



UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS
DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 14

Hedonic House Pricing with Language-Model and Vision Features

Evidence from Quebec Real-Estate Listings

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 2.1

Abstract

Hedonic house-price models are bounded by the characteristics an analyst can measure, yet much of a home's value is recorded only in the free-text description and photographs of its listing. We treat general-purpose AI models as a feature-engineering technology for hedonic analysis: a large language model converts each description into structured property specifications, and a vision-language model converts the façade photograph into visual attributes, both under strict JSON schemas that make the outputs auditable econometric variables. Using 19,123 Quebec house listings, the two AI layers raise the R^2 of an already rich log-linear hedonic specification with municipality fixed effects from 0.717 to 0.755, 3.0 of the 3.8 points coming from the façade photograph, and they raise the out-of-sample R^2 of every estimator from OLS to gradient boosting and neural networks (best ensemble: R^2 of 0.779, median absolute error 14.0%). The vision layer adds more to the linear model than switching from OLS to a random forest does. The extracted attributes carry economically sensible implicit prices—2.3% per point of curb appeal, 3–4% for brick or stone façades—that survive municipality-clustered inference, a cluster bootstrap, postal-code fixed effects, and a control for the language of the description, which itself carries a 7% conditional premium for English listings. We also document that for-sale-by-owner descriptions embed the asking price, so that the text layer's gain is credible only on the brokered platform, and that spatially grouped cross-validation lowers all accuracy levels by three to five points while preserving the AI increments.

Keywords: hedonic pricing, house prices, large language models, computer vision, text as data, automated valuation, Quebec

JEL Classification: R31, C21, C55, R21

1. Introduction

The hedonic price framework decomposes the market value of a heterogeneous good into the implicit prices of its characteristics (Lancaster, 1966; Rosen, 1974). In housing markets it remains the workhorse tool for appraisal, index construction, property-tax assessment, and policy analysis. Yet the empirical performance of hedonic models is bounded by the richness of the characteristics an analyst can measure. A large share of what determines a home’s value—its state of repair, the quality of its finishes, its curb appeal, the character of its street—is recorded not in structured fields but in the free-text description and the photographs that accompany each listing. These modalities have long been discarded because they were costly to encode at scale, leaving hedonic estimates exposed to omitted-quality bias and hedonic predictions short of their attainable accuracy.

This paper treats general-purpose artificial-intelligence models as a *feature-engineering technology* for hedonic analysis and quantifies the explanatory power that technology adds. We assemble a cross-section of 19,123 single-family houses and townhouses listed for sale in Quebec on 3 July 2026, pooling the brokerage-listed MLS inventory from Realtor.ca with the for-sale-by-owner inventory from DuProprio. On top of standard structural, locational, and amenity attributes, we build two machine-generated feature layers. First, a large language model (LLM) reads each free-text description and returns a fixed set of property specifications—state of repair, renovation of the kitchen and bathroom, flooring and countertop materials, heating system, view, and income potential—under a strict JSON schema with closed enumerations. Second, a vision-language model inspects the façade photograph of each property and returns visual attributes: overall condition, curb appeal, renovation level, exterior materials, landscaping, setting, and a luxury indicator. Because both extractions are schema-constrained, their outputs enter the hedonic regression as ordinary tabulated regressors with interpretable coefficients rather than as latent scores.

Our contribution is threefold. First, we show that AI-based feature engineering delivers gains in explanatory power over an already rich hedonic specification. In our preferred specification—a log-linear model with 195 municipality fixed effects, property-type fixed effects, and fifteen keyword amenities, room-level and listing-quality controls—the two machine-generated layers raise in-sample R^2 from 0.717 to 0.755, and they raise the out-of-sample R^2 of every estimator we consider, from ordinary least squares to gradient boosting and an embedding-based neural network. Of the two layers, the façade-vision layer is by far the more informative: it adds 3.0 points of in-sample R^2 against 0.8 for the text layer, and 5.8 (2.7) points of out-of-sample R^2 for OLS (gradient boosting), more than the gain from replacing OLS by a random forest. Second, we show that the extracted attributes carry interpretable implicit prices—a one-point increase in visual curb appeal is associated with a 2.3% higher asking price, a visible luxury cue with 5.2%, brick or stone façades with premia of 3–4% relative to the omitted materials, and a renovated

kitchen (read from the text) with 2.6%—and that these associations survive municipality-clustered inference, a cluster bootstrap, finer postal-code (RTA) fixed effects, and a control for the language of the description. Third, we document a complete and reproducible pipeline from raw scraped listings to an analysis-ready dataset, including rules-based outlier cleaning, unit harmonization that reconstructs living area from room-level dimensions, and the schema-constrained extraction protocol, with a total AI cost of about \$60 for the full sample—roughly a third of a cent per listing.

The revision reported here (version 2.1) also confronts three threats that a skeptical reader should raise against any “AI features” exercise, and we believe the paper is more useful for addressing them explicitly. The first is *information leakage*. Every DuProprio description in our data embeds a structured block that includes the asking price itself (“Asking Price \$599,000”), so the language model read the dependent variable for 19.7% of the sample. The schema contains no price field, but the text-layer gain is five times larger on DuProprio than on Realtor.ca, whose descriptions never contain the asking price, and we therefore report all gains separately by platform and treat the Realtor.ca estimates as the clean benchmark. The vision model never sees the price. The second is *language*. Forty-three percent of the descriptions are in English, and English-language listings are priced about 7% higher than French ones conditional on all other attributes and on location; the AI implicit prices are, however, essentially unchanged when language is controlled for, and the gains from both layers are present within the French and within the English subsamples. The third is *spatial dependence in validation*. Random cross-validation folds let a model learn from a neighbour of each held-out house; when folds are formed by forward sortation area instead, all estimators lose three to five points of out-of-sample R^2 , but the increments from the text and vision layers are preserved.

Our main findings can be summarized as follows. The preferred specification explains 75.5% of the variation in log asking prices when living area is observed and 72.5% on the full sample when it is not. The living-area elasticity is 0.36, each additional bathroom is associated with a 7.6% premium, and the age profile traces the familiar depreciation-then-heritage pattern. The municipality effects range from about +231% in Westmount to about −30% in remote municipalities relative to the pooled group of small municipalities, and order the province’s markets as expected. On the identical feature set, an equally weighted blend of a robust embedding-based neural network and gradient boosting attains an out-of-sample R^2 of 0.779 with a median absolute percentage error (MdAPE) of 14.0% under random folds. We interpret all coefficients as *predictive associations* rather than causal effects: listing content is chosen by sellers and their agents and correlates with unobserved determinants of value, and the dependent variable is an asking price, two issues we discuss at length in Section 7.

Beyond the Quebec application, the exercise speaks to a measurement problem in housing economics. Text- and image-derived quality measures have until now required bespoke supervised

models trained on labelled data (Nowak and Smith, 2017; Shen and Ross, 2021; Johnson et al., 2020; Lindenthal and Johnson, 2021). General-purpose foundation models with schema-constrained decoding lower the fixed cost of that measurement step to nearly zero, produce attributes that are human-readable rather than latent, and can be audited field by field. Our results suggest that this technology shifts the effective characteristics frontier of hedonic analysis by enough to matter for appraisal, automated valuation, and assessment applications—provided that the text fed to the model is screened for the outcome it is meant to predict.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the data, the cleaning rules, and the content of the descriptions. Section 4 presents the feature-engineering pipeline and the econometric specification. Section 5 reports the main results. Section 6 maps the predictive frontier across estimators. Section 7 reports the robustness analyses and discusses limitations. Section 8 concludes.

2. Related Literature

This paper connects four strands of research.

The first is the classical **hedonic pricing** literature. Building on Lancaster (1966), Rosen (1974) formalized the market for differentiated goods as a market for their underlying characteristics, providing the theoretical foundation for interpreting regression coefficients as implicit prices. The empirical housing literature that followed is vast; Sirmans et al. (2005) survey the characteristics most commonly included in hedonic house-price models and document substantial heterogeneity in specification, Malpezzi (2003) provides a critical review of functional-form and identification choices, and Hill (2013) surveys the use of hedonic methods for house-price measurement. A recurring theme is omitted-variable bias from unmeasured quality: the state of repair, the standard of finishes, and the visual quality of a dwelling are priced by the market but rarely recorded in the structured fields available to the econometrician. Nowak and Smith (2020) show that this omission is not innocuous even for aggregate statistics, since quality-adjusted repeat-sales indexes built with information from listing remarks differ materially from conventional ones. Our contribution speaks directly to this measurement problem. Our log-linear specification, with logged continuous attributes and a dependent variable in logs, follows the mainstream of this literature, and we read indicator coefficients with the Halvorsen and Palmquist (1980) transformation; the small-sample refinement of Kennedy (1981) is immaterial at our precision.

The second strand concerns **spatial structure, inference, and prediction**. Can (1992) and Anselin (1988) emphasize that unobserved locational amenities induce spatial dependence that, if ignored, biases coefficient estimates and understates uncertainty; we absorb the first-order component of this problem through municipality fixed effects, re-estimate with finer forward

sortation area (RTA) effects, and cluster inference at the level of the fixed effects (Cameron and Miller, 2015; Abadie et al., 2023), complementing the analytical standard errors with a cluster bootstrap because our clusters are of very unequal size (MacKinnon and Webb, 2017). The same dependence bears on validation. Bourassa et al. (2010) show that spatial structure that is not exploited by the model surfaces in prediction errors, Pace and Hayunga (2020) find that tree ensembles recover information left in the residuals of hedonic and spatial models, and Roberts et al. (2017) document that random cross-validation is optimistic whenever observations are spatially clustered, recommending blocked (spatially grouped) folds instead. We report both. On the prediction side, Bin (2004) compares parametric and semi-parametric hedonic predictors, Mullainathan and Spiess (2017) frame machine-learning methods as estimators of the conditional expectation $\mathbb{E}[P \mid \mathbf{z}]$ suited to prediction tasks, and Kok et al. (2017) document the rise of automated valuation models (AVMs) in industry. This literature generally finds that flexible learners beat linear hedonics on accuracy at some cost in interpretability; our estimator ladder (Section 6) reproduces that pattern on a common feature set, and our focus is on how much of the gap is closed not by a better *estimator* but by better *measurement*.

The third strand is the use of **unstructured listing content**—text and images—in real-estate valuation. Gentzkow et al. (2019) survey text-as-data methods in economics, and Dell (2025) and Korinek (2023) survey, respectively, deep learning and generative models as research tools for economists. In real estate specifically, Nowak and Smith (2017) show that words in MLS remarks predict transaction prices and reduce hedonic pricing errors; Nowak et al. (2021) show that the price-relevant vocabulary is not stable across markets or over time, a warning against transferring word lists; and Shen and Ross (2021) extract quality signals from property descriptions with supervised machine learning, finding that the text carries information about unobserved quality that is capitalized into prices. Zhang et al. (2024) benchmark text encoders for price prediction and find that textual and non-textual features are complementary, and a companion working paper on Quebec listings (Boucher, 2026) documents that semantic embeddings of descriptions carry out-of-sample price information and that the language of the description is a first-order confounder in a bilingual market—a finding we build on here. On the image side, Ahmed and Moustafa (2016) combine visual and textual features for price estimation; Bessinger and Jacobs (2016) and Poursaeed et al. (2018) show that learned measures of exterior and interior quality carry independent predictive power; Naik et al. (2017) and Glaeser et al. (2018) use street-level imagery to measure the physical appearance of neighbourhoods and houses and study how appearance relates to incentives and change; Law et al. (2019) and Kang et al. (2021) combine street-view and satellite imagery with structured attributes to predict prices and appreciation; Zhang and Dong (2018) price street-visible greenery from street-view images; Johnson et al. (2020) construct a curb-appeal index from Google Street View images and estimate its capitalization into sale prices; and Lindenthal and Johnson

(2021) classify architectural styles from photographs and document style premia. Relative to this work, our approach differs in the *technology of measurement*: rather than training bespoke supervised models on labelled data—which requires labels and task-specific engineering, and yields opaque scores—we use general-purpose large language and vision–language models with schema-constrained decoding to produce a fixed set of interpretable, human-readable attributes (“renovated kitchen,” “brick façade,” a 1–10 curb-appeal score) at a marginal cost of a fraction of a cent per listing, and we embed those attributes in a transparent hedonic regression whose coefficients can be read as implicit prices. The reliability of such zero-shot annotation is itself an active research question; [Gilardi et al. \(2023\)](#) find that general-purpose LLMs match or exceed crowd workers on standard text-annotation tasks, but the extraction error of any particular field must be assessed, which we do by confronting the machine-generated attributes with independent sources (Section 5.3).

The fourth strand is the **microstructure of listing markets**, which frames two institutional features of our data. Our sample pools a brokered MLS platform (Realtor.ca) with a for-sale-by-owner platform (DuProprio); [Hendel et al. \(2009\)](#) compare exactly these two market forms and document differences in seller composition, marketing, and outcomes, motivating our platform indicator and platform-specific robustness checks. And because we observe asking rather than transaction prices, the interpretation of our dependent variable rests on the list-price literature: the asking price is a strategic choice that anchors search and bargaining ([Horowitz, 1992](#); [Knight, 2002](#); [Han and Strange, 2015](#)), and [Han and Strange \(2016\)](#) show that it acts simultaneously as a commitment device and as a signal in a directed-search environment. We return to this limitation in Section 7.5.

3. Data

3.1. Data Sources

We assemble a cross-section of residential listings from two complementary Quebec platforms, both retrieved on 3 July 2026 with web-scraping actors. The first, **Realtor.ca**, is the national Multiple Listing Service (MLS) portal and provides broker-listed inventory (27,302 raw records). The second, **DuProprio**, is the dominant for-sale-by-owner (FSBO) platform in Quebec (7,693 raw records). Pooling the two sources broadens coverage across price segments and sale channels—brokered and FSBO listings differ systematically in seller composition and marketing ([Hendel et al., 2009](#))—and a source indicator absorbs any systematic level difference between them. The price we observe is the *asking* price; Section 7.5 discusses the interpretation of hedonic coefficients estimated on list rather than transaction prices ([Horowitz, 1992](#); [Knight, 2002](#); [Han and Strange, 2016](#)).

Each record is a structured JSON object containing a price, an address with a postal code, a

property-attribute block, a free-text description, a list of features, and an array of image URLs. The two sources differ in their auxiliary fields: Realtor.ca records include room-by-room dimensions but no geographic coordinates, whereas DuProprio records include coordinates and captioned images but no room block. We construct a unique key by concatenating the source and the listing identifier. Because the data form a single-date cross-section, no time dimension is modelled and no forward-in-time holdout is possible.

3.2. Sample Construction

We restrict attention to *houses*, defined as single-family dwellings and townhouses. On Realtor.ca this corresponds to building types *House* and *Row/Townhouse*; on DuProprio it corresponds to the property types *Home*, *Bungalow*, *Semi Detached*, *Intergenerational*, *Country Home*, *Cottage*, *Villa*, and *Townhouse*. Condominiums, income properties (duplex through quintuplex), vacant land, farms, commercial properties, and mobile homes are excluded. After requiring a non-missing, positive price, the pooled sample contains 19,273 houses.

3.3. Outlier Cleaning

Scraped prices contain parsing artifacts spanning several orders of magnitude (the raw minimum and maximum are \$899 and \$368 billion). We apply conservative, transparent bounds rather than model-based trimming: prices are retained in the \$50,000–\$15,000,000 range; living area in 300–20,000 ft²; lot area in 200–5,000,000 ft²; bedrooms in 0–15 and bathrooms in 0–12. When a single attribute violates its bound it is set to missing (and later imputed) rather than deleting the observation, preserving sample size. Finally, we trim the price-per-square-foot distribution at its 0.5th and 99.5th percentiles (approximately \$73–\$1,276/ft²) among listings with a measured area, removing residual price- or area-parsing errors. The cleaned sample contains **19,123 houses**, of which 80.3% originate from Realtor.ca (Table A2 in the Appendix summarizes the flow).

3.4. Content and Language of the Descriptions

Because the language model reads the description verbatim, what the description contains matters for the interpretation of the text layer. The two platforms differ sharply. A Realtor.ca description is a free-text paragraph written by the listing broker (mean length about 540 characters); only 2.5% of these descriptions contain any dollar amount, and fewer than 0.3% mention the municipal assessment. A DuProprio “description,” as delivered by the crawler, concatenates the owner’s comments with the platform’s structured property sheet rendered as text: room counts, living and lot area, year of construction, exterior cladding, heating, inclusions, and—in every one of the 3,767 DuProprio listings in the cleaned sample—the asking price itself (“Asking Price \$499,999”), which

equals the recorded list price in 3,761 cases; 46.6% of DuProprio descriptions also mention the municipal assessment. The language model was therefore shown the dependent variable for 19.7% of the sample. The extraction schema has no price field, and the prompt instructs the model to report only what the text states about the property, so leakage can operate only indirectly (a model that sees a high price may, for instance, lean towards “like-new” when the condition is ambiguous). We treat this as a first-order threat to the text layer and address it in Sections 7.3 and 7.5: all text-layer gains are reported separately for Realtor.ca, where the text contains no price, and for DuProprio. The vision model receives only the image URL and never sees any text or price.

Quebec is a bilingual listing market. Classifying each description from counts of French and English stop words (a ratio rule of 1.5, as in Boucher, 2026), 44.6% of descriptions are French, 43.1% English, and 12.3% “mixed”—mostly DuProprio descriptions that juxtapose an automated English translation and the French original. Within Realtor.ca, 57.0% of descriptions are French and 41.9% English, and the two groups are priced very differently: the median asking price of English-language Realtor.ca listings is \$749,000 against \$524,900 for French ones, a raw gap of 0.37 log points that falls to 0.21 within municipalities. English descriptions concentrate in the Montreal West Island and other affluent anglophone municipalities, so language proxies for location and for the clientele targeted by the broker. Because the LLM processes both languages with the same schema, language could also affect the extraction itself. We therefore add language indicators to the hedonic specification and re-estimate the gains from both AI layers within each language group (Section 7.2).

3.5. Summary Statistics

Table 1 reports descriptive statistics for the key variables. The median list price is \$599,900. The distribution of living area is right-skewed, motivating the logarithmic transformation adopted in Section 4. Living area is reported for only a minority of Realtor.ca listings; where it is missing we reconstruct it from room-level dimensions (Section 4.2), and where neither is available we rely on the area-free specification. The room count is available only for Realtor.ca listings (the DuProprio room block is absent and coded as zero rooms, with the source indicator absorbing the level).

Table 2 previews the machine-generated attributes described in Section 4: roughly a quarter of listings are described as recently renovated, the mean visual condition and curb-appearance scores are 7.4 and 7.2 on a 10-point scale, and 43% of façades display a visual luxury cue. The text layer covers all 19,123 listings and the vision layer all 19,123 photographs, of which 95.8% were recognized by the model as exterior shots.

Table 1. Summary Statistics

Variable	<i>N</i>	Mean	Std. Dev.	<i>p</i> ₂₅	Median	<i>p</i> ₇₅
Price (CAD)	19,123	764,708	689,118	439,000	599,900	849,000
Living area (ft ²)	9,446	1,730	990	1,128	1,470	2,026
Lot area (ft ²)	19,123	71,429	309,209	6,000	10,686	34,978
Bedrooms	19,123	3.5	1.2	3.0	3.0	4.0
Bathrooms	19,123	1.8	0.8	1.0	2.0	2.0
Parking spaces	19,123	4.0	4.4	0.0	3.0	6.0
Storeys	19,123	1.6	1.6	1.0	2.0	2.0
Age (years)	19,123	42.5	28.0	24.5	40.0	53.0
Rooms (Realtor)	19,123	10.1	6.5	7.0	11.0	14.0
Photos	19,123	42	18	30	40	51

Notes: Descriptive statistics for the cleaned sample of 19,123 single-family and townhouse listings (Realtor.ca and DuProprio), scraped on 3 July 2026. 80% of listings originate from Realtor.ca. Living area is the reported area, or reconstructed from room dimensions where missing. Continuous variables are bounded by the outlier rules described in Section 3; price-per-square-foot is trimmed at the 0.5th and 99.5th percentiles.

Table 2. Machine-Generated Feature Summary

Text (LLM)	Share	Vision (façade)	Mean / share
Recently renovated	26%	Condition score (1–10)	7.4
Renovated kitchen	17%	Curb appeal (1–10)	7.2
Hardwood floor	16%	Luxury cue	43%
Quartz/granite counter	7%	Garage visible	46%
New construction	14%	Brick façade	47%
Income property	12%	Waterfront setting	7%
Water view	24%	Exterior shot detected	96%

Notes: Prevalence of selected machine-generated attributes across the 19,123 cleaned listings. Text attributes are extracted from the description by gpt-4o-mini; vision attributes from the façade photograph by claude-haiku-4-5. Scores are 1–10 means; the remainder are shares.

4. Methodology

Our empirical strategy has two parts: a feature-engineering pipeline that turns raw listings into an analysis-ready design matrix (Sections 4.2–4.5), and a hedonic regression that maps those features to price (Section 4.6).

4.1. Theoretical Framework

Following Rosen (1974), a house is a bundle of K characteristics $\mathbf{z} = (z_1, \dots, z_K)$, and the market clears through a hedonic price function $P(\mathbf{z})$ that equates the marginal bid of buyers to the marginal offer of sellers. A buyer with income y and preferences $U(x, \mathbf{z})$ over a numéraire good x and housing attributes chooses $\mathbf{z}$ to solve

$$\max_{x, \mathbf{z}} U(x, \mathbf{z}) \quad \text{s.t.} \quad y = x + P(\mathbf{z}), \quad (1)$$

which yields the first-order condition that, at the optimum, the implicit (*marginal*) price of each attribute equals the buyer’s marginal willingness to pay:

$$p_k(\mathbf{z}) \equiv \frac{\partial P(\mathbf{z})}{\partial z_k} = \frac{\partial U / \partial z_k}{\partial U / \partial x}, \quad k = 1, \dots, K. \quad (2)$$

Equation (2) is the object of interest: the slope of the hedonic surface in the direction of attribute z_k . The empirical task is to recover $P(\mathbf{z})$ —and hence its gradient—from observed prices and characteristics. The linear model of Section 4.6 imposes a parametric form on $P(\cdot)$ and reads the implicit prices off its coefficients; the machine-learning estimators of Section 6 approximate $P(\cdot)$ nonparametrically, trading the closed-form gradient for predictive flexibility. Two caveats apply throughout. The observed price is an asking price, so $P(\cdot)$ is the hedonic function of sellers’ offers rather than of transactions; and the machine-generated attributes are measured with error and chosen by the seller, so their coefficients are predictive associations rather than structural marginal prices (Section 7.5).

4.2. Structural and Locational Attributes

Structural attributes are extracted by deterministic rules. Living area and lot area are reported in heterogeneous units ("1535 sqft", "1,333.2 ft²", "11 561.52 ft²", "961.6 m²"); we parse each string and convert all areas to square feet. For Realtor.ca listings that report room-by-room dimensions (e.g. "23 ft ,4 in x 13 ft ,8 in") but no aggregate living area, we sum the parsed room areas and calibrate the total against listings where both are observed; the median ratio of reported area to summed-room area is 0.98, so the reconstructed value is used directly to fill

missing living area. Age is computed as 2026 – year built, entered together with its square to allow a non-monotone depreciation profile. Bedrooms, bathrooms, parking spaces, and storeys are taken from the attribute block. Missing values are imputed by the within-property-type median, and each imputed field is accompanied by a “has_” indicator so that the regression can distinguish observed from imputed values.

Location is captured by **municipality fixed effects**. City names are normalized (Realtor.ca appends neighbourhood qualifiers in parentheses, which we strip), and municipalities with fewer than 20 listings are pooled into a residual category, leaving 195 city effects; the pooled group of small municipalities is the omitted (reference) category, so every reported municipality effect is a premium or discount relative to a listing in a small municipality. In robustness checks we replace the municipality effects by **forward sortation area (RTA) effects**, the first three characters of the postal code, which are available for all but one listing: 419 RTAs appear in the sample, of which 250 have at least 20 listings and receive their own effect. Property-type fixed effects and a source indicator complete the locational and platform controls.

4.3. Rule-Based Amenities and Room Detail

Fifteen amenity indicators are detected by bilingual (French/English) keyword matching over the concatenated feature list, description, and title: swimming pool, waterfront, garage, fireplace, central air conditioning or heat pump, finished basement, proximity to water, cul-de-sac, proximity to a golf course, ski hill, park, school, highway, and public transit, and the absence of rear neighbours. From the Realtor.ca room block we additionally construct room-detail indicators—presence of a primary bedroom, a basement, an office, a family room, a workshop, a wine cellar, a sunroom, a mezzanine, a library, and a laundry room—together with the room and storey counts. Two listing-quality signals, the number of photographs (capped at 60) and the presence of a virtual tour, proxy for the marketing tier of the listing. Finally, following the evidence of Section 3.4, two language indicators (English, mixed; French omitted) and the description length are available as controls.

4.4. Language-Model Specifications from Text

Free-text descriptions contain valuation-relevant information that resists keyword extraction: whether the home is turnkey or a fixer-upper, whether the kitchen or bathroom has been renovated, the flooring and countertop materials, the heating system, the presence of a water view, and income potential. We extract these as a fixed set of structured specifications using a large language model (gpt-4o-mini, OpenAI, called through the Chat Completions API in July 2026) with *schema-constrained decoding*: the request specifies a strict JSON schema with closed enumerations and required fields, so the API validates the output at generation time and the model returns exactly the

expected fields in the expected value domains (Table A1). Decoding temperature is set to zero, and the system instruction, written in French, constrains the model to rely only on the text and to code absent information as `false`, `unknown`, or an empty list. Each call consumes roughly 600 input and 110 output tokens; the extraction is run in parallel, is resumable through an on-disk cache keyed by the listing identifier, and retries transient errors with exponential backoff, for a total cost of about \$3 across the sample. Descriptions are truncated at 4,000 characters, which affects only the longest DuProprio property sheets.

4.5. Vision Attributes from the Façade Photograph

To capture the visual quality that neither structured fields nor text convey, we analyze one photograph per listing—the façade—with a vision-language model (`claude-haiku-4-5`, Anthropic, Messages API, July 2026). For Realtor.ca the first image is the exterior shot; for DuProprio, whose images are captioned, we select the first image whose caption is not a room name. The model returns twelve fields under a strict JSON schema: an exterior-shot flag, a visual building type, an overall condition score (1–10), a curb-appeal score (1–10), a renovation level, the list of exterior materials, landscaping quality, a visible-garage flag, the visual storey count, the setting (dense urban, suburban, rural, waterfront, wooded), a photo-quality score, and a luxury indicator. The system instruction directs the model to ground each score in concrete criteria (surface condition, materials, landscaping) rather than taste, and to flag non-exterior images. Images are passed by URL without local download; each call consumes roughly 2,470 input and 100 output tokens, for a total cost of about \$57 across the sample. As with the text layer, the pass is parallelized, resumable, and retried. Unlike the text layer, the vision calls did not set a sampling temperature and therefore ran at the API default, so repeated calls could return slightly different scores for borderline images; we did not re-run the extraction to measure this test–retest variability (see Section 7.5).

From schema fields to regressors. Imposing a typed schema with enumerations, required fields, and `additionalProperties:false` moves validation into the API layer, so the outputs are tabulated variables that require no free-text post-processing. They are then mapped to numeric regressors: booleans to 0/1; the six-level text condition scale to an ordinal score 0–5 (*to renovate* = 0, ..., *new* = 5, with *unknown* placed at the “good” midpoint of 2); the five-level visual renovation scale to 0–4 and landscaping to 0–3; multi-valued fields (flooring, heating, exterior materials) and the setting to one indicator per category. The twelve vision fields thus generate 21 regressors (two scores, the photo-quality score, two ordinal scales, three indicators, five setting indicators, and eight material indicators), and the text schema 18 regressors. A listing whose photograph is not an exterior shot keeps the model’s best-effort scores and is flagged by the exterior indicator. Model identifiers are API aliases that resolve to the provider’s current snapshot; we record the alias and the extraction

date, and the raw JSON outputs are archived so that every regressor can be regenerated from them without new API calls.

4.6. Econometric Specification

We adopt a semi-logarithmic functional form, the standard workhorse in the hedonic literature (Malpezzi, 2003). Collect the attributes of listing i into disjoint blocks,

$$\mathbf{X}_i = \left(\underbrace{\ln \text{Area}_i, \ln \text{Lot}_i}_{\text{size (logged)}}, \underbrace{\mathbf{s}_i}_{\text{structural}}, \underbrace{\mathbf{a}_i}_{\text{amenities}}, \underbrace{\mathbf{r}_i}_{\text{rooms}}, \underbrace{\mathbf{m}_i}_{\text{listing}}, \underbrace{\mathbf{g}_i}_{\text{LLM text}}, \underbrace{\mathbf{v}_i}_{\text{vision}} \right) \quad (3)$$

and estimate, by ordinary least squares,

$$\begin{aligned} \ln P_i = & \alpha + \beta \ln \text{Area}_i + \gamma \ln \text{Lot}_i + \mathbf{s}_i^\top \boldsymbol{\delta}_s + \mathbf{a}_i^\top \boldsymbol{\delta}_a + \mathbf{r}_i^\top \boldsymbol{\delta}_r + \mathbf{m}_i^\top \boldsymbol{\delta}_m \\ & + \underbrace{\mathbf{g}_i^\top \boldsymbol{\delta}_g}_{\text{LLM text}} + \underbrace{\mathbf{v}_i^\top \boldsymbol{\delta}_v}_{\text{vision}} + \underbrace{\mu_{c(i)}}_{\text{city FE}} + \underbrace{\tau_{t(i)}}_{\text{type FE}} + \varepsilon_i, \end{aligned} \quad (4)$$

where $c(i)$ and $t(i)$ index the municipality and property type of listing i , and μ_c , τ_t are the corresponding fixed effects that sweep out location- and type-specific price levels. The source indicator is part of $\mathbf{m}_i$. In matrix form $\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ with $\hat{\boldsymbol{\theta}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}$. The contribution of a feature layer is measured by the change in R^2 when its block is added to the specification that contains all preceding blocks; the order is size and structural, amenities, rooms and listing signals, text, vision, so the text layer is credited only with what the keyword amenities do not already capture, and the vision layer only with what neither the text nor the structured fields capture.

Interpretation of the coefficients. Because the size variables enter in logs, β and γ are *elasticities*: $\partial \ln P / \partial \ln \text{Area} = \beta$. For an indicator regressor d with coefficient b , the percentage price effect is not b but

$$\% \Delta P = (e^b - 1) \times 100, \quad (5)$$

the Halvorsen and Palmquist (1980) correction we apply throughout (for small b , $e^b - 1 \approx b$; the Kennedy, 1981 adjustment $e^{b - \frac{1}{2} \widehat{\text{Var}}(b)} - 1$ moves none of our reported effects by more than 0.1 percentage point). For an ordinal score such as curb appeal, b is the log-price difference associated with one step of the scale, holding the other attributes fixed. Under Equation (4) the implicit price of a continuous attribute z_k from (2) is $p_k = \delta_k P / z_k$, so the semi-log form lets the *dollar* marginal price scale with the price level while keeping the *percentage* effect constant.

Two-part coverage design. Because living area is reported for only about 38% of Realtor.ca listings, we estimate two nested versions of (4) and combine their predictions:

$$\widehat{\ln P_i} = \begin{cases} \widehat{\ln P_i^A} & \text{if area is measured (Model A, rich),} \\ \widehat{\ln P_i^B} & \text{otherwise (Model B, broad; } \beta \equiv 0). \end{cases} \quad (6)$$

Model A includes $\ln \text{Area}$ and is fit on the 9,446 listings with a measured or reconstructed area; Model B drops it and is fit on the full cleaned sample, so every listing can be priced.

Inference. House-price residuals are strongly heteroskedastic, and listings within a municipality share unobserved local shocks that induce within-cluster error correlation. Our preferred standard errors are therefore clustered at the municipality level (195 clusters), the same level as the fixed effects (Cameron and Miller, 2015; Abadie et al., 2023):

$$\widehat{\text{Var}}_{\text{CL}}(\hat{\boldsymbol{\theta}}) = (\mathbf{Z}^\top \mathbf{Z})^{-1} \left(\sum_{c=1}^C \mathbf{Z}_c^\top \hat{\boldsymbol{\varepsilon}}_c \hat{\boldsymbol{\varepsilon}}_c^\top \mathbf{Z}_c \right) (\mathbf{Z}^\top \mathbf{Z})^{-1}, \quad (7)$$

where $\mathbf{Z}_c$ and $\hat{\boldsymbol{\varepsilon}}_c$ stack the regressors and residuals of cluster c . Cluster sizes are very unequal—Montreal alone contributes about one listing in twenty, and the pooled residual category of small municipalities, which is treated as a single cluster, holds about one listing in four (5,006 listings from 999 municipalities)—a configuration in which the asymptotic cluster-robust variance can be unreliable (MacKinnon and Webb, 2017). We therefore also report a pairs cluster bootstrap that resamples municipalities with replacement ($B = 299$) and takes the 2.5th and 97.5th percentiles of the resampled coefficients as a confidence interval. As a complement we compute heteroskedasticity-consistent (HC3) standard errors (White, 1980; MacKinnon and White, 1985); Section 7.1 shows that the three approaches agree on the significance of the property-level AI attributes and differ mainly for keyword amenities with a strong spatial component. When RTA fixed effects replace municipality effects, standard errors are clustered by RTA.

Retransformation. Predicting the *level* of price requires undoing the log. Because $\mathbb{E}[P | \mathbf{Z}] = \exp(\mathbf{Z}\boldsymbol{\theta}) \mathbb{E}[e^\varepsilon] \neq \exp(\mathbf{Z}\boldsymbol{\theta})$, naive exponentiation is biased downward. Under $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ we use the closed-form correction, and more generally Duan’s nonparametric smearing estimate (Duan, 1983):

$$\hat{P}_i = \exp(\widehat{\ln P_i}) \cdot \underbrace{\exp(\frac{1}{2}\hat{\sigma}^2)}_{\text{normal}} \quad \text{or} \quad \hat{P}_i = \exp(\widehat{\ln P_i}) \cdot \underbrace{\frac{1}{n} \sum_{j=1}^n e^{\hat{\varepsilon}_j}}_{\text{smearing } \hat{\phi}}. \quad (8)$$

The out-of-sample accuracy metrics of Section 6 are computed on $\ln P$ (for R^2) and on the ratio of predicted to actual price (for the median absolute percentage error), neither of which depends on the retransformation constant.

Out-of-sample validation. All predictive comparisons use five-fold cross-validation with a common seed, so that every estimator and every feature set is evaluated on identical held-out folds. Under *random* folds a held-out house typically has close neighbours in the training folds, which flatters any model that can exploit spatial proximity (Roberts et al., 2017). We therefore repeat the exercise with *spatially grouped* folds, assigning entire RTAs to folds so that no held-out listing shares its RTA with a training listing; because many small municipalities consist of a single RTA, their municipality indicator is then unobserved in training and the model prices them at the reference level. The spatial design is a deliberately harsh test of how the estimators and feature layers behave for a new area; the random design corresponds to the operational use of an AVM that prices a new listing in a market it already covers.

5. Results

5.1. Baseline Estimates

Table 3 reports the two hedonic specifications estimated on the full feature set (structural, amenity, room-detail, listing-signal, LLM text-derived, and vision-derived attributes). Model A, which includes living area and is estimated on the 9,446 listings with a measured area, explains 75.5% of the variation in log prices ($R^2 = 0.755$, adjusted 0.748). Model B, the area-free specification estimated on all 19,123 listings, explains 72.5% ($R^2 = 0.725$, adjusted 0.721). Adding the machine-generated feature layers in sequence raises Model A’s R^2 from 0.717 (structural, amenity, room, and listing attributes only) to 0.725 with the text layer and to 0.755 with the vision layer: 3.8 points in total, of which 0.8 come from the text and 3.0 from the façade photograph. The text increment is small because the fifteen keyword amenities already capture the most frequently mentioned features (garage, pool, fireplace, air conditioning), so the LLM is credited only with what keyword matching misses; Section 7.3 shows that on Realtor.ca, where the text contains no price, the increments are 0.65 and 2.4 points.

The estimated implicit prices are economically sensible and, with few exceptions, statistically significant with standard errors clustered by municipality. In Model A, the living-area elasticity is 0.36 (a 10% larger home commands roughly a 3.6% higher price) and the lot-area elasticity is small (0.027) but precisely estimated. Each additional **bathroom** is associated with a 7.6% premium and each parking space with about 1%. **Central air conditioning** carries a 5.5% premium and **public-transit** proximity 9.4%, although the transit coefficient—like other keyword amenities with

a strong spatial component—is estimated much less precisely once errors are allowed to correlate within municipalities. **Waterfront** access is associated with a 6.0–6.4% premium, significant at conventional levels in the full-sample model. Age enters with a negative linear term (-0.4% per year) and a small positive quadratic (2.6×10^{-5}), so the price–age profile bottoms out at about 70 years and rises for older, heritage-type houses.

5.2. Text- and Vision-Derived Attributes

The two AI feature layers contribute distinct signal.

Text (LLM). Beyond the keyword amenities, the structured specifications carry independent information about finish and condition: a **renovated kitchen** is associated with a 2.6% premium, **hardwood flooring** with 2.4%, and the ordinal **condition** scale is positive and significant in both models (1.6% per step). Listings the model reads as **income properties** are listed about 1.9% below otherwise comparable houses. The **new-construction** indicator, by contrast, is indistinguishable from zero once age, the visual renovation level, and the condition scales are controlled for: the premium a naive specification would attribute to the label “new” is absorbed by the measured quality attributes.

Vision (façade). The image-derived attributes are jointly significant and add the larger increment to fit. Holding all structural, amenity, and text attributes fixed, a one-point increase in **curb appeal** (1–10) is associated with a 2.3% higher price, and a one-step increase in the visual **renovation level** with 2.4%. A visible **luxury** cue adds 5.2%, a **visible garage** 4.7%, and each step of **landscaping** quality 1.5%. Exterior materials price as expected: **brick** (+3.4%) and **stone** (+3.7%) command premia, while **vinyl** siding carries a 4.4% discount; because a façade can combine several materials, each indicator is measured relative to façades without that material. The visual **rural** setting is associated with an 11.5% discount relative to the other settings even within municipalities, which indicates that the model picks up within-municipality variation in density that the fixed effects cannot. The visual condition score, by contrast, is not significant in Model A once curb appeal and the renovation level are included; the three scores are highly collinear (the model rates condition, curb appeal and renovation level from the same cues), and their individual coefficients should be read as a joint “visual-quality” block rather than as three separate prices. That these visual measures remain significant after conditioning on the full text and structural set, with inference clustered at the municipality level, indicates that the façade photograph carries price-relevant information available in no other field.

Table 3. Hedonic Regression Results (Text + Vision Features)

	(1) Model A <i>rich</i>	(2) Model B <i>broad</i>
ln(Living area)	0.358*** (0.012)	
ln(Lot area)	0.027*** (0.006)	0.040*** (0.005)
Bedrooms	0.014*** (0.005)	0.032*** (0.003)
Bathrooms	0.073*** (0.007)	0.113*** (0.007)
Parking	0.011*** (0.001)	0.015*** (0.002)
Age	−0.004*** (0.001)	−0.005*** (0.001)
<i>Amenities & text (LLM)</i>		
Central A/C	0.053*** (0.018)	0.020 (0.018)
Public transit	0.090** (0.046)	0.058 (0.039)
Waterfront	0.058* (0.033)	0.062*** (0.019)
Condition (LLM, ord.)	0.016*** (0.006)	0.016*** (0.005)
Reno. kitchen (LLM)	0.026*** (0.007)	0.027*** (0.010)
Hardwood floor (LLM)	0.024** (0.009)	0.030*** (0.011)
New construction (LLM)	−0.000 (0.013)	−0.016 (0.014)
Income property (LLM)	−0.019** (0.009)	−0.013* (0.008)
<i>Vision (façade)</i>		
Curb appeal (1–10)	0.023*** (0.009)	0.016 (0.013)
Renovation level (ord.)	0.024** (0.010)	0.020** (0.009)
Luxury (visual)	0.051*** (0.012)	0.062*** (0.009)
Garage visible	0.046*** (0.006)	0.047*** (0.007)
Brick façade	18 0.033*** (0.009)	0.044*** (0.011)
Stone façade	0.036*** (0.009)	0.047*** (0.007)

5.3. Validation of Machine-Generated Features

Before relying on the AI-generated attributes we assess them in two ways. First, we check that the vision scores order prices in the expected direction. Figure 1 plots the median list price against the vision model’s curb-appeal and condition scores among listings with a detected exterior shot. Both relationships are monotone: median price rises from the lowest to the highest visual scores, with no reliance on the regression to impose the pattern. Because these scores are produced by a model that never sees the price, the monotonicity shows that the vision layer extracts a price-correlated quality signal rather than noise; it does not, of course, establish that the signal is *quality* rather than, say, the marketing budget of the seller (Section 7.5).

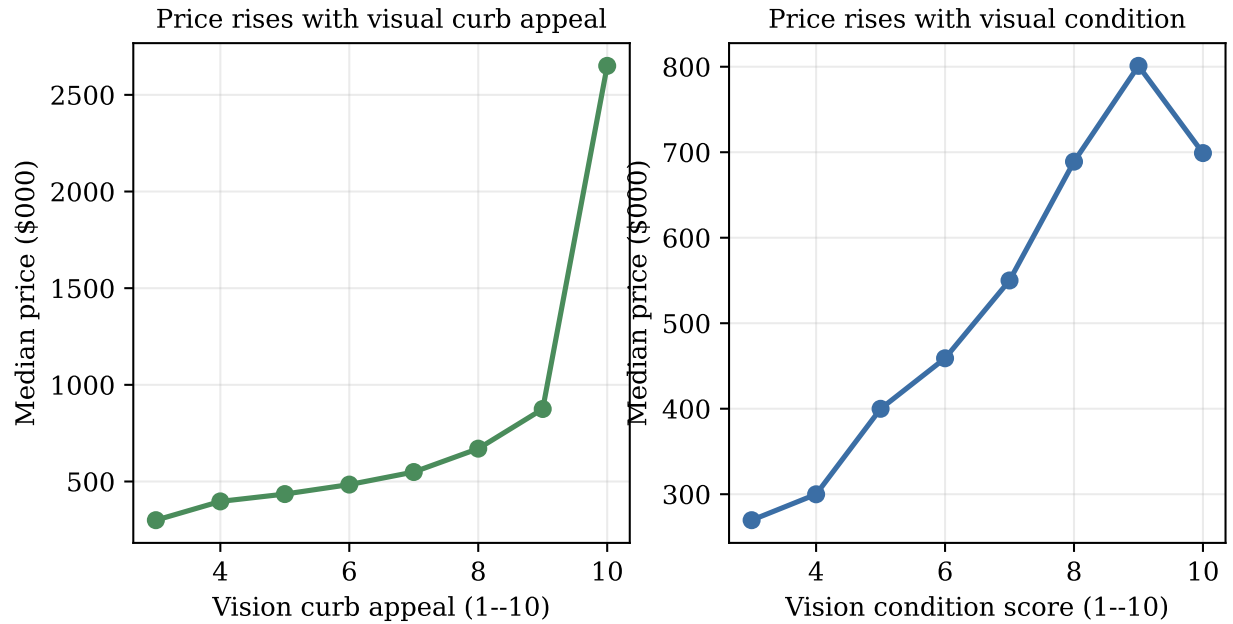


Figure 1. Validation of the vision layer. Median list price rises monotonically with the model’s curb-appeal (left) and condition (right) scores, among listings with a detected façade shot. The scores are generated without access to price.

Second, because we do not have human labels, we measure extraction error indirectly by confronting the machine-generated attributes with independent sources (Table 4). Where the LLM and the bilingual keyword rule read the same text for the same binary feature, agreement is very high for concrete, unambiguous amenities—garage (agreement 98%, Cohen’s $\kappa = 0.96$), fireplace ($\kappa = 0.91$), swimming pool ($\kappa = 0.81$)—and lower for features whose definition is fuzzy or whose keywords are ambiguous: central air conditioning ($\kappa = 0.65$; the keyword rule also fires on “air exchanger”), finished basement ($\kappa = 0.56$; the keyword rule fires on any mention of “basement”), and waterfront ($\kappa = 0.50$). For waterfront the LLM flags 22.5% of listings against 9.2% for the rule, with a recall of 98% but a precision of only 40% relative to the rule, which suggests that

the model treats water *access* or *views* as waterfront; the regression includes both variables, and the keyword waterfront is the one whose coefficient we report. Across modalities, the vision model’s visible-garage flag agrees with the text-based garage indicator for 68% of exterior shots ($\kappa = 0.35$)—a garage need not be visible from the front—and its storey count matches the structured storey field exactly for 60% of listings and within half a storey for 89%. Two fields are clearly unreliable: the visual waterfront setting ($\kappa = 0.39$ against the keyword) and the visual building type, which labels 12.7% of façades as row houses when only 3.0% of listings are townhouses (precision 18%) and 14.5% as “chalets.” We accordingly use the visual building type and setting only as controls and base no conclusion on them, and we read the reported coefficients of the vision block as prices of *what the model sees* rather than of verified physical attributes.

Table 4. Agreement Between Machine-Generated Attributes and Independent Sources

Attribute	<i>N</i>	Share (AI)	Share (ref.)	Agreement	κ	Precision	Recall
<i>Panel A: LLM text layer vs. bilingual keyword rule on the same text</i>							
Garage	19,123	40.8%	41.0%	98.2%	0.96	98.0%	97.6%
Swimming pool	19,123	17.9%	23.9%	93.5%	0.81	98.8%	73.9%
Waterfront	19,123	22.5%	9.2%	86.2%	0.50	39.8%	97.7%
Fireplace	19,123	21.9%	23.6%	96.7%	0.91	96.3%	89.4%
Central A/C or heat pump	19,123	11.5%	18.9%	90.8%	0.65	92.5%	56.0%
Finished basement	19,123	35.5%	31.9%	80.5%	0.56	67.5%	75.0%
<i>Panel B: Façade vision layer vs. text or structured fields (exterior shots only)</i>							
Garage visible vs. garage (LLM)	18,314	47.5%	41.0%	67.6%	0.35	59.0%	68.4%
Garage visible vs. garage (keyword)	18,314	47.5%	41.1%	67.4%	0.34	59.0%	68.1%
Waterfront setting vs. waterfront (keyword)	18,314	6.5%	9.1%	91.1%	0.39	51.7%	37.2%
Row house vs. townhouse (listing type)	18,314	12.7%	3.0%	89.0%	0.26	18.5%	79.0%

Notes: “Share (AI)” is the prevalence of the attribute according to the AI layer, “Share (ref.)” according to the reference source; “Agreement” is the share of listings on which the two coincide, κ is Cohen’s kappa, and precision and recall treat the reference as ground truth (it is not: keyword rules miss negations and synonyms, and a garage need not be visible on the façade photograph). Visual storey count agrees exactly with the structured storey field for 60.5% of 18,065 listings and within half a storey for 88.7%.

5.4. Location Fixed Effects

The 195 municipality fixed effects absorb the dominant locational component of price. Relative to the pooled group of small municipalities, the largest positive premia are estimated for **Westmount** ($\approx +231\%$), **Hampstead** ($+173\%$), and **Mont-Royal** ($+169\%$), followed by remote municipalities at the low end carrying discounts of 23–30%. Figure 2 plots the extreme effects. The magnitude and ordering align with well-known patterns in the Quebec housing market and lend face validity to the specification; one apparent anomaly, Sainte-Anne-des-Plaines ($+143\%$), is an artefact of the two-part design: only two of that municipality’s 33 listings report a living area and enter Model A, one of them a \$5.5 million estate on a 100-acre lot, so the effect is identified from two observations.

Municipality effects estimated on a single-date cross-section with sparse cells should not be read as permanent location premia.

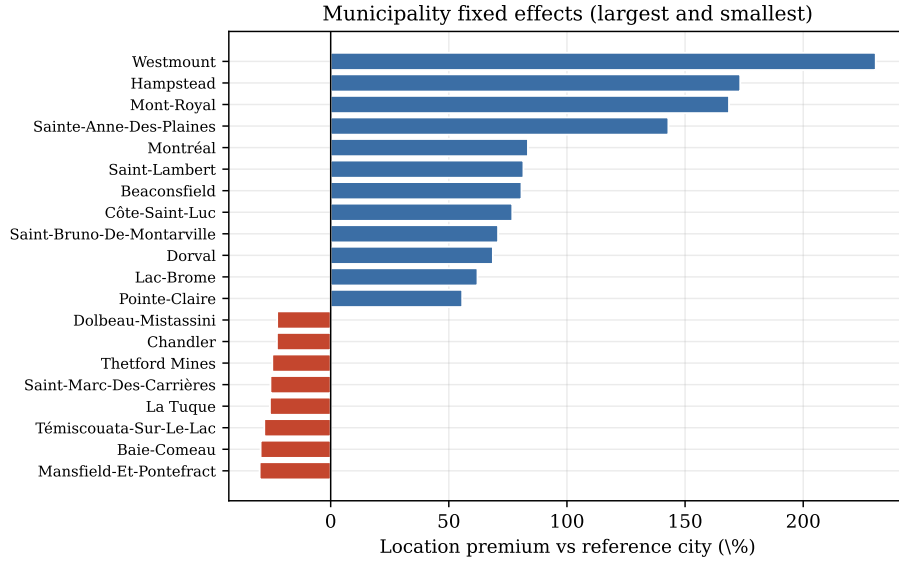


Figure 2. Estimated municipality fixed effects (Model A), the twelve largest and eight smallest, expressed as percentage premia relative to the pooled group of municipalities with fewer than 20 listings.

6. Predictive Benchmark: From OLS to Deep Learning

The hedonic OLS of Section 5 is designed for interpretation: its coefficients are implicit prices. It is not, however, the most accurate predictor of price. To map the accuracy frontier—and to gauge how much curvature and interaction the linear model leaves on the table—we benchmark a ladder of estimators of increasing flexibility on the *identical* feature set, under a common five-fold cross-validation protocol. All rows of Table 5 share the same folds and random seed, so differences reflect the estimator, not the split.

6.1. The Estimator Ladder

(1) Ordinary least squares. The linear hedonic of Equation (4), with one-hot city, property-type, and source indicators and within-property-type median imputation of missing continuous attributes. It is the interpretable baseline and, mechanically, the least flexible: it imposes constant marginal effects and additive separability.

(2) Random forest. A bagged ensemble of regression trees (Breiman, 2001; Mullainathan and Spiess, 2017). Trees capture non-linearities and low-order interactions automatically and handle

missing values natively, but averaging deep decorrelated trees tends to over-smooth sharp price thresholds.

(3) Gradient boosting. A histogram-based gradient-boosted tree ensemble (600 boosting iterations, learning rate 0.05, up to 63 leaves, L_2 regularization). Boosting fits trees sequentially to residuals, which lets it model both smooth curvature (age, area) and abrupt thresholds; like the forest it ingests missing values directly. It is the strongest single model in our comparison.

(4) Neural networks. A plain multilayer perceptron and a robust residual network with entity embeddings, described in Section 6.5. Neural networks are attractive when high-cardinality categoricals (here, 195 municipalities) can be represented as dense learned vectors rather than sparse indicators.

(5) Blend. An equally weighted average of the robust neural network and the gradient-boosting predictions. Because the two models make partly decorrelated errors—the network smooths through embeddings while the boosted trees cut sharp thresholds—their average cancels idiosyncratic noise, a standard model-averaging result. The weight of one half was fixed *a priori*; Section 6.3 reports how the result varies with the weight and under a nested choice of the weight.

6.2. Formal Definitions

Let $y_i = \ln P_i$ and let $\mathbf{x}_i$ denote the feature vector of Equation (3). Each estimator approximates the conditional mean $f(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$.

Linear (OLS). $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}$ with the closed form of Section 4.6; categoricals enter as one-hot indicators.

Random forest (Breiman, 2001). An average of B regression trees, each grown on a bootstrap resample with a random feature subset at every split:

$$\hat{f}_{\text{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x}; \Theta_b), \quad (9)$$

where Θ_b denotes the randomness (resample and split subsets) of tree b ; we use $B = 400$ trees, half of the features at each split, and a minimum leaf size of three.

Gradient boosting (Friedman, 2001). A forward stagewise additive expansion built by fitting each new tree to the negative gradient of the loss:

$$F_M(\mathbf{x}) = F_0 + \nu \sum_{m=1}^M f_m(\mathbf{x}), \quad f_m = \arg \min_f \sum_{i=1}^n (-g_{mi} - f(\mathbf{x}_i))^2, \quad g_{mi} = \frac{\partial L(y_i, F_{m-1}(\mathbf{x}_i))}{\partial F_{m-1}(\mathbf{x}_i)} \quad (10)$$

with learning rate ν , squared-error loss L , and L_2 -regularized histogram trees f_m ; missing values are routed to a learned default direction at each split.

Neural network (robust). High-cardinality categoricals are mapped to learned embeddings (Guo and Berkahn, 2016), concatenated with the batch-normalized numeric block and projected to a d -dimensional trunk:

$$\mathbf{h}_0 = \mathbf{W}_0 \left[\bigoplus_{c \in \{\text{city, type, src}\}} \mathbf{E}_c(\text{cat}_{ci}) \parallel \text{BN}(\mathbf{x}_i^{\text{num}}) \right]. \quad (11)$$

The trunk passes through $L = 3$ pre-activation residual blocks (He et al., 2016; Gorishniy et al., 2021),

$$\mathbf{h}_\ell = \mathbf{h}_{\ell-1} + \text{Drop}\left(\mathbf{W}_2^{(\ell)} \text{Drop}\left(\phi\left(\mathbf{W}_1^{(\ell)} \text{BN}(\mathbf{h}_{\ell-1})\right)\right)\right), \quad \phi = \text{GELU}, \quad (12)$$

and a normalization head yields $\hat{y}_i = \mathbf{w}^\top \phi(\text{BN}(\mathbf{h}_L)) + b$. Parameters minimize the Huber (smooth- L_1) loss, robust to price outliers,

$$L_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq \delta, \\ \delta(|r| - \frac{1}{2}\delta), & |r| > \delta, \end{cases} \quad r_i = \hat{y}_i - y_i, \quad \delta = 1, \quad (13)$$

on the standardized target, optimized with AdamW (decoupled weight decay) (Loshchilov and Hutter, 2019) under a cosine-annealed learning rate $\eta_t = \frac{1}{2}\eta_0(1 + \cos(\pi t/T))$, gradient-norm clipping, and early stopping on a 10% validation split of each training fold. Three seeds are averaged per fold.

Blend. A convex combination of the two strongest models in log space,

$$\widehat{\ln P}^{\text{blend}} = w \widehat{\ln P}^{\text{GBM}} + (1 - w) \widehat{\ln P}^{\text{NN}}, \quad w = 0.5. \quad (14)$$

Evaluation. With $\hat{y}_{-k(i)}$ the prediction for i from the model trained on the folds excluding i 's fold, we report out-of-sample fit and level accuracy,

$$R_{\text{OOS}}^2 = 1 - \frac{\sum_i (y_i - \hat{y}_{-k(i)})^2}{\sum_i (y_i - \bar{y})^2}, \quad \text{MdAPE} = \text{median}_i \left| e^{\hat{y}_{-k(i)} - y_i} - 1 \right|, \quad (15)$$

together with the share of listings predicted within $\pm 15\%$ and $\pm 25\%$ of the asking price. The MdAPE is computed on the ratio of predicted to actual price and is therefore unaffected by the retransformation constant of Equation (8).

6.3. Results

Table 5. Predictive Performance: From OLS to Deep Learning

Estimator	R_{OOS}^2	Median AE	Within $\pm 15\%$	Within $\pm 25\%$
<i>Panel A: Structural + amenity + LLM-text features</i>				
OLS (linear hedonic)	0.628	17.9%	43%	64%
Random Forest	0.677	17.9%	43%	64%
Neural network (plain MLP)	0.711	16.1%	47%	69%
Neural network (robust ResNet)	0.741	15.2%	49%	70%
Gradient boosting (HistGBM)	0.743	15.4%	49%	70%
Blend (NN + GBM)	0.760	14.7%	51%	72%
<i>Panel B: + Vision (façade) features</i>				
OLS (linear hedonic)	0.686	16.9%	45%	67%
Random Forest	0.704	17.2%	45%	66%
Neural network (robust ResNet)	0.747	14.8%	51%	71%
Gradient boosting (HistGBM)	0.770	14.5%	51%	71%
Blend (NN + GBM)	0.779	14.0%	53%	73%

Notes: Out-of-sample performance from 5-fold cross-validation with a common seed (identical folds across rows). R_{OOS}^2 is on $\ln(\text{Price})$; “Median AE” is the median absolute percentage error in levels (Equation 15). Panel A uses structural, amenity, room-detail, listing-signal, and LLM text-derived features; Panel B adds the façade vision layer (21 regressors derived from the twelve schema fields). City, type, and source enter as one-hot indicators for the linear and tree models and as learned embeddings for the neural networks. The tree and neural models handle missing living area natively; OLS uses within-type median imputation. The blend weights the two models equally (weight fixed a priori). The best model is the text + vision blend. Entries may differ by ± 0.001 from Tables 6 and 10, which come from separate runs.

Table 5 reports the ladder in two panels. In Panel A (structural, amenity, room, listing, and text features) out-of-sample R^2 rises monotonically with flexibility: from 0.628 for OLS, to 0.677 for the random forest, 0.711 for the plain network, and a near-tie at the top between the robust network (0.741) and gradient boosting (0.743); their blend attains $R_{\text{OOS}}^2 = 0.760$. Panel B adds the façade

vision layer. Vision lifts every estimator—OLS gains 5.8 points ($0.628 \rightarrow 0.686$), the random forest 2.7 ($0.677 \rightarrow 0.704$), the robust network 0.6 ($0.741 \rightarrow 0.747$), and gradient boosting 2.7 ($0.743 \rightarrow 0.770$)—and the text + vision blend attains $R^2_{\text{OOS}} = 0.779$ with a median absolute percentage error of 14.0%, the best model in the study.

Three lessons follow. First, on the full feature set roughly nine points of out-of-sample R^2 separate the interpretable linear model (0.686) from the predictive frontier (0.779)—the price of interpretability, and a measure of the curvature and interactions the hedonic form omits. Notably, the vision layer buys the linear model more accuracy (+5.8 points) than switching from OLS to a random forest does (+4.9 points in Panel A), so better measurement rivals better estimation. Second, on this tabular problem gradient boosting and a carefully regularized neural network are statistically comparable, consistent with the broader evidence that tree ensembles remain strong baselines on tabular data (Grinsztajn et al., 2022); the gains come from *combining* them. Third, the façade-vision layer is not redundant with the text and structural fields: it is the more informative of the two AI layers for every estimator.

The equal-weight blend is not the best achievable combination, but it is close. Varying the weight w on gradient boosting in Equation (14) over $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ gives out-of-sample R^2 of 0.771, 0.776, 0.779, 0.780, and 0.780 for the text + vision models, so $w = 0.5$ is within 0.002 of the best weight on the grid; choosing w in a nested fashion—on the four training folds, then applied to the fifth—returns $R^2_{\text{OOS}} = 0.775$, slightly below the fixed equal weight, because the fold-specific optimal weights are noisy. For the text-only models the equal weight is exactly the best on the grid (0.760). We therefore keep the fixed equal weight and do not tune it on the test predictions.

6.4. Feature Ablation and Importance

Table 6 and Figure 3 decompose the accuracy gains layer by layer. Reading down the table, the room-detail and listing-signal block adds the largest increment of all (9.8 points for OLS, 3.7 for gradient boosting), which reflects both the information in the Realtor.ca room block and the fact that this block carries the platform’s identity; the LLM text layer adds 0.9 and 1.4 points; and the vision layer adds 5.8 and 2.7 points, the largest increment of the two AI layers and the second largest overall. The pattern is monotone: no layer degrades fit. Under spatially grouped folds (Table 10, Section 7.4) all levels fall by three to five points, but the increments from the text and vision layers are essentially unchanged.

Figure 4 aggregates the gradient-boosting permutation importances into feature blocks. Structural attributes and location fixed effects dominate, as expected, but the vision block accounts for 13% of total importance—more than the rule-based amenities and the LLM text block combined—so the façade photograph is a first-order, not marginal, source of price information in the boosted model. Figure 5 shows that the blend is well calibrated across the price range, with an approximately

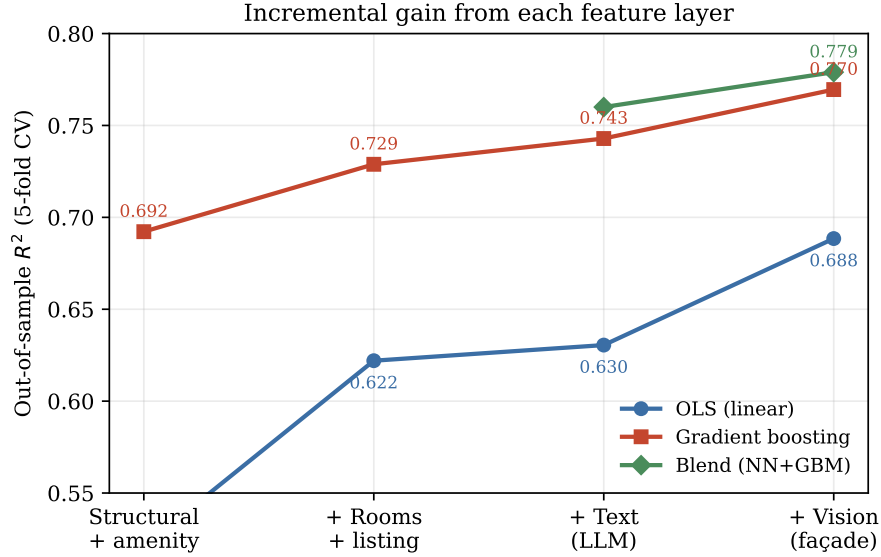


Figure 3. Out-of-sample R^2 as feature layers are added (nested, left to right), random five-fold cross-validation. Each layer improves both the linear and boosted models; the façade-vision layer adds about +5.8 points for OLS and +2.7 for gradient boosting, the text layer +0.9 and +1.4.

Table 6. Feature Ablation: Out-of-Sample R^2 by Feature Layer

Feature set	# feats	OLS	GBM	Blend
Structural + amenity	24	0.521	0.692	—
+ Rooms + listing	38	0.619	0.729	—
+ Text (LLM)	56	0.628	0.743	0.760
+ Vision (façade)	77	0.686	0.770	0.779

Notes: Each row adds a feature layer to the row above (nested). Entries are out-of-sample R^2 on $\ln(\text{Price})$ from 5-fold cross-validation with a common seed. “# feats” counts numeric/indicator features before the city, type, and source fixed effects (added in every specification). The blend averages the robust neural network and gradient boosting and is reported where both were trained. GBM handles missing living area natively; OLS uses within-type median imputation. The gain from a layer is the difference between consecutive rows; the room and listing block includes the Realtor.ca room detail and therefore also carries platform identity.

symmetric error distribution centred near zero.

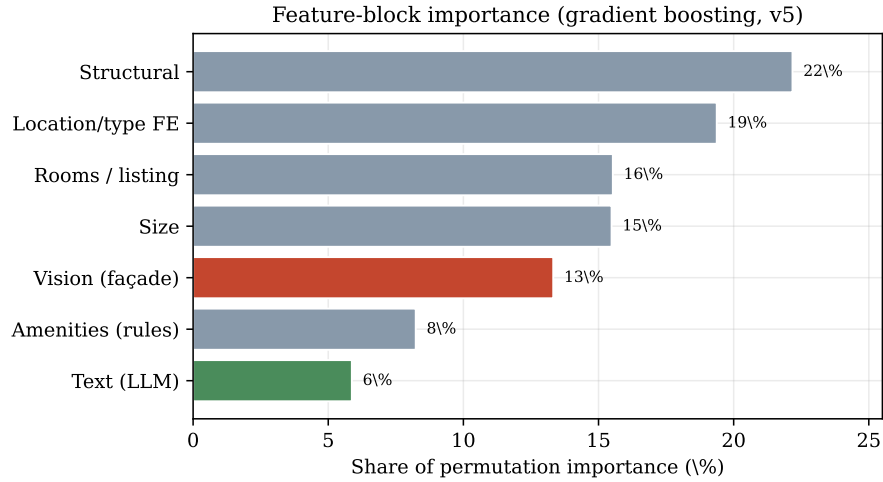


Figure 4. Share of gradient-boosting permutation importance by feature block (text + vision model). The vision block accounts for 13% of total importance, exceeding the rule-based amenity and LLM-text blocks.

6.5. A Robust Tabular Neural Network

Figure 6 details the robust network. High-cardinality categoricals enter as learned **entity embeddings**: each municipality, property type, and source is mapped to a dense vector (dimension $\approx 1.6C^{0.56}$ in cardinality C , so about 30 dimensions for the 195 municipalities). The embeddings are concatenated with the batch-normalized numeric block and projected to a 256-dimensional trunk, which passes through three **residual blocks** (BatchNorm $\rightarrow$ Linear₅₁₂ $\rightarrow$ GELU $\rightarrow$ Dropout $\rightarrow$ Linear₂₅₆ $\rightarrow$ Dropout, with a skip connection), followed by a normalization head that outputs $\ln(\text{Price})$.

Robustness comes from the training recipe rather than raw capacity: a **Huber (smooth- L_1) loss** that damps the influence of price outliers; **AdamW** with decoupled weight decay and a **cosine-annealed** learning rate; **gradient-norm clipping** at 5; **dropout** and batch normalization throughout; **missingness indicators** appended to the numeric block so the network can exploit the non-random pattern of missing fields; and **early stopping** on a held-out 10% of each training fold. Finally, three networks with different random seeds are trained per fold and **averaged**, which reduces prediction variance. The result (0.741 with text, 0.747 with text and vision) is competitive with gradient boosting and, once blended with it, produces the best model in the study.

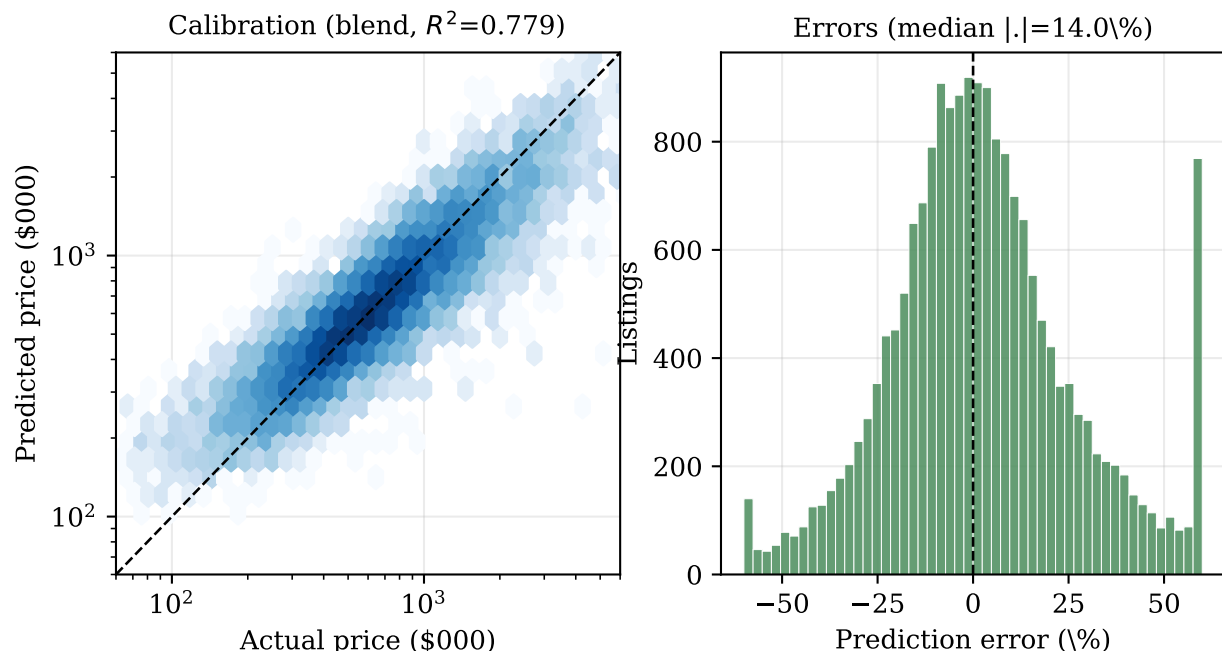


Figure 5. Out-of-sample calibration of the text + vision blend. Left: predicted versus actual price (log–log; dashed line is the 45° identity). Right: distribution of percentage prediction errors, with median absolute error of 14.0%.

7. Robustness and Limitations

7.1. Inference and Subsample Stability

Several features of the design guard against spurious results. The two-part specification (Section 4.6) ensures that estimates are not driven by the non-random availability of living area: coefficients on attributes common to both models retain their sign and rough magnitude whether or not area is conditioned on. The “has_” indicators accompanying each imputed field let the regression separate observed from imputed information, so median imputation does not silently attenuate the size coefficients. The price-per-square-foot trim removes the residual parsing errors that would otherwise dominate the tails, and the municipality fixed effects absorb the first-order spatial confound emphasized by [Can \(1992\)](#) and [Anselin \(1988\)](#).

Inference. Because listings within a municipality share unobserved local shocks, our preferred standard errors are clustered at the municipality level (195 clusters; [Cameron and Miller, 2015](#)), the same level as the fixed effects. Clustering matters exactly where theory predicts: the text- and vision-derived quality attributes—which vary house by house within a municipality—are essentially unaffected (curb appeal, renovation level, luxury, garage, materials, landscaping, and

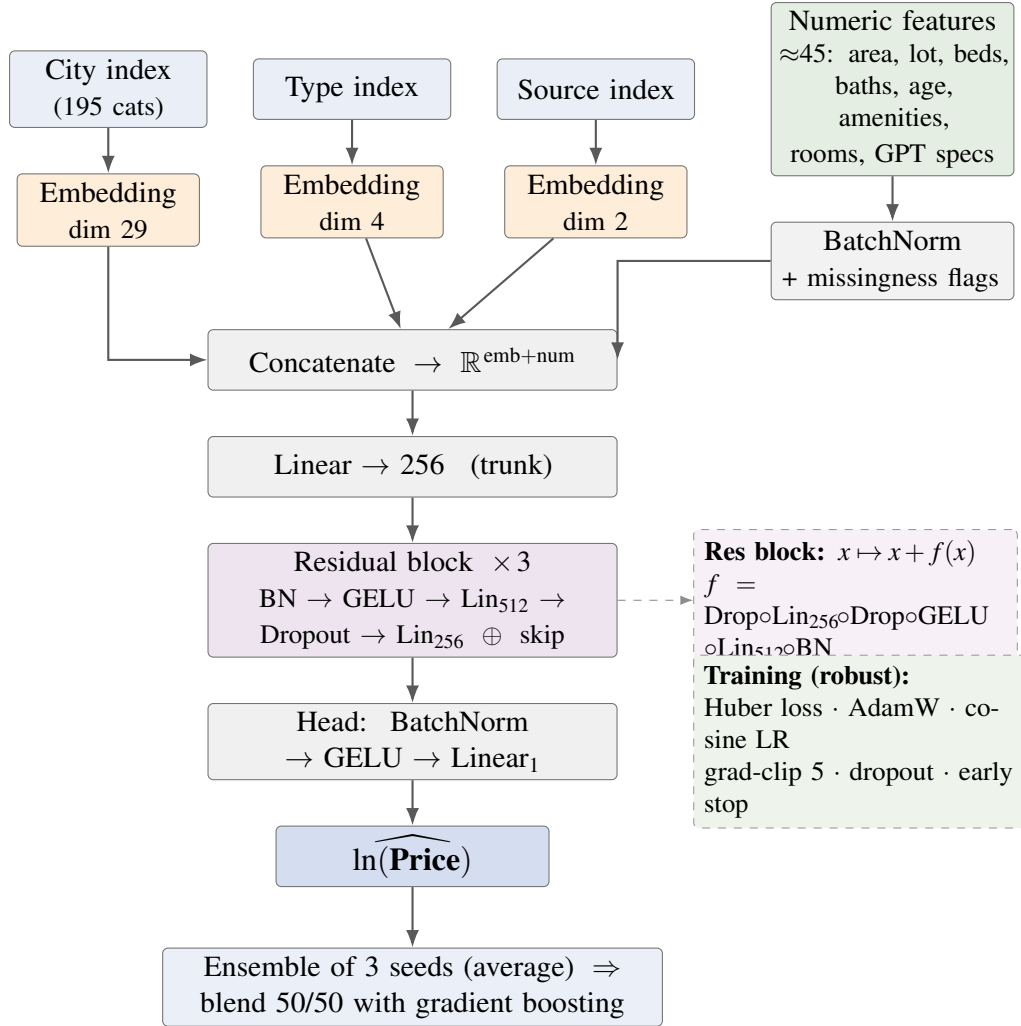


Figure 6. Architecture of the robust tabular neural network. High-cardinality categoricals enter as learned embeddings; numeric features are batch-normalized and augmented with missingness indicators. The concatenated representation is projected to a 256-dimensional trunk, refined by three residual blocks, and mapped to $\ln(\text{Price})$. Three seeds are ensembled per fold; the final model blends this network equally with gradient boosting.

the rural-setting discount all remain significant at the 5% level or better in Model A, as do the LLM condition scale, renovated kitchen, and hardwood indicators), whereas keyword amenities with a strong spatial component lose precision (public-transit proximity falls to the 5% boundary in Model A and is insignificant in Model B; central air conditioning is insignificant in Model B; the curb-appeal score, precisely estimated in Model A, loses significance in the broader Model B). Heteroskedasticity-robust (HC3) standard errors (White, 1980; MacKinnon and White, 1985) yield the same qualitative pattern with uniformly smaller intervals. The last column of Table 7 reports 95% intervals from a pairs cluster bootstrap that resamples the 195 municipalities with replacement; they are wider than the analytical ones, as expected with clusters of very unequal size (MacKinnon and Webb, 2017), and they exclude zero for curb appeal, the renovation level, the luxury cue, the visible garage, brick, stone, vinyl, the rural setting, the LLM condition scale, and the renovated kitchen; the intervals for landscaping and hardwood flooring include zero by a small margin. This asymmetry between property-level and location-level regressors supports our main claim: the AI-generated attributes capture within-market, property-level quality variation rather than repackaged location effects.

Table 8 probes the stability of the predictive blend across subsamples. Three patterns stand out. First, accuracy is comparable whether living area is observed (median error 13.5%) or imputed (14.4%), so the area-free model prices the incomplete listings almost as well as the rich model. Second, the model fits the broker-listed Realtor.ca inventory much better ($R^2_{\text{OOS}} = 0.80$) than the smaller for-sale-by-owner DuProprio segment (0.58), though median errors are identical at 14.0%; the gap is consistent with the compositional and marketing differences between brokered and FSBO platforms documented by Hendel et al. (2009), and with the absence of the room block for DuProprio listings. Third, the median error is stable across price terciles (11–16%), lowest in the mid-market and modestly higher in the tails, where idiosyncratic and luxury attributes are harder to capture. (Within a single tercile R^2 is uninformative because the price range is restricted; the median error is the comparable metric.)

7.2. Language of the Description

Section 3.4 showed that English-language listings are substantially more expensive than French ones. Column (2) of Table 7 adds the two language indicators and the description length to Model A. The English indicator is large and precisely estimated: an English description is associated with a 7.2% higher asking price conditional on every structural, amenity, text, and vision attribute and on the municipality (11.8% in Model B), and the mixed-language indicator with 4.6%. We read this as a location and clientele effect within municipalities—English listings cluster in the most expensive parts of large municipalities such as Montreal, and brokers write in the language of the clientele they target—rather than as a price of the language itself. What matters for our purpose is

Table 7. Robustness of the AI-Derived Implicit Prices (Model A)

	(1) Baseline	(2) +Language	(3) RTA FE +Language	(4) Realtor.ca only RTA FE +Lang.	(5) Cluster bootstrap 95% CI
Curb appeal (1–10)	0.023*** (0.009)	0.022** (0.009)	0.016*** (0.006)	0.021** (0.010)	[0.001, 0.035]
Renovation level (ord.)	0.024** (0.010)	0.024*** (0.009)	0.029*** (0.009)	0.041** (0.018)	[0.010, 0.046]
Luxury (visual)	0.051*** (0.012)	0.051*** (0.012)	0.045*** (0.010)	0.042*** (0.015)	[0.019, 0.066]
Garage visible	0.046*** (0.006)	0.046*** (0.006)	0.037*** (0.006)	0.041*** (0.009)	[0.029, 0.055]
Brick façade	0.033*** (0.009)	0.032*** (0.009)	0.031*** (0.010)	0.024 (0.017)	[0.001, 0.044]
Stone façade	0.036*** (0.008)	0.037*** (0.008)	0.037*** (0.009)	0.030*** (0.010)	[0.022, 0.049]
Vinyl façade	−0.045*** (0.015)	−0.044*** (0.015)	−0.047*** (0.018)	−0.051** (0.021)	[−0.067, −0.007]
Rural setting	−0.122*** (0.032)	−0.120*** (0.033)	−0.122*** (0.042)	−0.119 (0.087)	[−0.176, −0.029]
Landscaping (ord.)	0.015** (0.006)	0.015** (0.006)	0.013** (0.006)	0.027*** (0.007)	[−0.000, 0.025]
Condition (LLM, ord.)	0.016*** (0.006)	0.017*** (0.006)	0.015*** (0.005)	0.012 (0.009)	[0.002, 0.024]
Reno. kitchen (LLM)	0.026*** (0.007)	0.025*** (0.007)	0.019** (0.009)	0.030** (0.012)	[0.010, 0.039]
Hardwood floor (LLM)	0.024** (0.009)	0.026*** (0.009)	0.016** (0.007)	−0.001 (0.013)	[−0.003, 0.036]
English description		0.070*** (0.011)	0.090*** (0.034)	0.082** (0.042)	
Mixed-language description		0.045*** (0.011)	0.067** (0.030)	0.181*** (0.042)	
Fixed effects	Municipality	Municipality	RTA	RTA	Municipality
Observations	9,446	9,446	9,446	5,781	9,446
R ²	0.755	0.757	0.764	0.802	

Notes: OLS estimates of Equation (4) on the measured-area subsample with the full control set (coefficients of the controls not shown). Column (1) is the baseline of Table 3. Column (2) adds indicators for English and mixed-language descriptions and the description length (thousands of characters). Column (3) replaces the 195 municipality fixed effects by 250 forward-sortation-area (RTA) fixed effects (RTAs with at least 20 listings; the remainder pooled). Column (4) restricts the sample to Realtor.ca listings, whose descriptions never contain the asking price. Standard errors in parentheses are clustered at the level of the fixed effects. Column (5) reports the 2.5th and 97.5th percentiles of a pairs cluster bootstrap (299 replications, resampling the 195 municipalities with replacement) for the baseline specification. ***, **, *: significance at 1%, 5%, 10%.

Table 8. Robustness: Blend Performance Across Subsamples

Subsample	N	R^2_{OOS}	Median AE	Within $\pm 15\%$
<i>Full sample</i>	19,123	0.779	14.0%	53%
Price tercile 1 (low)	6,375	n/a [†]	15.6%	49%
Price tercile 2 (mid)	6,412	n/a [†]	11.0%	64%
Price tercile 3 (high)	6,336	n/a [†]	16.4%	47%
Realtor.ca	15,356	0.802	14.0%	53%
DuProprio	3,767	0.579	14.0%	54%
Measured area	9,446	0.773	13.5%	55%
Imputed area	9,677	0.783	14.4%	52%

Notes: Out-of-sample performance of the text + vision blend (Table 5, Panel B) computed on subsamples of the same 5-fold cross-validated predictions. Price terciles are defined on the actual list price. “Measured area” are listings with a reported or room-reconstructed living area (priced by Model A); “Imputed area” listings are priced by the area-free model. [†]Within a single price tercile the range of $\ln P$ is heavily restricted, so R^2 —which normalizes by within-group variance—is uninformative; the median absolute error is the comparable accuracy metric across terciles.

that the AI-derived implicit prices are unchanged to the third decimal when language is controlled for: curb appeal moves from 0.023 to 0.022, the luxury cue stays at 0.051, brick at 0.032–0.033, vinyl at -0.044 , and the LLM condition and kitchen coefficients at 0.017 and 0.025. The language indicators also leave the explanatory gain of the AI layers intact (Panel A of Table 9: $0.719 \rightarrow 0.727 \rightarrow 0.757$ with language controls, against $0.717 \rightarrow 0.725 \rightarrow 0.755$ without).

The gains are present within each language group. Re-estimating the nested ablation of Model A separately on French and English descriptions (Table 9, Panel A), the text layer adds 1.5 points of R^2 for French listings and 0.6 for English ones, and the vision layer 3.4 and 2.2 points; out of sample, gradient boosting gains 1.5 points from the text in both groups and 3.1 and 2.6 from vision (Panel C). The smaller text gain for English listings is consistent with a specification that already fits English listings better (base R^2 of 0.763 against 0.704), so that less is left for the text to explain, rather than with a weaker extraction in English. Adding the language indicators to the gradient-boosting feature set raises its out-of-sample R^2 from 0.770 to 0.774, a modest gain that confirms language is informative but largely redundant with location and the other attributes.

7.3. Information Leakage Through the Text

Because every DuProprio description contains the asking price, the clean test of the text layer is the Realtor.ca subsample. The contrast in Panel A of Table 9 is stark. On Realtor.ca listings the text layer adds 0.6 points of in-sample R^2 to Model A and the vision layer 2.4 points; on DuProprio listings the increments are 3.3 and 4.1 points. Out of sample (Panel C), gradient boosting gains 1.2

Table 9. Gains from the AI Feature Layers by Language, Platform, and Presence of a Price in the Text

Subsample	N	R^2 base	ΔR^2 text (pp)	ΔR^2 vision (pp)
<i>Panel A: In-sample R^2, Model A (measured area), municipality FE</i>				
All listings	9,446	0.717	+0.8	+3.0
French descriptions	2,619	0.704	+1.5	+3.4
English descriptions	4,815	0.763	+0.6	+2.2
Mixed / bilingual	2,012	0.579	+3.2	+4.8
Realtor.ca (no asking price in text)	5,781	0.779	+0.6	+2.4
DuProprio (asking price in text)	3,665	0.564	+3.3	+4.1
Realtor.ca, no dollar amount in text	5,640	0.780	+0.6	+2.4
All listings, RTA fixed effects	9,446	0.722	+0.7	+3.2
<i>Panel B: Out-of-sample R^2 (5-fold, random), OLS, full sample</i>				
All listings	19,123	0.619	+0.8	+5.8
French	8,529	0.622	+1.2	+4.8
English	8,250	0.562	+0.7	+7.9
Realtor.ca	15,356	0.701	+0.6	+3.9
DuProprio	3,767	-0.083	+2.8	+22.8
<i>Panel C: Out-of-sample R^2 (5-fold, random), gradient boosting, full sample</i>				
All listings	19,123	0.729	+1.4	+2.7
French	8,529	0.680	+1.5	+3.1
English	8,250	0.737	+1.5	+2.6
Realtor.ca	15,356	0.748	+1.2	+2.6
DuProprio	3,767	0.564	+3.1	+2.8

Notes: Panel A re-estimates the nested in-sample ablation of Model A (structural, amenity, room, and listing controls with municipality and type fixed effects; then + LLM text layer; then + façade vision layer) on each subsample; “ ΔR^2 ” is the gain, in percentage points, from adding the layer. Language is classified from French and English stop-word counts (ratio rule of 1.5; “mixed” listings are mostly DuProprio descriptions that concatenate an English translation and the French original). Every DuProprio description contains the asking price (“Asking Price \$...”); 2.5% of Realtor.ca descriptions contain some dollar amount (usually a municipal assessment, a renovation budget, or a tax; the 141 Model A listings concerned are too few for a separate regression). Panels B and C evaluate the full-sample cross-validated predictions (Table 6, identical folds) within each subsample; the base is the structural + amenity + rooms + listing specification. Within-subsample R^2 values can be negative (DuProprio, OLS) when the pooled model is miscalibrated for the subsample.

points from the text on Realtor.ca and 3.1 on DuProprio. Three remarks temper the comparison. First, the DuProprio base is much weaker (in-sample R^2 of 0.564 against 0.779) because the room block is missing, so every additional block has more left to explain there; the vision layer, which never sees any text, also gains more on DuProprio (4.1 against 2.4 points), and the ratio of the DuProprio to the Realtor.ca gain is about 1.7 for vision against 5 for text. Second, the DuProprio “description” also embeds structured fields (cladding, heating, inclusions) that the LLM can read verbatim and that are genuinely informative, so part of the excess text gain is legitimate. Third, the asking price itself cannot be reproduced by any field of the schema; what the price can do is tilt the model’s reading of ambiguous fields (condition, turnkey, luxury-type highlights) towards the expensive or the cheap end. We cannot separate these channels without re-running the extraction on descriptions purged of the price, which we recommend as the first step of any replication and which the archived raw text makes possible at a cost of about \$1. Until then, we regard the Realtor.ca increments—0.6 points of in-sample R^2 and 0.6 (OLS) to 1.2 (gradient boosting) points out of sample for the text layer—as the credible estimate of what the LLM specifications add beyond keyword amenities, and we note that none of the paper’s conclusions about the vision layer is affected, since the vision model is exposed neither to the text nor to the price. Column (4) of Table 7 re-estimates Model A on Realtor.ca listings alone with RTA fixed effects and language controls: curb appeal (0.021), the renovation level (0.041), the luxury cue (0.042), the visible garage (0.041), stone (0.030), vinyl (−0.051), landscaping (0.027), and the renovated kitchen (0.030) retain their magnitude and significance, whereas brick, the rural setting, the LLM condition scale, and hardwood flooring keep their sign but lose significance in this smaller sample of 5,781 listings.

7.4. Finer Location Controls and Spatial Validation

Municipalities are coarse spatial units: Montreal, Quebec City, Laval, and Gatineau each contain neighbourhoods whose price levels differ by a factor of two or more. Column (3) of Table 7 replaces the 195 municipality effects by 250 RTA effects. In-sample fit improves (Model A R^2 from 0.755 to 0.764, Model B from 0.725 to 0.726), and the AI coefficients are stable: curb appeal 0.016, renovation level 0.029, luxury 0.045, garage 0.037, brick 0.031, stone 0.037, vinyl −0.047, rural −0.122, LLM condition 0.015, renovated kitchen 0.019, hardwood 0.016, all significant at the 5% level with standard errors clustered by RTA. The in-sample increments from the two AI layers under RTA effects are 0.7 and 3.2 points (Table 9, last row of Panel A). Finer location controls therefore do not absorb the AI-measured quality, which is what one expects if the layers measure property-level rather than neighbourhood-level attributes.

Random cross-validation folds let every held-out listing borrow strength from neighbours in the training folds. Table 10 repeats the nested ablation with folds formed by RTA, so that an entire postal area is predicted from a model that has never seen it. All levels fall, as they must:

the full text + vision OLS drops from 0.686 to 0.638 and gradient boosting from 0.770 to 0.734, with median absolute errors rising by about 1.5 percentage points. The increments from the AI layers, however, are preserved almost exactly—the text layer adds 0.7 (OLS) and 1.3 (gradient boosting) points under spatial folds against 0.9 and 1.4 under random folds, and the vision layer 7.0 and 3.1 points against 5.8 and 2.7. If anything, the vision layer is *more* valuable when location is poorly measured, which is consistent with the façade carrying information about the neighbourhood (density, setting, housing vintage) as well as about the house. We regard the random-fold numbers as the relevant ones for an AVM operating within its coverage area, and the spatial-fold numbers as the relevant ones for extrapolation to new areas.

Table 10. Random versus Spatial (RTA-Grouped) Cross-Validation

Estimator	Random 5-fold				Spatial 5-fold (RTA groups)			
	Base	+Text	+Vision	MdAPE	Base	+Text	+Vision	MdAPE
OLS (linear hedonic)	0.619	0.628	0.686	16.9%	0.561	0.568	0.638	18.5%
Gradient boosting	0.729	0.743	0.770	14.5%	0.690	0.703	0.734	15.9%

Notes: Out-of-sample R^2 on $\ln(\text{Price})$ and median absolute percentage error (MdAPE, for the full text + vision specification). “Base” is the structural + amenity + rooms + listing specification; the next two columns add the LLM text layer and the façade vision layer. In the spatial design the 420 forward sortation areas (RTA, first three characters of the postal code) are assigned to five folds by GroupKFold, so no listing is predicted from a model that has seen a listing of the same RTA; municipality indicators of RTAs held out in their entirety are unobserved in training and the model falls back on the reference category. Random-fold entries may differ by ± 0.001 from Table 6, which comes from a separate run. $N = 19,123$.

7.5. Limitations

Seven limitations qualify the interpretation of our estimates.

Asking prices. The dependent variable is the asking price of properties listed for sale, not the transaction price. The list price is a strategic choice that anchors buyer search and bargaining (Horowitz, 1992; Knight, 2002; Han and Strange, 2015, 2016): it is set with sale in mind and is highly correlated with the eventual transaction price, but list-to-sale ratios vary systematically with market conditions and overpricing is more common in thin segments. Two implications follow. First, our implicit prices measure how sellers and their agents *price* attributes into the ask, which converges to buyers’ valuation only insofar as the market disciplines the ask; the economically sensible magnitudes and the monotone validation of Section 5.3 suggest the discipline is substantial. Second, any attribute that shifts the list-to-sale ratio itself (e.g., attributes of hard-to-sell properties) will carry a coefficient that mixes valuation and listing strategy. Replicating the analysis on registry transaction prices is the natural next step.

Price in the text. For the 19.7% of listings from DuProprio, the language model read the asking

price as part of the description. The text-layer gains on that platform are therefore an upper bound, and the Realtor.ca gains the credible estimate (Section 7.3). The vision layer is unaffected.

Non-causality and content endogeneity. Descriptions and photographs are marketing instruments chosen by sellers and agents. Richer descriptions, more photographs, and higher visual quality correlate with higher-end listing agents and more valuable properties (Nowak and Smith, 2017). The coefficients are therefore *predictive* associations, not causal effects; they should not be read as the price change from physically adding a given attribute. We control for two observable dimensions of marketing effort (photo count and virtual tours) and for the photo quality score, but cannot rule out that residual marketing quality loads on the AI-derived attributes, and the three visual quality scores are collinear enough that only their joint contribution is well identified.

Machine-generated features and their reproducibility. LLM and vision outputs contain errors. Strict schemas and closed enumerations bound the *form* of the output but not its accuracy, and the agreement analysis of Section 5.3 shows that accuracy varies widely across fields: high for concrete amenities, low for waterfront, the visual setting, and the visual building type. We have no human-labelled validation set; a few hundred manually coded listings would allow field-level precision and recall against ground truth and are the obvious next step. Classical measurement error in the extracted scores biases their coefficients towards zero, which makes our implicit prices conservative, but errors correlated with price (through the leakage channel above, or because the model rates expensive-looking houses more generously on every dimension) need not. The text extraction ran at temperature zero, the vision extraction at the API default temperature, and both model identifiers are provider aliases rather than pinned snapshots; the raw outputs are archived, so the regressors are exactly reproducible from them, but a fresh extraction with a later model version could return different scores, and we have not measured test–retest variability. The vision layer judges each property from a single façade photograph and ignores interiors.

Language. The LLM processes French and English descriptions with the same French-language instruction and schema. We find no evidence that the extraction is weaker in either language and the implicit prices are robust to language controls, but the English indicator itself carries a large conditional premium that we can interpret only as a proxy for unobserved location and clientele.

Selection. Listed properties are not a random sample of the housing stock, and listings with missing area or room detail are not missing at random (they tend to be less complete listings, and all DuProprio listings lack the room block). The “has_” indicators and the two-part design mitigate, but do not eliminate, the second concern, and the markedly worse fit for DuProprio shows that the pooled model is not equally adequate on both platforms.

Cross-section. The data are a single-date snapshot, so the model captures no market dynamics, no forward-in-time holdout is possible, and the location premia are fixed as of 3 July 2026. Municipalities with fewer than 20 listings are pooled, reducing spatial resolution where data are sparse,

and some municipality effects in Model A rest on very few observations (Section 5.4).

8. Conclusion

We have estimated a hedonic price model for Quebec houses that augments the standard set of structural, locational, and amenity attributes with two machine-generated feature layers: structured specifications extracted from free-text descriptions by a large language model, and visual attributes extracted from façade photographs by a vision–language model. Schema-constrained decoding renders these outputs as ordinary, auditable econometric variables, obtained at a marginal cost of a fraction of a cent per listing. The central message is methodological: general-purpose AI models, disciplined by strict output schemas, function as a low-cost feature-engineering technology that measurably expands the characteristics set available to hedonic analysis—and the discipline must extend to the input, since the text handed to the model may contain the very price one is trying to explain.

The gains are economically meaningful but unequal across the two layers. Together they add 3.8 points of in-sample R^2 to an already rich log-linear specification (0.717 to 0.755 with living area observed), of which 3.0 come from the façade photograph and 0.8 from the text; they raise the out-of-sample accuracy of every estimator in our ladder, from OLS to gradient boosting and an embedding-based residual neural network; and the best blend attains an out-of-sample R^2 of 0.779 with a median absolute percentage error of 14.0% under random folds, 0.734 for gradient boosting when entire postal areas are held out. The vision layer adds more to the linear model than switching from OLS to a random forest does. The extracted attributes carry interpretable implicit prices—curb appeal, renovation level, luxury cues, and exterior materials all price sensibly—that survive municipality-clustered inference, a cluster bootstrap, finer postal-code fixed effects, and a control for the language of the description. The text layer’s contribution, by contrast, is modest once keyword amenities are already in the model, and part of its measured gain on the for-sale-by-owner platform reflects the asking price embedded in those descriptions rather than information about the house.

Three extensions are natural. First, the same schema-constrained approach scales from a single façade image to a small panel of interior photographs (kitchen, bathroom, living space), which our design already accommodates; a repeat extraction of a subsample at a fixed temperature, and a human-labelled validation set, would also quantify the extraction error that we could only measure indirectly. Second, the DuProprio coordinates and the postal codes of all listings permit continuous spatial features—distance to the urban core, to water, or to amenities—that would refine the geography beyond discrete area effects. Third, and most important for external validity, linking listings to registry transaction prices would establish how much of the AI-measured quality premium

survives bargaining. For appraisal, automated valuation, and property-tax assessment—applications that consume hedonic models at scale—the practical implication is that the unstructured majority of a listing’s information content is no longer out of reach, provided that it is read with the same care as any other regressor.

Declarations

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest. The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability. The listing data were collected from publicly accessible web platforms (Realtor.ca and DuProprio) on 3 July 2026. The full data-construction code, the extraction schemas, and the derived analysis dataset (excluding copyrighted listing text and images), together with the raw JSON outputs of the two AI models, are available from the author upon reasonable request. The robustness analyses added in version 2.1 are produced by a single script (`scripts/18_revision_language_leakage_spatial.py`) from the archived extractions, without new API calls.

Use of AI models. The large language model `gpt-4o-mini` and the vision–language model `claude-haiku-4-5` are the object of study of this paper and were used exclusively as documented in Section 4 to extract structured features from listing text and photographs. They were not used to generate the analysis or the conclusions of the paper. An AI coding assistant was used in the October 2026 revision, under the author’s direction, to check the manuscript for internal consistency and to run the additional robustness analyses; all results were verified by the author.

References

- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge**, “When should you adjust standard errors for clustering?,” *Quarterly Journal of Economics*, 2023, 138 (1), 1–35.
- 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 (NCTA)” 2016, pp. 62–68.
- Anselin, Luc**, *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer Academic Publishers, 1988.
- Bessinger, Zachary and Nathan Jacobs**, “Quantifying curb appeal,” in “2016 IEEE International Conference on Image Processing (ICIP)” 2016, pp. 4388–4392.
- Bin, Okmyung**, “A prediction comparison of housing sales prices by parametric versus semi-parametric regressions,” *Journal of Housing Economics*, 2004, 13 (1), 68–84.
- Boucher, Simon-Pierre**, “Decoding real estate descriptions: Semantic embeddings and hedonic pricing of residential properties in Quebec,” UQO Working Paper 2, Université du Québec en Outaouais, Département des sciences administratives 2026.
- Bourassa, Steven C., 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.
- Breiman, Leo**, “Random forests,” *Machine Learning*, 2001, 45 (1), 5–32.
- Cameron, A. Colin and Douglas L. Miller**, “A practitioner’s guide to cluster-robust inference,” *Journal of Human Resources*, 2015, 50 (2), 317–372.
- Can, Ayse**, “Specification and estimation of hedonic housing price models,” *Regional Science and Urban Economics*, 1992, 22 (3), 453–474.
- Dell, Melissa**, “Deep learning for economists,” *Journal of Economic Literature*, 2025, 63 (1), 5–58.
- Duan, Naihua**, “Smearing estimate: a nonparametric retransformation method,” *Journal of the American Statistical Association*, 1983, 78 (383), 605–610.
- Friedman, Jerome H.**, “Greedy function approximation: a gradient boosting machine,” *Annals of Statistics*, 2001, 29 (5), 1189–1232.

- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy**, “Text as data,” *Journal of Economic Literature*, 2019, 57 (3), 535–574.
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli**, “ChatGPT outperforms crowd workers for text-annotation tasks,” *Proceedings of the National Academy of Sciences*, 2023, 120 (30), e2305016120.
- Glaeser, Edward L., Michael Scott Kincaid, and Nikhil Naik**, “Computer vision and real estate: Do looks matter and do incentives determine looks,” Working Paper 25174, National Bureau of Economic Research 2018.
- Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko**, “Revisiting deep learning models for tabular data,” in “Advances in Neural Information Processing Systems (NeurIPS) 34” 2021. arXiv:2106.11959.
- Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux**, “Why do tree-based models still outperform deep learning on tabular data?,” in “Advances in Neural Information Processing Systems (NeurIPS) 35” 2022.
- Guo, Cheng and Felix Berkhahn**, “Entity embeddings of categorical variables,” *arXiv preprint*, 2016, *arXiv:1604.06737*.
- Halvorsen, Robert and Raymond Palmquist**, “The interpretation of dummy variables in semilogarithmic equations,” *American Economic Review*, 1980, 70 (3), 474–475.
- Han, Lu and William C. Strange**, “The microstructure of housing markets: Search, bargaining, and brokerage,” in “Handbook of Regional and Urban Economics,” Vol. 5, Elsevier, 2015, pp. 813–886.
- **and** —, “What is the role of the asking price for a house?,” *Journal of Urban Economics*, 2016, 93, 115–130.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun**, “Deep residual learning for image recognition,” in “IEEE Conference on Computer Vision and Pattern Recognition (CVPR)” 2016, pp. 770–778.
- Hendel, Igal, Aviv Nevo, and François Ortalo-Magné**, “The relative performance of real estate marketing platforms: MLS versus FSBOMadison.com,” *American Economic Review*, 2009, 99 (5), 1878–1898.
- Hill, Robert J.**, “Hedonic price indexes for residential housing: A survey, evaluation and taxonomy,” *Journal of Economic Surveys*, 2013, 27 (5), 879–914.

- Horowitz, Joel L.**, “The role of the list price in housing markets: Theory and an econometric model,” *Journal of Applied Econometrics*, 1992, 7 (2), 115–129.
- Johnson, Erik B., Alan Tidwell, and Sriram V. Villupuram**, “Valuing curb appeal,” *Journal of Real Estate Finance and Economics*, 2020, 60 (1-2), 111–133.
- Kang, Yuhao, Fan Zhang, Wenzhe Peng, Song Gao, Jinhong 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.
- Kennedy, Peter E.**, “Estimation with correctly interpreted dummy variables in semilogarithmic equations,” *American Economic Review*, 1981, 71 (4), 801.
- Knight, John R.**, “Listing price, time on market, and ultimate selling price: Causes and effects of listing price changes,” *Real Estate Economics*, 2002, 30 (2), 213–237.
- Kok, Nils, Eija-Leena Koponen, and Carmen Adriana Martínez-Barbosa**, “Big data in real estate? From manual appraisal to automated valuation,” *Journal of Portfolio Management*, 2017, 43 (6), 202–211.
- Korinek, Anton**, “Generative AI for economic research: Use cases and implications for economists,” *Journal of Economic Literature*, 2023, 61 (4), 1281–1317.
- Lancaster, Kelvin J.**, “A new approach to consumer theory,” *Journal of Political Economy*, 1966, 74 (2), 132–157.
- Law, Stephen, Brooks Paige, and Chris Russell**, “Take a look around: Using street view and satellite images to estimate house prices,” *ACM Transactions on Intelligent Systems and Technology*, 2019, 10 (5), 1–19.
- Lindenthal, Thies and Erik B. Johnson**, “Machine learning, architectural styles and property values,” *Journal of Real Estate Finance and Economics*, 2021, 71 (3), 353–384.
- Loshchilov, Ilya and Frank Hutter**, “Decoupled weight decay regularization,” in “International Conference on Learning Representations (ICLR)” 2019. arXiv:1711.05101.
- MacKinnon, James G. and Halbert White**, “Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties,” *Journal of Econometrics*, 1985, 29 (3), 305–325.
- **and Matthew D. Webb**, “Wild bootstrap inference for wildly different cluster sizes,” *Journal of Applied Econometrics*, 2017, 32 (2), 233–254.

- Malpezzi, Stephen**, “Hedonic pricing models: a selective and applied review,” in Tony O’Sullivan and Kenneth Gibb, eds., *Housing Economics and Public Policy*, Blackwell Science, 2003, pp. 67–89.
- Mullainathan, Sendhil and Jann Spiess**, “Machine learning: an applied econometric approach,” *Journal of Economic Perspectives*, 2017, 31 (2), 87–106.
- Naik, Nikhil, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, and César A. Hidalgo**, “Computer vision uncovers predictors of physical urban change,” *Proceedings of the National Academy of Sciences*, 2017, 114 (29), 7571–7576.
- Nowak, Adam and Patrick S. Smith**, “Textual analysis in real estate,” *Journal of Applied Econometrics*, 2017, 32 (4), 896–918.
- Nowak, Adam D. and Patrick S. Smith**, “Quality-adjusted house price indexes,” *American Economic Review: Insights*, 2020, 2 (3), 339–356.
- , **Bradley S. Price, and Patrick S. Smith**, “Real estate dictionaries across space and time,” *Journal of Real Estate Finance and Economics*, 2021, 62 (1), 139–163.
- Pace, R. Kelley and Darren Hayunga**, “Examining the information content of residuals from hedonic and spatial models using trees and forests,” *Journal of Real Estate Finance and Economics*, 2020, 60 (1–2), 170–180.
- Poursaeed, Omid, Tomas Matera, and Serge Belongie**, “Vision-based real estate price estimation,” *Machine Vision and Applications*, 2018, 29 (4), 667–676.
- Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann**, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography*, 2017, 40 (8), 913–929.
- Rosen, Sherwin**, “Hedonic prices and implicit markets: product differentiation in pure competition,” *Journal of Political Economy*, 1974, 82 (1), 34–55.
- Shen, Lily and Stephen L. Ross**, “Information value of property description: A machine learning approach,” *Journal of Urban Economics*, 2021, 121, 103299.
- Sirmans, Stacy G., David A. Macpherson, and Emily N. Zietz**, “The composition of hedonic pricing models,” *Journal of Real Estate Literature*, 2005, 13 (1), 1–44.

White, Halbert, “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity,” *Econometrica*, 1980, 48 (4), 817–838.

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

Zhang, Yonglin and Rencai Dong, “Impacts of street-visible greenery on housing prices: Evidence from a hedonic price model and a massive street view image dataset in Beijing,” *ISPRS International Journal of Geo-Information*, 2018, 7 (3), 104.

Appendices

A. Machine-Generated Feature Definitions

Table [A1](#) lists the fields extracted by the language and vision models under their respective strict JSON schemas, with the value domains as specified in the extraction scripts (the enumerations are in French in the schema; they are translated here). All fields use closed enumerations, integers, or booleans; absent information is coded as false, unknown, or an empty list. The mapping from fields to regressors is described in Section [4.5](#).

B. Sample Construction Flow

C. Language Classification

Each description is classified from the counts e and f of English and French stop words it contains (articles, prepositions, pronouns, and auxiliaries such as *the*, *and*, *with*, *is* and *le*, *la*, *les*, *et*, *avec*, *est*). A description is “English” if $e \geq 1.5f$, “French” if $f \geq 1.5e$, and “mixed” otherwise or when neither count is positive. On Realtor.ca the mixed class is small (2.6%) and consists mostly of very short descriptions; on DuProprio it is large (51.7%) because the crawler returns the automated English translation followed by the original French comments. Table [9](#) reports the gains from the AI layers within each class, and Table [7](#) the effect of adding the language indicators to the hedonic regression. The language variable is produced by `scripts/18_revision_language_leakage_spatial.py`.

Table A1. Fields Extracted by the AI Models

Layer	Field	Domain
<i>Panel A: Text (LLM, gpt-4o-mini; 20 fields, 18 regressors)</i>		
Condition	condition	to renovate / original / good / renovated / like-new / new / unknown
Turnkey; recently renovated	turnkey, recently_renovated	boolean
Renovated kitchen / bathroom	renovated_kitchen, renovated_bathroom	boolean
Flooring (multi-valued)	flooring	hardwood / ceramic / floating / carpet / vinyl / concrete / other
Countertop	countertop	granite / quartz / marble / laminate / wood / other / unknown
Heating (multi-valued)	heating	electric / baseboards / heat pump / wood / gas / oil / radiant / geothermal / other
View	view	water / mountain / city / none / unknown
Amenities	garage, pool, waterfront, finished_basement, fireplace, central_ac	boolean
New construction; income property	new_construction, income_property	boolean
Large lot; quiet area	large_lot, quiet_area	boolean
Highlights	highlights	3–6 free-text keywords (not used in the regressions)
<i>Panel B: Vision (claude-haiku-4-5, façade; 12 fields, 21 regressors)</i>		
Exterior shot	is_exterior	boolean
Building type (visual)	building_type_visual	detached / semi-detached / row / condo building / cottage / land / other
Condition score	condition_score	integer 1–10
Curb appeal	curb_appeal	integer 1–10
Renovation level	renovation_level	to renovate / dated / standard / renovated / high-end / unknown
Exterior materials (multi-valued)	exterior_materials	brick / stone / vinyl / wood / aluminum / stucco / roughcast / concrete / other
Landscaping	landscaping	none / minimal / maintained / manicured
Garage visible	has_garage_visible	boolean
Storeys (visual)	storeys_visual	1 / 1.5 / 2 / 3+ / unknown
Setting	setting	dense urban / suburban / rural / waterfront / wooded / unknown
Photo quality	photo_quality	integer 1–10
Luxury indicator	curb_luxury	boolean

Notes: Both extractions use schema-constrained decoding with closed enumerations, required fields, and `additionalProperties:false`. The text layer reads the free-text description (truncated at 4,000 characters) at temperature zero; the vision layer analyzes one façade photograph per listing at the API’s default temperature. Model identifiers are the API aliases current in July 2026. Fields are mapped to numeric regressors (booleans to 0/1, ordinal scales to integers, multi-valued fields and categories to indicators); the text condition scale is coded 0–5 with “unknown” at 2, the visual renovation level 0–4 with “unknown” at 2, and landscaping 0–3.

Table A2. From Raw Records to Estimation Sample

Step	Records
Realtor.ca raw	27,302
DuProprio raw	7,693
Houses (single-family + townhouse) with price	19,273
After outlier cleaning (price bounds, price-per-ft ² trim)	19,123
of which Realtor.ca / DuProprio	15,356 / 3,767
of which with measured living area (Model A)	9,446
of which with a detected exterior photograph	18,314

Notes: See Section 3. “Measured living area” means reported area or area reconstructed from room-level dimensions. Both AI layers cover all 19,123 cleaned listings.