On the Performance of LLMs for Real Estate Appraisal

Margot Geerts[1] (✉), Manon Reusens[1], Bart Baesens[1,3], Seppe vanden Broucke[2,1], and Jochen De Weerdt[1]

[1] LIRIS, KU Leuven, Leuven, Belgium margot.geerts@kuleuven.be
[2] Department of Business Informatics and Operations Management, Ghent University, Ghent, Belgium
[3] Department of Decision Analytics and Risk, University of Southampton, Southampton, UK

Abstract. The real estate market is vital to global economies but suffers from significant information asymmetry. This study examines how Large Language Models (LLMs) can democratize access to real estate insights by generating competitive and interpretable house price estimates through optimized In-Context Learning (ICL) strategies. We systematically evaluate leading LLMs on diverse international housing datasets, comparing zero-shot, few-shot, market report-enhanced, and hybrid prompting techniques. Our results show that LLMs effectively leverage hedonic variables, such as property size and amenities, to produce meaningful estimates. While traditional machine learning models remain strong for pure predictive accuracy, LLMs offer a more accessible, interactive and interpretable alternative. Although self-explanations require cautious interpretation, we find that LLMs explain their predictions in agreement with state-of-the-art models, confirming their trustworthiness. Carefully selected in-context examples based on feature similarity and geographic proximity, significantly enhance LLM performance, yet LLMs struggle with overconfidence in price intervals and limited spatial reasoning. We offer practical guidance for structured prediction tasks through prompt optimization. Our findings highlight LLMs' potential to improve transparency in real estate appraisal and provide actionable insights for stakeholders.

Keywords: Large Language Models · Real Estate Appraisal · In-Context Learning.

1 Introduction

Global real estate, valued at $379.7 trillion in 2022, represents the world's largest wealth store, with residential properties constituting the majority[4]. The real estate market plays a crucial role in economies worldwide, impacting homeowners,

[4] https://www.savills.com/impacts/market-trends/the-total-value-of-global-real-estate-property-remains-the-worlds-biggest-store-of-wealth.html

investors, and governments. Accurate price estimations are vital for all stakeholders, from home buyers facing affordability challenges in Europe[5] and the U.S.[6] to China's slowing market[7]. Access to reliable price data helps ensure informed decision-making and supports a stable and sustainable market across regions. Nevertheless, real estate valuation remains opaque and unevenly accessible, contributing to information asymmetry between buyers and sellers [16]. Sellers inherently have superior knowledge of the local market and the property's condition as opposed to buyers. While potential buyers can call upon a real estate broker or other experts, this asymmetry is difficult to eliminate. Some argue that a data-driven house price prediction approach can help real estate stakeholders, including buyers, by informing their decisions [1,18]. However, this approach requires advanced Machine Learning (ML) expertise, extensive manual data processing and access to a substantial dataset, which may not be readily available to the average home buyer. Large Language Models (LLMs) present a promising solution to address this information asymmetry [34]. Trained on vast and diverse datasets encompassing a significant portion of Internet knowledge [5], these models have the potential to uncover meaningful insights and patterns, including those relevant to real estate [9]. Recently, LLMs have been proven to excel in structured prediction tasks with In-Context Learning (ICL), enabling them to approximate regression problems without explicit training [32]. This makes them a promising tool for ad hoc house price prediction, reducing the barriers to data-driven real estate insights. By leveraging their extensive training, LLMs could bridge knowledge gaps, offering nuanced perspectives and data-driven guidance in this complex domain. This marks a key step towards democratizing access to real estate appraisal insights and enhancing transparency for a diverse range of stakeholders. Reducing information asymmetry improves price accuracy, benefiting both buyers, who avoid overpaying, and sellers, who experience faster sales due to improved liquidity and as such receiving fair market value [16]. Additionally, investors, financial institutions, and policymakers can make more informed decisions, leading to improved investment strategies, risk assessments, and more effective tax and policy frameworks.

In this paper, we assess LLMs' potential for improving accessibility to real estate appraisal, or valuation, by answering four Research Questions (RQs):

RQ1 How effectively can prompt engineering techniques optimize LLM performance for house price prediction and what is the most effective prompt?

RQ2 Can LLMs generate sufficiently accurate house price estimates to serve as viable alternatives to traditional ML models?

RQ3 How reliably do LLMs estimate price intervals for real estate appraisal, and how does this compare to traditional ML approaches?

[5] https://ec.europa.eu/eurostat/web/interactive-publications/housing-2023

[6] https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/

[7] https://www.imf.org/en/News/Articles/2024/02/02/cf-chinas-real-estate-sector-managing-the-medium-term-slowdown

RQ4 What features do LLMs prioritize in their house price prediction processes, and how do these align with traditional valuation methodologies?

To address these questions, we investigate different prompting approaches with ICL and evaluate a wide range of pre-trained LLMs on various housing datasets worldwide. In the context of the house price prediction task, we scrutinize the capabilities of LLMs in three dimensions: the accuracy of price predictions, the delineation of price intervals, and their explanatory capacity. Our contributions can be summarized as follows:

1. We demonstrate that optimizing prompt design significantly improves LLM performance in house price prediction. Carefully selecting in-context examples based on feature similarity and geographic proximity enhances accuracy and adaptability across different housing markets.
2. We show that LLMs can generate sufficiently accurate house price estimates, approaching the performance of traditional ML models. While they do not surpass ML models in predictive accuracy, their accessibility, interpretability, and flexibility make them valuable for real estate stakeholders.
3. We identify overconfidence in price intervals as a key limitation of LLM-based valuation. LLMs consistently underestimate price uncertainty, producing narrower prediction ranges that fail to capture real market values.
4. We find that LLMs prioritize hedonic property features effectively but struggle with spatial and temporal reasoning. Despite leveraging variables like property size and amenities, they undervalue the role of location and time.
5. We help reduce information asymmetry in the real estate market by providing concrete guidelines for harnessing LLMs to support informed decision-making among buyers, sellers, financial institutions and policymakers.

2 Related Work

House price prediction is typically framed as a supervised learning problem involving tabular data. Automated Valuation Models (AVMs) are trained on datasets $D = \{(X_i, y_i)\}_{i=1}^{n}$ where $X_i \in \mathbb{R}^m$ are the m-dimensional features of property i, and $y_i \in \mathbb{R}$ is its price. The objective is to minimize prediction error $L(\hat{y}, y)$, using loss functions like mean squared error (MSE). Features are broadly categorized into hedonic attributes, i.e. structural attributes (e.g., size, number of rooms), and locational factors (e.g., coordinates, proximity to amenities). Recent research has shifted from hedonic regression models [2] to modern ML and deep learning approaches [7,17,18], with increasing focus on interpretability, including Shapley values [29] and uncertainty quantification techniques such as conformal prediction for prediction intervals [1,12].

LLMs for data science have revolutionized predictive modeling with unstructured textual data. In house price prediction, Natural Language Processing (NLP) extracts insights from property descriptions, market reports, and reviews,

converting them into structured features [30,40]. Building on recent advancements in NLP, pre-trained LLMs are increasingly used in data science applications with tabular data [35], such as geospatial interpolation [24], Point of Interest recommendation [6], and time series analysis [14]. Despite their growing adoption, research on LLMs for real estate appraisal remains limited, with a recent study focusing only on rental price prediction [3]. Our work addresses this gap by examining more robust prompting strategies, evaluating diverse datasets, and integrating interpretability, thereby offering new insights into real estate appraisal with LLMs.

In-Context Learning (ICL) is an inference-time technique where the model, without updating its parameters, generalizes from provided examples. Specifically, LLMs are provided with K examples as context and are then tasked with completing a new example by leveraging the patterns and information from the preceding ones [22]. [32] show how LLMs can perform regression tasks when provided in-context examples. LLMs' ability to handle tabular data is demonstrated through frameworks like TabLLM for data-efficient classification [11], while the Meta-ICL framework further enhances ICL efficiency [4].

3 Methodology

3.1 Large Language Models for House Price Prediction

To effectively prompt LLMs with tabular housing data, we follow the guidelines set by [11] for manual data serialization in zero- and few-shot learning settings. Our goal is not to replace traditional ML models but to evaluate LLMs' capabilities in estimating house prices and identify the optimal prompt.

Figure 1 illustrates the prompting strategy used in the experiments. To determine the optimal prompt for house price prediction, we evaluate twelve different strategies combining various building blocks. Specifically, we incorporate market reports and in-context examples on top of the zero-shot baseline containing the task definition and property description. Our prompting optimization strategy uses ICL to enforce two pillars that are relevant in real estate: regional real estate market dynamics, which are often accounted for in AVMs by explicitly modeling temporal effects [17], and comparable property valuation examples [18], essentially delineating housing submarkets [2]. The market reports provide context on regional house price indices from the preceding month or quarter. The labeled examples from the training data are selected based on either haversine distance (geographic proximity) or cosine distance (similar hedonic features) and limited to three or ten examples. Finally, we test a combination of ten examples evenly split between geographic and hedonic neighbors. This combination of geographic and characteristic-based similarity has been shown to be effective in hedonic price modeling in prior work [27]. Based on these results, we extend the interaction with the LLM by maintaining the conversation history. Using the best-ranked prompt configuration across datasets and LLMs, the LLM generates a price interval with a 90% target coverage. While methods like conformal

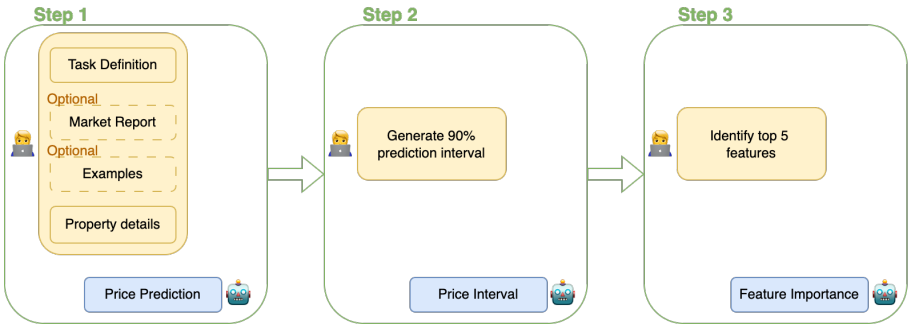

Fig. 1: Overview of the LLM prompting methodology for house price prediction. Step 1: The model receives a structured prompt containing the task definition, an optional market report, optional ICL examples, and details of the target property-forming the basis for prompt optimization. It then predicts the property price. Step 2: The model generates a 90% prediction interval. Step 3: The model identifies the five most important features. This approach enables price estimation, uncertainty quantification, and interpretability in real estate appraisal.

prediction require access to the internal LLM structure, which is unavailable in standard API interactions, direct prompting aligns with how real estate practitioners would use LLMs. Finally, we assess explainability by asking the model to identify the top five features influencing its prediction. The full prompt template is shown in Appendix A.

Since LLMs struggle with raw geographic coordinates due to tokenization of prompts, we follow [24] in reverse geocoding the coordinates into full addresses using the Nomatim API[8] to OpenStreetMap[9], incorporating both representations into the prompt.

Given the importance of accessibility in this study, we include a range of the most recent pre-trained LLMs comprising both open-source and closed-source models of varying sizes and architectural design. In our experiments, we use Llama 3.2:3B [25], Llama 3.1:70B [5], and GPT-4o-mini [26]. These models were selected to balance scale (number of parameters), provider diversity, and accessibility constraints. All models were prompted with a seed of 0 and temperature of 0 to ensure reproducibility. All code and data used for the experiments is available via `https://github.com/margotgeerts/LLM4RealEstate`. More information on the checkpoints and computing environment used can be found in Appendix B.

3.2 ML baselines

We consider two common house price prediction baselines: k-Nearest Neighbor (kNN) regression and Gradient Boosted Trees (GBT). These baselines are cho-

[8] `https://nominatim.org/`
[9] `https://www.openstreetmap.org/`

sen based on their conceptual relevance to our LLM-approach and their strong empirical performance in real estate appraisal. First, kNN serves as a natural baseline because it predicts a property's price based on an interpolation of its nearest neighbors. This aligns well with our LLM prompting strategy, which provides the model with similar properties as context. Comparing LLMs to kNN allows us to determine whether LLMs can extract deeper insights beyond simple interpolation. To ensure a fair comparison, we match the LLM prompt settings with $k = \{3, 10\}$ neighbors, using haversine distance (geographic proximity), cosine distance (hedonic similarity), or a combination of both for ten examples. Second, we include GBTs—specifically LightGBM (LGBM) [15]—as they remain the state-of-the-art (SOTA) choice for structured tabular data and have consistently outperformed deep learning methods in house price prediction tasks [8]. Unlike kNN, LGBM learns complex, nonlinear relationships in data, allowing us to benchmark whether LLMs can approach fully optimized ML models that have access to structured training data. We use LGBM with default parameters to ensure a fair, out-of-the-box comparison that mirrors how practitioners might deploy an ML model without extensive hyperparameter tuning. Other ML methods, such as linear regression or support vector machines (SVMs), were excluded as they generally underperform compared to GBTs on tabular data. Similarly, deep learning models such as Graph Neural Networks, while promising, have not yet demonstrated consistent superiority over GBTs in house price prediction [8].

To compare prediction intervals between LLMs and LGBM, we use Conformal Prediction (CP). CP is a framework that provides valid prediction intervals without assuming any specific model, offering a distribution-free method for uncertainty quantification in ML tasks [33]. This approach works by using the observed data to "conform" the model's predictions, ensuring that the true value lies within the predicted interval with a specified confidence level. Since house prices exhibit temporal trends that violate the assumption of data exchangeability (i.e., the data distribution is not independent and identically distributed over time), we apply a CP procedure designed to handle such distribution shifts, known as EnbPI [37], via the MAPIE library [31]. Finally, we use the SHapley Additive exPlanations (SHAP) values [20] to compare the LLM and LGBM explanations. Considering that the LLMs are provided with both coordinates and address, we adjust for this by aggregating the SHAP values corresponding to the two coordinate features (X-Y). Subsequently, we rank all features based on the mean absolute SHAP value computed across the test instances.

3.3 Evaluation metrics

To evaluate predictive performance, we report the Mean Absolute Percentage Error (MAPE), consistent with prior work [18], due to space limitations, and the standard deviation of the Percentage Error (PE) across test observations. Prediction intervals are assessed based on actual coverage (percentage of instances where the true price is within the predicted interval) and the Mean Prediction Interval Width (MPIW), which measures precision. Valid intervals should achieve coverage close to 90%, with narrower MPIWs indicating higher

precision. Feature importance is compared by evaluating the top five features prioritized by LLMs and SHAP-based rankings from LGBM. With this comparison, we do not attempt to evaluate the ground truth correctness of these explanations. Our objective is to examine the degree of alignment between two distinct paradigms: LLMs and GBTs. Validating the intrinsic accuracy or faithfulness of the LLM explanations would require expert assessments or adversarial testing, which guide important directions for future work.

3.4 Datasets

To generalize LLM performance for real estate appraisal, we selected four real-world housing datasets located in various geographic areas. The datasets from King County, USA (from Kaggle[10]), Flanders, Belgium (proprietary), and Beijing, China (from Kaggle[11]) contain property transactions, while the dataset from Barcelona, Spain [28] contains property listings. Despite the subtle distinction between listings and transactions, we treat them similarly. A 60:20:20 train-validation-test split is used, and results are reported based on a random subset of 1000 test examples per dataset. Table 1 summarizes key statistics.

Table 1: Summary statistics of the housing datasets used for evaluation.

	King County	Flanders	Barcelona	Beijing
Train size	12914	174135	25714	117349
Validation size	4349	59188	12339	37045
Test size	3569	55125	23295	41301
No. variables	14	16	35	17
Min. price	$75 000	€34 280	€37 000	¥1 270 021
Max. price	$7 700 000	€970 512	€4 866 000	¥11 000 094
Min. date	2014-05-02	2015-01-04	2018-03-01	2015-01-01
Max. date	2015-05-27	2023-05-24	2018-12-01	2017-12-31

4 Results & Discussion

This section examines the effectiveness of large language models (LLMs) in house price prediction, focusing on both prompt optimization and comparisons with traditional ML approaches. First, we analyze the impact of different prompt engineering strategies on LLM performance, identifying the most effective techniques for improving prediction accuracy (**RQ1**). Next, we compare LLM-based predictions with traditional ML baselines, evaluating their absolute accuracy, interval estimates, and feature prioritization (**RQ2**–**RQ4**). Finally, we synthesize

[10] https://www.kaggle.com/datasets/astronautelvis/kc-house-data
[11] https://www.kaggle.com/datasets/ruiqurm/lianjia

these findings into practical guidelines, offering insights on when and how LLMs can be effectively deployed for real estate appraisal.

4.1 Optimizing LLMs with Prompt Engineering

Figure 2 summarizes the MAPE scores across all prompting strategies, models, and datasets. Generally, prompting strategies with more labeled examples lead to more accurate predictions, with ten mixed examples (`10 ex. mixed`) emerging as the best-performing strategy in most datasets and models. This suggests that combining geographically near and hedonic similar properties provides a balanced context that enhances LLM predictions. This approach aligns with prior work in hedonic price modeling, where integrating geographic and characteristic-based similarity has been shown to effectively capture local housing market dynamics [27]. Ten-shot prompting strategies consistently outperform three-shot prompts, independent of the selection method, indicating that LLMs benefit from comprehensive contextual information. Only for the Beijing dataset, the performance is sensitive to the specific method for example selection.

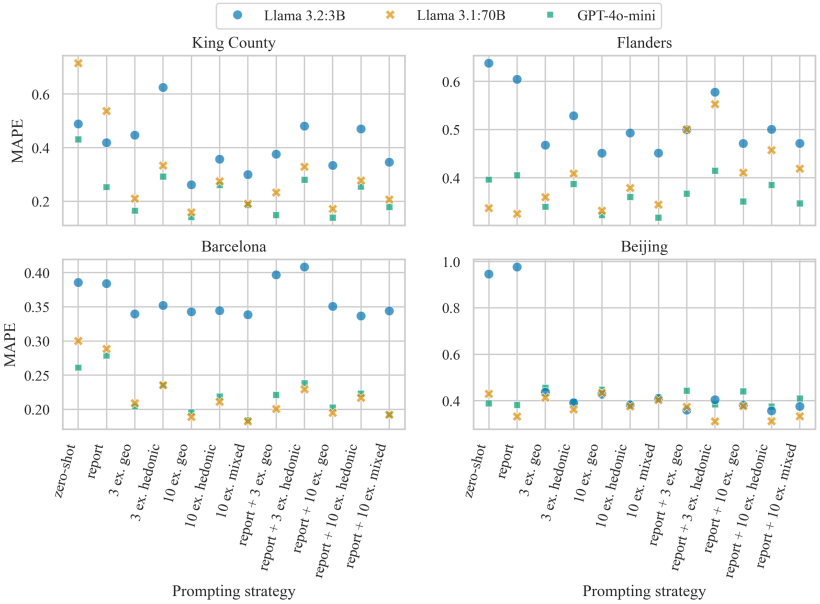

Fig. 2: Including 10 mixed examples (geographic and hedonic similarity) provides the best results overall and zero-shot prompting the worst. GPT-4o-mini generally outperforms the other models. This figure shows the results for the twelve different prompting strategies across all four datasets.

While `10 ex. mixed` performs best on average, the optimal strategy can vary by dataset. King County benefits from combining market reports with ten geographic neighbors (`report + 10 ex. geo`), Flanders and Barcelona favor ten mixed examples (`10 ex. mixed`), and Beijing performs best with market reports and three hedonic examples (`report + 3 ex. hedonic`). While a mixed example selection approach generally ranks best, geographic neighbors provide typically more useful contextual information than hedonic examples. Again, the Beijing dataset deviates from this observation and ranks hedonic examples higher. In most regions, spatial correlations in house prices are prominent, while Beijing's market may be influenced by higher levels of heterogeneity leading to property characteristics being more important in combination with broader economic trends. Another reason could be that the spatial structure in Beijing's house prices may be more complex than geographic proximity.

While market reports consistently improve zero-shot prompting, they are rarely effective on their own. Notably, in Beijing, monthly, city-specific reports ($\approx$ 501 characters in length) improve results across all prompting strategies, reinforcing the importance of market trends in this region and underscoring the value of fine temporal and spatial resolution. In the King County dataset, its US-wide, monthly market reports ($\approx$ 987 characters), which include quarterly sub-regional breakdowns of market trends, also benefit price prediction in most cases. In contrast, Barcelona and Flanders see limited or no gains: for both regions, quarterly, country-level reports ($\approx$ 1 817 characters for Barcelona, $\approx$ 1 566 characters for Flanders) are available, with Flanders additionally comparing neighboring countries and Barcelona highlighting only regions with the largest changes. This pattern suggests that higher temporal granularity (monthly vs. quarterly), tighter geographic specificity (city vs. national), and even concise report length materially enhance an LLM's ability to incorporate dynamic market trends. Given that the King County and especially the Beijing dataset present significant temporal trends in house prices, these results show that market reports provide essential context for LLMs to capture temporal dynamics and thereby improve predictive performance.

Ranking the LLMs across datasets and prompting strategies reveals that GPT-4o-mini outperforms the other models, with Llama 3.1:70B as a close second. While GPT-4o-mini still achieves reasonable results with zero-shot prompts, Llama 3.2:3B consistently performs worst. This might be due to low-parameter models containing less world knowledge compared to high-parameter models, making them more dependent on prompts with richer context for accurate predictions. While larger models, Llama 3.1:70B and GPT-4o-mini, generally outperform smaller models, LLM performance also slightly varies across datasets. Llama 3.2:3B performs comparably to larger models on the Beijing dataset. This anomaly could be again due to the volatile market dynamics which makes all LLMs struggle equally. GPT-4o-mini performs best in King County and Flanders, whereas Llama 3.1:70B outperforms it in Barcelona and Beijing. This performance disparity could be attributed to GPT-4o-mini's stronger tailoring to the US and Western European contexts, where its training data and fine-tuning

are more focused [10]. In contrast, Llama 3.1, designed for multilingual text generation, appears to perform better for non-English or geographically diverse contexts, making it more adept at handling the varied linguistic and cultural features found in regions such as Beijing and Barcelona [25]. Furthermore, GPT-4o-mini exhibits the least variability across prompting strategies, making it the most stable performer overall. While its advantage over Llama models is not always substantial, it consistently follows structured formatting, reducing output inconsistencies. Llama models, particularly Llama 3.2:3B, sometimes struggle to conform to the output format, which can contribute to higher prediction errors.

Despite some dataset-specific variations, `10 ex. mixed` emerges as the most robust prompting strategy, ranking the highest when averaged over models and datasets. This approach effectively balances geographic and hedonic information, making it a strong default choice when selecting prompting strategies for house price estimation. While market reports may further enhance performance in markets with strong temporal trends, their usefulness varies by region. Therefore, a hybrid approach—prioritizing mixed examples and potentially incorporating market reports when relevant—offers the most generalizable strategy.

4.2 Positioning LLM Performance Relative to ML Baselines

Prediction accuracy Table 2 compares the performance of LLMs with baseline methods using MAPE across datasets. The table shows the LLM results for the best-ranked prompt strategy `10 ex. mixed` and corresponding setting for kNN, alongside a SOTA GBT model with (LGBM) and without coordinates (LGBM ∅ XY). Generally, high-parameter LLMs outperform kNN, indicating they leverage labeled examples through ICL more effectively than kNN's simple interpolation. The Beijing dataset presents a particular challenge, with all LLMs performing worse than kNN. This is due to the strong temporal trends as established earlier, and can be mitigated with a market report (`report + 10 ex. mixed`) which results in a decrease in MAPE from 0.4022 to 0.3322 for Llama 3.1:70B, effectively outperforming kNN (0.3810).

Table 2: LLMs generally outperform kNN and get competitive to SOTA models. Comparison of MAPE and PE Standard Deviation between LLMs with `10 ex. mixed` prompt and baseline models.

	King County	Flanders	Barcelona	Beijing
Llama 3.2:3B	0.2995 ± 0.3282	0.4511 ± 0.5828	0.3383 ± 0.4211	0.4092 ± 0.1320
Llama 3.1:70B	0.1905 ± 0.2072	0.3440 ± 0.4997	<u>0.1825</u> ± 0.2044	<u>0.4022</u> ± 0.1108
GPT-4o-mini	<u>0.1861</u> ± 0.1925	<u>0.3170</u> ± 0.4782	0.1842 ± 0.2004	0.4125 ± 0.1117
kNN	0.2105 ± 0.2113	0.3207 ± 0.4380	0.2638 ± 0.4070	0.3810 ± 0.1292
LGBM ∅ XY	0.2391 ± 0.3220	0.3136 ± 0.4988	0.1936 ± 0.2825	0.2427 ± 0.2031
LGBM	**0.1378** ± 0.1611	**0.2625** ± 0.4170	**0.1556** ± 0.1780	**0.1056** ± 0.0840

Comparing LLMs with LGBM ∅ XY, we see that LLMs show comparable performance. This indicates that LLMs are effective in extracting hedonic patterns from real estate pricing data. Finally, comparing LLMs with the SOTA LGBM models, we do not expect LLMs to perform better, but we see that LLMs can get relatively close without having access to the full dataset. In Flanders and Barcelona, GPT-4o-mini's MAPE is around 20% higher than LGBM, while in King County the difference is 35%. However, with the optimal strategy (`report + 10 ex. geo`), the MAPE in King County improves to 0.1390, completely matching LGBM's performance.

Interestingly, the Beijing dataset deviates in the baseline performance opposed to other datasets as well. It sees a great decrease in MAPE between kNN and LGBM ∅ XY, likely due to LGBM's ability to take advantage of temporal features. In addition, adding the geographic coordinates as predictors to LGBM reduces the MAPE from 0.2427 to 0.1056 for Beijing. While LLMs struggle with incorporating spatial relationships through neighboring examples, LGBM succeeds in deciphering the spatial structure in the dataset and significantly improves predictions. This strengthens our findings that LLMs can learn hedonic pricing patterns, but require more advanced techniques when the dataset is characterized by unconventional spatial structures and strong temporal dynamics.

The PE Standard Deviation shows LLMs have error variability comparable to baselines, though Llama 3.2 exhibits greater fluctuation. Overall, LLMs surpass kNN and show competitive performance compared to SOTA models, particularly in extracting hedonic patterns from real estate pricing data.

Price Intervals To address **RQ3**, we assess LLM prediction intervals using the `10. mixed` prompt. Table 3 reports two metrics: coverage (percentage of true prices within intervals) and MPIW (Mean Prediction Interval Width).

Table 3: Prediction interval quality measured by Coverage (Cov.), percentage of true prices in test sample within intervals, and MPIW, Mean Prediction Interval Width. As we enforce LLMs to produce intervals around their predicted price, we included respectively 949, 872, 960, and 945 intervals for Llama 3.2:3B and 998 for Llama 3.1:70B on the Flanders dataset and 999 on the Beijing dataset.

	King County		Flanders		Barcelona		Beijing	
	Cov.	MPIW	Cov.	MPIW	Cov.	MPIW	Cov.	MPIW
Llama 3.2:3B	39.6	220 289	36.7	193 447	46.8	262 199	<u>10.8</u>	1 625 475
Llama 3.1:70B	<u>57.5</u>	<u>182 823</u>	<u>51.4</u>	<u>156 658</u>	<u>64.0</u>	<u>151 641</u>	3.6	<u>1 093 476</u>
GPT-4o-mini	35.5	**98 319**	25.8	**65 488**	40.3	**74 444**	1.2	**514 394**
LGBM	**90.5**	316 293	**90.5**	317 476	**86.2**	210 681	**85.1**	1 900 473

LLMs generate narrower intervals but often miss the 90% coverage target, showing overconfidence [36]. In contrast, conformal prediction enables LGBM to

achieve near-target coverage but with wider intervals, illustrating the trade-off between coverage and precision. GPT-4o-mini produces the narrowest intervals but consistently underperforms on coverage, while Llama 3.1 offers the best balance across datasets. The Beijing dataset proves particularly difficult, with LLMs showing extremely low coverage and LGBM struggling despite conformal adjustments, likely due to the dataset's temporal trends. Despite adjusting for this distribution shift, this still influences predictions. LLMs may also lack geographical knowledge or show regional biases in Beijing [23]. Overall, it is clear that LLMs struggle with producing calibrated prediction intervals, but advanced techniques like conformal prediction or iterative prompting [38,39] that would be necessary to mitigate this problem, make it less evident for real estate practitioners to leverage LLM-based solutions.

Feature Importance We compare LLM-generated feature explanations to SHAP values from LGBM in Figure 3, which shows the Venn diagrams of the top five features for all datasets. GPT-4o-mini generally aligns with LGBM on hedonic features, supporting their ability to extract property-related pricing patterns. This alignment suggests that LLM-generated explanations are not only consistent with established ML models but also offer a degree of trustworthiness, as they reflect key predictive drivers identified through robust, model-agnostic interpretability methods like SHAP.

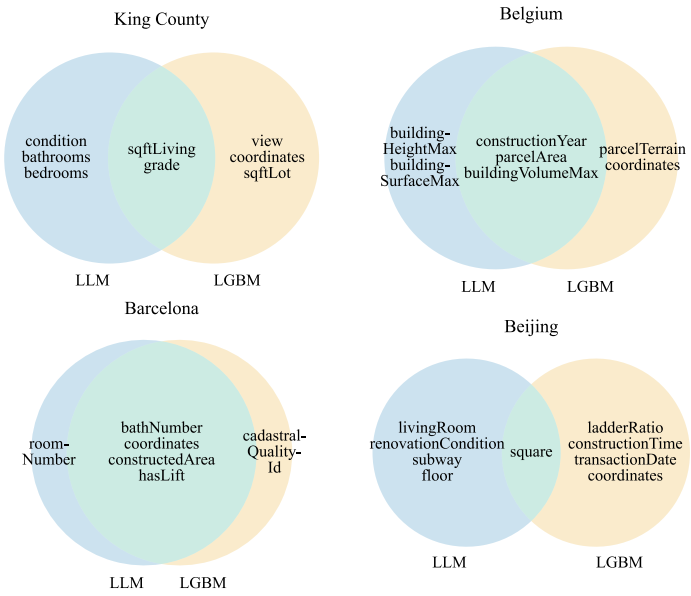

Fig. 3: LLMs generally align with LGBM on the importance of hedonic variables. Comparison of top five features between GPT-4o-mini and LGBM.

However, LGBM consistently ranks locational features, particularly coordinates, among its top predictors, while LLMs do not prioritize them, despite receiving full addresses and coordinates. The only exception is in the Barcelona dataset. This suggests that LLMs struggle with spatial reasoning, likely due to tokenization issues with coordinates and difficulty mapping addresses to house price patterns [19,24]. Additionally, in accordance to the previous analyses, LGBM prioritizes temporal features in the Beijing dataset, while GPT-4o-mini does not. Similar to LLMs' issues with interpreting coordinates, LLMs might struggle with dates and generalizing temporal trends in the data. Recent research has focused on improving temporal generalization of LLMs [13]. Although these limitations in spatial and temporal reasoning may explain the performance gap, caution is warranted since LLM self-explanations are not always reliable [21]. Appendix C confirms the findings for Llama3.1:70b.

4.3 Practical implications

Our findings indicate that LLMs offer a low-barrier alternative to traditional ML solutions for real estate appraisal. With as few as ten examples, LLMs can provide reasonable price estimates, making them valuable for quick, accessible, and interactive valuations. We recommend structuring prompts with ten properties that are geographically near and share similar hedonic characteristics, while incorporating a market report in cases with strong temporal trends. LLMs are particularly suited for decision-support systems, especially for non-technical users and low-data environments. Private buyers and sellers, for instance, can use LLMs to gain insight into fair market prices, while creditors, real estate agents, and investors may rely on them for preliminary valuations that can later be refined with expert assessments or ML-based approaches.

Table 4 summarizes the key advantages and limitations of LLMs compared to ML models. While LLMs work out of the box, we identify three main limitations: they struggle with spatial reasoning, failing to properly integrate location effects; they exhibit weak temporal understanding, making it difficult to learn price trends over time; and they are overconfident in uncertainty estimation, often producing narrow and unreliable price intervals. Though market reports improve temporal generalization capabilities, fully addressing these weaknesses requires more advanced techniques, such as retrieval-augmented generation or fine-tuning [41]. These approaches introduce additional complexity that can undermine the accessibility and immediacy that make LLMs attractive alternatives to traditional ML methods. Similarly, reliable uncertainty estimation requires post-processing techniques that are challenging, especially with closed-source models [38,39]. Given these trade-offs, LLMs are best suited for fast, accessible valuations rather than high-accuracy, large-scale appraisals where calibrated uncertainty estimates are essential.

Table 4: Comparison of LLMs and ML models for real estate appraisal

Aspect	LLMs	ML Models
Accessibility		
Ease of Use	Works out of the box	Requires training & tuning
User Input	Natural Language via interface	Structured data through code
Data Requirements		
Data Needs	Few examples needed	Large structured dataset
Feature Handling	Implicit understanding	Manual selection required
Unstructured Data	Can use reports & text	Limited to structured inputs
Model Capabilities		
Accuracy	Competitive, slightly below ML	State-of-the-art
Geospatial Data	Limited handling	Explicitly modeled
Temporal Trends	Struggles with time patterns	Can model time effects
Prediction Intervals	Often overconfident	More calibrated
Explainability	Text-based, intuitive	SHAP & feature importance
Practical Deployment		
Interactivity	Can take user feedback	Static predictions
Computation	Free/low-cost web, API, local	Local CPU/GPU
Best Use Case	Quick, flexible valuations	Accurate, large-scale modeling

5 Conclusion

This study investigated how prompt engineering techniques optimize LLM performance for real estate appraisal (**RQ1**) and whether LLMs can serve as viable alternatives to traditional ML models (**RQ2-4**). Our results show that LLMs, when prompted using In-Context Learning with just ten real estate examples selected based on geographic and hedonic similarity (**RQ1**), can generate competitive price estimates (**RQ2**), making them a practical tool for real estate valuation. However, their spatial reasoning and temporal generalization capabilities remain limited, affecting the reliability of predicted price intervals (**RQ3**) compared to structured ML models. Nevertheless, LLMs align with ML models in explaining predictions based on property characteristics (**RQ4**), reinforcing their ability to capture hedonic valuation patterns.

By improving accessibility to property appraisal, LLMs help reduce information asymmetry in real estate transactions. While ML models remain more accurate in structured, large-scale applications, LLMs provide an interactive and intuitive alternative, particularly for non-technical users who need quick and interpretable price estimates. These findings align with **RQ2** and **RQ4**, highlighting that while LLMs can extract meaningful hedonic features, they require further refinement to fully capture spatial and temporal trends.

In summary, LLMs show potential for accurate and explainable price predictions. Future work should explore more recent LLMs with enhanced reasoning ca-

pabilities, alongside diverse prompting techniques such as chain-of-thought and self-consistency, or retrieval-augmented generation to further improve performance and robustness. Additionally, a systematic investigation into scaling LLMs with large-scale datasets is needed to provide deeper insights into data efficiency and generalization, though our results indicate that larger model sizes generally yield better accuracy. Exploring alternative geographic encoding strategies could also address current spatial reasoning limitations. Collectively, these directions offer a clear roadmap to enhance LLM trustworthiness and reliability, advancing their practical application in real estate appraisal.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bastos, J.A., Paquette, J.: On the uncertainty of real estate price predictions. J Prop Res pp. 1–19 (2024). https://doi.org/10.1080/09599916.2024.2403998
2. Bourassa, S.C., Hoesli, M., Peng, V.S.: Do housing submarkets really matter? J Hous Econ **12**, 12–28 (2003). https://doi.org/10.1016/S1051-1377(03)00003-2
3. Chen, T., Si, S.: Predicting rental price of lane houses in shanghai with machine learning methods and large language models. arXiv preprint arXiv:2405.17505 (2024)
4. Coda-Forno, J., Binz, M., Akata, Z., Botvinick, M., Wang, J., Schulz, E.: Meta-in-context learning in large language models. In: Advances in Neural Information Processing Systems. vol. 36, pp. 65189–65201 (2023)
5. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
6. Feng, S., Lyu, H., Li, F., Sun, Z., Chen, C.: Where to move next: Zero-shot generalization of llms for next poi recommendation. In: IEEE Conference on Artificial Intelligence (2024). https://doi.org/10.1109/CAI59869.2024.00277
7. Geerts, M., vanden Broucke, S., De Weerdt, J.: A survey of methods and input data types for house price prediction. ISPRS Int J Geo-Inf **12**, 200 (2023). https://doi.org/10.3390/ijgi12050200
8. Geerts, M., vanden Broucke, S., De Weerdt, J.: Graph neural networks for house price prediction: do or don't? Int J Data Sci Anal (2024). https://doi.org/10.1007/s41060-024-00682-y
9. Gloria, B., Melsbach, J., Bienert, S., Schoder, D.: Real-gpt: Efficiently tailoring llms for informed decision-making in the real estate industry. J Real Estate Portf Manag (2024). https://doi.org/10.1080/10835547.2024.2372748
10. Guo, H., Venkit, P.N., Jang, E., Srinath, M., Zhang, W., Mingole, B., Gupta, V., Varshney, K.R., Sundar, S.S., Yadav, A.: Hey gpt, can you be more racist? analysis from crowdsourced attempts to elicit biased content from generative ai. arXiv preprint arXiv:2410.15467 (2024)
11. Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., Sontag, D.: Tabllm: Few-shot classification of tabular data with large language models. In: 26th International Conference on Artificial Intelligence and Statistics. vol. 206, pp. 5549–5581. PMLR (2023)

12. Hjort, A., Williams, J.P., Pensar, J.: Clustered conformal prediction for the housing market. In: 13th Symposium on Conformal and Probabilistic Prediction with Applications. Proceedings of Machine Learning Research, vol. 230, pp. 366–386. PMLR (2024)
13. Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J.Y., Shi, X., Chen, P.Y., Liang, Y., Li, Y.F., Pan, S., Wen, Q.: Time-llm: Time series forecasting by reprogramming large language models. In: 12th International Conference on Learning Representations (2023)
14. Jin, R., Xu, Q., Wu, M., Xu, Y., Li, D., Li, X., Chen, Z.: Llm-based knowledge pruning for time series data analytics on edge-computing devices. arXiv preprint arXiv:2406.08765 (2024)
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: 31st International Conference on Neural Information Processing Systems. vol. 30, pp. 3146–3154 (2017)
16. Kurlat, P., Stroebel, J.: Testing for information asymmetries in real estate markets. Rev Financ Stud **28**, 2429–2461 (2015). https://doi.org/10.1093/rfs/hhv028
17. Lee, H., Jeong, H., Lee, B., Lee, K.D., Choo, J.: St-rap: A spatio-temporal framework for real estate appraisal. In: 32nd ACM International Conference on Information and Knowledge Management. pp. 4053–4058 (2023)
18. Li, C., Wang, W., Du, W., Peng, W.: Look around! a neighbor relation graph learning framework for real estate appraisal. In: Advances in Knowledge Discovery and Data Mining (2024). https://doi.org/10.1007/978-981-97-2238-9_1
19. Li, F., Hogg, D.C., Cohn, A.G.: Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In: AAAI Conference on Artificial Intelligence. vol. 38, pp. 18500–18507 (2024). https://doi.org/10.1609/aaai.v38i17.29811
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: 31st International Conference on Neural Information Processing Systems. vol. 30, pp. 4768–4777 (2017)
21. Madsen, A., Chandar, S., Reddy, S.: Are self-explanations from large language models faithful? In: Findings of the Association for Computational Linguistics. pp. 295–337 (2024). https://doi.org/10.18653/v1/2024.findings-acl.19
22. Manikandan, H., Jiang, Y., Kolter, Z.: Language models are weak learners. In: Advances in Neural Information Processing Systems. vol. 36, pp. 50907–50931 (2023)
23. Manvi, R., Khanna, S., Burke, M., Lobell, D., Ermon, S.: Large language models are geographically biased. arXiv preprint arXiv:2402.02680 (2024)
24. Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D.B., Ermon, S.: GeoLLM: Extracting geospatial knowledge from large language models. In: 12th International Conference on Learning Representations (2024)
25. Meta: Llama 3.2: Revolutionizing edge ai and vision with open, customizable models (2024), https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/
26. OpenAI: Hello gpt-4o (2024), https://openai.com/index/hello-gpt-4o/
27. Ozhegov, E.M. and Ozhegova, A.: Distance in geographic and characteristics space for real estate pricing. Int J Hous Mark Anal **15**, 938–952 (2022). https://doi.org/10.1108/IJHMA-04-2021-0041
28. Rey-Blanco, D., Arbues, P., Lopez, F., Paez, A.: A geo-referenced micro-data set of real estate listings for spain's three largest cities. Environ Plan B: Urban Anal City Sci **51**, 1369–1379 (2024). https://doi.org/10.1177/23998083241242844

29. Rico-Juan, J.R., de La Paz, P.T.: Machine learning with explainability or spatial hedonics tools? an analysis of the asking prices in the housing market in alicante, spain. Expert Syst Appl **171** (2021). https://doi.org/10.1016/j.eswa.2021.114590
30. Shen, L., Liu, Q., Chen, G., Ji, S.: Text-based price recommendation system for online rental houses. Big Data Min Anal **3**, 143–152 (2020). https://doi.org/10.26599/BDMA.2019.9020023
31. Taquet, V., Blot, V., Morzadec, T., Lacombe, L., Brunel, N.: Mapie: an open-source library for distribution-free uncertainty quantification. arXiv preprint arXiv:2207.12274 (2022)
32. Vacareanu, R., Negru, V.A., Suciu, V., Surdeanu, M.: From words to numbers: Your large language model is secretly a capable regressor when given in-context examples. arXiv preprint arXiv:2404.07544 (2024)
33. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005). https://doi.org/10.1007/978-3-031-06649-8
34. Weiss, M., Rahaman, N., Wuthrich, M., Bengio, Y., Li, L.E., Schölkopf, B., Pal, C.: Redesigning information markets in the era of language models. In: First Conference on Language Modeling (2024)
35. Wu J., Hou M.: An Efficient Retrieval-Based Method for Tabular Prediction with LLM. In: 31st International Conference on Computational Linguistics. pp. 9917–9925 (2025)
36. Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., Hooi, B.: Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In: 12th International Conference on Learning Representations (2024)
37. Xu, C., Xie, Y.: Conformal prediction for time series. IEEE Trans Pattern Anal Mach Intell **45**, 11575–11587 (2023). https://doi.org/10.1109/TPAMI.2023.3272339
38. Yadkori, Y.A., Kuzborskij, I., György, A., Szepesvari, C.: To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty. In: Advances in Neural Information Processing Systems. vol. 37, pp. 58077–58117 (2024)
39. Ye, F., Yang, M., Pang, J., Wang, L., Wong, D., Yilmaz, E., Shi, S., Tu, Z.: Benchmarking llms via uncertainty quantification. In: Advances in Neural Information Processing Systems. vol. 37, pp. 15356–15385 (2024)
40. Zhang, H., Li, Y., Branco, P.: Describe the house and i will tell you the price: House price prediction with textual description data. Nat Lang Eng **30**, 661–695 (2024). https://doi.org/10.1017/S1351324923000360
41. Zhang, Y., Wang, Z., He, Z., Li, J., Mai, G., Lin, J., Wei, C., Yu, W.: Bb-geogpt: A framework for learning a large language model for geographic information science. Inf Process Manag **61** (2024). https://doi.org/10.1016/J.IPM.2024.103808