio of the indication to the asking price (*market alignment*: median ratio, median absolute percentage deviation, share within 20%). Confidence intervals come from 499 cluster bootstrap replications by municipality, because listings in one municipality share an assessor and a market. A single depreciation or level parameter is calibrated against the assessor benchmark on five metropolitan areas and evaluated on the Québec City area as a geographic hold-out; prices are never used for calibration. Recorded sale prices after listing are used for a small market-alignment test,

labelled as such. Everything that is a scenario (renovation, grade shift, external obsolescence) is labelled a scenario. All experiments, including those that could not be run for lack of data (human expert agreement, human-in-the-loop gains), are listed in the pre-registered plan with their status.

7. Results I: what listing media reveal about the building

7.1. What sellers photograph

Before asking how well attributes are extracted, we ask what there is to extract. Figure 3 classifies every photograph of the sample by the scene the small extractor assigns to it and by its position in the seller’s order. The first photograph is an exterior front view in more than four cases out of five; living and dining rooms dominate positions three to ten; kitchens appear from position four onward; bedrooms, bathrooms and basements are pushed to the end of the sequence. Within the first ten photographs analysed by the mid-size extractor, 95.0% of listings show at least one exterior view and 67.8% a kitchen, but only 19.6% a bathroom and 3.1% a basement (Table 9); the median position of the first basement photograph, when there is one within twelve, is 9, against 1 for the front exterior and 7 for the kitchen. Figure 4 shows how the share of listings with each room rises with the number of photographs analysed: kitchens saturate quickly, bathrooms slowly, and basements hardly at all within the budget we could afford.

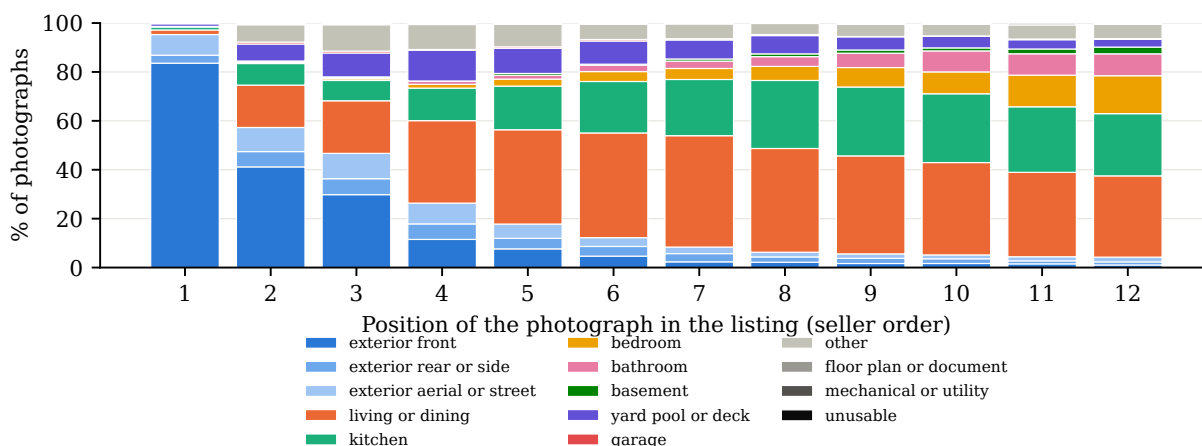


Figure 3. Scene of each photograph by its position in the seller’s order (small extractor, up to twelve photographs per listing, 14,245 photographs). The seller’s sequence is a presentation choice: exterior first, principal rooms next, service rooms last.

Table 9. Which rooms are shown within the first ten photographs, and how listings that show them differ (mid-size extractor)

Room	Shown (%)	Median price, shown	Median price, not shown	Mean age, shown	Mean age, not shown	Condition, shown	Condition, not shown
basement	3.1	494,000	548,000	49	42	3.86	3.98
kitchen	67.8	498,900	649,000	41	44	3.95	4.03
bathroom	19.6	469,000	559,000	42	42	3.96	3.98
exterior	95.0	539,000	629,900	42	45	3.97	4.33

Note: “Condition” is the composite visual condition score (1–5). Among listings whose broker field describes the basement as fully finished, 2.1% show it within ten photographs (n=284).

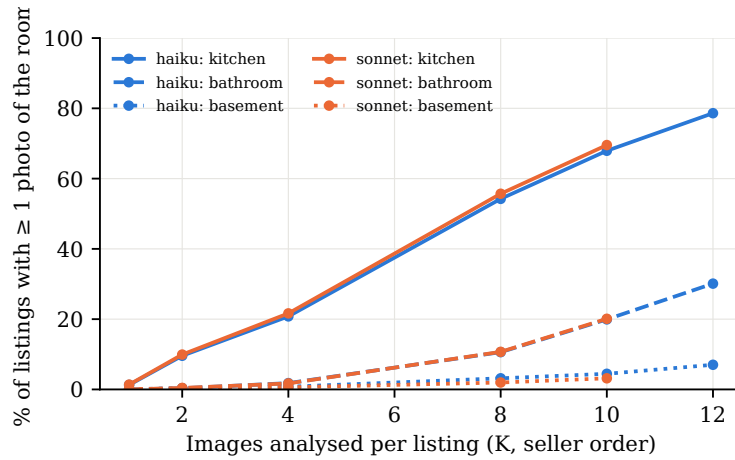


Figure 4. Share of listings with at least one photograph of the kitchen, the bathroom and the basement among the first K photographs, by extractor.

This is the missing-not-at-random structure that the framework anticipated, and it has a direct consequence for the cost approach: basement finish, which adds up to a fifth of the hard cost of a bungalow in the schedule, cannot be read from photographs in 3.1% of listings and must come from the broker’s field or the text, while finish grade is read from kitchens and living rooms and bathroom grade is unavailable for four listings in five. Table 9 also shows that the rooms left out are not left out at random: listings that show a basement ask a lower median price and are older than those that do not, and listings that do not show a kitchen ask more. The pattern is consistent with a presentation logic in which sellers of expensive or newer houses let the principal rooms and the exterior speak and omit service spaces, and with the warning in [Almaas et al. \(2026\)](#) that image selection is strategic.

7.2. Agreement of visual attributes with reference fields

Table 10 and Figure 5 report coverage and agreement for the attributes that have a reference. The references are broker fields, except for storeys (the roll) and building type (the listing’s own type).

Three facts stand out.

Table 10. Coverage and agreement of extracted attributes with reference fields (first ten photographs)

Extractor	Channel	Attribute vs reference	Ref. n	Coverage (%)	Compared	Agreement (%)	κ
haiku	visual	cladding_primary in broker siding set	105	92.4	97	75.3	–
haiku	visual	roof_covering vs broker roofing	435	84.6	368	95.7	0.73
haiku	visual	basement_finish vs broker basement	344	2.6	9	77.8	0.53
haiku	visual	garage present (visual) where broker reports a garage	265	97.0	257	91.4	0.00
haiku	visual	garage type (attached/detached) vs broker	222	88.7	197	90.9	0.75
haiku	visual	storeys_visible vs roll storeys (± 0.5 tolerated)	1,199	95.2	1,142	86.1	0.49
haiku	visual	building_type_visual vs listing type	95	82.1	78	67.9	0.37
haiku	text	Text: cladding in broker siding set	105	100.0	105	91.4	–
haiku	text	Text: roof vs broker roofing	435	99.8	434	98.6	0.92
haiku	text	Text: basement vs broker basement	344	100.0	344	89.5	0.72
haiku	text	Text: storeys vs roll storeys (± 0.5)	1,199	40.1	481	82.1	0.63
haiku	text	Text: year built vs roll year (± 1)	1,200	63.3	760	94.9	–
haiku	text	Text: building footprint $\times$ storeys vs roll floor area (± 15 %)	1,200	31.8	381	55.1	–
sonnet	visual	cladding_primary in broker siding set	105	92.4	97	84.5	–
sonnet	visual	roof_covering vs broker roofing	435	87.1	379	95.8	0.73
sonnet	visual	basement_finish vs broker basement	344	2.0	7	71.4	0.36
sonnet	visual	garage present (visual) where broker reports a garage	265	97.7	259	93.4	0.00
sonnet	visual	garage type (attached/detached) vs broker	222	91.0	202	94.1	0.85
sonnet	visual	storeys_visible vs roll storeys (± 0.5 tolerated)	1,199	94.9	1,138	83.5	0.51
sonnet	visual	building_type_visual vs listing type	95	80.0	76	86.8	0.74
sonnet	text	Text: cladding in broker siding set	105	100.0	105	94.3	–
sonnet	text	Text: roof vs broker roofing	435	100.0	435	98.9	0.93
sonnet	text	Text: basement vs broker basement	344	99.7	343	88.9	0.70
sonnet	text	Text: storeys vs roll storeys (± 0.5)	1,199	55.1	661	82.6	0.62
sonnet	text	Text: year built vs roll year (± 1)	1,200	63.9	767	95.0	–
sonnet	text	Text: building footprint $\times$ storeys vs roll floor area (± 15 %)	1,200	38.3	460	57.4	–

Note: coverage is the share of listings with a reference for which the extractor returns an informative value; agreement is computed on informative cases only. “Cladding in broker siding set” counts the primary visual cladding as agreeing when it belongs to the set of materials the broker lists. Storeys tolerate a half-storey difference; the strict agreement is reported in the text. Text rows are the language model reading the masked text, which contains the broker fields themselves; their high agreement is a check on the masking and parsing, not evidence about images. κ is Cohen’s kappa; it is undefined when the reference has a single class.

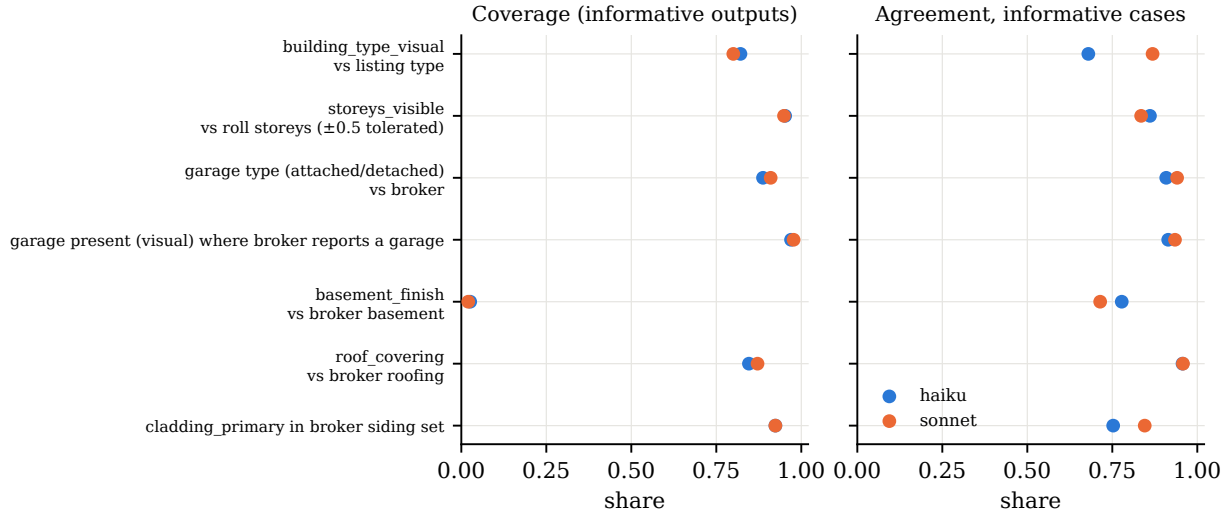


Figure 5. Coverage (left) and agreement with the reference (right) of visual attributes, by extractor, first ten photographs.

First, materials and equipment that are visible from the street are extracted reliably. The mid-size extractor returns an informative roof covering for 87.1% of listings with a broker roofing field and agrees with it in 95.8% of those cases ($\kappa = 0.73$, $n=379$); it names the garage type (attached or detached) for 91.0% of listings whose broker field has one and agrees in 94.1% ($\kappa = 0.85$); its primary cladding belongs to the broker’s list of materials in 84.5% of the 97 comparable cases; and its building type agrees with the listing’s type in 86.8% of the 76 cases where the type is informative ($\kappa = 0.74$). The confusion matrices (Figure 6) show where the residual disagreement lies: membranes on low-slope roofs are read as shingles when only a façade is visible, and the “other” categories absorb cases the broker coded with local vocabulary.

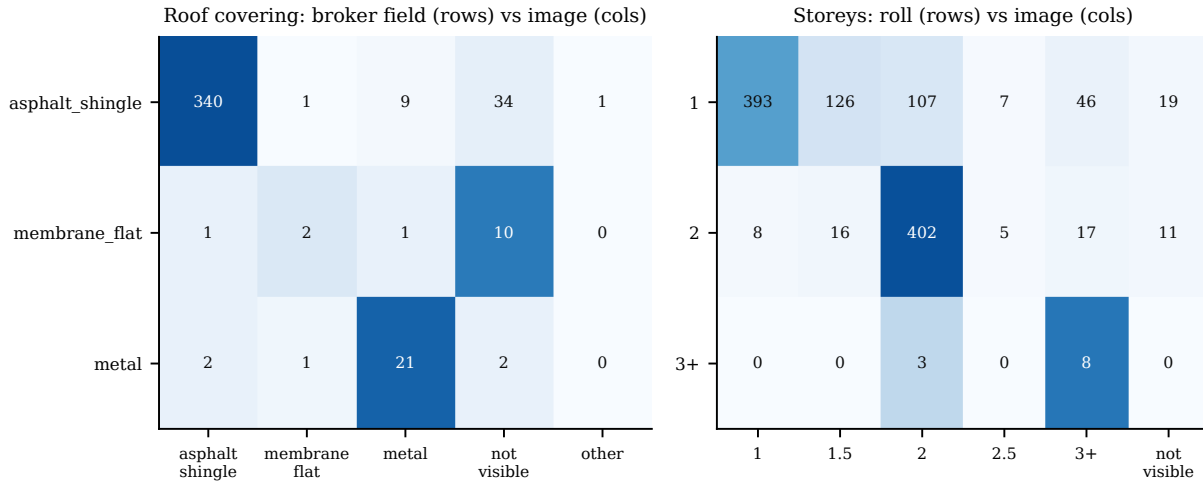


Figure 6. Confusion matrices for roof covering (broker field vs image) and storeys (roll vs image), mid-size extractor; cells give counts, shading row-normalised shares.

Second, storeys are a different kind of attribute. The image gives a count in 94.9% of listings, which agrees with the roll within a half-storey in 83.5% of cases but exactly in only 70.6% ($\kappa = 0.51$). The roll records whole storeys only; the image sees split levels, one-and-a-half-storey cottages and walk-out basements that the assessor codes as one or two. The text channel, which reads the broker’s own storey count, agrees with the roll no better (82.6% within a half-storey, 77.3% exactly), so the disagreement is between definitions as much as between observers. This is why the administrative value overrides the visual one in B4, and why half-storey multipliers in the cost schedule are a source of uncertainty we vary.

Third, the interior attributes that matter most for cost have the thinnest references. Basement finish is informative in the image channel for 2.0% of listings with a broker basement field—the missingness of Section 7.1—and in those few cases agrees in 71.4%. Kitchen counters and cabinets have a usable broker reference for a dozen listings only, too few to report. Finish grades and condition have no reference at all; they are evaluated in Section 7.5.

7.3. Information content of the channels

Table 11 compares, attribute by attribute, how often the image channel and the text channel return an informative value, how often both do, and how often they agree when they do. The channels are complementary in a structured way. Exterior attributes are an image matter: cladding is informative from images for 95% of listings and from text for 30%, and the channels agree in 62% of the cases where both speak; roof covering 82% against 50%, agreeing in 95%. Basement finish is a text matter: images 3%, text 74%, agreement when both 80%. Kitchen grade is mostly an image matter (68% against 19%) but the two channels agree only in 58% of the cases where both grade the

kitchen: the text’s grade is the broker’s adjective, the image’s grade is the counter and the cabinets, and they measure different things. Garage and building type are available from both and agree in 83% and 89% respectively. The either-channel coverage column is what the multimodal baseline B3 works with; for the attributes that enter the cost model it exceeds 70% everywhere except bathroom grade.

Table 11. Information content of the image and text channels (mid-size extractor, first ten photographs)

Attribute	Images (%)	Text (%)	Either (%)	Images only (%)	Text only (%)	Both n	Agreement (%)
building_type	95.0	66.4	99.2	32.8	4.2	746	88.7
storeys	94.9	55.1	97.6	42.5	2.7	629	77.9
cladding	94.9	29.6	96.4	66.8	1.5	337	62.3
roof_covering	82.2	50.2	88.8	38.5	6.5	525	94.9
garage	94.2	65.2	97.2	32.1	3.1	745	83.1
basement_finish	3.1	73.5	74.1	0.6	71.0	30	80.0
kitchen_grade	67.8	19.4	73.9	54.5	6.2	159	58.5
bathroom_grade	19.6	15.9	32.1	16.2	12.5	41	53.7
kitchen_counters	67.7	10.2	71.1	60.8	3.4	82	91.5
flooring_main	88.7	52.7	93.8	41.1	5.1	571	73.9

The text channel also reads two administrative quantities when brokers state them: the year built, which agrees with the roll within a year in 95.0% of the 767 listings where it is stated, and the building footprint, which, multiplied by the roll’s storey count, falls within 15% of the roll’s floor area in 57.4% of the 460 listings that give dimensions, with a median ratio of 1.020 and a median absolute deviation of 10.5%. The footprint result is a reminder that the roll’s floor-area concept (sum of above-grade floor areas) and the broker’s building dimensions describe the same house differently, and that neither should be inferred from a photograph.

7.4. Image budget

Because aggregation is deterministic, truncating the photograph set to the first K in seller order measures what an analyst gets for a smaller image budget without new model calls (Table 12, Figure 7). With a single photograph, almost always the façade, exterior attributes are nearly complete but no finish grade exists. With four photographs the small extractor has a kitchen for 20.8% of listings and a composite grade for 57.8%; with eight, 54.2% and 86.1%; with twelve, 78.6% and 94.3%. The grade computed from eight photographs equals the one computed from twelve in most listings where the latter exists, and the composite condition moves by a median of a few hundredths of a point between eight and twelve. The basement share rises from 3.2% at eight to 7.0% at twelve: more photographs help, but at the seller’s pace, not the analyst’s. The practical

implication for a mass-appraisal deployment is that about eight photographs capture what the listing will reveal about grade and condition, and that no budget captures the basement.

Table 12. Attribute availability and stability by number of photographs analysed (K , seller order)

Extractor	K	Exterior (%)	Kitchen (%)	Bathroom (%)	Basement (%)	Grade avail. (%)	Grade = full (%)	Cond. MAD
haiku	1	95	1	0	0	3	2	0.12
haiku	2	96	10	0	0	28	19	0.10
haiku	4	98	21	2	1	58	43	0.06
haiku	8	98	54	11	3	86	79	0.03
haiku	10	98	68	20	4	91	88	0.01
haiku	12	98	79	30	7	94	100	0.00
sonnet	1	94	1	0	0	3	3	0.20
sonnet	2	95	10	0	0	29	21	0.16
sonnet	4	97	22	2	1	59	44	0.10
sonnet	8	97	56	11	2	87	82	0.02
sonnet	10	98	70	20	3	91	100	0.00

Note: “Grade = full” is the share of listings with a composite grade at the full budget whose grade at K is identical; “Cond. MAD” is the median absolute difference of the composite condition at K from its full-budget value.

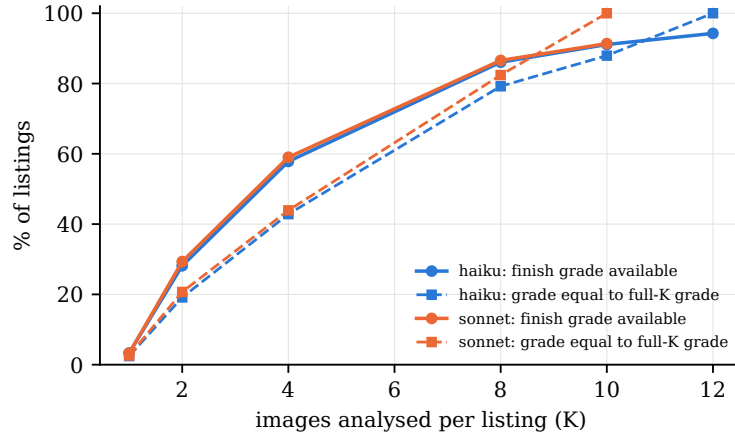


Figure 7. Availability and stability of the composite finish grade as a function of the number of photographs analysed.

7.5. Reliability of judgment attributes across extractors

Finish grades and condition scores have no external reference in our data, and expert labels do not exist (Section 11). What can be measured is whether different extractors, given the same photographs, reach the same judgments; disagreement between them bounds from below the measurement error of each. Table 13 and Figure 8 report pairwise agreement, Cohen’s κ and Krippendorff’s nominal α across the tiers on the common photographs, and the correlation of

condition scores. At the level of the individual photograph the small and mid-size extractors assign the same scene to 88.2% of 11,916 images ($\kappa = 0.89$). At the property level, factual attributes are reliable—roof covering agrees in 97.2% of listings where both are informative ($\kappa = 0.85$), with $\alpha = 0.85$ across the tiers—while judgment attributes are not: the kitchen grade agrees in 72.0% ($\kappa = 0.55$, $\alpha = 0.55$) and the composite grade in 74.2% ($\alpha = 0.56$). The composite condition scores of the two extractors correlate at $r = 0.88$ with a median absolute difference of about two tenths of a point on the five-point scale, and the smaller extractor is slightly more generous on average. By the conventional thresholds for coding reliability, the factual attributes would be acceptable as reference labels and the judgment attributes would not; the latter are usable as ordinal proxies whose scale depends on the extractor, which is exactly how the cost model treats them, through a condition factor whose strength is a sensitivity parameter.

Table 13. Cross-extractor agreement on the common first ten photographs

Attribute	n	α	agree. (%) haiku–sonnet	κ haiku–sonnet	r haiku–sonnet
building type visual	1169	0.79	92	0.79	–
storeys visible	1169	0.77	85	0.77	–
cladding primary	1169	0.68	75	0.68	–
roof covering	1169	0.85	97	0.85	–
garage visible	1169	0.84	89	0.84	–
basement finish	1169	0.54	88	0.54	–
kitchen grade	1169	0.55	72	0.55	–
bathroom grade	1169	0.60	74	0.60	–
kitchen counters	1169	0.73	83	0.73	–
kitchen cabinets	1169	0.62	73	0.63	–
flooring main	1169	0.82	89	0.82	–
renovation evidence	1169	0.49	69	0.49	–
grade composite	1169	0.56	74	0.57	–
condition composite	1169	–	–	–	0.88
exterior condition	1169	–	–	–	0.81
interior condition	1169	–	–	–	0.86
photo quality mean	1169	–	–	–	0.74

Note: α is Krippendorff’s nominal alpha across all available tiers; pairwise statistics are computed on listings where both extractors return an informative value. For continuous scores the pairwise statistic is the Pearson correlation.

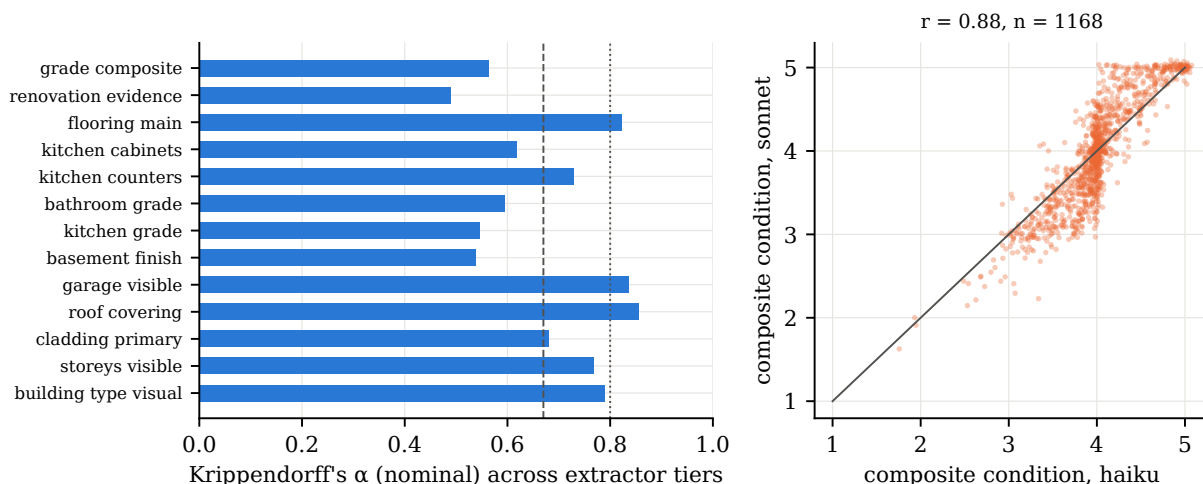


Figure 8. Left: Krippendorff's α across extractor tiers by attribute, with the 0.67 and 0.80 conventions marked. Right: composite condition scores of the two full-sample extractors for the same listings.

7.6. Confidence and abstention

The schema asks each extractor for a confidence in its scene-relevant attributes. Figure 9 and Table 14 relate stated confidence, per exterior image, to agreement with the broker's roofing and siding fields. Stated confidence is informative but compressed: almost all values lie between 0.7 and 0.9, agreement rises with confidence within that range, and the models are slightly under-confident on these easy exterior facts. Selective prediction confirms the usable direction: keeping only images with confidence at least 0.8 retains 72% of the mid-size extractor's exterior images and raises roof agreement from 95.0% to 96.8%; a 0.9 threshold retains only 5% of images. The abstention mechanism—answering `not_visible`—is used heavily and appropriately for interior attributes on exterior photographs and for roofs seen from the street, as the coverage columns of Table 10 show. We did not find stated confidence to be a reliable guide for the judgment attributes, whose disagreement across extractors is largely unrelated to either extractor's confidence; a review rule based on confidence alone would flag a large share of listings without targeting the uncertain ones (Section 9.6).

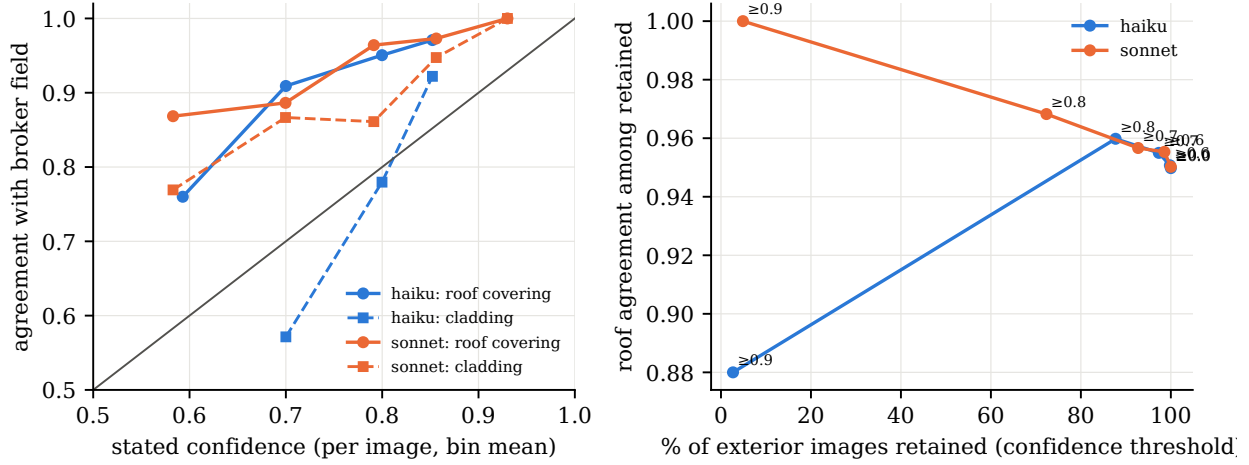


Figure 9. Left: agreement with broker fields by stated-confidence bin (per exterior image). Right: selective prediction—roof agreement among images retained at increasing confidence thresholds.

Table 14. Stated confidence and agreement with broker fields, per exterior image

Extractor	Confidence bin	Images	Mean conf.	Roof n	Roof agreement	Cladding n	Cladding agreement
haiku	(-0.001, 0.6]	156	0.59	25	0.76	0	–
haiku	(0.6, 0.7]	371	0.70	88	0.91	14	0.57
haiku	(0.7, 0.8]	1626	0.80	445	0.95	118	0.78
haiku	(0.8, 0.9]	1122	0.85	378	0.97	77	0.92
sonnet	(-0.001, 0.6]	318	0.58	76	0.87	13	0.77
sonnet	(0.6, 0.7]	533	0.70	132	0.89	15	0.87
sonnet	(0.7, 0.8]	1565	0.79	501	0.96	108	0.86
sonnet	(0.8, 0.9]	929	0.86	333	0.97	76	0.95
sonnet	(0.9, 1.0]	3	0.93	3	1.00	1	1.00

8. Results II: the cost indication

8.1. Decomposition and the effect of adding channels

Table 15 reports, for the mid-size extractor and both unit-cost anchors, the median cost indication under the five baselines, its decomposition and its alignment with the two benchmarks; Figure 10 shows the decomposition and Figure 11 the distributions of the two alignment ratios.

Table 15. Cost indications by information set and unit-cost anchor (mid-size extractor, $n = 1,200$)

Level	Anchor	Median V (\$)	Land share	Median D	Assessor ratio	Assessor MdAPE	Price ratio	Price MdAPE [95% CI]	Within 20%
B0	B	407,545	0.40	0.62	0.68	43.9	0.75	30.1 [26.2, 33.9]	34
B0	A	286,300	0.56	0.62	0.35	65.5	0.54	46.5 [42.7, 50.4]	6
B1	B	464,934	0.33	0.48	0.94	43.6	0.86	26.2 [24.5, 29.1]	40
B1	A	332,103	0.47	0.48	0.53	48.2	0.62	38.4 [33.1, 41.9]	15
B2	B	492,981	0.31	0.38	1.03	44.8	0.91	26.3 [25.0, 28.6]	38
B2	A	337,322	0.45	0.38	0.55	45.1	0.63	37.0 [31.8, 40.1]	19
B3	B	514,919	0.30	0.38	1.12	48.2	0.95	27.9 [26.2, 30.1]	37
B3	A	361,134	0.43	0.38	0.63	39.3	0.68	32.5 [28.6, 35.9]	27
B4	B	523,731	0.29	0.38	1.16	49.5	0.98	28.0 [26.3, 30.4]	37
B4	A	365,612	0.42	0.38	0.63	38.1	0.68	31.9 [27.9, 35.5]	28

Note: “Assessor ratio” is the median of depreciated improvements over the roll’s building value indexed with the metropolitan construction price index from the roll’s market-conditions date to 2026-Q2; “Price ratio” is the median of the full indication over the asking price. MdAPE is the median absolute percentage deviation of the ratio from one; the bracketed interval is a 95% cluster bootstrap by municipality (499 replications). Anchor B: Québec 2026 industry guides; anchor A: Altus Montréal hard costs.

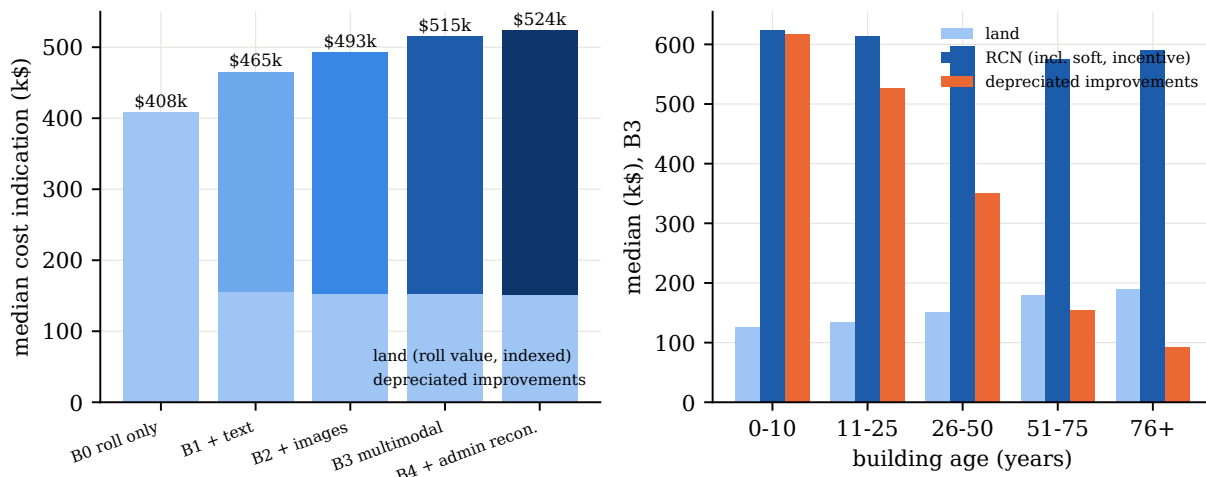


Figure 10. Left: median cost indication by information set, split into indexed land value and depreciated improvements (anchor B). Right: median land, replacement cost new and depreciated improvements by building-age band under B3.

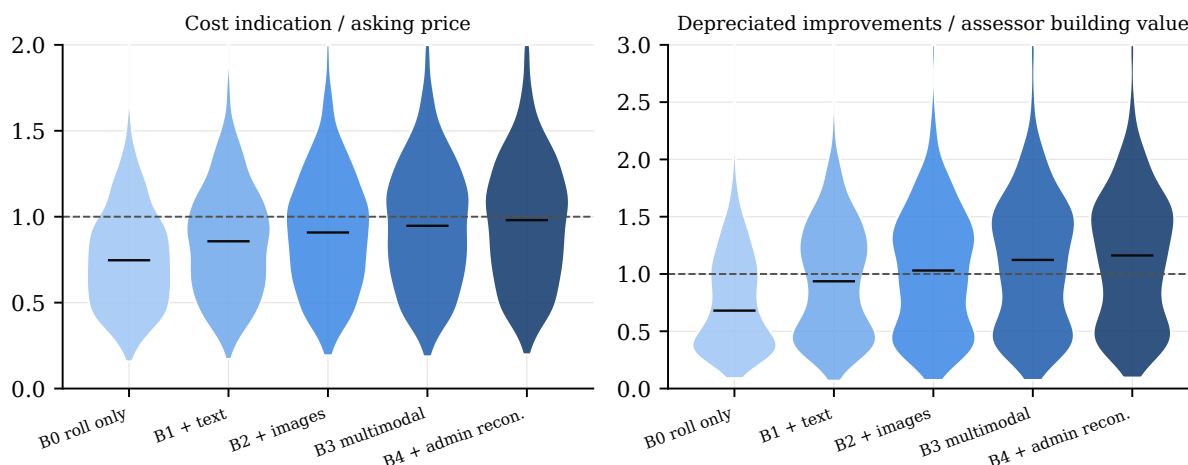


Figure 11. Distributions of the market-alignment ratio (left) and the assessor-alignment ratio (right) by information set, anchor B. Horizontal bars are medians; the dashed line marks a ratio of one.

Under the Québec anchor, the roll-only baseline B0 produces a median indication of \$407,545 with land accounting for 0.40 of it, a median depreciation of 0.62, a ratio to the indexed assessor building value of 0.68 and a ratio to the asking price of 0.75 (MdAPE 30.1%, 95% interval 26.2–33.9). Adding the text channel (B1) lifts the median indication to \$464,934; adding the image channel instead (B2) lifts it to \$492,981; the multimodal baseline B3 reaches \$514,919 with depreciation 0.38, an assessor ratio of 1.12 and a price ratio of 0.95 (MdAPE 27.9%, interval 26.2–30.1); B4 adds little (\$523,731, price ratio 0.98). The share of listings whose indication falls within 20% of the asking price rises from 34% under B0 to 37% under B3.

Three readings follow. The channels act almost entirely through depreciation: the median effective age falls from 37 years under B0 to 23 under B3, while the median replacement cost new moves from \$538,001 to \$598,525, a change driven by basements and garages read from the text and by the grade read from kitchens. The improvement in market alignment is real but modest and concentrated in the level of the indication rather than in its dispersion: the median ratio moves from three quarters to close to one, the MdAPE falls by a few points, and the bootstrap intervals of B0 and B3 do not overlap, but the dispersion of the ratio around its median—visible in Figure 11—is essentially unchanged. And the images are worth about as much as the text and the two together are worth more than either: B2 and B1 are close, B3 improves on both, which is what one expects when the image channel carries the exterior and the kitchen and the text channel carries the basement.

The anchor matters more than any channel. Under anchor A the same B3 indication has a median of \$361,134, an assessor ratio of 0.63 and a price ratio of 0.68; the -29.9% difference between anchors dwarfs the +7.3% of the largest depreciation scenario. Whatever listing media contribute to the specification of the building, the published unit cost sets the level of the indication.

8.2. Effect of visually inferred grade and condition

Figure 12 isolates what the image channel changes relative to the roll-only baseline, listing by listing. The composite finish grade shifts replacement cost by construction: a basic grade lowers depreciated improvements relative to the standard default, an upgraded or luxury grade raises them, with medians that follow the spread of the schedule. The condition score moves depreciation by several points in either direction, with the larger moves for older houses as Proposition 2 predicts, since the condition factor multiplies chronological age. The extractors rate most listings above average—the composite condition is centred near four on the five-point scale—so the image channel mostly *reduces* depreciation relative to B0’s neutral assumption; whether this reflects genuinely well-kept housing stock, selective photography of the best rooms, or a generous scale is exactly what an expert annotation study would settle and what the cross-extractor disagreement of Section 7.5 leaves open.

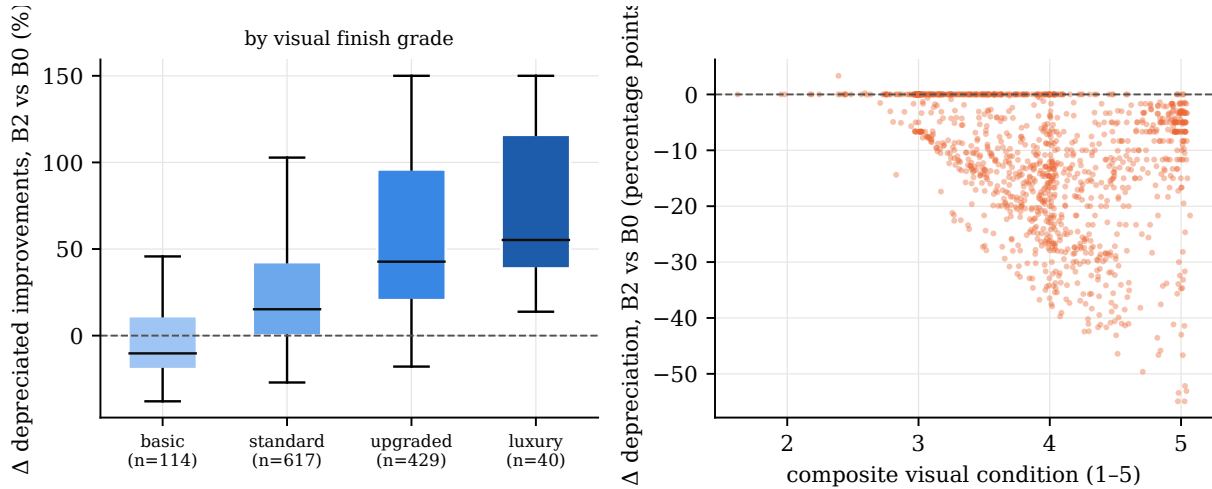


Figure 12. Listing-level effect of the image channel (B2 vs B0). Left: change in depreciated improvements by composite finish grade. Right: change in the depreciation rate against the composite visual condition.

8.3. Heterogeneity by age, price tier and region

Table 16 and Figure 13 report the alignment ratios by building-age band, asking-price quartile, metropolitan area and roll vintage under B3. The dominant pattern is an age gradient that no channel removes: for houses built in the last ten years the median depreciated improvements exceed the indexed assessor building value by half (ratio 1.53) and the indication exceeds the asking price (ratio 1.26); for houses older than 75 years the improvements are worth 0.40 of the assessor’s figure and the indication is 0.54 of the price. The 26–50-year band is where the schedule aligns best (price MdAPE 17.2%). Read through Proposition 2 and the depreciation literature, the gradient says that a straight-line age–life schedule with a 60-year life depreciates old houses faster than either assessors

or the market do and leaves too little depreciation—or too high a unit cost—for new ones. The cheapest quartile is systematically under-indicated (price ratio 0.81) and the three upper quartiles are close to one; across metropolitan areas the price ratio ranges from 0.89 in Montréal, where land is a larger share of value and our uniform regional level is most likely too low, to above one in Gatineau and Trois-Rivières. The roll-vintage split shows that indexing the assessor’s building value by the construction price index does not fully reconcile rolls referenced in 2022, 2023 and 2024; the 2024 rolls, which incorporate the post-2021 boom, yield the lowest assessor ratio.

Table 16. Alignment of the B3 indication by subgroup (mid-size extractor, anchor B)

Group	Value	n	Assessor ratio	Assessor COD	Price ratio	Price MdAPE B3	Price MdAPE B0
cma	gatineau	184	1.15	38.2	1.08	27.8	21.8
cma	montreal	203	0.84	50.1	0.89	23.8	30.1
cma	other	201	1.04	50.4	0.88	35.7	34.7
cma	quebec	198	1.17	39.6	0.95	24.4	26.3
cma	saguenay	127	1.27	42.0	0.94	24.8	41.8
cma	sherbrooke	175	1.20	40.5	0.91	27.5	30.0
cma	trois-rivieres	112	1.43	35.9	1.03	30.4	34.6
roll vintage	2022	267	1.14	38.1	1.05	27.6	23.6
roll vintage	2023	718	1.15	43.3	0.93	28.2	33.1
roll vintage	2024	215	0.90	59.0	0.92	26.4	29.7
age band	0-10	233	1.53	22.0	1.26	30.1	16.6
age band	11-25	256	1.48	23.0	1.19	22.0	14.6
age band	26-50	276	1.14	31.7	0.97	17.2	32.3
age band	51-75	241	0.58	54.6	0.68	33.1	49.2
age band	76+	194	0.40	43.4	0.54	46.1	47.6
price tier	Q1	308	0.74	57.4	0.81	30.0	32.5
price tier	Q2	333	1.19	39.4	0.99	26.2	28.5
price tier	Q3	296	1.26	37.4	1.04	26.2	30.0
price tier	Q4	263	1.30	39.5	0.99	28.8	30.3

Note: COD is the coefficient of dispersion of the assessor ratio (mean absolute deviation from the median, in percent of the median), the ratio-study statistic of mass-appraisal standards. The last column gives the B0 price MdAPE for the same subgroup.

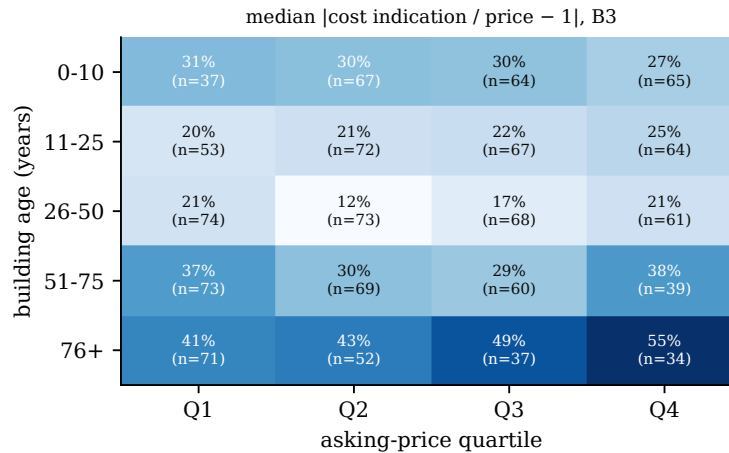


Figure 13. Median absolute deviation of the B3 indication from the asking price by building-age band and asking-price quartile.

8.4. Calibration against the assessor and geographic hold-out

The schedule's parameters are public figures and declared assumptions, not estimates, and the previous subsections show them to be jointly off in level and in age slope. To learn how far, without ever using prices, we calibrate one parameter at a time against the assessor benchmark on the five metropolitan areas other than Québec City and evaluate on Québec City (Table 17). Two calibrations are informative. Adjusting only the unit-cost level so that the median assessor ratio equals one on the calibration areas requires a factor of 0.90 under anchor B, an implied standard hard cost of \$272 per square foot, and 1.75 under anchor A, an implied \$306; both land near the lower edge of the Québec industry band and far above the Montréal hard-cost band. On the Québec City hold-out the level-calibrated B3 indication has an assessor ratio of 1.06 (anchor B) and 1.07 (anchor A) and a price MdAPE of 21.5% and 22.3%, against 24.4% with the public parameters: the level transfers. Adjusting only the economic life, by contrast, drives the parameter to implausible values under anchor A (200 years) because the level, not the slope, is what is wrong there; under anchor B the roll-only baseline asks for a life of 90 years and the multimodal baseline for 50, a reminder that the condition adjustment and the life are substitutes. These calibrations are diagnostics of the schedule, not part of the proposed method; the paper's headline indications remain those with public parameters.

Table 17. Calibration of one schedule parameter against the assessor benchmark and transfer to the Québec City hold-out

Anchor	Level	Set	Parameters	n	Median D	Assessor ratio	Assessor MdAPE	Price ratio	Price MdAPE
B	B0	calibration	public	1002	0.60	0.68	45.0	0.74	31.1
B	B0	calibration	life-only (90 y)	1002	0.40	1.00	31.1	0.89	21.8
B	B0	calibration	level-only (x1.5 to standard approx. 453/sq ft hard cost)	1002	0.60	1.01	48.9	0.94	34.9
B	B0	calibration	joint (x1.8, life 50 y)	1002	0.72	0.99	48.3	0.94	38.5
B	B0	holdout	public	198	0.63	0.71	38.5	0.76	26.3
B	B0	holdout	life-only (90 y)	198	0.42	1.04	23.1	0.91	16.4
B	B0	holdout	level-only (x1.5 to standard approx. 453/sq ft hard cost)	198	0.63	1.07	42.8	0.94	28.2
B	B0	holdout	joint (x1.8, life 50 y)	198	0.76	1.10	43.7	0.93	33.1
B	B3	calibration	public	1002	0.38	1.10	48.5	0.95	28.4
B	B3	calibration	life-only (50 y)	1002	0.45	0.98	51.1	0.88	31.6
B	B3	calibration	level-only (x0.9 to standard approx. 272/sq ft hard cost)	1002	0.38	1.00	43.4	0.89	26.1
B	B3	calibration	joint (x0.9, life 60 y)	1002	0.38	1.00	43.4	0.89	26.1
B	B3	holdout	public	198	0.36	1.17	45.9	0.95	24.4
B	B3	holdout	life-only (50 y)	198	0.43	1.05	44.1	0.91	25.4
B	B3	holdout	level-only (x0.9 to standard approx. 272/sq ft hard cost)	198	0.36	1.06	40.1	0.90	21.5
B	B3	holdout	joint (x0.9, life 60 y)	198	0.36	1.06	40.1	0.90	21.5
B	B4	calibration	public	1002	0.38	1.15	49.9	0.98	28.6
B	B4	calibration	life-only (50 y)	1002	0.45	1.04	51.2	0.91	31.9
B	B4	calibration	level-only (x0.85 to standard approx. 257/sq ft hard cost)	1002	0.38	0.99	41.2	0.88	25.1
B	B4	calibration	joint (x1.0, life 50 y)	1002	0.45	1.04	51.2	0.91	31.9
B	B4	holdout	public	198	0.36	1.21	45.8	0.97	24.1
B	B4	holdout	life-only (50 y)	198	0.43	1.13	44.8	0.94	25.8
B	B4	holdout	level-only (x0.85 to standard approx. 257/sq ft hard cost)	198	0.36	1.05	35.4	0.90	20.4
B	B4	holdout	joint (x1.0, life 50 y)	198	0.43	1.13	44.8	0.94	25.8
A	B0	calibration	public	1002	0.60	0.34	65.8	0.53	47.1
A	B0	calibration	life-only (200 y)	1002	0.18	0.72	29.2	0.70	30.6
A	B0	calibration	level-only (x2.2 to standard approx. 385/sq ft hard cost)	1002	0.60	0.75	47.2	0.79	31.6
A	B0	calibration	joint (x2.2, life 75 y)	1002	0.48	0.94	39.4	0.88	27.3
A	B0	holdout	public	198	0.63	0.38	62.4	0.56	43.6
A	B0	holdout	life-only (200 y)	198	0.19	0.75	27.3	0.73	27.4
A	B0	holdout	level-only (x2.2 to standard approx. 385/sq ft hard cost)	198	0.63	0.83	41.2	0.80	28.3
A	B0	holdout	joint (x2.2, life 75 y)	198	0.51	1.00	36.1	0.90	22.6
A	B3	calibration	public	1002	0.38	0.62	40.5	0.67	33.1
A	B3	calibration	life-only (200 y)	1002	0.11	0.95	18.7	0.84	18.7
A	B3	calibration	level-only (x1.75 to standard approx. 306/sq ft hard cost)	1002	0.38	1.00	43.5	0.89	26.6
A	B3	calibration	joint (x1.5, life 75 y)	1002	0.30	1.00	31.3	0.87	21.5
A	B3	holdout	public	198	0.36	0.65	35.2	0.70	30.6
A	B3	holdout	life-only (200 y)	198	0.11	0.95	20.1	0.84	17.4
A	B3	holdout	level-only (x1.75 to standard approx. 306/sq ft hard cost)	198	0.36	1.07	40.7	0.90	22.3
A	B3	holdout	joint (x1.5, life 75 y)	198	0.29	1.04	27.9	0.88	17.6
A	B4	calibration	public	1002	0.38	0.63	38.8	0.68	32.3
A	B4	calibration	life-only (200 y)	1002	0.11	0.97	18.1	0.85	17.7
A	B4	calibration	level-only (x1.7 to standard approx. 298/sq ft hard cost)	1002	0.38	1.00	41.6	0.89	26.0
A	B4	calibration	joint (x1.7, life 60 y)	1002	0.38	1.00	41.6	0.89	26.0
A	B4	holdout	public	198	0.36	0.67	33.8	0.71	29.4
A	B4	holdout	life-only (200 y)	198	0.11	0.96	21.6	0.85	16.9
A	B4	holdout	level-only (x1.7 to standard approx. 298/sq ft hard cost)	198	0.36	1.06	39.1	0.90	22.3
A	B4	holdout	joint (x1.7, life 60 y)	198	0.36	1.06	39.1	0.90	22.3

Note: calibration set = Montréal, Gatineau, Sherbrooke, Trois-Rivières, Saguenay and other areas ($n = 1002$); hold-out = Québec City area ($n = 198$). The calibration target is a median assessor ratio of one; prices are never used.

8.5. Market alignment, production benchmarks and recorded sales

Table 18 places the indications next to the asking price and next to the three production estimates that Vrai Prix stores for each listing, which are market-value estimates and are shown for context only. The Vrai Prix reconciled estimate sits at 0.94 of the asking price with an MdAPE of 11.5% and 73% of listings within 20%; its production “cost estimate”, whose formula we could not inspect, at 0.95 and 16.3%. Our B3 indication, which uses no price information at all, reaches a median ratio of 0.95 with an MdAPE of 27.9% and 37% within 20%, and its median ratio to the production cost estimate is 1.00. A cost approach with public parameters is thus twice as dispersed around asking prices as a hedonic model trained on half a million sales; it is not a competing price predictor, and the comparison is reported so that readers can see the size of the gap, not to close it.

Table 18. Market alignment of the indications and of the Vrai Prix production estimates (mid-size extractor, anchor B)

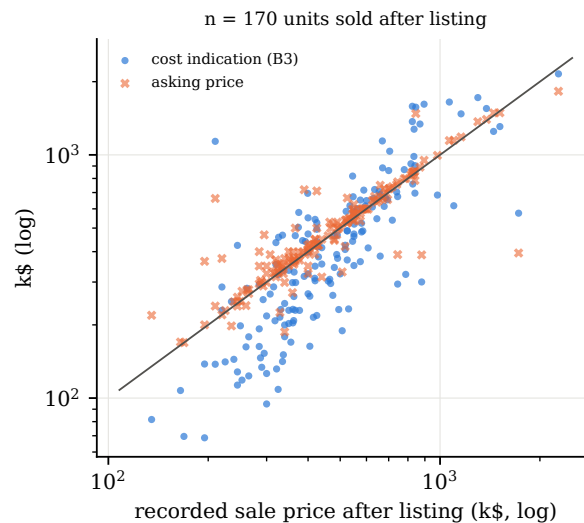
Indication	n	Median ratio to price	MdAPE	P10	P90	Within 20%	Ratio to VP cost est.
B0	1200	0.75	30.1	0.40	1.19	34	0.75
B1	1200	0.86	26.2	0.44	1.32	40	0.92
B2	1200	0.91	26.3	0.46	1.38	38	0.97
B3	1200	0.95	27.9	0.46	1.44	37	1.00
B4	1200	0.98	28.0	0.48	1.46	37	1.03
Benchmark: vp_est	1200	0.94	11.5	0.72	1.16	73	–
Benchmark: vp_cost_est	1200	0.95	16.3	0.67	1.30	59	–
Benchmark: roll_model_est	1200	0.93	10.7	0.72	1.12	75	–

Note: ratios to the asking price; the production benchmarks (“vp_est” reconciled estimate, “vp_cost_est” production cost estimate, “roll_model_est” pre-computed hedonic estimate) are market-value estimates shown for context. “Ratio to VP cost est.” is the median ratio of our indication to the production cost estimate.

For the 146 frame listings whose roll unit recorded a sale after the listing appeared (processed with the two cheaper extractors), the B3 indication has a median ratio to the recorded sale price of 0.84 and an MdAPE of 28.7%, against 0.70 and 34.2% for B0; the asking price exceeded the sale price by a median of 1.025 in this set (Table 19, Figure 14). The sample is small and the sellers who closed quickly are not a random draw of the frame; we report it because it is the only comparison with a transaction rather than an asking price, and we label it as a market-alignment statistic.

Table 19. Alignment with recorded sale prices for units sold after listing

Set	Level	n	Median ratio to sale	MdAPE	Within 20%	Asking / sale (median)
sold supplement	B0	146	0.70	34.2	26	1.025
sold supplement	B3	146	0.84	28.7	40	1.025
sample	B0	24	0.76	29.7	25	1.019
sample	B3	24	0.89	29.0	38	1.019

**Figure 14.** B3 cost indication and asking price against the recorded sale price for units sold after listing (log scales).

9. Robustness, uncertainty, missingness and failure cases

9.1. Sensitivity of the indication to schedule and depreciation assumptions

Table 20 and Figure 15 vary each assumption of the schedule one at a time around the B3 baseline. The ordering is instructive. The unit-cost anchor dominates (-29.9% for anchor A). Within anchor B, the economic life (-7.1% at 50 years, +7.3% at 75), the regional level (-6.7% and +6.1% for $\pm 10\%$), the external-obsolescence scenario (-7.3%), a flat storey multiplier (-7.2%), the strength of the condition adjustment (-5.6% at half, +5.2% at one and a half) and whether the roll land value is indexed to the valuation date (-7.0% when it is not) each move the median indication by five to seven percent; the grade band, the mark-ups and the add-ons move it by two to five percent. Removing depreciation altogether raises the indication by +50.5% and takes its ratio to the asking price to 1.43—replacement cost new plus land exceeds the asking price for the median house, which is the

normal state of a market in which existing houses sell below reproduction, and which confirms that the question is the size of depreciation, not the existence of a gap.

Table 20. One-at-a-time sensitivity of the B3 indication (mid-size extractor)

Scenario	Median <i>V</i> (\$)	Δ vs baseline (%)	Median <i>D</i>	Assessor ratio	Price ratio	Price MdAPE
baseline (anchor B, Québec 2026 guides)	514,919	+0.0	0.38	1.12	0.95	27.9
anchor A (Altus Montréal hard costs)	361,134	-29.9	0.38	0.63	0.68	32.5
non-metro regions –15 % (Dessins Drummond 2026)	485,055	-5.8	0.38	1.04	0.90	26.2
storey multiplier flat (1.0 all)	477,991	-7.2	0.38	1.02	0.90	26.8
regional level 0.90	480,227	-6.7	0.38	1.02	0.89	25.6
regional level 1.10	546,294	+6.1	0.38	1.22	1.01	30.2
relative growth since 2023 (CMA BCPI / Montréal BCPI)	503,408	-2.2	0.38	1.09	0.94	27.2
grade costs: low band	491,794	-4.5	0.38	1.05	0.91	26.5
grade costs: high band	541,170	+5.1	0.38	1.21	1.00	30.1
soft costs 5 %	498,372	-3.2	0.38	1.07	0.92	26.3
soft costs 15 %	531,069	+3.1	0.38	1.17	0.98	29.0
entrepreneurial incentive 0 %	486,981	-5.4	0.38	1.04	0.90	25.9
entrepreneurial incentive 15 %	538,116	+4.5	0.38	1.19	0.99	29.3
add-ons –30 %	501,932	-2.5	0.38	1.09	0.93	26.7
add-ons +30 %	523,369	+1.6	0.38	1.15	0.96	28.4
economic life 50 y	478,254	-7.1	0.45	0.99	0.88	30.4
economic life 75 y	552,293	+7.3	0.30	1.26	1.02	24.8
condition→effective age: half strength	485,852	-5.6	0.45	1.00	0.88	28.7
condition→effective age: 1.5× strength	541,831	+5.2	0.29	1.24	1.01	27.2
external obsolescence 10 %	477,108	-7.3	0.44	1.01	0.89	25.4
land not indexed	479,117	-7.0	0.38	1.12	0.89	28.7
land half indexed	499,105	-3.1	0.38	1.12	0.92	28.2
no depreciation (RCN + S + L)	775,115	+50.5	0.00	1.98	1.43	43.1
no depreciation, land not indexed	741,790	+44.1	0.00	1.98	1.36	36.9

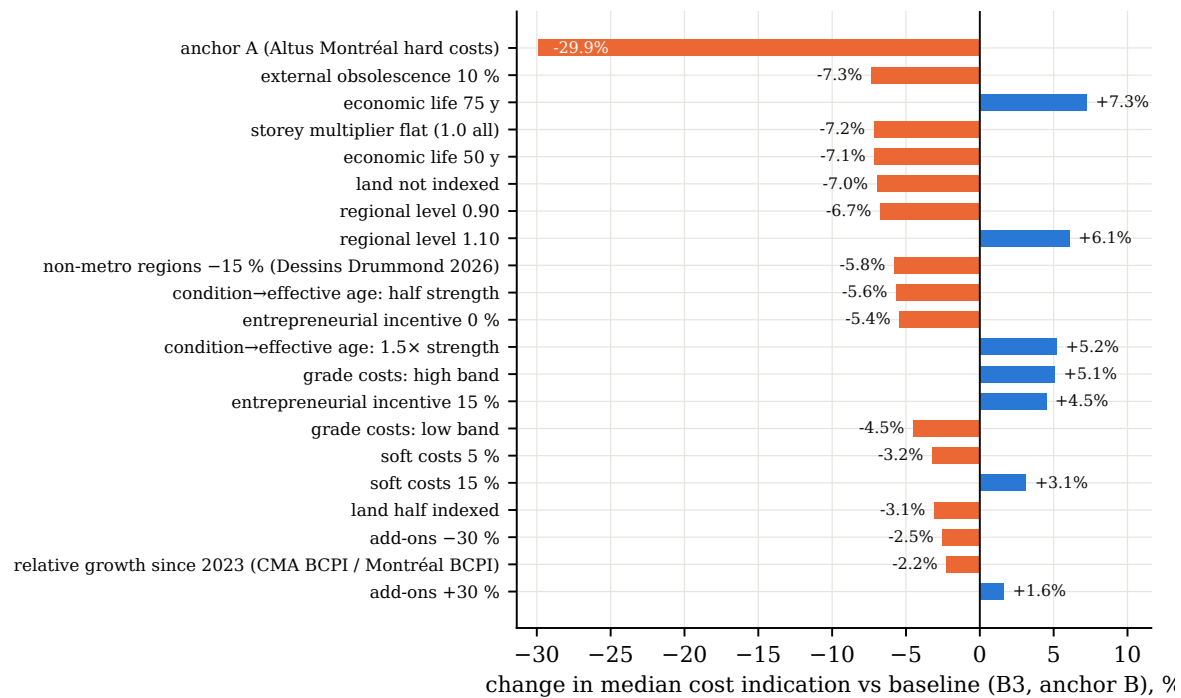


Figure 15. Change in the median B3 indication under each alternative assumption, relative to the anchor-B baseline.

9.2. Counterfactual scenarios

Table 21 reports labelled counterfactuals applied to every listing: a one-level change in the composite grade, a one-point change in the composite condition, extensive renovation evidence, a finished basement and an attached garage. Among listings whose indication changes, a one-level grade shift moves the indication by about +8.8% up or -8.4% down, a one-point condition shift by +13.6% up or -10.9% down, and renovation evidence by +12.1% for the 48% of listings not already read as extensively renovated; a finished basement adds +4.3% for the 49% of listings whose basement is not already recorded as finished and a garage +3.8% for the 64% without one. These magnitudes are the stakes of an extraction error on each attribute, and they are of the same order as the agreement gaps measured in Section 7: a kitchen misgraded by one level costs about as much as the whole image channel adds.

Table 21. Counterfactual scenarios on the B3 indication (labelled scenarios, not predictions)

Scenario (counterfactual, labelled)	Listings affected (%)	Median ΔV , affected (%)	Median Δ improvements, affected (%)	Median ΔV , affected (\$)
grade +1 level	96	+8.8	+12.6	+42,113
grade -1 level	90	-8.4	-12.6	-45,322
condition +1	70	+13.6	+21.8	+67,290
condition -1	66	-10.9	-17.3	-65,254
extensive renovation evidence	48	+12.1	+18.7	+54,062
basement finished	49	+4.3	+7.0	+15,905
attached garage added	64	+3.8	+7.1	+12,803

9.3. Monte Carlo propagation of attribute and parameter uncertainty

The Monte Carlo layer draws, for each of 400 replications, a grade and a condition around the extracted values with a spread that grows with the extractor’s stated uncertainty and with the absence of an informative observation, together with unit costs within the published band, an economic life between 50 and 75 years, mark-ups and land indexing within their ranges and a regional level within $\pm 10\%$. Figure 16 and Table 22 summarise the result. The 10–90% interval has a median relative width of 26% of the indication (interquartile range 22–39%), and one-at-a-time perturbations rank the sources of that width: grade (8.8% for a one-level shift), mark-ups (7.8%), condition (6.4%), regional level (6.4%), land indexing (5.7%), unit-cost band (5.4%) and economic life (2.3%). The intervals are honest about what they contain and silent about what they do not: they cover the asking price for 28% of listings and the indexed assessor total for 28%, far below the nominal 80%. The under-coverage is not a defect of the propagation but a measurement of what it omits—the dispersion of market prices around any cost indication, the error of the assessor’s own land and building split, and the latent attributes—and it is the quantitative reason the paper describes the indication as a screening tool.

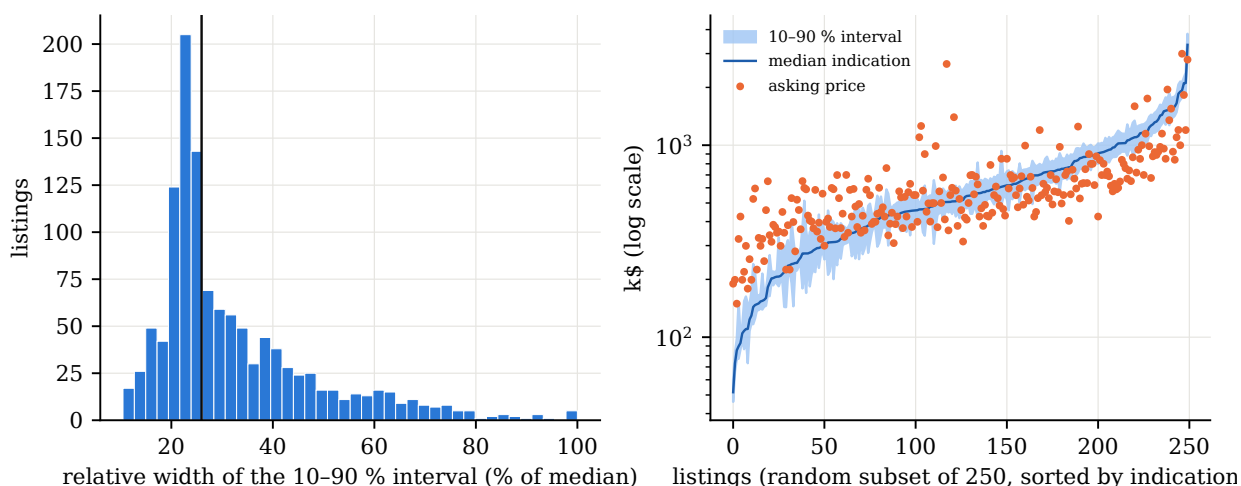


Figure 16. Left: distribution of the relative width of the 10–90% Monte Carlo interval. Right: intervals, median indications and asking prices for a random subset of 250 listings sorted by indication.

Table 22. One-at-a-time contributions to the uncertainty of the indication (median absolute change, B4)

Factor (one at a time)	Median $ \Delta V $ (%)
unit cost band	5.4
economic life 50	2.3
soft+EI high	7.8
land not indexed	5.7
regional 1.10	6.4
grade +1	8.8
condition +1	6.4

9.4. Missing-not-at-random photographs: a bounding simulation

Section 7.1 documented the omissions sellers actually make. To bound the bias that *additional* strategic omission would cause, we simulate sellers who drop kitchen, bathroom and basement photographs with a probability falling from 0.8 for rooms the extractor rates in the worst condition to 0.05 for rooms rated like new, re-aggregate the remaining photographs and recompute the B2 indication (40 draws). The simulation removes 5% of photographs, lowers the share of listings with a composite grade from 91% to 91%, shifts the composite condition by +0.01 points on average, moves the grade up for 4.2% of listings and down for 1.3%, and changes the median indication by +0.0%. Because the rooms most likely to be hidden are those in the worst condition, omission flatters the remaining evidence; the effect on the indication is small in the median, because the condition factor acts on depreciation and the default grade absorbs missing kitchens, but it is one-directional,

and no re-weighting can recover information that was never photographed. The appropriate remedy is institutional—mandatory room coverage in listings, or an inspection—rather than statistical.

9.5. Failure cases

Table 23 classifies the disagreements between the image channel and the broker’s roofing field using the visual cues the extractor cited, and lists the most frequent cladding disagreements. Roof disagreements (52 of 1,045 compared exterior images) come mainly from low-slope roofs whose membrane cannot be seen from the street and from metal-versus-shingle look-alikes, with smaller shares attributable to distance and to partial visibility; cladding disagreements (24 of 213) are dominated by stone read where the broker lists brick or vinyl and by wood read where the broker lists vinyl or brick—material pairs that look alike in photographs—and by brokers listing a material the image shows only on a side elevation. These are the error modes the ontology anticipated, and they argue for two design rules in deployment: attributes should be reported with the scene they were read from, and a broker’s own material field should be preferred to the image when the two disagree and the image’s confidence is not high. Beyond attribute errors, three systematic failure cases emerged. Split-level and one-and-a-half-storey houses are counted differently by the roll, the broker and the image, and the schedule’s storey multiplier makes the choice consequential. Staged or virtually staged interiors, flagged by the extractors in a minority of listings, inflate finish grades and condition; the share is small but not negligible in the upper price quartile. And new construction advertised with renderings or model-home photographs is graded on an idealised building; the roll’s year built catches most such cases, but the image channel does not know that it is looking at a drawing.

Table 23. Taxonomy of image–broker disagreements on roof covering and cladding (mid-size extractor)

Attribute	Category of disagreement	n	Share (%)
roof covering	roof not fully visible / angle	4	8
roof covering	snow or weather	1	2
roof covering	flat / low-slope confusion	13	25
roof covering	metal vs shingle look-alike	14	27
roof covering	distance / resolution	6	12
roof covering	other	14	27
cladding	stone (image) vs brick/vinyl (broker)	5	21
cladding	wood (image) vs vinyl (broker)	4	17
cladding	wood (image) vs brick (broker)	3	12
cladding	vinyl (image) vs wood (broker)	2	8
cladding	wood (image) vs brick/stucco_or_acrylic (broker)	1	4
cladding	wood (image) vs stone (broker)	1	4
cladding	aluminum_or_steel (image) vs brick/vinyl (broker)	1	4
cladding	fiber_cement (image) vs stone (broker)	1	4

Note: roof categories are assigned from the cue words in the extractor’s one-line evidence statement, in the order listed; cladding rows give the most frequent (image, broker) pairs among disagreements.

9.6. Review load under confidence-based triage

A deployment that routes uncertain listings to a human must decide what “uncertain” means. Table 24 shows the share of listings flagged by rules that combine a mean-confidence threshold with the absence of a finish grade or the presence of a visible defect. With the mid-size extractor, a threshold of 0.7 flags 40% of listings and 0.8 flags 62%; the rules are driven by the defect and missing-grade components rather than by confidence, which barely varies. Since no human review was performed, the table quantifies the cost of review rules, not their benefit; the annotation protocol in Appendix D is designed to measure the benefit.

Table 24. Share of listings flagged for review under confidence-based rules

Extractor	Rule	Flagged (%)	No grade (%)	Defects (%)
haiku	conf<0.7 or no grade or defects	30.1	11.2	21.8
haiku	conf<0.75 or no grade or defects	30.9	11.2	21.8
haiku	conf<0.8 or no grade or defects	49.2	11.2	21.8
haiku	conf<0.85 or no grade or defects	99.7	11.2	21.8
sonnet	conf<0.7 or no grade or defects	40.3	11.0	32.8
sonnet	conf<0.75 or no grade or defects	43.2	11.0	32.8
sonnet	conf<0.8 or no grade or defects	61.6	11.0	32.8
sonnet	conf<0.85 or no grade or defects	96.6	11.0	32.8

9.7. Replication across extractor tiers

Every cost result above was recomputed with the small extractor’s attributes (Table 25). The level progression from B0 to B3 and the alignment statistics are close across tiers, which says that the cost indication is not sensitive to which of the two full-sample extractors specified the building, even though their judgment attributes agree only moderately: the schedule’s coarse grades and the condition factor absorb much of the disagreement. The largest extractor was run on a 200-listing subsample for the cross-model table and is not used for cost results.

Table 25. Cost indications by extractor tier and information set (anchor B)

Extractor	Level	Median V (\$)	Median D	Price ratio	Price MdAPE	Within 20%	Assessor ratio	Assessor MdAPE
haiku	B0	407,545	0.62	0.75	30.1	34	0.68	43.9
haiku	B1	474,163	0.49	0.87	26.6	38	0.94	43.7
haiku	B2	506,170	0.39	0.93	27.2	39	1.06	42.9
haiku	B3	525,689	0.39	0.96	27.7	36	1.14	46.7
haiku	B4	539,473	0.39	0.99	28.3	36	1.19	48.1
sonnet	B0	407,545	0.62	0.75	30.1	34	0.68	43.9
sonnet	B1	464,934	0.48	0.86	26.2	40	0.94	43.6
sonnet	B2	492,981	0.38	0.91	26.3	38	1.03	44.8
sonnet	B3	514,919	0.38	0.95	27.9	37	1.12	48.2
sonnet	B4	523,731	0.38	0.98	28.0	37	1.16	49.5

10. Discussion

10.1. What the results mean for the cost approach

The cost approach has always been limited by the cost of specifying the building. The results say that listing media relax that constraint for some inputs and not for others, and that the pattern is set by sellers rather than by models. Exterior materials, roof covering, garage and building type can be read from a façade photograph with agreement against broker fields in the nineties; kitchens are usually shown and their counters and cabinets are read with high cross-extractor agreement even if their *grade* is a judgment on which extractors differ; bathrooms are shown in one listing in five and basements in one in thirty, so that two of the three largest interior cost items must come from the broker’s text. For a mass-appraisal office this means that the marginal information in photographs, relative to the broker fields it already has, is concentrated in condition and renovation evidence—the inputs to effective age—and in the finish grade of the principal rooms. That is where the cost indication responded: through depreciation, whose median falls from 0.62 to 0.38 between the roll-only and the multimodal baselines, with a modest gain in alignment with asking prices and a larger gain in level.

The second lesson is about the schedule rather than the images. The level of the indication depends more on the published unit cost than on anything extracted, and the two published anchors for Québec in 2026 differ by a third. A calibration against assessor values on five metropolitan areas points, under either anchor, to a standard hard cost near \$272–\$306 per square foot for a two-storey house and transfers to the Québec City hold-out; the Montréal hard-cost band understates what assessors and the market imply, and the Québec industry band is approximately right. The age gradient of the alignment ratios adds that a 60-year straight-line life depreciates old housing faster than either assessors or buyers do, in line with the net rates of one to two percent a year that the depreciation literature has long reported; a cost approach from listing media would do better with a slower, condition-driven schedule, and the condition information the images provide is exactly what such a schedule needs.

10.2. Transparency, human oversight and professional practice

The pipeline is interpretable in the plain sense that every number in the indication can be traced to a visible attribute, a declared rule or a sourced parameter; nothing is learned from prices. That is a property appraisers and assessors can audit, and it is the reason the paper insisted on a cost identity rather than a price model: a hedonic model with image embeddings would align better with prices and tell a reviewer nothing about why. The price of transparency is that the schedule's errors are the method's errors—the level and age-slope problems above are visible precisely because nothing was fitted to hide them—and that the indication is twice as dispersed around asking prices as a hedonic model trained on half a million sales. We see three uses in practice. As a *screening* tool, the indication and its interval flag listings whose price is far from replacement cost plus land, or whose photographs contradict the broker's claims, for a reviewer's attention. As a *specification* tool, the extracted ontology—with its evidence class and status flags—is a pre-filled inspection form that an appraiser verifies on site rather than fills from scratch. As a *reconciliation* input, a cost indication whose assumptions are explicit is what valuation standards ask the appraiser to weigh against the comparison and income approaches; the paper supplies the indication and its uncertainty, and leaves the weighing to the professional.

Human oversight is not optional in this design, and the results say where it should go. Confidence-based triage flags many listings without targeting the uncertain ones, because stated confidence is compressed and uninformative for judgment attributes. Review should instead be triggered by the status flags—text and image contradicting each other on basement, garage or storeys—by the absence of interior evidence, by visible defects, and by the age bands where the schedule is least reliable. The expert annotation protocol in Appendix D is designed to measure both the reliability of the judgment attributes and the gain from such targeted review.

10.3. Scalability and cost

At batch prices the three extractor tiers processed 1,200 listings and 14,245 photographs for about US\$99; the cheapest tier alone, which replicates the cost results (Section 9.7), costs a few dollars per thousand listings. Processing the 7,745-listing frame, or the roughly 40,000 house listings active on the companion platform at any time, is therefore inexpensive; what does not scale is the evidence. Eight photographs capture what a listing will reveal, and no budget reveals the basement. The binding constraints on a deployment are rights and representativeness rather than compute: one large franchise network’s media host disallows automated retrieval and was excluded, removing two fifths of the frame and its more urban, higher-land-value listings; terms of use vary by host; and listings are a selected sample of the stock. A production system would need licensed access to media and a sampling design that corrects for both selections.

10.4. Relation to the literature

Relative to the image-based valuation literature, the paper changes the target from price to specification and the model from a learned predictor to an audited chain, and it adds a measurement of what sellers photograph that the price literature has noted ([Almaas et al., 2026](#)) but not quantified by room and position. Relative to the depreciation literature, it shows that a condition score read from photographs moves effective age in the direction and with the age-dependence that theory predicts, and that its scale is extractor-dependent, which is a measurement-error statement of the kind [Schennach \(2016\)](#) would recognise. Relative to the construction-cost literature, it supplies the missing first step—obtaining the specification of an existing house from public evidence—and shows that the second step is still dominated by the level of the published schedule. And relative to the vision–language literature, it finds the well-known pattern: factual, visible attributes are reliable and abstention works, while judgment attributes are reliable only up to a scale that differs across models.

11. Limitations, ethics and future research

11.1. What the data cannot establish

No human expert labels. The study has no annotations by appraisers or building technologists. Agreement is measured against broker fields, which are claims, and against the roll, which records few of the attributes of interest; judgment attributes are evaluated by cross-extractor agreement only, which bounds their reliability from below but says nothing about their validity. Everything the paper says about finish grade and condition should be read as “what three vision–language models

report”, not as “what the house is”. The protocol for the expert study that should follow is specified in Appendix D and pre-registered with the code.

Latent attributes. Foundation integrity, hidden moisture or mould, structural soundness, permit status, code compliance, tenure and servitudes are not inferable from listing media and were never inferred. The Monte Carlo intervals are conditional on their absence, and their under-coverage of both prices and assessor values is partly the measure of that conditioning. A house that looks like new and is sitting on a failing foundation receives the same indication as its sound neighbour.

No cross-city cost levels. The Statistics Canada index gives temporal change by metropolitan area but not level differences; the two published anchors describe Montréal and “Québec” and were issued by an industry software vendor’s secondary reporting and by builder-facing guides, not by a cost manual. The regional level is therefore a sensitivity parameter, and the Montréal alignment ratios, where land is a larger share of value, are the least reliable.

Assessor values are not truth. The roll’s building value is the assessor’s dated, cost-based indication reconciled with market evidence, equilibrated and subject to its own error; indexing it with a construction price index to the valuation date is an approximation that the roll-vintage split shows to be imperfect. Alignment with it is a benchmark comparison, not an accuracy measure.

Asking prices are not transactions. Market alignment is measured against asking prices for the main sample and against recorded sales for a small, self-selected set of quickly sold units. Asking prices exceeded recorded sales by a median of two to three percent in that set, and the cost indication was closer to the sale than to the ask; the main results would shift accordingly but the small sample does not allow more than that statement.

Representativeness. The frame is the set of house listings active on 67 connectors over ten weeks, matched within 15 m to a roll unit, with at least four photographs, excluding two classified-advertisement platforms and one franchise network whose media host disallows retrieval. The excluded network’s listings are more urban and have higher land values; listings are not the housing stock; and the stratified sample over-represents the smaller metropolitan areas by design. Design weights are stored and the paper makes no claim about Québec housing as a whole.

One provider family. The three extractors are tiers of one provider’s model family, chosen for comparability of prompts and schemas. Cross-extractor agreement therefore understates the disagreement that models of different families would show, and the results do not establish that any particular model is required.

Production component not inspected. The Vrai Prix production “cost estimate” used as a context benchmark has a generating code we could not inspect during the study; it is reported as an observed output of unknown formula.

11.2. Ethics and rights

The study handled copyrighted media by reference only, excluded hosts that disallow automated retrieval, masked all value information before any model saw a text, reproduced no photograph, and reports only aggregates. Residual uncertainties remain about the terms of use of individual broker sites and about the status of broker descriptions as copyrighted works; the data-rights document released with the code records them. Two further ethical questions concern deployment rather than research. A cost indication from listing media could be used to pressure sellers or to screen neighbourhoods; because the indication is explicit about its inputs, misuse is at least auditable, but the fairness analysis the paper could perform is limited to subgroups by price, age and region, not by the characteristics of residents, which the data do not and should not contain. And an extraction schema that asked about occupants or belongings would be a privacy violation; ours asks only about the building, and the one free-text field is instructed to name building cues only.

11.3. Future research

Four studies follow directly. First, the expert annotation study of Appendix D, which would convert cross-extractor agreement into validity statements and measure the gain from targeted review. Second, a depreciation schedule estimated, not assumed—condition-driven, slower than straight-line, and calibrated against assessor values and transactions jointly with an explicit model of their errors—which the age gradient of Section 8.3 shows to be the largest remaining lever. Third, access to the detailed cost schedules of the provincial manual, or to a licensed cost manual, which would replace the two industry anchors with a professional one and allow the regional level to be measured rather than varied. Fourth, a study of photograph selection as an economic choice: the room-by-position evidence of Section 7.1 is the raw material for a model of what sellers choose to show, which would turn the missing-not-at-random structure from a nuisance into information.

12. Conclusion

This paper asked whether the photographs and text that sellers publish can supply the specification of the improvements that the cost approach to value requires, and what a transparent cost indication built on that specification is worth. It formalised an extraction-to-valuation chain with five evidence classes, deterministic aggregation and reconciliation rules, a fully sourced schedule and propagated

uncertainty, and it ran the chain on 1,200 Québec house listings matched to their assessment-roll unit, with 14,245 photographs read by three vision–language model tiers and every parameter either published or declared.

The answer has three parts. Visible exterior facts are read reliably and abstention works: roof covering, garage type, cladding and building type agree with broker fields in the mid-eighties to mid-nineties with coverage above 85%, and disagreements concentrate on material look-alikes and roofs that cannot be seen from the street; storeys agree with the roll only loosely because the roll, the broker and the image count half-storeys differently. The binding constraint is what sellers show: in the first ten photographs, basements appear in one listing in thirty and bathrooms in one in five, so the largest interior cost items come from the broker’s text, while condition and renovation evidence—the inputs to effective age—are what the images uniquely add. And the cost indication responds to that evidence mostly through depreciation: between a roll-only baseline and the multimodal baseline the median depreciation falls from 0.62 to 0.38, the ratio of the indication to the asking price moves from 0.75 to 0.95 and its median absolute deviation from 30.1% to 27.9%—a real, replicated, but modest gain, next to a one-third difference between two published unit-cost anchors and an age gradient that says the assumed depreciation schedule is too steep for old houses and too lenient for new ones. A calibration against assessor values on five metropolitan areas transfers to a geographic hold-out and implies a standard hard cost near \$300 per square foot, consistent with 2026 Québec industry guides.

The paper is deliberately not an image-enhanced price model. No price enters the indication; the photographs specify the building, the schedule prices it, and the gap to the market is reported and interpreted rather than minimised. What the data cannot establish is stated as plainly as what they can: there are no expert labels, hidden defects are unobservable, cross-city cost levels are not published, assessor values are judgments, asking prices are not transactions, and the frame excludes a large network for rights reasons. The ontology, prompts, schedule, code, pre-registered plan and annotation protocol are released so that the next study—with experts in the loop and a depreciation schedule estimated rather than assumed—can start where this one stops.

References

- Ahmed, Eman H. and Mohamed Moustafa**, “House Price Estimation from Visual and Textual Features,” in “Proceedings of the 8th International Joint Conference on Computational Intelligence” SCITEPRESS - Science and Technology Publications 2016, pp. 62–68.
- Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan**, “Flamingo: A Visual Language Model for Few-Shot Learning,” in “Advances in Neural Information Processing Systems 35” Neural Information Processing Systems Foundation, Inc. (NeurIPS) 2022, pp. 23716–23736.
- Almaas, Vebjørn, Are Oust, and Arne Johan Pollestad**, “Giving Eyes to Automated Valuation Models: Computer Vision Assessment of Interior Home Condition,” *The Journal of Real Estate Finance and Economics*, 2026.
- Angelopoulos, Anastasios N. and Stephen Bates**, “Conformal Prediction: A Gentle Introduction,” *Foundations and Trends® in Machine Learning*, 2023, 16 (4), 494–591.
- Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh**, “VQA: Visual Question Answering,” in “2015 IEEE International Conference on Computer Vision (ICCV)” IEEE 2015, pp. 2425–2433.
- Artstein, Ron and Massimo Poesio**, “Inter-Coder Agreement for Computational Linguistics,” *Computational Linguistics*, 2008, 34 (4), 555–596.
- Baur, Katharina, Markus Rosenfelder, and Bernhard Lutz**, “Automated real estate valuation with machine learning models using property descriptions,” *Expert Systems with Applications*, 2023, 213, 119147.
- Bency, Archith J., Swati Rallapalli, Raghu K. Ganti, Mudhakar Srivatsa, and B. S. Manjunath**, “Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery,” in “2017 IEEE Winter Conference on Applications of Computer Vision (WACV)” IEEE 2017, pp. 320–329.
- Benfield, Justin D., Christopher L. Cain, and Ken H. Johnson**, “On the Relationship Between Property Price, Time-on-Market, and Photo Depictions in a Multiple Listing Service,” *The Journal of Real Estate Finance and Economics*, 2011, 43 (3), 401–422.

- Bogin, Alexander N. and William M. Doerner**, “Property Renovations and Their Impact on House Price Index Construction,” *Journal of Real Estate Research*, 2019, 41 (2), 249–284.
- Bokhari, Sheharyar and David Geltner**, “Characteristics of Depreciation in Commercial and Multifamily Property: An Investment Perspective,” *Real Estate Economics*, 2018, 46 (4), 745–782.
- Bound, John, Charles Brown, and Nancy Mathiowetz**, “Measurement Error in Survey Data,” in “Handbook of Econometrics,” Elsevier, 2001, pp. 3705–3843.
- Bourassa, Steven, Eva Cantoni, and Martin Hoesli**, “Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods,” *Journal of Real Estate Research*, 2010, 32 (2), 139–160.
- Case, Bradford, John Clapp, Robin Dubin, and Mauricio Rodriguez**, “Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models,” *The Journal of Real Estate Finance and Economics*, 2004, 29 (2), 167–191.
- Chau, K. W., S. K. Wong, and C. Y. Yiu**, “Adjusting for Non-Linear Age Effects in the Repeat Sales Index,” *The Journal of Real Estate Finance and Economics*, 2005, 31 (2), 137–153.
- Clapp, John M. and Carmelo Giaccotto**, “Residential Hedonic Models: A Rational Expectations Approach to Age Effects,” *Journal of Urban Economics*, 1998, 44 (3), 415–437.
- Cohen, Jacob**, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, 1960, 20 (1), 37–46.
- , “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit,” *Psychological Bulletin*, 1968, 70 (4), 213–220.
- Coulson, N. Edward and Daniel P. McMillen**, “Estimating time, age and vintage effects in housing prices,” *Journal of Housing Economics*, 2008, 17 (2), 138–151.
- Crosby, Neil**, “Valuation accuracy, variation and bias in the context of standards and expectations,” *Journal of Property Investment & Finance*, 2000, 18 (2), 130–161.
- Dagdelen, John, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain**, “Structured information extraction from scientific text with large language models,” *Nature Communications*, 2024, 15 (1).
- Davis, Morris A. and Jonathan Heathcote**, “The price and quantity of residential land in the United States,” *Journal of Monetary Economics*, 2007, 54 (8), 2595–2620.

- **and Michael G. Palumbo**, “The price of residential land in large US cities,” *Journal of Urban Economics*, 2008, 63 (1), 352–384.
- Despotovic, Miroslav, David Koch, Sascha Leiber, Mario Döller, Muntaha Sakeena, and Matthias Zeppelzauer**, “Prediction and analysis of heating energy demand for detached houses by computer vision,” *Energy and Buildings*, 2019, 193, 29–35.
- Diewert, Erwin and Chihiro Shimizu**, “Alternative Approaches to Commercial Property Price Indexes for Tokyo,” *Review of Income and Wealth*, 2017, 63 (3), 492–519.
- Dubé, Jean, Diègo Legros, 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.
- Elfaki, Abdelrahman Osman, Saleh Alatawi, and Eyad Abushandi**, “Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey,” *Advances in Civil Engineering*, 2014, 2014, 1–11.
- Francke, Marc K. and Alex M. van de Minne**, “Land, Structure and Depreciation,” *Real Estate Economics*, 2017, 45 (2), 415–451.
- Gal, Yarin and Zoubin Ghahramani**, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” 2015. arXiv preprint 1506.02142.
- Geerts, Margot, Manon Reusens, Bart Baesens, Seppe vanden Broucke, and Jochen De Weerd**, “On the Performance of LLMs for Real Estate Appraisal,” in “Lecture Notes in Computer Science,” Springer Nature Switzerland, 2026, pp. 195–211.
- Glaeser, Edward L, Joseph Gyourko, and Raven E Saks**, “Why Have Housing Prices Gone Up?,” *American Economic Review*, 2005, 95 (2), 329–333.
- Glaeser, Edward, Michael Scott Kincaid, and Nikhil Naik**, “Computer Vision and Real Estate: Do Looks Matter and Do Incentives Determine Looks,” NBER Working Paper 25174, National Bureau of Economic Research 2018.
- Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger**, “On Calibration of Modern Neural Networks,” 2017. arXiv preprint 1706.04599.
- Gyourko, Joseph and Albert Saiz**, “CONSTRUCTION COSTS AND THE SUPPLY OF HOUSING STRUCTURE*,” *Journal of Regional Science*, 2006, 46 (4), 661–680.

- Günaydın, H Murat and S Zeynep Doğan**, “A neural network approach for early cost estimation of structural systems of buildings,” *International Journal of Project Management*, 2004, 22 (7), 595–602.
- Hallgren, Kevin A.**, “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial,” *Tutorials in Quantitative Methods for Psychology*, 2012, 8 (1), 23–34.
- Harding, John P., Stuart S. Rosenthal, and C.F. Sirmans**, “Depreciation of housing capital, maintenance, and house price inflation: Estimates from a repeat sales model,” *Journal of Urban Economics*, 2007, 61 (2), 193–217.
- Hausman, Jerry**, “Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left,” *Journal of Economic Perspectives*, 2001, 15 (4), 57–67.
- Hayes, Andrew F. and Klaus Krippendorff**, “Answering the Call for a Standard Reliability Measure for Coding Data,” *Communication Methods and Measures*, 2007, 1 (1), 77–89.
- Heckman, James J.**, “Sample Selection Bias as a Specification Error,” *Econometrica*, 1979, 47 (1), 153.
- Hill, Robert J.**, “HEDONIC PRICE INDEXES FOR RESIDENTIAL HOUSING: A SURVEY, EVALUATION AND TAXONOMY,” *Journal of Economic Surveys*, 2013, 27 (5), 879–914.
- Howell, Junia and Elizabeth Korver-Glenn**, “The Increasing Effect of Neighborhood Racial Composition on Housing Values, 1980–2015,” *Social Problems*, 2021, 68 (4), 1051–1071.
- Hulten, Charles R. and Frank C. Wykoff**, “The estimation of economic depreciation using vintage asset prices,” *Journal of Econometrics*, 1981, 15 (3), 367–396.
- Hwang, Seokyon**, “Dynamic Regression Models for Prediction of Construction Costs,” *Journal of Construction Engineering and Management*, 2009, 135 (5), 360–367.
- Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung**, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, 2023, 55 (12), 1–38.
- Johnson, Erik B, Alan Tidwell, and Sriram V Villupuram**, “Valuing Curb Appeal,” *The Journal of Real Estate Finance and Economics*, 2020, 60 (1-2), 111–133.
- 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 preprint 2207.05221.
- Kang, Yuhao, Fan Zhang, Wenzhe Peng, Song Gao, Jinneng Rao, Fabio Duarte, and Carlo Ratti**, “Understanding house price appreciation using multi-source big geo-data and machine learning,” *Land Use Policy*, 2021, *111*, 104919.
- Kapoor, Sayash and Arvind Narayanan**, “Leakage and the reproducibility crisis in machine-learning-based science,” *Patterns*, 2023, *4* (9), 100804.
- Kermani, Amir and Francis Wong**, “Racial Disparities in Housing Returns,” NBER Working Paper 29306, National Bureau of Economic Research 2021.
- Kim, Gwang-Hee, Sung-Hoon An, and Kyung-In Kang**, “Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning,” *Building and Environment*, 2004, *39* (10), 1235–1242.
- Knight, John R. and C.F. Sirmans**, “Depreciation, Maintenance, and Housing Prices,” *Journal of Housing Economics*, 1996, *5* (4), 369–389.
- Koch, David, Miroslav Despotovic, Muntaha Sakeena, Mario Döller, and Matthias Zeppelzauer**, “Visual Estimation of Building Condition with Patch-level ConvNets,” in “Proceedings of the 2018 ACM Workshop on Multimedia for Real Estate Tech” ACM 2018, pp. 12–17.
- , —, **Simon Thaler, and Zedong Zhang**, “From platforms to price: the impact of condition prediction using computer vision on real estate pricing models,” *Journal of European Real Estate Research*, 2026, *19* (1), 116–139.
- Kok, Nils, Eija-Leena Koponen, and Carmen Adriana Martínez-Barbosa**, “Big Data in Real Estate? <i>From Manual Appraisal to Automated Valuation</i>,” *The Journal of Portfolio Management*, 2017, *43* (6), 202–211.
- Koo, Terry K. and Mae Y. Li**, “A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research,” *Journal of Chiropractic Medicine*, 2016, *15* (2), 155–163.
- Krause, A., A. Martin, and M. Fix**, “Uncertainty in automated valuation models: Error-based versus model-based approaches,” *Journal of Property Research*, 2020, *37* (4), 308–339.

- Kucklick, Jan-Peter and Oliver Müller**, “Tackling the Accuracy-Interpretability Trade-off: Interpretable Deep Learning Models for Satellite Image-based Real Estate Appraisal,” *ACM Transactions on Management Information Systems*, 2023, 14 (1), 1–24.
- Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell**, “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,” 2016. arXiv preprint 1612.01474.
- Landis, J. Richard and Gary G. Koch**, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, 1977, 33 (1), 159.
- Law, Stephen, Brooks Paige, and Chris Russell**, “Take a Look Around,” *ACM Transactions on Intelligent Systems and Technology*, 2019, 10 (5), 1–19.
- Levitt, Steven D. and Chad Syverson**, “Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions,” *Review of Economics and Statistics*, 2008, 90 (4), 599–611.
- Li, Yifan, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen**, “Evaluating Object Hallucination in Large Vision-Language Models,” in “Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing” Association for Computational Linguistics 2023, pp. 292–305.
- Lindenthal, Thies and Erik B. Johnson**, “Machine Learning, Architectural Styles and Property Values,” *The Journal of Real Estate Finance and Economics*, 2025, 71 (3), 353–384.
- Lipton, Zachary C.**, “The Mythos of Model Interpretability,” *Queue*, 2018, 16 (3), 31–57.
- Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee**, “Visual Instruction Tuning,” in “Advances in Neural Information Processing Systems 36” Neural Information Processing Systems Foundation, Inc. (NeurIPS) 2023, pp. 34892–34916.
- Malpezzi, Stephen**, “Hedonic Pricing Models: A Selective and Applied Review,” 2002.
- , **Larry Ozanne, and Thomas G. Thibodeau**, “Microeconomic Estimates of Housing Depreciation,” *Land Economics*, 1987, 63 (4), 372.
- 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.
- McHugh, Marry L.**, “Interrater reliability: the kappa statistic,” *Biochemia Medica*, 2012, pp. 276–282.

- McMillen, Daniel and Ruchi Singh**, “Assessment Regressivity and Property Taxation,” *The Journal of Real Estate Finance and Economics*, 2020, 60 (1-2), 155–169.
- Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan**, “A Survey on Bias and Fairness in Machine Learning,” *ACM Computing Surveys*, 2022, 54 (6), 1–35.
- Mosqueira-Rey, Eduardo, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal**, “Human-in-the-loop machine learning: a state of the art,” *Artificial Intelligence Review*, 2023, 56 (4), 3005–3054.
- Naik, Nikhil, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo**, “Streetscore – Predicting the Perceived Safety of One Million Streetscapes,” in “2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops” IEEE 2014, pp. 793–799.
- Nordstrom, Ardyn, Morgan Nordstrom, and Matthew D. Webb**, “Using Images as Covariates: Measuring Curb Appeal with Deep Learning,” 2024. arXiv preprint 2403.19915.
- Nouriani, Ali and Lance Lemke**, “Vision-based housing price estimation using interior, exterior & satellite images,” *Intelligent Systems with Applications*, 2022, 14, 200081.
- Nowak, Adam and Patrick Smith**, “Textual Analysis in Real Estate,” *Journal of Applied Econometrics*, 2017, 32 (4), 896–918.
- Oust, Are, Sjur Westgaard, Jens Erik Waage, and Nahome Kidane Yemane**, “Assessing the Explanatory Power of Dwelling Condition in Automated Valuation Models,” *Journal of Real Estate Research*, 2025, 47 (1), 1–27.
- Pace, R. Kelley and Darren Hayunga**, “Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests,” *The Journal of Real Estate Finance and Economics*, 2020, 60 (1-2), 170–180.
- PAGLIN, MORTON and MICHAEL FOGARTY**, “EQUITY AND THE PROPERTY TAX: A NEW CONCEPTUAL FOCUS,” *National Tax Journal*, 1972, 25 (4), 557–565.
- Poursaeed, Omid, Tomáš Matera, and Serge Belongie**, “Vision-based real estate price estimation,” *Machine Vision and Applications*, 2018, 29 (4), 667–676.
- Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever**, “Learning Transferable Visual Models From Natural Language Supervision,” 2021. arXiv preprint 2103.00020.

- Rosen, Sherwin**, “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,” *Journal of Political Economy*, 1974, 82 (1), 34–55.
- Rosiers, Francois Des, Antonio Lagana, and Marius Theriault**, “Size and proximity effects of primary schools on surrounding house values,” *Journal of Property Research*, 2001, 18 (2), 149–168.
- Rosiers, François Des, Marius Thériault, Yan Kestens, and Paul Villeneuve**, “Landscaping and House Values: An Empirical Investigation,” *Journal of Real Estate Research*, 2002, 23 (1-2), 139–162.
- RUBIN, DONALD B.**, “Inference and missing data,” *Biometrika*, 1976, 63 (3), 581–592.
- Rudin, Cynthia**, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, 2019, 1 (5), 206–215.
- Schennach, Susanne M.**, “Recent Advances in the Measurement Error Literature,” *Annual Review of Economics*, 2016, 8 (1), 341–377.
- Seiler, Michael, Poornima Madhavan, and Molly Liechty**, “Toward an Understanding of Real Estate Homebuyer Internet Search behavior: An Application of Ocular Tracking Technology,” *Journal of Real Estate Research*, 2012, 34 (2), 211–242.
- Shehbaz, Muhammad Assad and Carlos Francisco Moreno-García**, “Structured Data Extraction from Real Estate Documents using Clustering, Classification, and Large Language Models,” 2026. arXiv preprint 2607.06012.
- Shen, Lily and Stephen Ross**, “Information value of property description: A Machine learning approach,” *Journal of Urban Economics*, 2021, 121, 103299.
- Shilling, James D., C.F. Sirmans, and Jonathan F. Dombrow**, “Measuring depreciation in single-family rental and owner-occupied housing,” *Journal of Housing Economics*, 1991, 1 (4), 368–383.
- Snow, Rion, Brendan O’Connor, Daniel Jurafsky, and Andrew Y. Ng**, “Cheap and fast—but is it good?,” in “Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP ’08” Association for Computational Linguistics 2008, p. 254.
- Somerville, C. Tsurriel**, “Residential Construction Costs and the Supply of New Housing: Endogeneity and Bias in Construction Cost Indexes,” *The Journal of Real Estate Finance and Economics*, 1999, 18 (1), 43–62.

- Thériault, Marius, François Des Rosiers, Paul Villeneuve, and Yan Kestens**, “Modelling interactions of location with specific value of housing attributes,” *Property Management*, 2003, 21 (1), 25–62.
- Wilhelmsson, Mats**, “House price depreciation rates and level of maintenance,” *Journal of Housing Economics*, 2008, 17 (1), 88–101.
- Yoshida, Jiro**, “The economic depreciation of real estate: Cross-sectional variations and their return implications,” *Pacific-Basin Finance Journal*, 2020, 61, 101290.
- You, Quanzeng, Ran Pang, Liangliang Cao, and Jiebo Luo**, “Image-Based Appraisal of Real Estate Properties,” *IEEE Transactions on Multimedia*, 2017, 19 (12), 2751–2759.
- Zeppelzauer, Matthias, Miroslav Despotovic, Muntaha Sakeena, David Koch, and Mario Döller**, “Automatic Prediction of Building Age from Photographs,” in “Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval” ACM 2018, pp. 126–134.
- Zhao, Yu, Haoxiang Gao, Jinghan Cao, and Shiqi Yang**, “Utilizing Large Language Models for Information Extraction from Real Estate Transactions,” 2024. arXiv preprint 2404.18043.
- Zurada, Jozef, Alan Levitan, and Jian Guan**, “A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context,” *Journal of Real Estate Research*, 2011, 33 (3), 349–388.

Appendices

A. Proofs

Proof of Proposition 1. Write the improvements component as $I_i = \mu_i A_i \kappa(g_i)(1 - D_i)$ with μ_i collecting the storey, attachment, regional and temporal multipliers and the mark-ups, all known given the record, and D_i held fixed. Under quadratic loss the Bayes estimate of I_i without the image is $\mu_i A_i (1 - D_i) \mathbb{E}_p[\kappa(g)]$ with expected loss $\mu_i^2 A_i^2 (1 - D_i)^2 \text{Var}_p(\kappa(g))$. Observing $\hat{g}_i$ with likelihood $q_{g\hat{g}} = \Pr(\hat{g} | g)$ yields the posterior $p(g | \hat{g}) \propto p_g q_{g\hat{g}}$, Bayes estimate $\mu_i A_i (1 - D_i) \mathbb{E}[\kappa(g) | \hat{g}]$ and expected loss $\mu_i^2 A_i^2 (1 - D_i)^2 \mathbb{E}_{\hat{g}} \text{Var}(\kappa(g) | \hat{g})$. The reduction is the stated expression by the law of total variance. If the rows of Q are equal, $\hat{g}$ is independent of g , the posterior equals the prior and the reduction is zero. For the symmetric channel in which $\hat{g} = g$ with probability π and is otherwise uniform over the other $G - 1$ grades, the posterior given $\hat{g} = v$ puts mass proportional to $p_v \pi$ on

v and $p_g(1 - \pi)/(G - 1)$ on $g \neq v$; as π increases the posterior concentrates and $\mathbb{E}_{\hat{g}} \text{Var}(\kappa \mid \hat{g})$ decreases monotonically to zero at $\pi = 1$, while for fixed π the reduction scales with $\text{Var}_p(\kappa(g))$, the between-grade dispersion of the schedule. The factor $(1 - D_i)^2$ is explicit. $\square$

Proof of Proposition 2. Before the cap and the renovation allowance, $D^{\text{phys}} = \text{age } \lambda(c)/\text{EL}$. If λ is affine on the relevant range with slope $\lambda' < 0$, then $\lambda(\hat{c}) = \lambda(c) + \lambda'\varepsilon$ and $\mathbb{E}[D^{\text{phys}}(\hat{c}) - D^{\text{phys}}(c) \mid c] = \text{age } \lambda'\delta(c)/\text{EL}$. A uniform $\delta > 0$ gives a negative bias proportional to age and inversely proportional to EL. If $\delta(c) = \delta_0 + \delta_1(c - \bar{c})$ with $\delta_1 < 0$ and $\mathbb{E}[\delta(c)] = \delta_0$, the mean bias is $\text{age } \lambda'\delta_0/\text{EL}$ and the reported $\lambda(\hat{c})$ has variance $(\lambda')^2(1 + \delta_1)^2 \text{Var}(c)$ plus noise, smaller than $(\lambda')^2 \text{Var}(c)$ when $-2 < \delta_1 < 0$: the cross-sectional dispersion of depreciation is compressed while its mean is unchanged when $\delta_0 = 0$. The cap and the renovation allowance only attenuate these effects. $\square$

B. Prompts and schemas

The system prompt for the per-image extraction (version wp20-prompts-1.0) reads:

You are assisting a real-estate appraiser who applies the cost approach. You see ONE photograph from a Québec residential for-sale listing. Report only what is visually observable in this photograph. Classify the scene, then fill every field: use “not_visible” when the feature is relevant to the scene but cannot be seen, and “not_applicable” when the field does not apply to this scene (e.g. kitchen fields on an exterior photo). Rules: (1) never infer hidden conditions (foundation, structure, mold behind finishes, permits, exact year or area); (2) condition scales describe wear and deterioration, finish-quality scales describe the cost level of the materials, do not confuse taste with quality; (3) a renovated or staged room is still rated on what is visible; (4) if the image is a floor plan, a document, a logo or unusable, say so and set other fields to “not_applicable”. Output JSON only, matching the schema.

The image schema (version wp20-schema-1.0) has 22 required fields: scene (14 classes), photo_quality (1–5), staging_or_rendering, building_type_visual, storeys_visible, cladding (list of up to three among brick, stone, vinyl, wood, aluminum or steel, stucco or acrylic, fiber cement, other), roof_form, roof_covering (asphalt shingle, metal, flat membrane, other), garage_visible (attached, detached, carport, none visible), exterior_condition (0–5), kitchen_grade and bathroom_grade (basic, standard, upgraded, luxury), kitchen_counters, kitchen_cabinets, flooring_main, basement_finish, interior_condition and interior_finish_quality (0–5), renovation_evidence, visible_defects (list), confidence (0–1) and a one-line evidence statement. Every categorical field admits not_visible and not_applicable. The text schema has 22 fields covering building type, storeys,

materials, garage, basement, grades, condition and renovation claims with items and years, year built and area claims, heating, pool, disclosed defects, a flag for any surviving value mention, and a confidence; every categorical field admits `not_stated`. Both schemas are released in machine-readable form with the code, together with the text system prompt.

C. Aggregation and reconciliation rules

Exterior attributes use exterior scenes only; categorical attributes take the confidence-weighted plurality among informative values with a confidence floor of 0.05; garage presence is positive if any exterior image reports a garage, carport or detached garage, “none visible” only if no image reports one and at least one reports none; ordinal grades take the weighted median of ordinal codes; conditions and finish quality take the confidence-weighted mean of non-zero scores; renovation evidence takes the maximum of (none, partial, extensive) with “unsure” ignored; defects are the union excluding “none visible”; the composite condition is the mean of exterior and interior condition when both exist; the composite grade is the rounded weighted average of kitchen grade (0.5), bathroom grade (0.3) and interior finish quality rescaled to 1–4 (0.2), each entering only when observed. Reconciliation gives precedence to the roll (storeys), then to informative visual values, then to informative text values, then to the B0 default; the status of each reconciled attribute (agree, contradict, visual only, text only, unknown) is stored.

D. Proposed expert annotation protocol

Three independent raters (two chartered appraisers and one building technologist, none involved in the schema design) code 300 listings drawn from the extraction sample with the same stratification, in two sessions—images only, then images and masked text—using a form with the property-level attributes of the ontology, a `not_visible` option, a three-level confidence and the indices of the images used as evidence. Condition scales are anchored with written descriptors and reference photographs assembled from licensed sources. Reliability is reported as Krippendorff’s nominal α and pairwise κ for nominal attributes, quadratic-weighted κ and ICC(2,1) for ordinal grades and conditions, with 0.67 as the minimum for reporting and 0.80 for use as reference labels. Disagreements are adjudicated by majority, three-way disagreements by a fourth senior appraiser, with a log. Each extractor tier is then treated as a fourth rater and its disagreements with the adjudicated labels are classified by error type (occlusion, staging, lighting, material look-alike, scene misclassification, hallucinated attribute). Time per listing is recorded for the review-cost analysis. Rater identities are kept outside the released data.

E. Institutional and statistical sources

- *Loi sur la fiscalité municipale*, RLRQ c. F-2.1, sections 14, 43–46, 263 (legisquebec.gouv.qc.ca, read 2026-10-20).
- *Manuel d'évaluation foncière du Québec*, structure document (Ministère des Affaires municipales et de l'Habitation, cdn-contenu.quebec.ca) and the ministry's page on *facteurs de rajustement* (quebec.ca, 2026-10-20). Part 3E was not consulted.
- Données Québec, « Rôle d'évaluation foncière du Québec », MAMH, licence CC-BY 4.0, annual (CKAN metadata, 2026-10-20).
- Statistics Canada, Table 18-10-0289-01, *Building construction price indexes, by type of building and division*, 2023=100, quarterly, Québec CMA (vector 1617911172), Montréal CMA (1617911556), Ottawa–Gatineau Ontario part (1617912324), 15-CMA composite (1617908100), single-detached house, division composite, latest 2026-Q2 (Web Data Service, 2026-10-20).
- Altus Group, *2026 Canadian Cost Guide*, released 2026-04-23 (press release); residential figures as tabulated by NerdWallet Canada ("How much does it cost to build a house", updated 2026; accessed 2026-10-20): Montréal \$155,000–\$215,000 per 1,000 sq ft, single-family with unfinished basement, land excluded.
- Soumission Rénovation, « Coût de construction d'une maison au Québec en 2026 », modified 2026-08-31; Dessins Drummond, « Combien coûte construire une maison au Québec en 2026? », updated 2026-09-24 (both accessed 2026-10-20).
- IAAO, *Standard on Mass Appraisal of Real Property*, revised 2025; Appraisal Institute of Canada, *CUSPAP 2024*; IVSC, *International Valuation Standards* effective 2025-01-31 (edition facts verified by web search 2026-10-20; texts not consulted in full).

F. Reproducibility

The repository released with the paper contains the configuration files (data rules, sample design, ontology, prompts, models, cost schedule), the ingestion, extraction, aggregation, cost-model, evaluation and visualization modules, the numbered scripts that run the pipeline, unit tests, the raw model responses (as JSON lines without images), every table and figure as generated, a machine-readable manifest with the hash of every artefact, the research log, the pre-registered experiment plan, the data-rights document and the literature matrix with the verification record of each reference. The two databases are read-only copies of the Vrai Prix production data and are not redistributed; the roll is public on Données Québec and the listing data can be requested from the author under the conditions described in the data-rights document. All numbers in the text are macros generated from the result files; the manuscript compiles with `latexmk`. The extraction runs can be reproduced

with the provider's batch interface and an API key, at a cost of about US\$99 at the prices in force in October 2026; model outputs will differ across model versions, which is one reason the raw responses are released.