

SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/

e-ISSN: 2636-8129
Publisher: Sakarya University

Vol. 9, No.3, 690-699, 2026
DOI:10.35377/saucis...1829206

SAKARYA UNIVERSITY

RESEARCH ARTICLE

Feature Extraction for Real Estate Images and Titles with LLMs

Afra Arslan [1*], **Tan Doruk Yetki** [1], **Arda Yücel** [1], **Hacer Turgut** [1], **Ömür Bali** [1], **Gülfem Işıklar Alptekin** [2], **Günce Keziban Orman** [2]

[1]iLab, Research and Development Center, Istanbul, 34736, Türkiye
[2]Galatasaray University, Computer Engineering Department, 34349 Istanbul, Türkiye

*Corresponding author:
aarslan@ilab.com.tr

Article History:
Received: 24 November 2025
Revised: 21 April 2026
Accepted: 05 May 2026
Published Online: 19 June 2026

Abstract
Images and titles often contain rich latent information about their associated objects, particularly on web-based platforms. Real estate websites provide a clear example, where listing images and titles provide important details that assist users in their decision-making. However, these unstructured elements cannot be directly utilized in downstream machine learning tasks, since their contextual meaning is not directly interpretable. This work aims to transform listing images and titles into structured, tabular representations, making them suitable for analytical and predictive modeling. To this end, we propose a modular framework based on state-of-the-art large language models. The framework incorporates ReAct, LLM-as-a-Judge, and few-shot prompting techniques. Its performance is evaluated on a real-world real estate dataset and compared with BERT and CLIP-based baselines. Experimental results demonstrate that our framework achieves up to a 44.26% improvement in recall for listing attributes, such as the presence of a balcony or the furnishing status of a property.

Keywords: CLIP, Feature extraction, Information retrieval, LLM, Real estate

1. Introduction

In the highly competitive landscape of online real estate, listing images and titles play a pivotal role in attracting potential buyers and driving engagement. As the initial point of interaction, these multimedia elements shape user perception and strongly influence decision-making. However, unlike structured data fields, images and titles are often presented in unstandardized formats, leading to inconsistencies in data quality. Such inconsistencies hinder systematic analysis and limit the ability to retrieve meaningful information. Consequently, this affects users' ability to identify suitable options and make informed choices.

Extracting structured data from unstructured textual and visual content has long been a challenging task. Nevertheless, recent advances in natural language processing and computer vision have led to remarkable progress. In the text domain, prompt engineering techniques have been explored to enhance the accuracy of large language models (LLMs), such as Google Bard and ChatGPT, for extracting structured trip information from e-mails [17]. Similarly, neural joint models have been proposed to extract biomedical entities from text [10], while pre-trained LLMs such as GPT-3 and LLaMA2 have been utilized to extract structured data in materials chemistry by combining named-entity recognition and relation extraction [2]. To address domain-specific challenges, MatSciBERT was introduced as a specialized model that improves information extraction performance in materials science [6]. More recently, LLMs have demonstrated strong performance in extracting structured attribute values directly from real estate-related textual data, highlighting their applicability in domain-specific information extraction tasks [9]. These studies collectively highlight LLMs' growing ability to generalize across domains and extract structured information without requiring task-specific model training.

In parallel, image-based extraction techniques have also progressed rapidly. Models such as AdaLN [5] use vision transformers and deep neural networks to classify architectural features from satellite data, while Vision GNN [7] represents image pixels as graph nodes to enhance reasoning over irregular objects. Cross-modal approaches such as CLIP [14] have further advanced large-scale image-text alignment by enabling

Cite as: A Arslan et al, "Feature extraction for real estate images and titles with llms " , *Sakarya University Journal of Computer and Information Sciences* 9 (3) 2026, 690-699. doi: 10.35377/saucis...1829206


This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License

semantically rich representations from both modalities. Standard evaluation metrics like precision, recall, and F1-score are now frequently complemented by benchmark suites such as GLUE and SuperGLUE [18], as well as calibration and human-in-the-loop evaluations [3].
Beyond general-purpose extraction, recent studies have also explored AI-driven applications specifically within the real estate domain. For instance, AI-based analyses have been utilized to investigate the economic impact of latent spatial and visual factors (e.g., Feng Shui elements), demonstrating the value of extracting implicit features from real estate data [11]. Similarly, autonomous-agent frameworks have been proposed for generating real estate marketing videos, highlighting the growing role of multimodal AI systems in property marketing [19]. These studies highlight both the growing importance of unstructured data in real estate and the need for robust methods to transform it into structured representations. However, these approaches primarily focus on content generation or high-level analysis, rather than the structured extraction and standardization of multimodal listing attributes required for downstream machine learning tasks.
Despite these advances, limited research has addressed the transformation and analysis of unstructured listing images and titles in the context of real estate. This gap is critical because the absence of standardized formats leads to inconsistent categorical labeling, limited automation, and reduced reliability in downstream machine learning applications. Existing methods still rely on manual annotation or simple extraction techniques, both of which are difficult to scale, prone to human error, and often focus on single modalities rather than jointly modeling textual and visual information.
To address these challenges, this paper proposes a novel framework that leverages LLMs to extract and organize categorical information from real estate listing images and titles. The framework integrates various prompting strategies, including few-shot prompting [1], LLM-as-a-Judge [21], and ReAct [20], to minimize errors and improve consistency. Its effectiveness is validated on a manually labeled dataset that captures image features (e.g., image type, logo presence) and title features (e.g., realtor information, occurrence of comment-related terms).
We further evaluate our framework through predictive experiments that serve as baselines. Title embeddings are generated using BERT [4] and applied to a binary logistic regression model, a binary SVM model, and a basic MLP model to predict categorical features. We also explored the impact of dimensionality reduction through Principal Component Analysis (PCA) [12] on BERT embeddings. Parallel experiments used image embeddings generated by CLIP under the same conditions. Results demonstrate that combining diverse prompting techniques significantly enhances categorization accuracy and consistency. Moreover, the structured data produced by our framework proves highly effective for downstream analytical and predictive applications.
The main contributions of this work are as follows:

- Introducing an LLM-based framework to extract and structure categorical information from unstandardized listing titles and images.
- Demonstrating the framework's effectiveness using a real-world real estate dataset.
- Analyzing the positive and negative effects of integrating diverse prompting strategies on accuracy and robustness.
- Manually labeling a testing dataset to effectively evaluate performance.
- Presenting baseline experiments using BERT and CLIP embeddings, both with and without PCA.

2. Dataset and Preprocessing

The main objective of this study is to identify predefined categories from listing images and titles. This process enables the transformation of unstructured listing information into a structured format that can be directly used in downstream machine learning tasks. This section provides a detailed overview of the listing dataset and its properties.

2.1. Dataset description

The dataset used in this study was collected from Emlakjet.com, a leading Turkish real estate platform that hosts a wide variety of property listings submitted by both individual users and professional realtors. The website attracts millions of visitors each month. An illustrative example of a listing is shown in Figure 1. The dataset includes two key modalities: (*i*) a cover image and (*ii*) a title of each listing. The dataset consists of 500 listing titles and 150 listing images. The data was collected from listings published in October 2023 and is geographically focused on the Kadıköy district of Istanbul, Turkey, ensuring consistency in regional characteristics while maintaining diversity in listing content.
One of the main challenges with this dataset is the reliability and quality of its content. For instance, the dataset may contain entries with low-quality images or vague or uninformative titles. To improve data quality and avoid redundancy, only unique and non-duplicate listings were included in the dataset. Additionally, listings with severely low-resolution or irrelevant images were excluded during the manual filtering process.



Figure 1. Representative image and title of a listing from Emlakjet

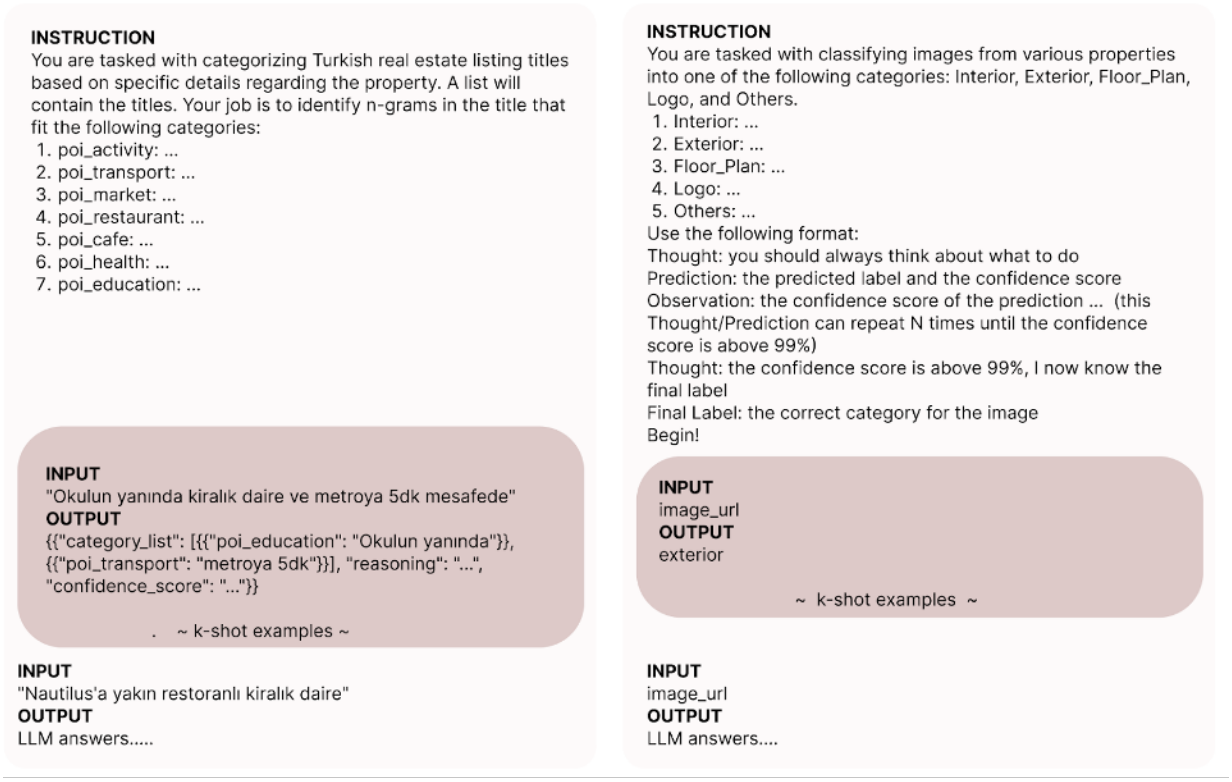

Figure 2. Prompts used in the proposed framework: the left panel presents the prompts for titles, while the right panel displays those used for images

2.2. Data manipulation

We contributed to the dataset in two ways: First, after manually examining a large number of listings, we developed a labeling schema for key categories that capture essential structural and semantic attributes, as well as domain expertise. For the titles, these categories include, but are not limited to, information about location, essential listing details, price, age, personal remarks from the owner, and points of interest mentioned in the title. For the images, we identified visual characteristics such as the presence of furniture, the perceived level of luxury, the amount of sunlight, and many more. This taxonomy was developed to reflect both objective (e.g., price, sunlight) and subjective (e.g., luxury perception) characteristics of real estate listings, reflecting the combined domain expertise of Emlakjet professionals and the authors.

Second, to enable quantitative evaluation and facilitate model training, we manually labeled a subset of the dataset, following the proposed scheme. During sampling, listings were randomly selected while explicitly excluding duplicate entries, ensuring that each sample represents a unique property. It is important to note that there is no inherent one-to-one correspondence between the selected titles and images, as each modality was sampled independently and treated as a separate source of information.

The annotation process was carried out by a single domain expert with knowledge of both the real estate domain and the labeling schema. Each listing was reviewed individually and assigned labels according to predefined category definitions. To ensure consistency, the annotator followed detailed written guidelines, and ambiguous cases were resolved through consultation with the research team. While single-annotator setups do not allow computation of inter-annotator agreement metrics, explicit definitions and a structured review process were adopted to mitigate subjectivity.

Before annotation and feature extraction, the title data underwent basic preprocessing steps, including normalization of Turkish characters and correction of typographical errors. Diversity was ensured across (*i*) property types, (*ii*) listing agents, and (*iii*) neighborhoods, all sampled from the same metropolitan region. These labeled examples serve as ground-truth annotations for two lines of experimentation: (*i*) feature extraction using LLM-based prompting techniques, and (*ii*) performance benchmarking using machine learning classifiers trained on textual and visual embeddings.

3. Methodology

For our extraction and classification tasks, we selected the GPT-4o-mini model, an LLM developed by OpenAI [13]. This model was chosen to achieve an optimal balance between cost-efficiency and performance. A key advantage of GPT-4o-mini is its ability to process both textual and visual inputs, making it particularly suitable for our joint analysis of listing titles and images. Among several alternatives -including GPT-3.5-turbo, GPT-4, and GPT-4o- GPT-4o-mini consistently offered the most favorable trade-off between inference cost and extraction accuracy. While larger models may provide marginal performance gains, their substantially higher computational cost would limit their practicality for scalable deployment. In contrast, GPT-4o-mini provides sufficiently accurate and consistent outputs across both modalities while maintaining a significantly lower cost profile. Based on these considerations, GPT-4o-mini was identified as the most suitable foundation for our proposed methodology.

To ensure reproducibility of the proposed framework, the prompt design and inference procedure were standardized across experiments. We employed a structured prompting strategy in which the model was explicitly instructed to generate outputs in a predefined JSON format. The prompts consisted of both system-level instructions, defining the task and constraints, and user-level inputs, providing the listing content (title or image). Detailed prompt templates and implementation specifics are available in our public repository [8]. To maintain consistency and allow

for standard model behavior, all parameters were kept at their library defaults: a sampling temperature of 0.7, no explicit limit on max tokens (allowing the model's native maximum), and no explicit request timeout, relying on the default behavior of the underlying OpenAI Python SDK. Structured output was enforced by injecting parser-derived format instructions into prompts and validating responses against Pydantic schemas (via custom JSON parsers). This approach constrained outputs to predefined JSON keys and data types. When the generated output did not conform to the expected format or contained parsing errors, a retry mechanism was applied, allowing up to 3 attempts per sample.

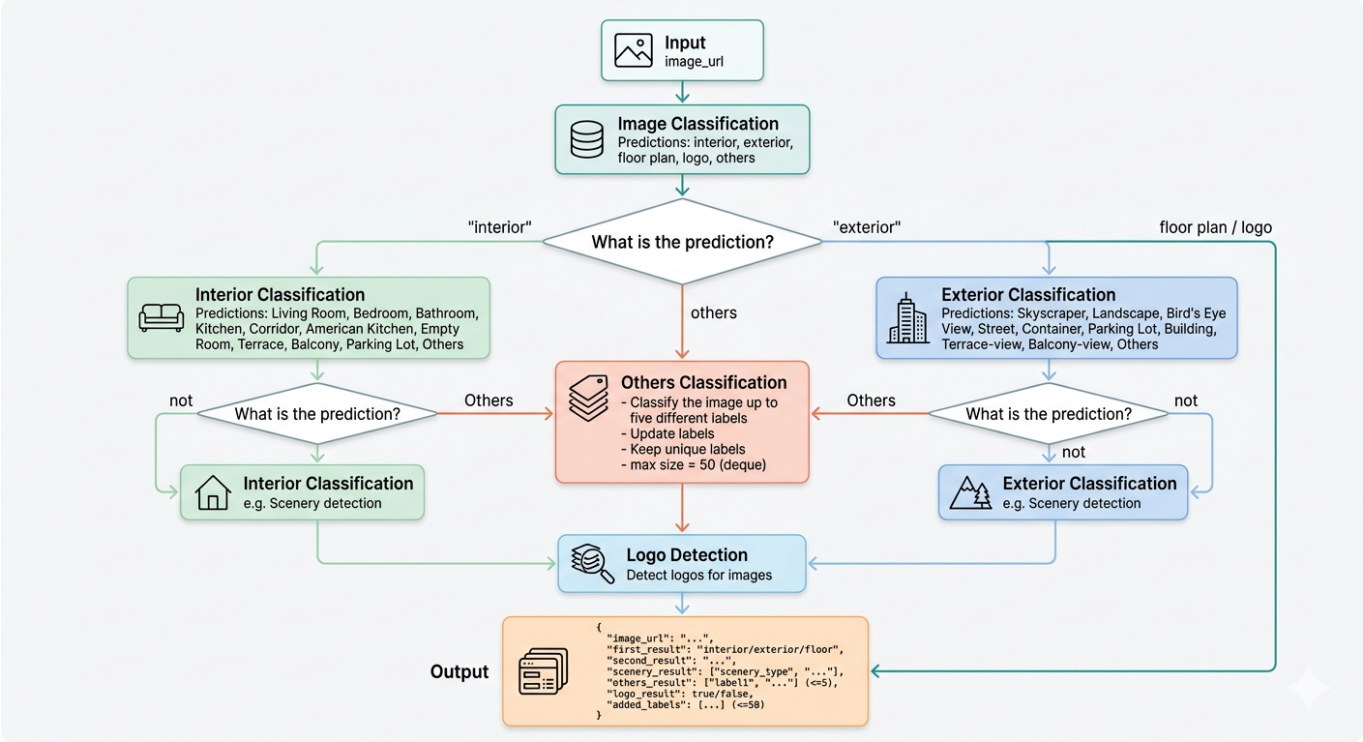

Figure 3. Hierarchical image classification pipeline

3.1. Image engineering

The primary difficulty in real estate images lies in effectively interpreting their content. In our framework, image content is transformed into predefined categories. To address the inherent variability of the data, we adopted two key prompting strategies: (*i*) few-shot prompting and (*ii*) ReAct prompting. Few-shot prompting provides the model with a small set of labeled examples to help its recognition of visual elements. In our case, the examples highlighted property types, architectural styles, and the presence of logos. By learning from these annotated samples, the LLM gains a better understanding of how to generate structured metadata from images. However, this approach can be less effective with low-quality images or outliers. To improve robustness, we incorporated ReAct prompting, which improves a language model's reasoning and decision-making through iterative sequences of thought, action, and observation. This process allows the model to gradually refine its understanding and produce more accurate outputs. We adapted this structure to enhance the model's understanding of real estate images. More specifically, we modified the standard thought-observation-action structure into a thought-observation-prediction structure, as shown in Figure 2. This modification strengthens the model's predictive capability and leads to a more coherent visual assessment. As a result, the framework achieved better performance in identifying property types and architectural design features, yielding more stable and accurate predictions overall.
In addition to prompt design, the image-based feature extraction pipeline follows a structured and hierarchical workflow, as illustrated in Figure 3. The process begins with a high-level classification (interior, exterior, floor plan, logo, others), followed by specialized sub-classification modules. Depending on the predicted category, the pipeline dynamically routes the input to interior, exterior, or floor plan/logo branches. For non-standard cases, a separate "others" classification module assigns up to five auxiliary labels. For images identified as interior or exterior, a second-stage classification is performed to detect category-specific classes. If the prediction at this stage is not sufficiently specific, the image is redirected to a fallback "others" classification module again. Finally, logo detection is applied as a post-processing step across all images.
The final output is a structured JSON representation that aggregates results from all stages, including the primary category, secondary classification, detected image attributes, auxiliary labels, and logo presence. This modular pipeline enables flexible and extensible feature extraction while maintaining consistency across different image types.

3.2. Title engineering

Real estate listing titles are typically short texts that may contain spelling errors. To address these issues, we first applied pre-processing to the input data using an LLM. Subsequently, we identified four key categories within the titles: (*i*) Points of Interest tagging, i.e., modes of transportation, potential activities near the listing, or educational centers, (*ii*) Real Estate Detail tagging, i.e., information about the property such as location, furnishing status, and building age, (*iii*) Other – Comments tagging, i.e., personal remarks or promotional adjectives provided by the seller or realtor, e.g 'Luxurious', and (*iv*) Rule-based tagging, an indication of whether the listing is for rent or for sale.
In addition to LLM-based tagging, the Rule-based tagging is incorporated to identify deterministic patterns such as transaction type (e.g., for sale, for rent) and property category (e.g., residence). This hybrid approach enables the system to combine the flexibility of LLMs with the precision

of rule-based methods. The outputs of these modules are then merged using an Output-combiner, which aggregates all extracted features into a unified intermediate representation. The overall pipeline of this hybrid tagging and validation process is illustrated in Figure 4.
Although LLMs can extract such information, they are also prone to hallucinations and misclassifications. To mitigate this issue, we adopted the LLM-as-a-Judge approach. In this approach, we designed a verification prompt for each category that asked the LLM to determine whether the predicted categories and their associated tokens were consistent with predefined guidelines. Any categories flagged as inconsistent were automatically disregarded.

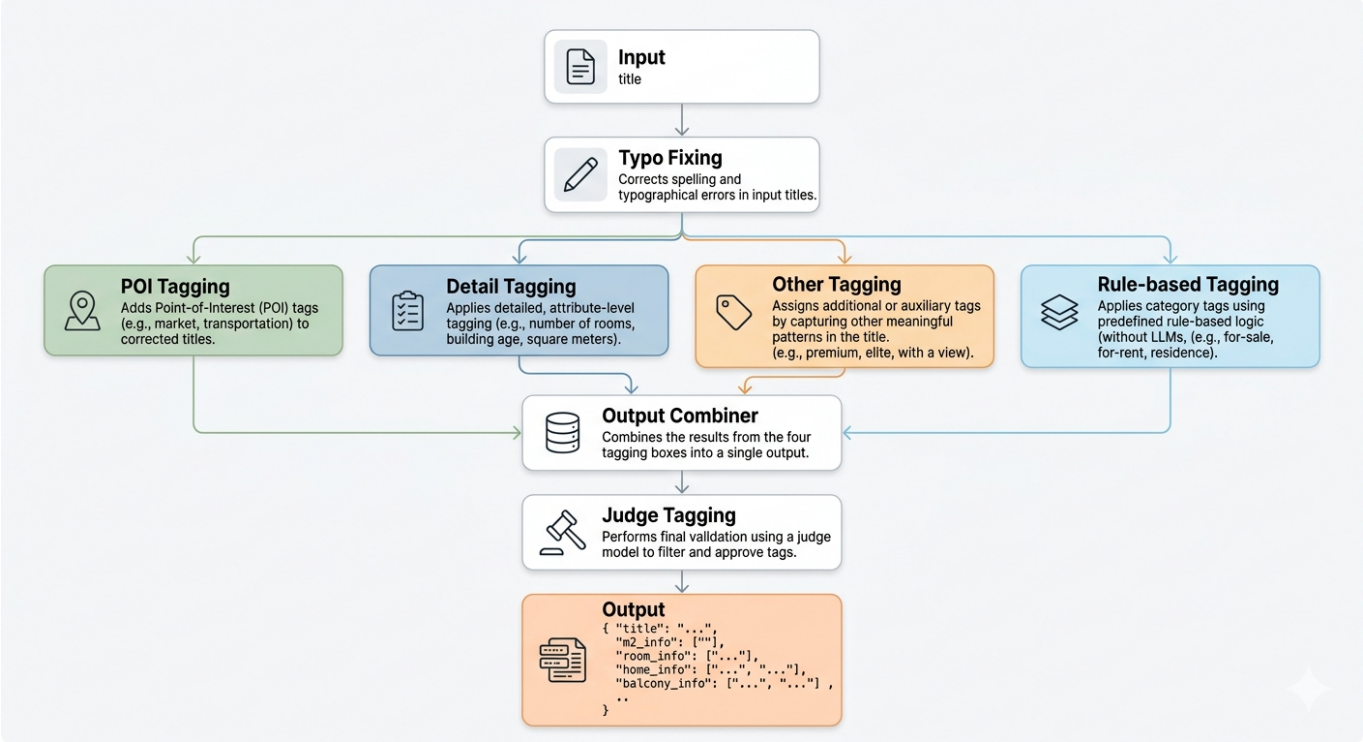

Figure 4. Tagging pipeline for real estate listing titles

4. Experiments and Results

This section presents the experimental setup and evaluation results. The experiments are designed to benchmark our proposed LLM-based extraction framework against both less restrictive LLM prompting techniques (i.e., without LLM-as-a-Judge or ReAct) and embedding-based baselines. Baselines. We compared our framework with two baseline models: (*i*) a Multilingual BERT embedding model applied to listing titles, and (*ii*) a CLIP embedding model applied to listing cover images.
For the title-based embedding baseline, the following tests were conducted to predict the predefined categories within our structured schema:

- Bert Embeddings fed into a Logistic Regression Model (B-LR)
- Bert Embeddings dimensionally reduced via PCA and fed into a Logistic Regression Model (PCA-B-LR)
- Bert Embeddings fed into an SVM Model (B-SVM)
- Bert Embeddings dimensionally reduced via PCA and fed into an SVM Model (PCA-B-SVM)
- Bert Embeddings fed into an MLP Model (B-MLP)
- Bert Embeddings dimensionally reduced via PCA and fed into an MLP Model (PCA-B-MLP)

For the image-based embedding baseline, we tested the following with CLIP embeddings:

- CLIP Embeddings fed into a Logistic Regression Model (C-LR)
- CLIP Embeddings dimensionally reduced via PCA and fed into a Logistic Regression Model (PCA-C-LR)
- CLIP Embeddings fed into an SVM Model (C-SVM)
- CLIP Embeddings dimensionally reduced via PCA and fed into an SVM Model (PCA-C-SVM)
- CLIP Embeddings fed into an MLP Model (C-MLP)
- CLIP Embeddings dimensionally reduced via PCA and fed into an MLP Model (PCA-C-MLP)

MLP models consisted of a single hidden layer with 64 neurons and ReLU activation, trained using the Adam optimizer with a learning rate of 0.001 and L2 regularization ($\alpha = 1e-4$). Early stopping was enabled to prevent overfitting.
Additionally, we compared our framework against multiple LLM-based configurations: (*i*) For titles, an LLM model with few-shot learning (FSLLM) was compared with the same model with the use of a judge prompt (FSLLM-J) , and (*ii*) For images, a few-shot learning LLM (FSLLM) and a few-shot learning LLM with ReAct (FSLLM-R) prompting.
Performance Evaluation. Evaluation of our primary LLM-based approach was performed on the manually labeled dataset. Performance was measured using precision [15], recall [16], and F1 scores as the primary metrics.

To guide our analysis, we formulated two RQs: (**RQ1**) How does our framework perform, on average, compared to the baselines? (**RQ2.**) How does its performance vary across individual predefined categories?

Experimental Setup. (*i*) To address RQ1, we conducted extraction and classification experiments using structured prompts to the LLM and evaluated its outputs to compute an overall performance score. (*ii*) For RQ2, we considered that some categories are underrepresented in the dataset, which makes an overall average insufficient to reflect true performance. Therefore, we calculated category-specific metrics to obtain a more granular evaluation.

For the embedding-based baseline experiments, we used 5-fold cross-validation. For each fold, the labeled data were split into training and test sets using a random split, with 80% of the samples allocated to training and 20% reserved for evaluation. Separate models were trained for each predefined category to predict its presence or absence, and evaluated using the same precision, recall, and F1 metrics. This procedure was repeated for each category in our structured format, and the final results were obtained by averaging performance across all folds.

For titles, we used a model that generates 768-dimensional vector embeddings for each listing title. To handle the multi-label nature of the categories, the previously mentioned baseline models were trained for each title category. Performance was assessed both per category and overall. In an additional experiment, BERT embeddings were reduced to 96 dimensions using PCA, retaining 95% of the cumulative variance, as illustrated in Figure 5. The target dimensionality was selected based on cumulative variance retention to ensure the preservation of information in the compressed features. The reduced embeddings were then used to retrain the same classifiers.

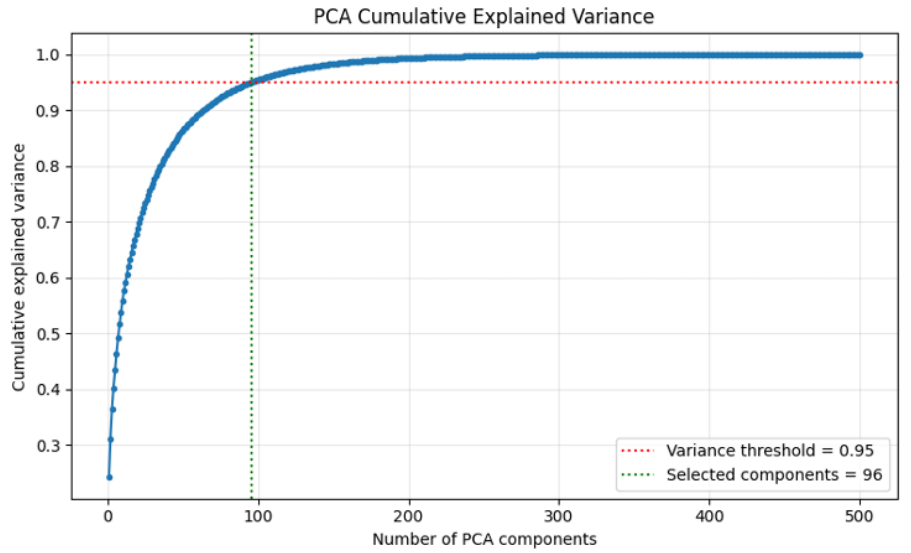

Figure 5. Title embedding dimensionality reduction

For images, CLIP embeddings of 512 dimensions were generated for each cover image. As with the title analysis, separate logistic regression models were trained for each image-specific category. In a parallel experiment, PCA was applied to reduce the embeddings to 81 dimensions, retaining 95% of the cumulative variance, as illustrated in Figure 6.

For evaluation, the manually labeled dataset was used. S Two experimental configurations were tested for each modality: FSLLM, which uses only the few-shot prompting, and FSLLM-J, which combines few-shot prompting with the LLM-as-a-Judge approach. For image analysis, imageA relies solely on few-shot prompting, whereas imageB integrates it with a ReAct-style strategy.

The precision and recall results for title experiments are summarized in Table 1, organized by predefined category. To ensure statistical significance, only categories with at least 10 samples were included, and the best-performing results for each category are highlighted in bold. Throughout the experiments we conducted, we observed that applying dimensionality reduction to the title embeddings was very rarely beneficial. In most cases, the application of PCA either degraded classification performance massively or resulted in nearly no change.

Amongst our six baseline approaches, B-LR is the strongest and most consistent, whilst also the most conventional and simple. It achieves high F1 scores on multiple important features such as ‘m2 info’, ‘room info’ and ‘location info’. The stability offered by logistic regression suggests that this approach is our strongest baseline. On the other hand, B-SVM has similar metrics on some features but fails to provide the stability offered by B-LR. The disappointing result in features such as ‘room info’ results in B-LR being superior. Similarly, B-MLP performs similarly to B-SVM as it also fails to generalize and underperforms in features such as ‘home info’ and ‘floor info’. However, applying PCA seems to only be beneficial when training and testing an MLP model, since there are improvements.

From purely a performance metric evaluation, our superior approach is FSLLM as it consistently offers the highest average performance across multiple features. Subsequently, the lack of a judge LLM results in a decrease in the cost of running the LLM. When the main goal is to maximize the performance metrics, FSLLM is clearly the best model. However, the introduction of the judge LLM reduces the likelihood of hallucinations and irrelevant responses in our pipeline. Whilst FSLLM-J underperforms slightly compared to FSLLM, the trade-off seems acceptable.

The outcomes of the image-based experiments are summarized in Table 2, following a consistent structure to facilitate direct comparison across models, feature types, and evaluation metrics. In contrast to our initial assumptions, applying PCA to image embeddings does not yield performance improvements across models. In most cases, PCA generally leads to performance degradation, particularly for neural architectures such as MLP and SVM. This suggests that dimensionality reduction may remove critical discriminative information for downstream classification tasks.

Across all evaluated approaches, the FSLLM-based models clearly demonstrate superior and more stable performance, significantly outperforming machine learning baselines (C-LR, C-SVM, and C-MLP) in most feature–metric combinations. When considering both the First Prediction (e.g.,

Table 1. Precision (Prec.), Recall (Rec.), and F1 scores for different models and features for listing titles

		M2 info	*Room info*	*Home info*	*Balcony info*	*Floor info*	*Furniture info*	*Home age*	*Location info*	*POI activity*	*POI transport*
	Prec.	0.85	**0.97**	0.32	0.70	0.30	0.50	0.64	0.93	0.28	0.62
B-LR	Rec.	**1.00**	0.95	0.41	0.90	0.43	0.67	0.74	0.88	0.42	0.80
	F1.	0.92	0.96	0.36	0.78	0.34	0.57	0.68	0.90	0.34	0.70
	Prec.	0.78	0.96	0.30	0.57	0.19	0.39	0.56	0.93	0.26	0.53
PCA-B-LR	Rec.	0.91	0.95	0.47	0.78	0.38	0.62	0.75	0.82	0.45	0.71
	F1.	0.84	0.96	0.37	0.65	0.26	0.48	0.64	0.87	0.33	0.60
	Prec.	0.89	0.27	0.31	0.71	0.35	0.47	0.62	0.91	0.27	0.67
B-SVM	Rec.	0.97	0.23	0.40	0.73	0.43	0.61	0.68	0.89	0.34	0.76
	F1.	0.93	0.25	0.35	0.72	0.38	0.53	0.64	0.90	0.30	0.71
	Prec.	0.80	0.95	0.27	0.50	0.22	0.41	0.49	0.92	0.26	0.55
PCA-B-SVM	Rec.	0.81	0.93	0.46	0.67	0.41	0.54	0.60	0.86	0.40	0.65
	F1.	0.80	0.94	0.34	0.57	0.28	0.46	0.53	0.89	0.31	0.58
	Prec.	0.81	0.95	0.17	0.72	0.23	0.58	0.59	0.90	0.43	0.82
B-MLP	Rec.	0.95	0.93	0.09	0.45	0.08	0.35	0.44	**0.97**	0.21	0.59
	F1.	0.87	0.94	0.11	0.55	0.12	0.43	0.51	0.93	0.26	0.68
	Prec.	0.64	0.80	0.37	0.81	0.11	0.73	0.52	0.90	0.24	0.85
PCA-B-MLP	Rec.	0.29	0.90	0.17	0.40	0.07	0.36	0.34	0.94	0.17	0.42
	F1.	0.40	0.85	0.23	0.51	0.08	0.48	0.38	0.92	0.17	0.56
	Prec.	0.98	0.96	0.52	0.92	**0.79**	0.73	0.80	**0.95**	0.53	0.90
FSLLM	Rec.	0.99	**0.99**	**0.69**	**1.00**	**0.81**	**0.99**	0.98	0.94	**0.60**	**0.92**
	F1.	0.98	**0.98**	0.59	0.96	**0.80**	0.84	0.88	**0.94**	**0.56**	**0.91**
	Prec.	**0.99**	**0.97**	**0.55**	**0.94**	0.69	**0.85**	**0.84**	0.94	**0.57**	**0.92**
FSLLM-J	Rec.	0.99	**0.99**	0.65	**1.00**	0.66	0.96	**0.99**	0.92	0.54	0.86
	F1.	**0.99**	**0.98**	**0.60**	**0.97**	0.67	**0.90**	**0.91**	0.93	0.55	0.88

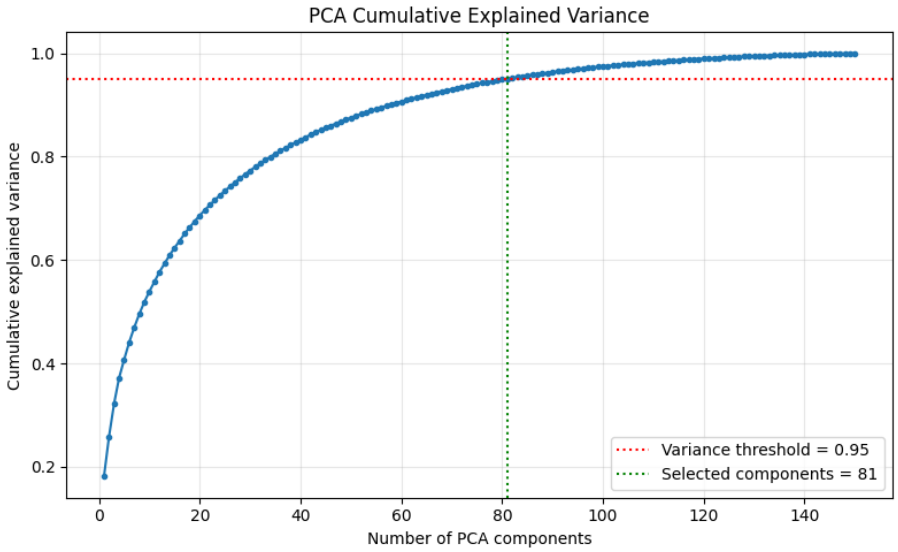

Figure 6. Image embedding dimensionality reduction

interior, exterior) and the Second Prediction (e.g., kitchen, living room) tasks together, FSLLM-based models consistently achieve the highest results. In particular, FSLLM-R reaches perfect macro scores (1.00 precision, recall, and F1) for First Prediction, indicating near-complete reliability in identifying the primary class of an image. This is a critical capability, as correctly identifying the fundamental category of a real estate image directly impacts downstream tasks. For the Second Prediction task, which is inherently more complex, FSLLM and FSLLM-R still outperform all baseline models by a considerable margin, achieving notably higher macro recall (up to 0.79) and macro F1 score (up to 0.74). These results show that FSLLM-based approaches achieve better generalization and class separation than traditional models.

A similar trend is observed in the Scenery and Logo detection tasks. While C-SVM achieves relatively high precision in isolated cases of scenery detection, its overall performance remains inconsistent, particularly in recall and F1 score. In the Logo detection task, the baseline C-MLP model demonstrates relatively strong macro-level performance; however, their micro-level consistency is limited. FSLLM-based models provide more balanced, robust performance across tasks. This balance is crucial in real-world applications, where missing relevant visual elements can be as problematic as incorrect detections.

When comparing FSLLM and FSLLM-R, both models exhibit very strong and competitive performance. FSLLM-R consistently provides

slight but meaningful improvements, particularly in First Prediction and micro-level metrics. While FSLLM already achieves high accuracy and balanced performance, FSLLM-R demonstrates enhanced reliability and consistency, suggesting that our prompting strategy contributes positively to overall predictive quality.

Overall, the experimental findings clearly demonstrate that FSLLM-based approaches outperform traditional baselines across most of the evaluated tasks, while also providing more consistent and reliable performance. These results strongly support the effectiveness of LLM-driven feature extraction and reasoning pipelines for complex textual and visual classification problems in the real estate domain. Furthermore, in some categories, near-perfect recall values can be explained by the relatively well-defined and less ambiguous nature of the corresponding features. Such outcomes are consistent with the intrinsic characteristics of these tasks and are more likely to reflect clearly separable class boundaries rather than indications of overfitting.

Table 2. Macro-Micro Precision (Prec.), Recall (Rec.), and F1 results for different models and features for listing images

			First prediction	*Second prediction*	*Scenery*	*Logo*
C-LR	Macro	Prec.	0.72	0.58	0.74	0.69
		Rec.	0.73	0.62	0.74	0.68
		F1.	0.73	0.60	0.74	0.68
	Micro	Prec.	0.96	0.76	0.81	0.68
		Rec.	0.96	0.76	0.81	0.68
		F1.	0.96	0.76	0.81	0.68
PCA-C-LR	Macro	Prec.	0.71	0.47	0.69	0.64
		Rec.	0.72	0.55	0.68	0.64
		F1.	0.71	0.49	0.68	0.63
	Micro	Prec.	0.93	0.65	0.78	0.64
		Rec.	0.93	0.65	0.78	0.64
		F1.	0.93	0.65	0.78	0.64
C-SVM	Macro	Prec.	0.71	0.58	**0.86**	0.67
		Rec.	0.72	0.63	0.76	0.67
		F1.	0.72	0.60	**0.79**	0.67
	Micro	Prec.	0.95	0.76	**0.87**	0.67
		Rec.	0.95	0.76	**0.87**	0.67
		F1.	0.95	0.76	**0.87**	0.67
PCA-C-SVM	Macro	Prec.	0.63	0.14	0.84	0.70
		Rec.	0.63	0.25	0.57	0.69
		F1.	0.62	0.16	0.56	0.69
	Micro	Prec.	0.82	0.29	0.79	0.70
		Rec.	0.82	0.29	0.79	0.70
		F1.	0.82	0.29	0.79	0.70
C-MLP	Macro	Prec.	0.58	0.49	0.68	**0.73**
		Rec.	0.57	0.55	0.68	**0.73**
		F1.	0.58	0.51	0.68	**0.73**
	Micro	Prec.	0.90	0.69	0.77	0.73
		Rec.	0.90	0.69	0.77	0.73
		F1.	0.90	0.69	0.77	0.73
PCA-C-MLP	Macro	Prec.	0.27	0.04	0.50	0.46
		Rec.	0.21	0.04	0.50	0.46
		F1.	0.23	0.04	0.49	0.46
	Micro	Prec.	0.43	0.07	0.57	0.47
		Rec.	0.43	0.07	0.57	0.47
		F1.	0.43	0.07	0.57	0.47
FSLLM	Macro	Prec.	0.87	0.71	0.70	0.57
		Rec.	0.87	0.72	0.78	0.55
		F1.	0.83	0.68	0.68	0.56
	Micro	Prec.	0.98	0.87	0.71	0.82
		Rec.	0.98	0.87	0.71	0.82
		F1.	0.98	0.87	0.71	0.82
FSLLM-R	Macro	Prec.	**1.00**	**0.77**	0.72	0.58
		Rec.	**1.00**	**0.79**	**0.80**	0.57
		F1.	**1.00**	**0.74**	0.71	0.57
	Micro	Prec.	**0.99**	**0.89**	0.73	**0.85**
		Rec.	**0.99**	**0.89**	0.73	**0.85**
		F1.	**0.99**	**0.89**	0.73	**0.85**

5. Discussion

The experimental results demonstrate that the proposed framework is effective in transforming unstructured listing data into structured representations across both textual and visual modalities. However, compared to title experiments, both precision and recall scores for images remained lower across some categories, such as Scenery. Despite this, our framework still achieved the highest overall average scores, indicating promising performance under the evaluated conditions. It is also important to highlight that, compared to listing titles, identifying the presence of specific categories in images represents a considerably more challenging task. Although the number of annotated samples was statistically sufficient, increasing the size of manually labeled data could further enhance the performance of our framework.

When analyzing the results for both listing titles and images, our approach consistently demonstrates superior performance in transforming unstructured data into structured representations. Performance differences between categories are expected, as each reflects varying degrees of subjectivity and user interest. For instance, Home Info in titles and Scenery in images exhibit unique behavioral patterns due to their contextual and perceptual diversity. Nevertheless, the combined use of both frameworks, tailored to category-specific goals, yields the most effective results for real-world practical applications.

Despite these promising results, several limitations should be considered. First, the size of the manually annotated dataset is relatively limited, particularly for the image-based experiments, which may affect the generalizability of the findings. Another point to note is that some of the predefined categories (e.g., luxury perception or scenery) involve subjective interpretation, which may introduce annotation bias. Additionally, since the dataset is collected from a single real estate platform within a specific geographic region, this may limit the external validity of the results. Lastly, the framework relies on a specific LLM (GPT-4o-mini), and performance may vary when different models or prompting strategies are applied.

6. Conclusion

This study presents a novel framework for transforming unstructured real-estate data into structured representations. Our framework uses state-of-the-art LLM technologies and supports multiple prompting strategies (Title prompts with few-shot, and an optional judge prompt; image prompts using few-shot and an optional ReACT strategy). We benchmarked our approach against strong embedding-based baseline models (BERT for titles and CLIP for images) and applied PCA to assess the effect of dimensionality reduction. We found that our approach with few-shot prompting for both the images and titles generally outperformed all baselines on a category basis.

Although the annotated dataset was relatively small, it provided sufficient evidence for the framework's scalability and generalization potential. Future work will involve expanding the labeled dataset and testing the framework with LLMs beyond the OpenAI ecosystem. In addition, future studies may explore cross-platform generalization by applying the framework to listings from different real estate platforms and geographic regions. Investigating more advanced multimodal architectures and hybrid approaches that combine LLM reasoning with fine-tuned vision models could further improve performance. Another promising direction is incorporating user interaction data to better understand which extracted features contribute most to engagement and conversion. In addition, the structured outputs generated by this framework can support downstream analyses, such as identifying which listing attributes most strongly influence user engagement and decision-making on real-estate platforms.

Article Information Form

Authors' Contributions: Afra Arslan: Contributed to the conceptualization and methodology of the study; developed software components; conducted formal analysis, investigation, and data curation; contributed to visualization; participated in writing—review and editing; took part in project administration; reviewed and approved the final manuscript; and accepts responsibility for the integrity of the work.

Tan Doruk Yetki: Contributed to the conceptualization and methodology; participated in software development, formal analysis, investigation, and data curation; contributed to visualization; participated in writing—original draft preparation; took part in project administration; reviewed and approved the final manuscript; and accepts responsibility for the integrity of the work.

Arda Yücel: Contributed to the conceptualization and methodology; supervised the study; participated in project administration; reviewed and approved the final manuscript; and accepts responsibility for the integrity of the work.

Hacer Turgut: Contributed to the conceptualization and methodology; participated in writing—review and editing; supervised the study; secured funding; reviewed and approved the final manuscript; and accepts responsibility for the integrity of the work.

Ömür Bali: Contributed to the conceptualization and methodology; supervised the study; reviewed and approved the final manuscript; and accepts responsibility for the integrity of the work.

Gülfem Işıklar Alptekin: Contributed to the conceptualization and methodology; participated in writing—original draft preparation and writing—review and editing; supervised the study; reviewed and approved the final manuscript; and accepts responsibility for the integrity of the work.

Günce Keziban Orman: Contributed to the conceptualization and methodology; participated in writing—original draft preparation and writing—review and editing; supervised the study; reviewed and approved the final manuscript; and accepts responsibility for the integrity of the work.

Acknowledgments: This work was supported by TUBITAK grant number 124E135.

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

Artificial Intelligence Statement: Artificial intelligence tools were used solely for grammar checking, language refinement, and improving the clarity of some sentences. The authors maintained full control over the structure, arguments, and scientific contributions of the manuscript.

Plagiarism Statement: This article has been scanned by iThenticate™.

References

[1] T. Brown et al., *Language models are few-shot learners*, Adv. Neural Inf. Process. Syst., **33** (2020), 1877–1901.
[2] J. Dagdelen et al., *Structured information extraction from scientific text with large language models*, Nat. Commun., **15**(1) (2024), 1418.

[3] S. Desai and G. Durrett, *Calibration of pre-trained transformers*, in Proc. 2020 Conf. Empirical Methods Nat. Lang. Process. (EMNLP), Association for Computational Linguistics, (2020), 295–302.
[4] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, *Language-agnostic BERT sentence embedding*, arXiv preprint arXiv:2007.01852, 2020. https://arxiv.org/abs/2007.01852
[5] Y. Guo, C. Wang, S. X. Yu, F. McKenna, and K. H. Law, *AdaLN: a vision transformer for multidomain learning and predisaster building information extraction from images*, J. Comput. Civ. Eng., **36**(5) (2022), 04022024.
[6] T. Gupta, M. Zaki, N. M. A. Krishnan, and Mausam, *MatSciBERT: a materials domain language model for text mining and information extraction*, Comput. Mater. Sci., **8**(1) (2022), 102.
[7] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, *Vision GNN: an image is worth graph of nodes*, Adv. Neural Inf. Process. Syst., **35** (2022), 8291–8303.
[8] ilab-core, *LLM-based real estate information extraction*, GitHub repository. https://github.com/ilab-core/llm-based-real-estate-information-extraction
[9] M. Kvet, M. Potočár, and S. Tatarka, *Real estate attribute value extraction using large language models*, IEEE Access, **13** (2025).
[10] F. Li, M. Zhang, G. Fu, and D. Ji, *A neural joint model for entity and relation extraction from biomedical text*, BMC Bioinformatics, **18** (2017), 1–11.
[11] W.-C. Liao, D. Y. Koh, G. Yin, and S. Zheng, *Feng Shui meets AI: Measuring its economic value in commercial real estate*, MIT Center for Real Estate Research Paper No. 25/08, Nov. 2025.
[12] S. Marukatat, *Tutorial on PCA and approximate PCA and approximate kernel PCA*, Artif. Intell. Rev., **56**(6) (2023), 5445–5477.
[13] OpenAI, *OpenAI*, online resource. https://openai.com
[14] A. Radford et al., *Learning transferable visual models from natural language supervision*, arXiv preprint arXiv:2103.00020, 2021. https://arxiv.org/abs/2103.00020
[15] G. Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill Book Company, New York, 1968.
[16] K. Spärck Jones, *A theoretical comparison of recall and precision*, J. Doc., **28**(1) (1972), 11–21.
[17] A. Vijayan, *A prompt engineering approach for structured data extraction from unstructured text using conversational LLMs*, in Proc. 2023 6th Int. Conf. Algorithms, Comput. Artif. Intell. (ACAI '23), Association for Computing Machinery, (2024), 183–189.
[18] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, *GLUE: a multi-task benchmark and analysis platform for natural language understanding*, in Proc. 7th Int. Conf. Learn. Represent. (ICLR), 2019.
[19] K. Wu, S. Yuan, C. Shen, L. Xu, and M. Chen, *AutoMV: An autonomous agent framework for real estate marketing video generation*, Proc. AAAI Conf. Artif. Intell., **39**(28) (2025), 29715–29717.
[20] S. Yao et al., *ReAct: Synergizing reasoning and acting in language models*, arXiv preprint arXiv:2210.03629, 2023. https://arxiv.org/abs/2210.03629
[21] L. Zheng et al., *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*, arXiv preprint arXiv:2306.05685, 2023. https://arxiv.org/abs/2306.05685