Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 242 (2024) 1082–1088

Procedia Computer Science

www.elsevier.com/locate/procedia

11th International Conference on Information Technology and Quantitative Management (ITQM 2024)

Real Estate with AI: An agent based on LangChain

Kasper Raupach Haurum[a], Ruiqi Ma[b,c,d], Wen Long[c,d,e,] *

[a] Business School, Aalborg University, Aalborg 9220, Denmark
[b] Sino-Danish College, University of Chinese Academy of Sciences, Beijing 100190, China
[c] Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China
[d] Key Laboratory of Big Data Mining & Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[e] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

Abstract

Recent developments in large language models (LLMs) have opened new avenues for the real estate industry. These models not only understand language but also function as intelligent agents, engaging with investors through open-ended conversations and influencing their decision-making. Utilizing unstructured data from a professional Danish real estate website, we developed a real estate AI agent in both English and Danish using LangChain and Pinecone. Through testing and evaluation, our agent has demonstrated superior professional and concise outputs compared to other LLMs like Doubao and ChatGPT 4 and shown excellent performance and effectiveness. Our work serves as a reference for AI in real estate investment-related research and proposes new solutions to the "unprofessional foundation " and " expensive consulting fee" problems encountered by ordinary investors in their investment decisions.

Keywords: Real Estate Decision-making; LLMs; LangChain; Text Data Mining

1. Introduction

Increased investment in real estate leads to higher demand. However, many investors struggle to make informed decisions due to a lack of professional knowledge, often referred to as cognitive bandwidth, and difficulties in accessing necessary materials, which are frequently financially inaccessible (such as the costs associated with hiring

* Corresponding author. Wen Long.
E-mail address: longwen@ucas.ac.cn

10.1016/j.procs.2024.08.199

an expert). Additionally, their decisions are often swayed by complex psychological factors and external information (Elster, 2016). Information asymmetry further places consumers at a disadvantage, especially when dealing with well-informed real estate agencies. In Denmark, for example, buyers usually seek legal advice primarily for insurance purposes during property transactions. This heavy reliance on legal counsel could increase the risk of financial losses, known as "lemons" (Akerlof, 1970), in high-cost real estate purchases.

Emerging technologies are enhancing financial accessibility, notably through the widespread adoption of Artificial Intelligence (AI), which has popularized smart chatbots. These bots provide a low-cost means for acquiring new knowledge and serve as effective decision-making aids in professional sectors. In the realm of real estate investment, where transactions are complex, there is a significant demand for intelligent decision-support systems to assist consumers facing tough decisions (Craswell, 1985). A Large Language Model (LLM) is essentially a neural network language model, equipped with a vast array of parameters—often billions of weights—trained on large volumes of unlabeled text through self-supervised or semi-supervised learning methods (Liu et al., 2021; Baevski et al., 2022). These models are renowned for their ability to produce high-quality English sentences (Villalobos, 2023). Although humans outperform LLMs in specific tasks according to BIG-bench tests (Srivastava, 2022), LLMs excel in processing speed. Google's Duplex technology exemplifies this advancement by facilitating human-like conversations, representing a significant evolution in data interaction that aids consumers in overcoming cognitive limitations and improving decision-making (Matias & Leviathan, 2018).

However, the accuracy and credibility of AI-generated content remain concerns (Maedche et al., 2019). Recent research has introduced a new method for constructing a domain-specific AI agent using the advanced GPT model, which is pretrained on a constantly updated professional corpus and further refined by human annotators to adapt to consumers' evolving needs. For instance, these AI agents offer personalized and timely assistance, tailored precisely to the user's requirements (Gnewuch et al., 2017). Additionally, experimental applications of AI agents in the therapy sector provide a fascinating glimpse into potential future uses (Lukac et al., 2023; Marchi et al., 2024). In the financial services sector, the application of such technology presents intriguing possibilities. Conversational agents, including those using LangChain, offer comprehensive solutions that streamline communication and facilitate decision-making for complex business challenges. Companies like JPMorgan Chase have developed AI assistants that are adept at understanding and navigating financial data, aiding employees in their decision-making processes (Veloso et al., 2021; Belhaj, M., & Hachaıch, 2021). By utilizing natural language processing to distill complex financial market information, these AI tools enable employees to make better-informed decisions, effectively expanding their cognitive bandwidth (Mullainathan, 2013).

We aim to tackle three significant challenges currently facing the real estate market: information asymmetry between buyers and suppliers, bounded rationality combined with information overload experienced by average buyers, and the formation and influence of preferences on decision-making. Our focus is on the consumer base in Denmark, necessitating the use of both Danish and English data sources to adequately serve both national and international audiences in the Danish real estate market. We leverage the capabilities of the latest LLMs to develop a proof-of-concept agent specifically for mortgage and finance-related issues. This agent distills complex financial and real estate information into user-friendly responses, tailored to individual queries. Our AI agent is designed to assist potential homeowners and investors by reducing decision-making risks in this intricate domain and is tailored to meet the unique needs and circumstances of each consumer.

The highlights of our work consist of three key aspects: Firstly, we tackle "bounded rationality" and "information overload" cognitive challenges (Berger, 2018) that are especially significant in the complex landscape of real estate, where factors such as insurance costs, property conditions, location, and transportation must be considered. Secondly, leveraging artificial intelligence in this manner enhances consumers' decision-making capabilities and saves them time and money that would otherwise be spent sifting through excessive amounts of information or hiring a professional to assist. Thirdly, we have created both English and Danish language agents and achieved cross-validation between them. Our work provides a reference for further research in building AI agents for minority languages.

2. Methodology

2.1. Data Collection and Preprocessing

A. Data collection. To acquire updated data, we focused on sources from 2018 or newer and scraped text from HTML pages and PDF files using the PyPDF2 and BeautifulSoup4 libraries. Our target sites included Finans Danmark, Bomae, Coface, Nationalbanken, Advodan, Forenetkredit, Jyske Realkredit, OnlineMortgageAdvisor, and Tjeklån, with content in both English and Danish. We chose unstructured data for its detailed explanations, expert insights, updates, and technical information. The collected text data underwent cleaning involving data masking and a custom function to remove stop words, symbols, and other non-essential content like time, numbers, and incomplete sentences (shorter than 10 characters). Pre-processing involved tokenization, lemmatization, and removal of punctuation. We also handled inconsistencies in text formatting and normalized the case of all text. To avoid noise and overlap, the data was divided into Danish and English subsets, resulting in 1765 English and 1414 Danish sentences. The experiment was to be conducted on Google Colab within a Python 3.9 virtual environment, and the process was dockerized for reproducibility and ease of deployment, whereas Docker being a service that automates the deployment of applications inside lightweight, portable containers, as shown by Merkel (2014).

B. Embedding. Embedding transforms high-dimensional word or phrase vectors into lower dimensions, better representing their semantic meanings and correlations for models to comprehend. Previous research has utilized embedding technologies such as Word2Vec, FastText, and Doc2Vec. However, emerging methods like SBERT, BERT, and the Ada model have enhanced efficiency and productivity and simplified interactions (Muennighoff, 2022). Consequently, we employ the 'text-embedding-ada-002' model to generate embeddings, converting text tokens into numerical vectors. The context length is 8192 and the embedding size is 1536 dimensions. Since we don't generate the full text, set max_tokens to 1. We use cosine similarity to measure of similarity in the direction of two vectors, which is calculated as follows.

$$\cos\theta = \frac{A \cdot B}{|A| \times |B|} \tag{1}$$

Where A and B are n-dimensional vectors corresponding between two texts.

C. Clustering. We employed K-means clustering, which can efficiently handle large-scale text clustering, to categorize dimensionality-reduced sentences into distinct groups. To obtain the optimal clustering number, we use elbow method, by calculating the sum of squares due to error (SSE), and when k reaches the true clustering number, the return to the aggregation degree obtained by k becomes rapidly smaller, and this is when the decrease rate in SSE undergoes an abrupt change. The inflection point is the optimal number of clusters, which is three.

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{p \in C} |p - m_i|^2 \tag{2}$$

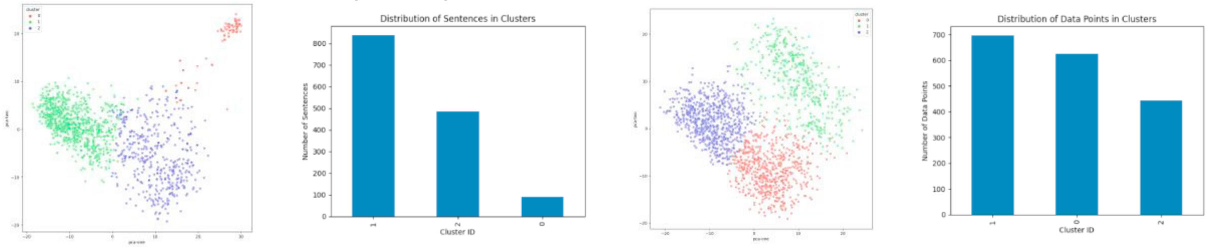
Where c_i is the i th cluster, p is the sample points, m_i is c_i's centroid.

K-Means starts by randomly selecting centroids $\mu_i, i = 1 \cdots k$, and then it labels each data point to the nearest centroid, as $label_i = \arg\min_{1 \le j \le k} \|x_i - \mu_j\|$. After that, the centroids are updated based on the assigned points' mean $\bar{\mu}_i$. This process repeats until convergence, resulting in distinct clusters. We then identified the top ten representative sentences for each cluster by calculating the Euclidean distance from each sentence to the centroid, the distance between two text vectors is calculated as follows.

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

Where x_i is the i element of text vector A and y_i is the i element of text vector B.

To facilitate a better comparison of the outputs from the three specified cluster groups for each language set, we used principal component analysis (PCA) to reduce the clustering results to two dimensions. In Figure 1, the distribution of the Danish embedding clusters shows a noticeable skewness compared to the English clusters. This skewness may stem from the Danish preprocessing, which did not include the removal of numerical values, leading to an outlier cluster of numerical values in Cluster 0, while other data were grouped together in Clusters 1 and 2. This suggests that recent LLMs are not as optimized for tokenization in minor languages as they are for English. This difference underscores the necessity of building AI agents separately for each language to avoid skewed data grouping and the introduction of noise during training.



(a) The clustering in DK language (b) The clustering in ENG language

Fig. 1. The distribution of embeddings over the three cluster groups in two languages.

After generating vector-based embeddings, we utilize Pinecone for storage and accessibility. AI models, such as LLMs, produce embeddings that are complex and multifaceted, making their management challenging. Pinecone, a specialized database, provides optimized storage and querying capabilities specifically for embeddings. Unlike traditional databases, vector databases like Pinecone offer the functionalities of standard databases but are uniquely equipped to handle vector embeddings, which traditional scalar-based databases cannot adequately support. In our Pinecone setting, we use "us-east1-gcp" (Google Cloud in the eastern United States, Virginia) environment for low-latency, high-availability and scalable cloud computing services.

2.2. Model Construction

To develop a chatbot capable of maintaining contextual interaction, we utilized ChatOpenAI from langchain.chat_models, initiating the temperature parameter based on the gpt-3.5-turbo deployed model. LangChain agents, central to our work, process conversational context represented in vectors to respond appropriately to user inquiries. These agents operate by transforming input through an embedding and tokenization pipeline, then formulating an output in a user-friendly format. By retrieving finance-related answers from the Pinecone vector database, the agent is trained to recognize and respond to specific prompts about loan applications and types. It matches the closest vectors to the query and formulates a response communicated back to the user. This process equips the agent to understand and respond to queries in real time. Figure 2 illustrates our model design framework.

Additionally, we allocated two separate programming environments for the Danish and English language data to set up two distinct agents. This approach allowed us to tailor strategies and methodologies to the unique demands of each language while maintaining the same intended output across both environments.

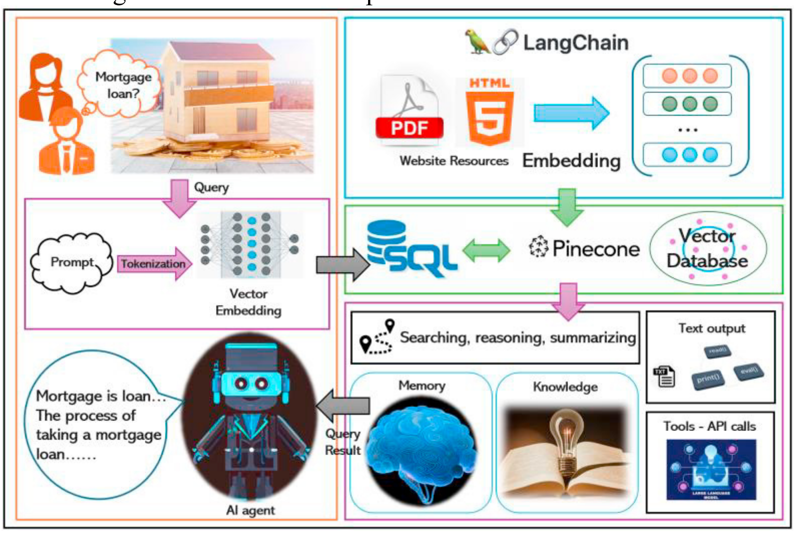


Fig. 2. Framework of real estate AI agent.

3. Test and Evaluation

Based on five design evaluation methods outlined by Hevner et al. (2004), we assessed the performance of our output by testing it with real estate-related questions. We compared the results with those of other LLMs and conducted human evaluations. Additionally, we performed ablation analysis by manipulating various module components to determine their individual impact on the overall performance. This comprehensive approach allowed us to validate the effectiveness and refine the functionality of our AI model in handling specific real estate queries.

3.1. Performance Test

The embeddings we generated exhibited an average cosine similarity of 76%, with the peak sentence cosine similarity reaching 87%. This high level of similarity indicates robust semantic consistency between sentences, demonstrating that their interrelationships are well-structured. Additionally, we selected Doubao, developed by TikTok, and ChatGPT 4 to evaluate the performance of our real estate AI agent. Due to page constraints, we only presented results from the English agent. Initially, we input the query, "What is the process of taking a Mortgage Loan?" and reviewed the top three most relevant documents provided by the agent. Subsequently, we configured our agent to generate the output. For comparative analysis, we pre-trained Doubao and ChatGPT 4 as professional real estate agents and input the same question to highlight the differences in their responses, as illustrated in Table 1.

Table 1. Comparison of the output performance

User prompt: **What is the process of taking Mortgage Loan?**	
The top 3 relevant sentences	[1] During the loan period, borrowers will make principal and interest payments to the mortgage bank, which will then transfer these amounts to investors. [2] Naturally, the loan process can look very different depending on what loan and bank you have chosen. [3] When mortgage bank has approved the loan amount, the funds can be disbursed and the purchase completed.
Our real estate agent	The process of taking a mortgage loan can vary depending on the loan and bank you have chosen. Generally, the process involves applying for the loan, submitting documentation such as income and credit history, getting approved for the loan amount, and then making principal and interest payments to the mortgage bank during the loan period. Once the mortgage bank has approved the loan amount, the funds can be disbursed, and the purchase completed. The money collected from the borrower's payments will fund the mortgage loan. It's important to research and compare different loan options and banks to find the best fit for your financial situation.
Doubao	The process generally includes the following steps: application, submission of materials, evaluation, approval, signing of contract, mortgage registration, and loan release.
ChatGPT 4	As a real estate professional, I recommend starting your mortgage process with a thorough review of your finances and securing pre-approval to understand your budget. Then, focus on finding a property that fits your needs and budget. Once you find it, make an offer and proceed with a formal loan application. Ensure all documentation is in order for the loan processing and closing phases. Remember, working with a trusted mortgage broker can streamline the process and potentially offer better rates.

In comparison, our agent demonstrates superior capability in understanding and synthesizing accurate and correct answers, rather than merely listing all possible answers as Doubao and ChatGPT 4 tend to do. Furthermore, we sourced input from Investopedia (an out-of-sample data source) and used the key takeaways from the corresponding answers as the ideal output for calculating the ROUGE score r (Lin, 2004). The results, as shown in Table 2, indicate that our agent performs the best overall. Specifically, in 1-gram prediction (ROUGE-1), our agent matches the performance of ChatGPT 4 and outperforms the other scores.

Table 2. Comparison of ROUGE score (Source of benchmark: https://www.investopedia.com/)

LLMs	ROUGE-1	ROUGE-2	ROUGE-L
Our agent	**0.7714**	**0.6346**	**0.7714**
Doubao	0.6857	0.3702	0.6286
ChatGPT 4	**0.7714**	0.5865	0.6571

3.2. Cross-validation Test

Though we train two LLMs to get more accurate and specific agent, it's still wonder whether the English input will work for the Danish agent (or vice versa) given that the tool of the LangChain supports the translation. Table 3 depicts the outputs both from Danish input and English input. Seen from our text with the Danish agent, which is pretrained by the Danish corpus, it also answers matching to the point with English text input.

Table 3. Cross-validation test for Danish agent

Input	Output
Hvad er et fast rente lån?	Et fast rente lån er en type lån, hvor renten er fastsat og ikke ændrer sig i løbet af lånets løbetid. Dette betyder, at låntageren vil have en fast månedlig betaling, som ikke vil ændre sig, uanset om renten stiger eller falder i markedet. Det kan være en fordel at vælge et fast rente lån, hvis man ønsker at have en forudsigelig og stabil månedlig betaling.
What is a fixed rate loan?	A fixed rate loan is a type of loan where the interest rate remains the same throughout the loan term or for a specified period of time. This means that the borrower's monthly payments will remain the same, providing stability and predictability in their budget. Fixed rate loans are commonly used for mortgages and home equity loans.

In Table 4, we further validate the responses generated by both the Danish and English agents when given similar inputs. In Danish, we tested with a query translated as "Can you describe how you acquire a loan from the bank?" and compared it to the previous output from the English agent for the query "What is the process of taking a mortgage loan?" Although the focus of the two outputs varies, they both include financial terms such as "interest", "period" and "bank", and ultimately provide nearly identical suggestions at the end. This demonstrates that even though these two agents are developed from different content and language corpora, they still perform similarly well. This suggests that answers to specialized questions remain consistent, irrespective of the language and corpus used.

Table 4. Output comparison between Danish and English AI agent

Input	Output
Kan du beskrive hvordan man låner fra banken?	Ja, selvfølgelig. Når man låner fra banken, så tager man et lån, som man skal betale tilbage med renter over en aftalt periode. Det kan være til køb af en bolig, en bil eller andre større investeringer. **Det er vigtigt at undersøge forskellige banker og deres lånevilkår, så man kan finde det bedste tilbud med den laveste rente og de bedste betingelser**.
What is the process of taking Mortgage Loan?	The process of taking a mortgage loan can … and **interest** payments to the mortgage **bank** during the loan **period**. Once … **It's important to research and compare different loan options and banks to find the best fit for your financial situation**.

3.3. Ablation Study

The LLMs utilize tools that integrate various logic chains and interact with external data sources to boost the agent's answering capabilities. We define four tools: the first, "Prefit for AI", is used for prompt-engineering and describes itself as an AI conversational bot focused on the Danish real estate market and banking and finance. The second, "Intermediate Answer", assists with search queries. The third, "Language Understanding", processes natural language inputs to generate appropriate responses. The fourth, "Text Summarization", condenses long documents into concise summaries. In an ablation study, we removed each tool individually to assess its impact on the framework, using performance test queries as benchmarks. The results, shown in Table 5, indicate that our agent primarily requires the "Intermediate Answer" tool to function effectively. Without it, ROUGE scores decrease significantly compared to when the other three tools are removed, particularly affecting 1-gram and long sequence performance.

Table 5. Ablation study of tools

Tools removal	Output	ROUGE-1	ROUGE-2	ROUGE-L
Prefit for AI	The process of taking … involves paying closing costs and fees.	0.964	0.673	0.893
Intermediate Answer	Taking a mortgage loan …the loan and bank you have chosen.	**0.893**	0.724	**0.857**
Language Understanding	The process of taking …taking out a mortgage loan.	0.964	0.714	0.929
Text Summarization	The process of taking …taking out a mortgage loan.	0.964	0.714	0.929

4. Conclusion and Future Work

In conclusion, we developed a conversational agent using the Large Language Model with support from OpenAI, Pinecone, and LangChain for real estate finance and mortgage matters. We also explored enhancing the agent's capacity by augmenting it with various tools through a chaining process. We discovered that the agent's ability to mitigate information asymmetry lies in its use of zero-shot learning to accurately answer queries about mortgage and finance issues, even when data is not directly available. This capability helps combat information asymmetry and aids users in navigating complex real estate matters, thereby reducing potential cognitive dissonance and information overload. Looking ahead, we recommend integrating a SQL database alongside the existing Pinecone NoSQL/Vector database. This setup would allow a SQL agent to process user data, linking it with the Pinecone database that manages finance and mortgage-related metadata. Additionally, we suggest that using TensorFlow instead of OpenAI's embedding could reduce preprocessing and coding costs (Nandy, 2018). Furthermore, reducing the dimensionality to 512, instead of maintaining 1536 dimensions, could help mitigate the curse of dimensionality.

Acknowledgements

The authors would like to express their deepest gratitude to Homechat.ai for providing the Danish real estate case study and Muhammad Raiyan Alam for contributing code about OpenAI that facilitated this research. This research was partly supported by the grants from National Natural Science Foundation of China (No.72231010).

References

[1] Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics* (pp. 235-251). Academic Press.

[2] Baevski, A., Hsu, W. N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022, June). Data2vec: A general framework for self-supervised learning in speech, vision and language. *In International Conference on Machine Learning* (pp. 1298-1312). PMLR.

[3] Belhaj, M., & Hachaıchi, Y. (2021). Artificial Intelligence, Machine Learning and Big Data in Finance Opportunities, Challenges, and Implications for Policy Makers.

[4] Craswell, R. (1985). Interpreting deceptive advertising. *BUL Rev.*, 65, 657.

[5] Elster, J. (2016). *Sour grapes*. Cambridge university press.

[6] Gnewuch, U., Morana, S., & Maedche, A. (2017, December). Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *ICIS* (pp. 1-13).

[7] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

[8] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

[9] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1), 857-876.

[10] Lukac, S., Dayan, D., Fink, V., Leinert, E., Hartkopf, A., Veselinovic, K., ... & Ebner, F. (2023). Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Archives of Gynecology and Obstetrics*, 308(6), 1831-1844.

[11] Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., ... & Söllner, M. (2019). AI-based digital assistants: Opportunities, threats, and research perspectives. *Business & Information Systems Engineering*, 61, 535-544.

[12] Marchi, F., Bellini, E., Iandelli, A., Sampieri, C., & Peretti, G. (2024). Exploring the landscape of AI-assisted decision-making in head and neck cancer treatment: a comparative analysis of NCCN guidelines and ChatGPT responses. *European Archives of Oto-Rhino-Laryngology*, 1-14.

[13] Matias, Y., & Leviathan, Y. (2018). Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone.

[14] Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. Linux j, 239(2), 2.

[15] Muennighoff, N. (2022). Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

[16] Mullainathan, S., & Shafir, E. (2013). Scarcity: Why having too little means so much. Macmillan.

[17] Nandy, A., & Biswas, M. (2017). *Reinforcement Learning: With Open AI, TensorFlow and Keras Using Python*. Apress.

[18] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arxiv preprint arxiv:2206.04615*.

[19] Veloso, M., Balch, T., Borrajo, D., Reddy, P., & Shah, S. (2021). Artificial intelligence research in finance: discussion and examples. *Oxford Review of Economic Policy*, 37(3), 564-584.

[20] Villalobos, P. (2023). Scaling Laws Literature Review. *Published online at epochai. org*.