Université du Québec en Outaouais

Département des sciences administratives

Working Paper No. 2

Decoding Real Estate Descriptions: Semantic Embeddings and Hedonic Pricing of Residential Properties in Quebec

Simon-Pierre Boucher

Département des sciences administratives
Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

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

October 2026

Version 1.2

Abstract

Listing descriptions written by real estate brokers contain information about housing quality that structured hedonic variables omit. We propose an interpretable way to use that information: sentence-transformer embeddings of 17,087 single-family listings from Quebec are projected, by cosine similarity, onto 20 researcher-defined reference descriptions (luxury, needs renovation, waterfront, motivated seller, and so on), and the 20 similarity scores enter a semi-logarithmic hedonic model as named regressors. Adding the semantic block to a structural model raises the adjusted R^2 from 0.452 to 0.511, a gain confirmed out of sample in random and spatially grouped cross-validation. Three qualifications follow. First, 40% of the descriptions are written in English and score systematically lower against the French references; the language of the listing alone explains 86% of the variance of the mean similarity and accounts for more than half of the apparent semantic gain. Net of language and of 415 forward-sortation-area fixed effects, the semantic block adds 2.0 percentage points of R^2 with the baseline encoder and 3.9 points with a multilingual encoder, and remains jointly significant with standard errors clustered by area. Second, the individual implicit price gradients are sensitive to the phrasing of the references and to the encoder: only a core of dimensions—Luxury, Land & Nature, Family-Friendly, Quiet & Peaceful, Needs Renovation—keeps its sign and significance across specifications, while others, including Motivated Seller, New Construction and Entry-Level, do not. Third, the coefficients are conditional associations in listing prices, not causal effects of wording. The paper documents what reference-based projection can and cannot deliver for hedonic inference and provides a replication pipeline.

Keywords: hedonic pricing, text as data, sentence embeddings, cosine similarity, listing descriptions, real estate valuation, Quebec

JEL Classification: R31, C45, C21, R21

1. Introduction

Hedonic pricing models have been the dominant empirical framework for estimating the implicit prices of housing characteristics since Lancaster (1966) and Rosen (1974), building on a quality-adjustment tradition that dates back to the earliest hedonic price indexes (Griliches, ed, 1971). The standard approach decomposes observed prices into the marginal contributions of structural attributes (bedrooms, bathrooms, lot size), locational factors (neighbourhood quality, proximity to amenities) and environmental characteristics (Sirmans et al., 2005). Decades of empirical work have refined these models, establishing which structural and spatial variables matter most (Sirmans et al., 2006; Malpezzi, 2003), clarifying the conditions under which hedonic estimates identify preference parameters (Ekeland et al., 2004; Kuminoff et al., 2013), and embedding the framework in the operational infrastructure of property taxation, mass appraisal and automated valuation models (Kok et al., 2017).

A persistent limitation of hedonic models is that they can only price attributes that the econometrician observes. Property listing descriptions—the free-text narratives written by real estate brokers to market properties—contain qualitative information that structured fields do not record: the quality of finishes, the setting of the lot, the motivation of the seller, the architectural character of a home. A listing that describes a "spa-like bathroom with heated marble floors" conveys information that a bathroom count does not, and two properties with identical structural records can differ enormously in the quality the text reveals. Pakes (2003) and Bajari and Benkard (2005) show that unobserved characteristics bias the implicit prices of the observed ones; listing narratives offer a direct and scalable measurement of some of the missing dimensions.

The opportunity to exploit this information has grown with the text-as-data movement in empirical economics (Gentzkow et al., 2019; Ash and Hansen, 2023), which has turned unstructured text into a standard measurement input in finance (Loughran and McDonald, 2016; García et al., 2023), monetary policy (Hansen et al., 2017) and labour markets (Marinescu and Wolthoff, 2020). The housing applications to date face a trade-off between semantic depth and interpretability. Keyword and vernacular methods (Haag et al., 2000; Nowak and Smith, 2017; Goodwin et al., 2014; Nowak et al., 2021) are interpretable but lexically brittle: they cannot recognise that "quartz countertops" and "premium stone surfaces" convey similar information. Sentiment and connotation measures (Goodwin et al., 2018; Zhu et al., 2023) compress multidimensional quality descriptions into one or two polarity scores. Topic models (Blei et al., 2003) capture latent themes whose content is difficult to name ex ante and unstable in short, formulaic texts. Machine learning pipelines built on rich text representations (Shen and Ross, 2021; Baur et al., 2023) extract substantially more price-relevant information, but their features are opaque—useful for prediction, unsuitable for inference about identifiable housing characteristics.

This paper proposes and evaluates a *concept-projection* approach to text-augmented hedonic pricing. Each listing's *PublicRemarks* field is encoded into a 384-dimensional vector with a sentence transformer (Reimers and Gurevych, 2019; Wang et al., 2020); 20 short reference descriptions, each written to embody one qualitative housing dimension, are encoded with the same model; the cosine similarity between a listing and each reference yields 20 named scalar features; and the features enter a semi-logarithmic hedonic model estimated by least squares with heteroskedasticity-robust standard errors. In the taxonomy of text-as-data methods, the procedure is a semantic generalisation of the dictionary method (Gentzkow et al., 2019; Loughran and McDonald, 2016): each reference plays the role of a word list, but the matching is performed in embedding space, so paraphrases and synonyms load on the intended dimension. The objective is not to claim that changing listing words changes prices, but to measure whether, and how much, the semantic content of listing narratives carries price-relevant information omitted from structured hedonic variables—and how far the individual projections can be trusted as implicit prices.

Quebec provides an informative setting. The provincial market spans luxury urban properties in Montréal and Québec City, suburban family homes, and waterfront properties in the Laurentians and the Eastern Townships. The listing corpus is bilingual: 58% of the descriptions in our sample are written in French and 40% in English, which creates an opportunity to test the cross-lingual behaviour of embedding-based features and, as we show, a confound that must be controlled. A well-established hedonic tradition using Quebec microdata (Des Rosiers et al., 2002; Dubé and Legros, 2014) supplies local benchmarks for the amenity values that text-based measures should recover.

The paper makes three contributions. First, it establishes that the semantic block carries information beyond structural variables. For 17,087 single-family listings, adding the 20 similarities and description length to a structural hedonic model raises the adjusted R^2 from 0.452 to 0.511, and the gain is confirmed out of sample: in five-fold cross-validation with random folds and with folds that hold out entire forward sortation areas, the semantic features raise the out-of-sample R^2 by about four percentage points and lower the median absolute percentage error of the price prediction. Second, it quantifies how much of this information is genuinely semantic. Because the references are written in French and 40% of the listings are in English, the language of the listing explains 86% of the variance of the mean similarity score and accounts for more than half of the apparent semantic gain; net of language indicators and of 415 forward-sortation-area fixed effects, with standard errors clustered by area, the block adds 2.0 percentage points of R^2 with the baseline encoder and 3.9 points with a multilingual encoder, and remains jointly significant in every specification and in each language subsample. Third, it evaluates the interpretability claim that motivates the method. The individual implicit price gradients depend on the phrasing of the references and on the encoder: perturbing the reference texts leaves the similarity scores almost unchanged (correlations

of 0.93–0.99) but can reorder the coefficients, and switching to a multilingual encoder reverses several signs. A core of dimensions—Luxury, Land & Nature, Family-Friendly, Quiet & Peaceful and Needs Renovation—keeps its sign and significance across all specifications and both encoders; others, including Motivated Seller, New Construction and Entry-Level, do not. We therefore read the semantic coefficients as conditional associations within a correlated block of quality signals, robust at the level of the block and of a few well-separated dimensions, and we discuss what this implies for the use of reference-based projections in appraisal and research.

Relative to the first version of this paper, the present revision adds the language analysis, fixed effects and clustered inference, a unit-corrected lot-size measure, a floor-area subsample, domain-level block tests, out-of-sample validation, a reference-phrasing sensitivity analysis and a multilingual encoder. All headline numbers of the first version are reproduced and retained as the baseline; the new analyses qualify their interpretation rather than replace them.

The remainder of the paper is organised as follows. Section 2 reviews the four literatures on which this work draws. Section 3 describes the method, the estimating equations and the scope of inference. Section 4 presents the data, including the language composition of the corpus. Section 5 reports the baseline results, Section 6 the robustness checks of the first version, and Section 7 the extended specifications and validation exercises added in this revision. Section 8 interprets the findings against the literature and states their limitations. Section 9 concludes.

2. Literature Review

This paper sits at the intersection of four literatures: the hedonic pricing tradition in housing economics, the text-as-data program in empirical economics, the emerging body of work on unstructured data in real estate, and the natural language processing literature on sentence embeddings. We review each in turn and then articulate the gap this paper fills.

2.1. Hedonic Pricing Theory and Housing Markets

The idea that the price of a differentiated good can be decomposed into the implicit prices of its characteristics predates modern housing economics: hedonic price indexes were computed for automobiles as early as the 1930s, and the approach was systematised for quality-adjusted price measurement by Griliches, ed (1971). Its theoretical foundations rest on Lancaster (1966), who reconceptualised consumer theory around the characteristics of goods rather than goods themselves, and Rosen (1974), who formalised the framework for differentiated product markets. In Rosen's model, the observed market price of a differentiated good reflects the equilibrium of supply and demand for its constituent characteristics, allowing the researcher to recover the implicit marginal prices of individual attributes through regression analysis. Palmquist (1984) extended the theoretical

framework by showing how second-stage supply and demand identification depends on exogenous variation in household characteristics and builder cost structures, while Epple (1987) demonstrated that under certain conditions the hedonic price function is unique and recoverable. The modern identification literature has clarified both the power and the limits of the approach: Ekeland et al. (2004) establish conditions under which preferences can be recovered from a single hedonic market, Bajari and Benkard (2005) develop estimators that accommodate unobserved product characteristics—the very problem that listing text addresses in our setting—and Kuminoff et al. (2013) survey how equilibrium sorting across housing markets shapes the interpretation of hedonic estimates. This literature underpins the interpretive caution we adopt in Section 3.8: cross-sectional hedonic coefficients are equilibrium price gradients, not structural willingness-to-pay parameters.

In the housing context, the hedonic approach yields the canonical specification:

$$\ln(P_i) = \alpha + \boldsymbol{\beta}'\mathbf{X}_i + \varepsilon_i \tag{1}$$

where P_i is the price of property i, $\mathbf{X}_i$ is a vector of structural, locational, and neighbourhood attributes, and $\boldsymbol{\beta}$ captures the implicit marginal prices. The semi-logarithmic functional form, recommended by Halvorsen and Pollakowski (1981) and widely adopted in the literature, allows coefficients to be interpreted as approximate percentage changes in price for a unit change in the characteristic. Cropper et al. (1988) evaluated alternative functional forms through Monte Carlo experiments, finding that simpler parametric specifications (linear and log-linear) outperform flexible Box-Cox transformations when variables are measured with error or important attributes are omitted.

The empirical literature on hedonic housing models is vast. Sirmans et al. (2005) surveyed the composition of hedonic models and identified bedrooms, bathrooms, lot size, age, square footage, and proximity to central business districts as the most frequently included variables. Sirmans et al. (2006) conducted a meta-analysis of 64 hedonic studies, finding that bathrooms, square footage, and lot size consistently exhibit the largest marginal effects. Malpezzi (2003) provided a comprehensive review of methodological issues, including functional form selection, spatial autocorrelation, and multicollinearity. Two extensions of the baseline framework are particularly relevant here. First, *quantile hedonics*: Zietz et al. (2007) show that the implicit prices of housing attributes vary systematically across the conditional price distribution, with many characteristics valued more highly in upper price segments, and McMillen (2008) uses quantile methods to decompose changes in the entire distribution of house prices. Our quantile analysis in Section 6 extends this logic to text-derived characteristics. Second, *spatial hedonics*: housing data are spatially dependent, and a long literature—Anselin (1988), Dubin (1988), and Can (1992) among the foundations—has developed specifications that absorb spatial autocorrelation in prices and errors; Dubé and Legros (2014) adapt these methods to the Canadian and Quebec microdata context. In applied work the

most common responses to spatial dependence are area fixed effects and standard errors clustered at the level at which the sampling or the unobserved shocks are structured (Cameron and Miller, 2015; Abadie et al., 2023), and the validation of predictive models on spatially dependent data requires held-out folds that respect that structure (Roberts et al., 2017); we adopt all three in Section 7. Closer to our setting still, a sustained research program using Quebec City transactions has quantified the capitalisation of property attributes ranging from landscaping to accessibility (Des Rosiers et al., 2002). In the amenity-capitalisation literature more broadly, Cheshire and Sheppard (2004) demonstrate school-quality capitalisation in local housing markets and Gibbons (2015) documents the discount associated with visual environmental disamenities, while Irwin (2002) and Tyrväinen et al. (2005) establish the positive value of open space and urban vegetation.

A persistent challenge in hedonic pricing is the omitted variable bias arising from unobserved quality attributes. Pakes (2003) argued that standard hedonic regressions yield biased implicit price estimates when important product characteristics are unobservable to the econometrician but observable to market participants. This concern is directly relevant to our study: listing descriptions may convey quality information—about finish materials, maintenance history, or neighbourhood ambiance—that is not captured by conventional structural variables but is fully observable to buyers.

More recently, machine learning methods have been applied to hedonic pricing and mass appraisal, including random forests (Hong et al., 2020), penalised variable selection (Yoo et al., 2012), and large-feature-set automated valuation (Kok et al., 2017). While these methods typically achieve superior predictive accuracy, they sacrifice the coefficient interpretability that is central to the hedonic framework's appeal for policy analysis and market valuation—a general tension in the adoption of machine learning by applied economists (Varian, 2014; Mullainathan and Spiess, 2017; Athey and Imbens, 2019; Dell, 2025). Our approach seeks to recover this interpretability while incorporating the rich informational content of textual data.

2.2. Text as Data in Economics

Our methodology is best understood as an instance of the broader *text-as-data* program in empirical economics, surveyed by Gentzkow et al. (2019) and, with a focus on modern deep-learning representations, by Ash and Hansen (2023). This literature organises text-based empirical work around a common two-step structure: first, represent documents as quantitative features; second, use those features in an economic model. The representation step has evolved from hand-built dictionaries—word lists whose counts proxy for a concept of interest, the workhorse of the finance literature on disclosure tone (Loughran and McDonald, 2011, 2016; García et al., 2023)—through topic models (Blei et al., 2003; Hansen et al., 2017) to dense neural embeddings.

Dictionary methods remain attractive to economists precisely because they are transparent: the researcher names the concept, and the mapping from text to feature is auditable. Their weakness

is lexical brittleness—a dictionary counts only the exact words it enumerates. Embedding-based representations solve the brittleness problem but break the transparency: a 384-dimensional vector has no nameable economic content. Our reference-description approach is designed to occupy the point between these extremes: like a dictionary, each feature is defined ex ante by the researcher and carries a name and an expected sign; like an embedding method, the mapping from text to feature is semantic rather than lexical, so paraphrases and synonyms load on the same dimension. In the taxonomy of Gentzkow et al. (2019), it is a supervised projection of an unsupervised representation, with the supervision supplied by domain knowledge rather than by labeled data. Applications elsewhere in economics illustrate the payoff of interpretable text features: Hansen et al. (2017) measure deliberation in central bank transcripts, Marinescu and Wolthoff (2020) show that the words of job postings materially affect matching in labour markets, and Kelly et al. (2021) measure the novelty of patents from the similarity of their text to earlier and later patents. The same literature documents that dictionaries are context-dependent: Loughran and McDonald (2011) show that a general-purpose negative word list misclassifies financial text, and Nowak et al. (2021) show that the words that predict house prices differ across markets and over time. A continuous dictionary inherits this dependence, which is why the present paper reports the sensitivity of its estimates to the phrasing of the references and to the encoder.

2.3. Text and Unstructured Data in Real Estate

The incorporation of textual data into real estate analysis represents a growing but still nascent literature. An early strand asked whether agent remarks matter at all: Haag et al. (2000) found that specific remark categories in MLS listings are associated with price and marketing-time differentials, and Pryce and Oates (2008) documented the systematic rhetorical strategies of real estate advertising, showing that listing language is a deliberate marketing instrument rather than neutral description. Goodwin et al. (2014) showed that broker vernacular affects both selling price and liquidity, and Goodwin et al. (2018) distinguished the positive and negative connotations of listing wording. Because listing text is strategically produced, its information content connects to the theory of voluntary disclosure (Grossman, 1981; Milgrom, 1981) and to the evidence that informed intermediaries exploit textual and informational advantages (Levitt and Syverson, 2008).

A second strand quantifies the price-relevant information in the text itself. Nowak and Smith (2017) were among the first to embed listing-comment features in formal hedonic estimation, showing that words in agent comments proxy for unobserved quality and materially reduce omitted variable bias in the estimates of other coefficients; Nowak et al. (2021) extend the approach to the construction of real estate dictionaries and show that their content varies across space and time. Shen and Ross (2021) pushed this agenda furthest to date: applying machine learning to property descriptions, they show that text contains substantial information about otherwise unobserved property quality

and that this information predicts prices beyond structured attributes. Sentiment-based approaches summarise the tone of housing narratives (Zhu et al., 2023), and recent automated-valuation work incorporates full description text into machine learning pipelines (Baur et al., 2023). Listing language also encodes neighbourhood characteristics beyond the property itself: Delmelle and Nilsson (2021) show that the text of property advertisements predicts neighbourhood socioeconomic composition and change, a reminder that text features may capture locational as well as structural variation—a theme central to our identification discussion in Section 3.8.

Text is not the only unstructured signal attached to a listing. A parallel literature extracts quality information from listing *photographs*: Glaeser et al. (2018) show that exterior appearance predicts prices and renovation returns, Poursaeed et al. (2018) estimate interior quality from listing photos and improve price prediction, and Law et al. (2019) combine street-view and satellite imagery with structured attributes. We view text and images as complementary members of the same family: both reveal quality dimensions that structured fields omit, and both raise the same interpretability challenge when processed by deep networks. Topic models offer one interpretable middle ground for text (Blei et al., 2003), but their latent factors are difficult to name ex ante and are unstable in short, formulaic documents such as listings; our reference-based projection provides an alternative whose dimensions are fixed and named by design.

The tension between prediction and interpretation in text-augmented hedonic models mirrors a broader debate in applied econometrics. Athey and Imbens (2019) provide a framework for thinking about when machine learning methods complement rather than replace traditional econometric approaches—specifically, they are most useful for constructing features (as in our reference-based approach) rather than for final-stage inference. Our work builds on these contributions while addressing the interpretability limitation that pervades existing NLP-based approaches in real estate economics.

2.4. Sentence Embeddings and Semantic Textual Similarity

Sentence embeddings map variable-length text into fixed-dimensional vector spaces where semantic similarity corresponds to geometric proximity. The development of contextual word representations through architectures such as ELMo (Peters et al., 2018) and subsequently BERT (Devlin et al., 2019) represented a paradigm shift from static word vectors (Mikolov et al., 2013) to context-dependent representations. However, these models require cross-encoding for pairwise comparison, which scales quadratically with corpus size.

The Sentence-BERT framework (Reimers and Gurevych, 2019) addressed this limitation by fine-tuning pre-trained transformer models using siamese and triplet network structures to produce embeddings optimised for cosine similarity comparisons. This approach enables efficient comparison of text pairs without requiring cross-encoding, which is computationally prohibitive for large-scale

applications. Subsequent work by Gao et al. (2021) proposed contrastive learning objectives (SimCSE) that further improved embedding quality, while Li et al. (2020) demonstrated that sentence embeddings suffer from anisotropy—a geometric degeneration where embeddings occupy a narrow cone in the vector space—and proposed whitening transformations as a remedy.

The all-MiniLM-L6-v2 model (Wang et al., 2020) is a distilled version of the MiniLM architecture that produces 384-dimensional embeddings. The model employs self-attention distillation, transferring knowledge from a larger teacher model to a compact student network. Despite its compact size (22.7M parameters), it achieves competitive performance on standard semantic textual similarity benchmarks while being substantially faster than BERT-base encoders (Wang et al., 2020; Muennighoff et al., 2023), which makes it suitable for encoding large corpora of property listings. The model was trained predominantly on English paraphrase data, and its treatment of French text is incidental rather than designed. Multilingual sentence encoders address this by distillation: Reimers and Gurevych (2020) train a student model so that a sentence and its translations receive the same embedding as the English teacher, which aligns the languages in a shared space; paraphrase-multilingual-MiniLM-L12-v2, used in Section 7, is such a model. Dedicated French models such as CamemBERT (Martin et al., 2020) or FlauBERT (Le et al., 2020) could further improve semantic resolution for domain-specific vocabulary. Whether a monolingual encoder applied to a bilingual corpus measures content or language is an empirical question that the present paper answers directly.

Our use of reference descriptions as fixed semantic anchors is also related to zero-shot classification, where label descriptions serve as textual anchors against which documents are scored without task-specific training (Yin et al., 2019). The difference is that we retain the continuous similarity scores as regression covariates rather than converting them into discrete class assignments, preserving the marginal-effect interpretation required by the hedonic framework.

Cosine similarity between two L_2-normalised embedding vectors $\mathbf{u}$ and $\mathbf{v}$ is defined as:

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}|| \cdot ||\mathbf{v}||} = \mathbf{u} \cdot \mathbf{v} \tag{2}$$

This measure ranges from -1 to 1, with higher values indicating greater semantic similarity. For normalised vectors, cosine similarity reduces to the dot product, enabling efficient computation via matrix multiplication. The geometric interpretation is straightforward: cosine similarity measures the angle between two vectors in the embedding space, with a value of 1 indicating identical direction (perfect semantic alignment) and 0 indicating orthogonality (semantic independence). Negative values, while theoretically possible, are rare in practice for sentence embeddings due to the positivity bias inherent in transformer representations (Ethayarajh, 2019).

2.5. Research Gap and Contribution

The literatures above leave a two-sided gap. On the economics side, the text-as-data toolkit offers either transparent-but-brittle dictionaries or powerful-but-opaque embeddings, and the housing applications to date have prioritised prediction over the estimation of named implicit prices. On the NLP side, interpretable machine learning has developed post hoc explanation tools (Ribeiro et al., 2016; Lundberg and Lee, 2017) and arguments for inherently interpretable models in high-stakes settings (Rudin, 2019), but these have not been adapted to the requirements of hedonic inference, where features must be defined ex ante, carry economic names, and enter a linear model with valid standard errors. Table 1 summarises how our approach compares with existing methods along five dimensions. The table records the interpretability of the reference approach as high because its features are named ex ante; whether the coefficients attached to those names are stable is a separate question, which Section 7 examines and which the existing literature has not addressed.

Table 1. Comparison of text-based approaches in hedonic pricing.

Method	Semantic Depth	Interpret-ability	Scalability	Multi-lingual	Key Reference
Keyword counts	Low	High	High	No	Nowak and Smith (2017)
Broker vernacular / connotation	Low	High	High	No	Goodwin et al. (2018)
ML on text features	Low	Moderate	High	No	Shen and Ross (2021)
Sentiment indices	Low	High	High	Yes	Zhu et al. (2023)
LDA topics	Moderate	Low	Moderate	No	Blei et al. (2003)
Neural embeddings (black-box)	High	None	Low	Limited	Baur et al. (2023)
PCA on embeddings	High	None	High	Yes	—
Reference cosine	**High**	**High**	**High**	**Yes**	**This paper**

Our reference-based cosine similarity approach addresses the depth–interpretability tension by using the semantic representations of transformer embeddings while producing features with a clear economic definition. Each similarity score measures the degree to which a property's description resembles a researcher-defined qualitative archetype, yielding features that are both semantically meaningful and directly usable as right-hand-side variables in a hedonic regression.

This approach contributes to a broader methodological trend in applied economics where machine learning tools are used for feature construction rather than final-stage estimation (Athey and Imbens, 2019; Mullainathan and Spiess, 2017; Gentzkow et al., 2019). By treating embeddings as an intermediate representation and projecting them onto economically meaningful axes, we preserve

the inferential advantages of OLS while exploiting the representational power of deep learning. The reference-based projection is also related to the concept of "concept bottleneck models" in interpretable machine learning (Koh et al., 2020), where high-dimensional representations are channeled through human-interpretable intermediate concepts before reaching the prediction stage.

3. Methodology

3.1. Overview

The framework transforms unstructured listing text into a set of interpretable features through a four-stage pipeline, illustrated in Figure 1: (1) extraction of the listing descriptions; (2) encoding into dense vector representations with a pre-trained sentence transformer; (3) computation of cosine similarities against a set of researcher-designed reference descriptions; and (4) estimation of hedonic price equations in which the similarities enter as named covariates.

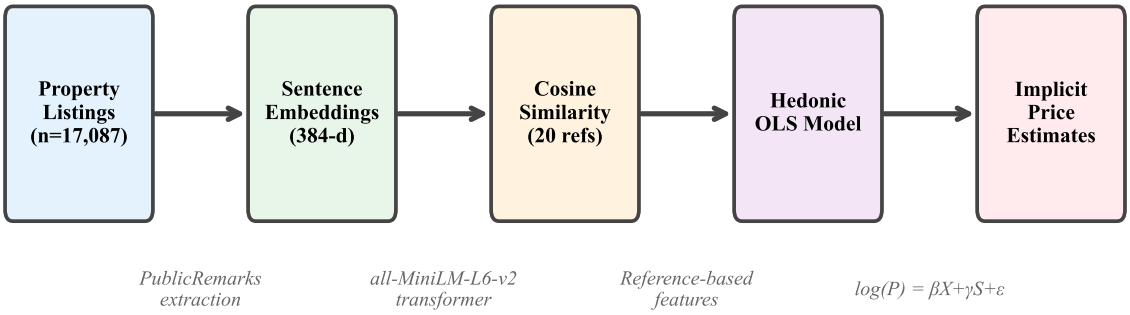


Figure 1. Methodological framework: from listing text to hedonic price estimates.
Notes: The four stages are (1) text extraction, (2) sentence embedding, (3) cosine similarity against 20 reference descriptions and (4) hedonic estimation.

3.2. Text Extraction

Each listing includes a *PublicRemarks* field containing the broker-written description. We apply minimal preprocessing: listings with descriptions shorter than 20 characters are excluded as uninformative, and listings with missing or non-positive prices are removed. We deliberately avoid stemming, lemmatisation or stop-word removal, because the sentence transformer is designed to process natural-language text and handles tokenisation internally. In addition to the embedding-based

features, we record the description length ℓ_i (number of characters) as a control: length may proxy for broker effort, property complexity or market conditions, and including it ensures that the semantic features capture content rather than verbosity. We also classify each description as French, English or mixed from counts of language-specific stop words (Section 4); the classification is used in the extended specifications of Section 7.

3.3. Sentence Embedding

Each description d_i is encoded into a unit-length vector with the all-MiniLM-L6-v2 sentence transformer (Wang et al., 2020; Reimers and Gurevych, 2019):

$$\mathbf{e}_i = f_{\boldsymbol{\theta}}(d_i) \in \mathbb{R}^{384}, \qquad \|\mathbf{e}_i\|_2 = 1. \tag{3}$$

The model has six transformer layers with 384 hidden dimensions; token embeddings are mean-pooled and L_2-normalised. The corpus of 17,087 descriptions is encoded in about 14 seconds on an Apple-silicon laptop. The model was trained predominantly on English paraphrase data; its behaviour on French text, and the consequences for a bilingual corpus, are examined directly in Section 7.1, where we also re-run the analysis with a multilingual encoder, paraphrase-multilingual-MiniLM-L12-v2, which was distilled to map translations of the same sentence to nearby vectors (Reimers and Gurevych, 2020).

3.4. Reference Description Design

3.4.1. Rationale

The central design choice is the construction of reference descriptions that serve as semantic anchors. Rather than reducing the 384-dimensional embedding by an unsupervised method (principal components, autoencoders), which yields features without economic names, we project each listing's embedding onto axes defined by the researcher. Each axis is a reference text that embodies one qualitative housing dimension.

The idea is analogous to factor-mimicking portfolios in asset pricing (Fama and French, 1993): just as a Fama–French factor is a portfolio constructed to load on a risk dimension, a reference description is a synthetic text constructed to load on a quality dimension. Within the text-as-data toolkit, the design is a semantic generalisation of the dictionary method (Loughran and McDonald, 2011; Gentzkow et al., 2019): a reference plays the role of a curated word list, but the matching operates in embedding space, so paraphrases and synonyms contribute without being enumerated. It is also related to anchor-based zero-shot classification, where label descriptions serve as fixed textual anchors (Yin et al., 2019), except that we retain the continuous similarity scores as regression

covariates instead of discretising them into class labels.

3.4.2. Reference Categories and Design Principles

We define 20 reference descriptions spanning six domains of housing quality (Table 2). Each reference is a synthetic paragraph of 150–260 characters written in French, the majority language of the corpus, listing the vocabulary and phrasing characteristic of its dimension; the verbatim texts are reproduced in Appendix A. Three principles guided their design: *semantic saturation* (each reference lists several synonyms and related phrases to cover its dimension broadly); *dimensional specificity* (each reference targets a single dimension); and *linguistic consistency* (all references use vocabulary typical of Quebec listings). The second principle is imperfectly met: several references share phrases—Land & Nature mentions waterfront access, Modern/Contemporary mentions new construction and energy efficiency, Premium Location mentions schools, parks and family neighbourhoods—and these overlaps, visible in Appendix A, matter for the interpretation of individual coefficients (Section 7).

Table 2. Reference description categories, domains, and expected price effects.

Domain	Category	Dimension	Expected Effect
Quality & Standing	Luxury	Premium finishes, prestige materials	Positive
	Entry-Level	Affordable, first-time buyer language	Negative
Condition	Renovated	Turnkey, recently updated	Positive
	Needs Renovation	Fixer-upper, requires work	Negative
Physical Attributes	Bright & Spacious	Luminosity, open concept, high ceilings	Positive
	Garage & Parking	Garage, parking facilities	Positive
	Finished Basement	Developed lower level	Positive
	Pool & Landscaping	Pool, outdoor amenities	Positive
Location & Setting	Land & Nature	Wooded lot, natural setting	Positive
	Panoramic View	Water, mountain, city views	Positive
	Premium Location	Near amenities, transit, schools	Positive
	Quiet & Peaceful	Rural, secluded, tranquil	Ambiguous
	Waterfront	Lake/river access, dock	Positive
Style & Character	Modern/Contemporary	Clean design, smart home, energy-efficient	Positive
	Heritage/Character	Period features, historic charm	Ambiguous
	New Construction	Newly built, never occupied	Positive
	Energy Efficient	Heat pump, solar, high insulation	Ambiguous
	Family-Friendly	Schools, parks, safe neighbourhood	Ambiguous
Market Signals	Income/Investment	Rental income, investment potential	Ambiguous
	Motivated Seller	Urgency, price reduction, distress	Negative

3.5. Cosine Similarity

For property i and reference j, with $\mathbf{r}_j$ the L_2-normalised embedding of the j-th reference,

$$s_{ij} = \mathbf{e}_i \cdot \mathbf{r}_j \in [-1, 1], \qquad j = 1, \ldots, 20. \tag{4}$$

Stacking the embeddings in $\mathbf{E} \in \mathbb{R}^{n\times 384}$ and the references in $\mathbf{R} \in \mathbb{R}^{20\times 384}$, the similarity matrix is $\mathbf{S} = \mathbf{E}\mathbf{R}' \in \mathbb{R}^{n\times 20}$, a single matrix product. A higher s_{ij} indicates that the listing uses language closer in meaning to reference j. Because every listing is compared with every reference, the 20 columns of $\mathbf{S}$ share whatever is common to the listing's embedding—its language, its register, its length—and are therefore strongly correlated; Sections 5 and 7 document and address this feature of the design.

3.6. Hedonic Specifications

Let $\mathbf{X}_i$ collect six structural variables (bedrooms, bathrooms, half-bathrooms, parking spaces, storeys and the lot-size field), ℓ_i the description length and $\mathbf{s}_i = (s_{i1}, \ldots, s_{i20})'$ the similarity vector. We estimate five nested specifications by ordinary least squares:

$$\text{Model A:} \quad \ln P_i = \alpha + \boldsymbol{\beta}'\mathbf{X}_i + \varepsilon_i, \tag{5}$$

$$\text{Model B:} \quad \ln P_i = \alpha + \boldsymbol{\beta}'\mathbf{X}_i + \delta\,\ell_i + \varepsilon_i, \tag{6}$$

$$\text{Model C:} \quad \ln P_i = \alpha + \boldsymbol{\beta}'\mathbf{X}_i + \boldsymbol{\gamma}'\mathbf{s}_i + \varepsilon_i, \tag{7}$$

$$\text{Model D:} \quad \ln P_i = \alpha + \boldsymbol{\beta}'\mathbf{X}_i + \delta\,\ell_i + \boldsymbol{\gamma}'\mathbf{s}_i + \varepsilon_i, \tag{8}$$

$$\text{Model E:} \quad \ln P_i = \alpha + \boldsymbol{\beta}'\mathbf{X}_i + \delta\,\ell_i + \boldsymbol{\gamma}'_{\text{sig}}\mathbf{s}_i^{\text{sig}} + \varepsilon_i, \tag{9}$$

where $\mathbf{s}_i^{\text{sig}}$ retains the similarities that are individually significant at the 5% level in Model D. All covariates are standardised to zero mean and unit variance on the estimation sample before estimation, so that each coefficient measures the change in log price associated with a one-standard-deviation change in the covariate. The semi-logarithmic form follows Halvorsen and Pollakowski (1981); Cropper et al. (1988) show that simple parametric forms are preferable to flexible transformations when attributes are measured with error or omitted, which is the case here. Standard errors are heteroskedasticity-consistent (HC3), the variant recommended by MacKinnon and White (1985).

Interpretation of the semantic coefficients. Because s_{ij} is standardised, the coefficient γ_j is the change in $\ln P$ for a one-standard-deviation increase in similarity to reference j, holding the other covariates fixed, and the corresponding proportional price change is

$$100 \times (e^{\gamma_j} - 1) \text{ per cent per standard deviation of } s_j. \tag{10}$$

Equivalently, since the standard deviations of the 20 similarities lie between 0.11 and 0.17 (Table 5), the coefficient per unit of cosine similarity is γ_j/σ_{s_j}, and a 0.10 increase in cosine similarity—roughly the distance between the median listing and the top quartile—changes log price by about $0.10\,\gamma_j/\sigma_{s_j}$. We report Equation (10) throughout and refer to it as the price impact per standard deviation. These are partial associations in a cross-section of listing prices; the identification discussion below explains why they are not causal effects of wording.

Joint test of the text block. The incremental contribution of the 21 text variables (length and the 20 similarities) is tested with

$$F = \frac{(\mathrm{SSR}_A - \mathrm{SSR}_D)/q}{\mathrm{SSR}_D/(n-k-1)}, \qquad q = 21,\ k = 27, \tag{11}$$

where SSR denotes the residual sum of squares, q the number of restrictions and k the number of regressors in Model D excluding the intercept. Equation (11) assumes homoskedasticity; in Section 7 we report heteroskedasticity- and cluster-robust Wald tests of the 20 semantic coefficients, overall and by domain, which do not.

3.7. Extended Specifications

The first version of this paper stopped at Models A–E. The present revision adds specifications that address three threats to interpretation: the language of the listing, unobserved location, and measurement error in the lot-size field. Let L_i be a pair of indicators for English and mixed-language descriptions, and let $g(i)$ denote the forward sortation area (FSA, the first three characters of the postal code) or the municipality of listing i. The extended model is

$$\ln P_i = \alpha_{g(i)} + \boldsymbol{\beta}'\mathbf{X}_i + \delta\,\ell_i + \boldsymbol{\lambda}'L_i + \boldsymbol{\gamma}'\mathbf{s}_i + \varepsilon_i, \tag{12}$$

estimated by the within transformation (all variables demeaned within g). The area effects α_g absorb location-specific price levels, so that $\boldsymbol{\gamma}$ is identified from variation among listings in the same FSA or municipality. Because listings in the same area share unobserved shocks and because the fixed effects define the sampling structure, standard errors are clustered by area (Cameron and Miller, 2015; Abadie et al., 2023); with 415 FSAs the cluster count is ample. In variants of Equation (12) the raw lot-size field is replaced by the logarithm of the lot area converted to square metres with an indicator for missing values, and the logarithm of the interior floor area is added on the subsample that reports it.

Out-of-sample validation. In-sample R^2 gains can overstate the value of added regressors. We therefore evaluate each specification by five-fold cross-validation (Hastie et al., 2009), with random folds and with folds that hold out entire FSAs so that no held-out listing shares an area with the training data (Roberts et al., 2017). For held-out predictions $\hat{y}_i$ of $y_i = \ln P_i$,

$$R^2_{\text{oos}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{13}$$

$$\text{MdAPE} = \text{median}_i \left| \frac{e^{\hat{y}_i} - P_i}{P_i} \right| \times 100, \tag{14}$$

together with the root mean squared error in log points. The median absolute percentage error is the accuracy measure conventionally used for automated valuation models.

3.8. Identification and Interpretation

The coefficients on the semantic similarities are *conditional associations*—implicit semantic price gradients—rather than causal effects of wording. This follows the modern reading of cross-sectional hedonic coefficients as equilibrium price gradients rather than structural willingness-to-pay parameters (Ekeland et al., 2004; Kuminoff et al., 2013), and it is reinforced by the way listing text is produced. The description is written by the broker after the asking price has been set, with knowledge of the property, of the price and of the targeted buyer segment. Text and price are therefore jointly determined by the same information set, and three channels could generate the observed correlations.

Information channel. Brokers describe observable attributes that affect prices but are absent from the structural fields. The similarities then proxy for genuine quality dimensions—finishes, condition, natural setting—and the coefficients have a hedonic interpretation as implicit prices of unmeasured quality.

Persuasion and positioning channel. Brokers choose language to match the price point they have set: a property priced for the luxury segment is described in luxury vocabulary, and a property priced to sell quickly is described with urgency cues. Under this channel the coefficient measures how price-driven wording is, not how wording drives prices.

Omitted-location channel. Listing language proxies for locational characteristics correlated with price. Family-oriented marketing characterises suburban markets; new-construction vocabulary characterises peripheral developments; the English language characterises the western part of the Montréal region and the Outaouais, where prices are higher. Fixed effects at the FSA or municipality level absorb between-area price differences but not within-area location quality.

The cross-sectional design cannot separate these channels, and the paper does not attempt to. The estimates are informative about the information content of broker narratives, about the segmentation

of the market along qualitative lines and about the predictive value of text for valuation. They do not establish that rewriting a description would change its price, and we avoid any language suggesting that semantic similarity "causes" or "determines" prices. Likewise, because prices are asking prices rather than transaction prices, the coefficients describe the pricing behaviour of sellers and brokers conditional on the text they wrote, not the valuation of buyers.

4. Data

4.1. Source and Sample Construction

The data are drawn from a database of residential listings in Quebec, Canada, collected from the Realtor.ca and Centris platforms in a single extraction completed on 7 January 2026. Realtor.ca is the public portal of the Canadian Real Estate Association, and Centris is the Quebec multiple listing service operated by the Québec Professional Association of Real Estate Brokers; together they cover the large majority of brokered residential listings in the province. The database is a snapshot of active listings, not a transaction file: prices are asking prices, and each property appears once.

The database contains 46,479 listings in five categories: houses (17,089), condominiums (8,586), land (8,506), rentals (8,240) and multiplexes (4,058). We restrict the sample to single-family houses to maintain a homogeneous product category, as is standard in the hedonic literature (Sirmans et al., 2005). After excluding listings with a description shorter than 20 characters or a non-positive price, the sample comprises 17,087 observations. Every listing carries geographic coordinates and a postal code; the first three characters of the postal code define 415 forward sortation areas (FSAs) with a median of 21 listings each, and the address field identifies 1,061 municipalities. These identifiers are used for the fixed effects, the clustering and the spatial cross-validation of Section 7.

4.2. Structural Variables

Table 3 reports descriptive statistics for the structural and textual variables.

The median asking price is $589,900 and the distribution is strongly right-skewed (mean $807,259, maximum $25 million); nine listings are priced below $50,000 and are almost certainly data errors or non-residential offers, and Section 6 shows that trimming the tails does not alter the results. The log transformation reduces the skewness (standard deviation of log price 0.70). The typical house has three bedrooms and two bathrooms (sample medians). The parking field averages 5.5 spaces because it counts all reported spots, including driveway and outdoor spaces. The lot-size field is recorded as a free-text quantity whose unit varies at the source: 61% of the values are in square metres, 32% in square feet, 2% in acres or hectares and 4% are missing. The first version of this paper used the numeric part of the field as reported; we retain that variable in the baseline

Table 3. Descriptive statistics for structural and textual variables (n = 17,087).

Variable	Mean	Std. Dev.	Min	P25	Median	P75	Max
Price ($)	807,259	900,313	501	399,000	589,900	889,900	25,000,000
log(Price)	13.32	0.70	6.22	12.90	13.29	13.70	17.03
Bedrooms	3.5	1.3	0	3	3	4	23
Bathrooms	1.8	0.9	0	1	2	2	18
Half-bathrooms	0.5	0.6	0	0	0	1	12
Parking spaces	5.5	5.1	0	2	4	7	152
Stories	1.4	0.7	0	1	2	2	2
Lot size (raw field, mixed units)	80,551	5,514,237	0	683	2,995	8,316	719,640,188
Description length (char.)	510	151	22	430	549	634	703

Notes: The sample is restricted to single-family houses with a positive listing price and a description of at least 20 characters. Descriptions are truncated at approximately 700 characters in the data export, which bounds the description-length variable. The raw lot-size field mixes units at the source (61

models for comparability and document in Section 7 that it is uninformative in-sample and harmful out of sample, and we replace it there by a unit-corrected lot area. An interior floor area is reported for 6,228 listings (36%) and is used in a subsample analysis. Descriptions average 510 characters with a standard deviation of 151 and are truncated at approximately 700 characters in the export, so the length variable is bounded from above.

Figure 2 displays the distributions of price and log price.

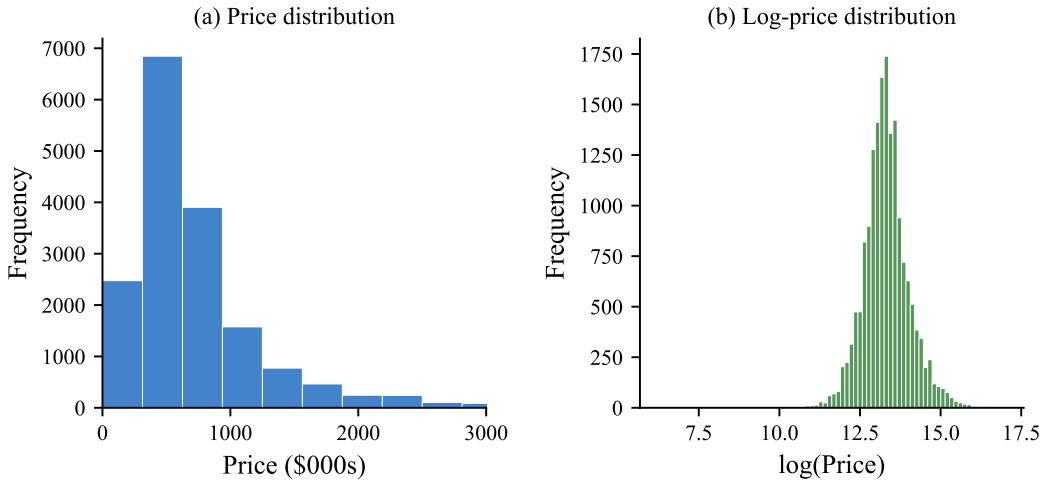


Figure 2. Distribution of listing prices and log prices (n = 17,087).
Notes: Left panel, asking prices; right panel, log asking prices. The log transformation substantially reduces the right skew.

4.3. Language of the Descriptions

The first version of this paper described the corpus as "predominantly in French with occasional English phrases". A systematic classification shows otherwise. We count French and English function words and domain terms in each description and classify a listing as French when the French count exceeds 1.5 times the English count, as English in the reverse case and as mixed otherwise. Table 4 reports the result: 57.6% of the descriptions are French, 39.7% are English and 2.7% are mixed, the last group consisting mostly of short boilerplate such as "See listing broker(s)". English listings are not a random subset: their median asking price is $749,900 against $499,900 for French listings, reflecting the geography of anglophone Quebec (the western part of the Montréal region, the Outaouais, parts of the Eastern Townships). Because the 20 references are written in French and the baseline encoder was trained mainly on English, an English description scores lower against every reference: the mean of the 20 similarities is 0.204 for English listings and 0.434 for French ones, and language alone explains 86% of the variance of the mean similarity. With the multilingual encoder the gap shrinks to 0.352 against 0.395 and the share explained by language falls to 53%. These facts drive much of the interpretation in Sections 5 and 7.

Table 4. Language of the listing descriptions and mean cosine similarity to the 20 French references.

						Mean similarity	
Language	***n***	**%**	**Length**	**Median price**	$\overline{\ln P}$	**MiniLM-L6**	**Multilingual**
French	9,843	57.6	521	$499,900	13.128	0.434	0.395
English	6,779	39.7	527	$749,900	13.605	0.204	0.352
Mixed / uninformative	465	2.7	33	$599,000	13.260	0.080	0.073
Share of the variance of the mean similarity explained by language (η^2)						0.862	0.525
Correlation of the mean similarity with the English indicator						−0.815	−0.189
Correlation of the mean similarity with description length						0.199	0.530
Share of the variance of the 20 similarities on their first principal component						0.862	0.522

Notes: Language is classified from French and English stop-word counts (ratio rule of 1.5; "mixed" listings are mostly short boilerplate such as "See listing broker(s)"). Length is the mean number of characters. MiniLM-L6 is the paper's baseline encoder (all-MiniLM-L6-v2); Multilingual is paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2020). The mean similarity is the average of the 20 cosine similarities of a listing.

4.4. Semantic Similarity Features

Table 5 reports the distribution of the 20 cosine similarities and their bivariate correlations with log price.

All bivariate correlations between similarity scores and log price are negative, between −0.29 and −0.17. The first version of this paper attributed this pattern to description length. The language

Table 5. Descriptive statistics for cosine similarity features and bivariate correlations with log(price).

Reference	**Mean**	**Std. Dev.**	**Min**	**Median**	**Max**	**Corr.** $\ln P$
Luxury	0.337	0.157	−0.030	0.376	0.692	−0.234
Entry-Level	0.313	0.167	−0.085	0.385	0.660	−0.270
Renovated	0.352	0.168	−0.056	0.411	0.753	−0.250
Needs Renovation	0.321	0.138	−0.083	0.367	0.652	−0.251
Bright & Spacious	0.357	0.153	−0.104	0.400	0.672	−0.228
Land & Nature	0.330	0.116	−0.021	0.337	0.701	−0.174
Panoramic View	0.316	0.120	−0.114	0.343	0.613	−0.185
Premium Location	0.321	0.141	−0.025	0.364	0.693	−0.289
Quiet & Peaceful	0.358	0.111	−0.065	0.376	0.630	−0.228
Income/Investment	0.358	0.126	−0.029	0.398	0.661	−0.275
Garage & Parking	0.310	0.112	−0.078	0.317	0.685	−0.210
Finished Basement	0.305	0.139	−0.069	0.324	0.708	−0.228
Modern/Contemporary	0.365	0.115	−0.044	0.382	0.671	−0.187
Heritage/Character	0.366	0.143	−0.090	0.399	0.715	−0.235
Energy Efficient	0.298	0.134	−0.058	0.327	0.648	−0.234
Motivated Seller	0.254	0.165	−0.151	0.326	0.586	−0.279
Family-Friendly	0.349	0.142	−0.080	0.371	0.722	−0.231
Waterfront	0.331	0.151	−0.117	0.370	0.728	−0.223
New Construction	0.372	0.122	−0.020	0.396	0.694	−0.240
Pool & Landscaping	0.348	0.158	−0.070	0.382	0.756	−0.240

Notes: Cosine similarity is computed between each listing embedding and the corresponding reference description embedding using the all-MiniLM-L6-v2 sentence transformer. Corr. ln P denotes the Pearson correlation with log listing price.

analysis identifies the main cause: English listings, which are more expensive, score lower against the French references on every dimension, so that each similarity is negatively correlated with price through the language composition of the corpus. Length plays a secondary role—the correlation between the mean similarity and description length is only 0.20—and the mixed-language boilerplate listings, which are short and score near zero, contribute to it. The multivariate specifications of Sections 5 and 7 control for length and, in the extended models, for language. Figure 3 displays the distributions of the similarities, which show substantial dispersion within each dimension.

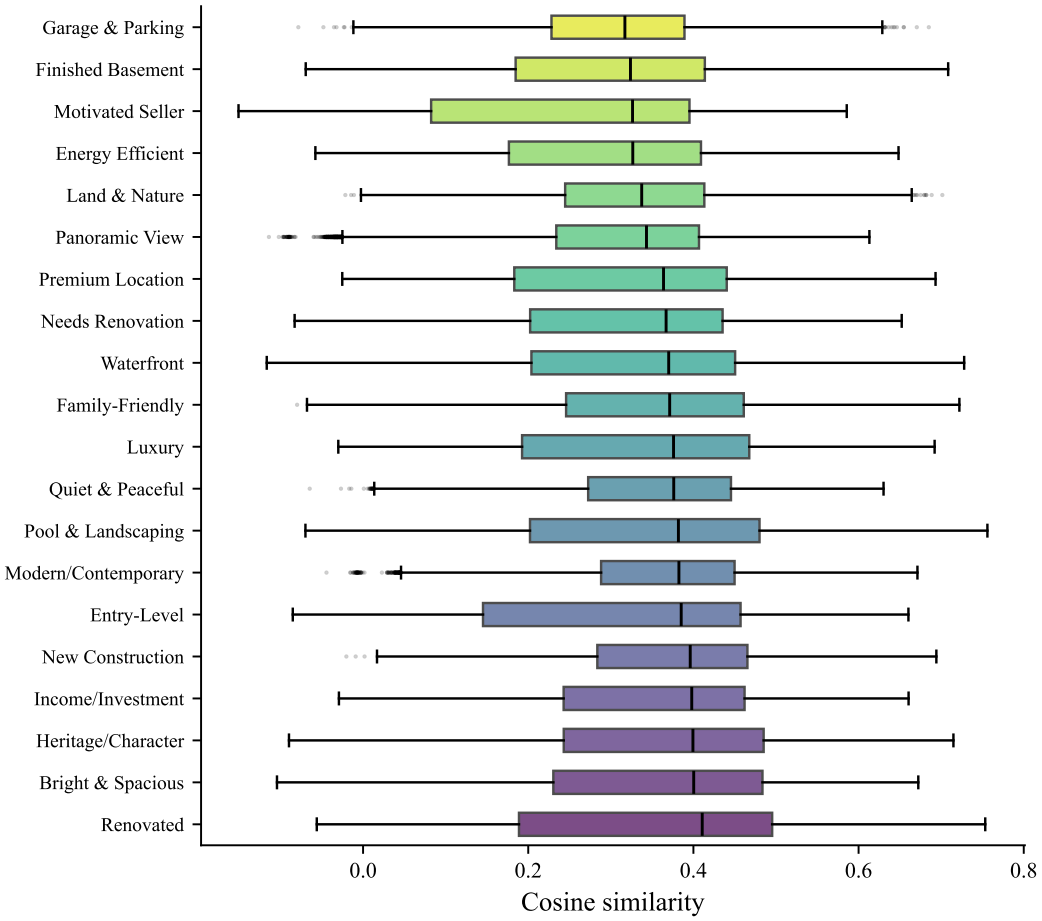


Figure 3. Distribution of cosine similarity scores across the 20 reference dimensions.
Notes: Each box plot displays the distribution of the cosine similarity between listing embeddings and the corresponding reference embedding, ordered by median similarity. Boxes span the interquartile range; whiskers extend to 1.5 times the interquartile range. The wide interquartile ranges, and the low lower quartiles of Motivated Seller and Entry-Level, partly reflect the two languages of the corpus (Table 4).

5. Baseline Results

This section reports the five nested specifications of Equations (5)–(9), which constitute the baseline of the paper and reproduce the estimates of its first version. The extended specifications that control

for language and location are presented in Section 7; readers should keep in mind, when reading the individual coefficients below, that several of them change materially once those controls are added.

5.1. Model Comparison

Table 6 reports goodness-of-fit statistics and Figure 4 visualises them.

Table 6. Hedonic model comparison (n = 17,087).

	Specification	R^2	Adj. R^2	AIC	BIC	k	ΔR^2 vs. A
A	Structural only	0.4525	0.4523	26,165	26,219	6	—
B	+ Description length	0.4647	0.4645	25,779	25,841	7	+0.012
C	+ Semantic similarities (20)	0.4962	0.4954	24,783	24,992	26	+0.044
D	Full model (B + C)	**0.5122**	**0.5115**	**24,232**	24,448	27	+0.060
E	Parsimonious (sig. only)	0.5118	0.5111	24,239	**24,425**	23	+0.059

Notes: k denotes the number of regressors excluding the intercept. Bold values indicate the best fit for each criterion. Models are estimated by OLS with HC3 robust standard errors.

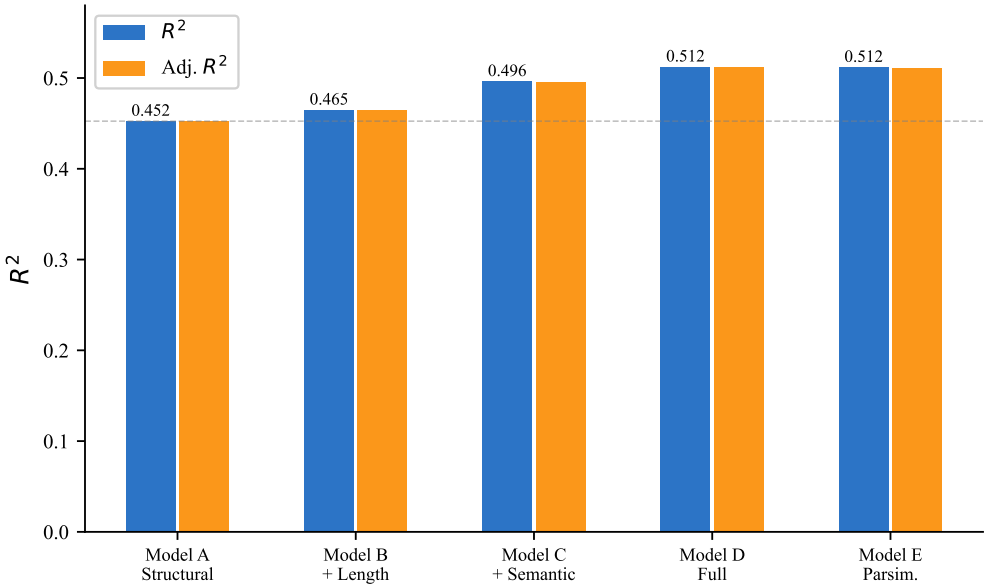


Figure 4. Goodness-of-fit comparison across hedonic model specifications.
Notes: R^2 and adjusted R^2 across the five nested specifications of Equations (5)–(9). The dashed line marks the structural baseline (Model A).

The structural model (A) explains 45.2% of the variation in log asking prices. This is modest by the standards of hedonic studies that include location variables, and Section 7 shows that area fixed effects alone raise the R^2 to about 0.65; Model A should be read as a deliberately sparse benchmark with no locational information. Adding description length (Model B) raises the R^2 by 1.2 percentage points, so the quantity of text is itself price-informative. Adding the 20 similarities without the

length control (Model C) raises it by 4.4 points, and the full model (D) by 6.0 points, to 0.512; the semantic block adds 4.8 points over Model B. The parsimonious model (E), which retains the 16 similarities individually significant at the 5% level in Model D, fits almost identically ($R^2 = 0.512$) and has the lowest BIC (24,425 against 24,448), so that the four dropped dimensions—Renovated, Income/Investment, Heritage/Character and Pool & Landscaping—contribute nothing independently once the others are included. Model E's coefficients are reported in Appendix B.

Figure 5 decomposes the R^2 of Model D in the order structural variables, length, semantics: 88.3%, 2.4% and 9.3% of the explained variance. The decomposition is sequential and therefore order-dependent; the semantic share is the increment of the block entered last, which is the most conservative attribution.

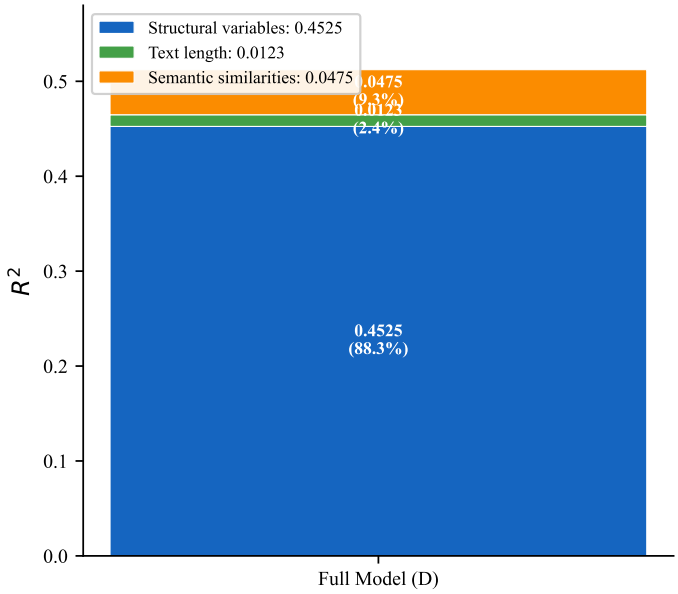


Figure 5. Sequential decomposition of the explained variance (R^2) of Model D.

Notes: Structural variables entered first account for 88.3% of the R^2 of Model D, description length entered second for 2.4%, and the 20 semantic similarities entered last for 9.3%. The shares depend on the order of entry.

5.2. Joint Significance

The F-statistic of Equation (11) for the 21 text variables (length and 20 similarities), comparing Model D with Model A, is $F(21,\ 17{,}059) = 99.53$, $p < 0.001$. The hypothesis that listing text contains no price-relevant information beyond the structural characteristics is rejected decisively. Equation (11) assumes homoskedastic errors, which the Breusch–Pagan test rejects (Section 6); the heteroskedasticity-robust Wald test of the 20 semantic coefficients alone gives $F = 54.3$, and the version clustered by FSA gives $F = 20.1$ (Table 11), both with $p < 0.001$.

5.3. Structural Variables

Figure 6 presents the structural coefficients of Model D with 95% confidence intervals.

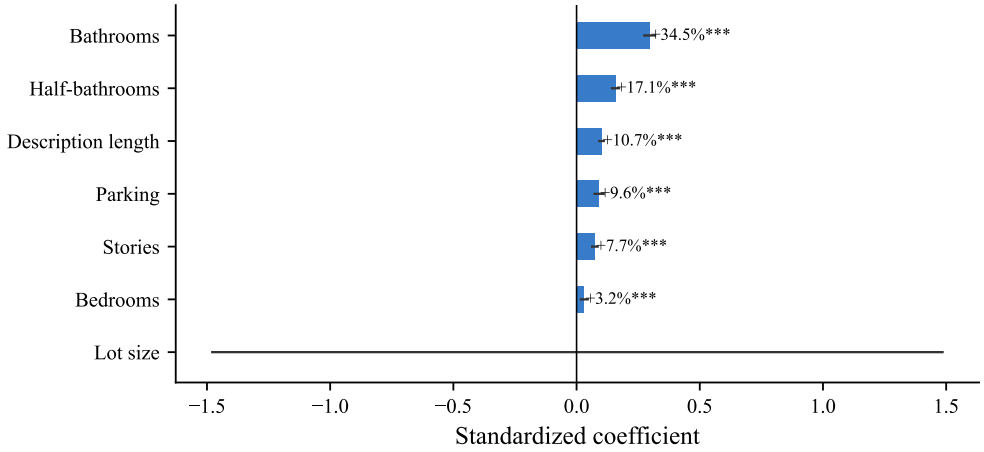


Figure 6. Structural coefficients of the full model (D), with 95% confidence intervals.
Notes: Each point is the percentage change in price associated with a one-standard-deviation increase in the variable (Equation 10). Error bars are 95% confidence intervals based on HC3 standard errors.

Bathrooms have the largest structural association (+34.5% per standard deviation), followed by half-bathrooms (+17.1%), description length (+10.7%), parking (+9.6%) and storeys (+7.7%); bedrooms add a modest +3.2%. The raw lot-size field is not significant ($t = 0.00$), which is the expected consequence of mixing square metres, square feet and acres in one variable (Section 4); the unit-corrected lot area used in Section 7 is strongly significant and, within FSAs, has the largest structural coefficient of all. The absence of a floor-area variable in the baseline is a notable omission relative to standard hedonic practice (Sirmans et al., 2005), and the floor-area subsample of Section 7 addresses it.

5.4. Semantic Dimensions

Table 7 reports the complete coefficient estimates of Model D and Figure 7 displays the semantic coefficients.

Three dimensions carry large positive associations. **Modern/Contemporary** (+16.4% per standard deviation, $p < 0.001$) is the largest: listings closer to the modern-design reference—contemporary architecture, clean lines, smart-home and energy-efficient construction—are the most expensive conditional on structure. **Luxury** (+14.2%) captures premium finishes, gourmet kitchens and spa-like bathrooms, quality variation that bathroom counts do not record. **Land & Nature** (+13.0%) captures wooded lots, landscaping, privacy and natural surroundings. **Waterfront** (+4.7%) and **Panoramic View** (+3.0%) are smaller and positive.

Table 7. Full model (D) coefficient estimates. HC3 robust standard errors. All variables standardised.

Variable	Coeff.	Std. Err.	t-stat	p-value	Impact (%)	
Panel A: Structural variables						
Bathrooms	0.2966	0.0110	26.86	< 0.001	+34.5	***
Half-bathrooms	0.1580	0.0072	21.92	< 0.001	+17.1	***
Description length	0.1013	0.0046	21.84	< 0.001	+10.7	***
Parking	0.0914	0.0094	9.68	< 0.001	+9.6	***
Stories	0.0744	0.0057	12.97	< 0.001	+7.7	***
Bedrooms	0.0314	0.0071	4.43	< 0.001	+3.2	***
Lot size	0.0033	0.7561	0.00	0.996	+0.3	
Panel B: Semantic similarities — positive price effects						
Modern/Contemporary	0.1517	0.0155	9.78	< 0.001	+16.4	***
Luxury	0.1330	0.0182	7.31	< 0.001	+14.2	***
Land & Nature	0.1225	0.0122	10.02	< 0.001	+13.0	***
Entry-Level	0.0991	0.0251	3.95	< 0.001	+10.4	***
Renovated	0.0501	0.0267	1.88	0.060	+5.1	
Waterfront	0.0462	0.0129	3.59	< 0.001	+4.7	***
Panoramic View	0.0300	0.0116	2.58	0.010	+3.0	**
Pool & Landscaping	0.0246	0.0161	1.52	0.128	+2.5	
Garage & Parking	0.0190	0.0086	2.19	0.028	+1.9	*
Heritage/Character	0.0103	0.0147	0.70	0.483	+1.0	
Panel C: Semantic similarities — negative price effects						
Family-Friendly	−0.1237	0.0127	−9.72	< 0.001	−11.6	***
New Construction	−0.1105	0.0176	−6.29	< 0.001	−10.5	***
Quiet & Peaceful	−0.0995	0.0142	−7.01	< 0.001	−9.5	***
Bright & Spacious	−0.0936	0.0185	−5.07	< 0.001	−8.9	***
Motivated Seller	−0.0890	0.0216	−4.12	< 0.001	−8.5	***
Needs Renovation	−0.0827	0.0170	−4.86	< 0.001	−7.9	***
Energy Efficient	−0.0812	0.0136	−5.95	< 0.001	−7.8	***
Premium Location	−0.0628	0.0139	−4.53	< 0.001	−6.1	***
Finished Basement	−0.0370	0.0147	−2.52	0.012	−3.6	*
Income/Investment	−0.0316	0.0162	−1.95	0.051	−3.1	

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Impact = $(e^{\hat{\beta}} - 1) \times 100\%$. n = 17,087; Adj. R^2 = 0.511.

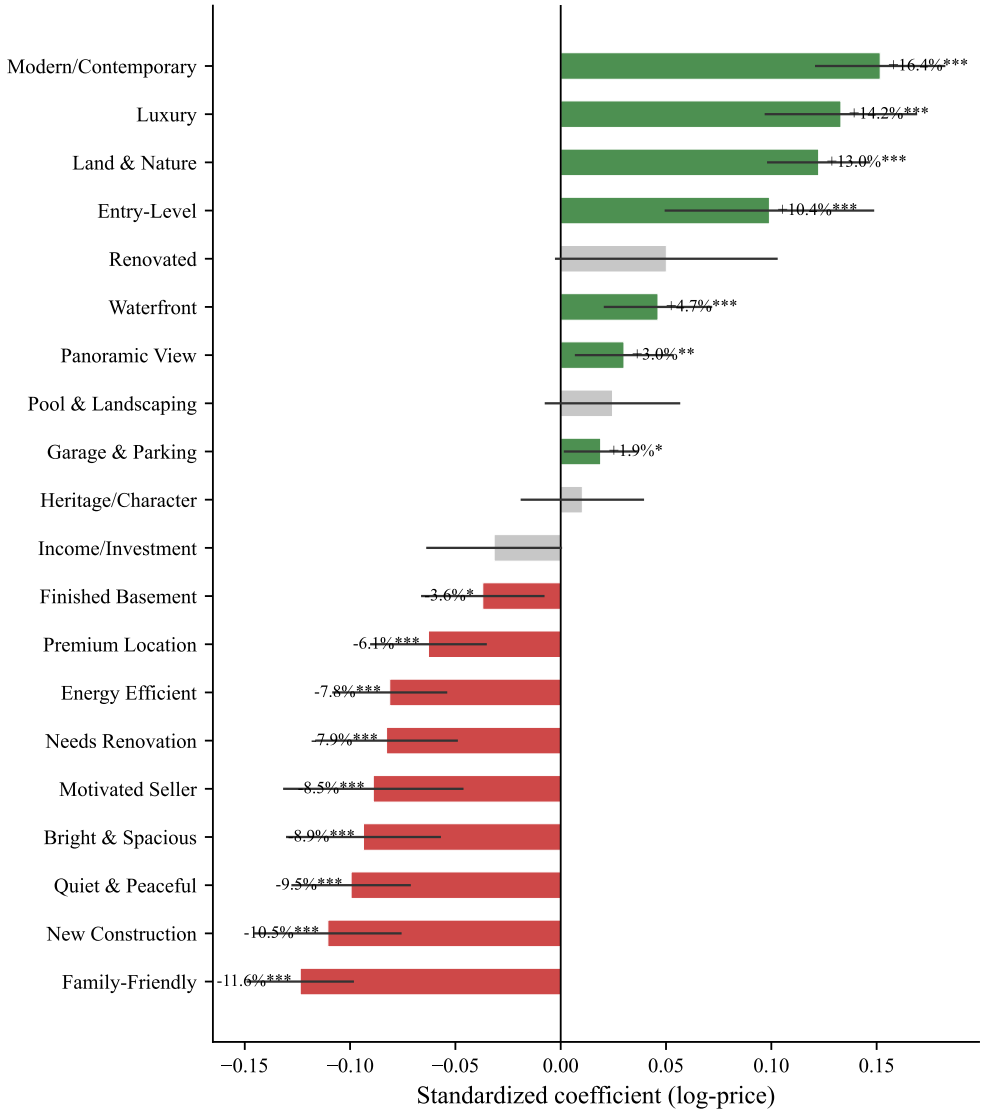


Figure 7. Price impact of the semantic similarity dimensions in Model D. Green bars indicate positive and red bars negative associations; grey bars are not significant at the 5% level. Error bars are 95% confidence intervals.

Several dimensions carry negative associations. **Family-Friendly** (−11.6%) and **New Construction** (−10.5%) are the largest; both describe segments—suburban family markets, peripheral new developments—whose price levels are lower, so that the language marks location rather than a disamenity. **Quiet & Peaceful** (−9.5%) likewise characterises rural and peripheral settings. **Motivated Seller** (−8.5%) and **Needs Renovation** (−7.9%) have the signs expected from the disclosure of urgency and of deferred maintenance. **Bright & Spacious** (−8.9%), **Energy Efficient** (−7.8%) and **Premium Location** (−6.1%) are negative against expectation, and **Entry-Level** (+10.4%) is positive against expectation. The first version of this paper did not discuss these four anomalies. They are the signature of the strong collinearity documented below: when 20 projections of the same embedding are correlated at 0.66–0.97, a coefficient measures the association with the part of a dimension that is orthogonal to the other nineteen, and that residual part need not carry the dimension's intended meaning. Entry-Level, for instance, is correlated at 0.97 with Motivated Seller, and its positive coefficient is the mirror image of Motivated Seller's negative one. Section 7 shows that Entry-Level, New Construction, Energy Efficient and Bright & Spacious change sign when the encoder is changed, whereas Luxury, Land & Nature, Family-Friendly, Quiet & Peaceful and Needs Renovation do not.

5.5. Semantic Profiles Across Price Quintiles

Figure 8 displays the mean similarity on each dimension by price quintile.

Lower-priced listings score higher on nearly every dimension. The first version attributed this gradient to longer descriptions; the language analysis of Section 4.3 shows that it is mainly the rising share of English listings in the upper quintiles. The relative differences across dimensions remain informative: from the bottom to the top quintile the steepest declines are for Motivated Seller, Entry-Level and Premium Location, and the smallest for Land & Nature, Modern/Contemporary and Panoramic View—the dimensions with positive coefficients in the regression.

5.6. Inter-Reference Correlations

Figure 9 presents the correlation matrix of the 20 similarities.

The correlations are uniformly positive and strong, from 0.66 to 0.97 with a mean of 0.84, and the first principal component of the 20 similarities accounts for 86% of their variance. The common component is, to a large extent, the language of the listing (Table 4): the first principal component is correlated at −0.82 with the English indicator and at only 0.20 with description length. The strongest pairwise correlations are between Entry-Level and Motivated Seller ($r = 0.97$), both characteristic of value-oriented French listings, and between Luxury and Renovated ($r = 0.95$), which share upgrade vocabulary. Individual semantic coefficients are therefore partial associations within a

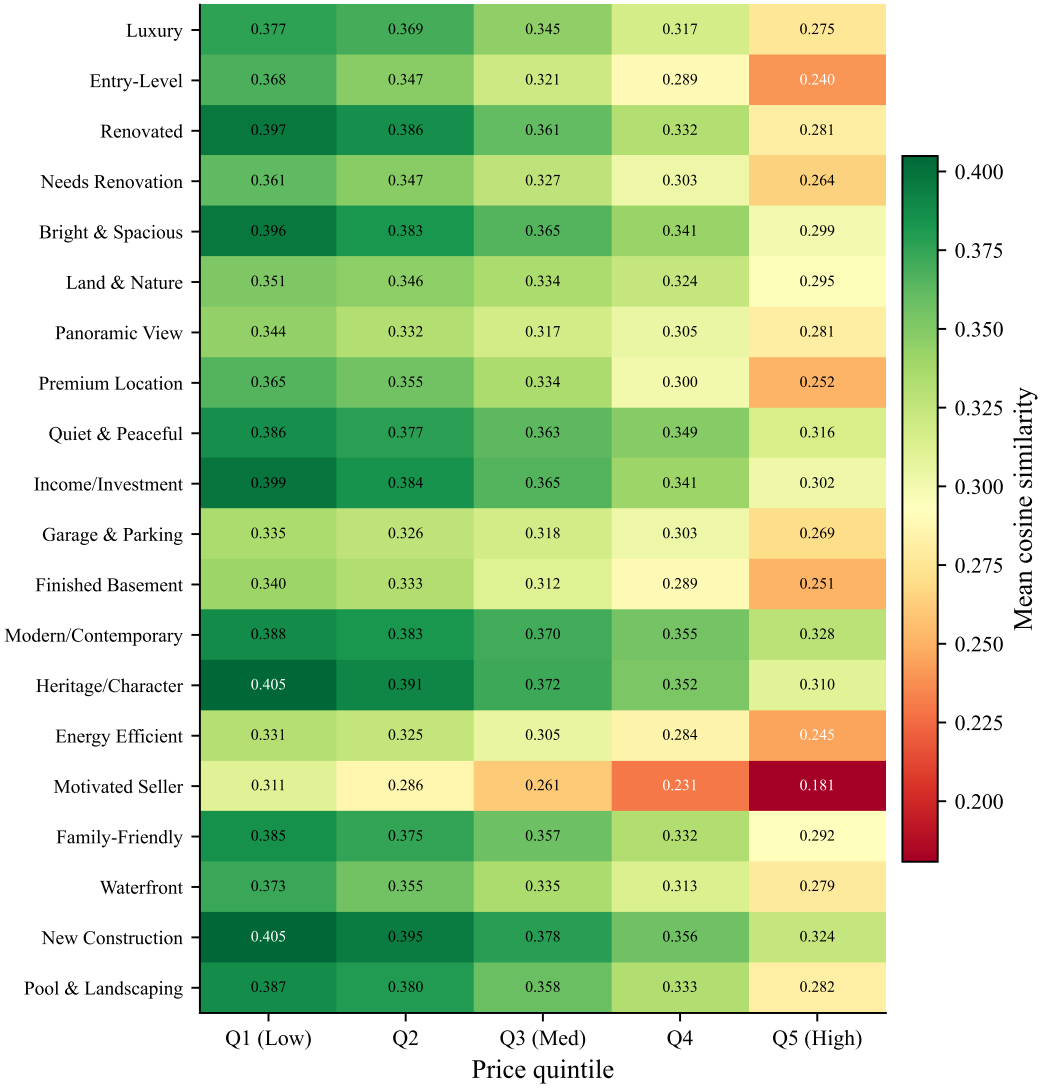


Figure 8. Mean cosine similarity by price quintile across the 20 reference dimensions.
Notes: Warmer colours indicate higher similarity. Each cell is the mean cosine similarity between the listings of a price quintile and the reference description.

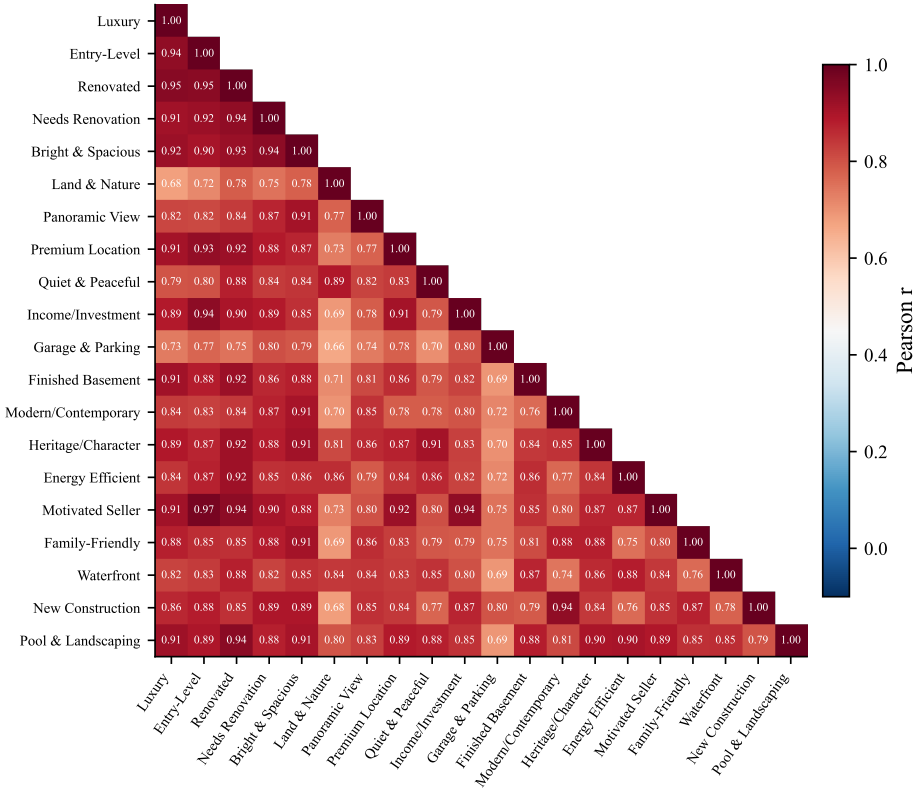


Figure 9. Pearson correlation matrix of the semantic similarity dimensions (lower triangle).
Notes: Correlations across all $n = 17{,}087$ listings. The uniformly strong positive correlations reflect a common component of the similarity scores—mainly the language of the listing, secondarily its length and register.

correlated block, and the block-level evidence is the most robust object of inference; Section 6 quantifies the consequences for standard errors and Section 7 for the stability of the coefficients.

5.7. Illustrative Bivariate Relationships and Residual Diagnostics

Figure 10 shows bivariate scatter plots for four dimensions; these relationships do not control for structure, length or language and are shown for illustration only. Figure 11 presents residual diagnostics for Model D. The residual-versus-fitted plot reveals heteroskedasticity, with larger dispersion at both ends of the fitted distribution, which motivates the HC3 standard errors; the normal quantile plot shows heavier-than-normal tails, especially in the upper tail, consistent with the right skew of housing prices even after the log transformation. Least squares remains consistent under these conditions and the robust standard errors remain valid.

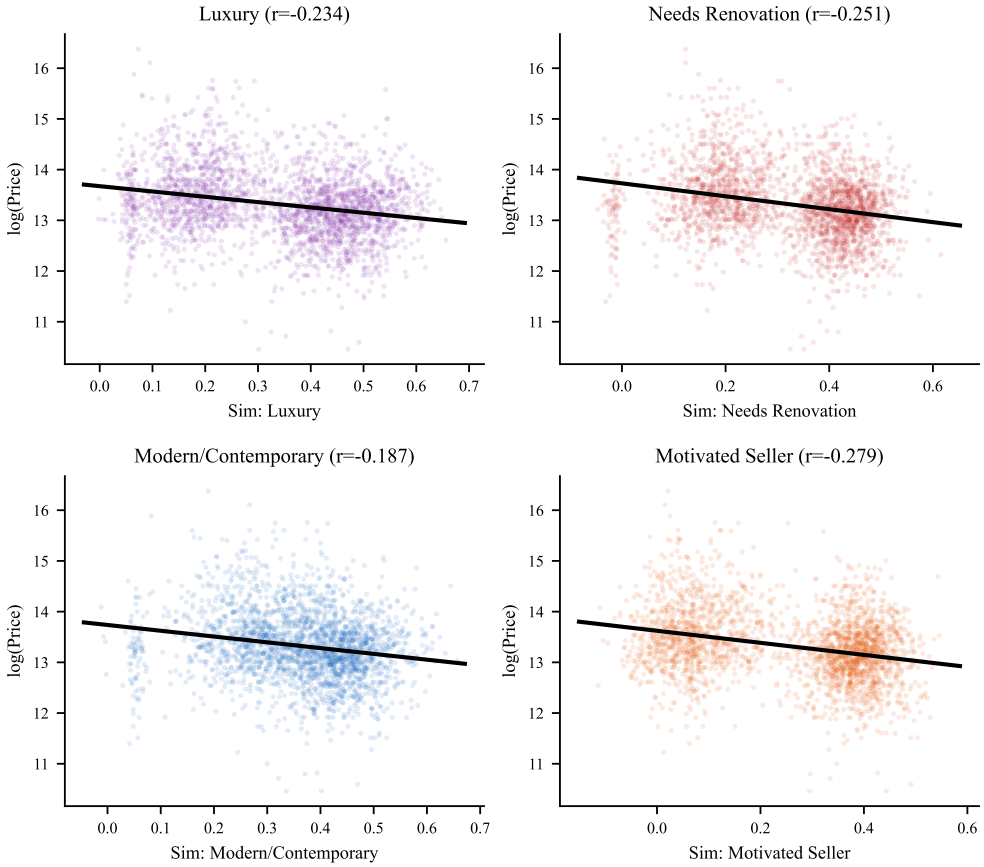


Figure 10. Bivariate relationships between selected semantic similarities and log price.
Notes: Lines are ordinary least squares fits. The relationships are bivariate and do not control for structural variables, description length or language.

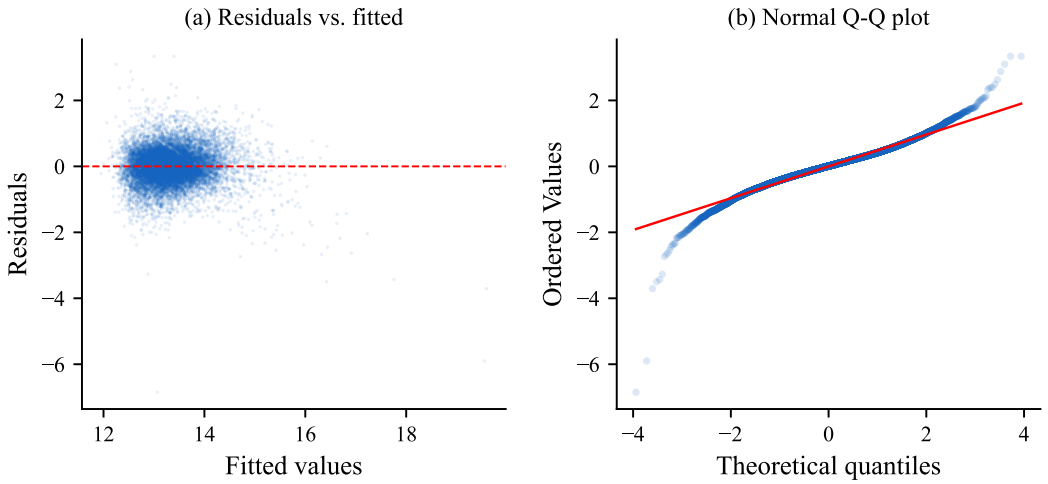


Figure 11. Residual diagnostics for Model D: (a) residuals against fitted values; (b) normal quantile plot.
Notes: Panel (a) plots least-squares residuals against fitted values and shows heteroskedasticity at the tails. Panel (b) compares the residual distribution with the normal and shows heavier-than-normal tails.

6. Robustness of the Baseline

This section reports the robustness checks of the first version of the paper, all re-executed by the replication pipeline: heteroskedasticity diagnostics, variance inflation factors, quantile regressions, trimming of extreme prices, bootstrap inference, regularised selection and a comparison with principal components of the embeddings. They address estimation and sampling issues within the baseline specification; the threats to interpretation raised by language, location and measurement are the subject of Section 7.

6.1. Heteroskedasticity

The Breusch–Pagan test (Breusch and Pagan, 1979) applied to Model D rejects homoskedasticity (LM = 768.0, $p < 0.001$), which justifies the HC3 standard errors used throughout. The pattern—larger residual variance at both tails of the fitted distribution (Figure 11, panel a)—is typical of housing price equations (Goodman and Thibodeau, 1995).

6.2. Variance Inflation Factors

The correlations among the similarities translate into substantial multicollinearity: the variance inflation factors of the 20 similarity variables in Model D average 17.2, 16 of them exceed 10, and the maxima are Renovated (41.6) and Entry-Level (40.7). This is a consequence of the design: all 20 features are projections of the same embedding and share its language and register. As O'Brien (2007) emphasises, a high VIF signals imprecision rather than bias and is not by itself grounds for respecification. Three facts bound the damage. Despite the inflated standard errors, 16 of the

20 coefficients are individually significant at 5%; the block-level evidence (joint tests, R^2 gain) is unaffected by collinearity within the block; and the coefficients are stable under trimming and bootstrap resampling. What collinearity does undermine is the attribution of effects to particular dimensions—a point that Section 7 makes concrete by showing that several coefficients depend on the reference phrasing and on the encoder.

6.3. Quantile Regressions

Table 8 reports quantile regressions of Model D (Koenker and Bassett, 1978; Koenker, 2005) at the 25th, 50th and 75th percentiles for the six dimensions with the largest baseline coefficients.

Table 8. Quantile regression coefficients for selected semantic dimensions.

Dimension	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	OLS	Pattern
Modern/Contemporary	0.148***	0.121***	0.155***	0.152***	High at both tails
Luxury	0.093***	0.130***	0.160***	0.133***	Increasing
Land & Nature	0.105***	0.098***	0.104***	0.122***	Stable
Family-Friendly	−0.121***	−0.105***	−0.108***	−0.124***	Mildly attenuating
Motivated Seller	−0.091***	−0.046*	−0.039	−0.089***	Attenuating
Needs Renovation	−0.090***	−0.065***	−0.044**	−0.083***	Attenuating

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Quantile regressions of Model D; standard errors follow the kernel-based estimator of Koenker and Bassett (1978).

Three patterns emerge. The Luxury coefficient increases monotonically with the conditional price level, from a 9.7% premium per standard deviation at $\tau = 0.25$ to 17.3% at $\tau = 0.75$, consistent with the finding of Zietz et al. (2007) that many attributes are valued more in upper segments. The Motivated Seller and Needs Renovation discounts attenuate at higher quantiles—the Motivated Seller coefficient falls from −8.7% at $\tau = 0.25$ to an insignificant −3.8% at $\tau = 0.75$. Land & Nature is stable across quantiles and Modern/Contemporary is larger at both tails than at the median. The New Construction coefficient, not shown in the table, becomes *more* negative at higher quantiles (−0.068, −0.079 and −0.142 at the three quantiles); the first version of this paper stated the opposite and used it to support a location-proxy reading of that coefficient. We withdraw that statement; the location-proxy reading is instead supported by the fixed-effects and multilingual results of Section 7.

6.4. Trimmed Prices

Re-estimating Model D after removing the 341 listings below the 1st or above the 99th percentile of price yields an adjusted R^2 of 0.486, lower than 0.511 because trimming removes price variation, with no sign reversal and all 16 significant similarities remaining significant. The largest coefficient changes are for Entry-Level (−0.024) and Motivated Seller (+0.023), both within the baseline confidence intervals. The results are not driven by the extreme prices noted in Section 4.

6.5. Bootstrap Inference

A nonparametric pairs bootstrap with 1,000 replications (Efron and Tibshirani, 1993), resampling the 17,087 listings with replacement, gives standard errors within 9% of the HC3 estimates for every semantic coefficient and within 5% for 15 of the 20; the percentile 95% intervals coincide closely with the HC3 intervals. Asymptotic inference is reliable under the observed heteroskedasticity. The bootstrap resamples listings independently and therefore does not address within-area dependence, which is handled by clustering in Section 7.

6.6. Regularised Selection

As an alternative to significance-based selection, Lasso (Tibshirani, 1996) and elastic-net (Zou and Hastie, 2005) regressions with five-fold cross-validated penalties retain only four similarities (Needs Renovation, Premium Location, Income/Investment and Family-Friendly) and shrink the other sixteen to zero. This is the textbook behaviour of ℓ_1 penalties under strong collinearity: one representative per correlated block is kept and its near-duplicates are discarded, so the retained variables stand for their blocks rather than for the "true" non-zero effects. The exercise confirms that the semantic block carries predictive signal that survives penalisation, and it reinforces the message that variable-by-variable attribution within the block is fragile.

6.7. Comparison with Principal Components

Table 9 compares the reference projections with alternative representations of the same text on the same sample.

Content matters more than quantity: length alone adds 1.2 points of R^2, content-based features 4.4–7.8 points. There is also a measurable price of interpretability: 20 principal components of the raw embeddings, which capture 49.5% of the embedding variance but have no economic names, fit better than the 20 reference similarities (ΔR^2 of +0.070 against +0.044 without the length control). The 20 similarities computed with the multilingual encoder (Section 7) close most of this gap (+0.069). The reference-based approach trades roughly two percentage points of in-sample fit for named features; Section 7 shows that the same ordering holds out of sample.

6.8. Summary

Table 10 summarises the checks of this section together with those of Section 7. Within the baseline specification, the joint explanatory power of the semantic block and the signs of the main coefficients are stable to the estimation method, to trimming and to the quantile considered. The diagnostics also delineate the limits of the evidence: the dimensions are strongly collinear, individual coefficients

Table 9. Comparison of text representation methods in the hedonic model. ΔR^2 is the improvement over Model A (structural only, $R^2 = 0.452$).

Text Representation	ΔR^2	Interpretability	Economic Usefulness
Description length only	+0.012	High	Low
PCA on embeddings (20 PC)	+0.070	None	Low
PCA on embeddings (20 PC) + length	+0.078	None	Low
Reference cosine (20 dim)	+0.044	High	High
Reference cosine (20 dim) + length	**+0.060**	**High**	**High**
Multilingual reference cosine (20 dim) + length	+0.069	High	High

Notes: All specifications include the six structural variables and are estimated on the same $n = 17{,}087$ sample. The 20 principal components capture 49.5% of the variance of the raw 384-dimensional embeddings. PCA achieves a larger fit improvement than the baseline reference projections but produces features with no economic interpretation; the reference-based approach trades roughly two percentage points of R^2 for named features. The multilingual row uses the similarities computed with paraphrase-multilingual-MiniLM-L12-v2 (Section 7) and closes most of the gap. None of these rows controls for the language of the listing; see Table 11 for specifications that do.

are measured imprecisely relative to the block, and regularised selection does not single out the same variables as significance testing.

7. Extended Specifications and Validation

The analyses in this section were added in the present revision. They address, in turn, the language of the listings, unobserved location and clustered inference, measurement of lot and floor area, the attribution of effects to domains, out-of-sample performance, the sensitivity of the estimates to the phrasing of the references, and the choice of encoder. Table 11 summarises the incremental explanatory power of the semantic block across all specifications; Table 12 and Figure 12 report the individual coefficients in the main ones.

7.1. The Language of the Listing

Section 4.3 established that 40% of the descriptions are in English and that English listings score about 0.23 lower on every French reference with the baseline encoder. Since English listings are also more expensive, every similarity is negatively correlated with price through language, and the regression coefficients of Model D are identified partly from the comparison of French with English listings. Adding indicators for English and mixed-language descriptions (specification D-L in Table 11) raises the R^2 from 0.512 to 0.518, and the English indicator itself carries a coefficient of +19.5% conditional on structure and length. More importantly, the incremental R^2 of the 20 similarities falls from 0.047 in Model D to 0.017 once language is controlled: about 63% of the

Table 10. Summary of robustness analyses.

Test	**Main Result Stable?**	**Notes**
HC3 robust standard errors	Yes	Baseline inference method
Breusch-Pagan test	Yes	Heteroskedasticity confirmed (LM = 768.0); HC3 justified
Variance inflation factors	Yes (block level)	Substantial collinearity (mean VIF 17.2; 16/20 > 10); joint inference unaffected, individual coefficients less precise
Bootstrap inference (1,000 rep.)	Yes	SE within 9% of HC3 (15/20 within 5%); CIs match
Trimmed prices (1st/99th pctile)	Yes	No sign reversals; all 16 significant dimensions remain significant
Quantile regression (τ = .25, .50, .75)	Yes	Luxury premium increasing in price; discounts attenuate at upper quantiles
Lasso / elastic net selection	Yes (block level)	Sparse selection of 4 block representatives; semantic signal survives penalisation
PCA comparison	Yes	PCA fits better (ΔR^2 +0.070 vs. +0.044) but is uninterpretable; reference approach preferred for inference
Added in the 2026 revision (Section 7)		
Language indicators	Yes (block level)	Semantic ΔR^2 falls from 0.047 to 0.017; block F = 22.6; Motivated Seller loses significance
FSA / municipality fixed effects	Yes (block level)	ΔR^2 0.020 / 0.016 within areas; clustered block F = 31.4 / 28.3
Clustered standard errors (FSA)	Yes	Semantic SEs 50% larger on average; 14/20 significant in Model D
Unit-corrected lot area, floor area	Yes (block level)	ΔR^2 0.013 (lot area) and 0.011 (floor-area subsample); Land & Nature falls to +4.2%
Language subsamples	Yes	French only ΔR^2 0.029; English only 0.024; core signs identical
Domain block tests	Yes	All six domains jointly significant; Market Signals weakest (p = 0.046 within FSAs)
Out-of-sample (random, FSA-grouped folds)	Yes	R^2_{oos} 0.489 → 0.528 (random), 0.484 → 0.522 (spatial); MdAPE 27.1% → 26.9%
Reference phrasing	Block yes, coefficients no	ΔR^2 0.042–0.054; coefficient correlations 0.02–0.81 with baseline
Multilingual encoder	Block yes, coefficients partly	ΔR^2 0.038–0.057; five core dimensions stable, several sign reversals

Table 11. Incremental explanatory power of the 20 semantic similarities across extended specifications.

Specification	***n***	R^2	**Within** R^2	ΔR^2 **(sem.)**	**Block** F	**Sig.**	**Cov.**
D: baseline (Table 6)	17,087	0.512	—	0.047	54.3	16/20	HC3
D, standard errors clustered by FSA	17,087	0.512	—	0.047	20.1	14/20	cluster
D-L: D + language indicators	17,087	0.518	—	0.017	22.6	16/20	HC3
D-L, standard errors clustered by FSA	17,087	0.518	—	0.017	17.6	15/20	cluster
D-L + FSA fixed effects	17,087	0.667	0.458	0.020	31.4	18/20	cluster
D-L + municipality fixed effects	17,087	0.710	0.456	0.016	28.3	14/20	cluster
D-L + FSA FE, unit-corrected lot area	17,087	0.725	0.552	0.013	25.5	14/20	cluster
D-L + FSA FE + floor area (subsample)	6,228	0.773	0.588	0.011	12.3	10/20	cluster
French listings only, D + FSA FE	9,843	0.599	0.413	0.029	24.4	15/20	cluster
English listings only, D + FSA FE	6,779	0.690	0.500	0.024	23.8	15/20	cluster
ML-D: multilingual encoder, as D	17,087	0.522	—	0.057	65.3	17/20	HC3
ML-D-L: multilingual encoder + language	17,087	0.538	—	0.038	54.0	16/20	HC3
ML-D-L + FSA fixed effects	17,087	0.687	0.489	0.039	70.6	15/20	cluster
ML-D-L + FSA FE, unit-corrected lot area	17,087	0.741	0.578	0.028	52.0	13/20	cluster

Notes: All specifications include the six structural variables and description length; "D-L" adds indicators for English and mixed-language listings; "ML" replaces the all-MiniLM-L6-v2 similarities by those computed with paraphrase-multilingual-MiniLM-L12-v2. Fixed effects are absorbed by within-group demeaning (415 forward sortation areas, FSA, or 1,061 municipalities); R^2 is then measured against the total variance of log price and "Within R^2" against the within-group variance. ΔR^2 (sem.) is the gain in R^2 from adding the 20 similarities to the otherwise identical specification. Block F is the robust Wald test (HC3 or clustered by FSA/municipality) that the 20 semantic coefficients are jointly zero; all $p < 0.001$. "Sig." counts similarities individually significant at the 5% level. The unit-corrected lot area converts the raw field to square metres (m^2, sq. ft., acres, hectares) and enters in logs with a missing-value indicator; the floor-area subsample is restricted to the 6,228 listings reporting interior floor area.

apparent semantic gain of the first version is attributable to the language composition of the corpus. The block remains jointly significant ($F = 22.6$ with HC3 standard errors, 17.6 when clustered by FSA; $p < 0.001$ in both cases), and 16 of the 20 coefficients remain individually significant.

The language subsamples confirm that the semantic information is not an artefact of the pooled comparison. Within French listings only ($n = 9{,}843$) and with FSA fixed effects, the block adds 0.029 to the R^2 and 15 coefficients are significant; within English listings only ($n = 6{,}779$) it adds 0.024, with 15 significant coefficients, even though the references are French and the encoder is monolingual. The signs of the core dimensions are the same in both languages (Table 12): Luxury and Land & Nature are positive and Family-Friendly, Quiet & Peaceful and Premium Location are negative in each subsample. Several other coefficients differ across languages, which is itself evidence that the monolingual projection does not measure the same thing in the two halves of the corpus.

7.2. Location Fixed Effects and Clustered Standard Errors

Specification D-L with 415 FSA fixed effects raises the R^2 to 0.667 (within R^2 0.458), and municipality fixed effects (1,061 groups) raise it to 0.710. Location thus explains far more of the variation in asking prices than the semantic block does—a useful corrective to the first version's suggestion that the text features were comparable in importance to neighbourhood controls. Within FSAs, the semantic block still adds 0.020 to the R^2 (0.016 within municipalities), with $F = 31.4$ and 18 of the 20 coefficients significant when standard errors are clustered by FSA. The English indicator falls from +19.5% to +9.1% within FSAs: about half of the English premium is the location of anglophone neighbourhoods, and half persists within them.

Clustering matters for inference but not for the conclusions. For the baseline Model D, clustering by FSA lowers the robust block F from 54.3 to 20.1 and the count of significant coefficients from 16 to 14; the standard errors of the semantic coefficients rise by 50% on average (between 22% and 107%). This is the expected consequence of within-area correlation of listing prices (Cameron and Miller, 2015), and we regard the clustered estimates as the appropriate ones for all specifications that use location (Abadie et al., 2023).

Table 12 shows how the individual coefficients respond. Five dimensions keep their sign and significance in every column: Luxury (+14.2% in the baseline, +10.7% within FSAs, +4.4% and +7.6% in the French and English subsamples), Land & Nature (+13.0%, +15.2%, +10.2%, +11.1%), Family-Friendly (−11.6%, −6.4%, −2.4%, −4.1%), Quiet & Peaceful (−9.5%, −11.7%, −9.4%, −5.3%) and Needs Renovation (−7.9%, −3.6%, insignificant in the French subsample, −3.7% in the English one). Modern/Contemporary and Waterfront remain positive and significant in all MiniLM specifications. Motivated Seller, by contrast, loses most of its baseline discount once language is controlled (−8.5% to −2.9%, insignificant) and is only marginally significant within FSAs (−4.0%):

Table 12. Semantic implicit price gradients (% per standard deviation) across specifications and encoders.

Dimension	**D**	**D-L**	**D-L + FSA FE**	**French only**	**English only**	**ML-D-L + FSA FE**
Luxury	14.2***	16.5***	10.7***	4.4***	7.6***	4.5**
Entry-Level	10.4***	15.6***	9.7***	0.7	10.3***	−4.9***
Renovated	5.1	5.8*	9.5***	4.3***	5.0*	−5.6***
Needs Renovation	−7.9***	−4.6**	−3.6*	−0.6	−3.7*	−4.5**
Bright & Spacious	−8.9***	−7.5***	−4.4*	−3.0**	−1.8	3.5*
Land & Nature	13.0***	16.2***	15.2***	10.2***	11.1***	22.1***
Panoramic View	3.0**	−1.4	−3.4*	0.7	−4.6***	0.3
Premium Location	−6.1***	−5.7***	−12.6***	−6.2***	−4.5***	−0.6
Quiet & Peaceful	−9.5***	−12.8***	−11.7***	−9.4***	−5.3***	−6.0***
Income/Investment	−3.1	−4.7**	2.9	2.1*	−1.6	4.7***
Garage & Parking	1.9*	3.3***	0.8	2.3***	−1.4	6.0***
Finished Basement	−3.6*	−5.4***	−4.9***	−3.4***	0.4	−10.3***
Modern/Contemporary	16.4***	14.8***	9.3***	7.6***	3.1*	−0.2
Heritage/Character	1.0	1.8	−2.6*	−1.1	−3.6*	4.3***
Energy Efficient	−7.8***	−6.1***	−5.1***	−2.0*	−3.9**	1.8*
Motivated Seller	−8.5***	−2.9	−4.0*	−1.3	−3.6*	0.4
Family-Friendly	−11.6***	−7.8***	−6.4***	−2.4***	−4.1***	−7.4***
Waterfront	4.7***	8.1***	15.5***	4.9***	9.4***	0.2
New Construction	−10.5***	−12.1***	−8.0***	−6.2***	−4.5**	7.2***
Pool & Landscaping	2.5	2.5	3.5*	2.4**	0.4	−11.9***
n	17,087	17,087	17,087	9,843	6,779	17,087
R^2	0.512	0.518	0.667	0.599	0.690	0.687
ΔR^2 (semantic block)	0.047	0.017	0.020	0.029	0.024	0.039

Notes: Entries are $100\,(e^{\hat{\gamma}_j} - 1)$ for a one-standard-deviation increase in similarity j (standard deviation computed on the estimation sample). D is the baseline full model (HC3 standard errors); D-L adds English and mixed-language indicators (HC3); "+ FSA FE" absorbs 415 forward-sortation-area fixed effects with standard errors clustered by FSA; the French-only and English-only columns re-estimate D with FSA fixed effects on each language subsample; ML-D-L replaces the encoder by paraphrase-multilingual-MiniLM-L12-v2. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

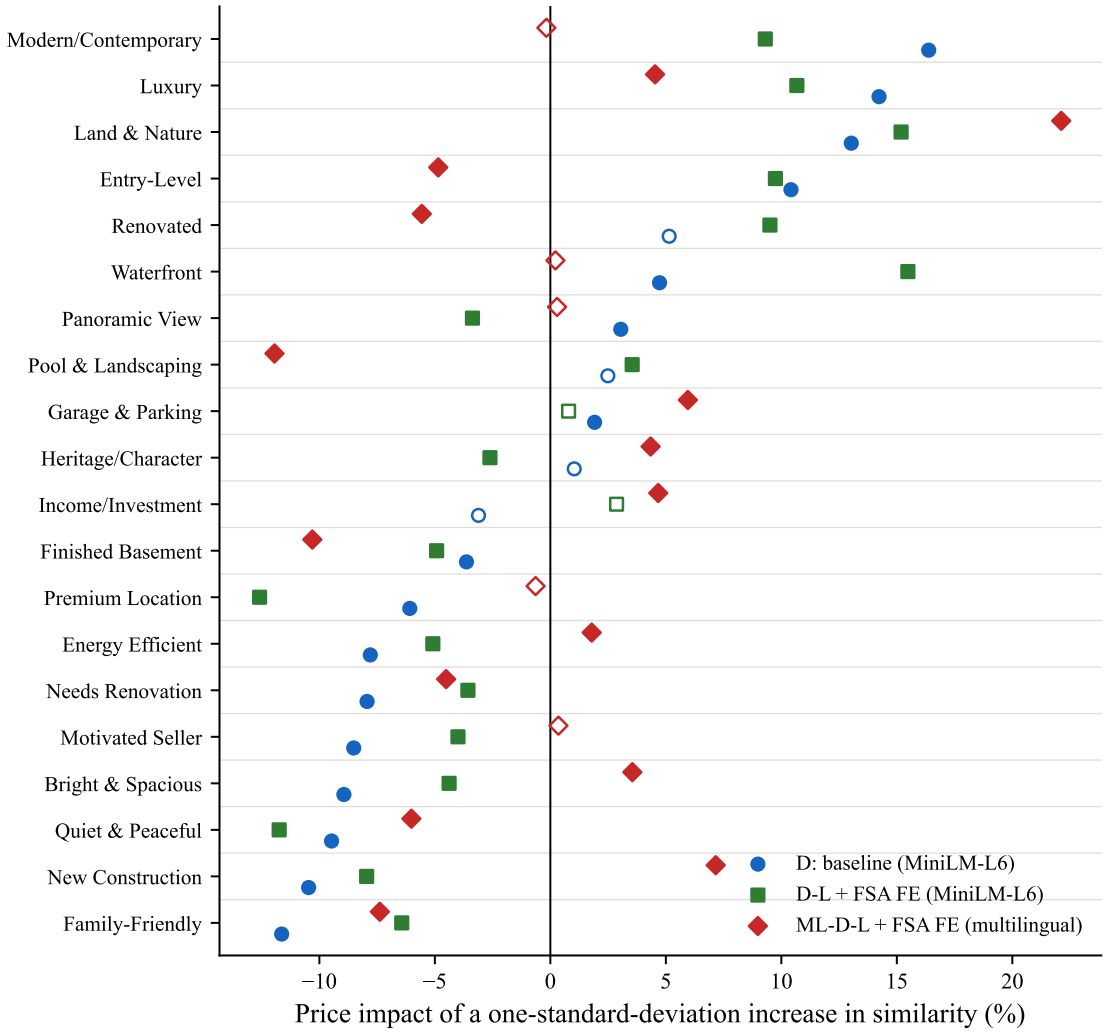


Figure 12. Semantic coefficients across specifications and encoders.

Notes: Price impact per standard deviation (Equation 10) of each dimension in the baseline Model D (HC3), in D-L with FSA fixed effects and FSA-clustered standard errors, and in the same specification with the multilingual encoder. Filled markers are significant at 5%; hollow markers are not. Dimensions are ordered by their baseline coefficient.

the urgency discount of the first version was in large part a language effect, Motivated Seller being the dimension with the largest French–English similarity gap (0.385 against 0.076). Panoramic View changes sign, and Premium Location becomes more negative within FSAs (−12.6%), which is consistent with its reference text (schools, parks, family neighbourhoods) marking suburban segments rather than premium ones.

7.3. Measurement of Lot and Floor Area

Replacing the raw lot-size field by the logarithm of the unit-corrected lot area raises the R^2 of the FSA specification from 0.667 to 0.725 (within R^2 from 0.458 to 0.552): a one-standard-deviation increase in log lot area is associated with a 40% higher asking price within an FSA, the largest structural coefficient in the model. The semantic block's increment falls to 0.013, and 14 coefficients remain significant. The Land & Nature coefficient falls from +15.2% to +4.2%: much of what the "large wooded lot" reference captured was the lot area itself, which the raw field had failed to measure. Waterfront and Pool & Landscaping, by contrast, keep or increase their coefficients (+10.9% and +11.6%), and Luxury (+10.4%) and Modern/Contemporary (+9.6%) are unchanged. On the 6,228 listings that report an interior floor area, adding its logarithm raises the R^2 to 0.773; the semantic block adds 0.011 and 10 coefficients remain significant. The semantic features therefore survive the inclusion of the two size variables that the baseline lacked or mismeasured, with a smaller incremental contribution.

7.4. Domain-Level Block Tests

Because the 20 dimensions are grouped into six domains (Table 2), Table 13 tests the joint significance of each domain's coefficients in specification D-L, with and without FSA fixed effects. Every domain is jointly significant in both specifications. Location & Setting is by far the strongest ($F = 67.8$ within FSAs), followed by Quality & Standing and Style & Character; Market Signals (Income/Investment and Motivated Seller) is the weakest and is only marginally significant within FSAs ($F = 3.1$, $p = 0.046$). Domain-level inference is less sensitive to the arbitrary allocation of shared vocabulary among references than dimension-level inference, and we recommend it as the primary unit of interpretation for reference-based designs.

7.5. Out-of-Sample Validation

Table 14 reports five-fold cross-validation with random folds and with folds that hold out entire FSAs (Equations 13–14). Two preliminary remarks. First, the raw lot-size field is not merely uninformative but harmful out of sample: with an extreme value in a held-out fold, the structural model's out-of-sample R^2 is −1.04; all other rows use the unit-corrected lot area, which also explains

Table 13. Domain-level block tests of the semantic dimensions.

		D-L (HC3)		D-L + FSA FE (clustered)	
Domain	q	F	p	F	p
Quality & Standing	2	51.60	< 0.001	18.77	< 0.001
Condition	2	4.19	0.015	7.85	< 0.001
Physical Attributes	4	13.16	< 0.001	8.51	< 0.001
Location & Setting	5	40.39	< 0.001	67.80	< 0.001
Style & Character	5	22.74	< 0.001	17.29	< 0.001
Market Signals	2	6.85	0.001	3.11	0.046
All 20 dimensions	20	22.61	< 0.001	31.44	< 0.001

Notes: Wald tests that the q coefficients of the dimensions belonging to each domain of Table 2 are jointly zero, under the D-L specification with HC3 standard errors and under D-L with 415 FSA fixed effects and standard errors clustered by FSA.

why the out-of-sample R^2 of the structural model (0.480) exceeds its in-sample counterpart in Table 6. Second, the sample is large relative to the number of regressors, so little shrinkage is expected and the in-sample gains should carry over.

Table 14. Out-of-sample performance: random and spatial five-fold cross-validation.

	Random folds			FSA-grouped folds		
Specification	R^2_{oos}	**RMSE**	**MdAPE**	R^2_{oos}	**RMSE**	**MdAPE**
A, raw lot-size field (for reference)	−1.036	1.003	27.5	−1.309	1.068	27.6
A: structural	0.480	0.507	27.3	0.475	0.509	27.3
B: A + description length	0.489	0.502	27.1	0.484	0.505	27.3
B-L: B + language indicators	0.523	0.485	26.9	0.518	0.488	27.0
D: B + 20 similarities	0.528	0.483	26.9	0.522	0.486	27.0
D-L: D + language indicators	0.534	0.480	26.6	0.528	0.483	26.7
B + 20 principal components of the embeddings	0.544	0.475	26.5	0.537	0.478	26.7
B-L + 20 principal components	0.544	0.475	26.5	0.537	0.478	26.7
ML-D: B + 20 multilingual similarities	0.534	0.480	26.8	0.528	0.483	26.8
ML-D-L: B-L + 20 multilingual similarities	0.551	0.471	26.4	0.545	0.474	26.5

Notes: Five-fold cross-validation on the n = 17,087 listings; predictions for each fold are made from a model estimated on the other four folds, and the metrics are computed on the pooled out-of-fold predictions (Equations 13–14). Random folds are drawn with seed 42; FSA-grouped folds hold out entire forward sortation areas, so that no held-out listing shares an FSA with the training data. RMSE is in log points; MdAPE is the median absolute percentage error of the price level. Except for the first row, the lot-size variable is the unit-corrected log lot area with a missing indicator; the raw field (mixed units, maximum 7.2×10^8) yields exploding out-of-fold predictions and a negative R^2. Principal components are estimated on the training folds only.

They do. Adding the 20 similarities to the structural-plus-length model raises the out-of-sample R^2 from 0.489 to 0.528 with random folds and from 0.484 to 0.522 with FSA-grouped folds, and lowers the median absolute percentage error from 27.1% to 26.9%. Adding language indicators to the structural-plus-length model achieves most of this on its own (0.523), and the similarities then add a further 0.011; the multilingual similarities add 0.028 over the language-augmented model and reach

the best performance of all specifications (0.551 random, 0.545 spatial, MdAPE 26.4%), slightly ahead of 20 principal components (0.544 and 0.537). The ordering of specifications is the same under spatial and random folds, and the spatial penalty is small (about 0.006 of R^2), so the predictive value of the text generalises to areas not seen in estimation. The gains in valuation accuracy are nevertheless modest in absolute terms—less than one percentage point of MdAPE—because the dominant source of error in these models is the absence of location and floor area, not the absence of text.

7.6. Sensitivity to the Phrasing of the References

The first version acknowledged that the 20 references embody one researcher's choices but did not quantify the consequence. Table 15 does so with three systematic perturbations: keeping only the first half of each reference's comma-separated phrases, only the second half, or reversing their order. The perturbed references are re-encoded and Model D re-estimated.

Table 15. Sensitivity of Model D to the phrasing of the reference descriptions.

Reference variant	Corr. of similarities	R^2	ΔR^2 (sem.)	Block F	Corr. of coefficients	Sign agreement	Sig.
Baseline (full references)	1.000	0.5122	0.048	54.3	1.000	100%	16/20
First half of each reference's phrases	0.970	0.5185	0.054	63.5	0.625	81%	15/20
Second half of each reference's phrases	0.929	0.5080	0.043	52.6	0.024	62%	11/20
Phrases in reverse order	0.987	0.5066	0.042	51.6	0.805	88%	11/20

Notes: Each reference text in Appendix Table A1 is split into its comma-separated phrases and re-encoded with all-MiniLM-L6-v2 after the indicated perturbation; Model D is then re-estimated with the 20 perturbed similarities. "Corr. of similarities" is the mean, over the 20 dimensions, of the correlation between perturbed and baseline similarity scores; "Corr. of coefficients" is the correlation between the 20 perturbed and baseline semantic coefficients; "Sign agreement" is the share of the 16 dimensions significant in the baseline whose coefficient keeps its sign; "Sig." counts dimensions significant at 5%. Block F is the HC3 Wald test of the 20 semantic coefficients.

The similarity scores are very stable—the perturbed and baseline scores correlate at 0.93 to 0.99 on average across dimensions—and so is the block-level information: the incremental R^2 of the block lies between 0.042 and 0.054 against 0.048 in the baseline, and the joint test is always decisive. The individual coefficients are not stable. Reversing the order of the phrases, which changes the texts least, yields coefficients correlated at 0.81 with the baseline and preserves 88% of the significant signs; keeping the first half yields a correlation of 0.63 and 81% sign agreement; keeping the second half yields a correlation of 0.02 and 62% sign agreement, with the Luxury coefficient turning negative and Land & Nature turning negative. The explanation is the one anticipated in Section 5: with 20 highly collinear regressors, a small change in the shared variance among references redistributes the block's explanatory power among its members, so that the vector of coefficients—though not their sum—depends on wording choices that have no economic content.

A reader who wishes to use the individual gradients of Table 7 should regard them as one draw from a distribution over admissible phrasings, whose dispersion this table approximates.

7.7. A Multilingual Encoder

Finally, we replace all-MiniLM-L6-v2 by paraphrase-multilingual-MiniLM-L12-v2, a model distilled so that translations of a sentence receive similar embeddings (Reimers and Gurevych, 2020), re-encode the references and the 17,087 descriptions, and recompute the 20 similarities ("ML" specifications in Table 11). The multilingual similarities correlate only moderately with the baseline ones (0.33 to 0.59 across dimensions), and they largely remove the language gap (Table 4). Their information content is greater: the block adds 0.057 to the R^2 of the structural-plus-length model, 0.038 once language is controlled, 0.039 within FSAs and 0.028 with the corrected lot area—roughly twice the increments obtained with the baseline encoder in the corresponding specifications—with block F-statistics of 52 to 71. Out of sample the multilingual specification is the best of all (Table 14).

The individual coefficients, however, differ from the baseline in ways that bear directly on the interpretability claim (last column of Table 12 and Figure 12). Luxury (+4.5%), Land & Nature (+22.1%), Family-Friendly (−7.4%), Quiet & Peaceful (−6.0%) and Needs Renovation (−4.5%) keep their signs and significance—the same five dimensions that were stable across the MiniLM specifications. Entry-Level becomes negative (−4.9%) and New Construction positive (+7.2%), which are the signs one would expect a priori and the opposite of the baseline; Energy Efficient and Bright & Spacious also change sign. Modern/Contemporary, Waterfront, Motivated Seller and Premium Location become insignificant, and Finished Basement and Pool & Landscaping become strongly negative. With a different encoder, the same reference texts define different directions in a different space, and only those dimensions whose vocabulary is both distinctive and strongly price-related survive the change. The multilingual encoder is the appropriate choice for this bilingual corpus, and we regard the ML specifications with language indicators and FSA fixed effects as the preferred estimates of this paper; but the lesson of the comparison is general. Reference-based projection delivers a robust, interpretable statement at the level of the block and of a few well-separated dimensions, not twenty robust implicit prices.

8. Discussion

8.1. What the Evidence Establishes

Three findings survive every specification in this paper. First, broker-written descriptions contain price-relevant information beyond the structural fields. The 20 semantic similarities are jointly significant in the baseline, within language subsamples, within forward sortation areas

and municipalities, with clustered standard errors, after correcting the lot-size measure, on the floor-area subsample, under every perturbation of the reference texts and with both encoders; and the information generalises out of sample, to random held-out listings and to held-out areas alike. Second, the magnitude of this information is smaller than the first version of the paper suggested and depends on what one controls for. The headline gain of six percentage points of R^2 over a structural model with no location variables shrinks to about two points once the language of the listing and area fixed effects are accounted for with the baseline encoder, and to about four points with a multilingual encoder; location explains far more of the variation in asking prices than text does. Third, the individual implicit price gradients are robust for a core of five dimensions—Luxury, Land & Nature, Family-Friendly, Quiet & Peaceful, Needs Renovation—and not for the others.

The interpretable dimensions tell a coherent story. Listings that use luxury vocabulary are priced higher conditional on structure, in every specification and with both encoders, with a gradient of roughly 4 to 14% per standard deviation depending on the controls; the quantile regressions show the gradient rising with the price level, consistent with Zietz et al. (2007) and McMillen (2008), and with the view of luxury finishes as positional goods whose marginal valuation rises with the quality of the surrounding bundle (Frank, 2007). Listings that describe a natural setting are priced higher; the gradient is large in the baseline (+13%) and falls to +4% once the lot area is measured correctly, which indicates that the "large wooded lot" reference was in part a proxy for lot size and in part a genuine amenity signal—a decomposition that structured data alone could not have provided. Using Quebec City transactions and site inspections, Des Rosiers et al. (2002) estimate landscaping premiums of roughly 4 to 12%; the text-based gradient, net of lot area, lies at the lower end of that range. Listings that need renovation are priced lower, with a gradient of 4 to 8%, the sign expected from the capitalisation of deferred maintenance (Nowak and Smith, 2017; Goodwin et al., 2014). Listings that market a family-friendly or quiet setting are priced lower, in every specification including those with area fixed effects, which means that these words mark lower-priced segments within areas as well as between them; the equilibrium-sorting logic of Kuminoff et al. (2013) and the evidence of Delmelle and Nilsson (2021) that advertisement text predicts neighbourhood composition both point the same way: listing language is as much about where a property is as about what it is.

8.2. What the Evidence Does Not Establish

The first version of this paper reported an 8.5% urgency discount for Motivated Seller language and compared it with the 3.7% premium that Levitt and Syverson (2008) find for agent-owned homes. The comparison does not survive the revision. Motivated Seller is the dimension with the largest French–English similarity gap, and its coefficient falls to an insignificant −2.9% once language is controlled, to −4.0% (marginally significant) within FSAs, and to zero with the multilingual encoder.

The listing-language literature does find that remarks signalling seller motivation are associated with lower prices and faster sales (Haag et al., 2000; Goodwin et al., 2014); our data do not allow us to confirm a robust magnitude for Quebec asking prices. Similarly, the counter-intuitive signs on New Construction (−10.5%) and Entry-Level (+10.4%) in the baseline reverse with the multilingual encoder (+7.2% and −4.9%), and the negative signs on Bright & Spacious and Energy Efficient reverse as well. These reversals are not evidence that one encoder is right and the other wrong; they are evidence that, in a block of 20 projections correlated at 0.66–0.97, the coefficient of any one dimension is defined by the vocabulary that happens not to be shared with the other nineteen, and that this residual vocabulary differs across encoders and across phrasings. The reference-phrasing experiment of Section 7 makes the same point from a different direction.

The design itself contributes to the problem. Appendix A shows that several references share phrases—Land & Nature includes "bord de l'eau, accès au lac" (Waterfront), Modern/Contemporary includes "construction neuve" and "écoénergétique" (New Construction, Energy Efficient), Premium Location includes "près des écoles, des parcs, quartier familial" (Family-Friendly). The principle of dimensional specificity stated in Section 3 was not fully respected, and these overlaps are a plausible mechanism for the sign anomalies: when Modern/Contemporary absorbs the positive valuation of new and efficient construction, New Construction and Energy Efficient are left with the residual association of peripheral new developments. A redesign with disjoint vocabularies, or an explicit orthogonalisation of the references in embedding space, would be the first step for future applications.

8.3. Language as a Confound in Multilingual Text-as-Data Applications

The most consequential finding of the revision is methodological. Reference-based projection, like any dictionary method, compares each document with a fixed anchor; when the corpus is bilingual and the anchors are monolingual, the comparison measures language before it measures content. In our data this single confound accounts for 86% of the variance of the mean similarity and for more than half of the apparent semantic gain, and it is correlated with price through the geography of the two language communities. The problem is not specific to housing: it arises whenever a text-as-data study pools documents written in different languages, or in different registers, and scores them against a single set of references (Gentzkow et al., 2019; Ash and Hansen, 2023). Three remedies are available and we have applied all three: control for language explicitly, estimate within language, and use an encoder trained to align languages (Reimers and Gurevych, 2020). The multilingual encoder is the most satisfactory because it removes the gap at the source and retains more information, but it also changes the individual coefficients, which shows that the choice of encoder is part of the specification and should be reported and varied as such.

8.4. Relation to Prior Findings

On the information content of text, our results agree with Nowak and Smith (2017), who show that words in agent comments proxy for unobserved quality, and with Shen and Ross (2021), who find substantial information about unobserved quality in property descriptions; our out-of-sample results quantify this information for Quebec asking prices at about four percentage points of R^2 and a fraction of a point of median absolute percentage error, net of language. On the magnitude of text gains relative to unrestricted representations, our principal-component and multilingual benchmarks reproduce the ordering reported by the prediction-oriented literature (Shen and Ross, 2021; Baur et al., 2023): unrestricted features fit better, and the reference approach pays a fit penalty of about two points for its names. On the stability of text features across contexts, Nowak et al. (2021) show that real estate dictionaries differ across space and time; our finding that the same references yield different coefficients across languages and encoders is the embedding-space analogue of their result. On distributional heterogeneity, the rising Luxury gradient agrees with Zietz et al. (2007) and McMillen (2008). On the parallel with listing photographs (Glaeser et al., 2018; Poursaeed et al., 2018), both modalities reveal quality that structured fields omit, and both face the attribution problem we document when their features are many and correlated.

8.5. Description Length

The positive coefficient on description length (+10.7% in the baseline, +7.9% within FSAs, +6.8% with the corrected lot area) is one of the more stable text effects in the paper. Conditional on what a description says and on the language it is written in, a longer description is associated with a higher asking price, which is consistent with length proxying for broker effort, property complexity or an abundance of features worth describing (Shen and Ross, 2021; Goodwin et al., 2014). The variable is truncated at approximately 700 characters, so the coefficient measures length variation within that range.

8.6. Practical Implications

For automated valuation, the results are encouraging but measured. Text features computed with a multilingual encoder improve out-of-sample R^2 by about three percentage points over a model with structure, length and language, and by about seven points over a structural model, at negligible computational cost; but the dominant sources of valuation error in these data are the absence of floor area and of fine location, and an AVM should address those first. Reference-based features have the advantage over principal components that they can be audited—an appraiser can read the reference that defines each one—but the disadvantage, documented here, that their individual weights are not stable to the phrasing of the references. For researchers, the recommended practice that emerges is

to (i) report the language composition of the corpus and control for it or use a language-aligned encoder; (ii) design references with disjoint vocabularies; (iii) report block- and domain-level tests as the primary inferential objects; (iv) report the sensitivity of individual coefficients to reference phrasing; and (v) validate out of sample with folds that respect the spatial structure of the data (Roberts et al., 2017). For brokers, the correlational nature of the estimates bears repeating: the coefficients describe how asking prices and descriptions co-vary, not what rewriting a description would do to the price a property fetches.

8.7. Limitations

Asking prices. The data are asking prices from a single snapshot of active listings. List and sale prices differ systematically, and the gap varies with market conditions and marketing time (Haurin et al., 2010); if the gap varies with the semantic content of the description—plausibly larger for listings that disclose urgency—the coefficients on sale prices would differ from those reported here. The estimates describe the pricing behaviour of sellers and brokers conditional on the text they wrote.

Identification. The cross-sectional design cannot separate the information, persuasion and omitted-location channels of Section 3.8, and the fixed effects absorb only between-area location; a spatial autoregressive or Durbin specification (LeSage and Pace, 2009; Dubé and Legros, 2014) would be needed to model the within-area dependence that the clustered standard errors only accommodate. Text is written after the asking price is set and by the same agent who sets it, so the estimates are associations in a jointly determined pair of outcomes. Within-property variation across successive listings, or platform experiments on listing text, would be required for causal statements; Alfano and Guarino (2022) point toward designs that exploit variation in listing composition.

Missing structural information. The baseline lacks floor area and measures lot size with mixed units. The revision corrects the second and addresses the first on a 36% subsample; on that subsample the semantic block's increment is about one percentage point of R^2, and a full-sample estimate with floor area would likely be of that order.

Encoder and language. The baseline encoder was trained mainly on English and the references are French; 40% of the corpus is English. We have documented the consequences and provided multilingual estimates, but a French-specific or domain-adapted model (Martin et al., 2020; Le et al., 2020) might extract more; Conneau et al. (2020) document material performance losses when general multilingual models are applied to specialised monolingual tasks.

Reference design. The 20 references are one operationalisation, with overlapping vocabularies, and the coefficients attached to them are sensitive to phrasing. The paper quantifies this dependence but does not resolve it; a principled procedure for designing or orthogonalising references is the natural next step.

Collinearity. The 20 similarities are correlated at 0.66–0.97 with variance inflation factors

averaging 17. Block- and domain-level conclusions are unaffected; dimension-level conclusions are, and we have restricted our claims accordingly.

Temporal stability. The data are a single cross-section from January 2026. The relation between text and prices may vary over the cycle—urgency language, for instance, may be more diagnostic when it is rarer—and longitudinal data would be needed to assess this.

9. Conclusion

This paper asked whether the free-text descriptions that brokers write for residential listings contain price-relevant information that structured hedonic variables omit, and whether that information can be extracted in a form that supports interpretation rather than only prediction. Using sentence embeddings of 17,087 single-family listings in Quebec, projected onto 20 researcher-defined reference descriptions, we find that the answer to the first question is yes and the answer to the second is a qualified yes.

The semantic block raises the adjusted R^2 of a structural hedonic model from 0.452 to 0.511 and improves out-of-sample prediction in random and spatially grouped cross-validation. But a large part of this gain reflects the language of the listing rather than its content: 40% of the descriptions are in English, they score systematically lower against French references, and they are written for more expensive properties. Net of language and of 415 area fixed effects, with standard errors clustered by area, the semantic block adds about two percentage points of R^2 with the baseline encoder and about four with a multilingual one, and it remains jointly significant in every specification, in each language subsample, after correcting the lot-size measure and after adding floor area. Five dimensions—Luxury, Land & Nature, Family-Friendly, Quiet & Peaceful and Needs Renovation—carry the same sign and remain significant across all specifications and both encoders. The others, including the urgency discount and the new-construction discount that the first version of this paper emphasised, do not: they depend on the phrasing of the references and on the encoder, as the sensitivity analyses show.

The methodological conclusion is twofold. Reference-based projection does what it promises at the level of the block and of a few well-separated dimensions: it turns an opaque embedding into named, auditable features whose joint information content is robust, and it does so at negligible cost. It does not deliver twenty stable implicit prices, because twenty projections of the same vector are too collinear for their individual coefficients to be invariant to wording and representation choices that have no economic content. The practical recommendations that follow—report and control for the language composition of the corpus or use a language-aligned encoder, design references with disjoint vocabularies, treat block- and domain-level tests as the primary inferential objects, report the sensitivity of individual coefficients to reference phrasing, and validate out of sample with spatially

grouped folds—apply to any application of continuous dictionaries in the text-as-data tradition (Gentzkow et al., 2019; Ash and Hansen, 2023), in housing or elsewhere: product descriptions in online marketplaces, job postings (Marinescu and Wolthoff, 2020), patent texts (Kelly et al., 2021), corporate disclosures (Loughran and McDonald, 2011; García et al., 2023).

Several extensions would strengthen the evidence. Transaction prices would replace the pricing behaviour of sellers by the valuation of buyers and allow the list-to-sale gap to be modelled as a function of text. Successive listings of the same property, or platform experiments, would offer within-property variation for causal statements, and buyer-side data on search behaviour would show whether the qualitative segments that brokers write to are the segments buyers search in (Piazzesi et al., 2020). A principled design of the references—disjoint vocabularies, orthogonalisation in embedding space, or data-driven anchors validated against expert labels—would address the attribution problem directly. A French-specific or domain-adapted encoder (Martin et al., 2020) and listing photographs (Glaeser et al., 2018; Poursaeed et al., 2018) would complete the measurement of quality from unstructured listing content. The replication pipeline accompanying the paper, which reproduces every number reported here from the raw database, is intended to make these extensions straightforward.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *The Quarterly Journal of Economics*, 2023, *138* (1), 1–35.

Alfano, Vincenzo and Massimo Guarino, "A Word to the Wise: Analyzing the Impact of Textual Strategies in Determining House Pricing," *Journal of Housing Research*, 2022, *31* (1), 88–112.

Anselin, Luc, *Spatial Econometrics: Methods and Models* Studies in Operational Regional Science, Dordrecht: Kluwer Academic Publishers, 1988.

Ash, Elliott and Stephen Hansen, "Text Algorithms in Economics," *Annual Review of Economics*, 2023, *15* (1), 659–688.

Athey, Susan and Guido W. Imbens, "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 2019, *11*, 685–725.

Bajari, Patrick and C. Lanier Benkard, "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach," *Journal of Political Economy*, 2005, *113* (6), 1239–1276.

Baur, Katharina, Markus Rosenfelder, and Bernhard Lutz, "Automated Real Estate Valuation with Machine Learning Models Using Property Descriptions," *Expert Systems with Applications*, 2023, *213*, 119147.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 2003, *3*, 993–1022.

Breusch, Trevor S. and Adrian R. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, 1979, *47* (5), 1287–1294.

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, Ayşe, "Specification and Estimation of Hedonic Housing Price Models," *Regional Science and Urban Economics*, 1992, *22* (3), 453–474.

Cheshire, Paul and Stephen Sheppard, "Capitalising the Value of Free Schools: The Impact of Supply Characteristics and Uncertainty," *The Economic Journal*, 2004, *114* (499), F397–F424.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, "Unsupervised Cross-Lingual Representation Learning at Scale," in "Proceedings of the 58th Annual Meeting of the ACL" 2020, pp. 8440–8451.

Cropper, Maureen L., Leland B. Deck, and Kenneth E. McConnell, "On the Choice of Functional Form for Hedonic Price Functions," *Review of Economics and Statistics*, 1988, *70* (4), 668–675.

Dell, Melissa, "Deep Learning for Economists," *Journal of Economic Literature*, 2025, *63* (1), 5–58.

Delmelle, Elizabeth C. and Isabelle Nilsson, "The Language of Neighborhoods: A Predictive-Analytical Framework Based on Property Advertisement Text and Mortgage Lending Data," *Computers, Environment and Urban Systems*, 2021, *88*, 101658.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in "Proceedings of NAACL-HLT" 2019, pp. 4171–4186.

Dubé, Jean and Diègo Legros, *Spatial Econometrics Using Microdata*, Wiley, 2014.

Dubin, Robin A., "Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms," *The Review of Economics and Statistics*, 1988, *70* (3), 466–474.

Efron, Bradley and Robert J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1993.

Ekeland, Ivar, James J. Heckman, and Lars Nesheim, "Identification and Estimation of Hedonic Models," *Journal of Political Economy*, 2004, *112* (S1), S60–S109.

Epple, Dennis, "Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products," *Journal of Political Economy*, 1987, *95* (1), 59–80.

Ethayarajh, Kawin, "How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings," in "Proceedings of EMNLP-IJCNLP" 2019, pp. 55–65.

Fama, Eugene F. and Kenneth R. French, "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 1993, *33* (1), 3–56.

Frank, Robert H., *Falling Behind: How Rising Inequality Harms the Middle Class*, University of California Press, 2007.

Gao, Tianyu, Xingcheng Yao, and Danqi Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in "Proceedings of EMNLP" 2021, pp. 6894–6910.

García, Diego, Xiaowen Hu, and Maximilian Rohrer, "The Colour of Finance Words," *Journal of Financial Economics*, 2023, *147* (3), 525–549.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Gibbons, Stephen, "Gone with the Wind: Valuing the Visual Impacts of Wind Turbines through House Prices," *Journal of Environmental Economics and Management*, 2015, *72*, 177–196.

Glaeser, Edward, Michael Scott Kincaid, and Nikhil Naik, "Computer Vision and Real Estate: Do Looks Matter and Do Incentives Determine Looks," NBER Working Paper 25174, National Bureau of Economic Research 2018.

Goodman, Allen C. and Thomas G. Thibodeau, "Age-Related Heteroskedasticity in Hedonic House Price Equations," *Journal of Housing Research*, 1995, *6* (1), 25–42.

Goodwin, Kimberly, Bennie Waller, and H. Shelton Weeks, "The Impact of Broker Vernacular in Residential Real Estate," *Journal of Housing Research*, 2014, *23* (2), 143–161.

Goodwin, Kimberly R., Bennie D. Waller, and H. Shelton Weeks, "Connotation and Textual Analysis in Real Estate Listings," *Journal of Housing Research*, 2018, *27* (2), 93–106.

Griliches, Zvi, ed., *Price Indexes and Quality Change: Studies in New Methods of Measurement*, Cambridge, MA: Harvard University Press, 1971.

Grossman, Sanford J., "The Informational Role of Warranties and Private Disclosure about Product Quality," *The Journal of Law and Economics*, 1981, *24* (3), 461–483.

Haag, Jerry T., Ronald C. Rutherford, and Thomas A. Thomson, "Real Estate Agent Remarks: Help or Hype?," *Journal of Real Estate Research*, 2000, *20* (1-2), 205–216.

Halvorsen, Robert and Henry O. Pollakowski, "Choice of Functional Form for Hedonic Price Equations," *Journal of Urban Economics*, 1981, *10* (1), 37–49.

Hansen, Stephen, Michael McMahon, and Andrea Prat, "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach," *The Quarterly Journal of Economics*, 2017, *133* (2), 801–870.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York: Springer, 2009.

Haurin, Donald R., Jessica L. Haurin, Taylor Nadauld, and Anthony Sanders, "List Prices, Sale Prices and Marketing Time: An Application to U.S. Housing Markets," *Real Estate Economics*, 2010, *38* (4), 659–685.

Hong, Jengei, Heeyoul Choi, and Woo sung Kim, "A House Price Valuation Based on the Random Forest Approach: The Mass Appraisal of Residential Property in South Korea," *International Journal of Strategic Property Management*, 2020, *24* (3), 140–152.

Irwin, Elena G., "The Effects of Open Space on Residential Property Values," *Land Economics*, 2002, *78* (4), 465–480.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 2021, *3* (3), 303–320.

Koenker, Roger, *Quantile Regression*, Cambridge University Press, 2005.

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

Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang, "Concept Bottleneck Models," in "Proceedings of ICML" 2020, pp. 5338–5348.

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

Kuminoff, Nicolai V., V. Kerry Smith, and Christopher Timmins, "The New Economics of Equilibrium Sorting and Policy Evaluation Using Housing Markets," *Journal of Economic Literature*, 2013, *51* (4), 1007–1062.

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.

Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab, "FlauBERT: Unsupervised Language Model Pre-Training for French," in "Proceedings of LREC" 2020, pp. 2479–2490.

LeSage, James P. and R. Kelley Pace, *Introduction to Spatial Econometrics*, Chapman & Hall/CRC, 2009.

Levitt, Steven D. and Chad Syverson, "Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions," *Review of Economics and Statistics*, 2008, *90* (4), 599–611.

Li, Bohan, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li, "On the Sentence Embeddings from Pre-Trained Language Models," in "Proceedings of EMNLP" 2020, pp. 9119–9130.

Loughran, Tim and Bill McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, 2011, *66* (1), 35–65.

_ and _, "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, 2016, *54* (4), 1187–1230.

Lundberg, Scott and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," arXiv preprint arXiv:1705.07874 2017.

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.

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.

Marinescu, Ioana and Ronald Wolthoff, "Opening the Black Box of the Matching Function: The Power of Words," *Journal of Labor Economics*, 2020, *38* (2), 535–568.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot, "CamemBERT: A Tasty French Language Model," in "Proceedings of the 58th Annual Meeting of the ACL" 2020, pp. 7203–7219.

McMillen, Daniel P., "Changes in the Distribution of House Prices over Time: Structural Characteristics, Neighborhood, or Coefficients?," *Journal of Urban Economics*, 2008, *64* (3), 573–589.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in "NeurIPS," Vol. 26 2013, pp. 3111–3119.

Milgrom, Paul R., "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, 1981, *12* (2), 380–391.

Muennighoff, Niklas, Nouamane Tazi, Loïc Magne, and Nils Reimers, "MTEB: Massive Text Embedding Benchmark," in "Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics" 2023, pp. 2014–2037.

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

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

Nowak, Adam D., Bradley S. Price, and Patrick S. Smith, "Real Estate Dictionaries Across Space and Time," *The Journal of Real Estate Finance and Economics*, 2021, *62* (1), 139–163.

O'Brien, Robert M., "A Caution Regarding Rules of Thumb for Variance Inflation Factors," *Quality & Quantity*, 2007, *41* (5), 673–690.

Pakes, Ariel, "A Reconsideration of Hedonic Price Indexes with an Application to PC's," *American Economic Review*, 2003, *93* (5), 1578–1596.

Palmquist, Raymond B., "Estimating the Demand for the Characteristics of Housing," *Review of Economics and Statistics*, 1984, *66* (3), 394–404.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep Contextualized Word Representations," in "Proceedings of NAACL-HLT" 2018, pp. 2227–2237.

Piazzesi, Monika, Martin Schneider, and Johannes Stroebel, "Segmented Housing Search," *American Economic Review*, 2020, *110* (3), 720–759.

Poursaeed, Omid, Tomáš Matera, and Serge Belongie, "Vision-based real estate price estimation," *Machine Vision and Applications*, 2018, *29* (4), 667–676.

Pryce, Gwilym and Sarah Oates, "Rhetoric in the Language of Real Estate Marketing," *Housing Studies*, 2008, *23* (2), 319–348.

Reimers, Nils and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in "Proceedings of EMNLP-IJCNLP" 2019, pp. 3982–3992.

_ and _, "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation," in "Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)" 2020, pp. 4512–4525.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" 2016, pp. 1135–1144.

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.

Rosiers, François Des, Marius Thériault, Yan Kestens, and Paul Villeneuve, "Landscaping and House Values: An Empirical Investigation," *Journal of Real Estate Research*, 2002, *23* (1-2), 139–162.

Rudin, Cynthia, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 2019, *1* (5), 206–215.

Shen, Lily and Stephen L. Ross, "Information Value of Property Description: A Machine Learning Approach," *Journal of Urban Economics*, 2021, *121*, 103299.

Sirmans, G. Stacy, Lynn MacDonald, David A. Macpherson, and Emily N. Zietz, "The Value of Housing Characteristics: A Meta Analysis," *Journal of Real Estate Finance and Economics*, 2006, *33* (3), 215–240.

Sirmans, Stacy, David Macpherson, and Emily Zietz, "The Composition of Hedonic Pricing Models," *Journal of Real Estate Literature*, 2005, *13* (1), 3–43.

Tibshirani, Robert, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, 1996, *58* (1), 267–288.

Tyrväinen, Liisa, Stephan Pauleit, Klaus Seeland, and Sjerp de Vries, "Benefits and Uses of Urban Forests and Trees," in Cecil C. Konijnendijk, Kjell Nilsson, Thomas B. Randrup, and Jasper Schipperijn, eds., *Urban Forests and Trees*, Berlin, Heidelberg: Springer, 2005, pp. 81–114.

Varian, Hal R., "Big Data: New Tricks for Econometrics," *The Journal of Economic Perspectives*, 2014, *28* (2), 3–28.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," in "Advances in Neural Information Processing Systems," Vol. 33 2020, pp. 5776–5788.

Yin, Wenpeng, Jamaal Hay, and Dan Roth, "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach," in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)" 2019, pp. 3912–3921.

Yoo, Sanglim, Jungho Im, and John E. Wagner, "Variable Selection for Hedonic Model Using Machine Learning Approaches: A Case Study in Onondaga County, NY," *Landscape and Urban Planning*, 2012, *107* (3), 293–306.

Zhu, Enwei, Jing Wu, Hongyu Liu, and Keyang Li, "A Sentiment Index of the Housing Market in China: Text Mining of Narratives on Social Media," *The Journal of Real Estate Finance and Economics*, 2023, *66* (1), 77–118.

Zietz, Joachim, Emily Norman Zietz, and G. Stacy Sirmans, "Determinants of House Prices: A Quantile Regression Approach," *The Journal of Real Estate Finance and Economics*, 2007, *37* (4), 317–333.

Zou, Hui and Trevor Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B*, 2005, *67* (2), 301–320.

Appendices

A. Reference Descriptions

Table A1 reproduces the verbatim French text of the 20 reference descriptions used to construct the semantic similarity features (Section 3). Publishing the exact reference texts makes the feature construction fully replicable and allows readers to assess the operationalisation of each qualitative dimension directly.

Table A1. The 20 reference descriptions (verbatim French text used for embedding).

Dimension	Reference description
Luxury	*Propriété de prestige haut de gamme avec finitions luxueuses, matériaux nobles, planchers de bois franc, comptoirs de quartz et granit, cuisine gastronomique, salle de bain spa avec douche en verre, domotique et système audio intégré. Résidence d'exception.*
Entry-Level	*Propriété abordable idéale pour premier acheteur, bon prix, petit budget, opportunité d'investissement, starter home, parfait pour débuter, prix compétitif, aubaine.*
Renovated	*Entièrement rénové, remis à neuf, nouvelles fenêtres, nouvelle toiture, plomberie et électricité refaites, cuisine et salle de bain rénovées, modernisé, mise à jour complète, rien à faire, clé en main, prêt à emménager.*
Needs Renovation	*À rénover, à rafraîchir, nécessite des travaux, bon potentiel, vendu tel quel sans garantie, handyman special, besoin de rénovation, fixer-upper, à mettre à son goût.*
Bright & Spacious	*Très lumineux, fenestration abondante, aires ouvertes, grands espaces, plafonds hauts, cathédrale, mezzanine, vaste salon, pièces spacieuses, beaucoup de rangement, walk-in.*
Land & Nature	*Grand terrain boisé, mature, aménagement paysager, piscine creusée, spa, terrasse, patio, cour arrière privée, jardin, haie de cèdres, intime, sans voisin arrière, vue sur la nature, bord de l'eau, accès au lac, rivière.*
Panoramic View	*Vue imprenable, vue panoramique, vue sur le fleuve, vue sur la montagne, vue sur la ville, skyline de Montréal, vue dégagée, vue spectaculaire, penthouse avec vue.*
Premium Location	*Emplacement de choix, quartier recherché, proche de tout, à distance de marche des commerces, restaurants, cafés, accès rapide au transport en commun, métro, autoroute, près des écoles, des parcs, quartier familial sécuritaire.*
Quiet & Peaceful	*Rue tranquille, cul-de-sac, quartier paisible, résidentiel, calme, intimité, retiré, campagne, nature, boisé, loin du bruit, environnement serein.*
Income/Investment	*Excellent investissement, revenu locatif, plex rentable, baux en cours, bon rendement, cash flow positif, logements loués, duplex triplex avec revenus, rapport qualité-prix.*
Garage & Parking	*Garage double, garage chauffé, stationnement intérieur, entrée de garage pavée, abri d'auto, grand garage, atelier dans le garage, espace de rangement au garage.*

Dimension	**Reference description**
Finished Basement	*Sous-sol entièrement aménagé, salle familiale au sous-sol, chambre supplémentaire, salle de cinéma, bureau, possibilité de logement au sous-sol, entrée indépendante, sous-sol avec salle de bain complète.*
Modern/Contemporary	*Design moderne, contemporain, architecture épurée, lignes droites, minimaliste, construction neuve, maison intelligente, écoénergétique, LEED, fenêtres panoramiques, toit plat.*
Heritage/Character	*Cachet d'époque, maison ancestrale, patrimoine, boiseries d'origine, moulures, foyer d'origine, plafonds de 10 pieds, charme victorien, pierre, brique.*
Energy Efficient	*Écoénergétique, thermopompe, géothermie, panneaux solaires, isolation supérieure, fenêtres Energy Star, chauffage radiant, faible consommation, certifié Novoclimat, réservoir d'eau chaude récent, coûts énergétiques bas.*
Motivated Seller	*Vendeur motivé, vente rapide, prix réduit, réduction de prix, succession, relocalisation, doit vendre, reprise de finance, offres multiples bienvenues, ne manquez pas cette occasion.*
Family-Friendly	*Maison familiale, quartier familial, parc pour enfants, cour clôturée, école à proximité, garderie, aire de jeux, voisinage sécuritaire, idéal pour famille avec enfants.*
Waterfront	*Bord de l'eau, accès au lac, vue sur le fleuve, rivière, quai privé, droits nautiques, plage, navigable, chalet au bord du lac, waterfront, pieds dans l'eau.*
New Construction	*Construction neuve, maison neuve, jamais habitée, modèle de démonstration, garantie GCR, livraison prochaine, choix de finitions, plans personnalisables, nouveau développement.*
Pool & Landscaping	*Piscine creusée chauffée, piscine hors terre, spa, cuisine extérieure, terrasse en composite, pergola, aménagement paysager professionnel, pavé uni, foyer extérieur.*

B. Parsimonious Model Estimates

Table A2 reports the complete coefficient estimates for the parsimonious specification (Model E), which retains the 16 semantic dimensions that are individually significant at the 5% level in the full model (Model D).

Table A2. Parsimonious model (E) coefficient estimates. HC3 robust standard errors. All variables standardised.

Variable	**Coeff.**	**Std. Err.**	p**-value**	**Impact (%)**	
Panel A: Structural variables					
Bathrooms	0.2962	0.0110	< 0.001	+34.5	***
Half-bathrooms	0.1581	0.0072	< 0.001	+17.1	***
Description length	0.1017	0.0046	< 0.001	+10.7	***
Parking	0.0914	0.0095	< 0.001	+9.6	***
Stories	0.0746	0.0058	< 0.001	+7.7	***
Bedrooms	0.0314	0.0071	< 0.001	+3.2	***
Lot size	0.0035	0.7525	0.996	+0.3	
Panel B: Semantic similarities (significant at 5% in Model D)					
Modern/Contemporary	0.1590	0.0152	< 0.001	+17.2	***
Luxury	0.1552	0.0158	< 0.001	+16.8	***
Land & Nature	0.1254	0.0119	< 0.001	+13.4	***
Entry-Level	0.0823	0.0248	< 0.001	+8.6	***
Waterfront	0.0470	0.0130	< 0.001	+4.8	***
Panoramic View	0.0272	0.0114	0.017	+2.8	*
Garage & Parking	0.0085	0.0076	0.262	+0.9	
Finished Basement	−0.0295	0.0138	0.033	−2.9	*
Premium Location	−0.0614	0.0137	< 0.001	−6.0	***
Energy Efficient	−0.0669	0.0130	< 0.001	−6.5	***
Needs Renovation	−0.0743	0.0154	< 0.001	−7.2	***
Motivated Seller	−0.0816	0.0196	< 0.001	−7.8	***
Bright & Spacious	−0.0836	0.0179	< 0.001	−8.0	***
Quiet & Peaceful	−0.0882	0.0116	< 0.001	−8.4	***
Family-Friendly	−0.1142	0.0110	< 0.001	−10.8	***
New Construction	−0.1279	0.0165	< 0.001	−12.0	***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Impact = $(e^{\hat{\beta}} - 1) \times 100\%$. n = 17,087; Adj. R^2 = 0.511.