Using Large Language Models in Real Estate Transactions: A Few-shot Learning Approach

Shiqi Yang [1], Yu Zhao [2], Haoxiang Gao [3]
sy3506@nyu.edu, yzqr.zhao@rotman.utoronto.ca haoxiang.gao@motional.com

Abstract

Large language models (LLMs) have demonstrated significant potential in transforming real estate transactions through advanced natural language processing capabilities. This paper proposes a method leveraging LLMs to extract information from real estate contracts and facilitate interactive querying by users. By emphasizing the use of few-shot learning, LLMs can effectively analyze contract documents and respond to specific questions from users regarding contract details. This paper explores the development of this method, its implementation in real estate workflows, and evaluates its effectiveness in improving efficiency and accuracy within the industry.

1 Introduction

Real estate transactions involve extensive documentation, including contracts that outline crucial terms and conditions. Extracting and interpreting information from these contracts is a labor-intensive task for real estate professionals. With the emergence of large language models (LLMs) such as GPT-3.5, there is an opportunity to automate and streamline this process through few-shot learning. This paper proposes a method using LLMs to extract information from real estate contracts and enable interactive querying by users, ultimately enhancing the efficiency and accuracy of real estate transactions.

2 Background and Related Work

Large language models (LLMs) possess the ability to comprehend and produce human-like text by leveraging extensive pre-existing data. These models operate using large neural network architectures, allowing users to input specific queries or instructions to prompt desired responses. LLMs have demonstrated successful applications across various domains, showcasing their versatility and effectiveness in tasks ranging from natural language understanding to text generation and information retrieval. [Yang *et al.*, 2023] and [Gao *et al.*, 2024] explore the applications of machine learning and large neural networks in autonomous driving. [Wang *et al.*, 2020] provides an overview of few-shot learning. [Xian *et al.*, 2017] provides an in-depth analysis of zero-shot learning. [Yu *et al.*, 2024c] provides an assessment of LLM models in medical question-answering tasks.

In addition to question-answer, classification problems are also important in the realm of real estate transactions. Classification algorithms can categorize text segments, such as identifying clauses related to liabilities, indemnities, or property descriptions. This helps organize a contract's content into structured forms that are easier to analyze and understand. [Yu *et al.*, 2024b] applies the transformer model to understand credit credit transactions and detect fraud. Traditionally, machine learning methods such as support vector machine or variants of the gradient boost decision tree models have achieved good results in such classification problems [Liu *et al.*, 2024] [Zheng *et al.*, 2024] [Li *et al.*, 2024]. LLM-based model provides embedding that can also be used to solve classification problems in natural language processing.

[Zhao and Liu, 2024] explains deep learning and LLM-based methods for information retrieval. Retrieval technology is highly relevant in this context because it can efficiently sift through vast amounts of contract data to find specific information required by real estate professionals. By quickly locating relevant sections within diverse contract documents, retrieval systems reduce the time spent manually searching for information, thereby accelerating the review process and improving productivity. This capability is particularly valuable in real estate transactions where timely access to accurate contract details can significantly impact decision-making and transaction outcomes.

In the context of law, [Wang *et al.*, 2023] explores the applications of deep learning in law. [Cui *et al.*, 2023] proposed a legal large language model named ChatLaw. [Trozze *et al.*, 2024] explores the applications of large language models in cryptocurrency securities legal cases. [Zhao and Gao, 2024] proposed that LLMs can be used to query real estate transaction specifics. [Zhao *et al.*, 2024] applied LLMs to analyze contracts.

3 Few-shot Learning and Fine Tuning

In the realm of utilizing large language models (LLMs) for real estate transactions, few-shot learning and fine-tuning are two key techniques that offer distinct advantages and challenges. Few-shot learning involves providing specific instructions or queries to LLMs to guide their responses, enabling targeted information retrieval and task completion. This method is advantageous because it allows users to interact with LLMs in a structured manner, ensuring that the generated outputs align with the desired objectives. Few-shot learning also facilitates the customization of LLM behavior for specific applications, making it a versatile approach in real estate settings.

However, few-shot learning has certain limitations. The effectiveness of few-shot learning heavily relies on the quality and specificity of the prompts provided. Ambiguous or poorly formulated prompts may lead to inaccurate or irrelevant responses from LLMs, diminishing the overall utility of the interaction. Additionally, few-shot learning may require a degree of manual intervention to craft appropriate queries, which can be time-consuming and may necessitate domain expertise to generate optimal prompts tailored to real estate tasks.

On the other hand, fine-tuning involves adapting pre-trained LLMs to specific tasks or domains by further training them on specialized datasets. This technique enhances the model's performance and relevance to particular applications, including real estate transactions. Fine-tuning allows LLMs to learn domain-specific nuances and vocabulary, resulting in more accurate and contextually relevant outputs for tasks like contract analysis and information extraction.

Despite its benefits, fine-tuning can be resource-intensive and requires annotated datasets for training, which may pose challenges in terms of data acquisition and computational resources. Moreover, fine-tuned models may suffer from overfitting if the training data is not representative of the broader real estate context, leading to limited generalization capabilities.

Method	Pros	Cons
Few-shot learning	• Simple and easy to implement. • Does not require retraining the model. • Provides control over the generated output.	• Limited to pre-trained model capabilities. • Relies on effective prompt design. • May require trial and error.
Fine-Tuning	• Allows model customization. • Improves performance on specific data. • Enables adaptation to new tasks.	• Requires access to large datasets. • Time-consuming and expensive. • May lead to overfitting.

Table 1: Summary of Pros and Cons: Prompting vs. Fine-Tuning

4 Retrieval-Augmented Generation

The integration of Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) offers a promising avenue to enhance the effectiveness of information retrieval. RAG essentially combines the robust language understanding capabilities of LLMs with a powerful retrieval mechanism. This approach enables the model to fetch relevant documents or data from a vast repository based on the input query before generating a response. The retrieval component acts as an initial filter that provides contextually relevant information which the generative component of the LLM uses to construct accurate and informative outputs. This methodology is particularly useful when dealing with complex real estate contracts, where the ability to quickly access specific contract clauses and related legal precedents can significantly streamline decision-making processes. Next, we outline the high-level design using RAG and defer implementation and empirical validation to future studies.

At a high level, applying RAG to real estate transactions involves three main steps. First, the LLM receives a query related to a specific aspect of a real estate contract. The query could pertain to terms of sale, disclosure requirements, or zoning regulations, among other topics. Next, the retrieval component of RAG springs into action, scouring through an extensive database of real estate documents and contracts to find relevant text snippets that match the query. Using a similar approach in [Zhao and Liu, 2024], this retrieval is based on semantic similarity, ensuring that the most pertinent information is selected to aid the response generation.

Once the relevant data is retrieved, it is passed along to the generative component of the LLM. Here, the model synthesizes the retrieved information, integrating it seamlessly with its pre-trained knowledge base to generate a coherent and contextually relevant response. This response could be a direct answer to a query, a summary of relevant contract provisions, or even advice on next steps in a transaction process. The power of RAG lies in its ability to bridge the gap between vast data repositories and the need for precise, context-aware outputs in real time.

4.1 RAG and Few-shot Learning

Few-shot learning, when viewed through the lens of Retrieval-Augmented Generation (RAG), can be understood as a process where the model leverages a small number of examples (the "few shots") retrieved from a large dataset to inform and enhance its response generation. In RAG, the retrieval component plays a crucial role by acting as an initial step that fetches contextually relevant examples or documents from a vast repository. These retrieved examples serve as the few-shot prompts that provide the model with concrete instances related to the query.

From a RAG perspective, few-shot learning involves the model using these retrieved examples to better understand the query context and generate a more accurate and relevant response. The retrieval system identifies the most pertinent examples based on semantic similarity or relevance to the input query, which are then passed to the generative component of the model. The generative model uses these examples as guiding points, integrating them with its pre-trained knowledge to produce a response that aligns closely with the information and patterns presented in the few-shot examples.

This approach is particularly powerful because it allows the model to dynamically access and utilize external knowledge, even when the pre-training data may not have fully covered the specific context of the query. By retrieving and incorporating relevant few-shot examples, the RAG system effectively augments the model's generative capabilities, enabling it to handle complex or specialized queries with greater accuracy and contextual awareness.

4.2 Document Indices

Indices are crucial for real estate transactions because they allow for the efficient and accurate retrieval of information that is often buried within a vast array of complex documents, such as contracts, zoning regulations, disclosures, and legal precedents. In real estate, where the stakes are high and legal requirements can vary significantly by state, county, or even municipality, indices help categorize and organize data according to specific criteria such as jurisdiction, document type, or date. This ensures that relevant information can be quickly accessed and is contextually appropriate for the specific transaction at hand. For example, when dealing with a property sale, having indices that filter documents by state-specific regulations or zoning laws ensures that all legal obligations are met, reducing the risk of errors or oversights. By streamlining the retrieval process, indices not only save time but also provide assurance that decisions are based on the most accurate and relevant information available, thereby enhancing the overall efficiency and legal compliance of real estate transactions.

1. **Document Type Index** A Document Type Index is essential in real estate transactions because it categorizes documents according to their specific types, such as purchase agreements, lease contracts, zoning ordinances, disclosure forms, and mortgage documents. Real estate transactions involve a wide range of documents, each serving a unique legal or financial purpose. By indexing these documents by type, the retrieval system can quickly locate the most relevant documents based on the nature of the query. For instance, if an agent needs to review the terms of a mortgage, the system can prioritize pulling up mortgage agreements and related financial documents, avoiding unnecessary retrieval of irrelevant documents like zoning regulations or property disclosures. This targeted approach ensures that professionals can access the precise documentation they need, facilitating smoother and more efficient transactions.

2. **Jurisdiction Index** The Jurisdiction or Municipality Index is critical in real estate transactions, given the significant variations in laws and regulations across different regions, even within the same state. Real estate transactions are heavily influenced by local ordinances, zoning laws, and property tax regulations, which can differ widely between cities, counties, and municipalities. By indexing documents according to jurisdiction or municipality, the retrieval system ensures that the information retrieved is precisely tailored to the location-specific legal framework. This is particularly important for developers, real estate attorneys, and agents who need to navigate local zoning laws or building codes. For instance, a query about permissible land use in a particular city will yield results that reflect the city's specific zoning regulations, rather than irrelevant information from other jurisdictions, thereby supporting accurate and legally compliant real estate practices.

3. **Subject Matter or Topic Index** A Subject Matter or Topic Index is indispensable in real estate, where professionals often need to find documents related to specific aspects of a transaction, such as environmental assessments, tax implications, property rights, or financing terms. Real estate involves numerous specialized topics, each requiring detailed legal and financial documentation. By categorizing documents based on their subject matter, this index enables users to quickly locate the exact information they need. For example, an investor researching environmental risks associated with a property would benefit from a system that can efficiently retrieve all relevant environmental reports, assessments, and related legal documents. This level of specificity ensures that professionals can address complex queries with confidence, backed by the most relevant and detailed information available.

4. **Contractual Clause or Section Index** A Contractual Clause or Section Index is particularly important in real estate, where contracts are often lengthy and complex, containing numerous clauses and provisions that are critical to the transaction. This index allows users to pinpoint and retrieve specific clauses within a contract, such as contingency provisions, financing terms, or force majeure clauses, which are often essential to negotiations and legal compliance. For example, if a buyer's agent needs to review the financing contingency in a purchase agreement, the system can directly retrieve that section of the contract, providing immediate access to the relevant details. This granular level of indexing streamlines the review process, ensuring that important contract terms are easily accessible and reducing the likelihood of overlooking critical legal elements.

In our few-shot learning experiment, we manually created examples and applied them to evaluate the model's performance. These handcrafted examples form the basis for our initial testing, allowing us to assess how well the model can handle specific queries with limited data.

Once the model is deployed, real examples can be collected from actual user interactions in production. Indices can then created by systematically organizing the data based on key attributes relevant to real estate transactions. This involves categorizing documents by type, such as contracts, zoning regulations, and disclosures, as well as by other critical factors like jurisdiction, date, and subject matter. Each document will be indexed according to these criteria, ensuring that the retrieval system can efficiently locate and access the most relevant information. Traditional machine learning models can also be used to produce such indices. [Yu *et al.*, 2024a] provides an example where traditional supervised machine learning models are adapted to handle financial tasks. The process may also involve leveraging metadata and contextual information to accurately tag and classify each document, enabling precise and contextually appropriate retrieval in response to user queries. By structuring the data in this way, the indices provide a robust framework for quickly finding the necessary documents within a vast repository, enhancing the overall effectiveness of the information retrieval system.

As these examples accumulate over time, the RAG dataset will naturally expand, incorporating a wider variety of real-world questions and answers. This continuous growth will enhance the model's accuracy and effectiveness, ensuring it becomes more robust and better suited to the evolving demands of its application.

5 Proposed Method

In this paper, we focus on experimenting with the prompting approach to leverage LLMs for real estate transactions. By exploring the use of prompts to extract information from real estate contracts and engage in interactive querying, we aim to assess the effectiveness and feasibility of this method in enhancing efficiency and accuracy within the real estate industry. Through empirical evaluation and analysis, we investigate how the prompting technique can be optimized and refined to address specific challenges and requirements in real estate transaction workflows. This experimental approach contributes to advancing the practical application of LLMs in real estate settings, shedding light on the benefits and considerations associated with leveraging prompting techniques for information retrieval and task-oriented interactions.

6 Experimentation and Evaluation Plan

To evaluate the effectiveness of using prompting techniques with large language models (LLMs) for real estate contracts analysis, we outline a structured plan with specific objectives, methodologies, and evaluation criteria.

6.1 Methodologies

The methodologies for conducting the experimentation and evaluation include:

1. **Dataset Selection**: Choose a diverse dataset of real estate contracts covering various types of transactions and containing different levels of complexity. Only sales contracts are included in this study. Other types of contracts commonly used in real estate transactions, such as lease agreements and agency agreements, are not within the scope of this analysis.
2. **Prompt Design**: Develop a set of specific prompts tailored to elicit information from real estate contracts. Prompts will include questions about property details (e.g., address, size, amenities), financial terms (e.g., purchase price, deposit amount), and transaction specifics (e.g., closing date, contingencies).
3. **Model Utilization**: Use the pre-trained GPT-3.5[OpenAI, 2023] model and implement the prompting process.
4. **Prompting Process**: The prompting process using the OpenAI API involves interacting with a pre-trained language model, such as GPT-3.5[OpenAI, 2023], by providing specific textual prompts or instructions to generate responses based on the input. For real estate contract analysis, prompts are formulated to ask specific questions or provide context related to desired information within the contracts. Once the prompt is submitted to the OpenAI API, the model processes the text and generates a response attempting to answer the query based on its learned knowledge and language understanding. The quality of the response depends on the clarity and specificity of the prompt, as well as the model's ability to comprehend and interpret the input accurately. Leveraging the OpenAI API for prompting facilitates more efficient contract analysis by automating information extraction tasks from real estate contracts.

6.2 Few Shot Learning

In this paper, we utilize the following questions and answers as exemplars in the context of few-shot learning. Few-shot learning is a machine learning paradigm aimed at training models to generalize from limited amounts of data, making it particularly valuable for scenarios where comprehensive training data is scarce or expensive to obtain. The questions presented here, along with their corresponding answers extracted from a contract, serve as foundational examples for training and evaluating models in understanding legal and contractual language with minimal supervision.

- **What is the address of the property?**
  - Property Address: 1 Main St, New York, New York
- **What is the deadline for the closing?**
  - Closing Date: January 1, 2022
- **Is it possible for the buyer or seller to delay the closing by a few days?**
  - No because time is of the essence specifically mentioned.
- **Is extrinsic evidence admissible?**
  - No. An entire agreement clause is present. Extrinsic evidence is generally not admissible.
- **Is the buyer or the seller responsible for fees associated with title transfer?**
  - Seller is responsible for the fee.
- **What damages are available if the buyer repudiates the contract?**
  - The liquidated damages clause specifies that the earnest money will be forfeited, and the seller cannot recover any additional damages.

These specific examples cover various critical aspects of real estate contracts and legal agreements. For instance, the question regarding the property address underscores the importance of extracting precise location details, essential for tasks like information retrieval or document summarization. Similarly, inquiries about deadlines and contractual constraints, such as the closing date and the irrevocable nature of time in contract fulfillment, demonstrate nuanced legal interpretations that models must learn to grasp accurately. Moreover, questions concerning admissible evidence and financial responsibilities shed light on legal principles related to evidence rules and financial obligations in property transactions. By employing these questions and answers within a few-shot learning framework, we aim to enhance models' capabilities in comprehending legal documents and real-world contracts efficiently and effectively.

6.3 Evaluation Criteria

Evaluating natural language answers generated by large language models (LLMs) in the context of real estate prompting poses unique challenges due to the complexity and variability of both the language used in real estate documents and the capabilities of LLMs. One significant difficulty is the ambiguity and context-dependency of language within real estate contracts. Real estate agreements often contain intricate legal terminology, nuanced conditions, and domain-specific jargon that require deep understanding to accurately extract and interpret. LLMs may struggle to comprehend the subtle nuances embedded in these documents, leading to potential inaccuracies or incomplete responses that are challenging to evaluate using standard metrics.

Another challenge arises from the diversity of prompts and questions posed in real estate transactions. Evaluating the effectiveness of LLM-generated answers requires assessing their relevance, correctness, and completeness across a range of query types, from specific property details to contractual deadlines and legal terms. Each prompt may require a tailored evaluation approach to capture the nuances of the information extracted, complicating the development of standardized evaluation methodologies. This variability in prompts can also impact the applicability of traditional NLP evaluation metrics, which may not fully account for the intricacies of real estate language and transactional contexts.

In this paper, we employ two distinct methods to evaluate the similarity between answers generated by a model and the expected answers. The first method utilizes the edit distance, which measures the minimum number of character edits (insertions, deletions, or substitutions) needed to transform one answer into another. We then normalize this distance by dividing it by the length of the longer answer. Mathematically, the normalized edit distance ED_norm is computed as:

$$\text{ED\_norm} = \frac{\text{edit\_distance}(A_{\text{model}}, A_{\text{expected}})}{\max(\text{len}(A_{\text{model}}), \text{len}(A_{\text{expected}}))}$$

where A_{model} and A_{expected} are the answers generated by the model and the expected answers, respectively. The edit_distance function calculates the number of edit operations required to transform one answer into the other, and $\text{len}(A)$ denotes the length of answer A.

The second method involves computing the cosine similarity between the answer embeddings produced by the paraphrase-MiniLM-L6-v2 model[Reimers and Gurevych, 2019]. Each answer is first converted into a dense vector representation using the embedding model. Let $\mathbf{v}_{\text{model}}$ and $\mathbf{v}_{\text{expected}}$ represent the embeddings of the model's answer and the expected answer, respectively. The cosine similarity cos_sim is then calculated as:

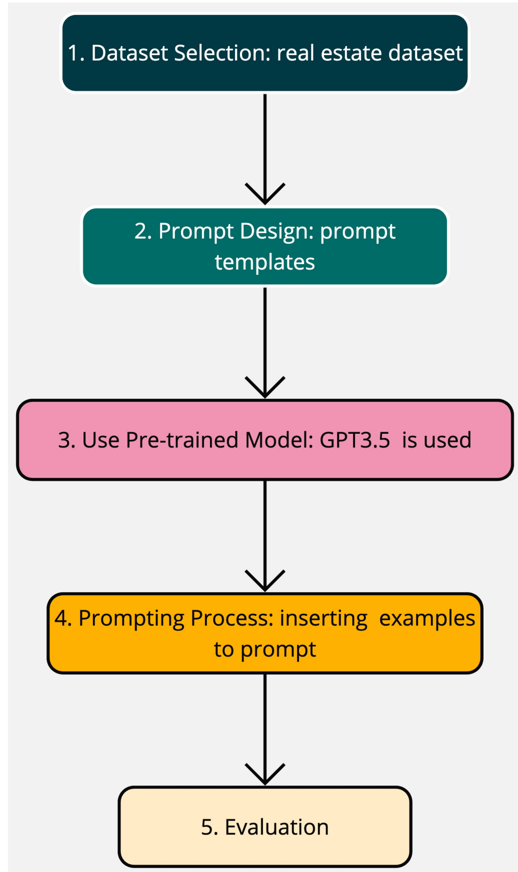

Figure 1: Data Flow

$$\text{cos\_sim} = \frac{\mathbf{v}_{\text{model}} \cdot \mathbf{v}_{\text{expected}}}{\|\mathbf{v}_{\text{model}}\| \times \|\mathbf{v}_{\text{expected}}\|}$$

where $\mathbf{v}_{\text{model}} \cdot \mathbf{v}_{\text{expected}}$ denotes the dot product of the two vectors, and $\|\mathbf{v}\|$ represents the Euclidean norm (or magnitude) of vector $\mathbf{v}$. This cosine similarity metric quantifies the directional agreement between the embeddings of the model's answer and the expected answer, providing a measure of their semantic similarity regardless of specific wording or length.

7 Results and Analysis

The results of zero-shot learning (ZSL) and few-shot learning (FSL) experiments are summarized in Table 2.

In the FSL scenario where examples were provided (Few-shot), the model achieved a semantic score of 0.999 and an accuracy rate of 98.82%. This indicates that when the model was provided with context consisting of explicit examples, it performed exceedingly well in understanding and responding to queries similar to those in the training set, demonstrating a robust ability to generalize from limited examples.

Table 2: Comparison of Zero-Shot Learning (ZSL) and Few-Shot Learning (FSL) Results

Example	Semantic Score	Diff Rate (%)	Accuracy (%)
Few-shot	0.999	1.18	98.82
Zero-shot	0.714	69.55	30.45

Conversely, in the ZSL setting where no examples were given (Zero-shot), the semantic score dropped significantly to 0.714, accompanied by a lower accuracy rate of 30.45%. This decrease in performance highlights the challenges of learning from ZSL, where the model struggles to generalize effectively without explicit examples. The difference rate of 69.55% underscores the importance of providing even a small number of examples to guide the learning process in few-shot scenarios. Overall, these results emphasize the efficacy of few-shot learning with examples compared to the difficulties encountered in zero-shot learning without explicit training instances.

8 Challenges and Future Directions

Despite the promising results, there are challenges associated with using LLMs in real estate transactions, such as data privacy concerns, legal compliance, and model limitations. Future research directions include addressing these challenges, optimizing prompting techniques, and integrating LLMs into real-world real estate workflows.

Just as few-shot learning (FSL) benefits from examples generated through RAG, zero-shot learning (ZSL) can be enhanced by using optimized prompts. The prompt-and-refine strategy used in [Song *et al.*, 2023] remain a promising direction for further exploration.

9 Conclusion

In conclusion, this paper explores the application of prompting tailored to real estate transactions, aiming to enhance large language model performance. Through experimental evaluation, the proposed method leveraging prompts demonstrates a substantial improvement over zero-shot learning in generating accurate and contextually relevant responses to real estate queries. The results highlight the effectiveness of incorporating structured prompts as input cues to guide the model's understanding and output in the context of real estate transactions, contract interpretation, and information retrieval.

References

[Cui *et al.*, 2023] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023.

[Gao *et al.*, 2024] Haoxiang Gao, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*, 2024.

[Li *et al.*, 2024] Shaojie Li, Xinqi Dong, Danqing Ma, Bo Dang, Hengyi Zang, and Yulu Gong. Utilizing the lightgbm algorithm for operator user credit assessment research. *Applied and Computational Engineering*, 75(1):36–47, July 2024.

[Liu *et al.*, 2024] Rui Liu, Xuanzhen Xu, Yuwei Shen, Armando Zhu, Chang Yu, Tianjian Chen, and Ye Zhang. Enhanced detection classification via clustering svm for various robot collaboration task. *arXiv e-prints*, pages arXiv–2405, 2024.

[OpenAI, 2023] OpenAI, 2023.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[Song *et al.*, 2023] Xingchen Song, Di Wu, Binbin Zhang, Zhendong Peng, Bo Dang, Fuping Pan, and Zhiyong Wu. ZeroPrompt: Streaming Acoustic Encoders are Zero-Shot Masked LMs. In *Proc. INTERSPEECH 2023*, pages 1648–1652, 2023.

[Trozze *et al.*, 2024] Arianna Trozze, Toby Davies, and Bennett Kleinberg. Large language models in cryptocurrency securities cases: can a gpt model meaningfully assist lawyers? *Artificial Intelligence and Law*, pages 1–47, 2024.

[Wang *et al.*, 2020] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[Wang *et al.*, 2023] Yixu Wang, Wenpin Qian, Hong Zhou, Jianfeng Chen, and Kai Tan. Exploring new frontiers of deep learning in legal practice: A case study of large language models. *International Journal of Computer Science and Information Technology*, 1(1):131–138, 2023.

[Xian *et al.*, 2017] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017.

[Yang *et al.*, 2023] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving, 2023.

[Yu *et al.*, 2024a] Chang Yu, Yixin Jin, Qianwen Xing, Ye Zhang, Shaobo Guo, and Shuchen Meng. Advanced user credit risk prediction model using lightgbm, xgboost and tabnet with smoteenn. *arXiv preprint arXiv:2408.03497*, 2024.

[Yu *et al.*, 2024b] Chang Yu, Yongshun Xu, Jin Cao, Ye Zhang, Yinxin Jin, and Mengran Zhu. Credit card fraud detection using advanced transformer model, 2024.

[Yu *et al.*, 2024c] Haoran Yu, Chang Yu, Zihan Wang, Dongxian Zou, and Hao Qin. Enhancing healthcare through large language models: A study on medical question answering. *arXiv preprint arXiv:2408.04138*, 2024.

[Zhao and Gao, 2024] Yu Zhao and Haoxiang Gao. Utilizing large language models for information extraction from real estate transactions. *arXiv preprint arXiv:2404.18043*, 2024.

[Zhao and Liu, 2024] Yu Zhao and Fang Liu. A survey of retrieval algorithms in ad and content recommendation systems, Jun 2024.

[Zhao *et al.*, 2024] Yu Zhao, Haoxiang Gao, and Shiqi Yang. Utilizing large language models to analyze common law contract formation, Jun 2024.

[Zheng *et al.*, 2024] Qi Zheng, Chang Yu, Jin Cao, Yongshun Xu, Qianwen Xing, and Yinxin Jin. Advanced payment security system:xgboost, catboost and smote integrated, 2024.