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

May 2026

Version 1.1

Abstract

This paper develops an interpretable NLP approach to hedonic pricing that projects sentence-transformer embeddings of property listing descriptions onto 20 researcher-defined semantic references, producing economically meaningful features. Using 17,087 single-family listings from Quebec, Canada, I show that adding cosine-similarity measures to a standard hedonic model increases adjusted R^2 from 0.452 to 0.511 ($F = 99.53$, $p < 0.001$). Modern/contemporary language is associated with +16.4% prices per standard deviation; luxury +14.2%; land/nature +13.0%. Renovation-need language is associated with −7.9% and seller urgency −8.5%. Quantile regressions reveal heterogeneous effects across the price distribution, and the estimates are robust to bootstrap inference and outlier trimming.

Keywords: hedonic pricing, NLP, sentence embeddings, cosine similarity, 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 the seminal contributions of Rosen (1974) and Lancaster (1966), building on a quality-adjustment tradition that dates back to the earliest hedonic price indexes (Griliches, ed, 1971). The standard approach decomposes observed transaction prices into the marginal contributions of structural attributes (bedrooms, bathrooms, lot size), locational factors (neighborhood 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 explain the largest shares of price variation (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 that price trillions of dollars of residential real estate (Kok et al., 2017).

Yet a persistent limitation of hedonic models is that they can only price attributes that the econometrician observes and measures. Property listing descriptions—the free-text narratives written by real estate agents to market properties—contain qualitative information that structured data fields fail to capture. These descriptions communicate the quality of finishes, the ambiance of a neighborhood, the motivation of the seller, the architectural character of a home, and the lifestyle a property affords. A listing that describes "spa-like master bathroom with heated marble floors" conveys quality information fundamentally different from one that states "bathroom needs updating." Both properties may have the same number of bathrooms, yet the hedonic price contribution differs enormously. As Pakes (2003) argued, standard hedonic regressions yield biased implicit price estimates when important characteristics are unobservable to the econometrician but observable to market participants—precisely the situation that listing text addresses. Bajari and Benkard (2005) formalize this unobserved-characteristics problem in a structural hedonic setting; listing narratives offer a direct, scalable measurement of the missing quality 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 input for measurement in finance (Loughran and McDonald, 2016), monetary policy (Hansen et al., 2017), and labor markets (Marinescu and Wolthoff, 2020). Despite this momentum, the hedonic pricing literature has been slow to incorporate unstructured text, and the approaches used to date face a fundamental tradeoff. Keyword and vernacular methods (Haag et al., 2000; Nowak and Smith, 2017; Goodwin et al., 2014) are interpretable but semantically shallow: they cannot recognize that "quartz countertops" and "premium stone surfaces" convey similar quality information. Sentiment and connotation measures (Goodwin et al., 2018; Zhu

et al., 2023) reduce multidimensional quality descriptions to one or two polarity scores. Topic models (Blei et al., 2003) capture latent themes but produce factors that are difficult to interpret and unstable in short, formulaic listing texts. Most recently, machine learning pipelines built on rich text representations (Shen and Ross, 2021; Baur et al., 2023; Devlin et al., 2019) extract substantially more price-relevant information, but the resulting features are opaque—effective for prediction but unsuitable for economic inference where the goal is to estimate the implicit prices of identifiable housing characteristics.

This paper proposes a *concept-projection* approach to text-augmented hedonic pricing. Rather than using transformer embeddings as black-box predictors, we project them onto economically interpretable semantic anchors. The framework proceeds in four stages. First, we encode each property listing's *PublicRemarks* field into a dense 384-dimensional vector using the all-MiniLM-L6-v2 sentence transformer (Wang et al., 2020). Second, we define 20 reference descriptions, each carefully designed to embody a specific qualitative housing dimension—luxury finishes, renovation status, architectural style, waterfront access, seller motivation, and others. Third, we compute the cosine similarity between each listing's embedding and each reference embedding, yielding 20 interpretable scalar features per property. Fourth, we integrate these similarity features into a semi-logarithmic hedonic pricing model estimated by OLS with heteroskedasticity-robust standard errors. In the taxonomy of text-as-data methods, the approach is a semantic generalization of the dictionary method (Gentzkow et al., 2019; Loughran and McDonald, 2016): each reference description plays the role of a word list, but the matching is performed in embedding space, so paraphrases, synonyms, and cross-lingual equivalents load on the intended dimension. The objective is not to claim that changing listing words mechanically changes prices, but to measure whether semantic content embedded in listing narratives captures price-relevant information omitted from structured hedonic variables.

Quebec provides a particularly interesting empirical context for this analysis. The province's real estate market exhibits substantial heterogeneity—from luxury urban properties in Montréal and Québec City to affordable family homes in suburban developments and waterfront properties in the Laurentians and Eastern Townships. Listing descriptions are predominantly in French with occasional English phrases, creating a bilingual corpus that tests the cross-lingual capabilities of the embedding model. The market spans urban, suburban, and peripheral segments with distinct price levels and qualitative characteristics, offering rich variation for semantic analysis. A well-established hedonic research tradition using Quebec microdata (Des Rosiers et al., 2002; Dubé and Legros, 2014) provides local benchmarks for the amenity values our text-based measures should—and, as we show, do—recover.

Our contributions are threefold. First, we demonstrate that semantic features derived from listing text capture economically significant information beyond what structural variables measure.

Adding the 20 cosine similarity variables to a standard hedonic specification increases the adjusted R^2 from 0.452 to 0.511 for 17,087 single-family homes ($F = 99.53$, $p < 0.001$). The semantic features contribute 9.3% of the full model's explained variance—comparable to the gains from adding neighborhood controls in typical hedonic studies (Sirmans et al., 2006).

Second, we estimate the implicit price gradients associated with qualitative housing dimensions that have been largely invisible to the hedonic literature. Listings semantically closer to modern/contemporary references are associated with 16.4% higher prices per standard deviation; luxury-oriented language with 14.2% premiums; and land/nature descriptions with 13.0% premiums. Conversely, renovation-need language is associated with 7.9% discounts and seller-urgency language with 8.5% discounts. Quantile regressions reveal that these associations are heterogeneous across the price distribution: luxury premiums are amplified at upper quantiles, consistent both with quality complementarities and with the broader finding that implicit prices vary across the conditional price distribution (Zietz et al., 2007; McMillen, 2008).

Third, the reference-based approach offers a methodological contribution that resolves the depth-interpretability tension in text-augmented hedonic models. Unlike PCA-based or neural-network-based approaches that yield opaque features, each of our similarity measures corresponds to a named, researcher-defined qualitative dimension with a clear coefficient estimate—an ex ante interpretable design in the spirit advocated by Rudin (2019) for high-stakes applications. This makes the results directly useful for real estate appraisal, market analysis, and automated valuation models (AVMs), and the methodology is portable to any domain where free-text descriptions accompany structured economic data. In doing so, the paper speaks simultaneously to three audiences: housing economists seeking to price unobserved quality, text-as-data practitioners seeking interpretable feature construction, and the interpretable machine learning community seeking economically grounded applications.

The remainder of the paper is organized as follows. Section 2 reviews the four literatures this work draws on. Section 3 describes the methodological framework, including the identification strategy. Section 4 presents the data and descriptive statistics. Section 5 reports the empirical results. Section 6 presents a comprehensive robustness analysis. Section 7 discusses interpretation, implications, and limitations. Section 8 concludes with directions for future research.

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 systematized for quality-adjusted price measurement by Griliches, ed (1971). Its theoretical foundations rest on Lancaster (1966), who reconceptualized consumer theory around the characteristics of goods rather than goods themselves, and Rosen (1974), who formalized 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 3.7: 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 neighborhood 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. Closer to our setting still, a sustained research program using Quebec City transactions has quantified the capitalization of property attributes ranging from landscaping to accessibility (Des Rosiers et al., 2002). In the amenity-capitalization literature more broadly, Cheshire and Sheppard (2004) demonstrate school-quality capitalization 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 neighborhood 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), penalized 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 organizes 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)—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, and Marinescu and Wolthoff (2020) show that the words of job postings materially affect matching in labor markets.

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. 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 summarize 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 neighborhood characteristics beyond the property itself: Delmelle and Nilsson (2021) show that the text of property advertisements predicts neighborhood 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 3.7.

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 optimized 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. For French text, the model's multilingual coverage—derived from its training on paraphrase data spanning multiple languages—provides adequate performance, though dedicated French models such as CamemBERT (Martin et al., 2020) or FlauBERT (Le et al., 2020) could potentially improve semantic resolution for domain-specific vocabulary.

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-normalized 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 normalized 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 prioritized 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 summarizes how our approach compares with existing methods along five dimensions.

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

Method	Semantic Depth	Interpretability	Scalability	Multilingual	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 resolves the depth-interpretability tension by leveraging the deep semantic representations of transformer embeddings while producing features with clear economic interpretation. 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. Methodological Overview

Our framework transforms unstructured listing text into a set of interpretable features through a four-stage pipeline, illustrated in Figure 1: (1) extraction and preprocessing of listing descriptions; (2) encoding into dense vector representations using a pre-trained sentence transformer; (3) computation of cosine similarities against a set of researcher-designed reference descriptions; and (4) integration of the resulting similarity features into a hedonic pricing model.

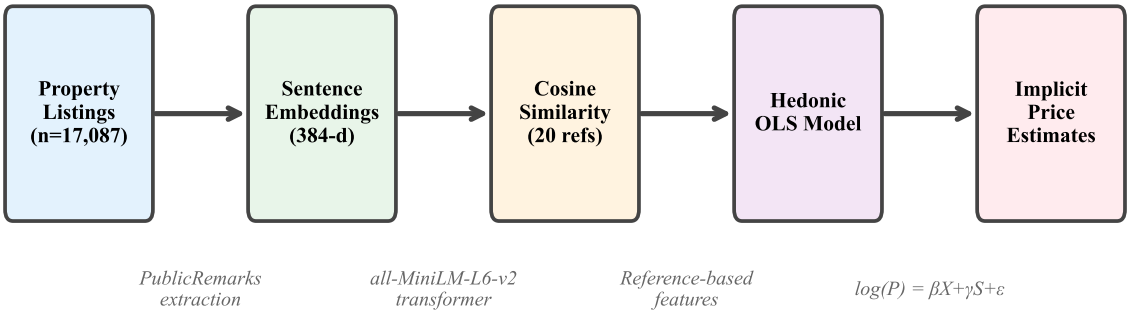


Figure 1. Methodological framework: from listing text to hedonic price estimates.
Notes: This figure illustrates the four-stage pipeline: (1) text extraction, (2) sentence embedding, (3) cosine similarity computation against 20 reference descriptions, and (4) integration into a hedonic pricing model.

3.2. Text Extraction and Preprocessing

Each property listing in our database includes a *PublicRemarks* field containing the agent-written description. We extract this field and 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 aggressive text preprocessing (stemming, lemmatization, stop-word removal) because the sentence transformer model is designed to process natural-language text and internally handles tokenization.

In addition to the embedding-based features, we compute a simple text length measure (character count of the description) as a control variable. Description length may proxy for listing effort, property complexity, or market conditions, and including it ensures that our semantic features capture content rather than verbosity.

3.3. Sentence Embedding

Each listing's description d_i is encoded into a 384-dimensional unit vector using the all-MiniLM-L6-v2 sentence transformer:

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

The model processes text through 6 transformer layers with 384 hidden dimensions, applying mean pooling over token embeddings and L_2 normalization to produce the final sentence embedding. The model handles both English and French text, which is important for our Quebec-based corpus where listings are predominantly in French with occasional English phrases.

Embeddings are computed in batches of 256 for computational efficiency. The entire corpus of 17,087 descriptions is encoded in approximately 14 seconds on consumer hardware (Apple M-series CPU), making the approach practical for large-scale applications.

3.4. Reference Description Design

3.4.1. Rationale

The key methodological innovation is the construction of reference descriptions that serve as semantic anchors. Rather than applying unsupervised dimensionality reduction (PCA, autoencoders) to the 384-dimensional embedding space—which would yield features without natural economic interpretation—we project each listing's embedding onto a set of researcher-defined semantic axes. Each axis is defined by a reference description that embodies a specific qualitative housing dimension.

This approach is analogous to the construction of factor-mimicking portfolios in asset pricing (Fama and French, 1993): just as Fama-French factors are portfolios designed to load on specific risk dimensions, our reference descriptions are synthetic texts designed to load on specific quality dimensions. Within the text-as-data toolkit, the design is a semantic generalization of the dictionary method (Loughran and McDonald, 2011; Gentzkow et al., 2019): each reference plays the role of a curated word list, but the matching operates in embedding space, so paraphrases and synonyms contribute to the intended dimension without being enumerated. It is likewise related to anchor-based zero-shot classification, in which label descriptions serve as fixed textual anchors (Yin et al., 2019), except that we retain the continuous similarity scores as regression covariates instead of discretizing them into class labels.

3.4.2. Reference Categories

We define 20 reference descriptions spanning six broad domains of housing quality, listed in Table 2. Each reference is a synthetic paragraph of 150–260 characters, written in French to match the listing

corpus, containing the vocabulary, phrasing, and semantic content characteristic of its dimension.

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 neighborhood	Ambiguous
Market Signals	Income/Investment	Rental income, investment potential	Ambiguous
	Motivated Seller	Urgency, price reduction, distress	Negative

3.4.3. Design Principles

The reference descriptions were designed following three principles. First, *semantic saturation*: each reference includes multiple synonyms, related phrases, and characteristic vocabulary to ensure broad coverage of its target dimension. Second, *dimensional specificity*: each reference targets a single qualitative dimension to minimize cross-loading. Third, *linguistic consistency*: all references are written in French using vocabulary typical of Quebec real estate listings, ensuring alignment with the embedding space of the listing corpus.

3.5. Cosine Similarity Computation

For each property i and reference j, we compute the cosine similarity:

$$s_{ij} = \text{sim}(\mathbf{e}_i, \mathbf{r}_j) = \mathbf{e}_i \cdot \mathbf{r}_j \tag{4}$$

where $\mathbf{r}_j$ is the L_2-normalized embedding of the j-th reference. The result is a similarity matrix $\mathbf{S} \in \mathbb{R}^{n \times 20}$.

Each $s_{ij} \in [-1, 1]$ measures the degree to which property i's description semantically resembles reference j. A higher score indicates that the listing text uses language closer in meaning to the reference description, capturing not just keyword overlap but deeper semantic alignment.

The full similarity matrix is computed as a single matrix multiplication $\mathbf{S} = \mathbf{E} \cdot \mathbf{R}'$, where $\mathbf{E} \in \mathbb{R}^{n \times 384}$ is the matrix of listing embeddings and $\mathbf{R} \in \mathbb{R}^{20 \times 384}$ is the matrix of reference embeddings. This operation is computationally trivial.

3.6. Hedonic Model Specification

We estimate a series of nested OLS models with White-Huber heteroskedasticity-robust standard errors (HC3 variant):

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

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

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

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

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

where $\mathbf{X}_i$ includes six structural variables (bedrooms, bathrooms, half-bathrooms, parking spaces, stories, and lot size), ℓ_i is the description length (character count), $\mathbf{s}_i$ is the full vector of 20 cosine similarities, and $\mathbf{s}_i^{\text{sig}}$ retains only the similarities that are individually significant at the 5% level in Model D. All independent variables are standardized (zero mean, unit variance) prior to estimation, so coefficients represent the effect of a one-standard-deviation change.

The choice of HC3 standard errors, recommended by MacKinnon and White (1985) for samples of moderate size, provides consistent inference under heteroskedasticity without requiring specification of the error variance structure.

The incremental contribution of semantic features is tested using an F-statistic:

$$F = \frac{(\text{SSR}_A - \text{SSR}_D)/q}{\text{SSR}_D/(n - k - 1)} \tag{10}$$

where q is the number of additional semantic variables and k is the total number of regressors in Model D.

3.7. Identification and Interpretation

The coefficients on the semantic similarity variables should be interpreted as *conditional associations*—implicit semantic price gradients—rather than causal effects. This stance 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). Three distinct channels could generate the observed correlations between listing language and prices, and the cross-sectional design cannot distinguish among them.

Information channel. Agents accurately describe observable property attributes that affect prices but are not captured by the structural variables in our model. Under this interpretation, the semantic coefficients recover the implicit prices of genuine quality dimensions: luxury finishes, renovation status, natural amenities, and so forth. The listing text serves as a proxy for unobserved quality, and the coefficients have a straightforward hedonic interpretation.

Persuasion channel. Listing language influences buyer perceptions and willingness to pay, independent of underlying property quality. Skillfully written descriptions could inflate perceived value, creating a price premium attributable to marketing rather than to the property itself. Under this interpretation, the semantic coefficients capture the return to agent effort and linguistic skill rather than housing quality.

Omitted-location channel. Listing language proxies for locational characteristics that are correlated with both the property's description and its price. For example, the negative coefficient on Family-Friendly language likely reflects the systematic association between family-oriented marketing and lower-priced suburban markets, rather than a negative valuation of schools and parks. Similarly, New Construction language may capture peripheral location rather than an intrinsic quality discount.

In practice, the observed coefficients likely reflect a mixture of all three channels. The information channel is most plausible for dimensions with clear quality content (Luxury, Needs Renovation, Waterfront), while the omitted-location channel is most relevant for dimensions with geographic content (Family-Friendly, New Construction, Quiet & Peaceful). The absence of explicit spatial controls in our baseline specification (see Section 7) means that some semantic coefficients absorb locational price variation.

Scope of inference. We interpret our estimates as measuring whether and how much semantic content in listing text is associated with price variation, conditional on observed structural characteristics. This is informative for automated valuation, market segmentation, and understanding the informational content of agent narratives, even without causal identification. We avoid language implying that semantic similarity "causes" or "determines" prices throughout the paper.

4. Data

4.1. Data Source and Sample Construction

The dataset is drawn from a database of residential property listings in Quebec, Canada, collected from the Realtor.ca and Centris platforms. Realtor.ca is the public-facing portal of the Canadian Real Estate Association (CREA), and Centris is the Quebec-specific MLS (Multiple Listing Service) platform operated by the Québec Professional Association of Real Estate Brokers (APCIQ). Together, these platforms cover the large majority of brokered residential listings in the province.

The full database contains 46,479 listings across five property categories: houses (17,089), condominiums (8,586), land (8,506), rentals (8,240), and multiplexes (4,058). For this study, we restrict the sample to single-family houses to maintain a homogeneous product category, following standard practice in the hedonic pricing literature (Sirmans et al., 2005). After excluding listings with missing or uninformative descriptions (fewer than 20 characters) and non-positive prices, the final sample comprises 17,087 observations.

4.2. Structural Variables

Table 3 presents descriptive statistics for the structural and textual variables used in the hedonic models.

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 (sq. ft.)	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 lot-size field is noisy (units are not standardized at the source), which motivates its cautious interpretation in the regressions.

The median listing price is $589,900, with substantial right-skew (mean $807,259, maximum $25 million). The log transformation reduces this skewness considerably (log-price standard deviation of 0.70). The typical house has 3 bedrooms and 2 bathrooms (sample medians). The parking-spaces field averages 5.5 with high variance; it records the total number of parking spots reported by the

agent, including driveway and outdoor spaces, which explains its large mean and range. The lot-size field is recorded without standardized units at the source and is correspondingly noisy—a limitation reflected in its imprecise coefficient estimate in the regressions. Listing descriptions average 510 characters (standard deviation of 151) and are truncated at approximately 700 characters in the data export, so the length variable is bounded from above.

Figure 2 displays the distributions of price and log-price, confirming the appropriateness of the semi-logarithmic specification.

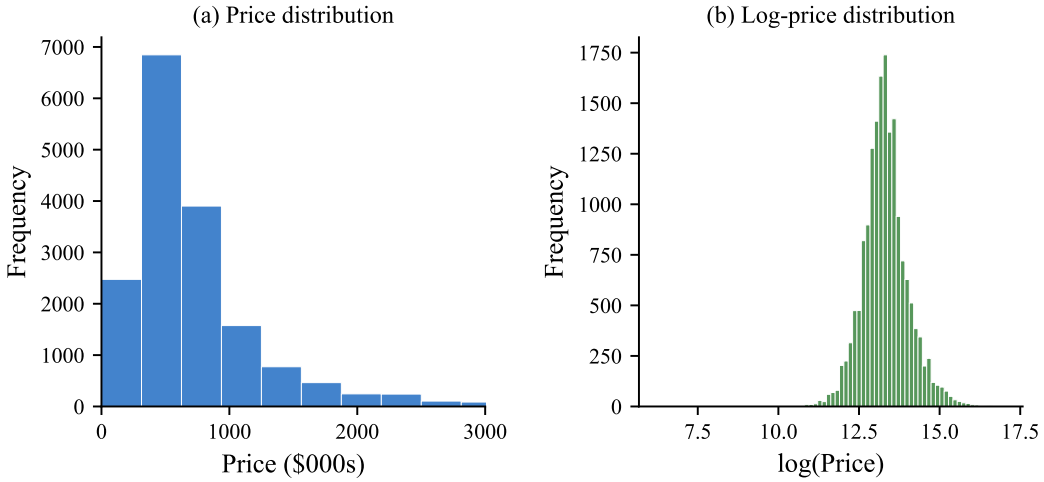


Figure 2. Distribution of listing prices and log-prices (n = 17,087).
Notes: Left panel shows the distribution of raw listing prices; right panel shows the distribution of log-transformed prices. The log transformation substantially reduces right-skewness.

4.3. Semantic Similarity Features

Table 4 reports the distribution of cosine similarity scores across the 20 reference dimensions, along with their bivariate correlations with log-price.

A notable pattern in Table 4 is that all bivariate correlations between similarity scores and log-price are negative, ranging from −0.29 to −0.17. This initially counterintuitive result arises because longer, more detailed descriptions—which characterize lower-priced listings where agents invest more effort in textual marketing—produce higher similarity scores with *all* references. This confound is addressed in the multivariate regression by controlling for description length, after which the theoretically expected positive and negative effects of different qualitative dimensions emerge clearly.

Figure 3 displays the distributions of cosine similarities, revealing substantial variation within each dimension.

Table 4. 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.

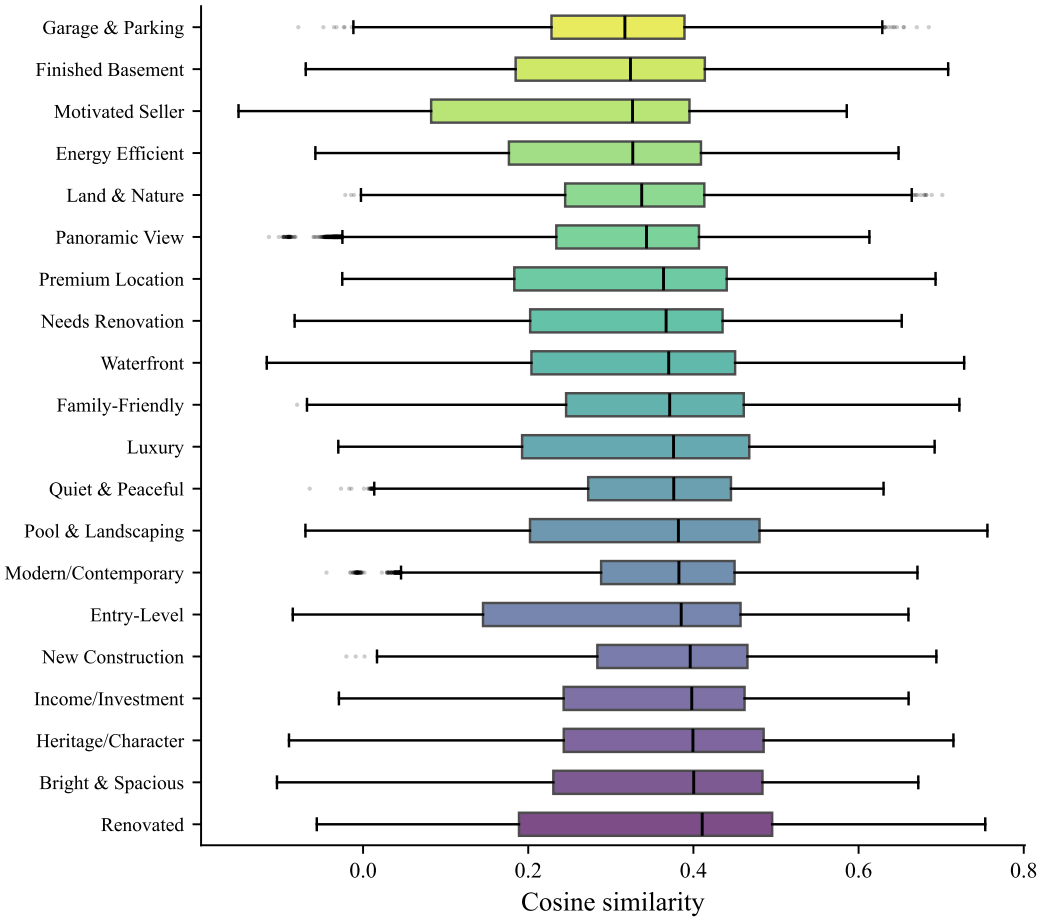


Figure 3. Distribution of cosine similarity scores across 20 reference dimensions.

Notes: Each box plot displays the distribution of cosine similarity between listing embeddings and the corresponding reference description embedding, ordered by median similarity. Boxes span the interquartile range; whiskers extend to 1.5 times the interquartile range.

5. Results

5.1. Model Comparison

Table 5 reports goodness-of-fit statistics for the five model specifications defined in Equations (5)–(9). Figure 4 visualizes the comparison.

Table 5. 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.

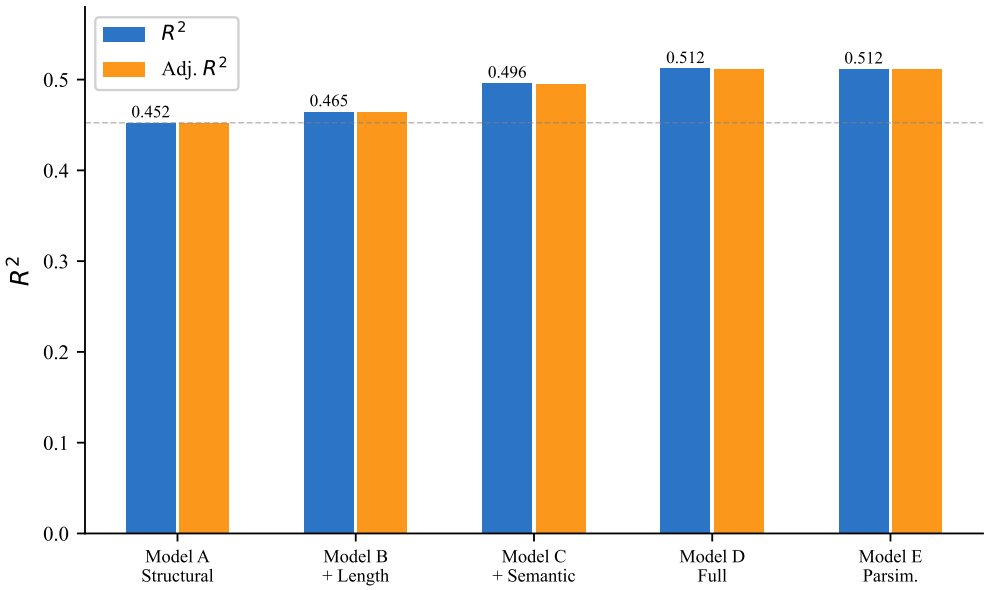


Figure 4. Goodness-of-fit comparison across hedonic model specifications.
Notes: This figure compares R^2 and adjusted R^2 across the five nested model specifications defined in Equations (5)–(9). The dashed line marks the structural baseline (Model A).

Several results merit discussion. The structural model (A) explains 45.2% of log-price variation, consistent with the typical range reported in the hedonic literature for single-family homes (Sirmans et al., 2005). Adding description length alone (Model B) improves the R^2 by 1.2 percentage points, confirming that the quantity of listing text is price-informative—an indirect quality signal.

Adding the 20 semantic similarity measures without controlling for length (Model C) yields a substantially larger improvement of 4.4 percentage points. This demonstrates that the *content* of the text matters beyond its mere length. The full model (D), which includes both text length and all semantic features, achieves an R^2 of 0.512—an improvement of 6.0 percentage points over the structural baseline.

The parsimonious model (E), retaining only the 16 individually significant similarity measures, achieves nearly identical fit (R^2 = 0.512, BIC = 24,425 vs. 24,448 for Model D). The lower BIC of Model E suggests it should be preferred on the basis of model parsimony. Four reference dimensions—Renovated, Income/Investment, Heritage/Character, and Pool & Landscaping—do not contribute independently after controlling for the other variables. The complete coefficient estimates for Model E are reported in Appendix Table A2.

Figure 5 decomposes the explained variance in the full model. Structural variables account for 88.3% of the total R^2, text length contributes 2.4%, and the semantic similarity measures add 9.3%.

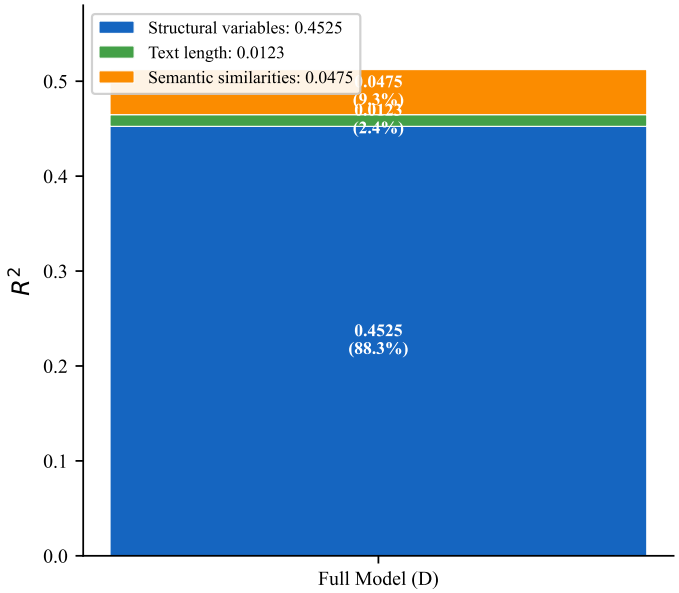


Figure 5. Decomposition of explained variance (R^2) in the full model (D).
Notes: Structural variables account for 88.3% of the total explained variance, description length contributes 2.4%, and the 20 semantic similarity measures add 9.3%.

5.2. Joint Significance Test

The F-statistic for the joint significance of all 21 text-based variables (length + 20 similarities) comparing Model D to Model A yields:

$$F = 99.53, \quad p < 0.001$$

This decisively rejects the null hypothesis that listing text contains no price-relevant information beyond structural characteristics. The test statistic is very large, reflecting the substantial improvement in model fit.

5.3. Structural Variable Estimates

Figure 6 presents the coefficient estimates for structural variables in the full model.

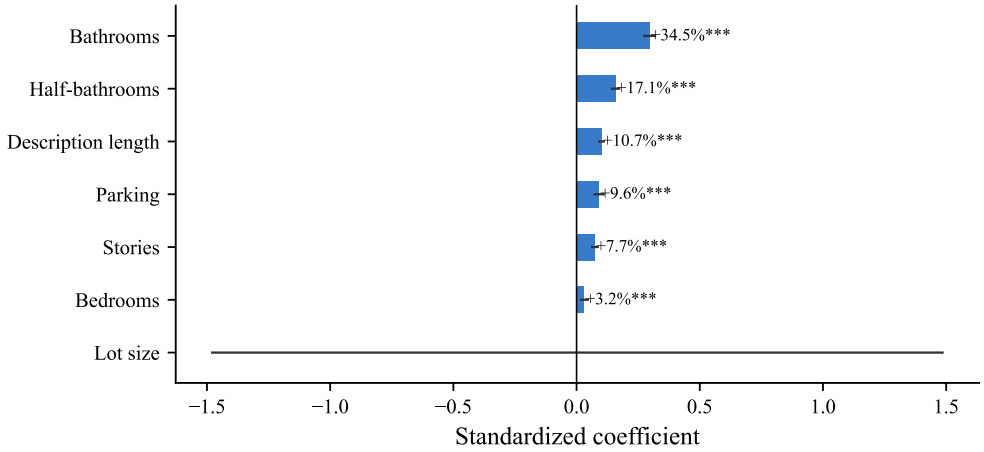


Figure 6. Structural variable coefficients in the full model (D), with 95% CI.
Notes: Coefficients represent the percentage change in price associated with a one-standard-deviation change in each structural variable. Error bars show 95% confidence intervals based on HC3 robust standard errors.

Bathrooms have the largest structural effect (+34.5% per standard deviation), followed by half-bathrooms (+17.1%), description length (+10.7%), parking (+9.6%), and stories (+7.7%). Bedrooms have a modest positive effect (+3.2%), and lot size is statistically insignificant—consistent with the noisy, unit-inconsistent measurement of this field in the raw data (Section 4).

5.4. Semantic Dimension Estimates

Table 6 reports the complete coefficient estimates from Model D. Figure 7 presents the semantic coefficients graphically.

5.4.1. Positive Price Associations

Three semantic dimensions exhibit large, positive, and highly significant associations with listing prices.

Modern/Contemporary (+16.4%, $p < 0.001$). Listings semantically closer to modern/contemporary reference descriptions—emphasizing contemporary architecture, minimalist design, smart home technology, and energy-efficient construction—are associated with the highest price premium. This

Table 6. Full model (D) coefficient estimates. HC3 robust standard errors. All variables standardized.

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.

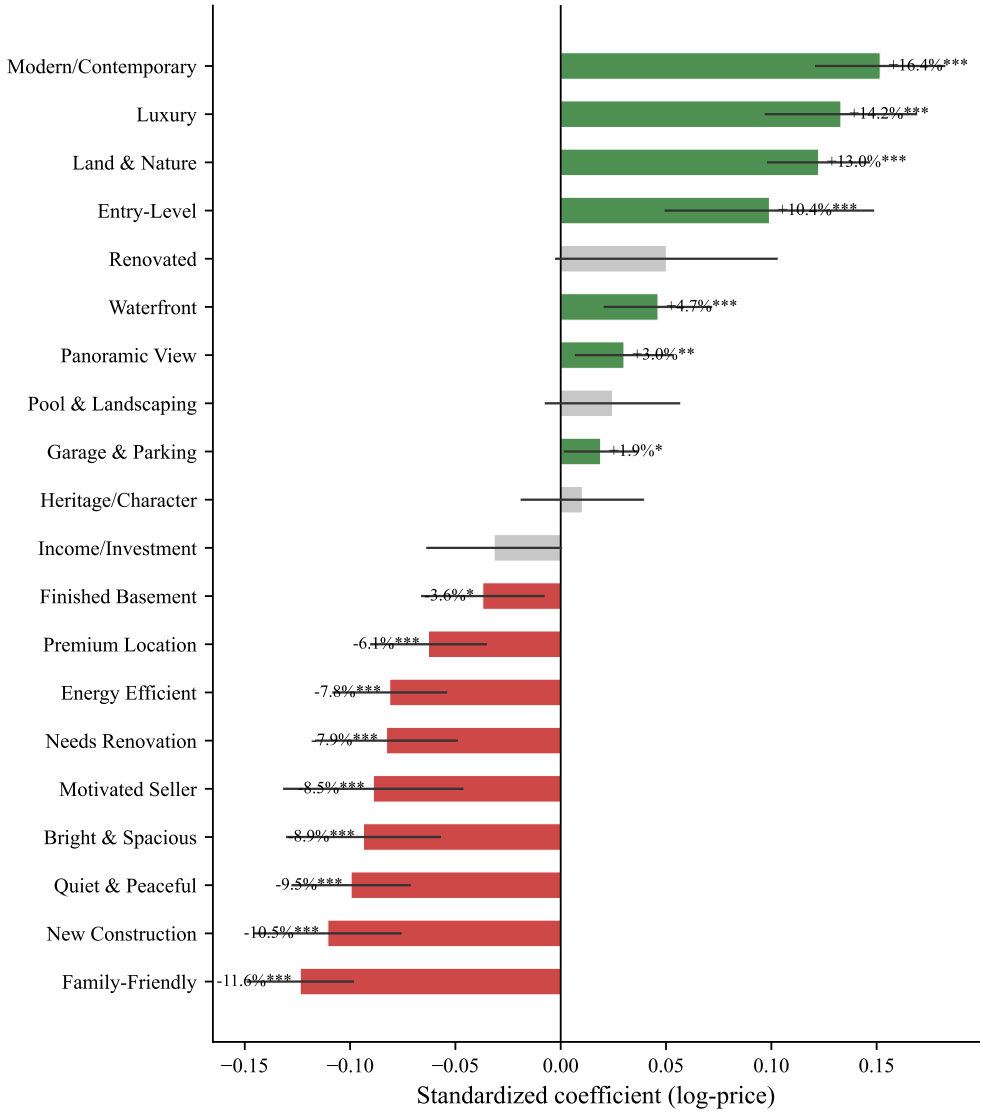


Figure 7. Hedonic price impact of semantic similarity dimensions (Model D). Green bars indicate positive effects; red bars indicate negative effects. Gray bars are not statistically significant at the 5% level. Error bars show 95% confidence intervals.

is consistent with revealed preferences for modern aesthetics and functional design in the Quebec housing market.

Luxury (+14.2%, $p < 0.001$). Language describing premium finishes (quartz, granite, hardwood), gourmet kitchens, spa-like bathrooms, and home automation systems is associated with a substantial price premium. This dimension captures quality variation that is invisible to standard structural variables: two properties with the same number of bathrooms can differ enormously in finish quality, and the listing text registers this difference.

Land & Nature (+13.0%, $p < 0.001$). Descriptions of wooded lots, professional landscaping, privacy, and access to natural surroundings are associated with a significant positive price gradient, consistent with the well-documented value of outdoor amenity space in residential markets.

Waterfront (+4.7%, $p < 0.001$) and **Panoramic View** (+3.0%, $p = 0.01$) capture location-specific amenities—access to water and scenic views—that are not reflected in standard structural variables but are associated with higher prices.

5.4.2. Negative Price Associations

Several dimensions exhibit negative coefficients that, upon closer examination, reflect market segmentation rather than value destruction per se (see Section 3.7 for the identification discussion).

Family-Friendly (−11.6%, $p < 0.001$). This is the most strongly negative dimension. Descriptions emphasizing proximity to schools, parks, and family amenities characterize lower-priced suburban markets. After controlling for structural characteristics, this language likely proxies for suburban location rather than indicating that family-oriented features reduce value. This coefficient should be interpreted with the caveat that it may absorb locational price variation in the absence of spatial controls.

New Construction (−10.5%, $p < 0.001$). The negative coefficient is counterintuitive at first glance, since "new" is typically a premium attribute. However, in Quebec's real estate geography, new residential developments are concentrated in peripheral suburban areas (e.g., Mirabel, Mascouche, Lachute) where land costs are lower. The "new construction" semantic dimension likely captures this locational sorting rather than an intrinsic quality discount; municipality fixed effects, once available, would disambiguate the two (Section 7).

Motivated Seller (−8.5%, $p < 0.001$). This is consistent with the information asymmetry literature (Levitt and Syverson, 2008): language signaling seller urgency ("must sell," "price reduced," "estate sale") reveals a weak bargaining position and is associated with a price discount relative to comparable properties with non-urgent descriptions.

Needs Renovation (−7.9%, $p < 0.001$). Descriptions indicating that a property requires work are associated with lower prices, consistent with the cost of deferred maintenance being capitalized into the listing price.

5.5. Semantic Profiles Across Price Quintiles

Figure 8 visualizes the mean cosine similarity for each reference dimension across price quintiles, revealing how the semantic "fingerprint" of listing descriptions varies with price level.

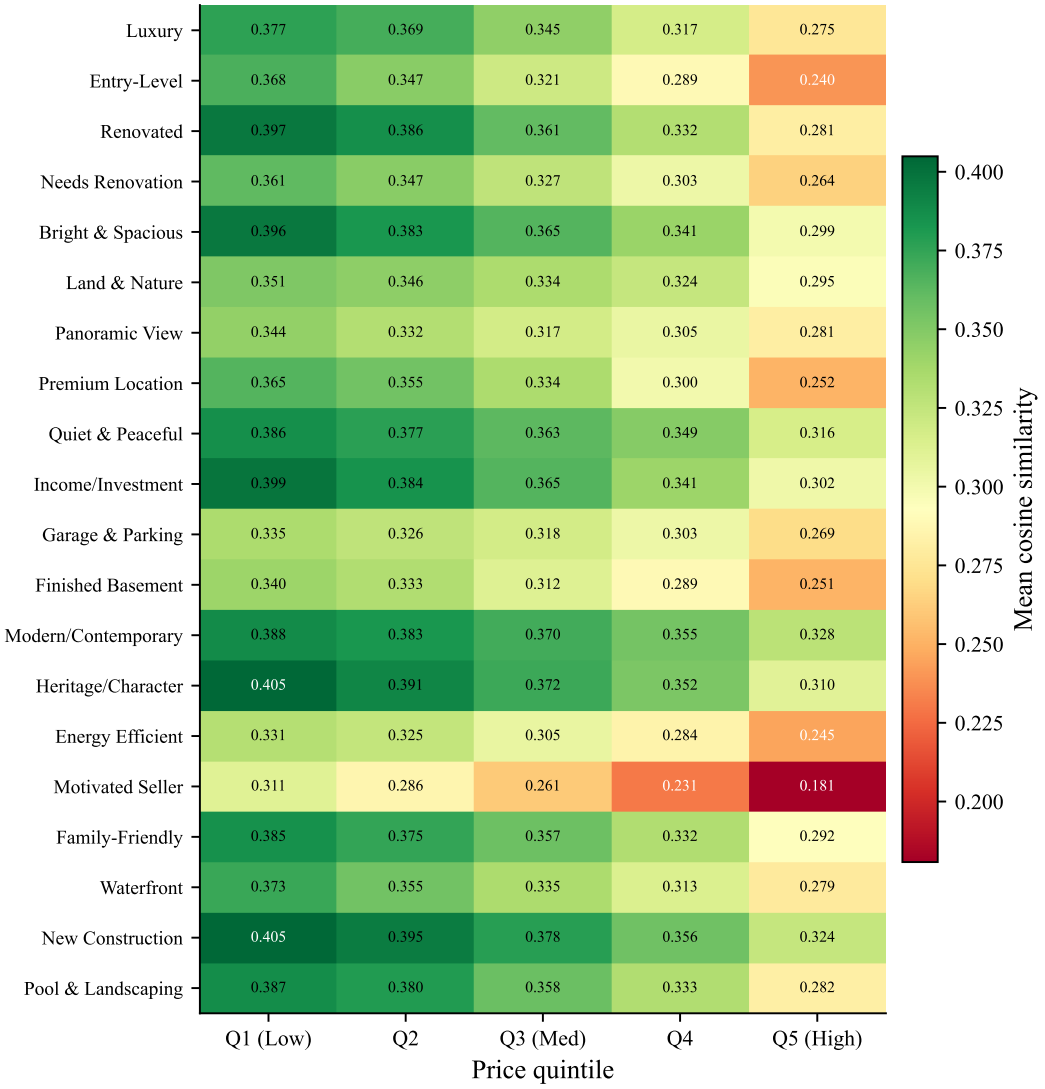


Figure 8. Mean cosine similarity by price quintile across 20 reference dimensions.
Notes: Warmer colors indicate higher similarity. Each cell reports the mean cosine similarity between listings in the given price quintile and the corresponding reference description.

The gradient reveals that lower-priced properties score higher on nearly all dimensions due to longer, more detailed descriptions. The relative differences across dimensions are nevertheless informative: moving from the bottom to the top quintile, the steepest declines occur for Motivated Seller, Entry-Level, and Premium Location, while Land & Nature, Modern/Contemporary, and Panoramic View decline the least—precisely the dimensions that carry positive premiums in the regression results.

5.6. Inter-Reference Correlations

Figure 9 presents the correlation matrix between the 20 similarity dimensions.

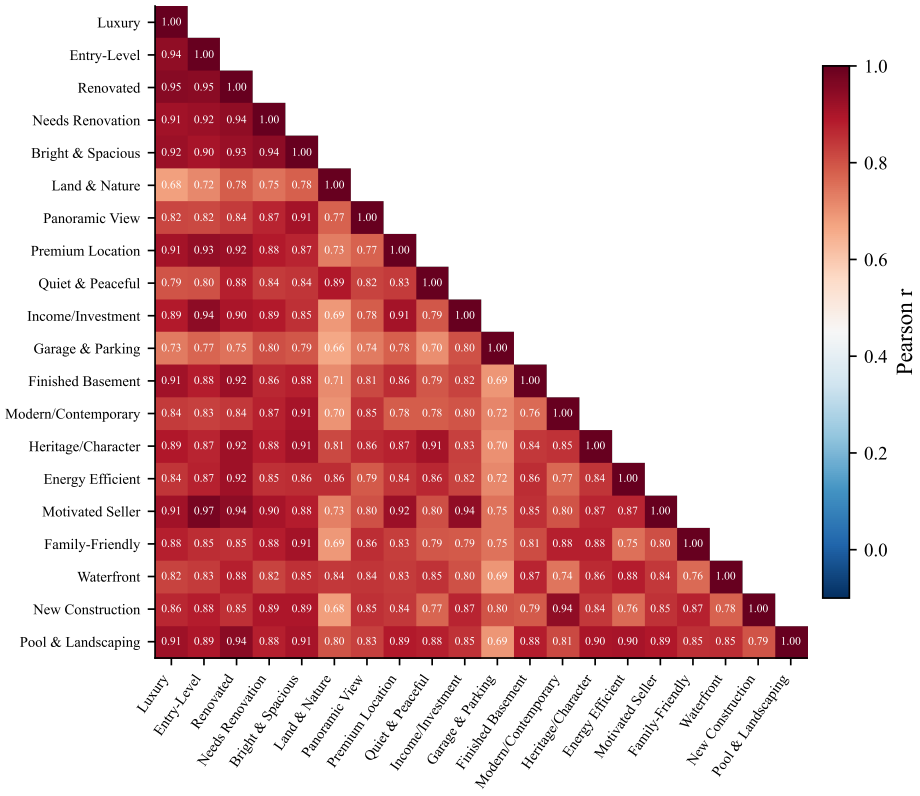


Figure 9. Pearson correlation matrix between semantic similarity dimensions (lower triangle).
Notes: Correlations are computed across all n = 17,087 observations. The strong positive correlations largely reflect the common influence of description length and verbosity on similarity scores.

Inter-reference correlations are strong and uniformly positive, ranging from 0.66 to 0.97 with a mean of 0.84. This common variation is driven primarily by description length and verbosity: detailed listings score higher on multiple dimensions simultaneously. The strongest correlations occur between Entry-Level and Motivated Seller (r = 0.97)—both characteristic of value-oriented listings—and between Luxury and Renovated (r = 0.95), which share upgrade-related vocabulary. This pronounced collinearity motivates two design features of the analysis: the inclusion of description length as a control, which absorbs much of the shared variation, and the multicollinearity diagnostics reported in Section 6, which quantify its consequences for inference. It also implies that individual semantic coefficients are best read as partial associations within a correlated block of quality signals rather than as isolated effects.

5.7. Illustrative Relationships

Figure 10 presents scatter plots for four selected dimensions, illustrating the heterogeneity in relationships between semantic similarity and price.

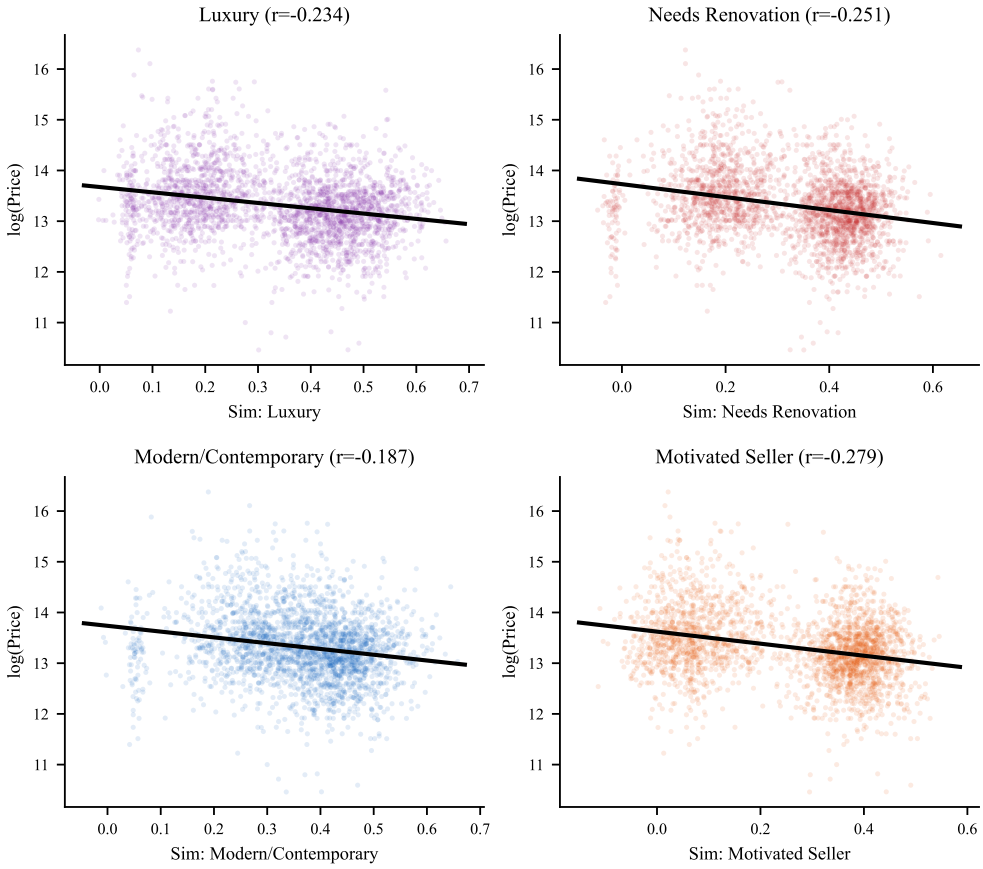


Figure 10. Bivariate relationships between selected semantic similarities and log-price.
Notes: Lines show ordinary least squares linear fit. These bivariate relationships do not control for structural variables or description length.

5.8. Residual Diagnostics

Figure 11 presents residual diagnostics for the full model, including a residual-versus-fitted plot and a normal Q-Q plot.

The residual plot (panel a) reveals mild heteroskedasticity, with residual variance increasing at the extremes of the fitted value distribution. This motivates our use of HC3 robust standard errors. The Q-Q plot (panel b) shows approximate normality in the central distribution with heavier-than-normal tails, particularly in the upper tail—consistent with the well-known right-skew of housing prices even after log transformation. The OLS coefficient estimates remain consistent under these conditions, and the HC3 standard errors provide valid inference.

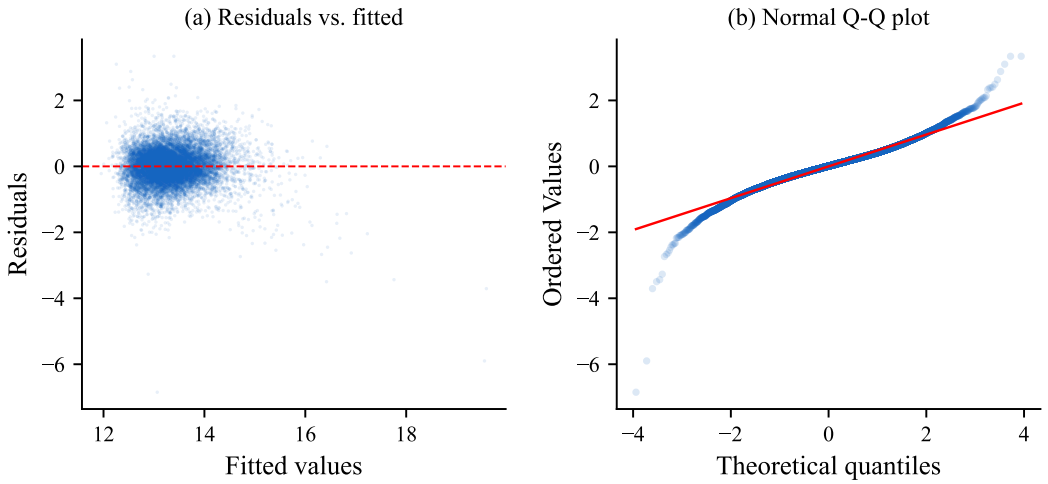


Figure 11. Residual diagnostics for Model D: (a) residuals vs. fitted values; (b) normal Q-Q plot.
Notes: Panel (a) plots OLS residuals against fitted values, revealing mild heteroskedasticity at the tails. Panel (b) compares the empirical residual distribution to the standard normal, showing heavier-than-normal tails.

6. Robustness Analysis

To assess the sensitivity of our findings to modeling assumptions, sample composition, and estimation techniques, we conduct a battery of robustness checks. These analyses address five potential concerns: heteroskedasticity, multicollinearity among semantic features, sensitivity to outliers, distributional heterogeneity across price quantiles, and variable selection stability. All statistics reported in this section are produced by the replication pipeline accompanying the paper.

6.1. Heteroskedasticity Diagnostics

The Breusch-Pagan test (Breusch and Pagan, 1979) applied to Model D strongly rejects the null hypothesis of homoskedasticity (LM = 768.0, $p < 0.001$). This result formally justifies our use of HC3 heteroskedasticity-consistent standard errors throughout. The pattern of heteroskedasticity—residual variance increasing at both tails of the fitted value distribution (Figure 11, panel a)—is typical of housing price models and reflects the greater price dispersion in both the lowest-quality and highest-quality segments of the market (Goodman and Thibodeau, 1995).

6.2. Variance Inflation Factors

The strong correlations among the 20 cosine similarity measures (Figure 9) translate into substantial multicollinearity by conventional standards: variance inflation factors (VIFs) computed for Model D average 17.2 across the similarity variables, and 16 of the 20 exceed the conventional threshold of 10 (O'Brien, 2007), with maxima for Renovated (41.6) and Entry-Level (40.7). This is an expected

consequence of the design: all 20 features are projections of the same embedding onto related quality concepts, and they share the common influence of description verbosity.

Three considerations indicate that this collinearity inflates standard errors without invalidating the analysis. First, as O'Brien (2007) emphasizes, a high VIF is not by itself grounds for respecification: it signals imprecision, not bias, and the relevant question is whether the affected coefficients remain informative. Despite VIF-inflated standard errors, 16 of the 20 similarity measures are individually significant at the 5% level, indicating that each retains sufficient unique variance for estimation. Second, the block-level evidence—the joint F-statistic of 99.53 and the 6-percentage-point R^2 gain—is unaffected by collinearity among the block's members. Third, the coefficient estimates are stable across the winsorized and bootstrap re-estimations reported below, which would not be the case if the estimates were fragile artifacts of a near-singular design. The practical implication, developed in Section 7, is that individual semantic coefficients should be interpreted as partial associations within a correlated block of quality signals, with block-level conclusions being the most robust.

6.3. Quantile Regression Analysis

To examine whether the implicit prices of semantic dimensions vary across the price distribution, we estimate Model D using quantile regression (Koenker and Bassett, 1978; Koenker, 2005) at the 25th, 50th, and 75th percentiles. Table 7 reports the coefficients for the six most economically significant semantic dimensions.

Table 7. 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. First, the Luxury premium *increases* monotonically with property value: a one-standard-deviation increase in luxury similarity is associated with a 9.7% premium at the 25th percentile but a 17.3% premium at the 75th percentile. This is consistent with quality complementarities—luxury features are valued more in properties that already occupy the upper market segment. Second, the negative effects of Motivated Seller and Needs Renovation language *attenuate* at higher quantiles (the Motivated Seller discount shrinks from 8.7% at $\tau = 0.25$ to a statistically insignificant 3.8% at $\tau = 0.75$), suggesting that urgency and condition discounts are proportionally smaller for high-value properties where buyer pools are less price-sensitive. Third,

the Modern/Contemporary and Land & Nature premiums are remarkably stable across quantiles, indicating that these dimensions carry value throughout the market rather than in a particular segment.

These heterogeneous effects across the conditional price distribution confirm that the OLS estimates represent averages across meaningfully different market segments, and they strengthen the economic interpretation of the semantic dimensions.

6.4. Sensitivity to Outliers

We re-estimate Model D after trimming prices below the 1st and above the 99th percentile, removing 341 observations from the tails of the price distribution. The trimmed model yields an adjusted R^2 of 0.486, somewhat lower than the full-sample estimate of 0.511—as expected, since trimming removes exactly the price variation that the model exploits. Importantly, no coefficient changes sign relative to the full model, and all 16 similarity measures that are significant in the full model remain significant in the trimmed specification. The largest coefficient shifts are for Entry-Level (−0.024) and Motivated Seller (+0.023), both within the confidence intervals of the full-model estimates. This stability confirms that our results are not driven by extreme observations in either tail of the price distribution.

6.5. Bootstrap Inference

To verify that our HC3 standard errors provide reliable inference, we estimate Model D using 1,000 nonparametric bootstrap replications (Efron and Tibshirani, 1993). For each replication, we resample $n = 17{,}087$ observations with replacement and re-estimate the model, obtaining the empirical distribution of each coefficient. Bootstrap standard errors are within 9% of the HC3 estimates for every semantic dimension—and within 5% for 15 of the 20—and the bootstrap 95% percentile confidence intervals closely match the HC3-based intervals. This concordance confirms the reliability of our asymptotic inference under the observed heteroskedasticity pattern.

6.6. Regularized Variable Selection

As an alternative to significance-based variable selection (Model E), we employ Lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) regularization with five-fold cross-validation over the standardized Model D design. At the cross-validated penalty, both procedures retain a sparse set of four similarity variables (Needs Renovation, Premium Location, Income/Investment, and Family-Friendly) and shrink the remainder to zero.

This aggressive selection is the textbook behavior of ℓ_1 penalties under strong collinearity: when regressors are correlated at 0.66–0.97, the Lasso selects one representative per correlated block and

discards near-duplicates, so the retained variables act as proxies for their blocks rather than as a list of the "true" non-zero effects (Zou and Hastie, 2005). The exercise therefore complements, rather than replicates, the significance-based selection in Model E: it confirms that the semantic block contains genuine predictive signal that survives penalization, while reinforcing the message of the VIF analysis that variable-by-variable attributions within the block are sensitive to the selection criterion. Conclusions in this paper that rely on individual dimensions (e.g., the Luxury gradient) are accordingly cross-checked against the quantile and bootstrap evidence above.

6.7. Comparison with Alternative Text Representations

To contextualize the predictive and inferential value of the reference-based approach, Table 8 compares it with alternative representations of the same text, estimated on identical samples. For each method, we report the ΔR^2 gain over the structural-only Model A together with an assessment of interpretability.

Table 8. 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**

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 reference projections but produces features with no economic interpretation; the reference-based approach trades roughly two percentage points of R^2 for coefficient-level interpretability (see Section 6).

Two results stand out. First, *what* the text says matters more than *how much*: description length alone adds 1.2 percentage points of R^2, whereas content-based features add 4.4–7.8 points. Second, there is a measurable price of interpretability. Twenty principal components of the raw embeddings—which capture 49.5% of the embedding variance but have no economic meaning—outperform the 20 reference similarities in pure fit (ΔR^2 of +0.070 vs. +0.044 without the length control). The reference-based approach deliberately trades roughly two percentage points of R^2 for features that support coefficient-level economic inference; for the inferential goals of hedonic analysis, we consider this trade worthwhile, while applications that only require prediction may prefer the unrestricted embedding space.

6.8. Summary of Robustness Findings

Table 9 summarizes the robustness checks.

Table 9. 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 penalization
PCA comparison	Yes	PCA fits better (ΔR^2 +0.070 vs. +0.044) but is uninterpretable; reference approach preferred for inference

The overall picture is that the paper's central findings—the joint explanatory power of semantic features and the sign and approximate magnitude of the main implicit price gradients—are stable across estimation methods, sample definitions, and distributional assumptions. The diagnostics also delineate the limits of the evidence: the semantic dimensions are strongly collinear, so individual coefficients are measured imprecisely relative to the block as a whole, and regularized selection does not single out the same variables as significance testing. Extensions that would further strengthen the analysis—spatial fixed effects, alternative embedding models, and systematic reference-description sensitivity analysis—are discussed in Sections 7 and 8.

7. Discussion

7.1. Interpretation of Key Findings

Our results establish that free-text property descriptions contain economically significant information that traditional hedonic variables fail to capture. The 6-percentage-point improvement in R^2 is substantial given that the baseline model already explains 45% of price variation with standard structural variables—a level consistent with the hedonic pricing literature (Sirmans et al., 2005). The semantic features contribute 9.3% of the total explained variance, suggesting that qualitative

housing attributes communicated through listing text represent a meaningful dimension of housing differentiation. To contextualize this magnitude, Sirmans et al. (2006) found that adding a full set of neighborhood controls to a structural-only model typically improves R^2 by 3–8 percentage points—our semantic features achieve comparable gains from a single text field.

The positive effects of **Modern/Contemporary** (+16.4%) and **Luxury** (+14.2%) language align with theoretical expectations and prior empirical work on quality premiums (Nowak and Smith, 2017; Goodwin et al., 2014). These dimensions capture aspects of housing quality—design aesthetics, material quality, technological amenities—that are simply not reflected in counts of bedrooms and bathrooms. A house with three bathrooms can range from a modest home with basic fixtures to a luxury property with spa-like en-suites; the listing description captures this variation. The quantile regression analysis (Section 6) reveals that the luxury premium is amplified in the upper portion of the price distribution, consistent with the notion that luxury is a positional good whose marginal value increases with baseline quality (Frank, 2007) and with the general finding of the quantile-hedonic literature that many quality attributes command larger implicit prices in upper market segments (Zietz et al., 2007).

The **Land & Nature** premium (+13.0%) reflects the well-documented value of natural amenities and outdoor space in residential markets (Irwin, 2002; Tyrväinen et al., 2005). This finding is particularly relevant in the Quebec context: using Quebec City transactions and site-inspection measures, Des Rosiers et al. (2002) estimate landscaping premiums ranging from roughly 4% (hedges, landscaped curbs) to about 12% (a landscaped patio) of house value. Our text-based measure recovers an amenity gradient of comparable direction and magnitude through the listing narrative alone, without site-level amenity data—evidence that the two measurement strategies tap the same underlying valuation.

The negative coefficient on **Family-Friendly** (−11.6%) deserves careful interpretation. This does not imply that schools and parks reduce property values; rather, it reflects the systematic association between family-oriented marketing language and lower-priced suburban markets. After controlling for structural characteristics, family-friendly language serves as a residual proxy for suburban location where prices are lower. This is precisely the equilibrium-sorting logic emphasized by Kuminoff et al. (2013): households sort across locations on price and amenities jointly, so language that marks a market segment inherits that segment's price level. The finding also echoes Delmelle and Nilsson (2021), who show that property-advertisement text is a strong predictor of neighborhood socioeconomic composition: listing language is as much about *where* a property is as about *what* it is. This illustrates the importance of interpreting hedonic coefficients as conditional marginal effects rather than causal estimates—a point emphasized by Pakes (2003) in the broader context of hedonic identification.

The strong negative effect of **Motivated Seller** (−8.5%) is consistent with the information

asymmetry literature. Levitt and Syverson (2008) showed that real estate agents selling their own homes achieve prices about 3.7% higher than when selling clients' homes, partly because they can conceal information about the urgency of the sale; our estimate implies that *explicit* urgency disclosure is associated with a discount more than twice that size, which is coherent with the idea that overt urgency language is a stronger signal of a weak outside option than the subtle cues agents suppress. This result connects to the broader theory of strategic disclosure (Grossman, 1981; Milgrom, 1981): in unraveling equilibria, transparency about motivation invites adverse inference, as buyers rationally read disclosed urgency as evidence of weak bargaining position. The listing-language literature reports the same qualitative pattern: remarks signaling seller motivation are associated with lower prices and faster sales (Haag et al., 2000; Goodwin et al., 2014).

The surprising negative coefficient on **New Construction** (-10.5%) warrants extended discussion. In most housing markets, "new" is a premium attribute. However, in Quebec's real estate geography, new residential construction is concentrated in peripheral suburban developments (e.g., Mirabel, Mascouche, Lachute) where land costs are lower. After controlling for structural characteristics, the "new construction" semantic dimension captures this locational sorting rather than a quality discount. This interpretation is supported by the quantile regression results, which show a diminished negative effect at higher quantiles—precisely the pattern expected if the coefficient proxies for peripheral location rather than an intrinsic quality penalty.

7.2. Relation to Prior Findings

Situating the estimates against the closest prior work sharpens both agreements and departures.

Agreement on the information content of text. Nowak and Smith (2017) find that words in agent comments proxy for unobserved quality and that omitting them biases other hedonic coefficients; we corroborate this at the level of model fit (a 6-point R^2 gain) and extend it by giving the text features named economic content. Shen and Ross (2021) conclude that property descriptions contain substantial information about otherwise unobserved quality; our results agree, and our decomposition (Figure 5) quantifies the same conclusion in variance terms. The early remark-category studies (Haag et al., 2000) and the vernacular studies (Goodwin et al., 2014, 2018) found sign patterns—premia for quality language, discounts for distress language—that match our forest plot dimension by dimension.

Agreement on distributional heterogeneity. The amplification of the Luxury premium at upper quantiles and the attenuation of condition/urgency discounts mirror the findings of Zietz et al. (2007), who document that most attribute prices rise with the conditional price level, and of McMillen (2008), who shows that coefficient heterogeneity—not just characteristics—drives changes in the price distribution.

A departure from the prediction-first literature. Studies that optimize prediction (Shen and Ross,

2021; Baur et al., 2023) report larger fit gains from unrestricted text representations than our 20 projections deliver; our own PCA benchmark reproduces this ordering (Table 8). The departure is deliberate: we accept a quantified fit penalty in exchange for coefficient-level interpretability, a trade that the prediction literature does not price.

A parallel with vision-based studies. Glaeser et al. (2018) and Poursaeed et al. (2018) extract quality from listing photographs and find, as we do with text, that unstructured listing content reveals price-relevant quality invisible to structured fields. That two different modalities produced by the same marketing process yield the same conclusion strengthens the case that structured hedonic variables systematically undermeasure quality—and suggests a natural extension combining both.

7.3. The Information Content of Agent Narratives

Our findings contribute to a broader literature on information production by market intermediaries. Real estate agents serve a dual role: they match buyers with properties and produce information about property characteristics through listing descriptions. The rhetoric of listings is deliberately crafted (Pryce and Oates, 2008), yet the economically significant coefficients on our semantic dimensions suggest that agents' textual descriptions contain genuine informational content—they are not merely "cheap talk" or undifferentiated marketing prose.

This interpretation is supported by two pieces of evidence. First, the signs of most semantic coefficients align with theoretical priors: luxury language commands premiums, renovation-need language is discounted, and urgency language penalizes prices. If listing descriptions were pure noise or uniform boilerplate, the semantic features would exhibit no systematic relationship with prices. Second, the stability of the coefficients across quantile regression, trimmed samples, and bootstrap inference (Section 6) indicates that the text-price associations are robust features of the data rather than artifacts of particular observations or distributional assumptions.

At the same time, the correlational nature of our estimates precludes definitive conclusions about the mechanism. The positive coefficient on Luxury language, for instance, could reflect: (a) agents accurately describing genuinely luxurious properties that command higher prices; (b) the language itself influencing buyer perceptions and willingness to pay; or (c) an omitted variable (e.g., neighborhood prestige) correlated with both luxury language and price. Disentangling these mechanisms would require quasi-experimental variation in listing text—an approach feasible with A/B testing data from real estate platforms but beyond the scope of the present study.

7.4. Description Length as a Quality Signal

The significant positive effect of description length (+10.7%) merits discussion. This finding may seem paradoxical given that bivariate correlations between similarity scores and price are

uniformly negative (reflecting the tendency for lower-priced listings to have longer descriptions). The resolution lies in the multivariate structure: once we control for *what* a description says (via the semantic similarities), *how much* it says becomes a positive signal. Longer descriptions, conditional on content, may indicate agent effort, property complexity, or a genuine abundance of features to describe. This is consistent with Shen and Ross (2021), who find that the informational content of property descriptions proxies for unobserved quality, and with the listing-effort interpretation in Goodwin et al. (2014). One caveat is that descriptions in our data are truncated at approximately 700 characters (Section 4), so the length variable measures verbosity only up to this bound; the estimated coefficient should be read as the effect of length variation within the observed range.

7.5. Comparison with PCA-Based Approaches

To contextualize our results, we compare the reference-based approach with the more standard PCA-based method on the same sample (Table 8). Applying PCA to the raw 384-dimensional embedding vectors and retaining 20 principal components—which together capture 49.5% of the embedding variance—yields a larger fit improvement than the 20 reference similarities: adding the components to the structural baseline raises R^2 by 7.0 percentage points, versus 4.4 points for the reference projections. This ordering is unsurprising: the principal components are the linear combinations of the embedding space that maximize retained variance, so any fixed set of 20 projections, including ours, is weakly dominated in pure fit.

What PCA cannot provide is economic content. A coefficient on "PC3" has no natural interpretation as a housing characteristic, cannot be communicated to appraisers or market participants, and is unstable across samples in its meaning even when stable in its fit. The reference-based approach deliberately trades roughly two percentage points of R^2 for features that are individually nameable, signable ex ante, and directly usable in the hedonic framework, where inference on implicit prices—not prediction—is the goal. This is the position argued on general grounds by Rudin (2019): when a model informs high-stakes decisions, inherent interpretability is preferable to post hoc explanation of a black box (Ribeiro et al., 2016; Lundberg and Lee, 2017). Researchers whose objective is purely predictive should use the unrestricted embedding space (or the full 384 dimensions with regularization); researchers who need interpretable implicit prices face the trade-off we quantify here.

7.6. Practical Implications

Our findings have several practical implications. For **automated valuation models (AVMs)**, incorporating semantic text features could meaningfully improve accuracy. The 6-percentage-point R^2 improvement suggests that AVMs relying solely on structured data—still the dominant design in

mass appraisal (Kok et al., 2017)—leave significant predictive power on the table. Moreover, our reference-based approach produces features that are computationally cheap to generate (requiring only a dot product between pre-computed embeddings) and stable across time.

For **real estate appraisers and agents**, the results highlight which qualitative dimensions of listing language are most strongly associated with price variation. Agents crafting listing descriptions can leverage these findings to understand how different linguistic strategies relate to market positioning. However, we caution that the relationship is correlational: changing the words in a listing description is unlikely to change the sale price if the underlying property attributes remain unchanged.

For **housing market researchers**, the reference-based framework offers a flexible tool for investigating qualitative dimensions of housing that have been difficult to operationalize. Researchers can define custom reference descriptions tailored to their specific research questions, making the approach portable to different markets, languages, and property types.

7.7. Limitations

Several limitations should be acknowledged, and we discuss each in turn along with potential avenues for resolution.

Listing prices vs. transaction prices. Our data consists of listing prices, not realized sale prices. List and sale prices differ systematically, and the gap varies with market conditions and marketing time (Haurin et al., 2010); the magnitude of this gap may correlate with our semantic measures (e.g., "motivated seller" listings may sell at a larger discount to listing price). If the listing-to-sale price ratio varies systematically with semantic content, our coefficient estimates may be biased. The direction of this bias is ambiguous: luxury language may be associated with a smaller listing premium (if agents of luxury properties price more accurately) or a larger one (if they price aspirationally). The availability of transaction-level data would allow estimation of the listing premium as a function of semantic content and would strengthen the analysis.

Cross-sectional identification. The cross-sectional nature of our data prevents causal inference. As discussed in Section 3.7, the relationship between listing language and price may reflect accurate quality description, strategic persuasion, or omitted variable correlation. While our robustness checks (Section 6) confirm the stability of the estimates across specifications, they do not address the fundamental identification concern. Future work using within-property variation—comparing successive listings for the same property with different agents or different textual strategies—could provide more credible causal identification, and recent work on textual strategies in housing (Alfano and Guarino, 2022) points toward research designs that exploit variation in listing composition; platform-level A/B experiments would be the gold standard.

Absence of spatial controls. Our hedonic model does not include explicit locational variables

(municipality fixed effects, distance to CBD, or spatial autoregressive terms), despite the well-established spatial dependence of housing prices (Anselin, 1988; Dubin, 1988; Can, 1992). Some semantic dimensions—particularly Family-Friendly, New Construction, and Quiet & Peaceful—likely capture locational variation as much as property-level attributes. Including municipality or census-tract fixed effects would absorb location-specific price levels, allowing the semantic coefficients to be interpreted more cleanly as within-location quality signals. However, this comes at the cost of eliminating between-location variation that the semantic features may legitimately capture. A spatial Durbin model (LeSage and Pace, 2009) would provide a formal framework for decomposing direct and indirect (spillover) effects of semantic content, and the microdata methods of Dubé and Legros (2014) are directly applicable to our setting.

Language model limitations. The all-MiniLM-L6-v2 model was primarily trained on English text. While it handles French adequately—the language of most Quebec listings—a model specifically fine-tuned on French real estate text could potentially produce more nuanced embeddings. The model may fail to capture domain-specific nuances: for instance, "plancher chauffant" (heated floor) and "radiant heating" carry identical meaning but may not embed identically across languages. Cross-lingual embedding models optimized for French, such as CamemBERT (Martin et al., 2020) or FlauBERT (Le et al., 2020), represent promising alternatives for future work. Conneau et al. (2020) document material performance losses when multilingual models are applied to specialized monolingual tasks, suggesting that our estimates may understate the true informational content of listing descriptions.

Reference description subjectivity. Our 20 reference descriptions represent one researcher's operationalization of qualitative housing dimensions. Alternative formulations could yield different similarity scores and potentially different hedonic estimates. While we designed the references following principled criteria (semantic saturation, dimensional specificity, linguistic consistency), the approach lacks a formal optimality criterion. A systematic sensitivity protocol—averaging similarities over multiple paraphrases per dimension, varying reference length, and comparing French-only against bilingual formulations—would quantify the dependence of the estimates on any single phrasing. One could also envision deriving reference descriptions empirically, for example by clustering listing embeddings and using cluster centroids as data-driven references, though this sacrifices the ex ante interpretability that motivated our approach. The full text of all 20 references is reproduced in Appendix Table A1 to make this dependence transparent and replicable.

Multicollinearity among similarity features. The strong correlations among similarity dimensions (0.66–0.97; Figure 9) introduce substantial multicollinearity: VIFs average 17.2 and exceed the conventional threshold of 10 for 16 of the 20 dimensions (Section 6). The consequences are visible in the regularized-selection exercise, where the Lasso retains only four block representatives rather than the sixteen individually significant variables. Individual coefficient magnitudes should

therefore be interpreted with caution: the "Luxury" premium of 14.2% includes variation that is shared with the Renovated and Bright & Spacious dimensions, and the data cannot fully apportion credit within such correlated blocks. The block-level conclusions—that semantic content jointly adds 6 percentage points of explanatory power with theoretically sensible signs—are unaffected. Ridge regression or principal component regression on the similarity features would provide bounds on the total effect of correlated quality dimensions, and reducing the number of references (or orthogonalizing them ex ante) is a natural design lever for future applications.

Temporal stability. Our analysis captures a single cross-section of listings. The relationship between textual content and prices may vary over time as market conditions change, buyer preferences evolve, and listing language conventions shift. During a seller's market, "motivated seller" language may carry a larger discount because it is rarer and more diagnostic; during a buyer's market, such language may be more common and less informative. Longitudinal analysis would be needed to assess the stability of semantic implicit prices across market cycles.

8. Conclusion

This paper demonstrates that property listing descriptions contain economically meaningful information that can be systematically extracted and integrated into hedonic pricing models. Using sentence embeddings and cosine similarity with 20 interpretable reference descriptions, we show that semantic features from listing text improve the explanatory power of a standard hedonic model by 5.9 percentage points (adjusted R^2: 0.452 $\rightarrow$ 0.511) for 17,087 single-family homes in Quebec, Canada. The estimates are stable under bootstrap inference, outlier trimming, and quantile-based heterogeneity analysis, while the accompanying diagnostics candidly document the strong collinearity among semantic dimensions and its implications for variable-by-variable interpretation.

Our approach bridges the hedonic pricing tradition and the text-as-data program in empirical economics (Gentzkow et al., 2019; Ash and Hansen, 2023) by producing features that satisfy the requirements of both: it generalizes the transparent dictionary method into embedding space while preserving the ex ante interpretability that inference on implicit prices requires (Rudin, 2019). They are semantically rich, capturing deep meaning rather than surface-level keyword overlap, and economically interpretable, with each feature corresponding to a named qualitative dimension carrying a clear coefficient estimate. Descriptions emphasizing luxury finishes (+14.2%), modern design (+16.4%), and natural settings (+13.0%) command significant premiums, while language signaling renovation needs (−7.9%), seller urgency (−8.5%), and family-oriented marketing (−11.6%) is associated with discounts. Importantly, the quantile regression analysis reveals that these effects are heterogeneous across the price distribution: the luxury premium is amplified for

high-value properties, while renovation and urgency discounts are attenuated, consistent with quality complementarities and differential buyer price sensitivity.

The general methodology—embedding domain-specific text, computing cosine similarities against researcher-designed references, and incorporating the resulting features into standard econometric models—is portable to any setting where free-text descriptions accompany structured economic data. Applications beyond real estate include:

- **E-commerce**: Product listing text on platforms such as Amazon or eBay contains quality and condition signals that could be incorporated into hedonic price models for used goods, electronics, or collectibles.
- **Labor economics**: Job posting descriptions embed information about workplace culture, benefits, and flexibility that are imperfectly captured by structured fields; reference-based similarity could quantify the implicit wage premiums associated with different workplace attributes (Marinescu and Wolthoff, 2020).
- **Innovation economics**: Patent abstracts describe inventions in varying degrees of technical novelty and commercial potential; semantic features could augment patent valuation models.
- **Corporate finance**: Earnings call transcripts and annual report narratives contain forward-looking information about firm prospects; reference-based sentiment dimensions could improve asset pricing models beyond the binary positive/negative sentiment used by Loughran and McDonald (2011).
- **Hospitality**: Hotel and Airbnb listing descriptions convey amenity and ambiance information that complements structured data on room type, location, and ratings.

In each case, the reference-based approach can translate qualitative textual information into quantitative features suitable for econometric analysis while preserving the interpretability required for inference on implicit prices.

Future research should extend this framework along several dimensions. First, applying the method to transaction prices rather than listing prices would provide cleaner identification of text-price relationships and enable estimation of the listing premium as a function of semantic content. Second, temporal analysis using panel data could reveal how the relationship between listing language and prices evolves over market cycles—for instance, whether luxury premiums are amplified in bull markets and attenuated in recessions. Third, spatial econometric extensions incorporating municipality fixed effects and spatial autoregressive structures (LeSage and Pace, 2009) could disentangle location-based from property-level semantic effects, addressing the concern that some dimensions (Family-Friendly, New Construction) proxy for suburban location. Fourth, multilingual embedding models specifically optimized for French—such as CamemBERT (Martin et al., 2020) or domain-adapted variants—together with TF-IDF, keyword, and sentiment baselines, would complete the comparison of text representations begun in Section 6. Fifth, a systematic

sensitivity analysis of the reference descriptions—multi-paraphrase averaging, bilingual variants, and length perturbations—would quantify the robustness of the semantic features to their exact phrasing. Sixth, the framework could be extended to buyer-side text (search queries, saved-listing patterns), connecting to the evidence that housing search is itself segmented along the qualitative dimensions buyers care about (Piazzesi et al., 2020), and to listing photographs, whose quality content parallels that of text (Glaeser et al., 2018; Poursaeed et al., 2018). Finally, causal identification could be strengthened through within-property variation—comparing successive listings for the same property by different agents—or through A/B testing partnerships with real estate platforms.

References

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.

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.

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.

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.

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.

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.

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.

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 operationalization 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 standardized.

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.