Terminology Enhanced Retrieval Augmented Generation for Spanish Legal Corpora

Patricia Martín-Chozas
Universidad Politécnica de Madrid
patricia.martin@upm.es

Pablo Calleja
Universidad Politécnica de Madrid
p.calleja@upm.es

Carlos Rodríguez Limón
Universidad Politécnica de Madrid
rodriguezlimoncarlos@gmail.com

Abstract

This paper intends to highlight the importance of reusing terminologies in the context of Large Language Models (LLMs), particularly within a Retrieval-Augmented Generation (RAG) scenario. We explore the application of query expansion techniques using a controlled terminology enriched with synonyms. Our case study focuses on the Spanish legal domain, investigating both query expansion and improvements in retrieval effectiveness within the RAG model. The experimental setup includes various LLMs, such as Mistral, LLaMA3.2, and Granite 3, along with multiple Spanish-language embedding models. The results demonstrate that integrating current neural approaches with linguistic resources enhances RAG performance, reinforcing the role of structured lexical and terminological knowledge in modern NLP pipelines.

1 Introduction

The increasing complexity of legal texts and the demand for efficient legal information retrieval have led to the exploration of different types of advanced Natural Language Processing (NLP) techniques.

At the European level, several initiatives can be found, such as the EUR-Lex platform[1], that provides access to European Union law, including treaties, legislation, case law, and legislative proposals, being a crucial resource for legal professionals seeking comprehensive legal information within the EU framework.

In terms of data standardisation, the European Case Law Identifier (ECLI) was created to standardize the citation of case law across Europe. By introducing a uniform identifier and a set of metadata, ECLI facilitates easier access and citation of European case law, enhancing the efficiency of legal information retrieval systems.

The EU-supported H2020 Lynx project incorporated both initiatives to create a knowledge-driven AI service platform. This platform was designed for content processing, enhancement, and analysis within the legal sector, with the main aim of aiding companies in efficiently tackling compliance challenges across different languages and legal systems (Schneider et al., 2022).

However, one of the key limitations of projects of that kind is the challenge of processing legal texts in low-resource languages, such as Spanish. Traditional NLP techniques, such as rule-based approaches and classical machine learning models, struggle with the complexity of legal language, which often includes long sentences, archaic terms, and jurisdiction-specific terminology.

In this project, different language resources were generated but, still, the data scarcity limited the effectiveness of semantic search and entity recognition, making legal information retrieval in Spanish less accurate and comprehensive compared to other European languages.

Currently, large language models (LLMs) have become an unprecedented Artificial Intelligence resource for processing and querying information. However, in critical domains such as the legal sector, these models cannot rely only on outdated training data or generate hallucinations (Magesh et al., 2024). Moreover, legal domain vocabulary and terminology evolve over time, requiring continuous adaptation.

In this context, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a promising solution to these issues, combining the strengths of retrieval-based and generative models to improve the accuracy and relevance of automated legal text processing.

This paper investigates the effectiveness of the combination of terminologies in the context of RAG scenarios. In particular, through a query expansion technique and a reranking process. This study has been performed over the Spanish legal domain, relying on Spanish language models for

[1] https://eur-lex.europa.eu/

the embedding process and LLMs such as LLaMA, Mistral and Granite. The experiments are available in an open repository[2].

2 Related Work

The rapid advancement of LLMs has boosted NLP applications across different domains. LLM-driven pipelines have demonstrated significant capabilities in text comprehension (Breton et al., 2025), generation (Satterfield et al., 2024), and data retrieval (Ganesan et al., 2024). Yet, LLMs present some limitations, the most notable being the lack of specific knowledge in specialised domains, hallucinations, and the high computational resources required for model updates (Zhao et al., 2023; Fan, 2024).

To address these challenges, significant efforts have been made in the implementation of RAG, leveraging its advantages to enhance the capabilities of LLMs in tasks that require a high level of recent and accurate knowledge, including question answering, AI4Science[3] and software engineering (Izacard and Grave, 2020; Shi et al., 2023). RAG systems help such tasks by providing the LLM with objective data from external resources to generate accurate responses (Lewis et al., 2020), using various fact-identification mechanisms (Asai et al., 2023; Thulke et al., 2021). Additionally, other studies have demonstrated that RAG can effectively reduce hallucinations in conversational tasks by providing the model with verified and contextualized information (Shuster et al., 2021).

Still, these limitations are specially relevant in fields such as medicine and law, where the reliability of information is essential. For instance, recent studies have shown that hallucinations in the legal domain are particularly frequent and concerning, with rates ranging between 69% and 88% in responses to legal queries made to some of the most advanced LLMs (Dahl et al., 2024). Previous research has shown that domain-specific pretraining significantly improves LLM performance in technical fields, such as law (Borgeaud et al., 2022). Additionally, recent advancements in model fine-tuning, such as Reinforcement Learning from Human Feedback (RLHF), have enhanced LLM adaptability to domain-specific language (Ouyang et al., 2022). Other approaches have started to apply this technique for the generation of documentation in Spanish public entities (Collado Alonso et al., 2024).

Despite the efforts mentioned above, there is room for research on the application of RAG models in the legal domain, particularly for languages with limited NLP resources such as Spanish. While advancements have been made in adapting LLMs to legal contexts, to the best of our knowledge, there are still few studies specifically focused on RAG-based experiments that leverage language resources for Spanish legal texts, which highlights the need for further exploration of such techniques to tackle the complexities of Spanish legal language.

3 Experiment

3.1 Methodology

The methodology of this experiment includes four key steps:

- **Knowledge Base Creation**: The legal corpora used in the RAG is segmented, processed, and stored in a vector database using FAISS (Facebook AI Similarity Search)[4] due to its ability to handle large amounts of data with high efficiency, providing outstanding performance in retrieval tasks based on semantic similarity.

- **Information Retrieval**: To enhance the retrieval capabilites of the RAG system, query expansion techniques are applied, including synonym and related-terms integration from existing language resources, as well as document reranking techniques using LLM models to improve search precision.

- **Prompt Engineering**: In order to ensure clear and contextualized interactions with LLMs, guides of good practices for prompt engineering have been followed (Phoenix and Taylor, 2024), including clarity, specificity, context and length.

- **Response Generation**: This step implements a LLM to generate the legal response. In this experiment, three well established models have been compared.

3.2 Data Selection

Given the interest of working with low resource, domain specific and small data, this experiment is

[2] https://github.com/oeg-upm/term-rag

[3] https://ai4sciencecommunity.github.io/

[4] https://ai.meta.com/tools/faiss/

focused on the Spanish labour law domain. The corpus used in the RAG is the Spanish Workers' Statute[5] which contains 1,568 sentences and 54,849 tokens. The terminology used for the query expansion step is a semi-automatically generated resource (Martín-Chozas et al., 2022) generated in the context of the Lynx project that contains 1,015 terms extracted from the same corpus, including main terms, synonyms, broader, narrower and related terms. The dataset employed for the evaluation (Calleja et al., 2021) was also generated in the Lynx project and includes 149 manually validated questions and answers from the same corpus.

3.3 Implementation

As depicted in the system architecture (Figure 1) the process is initiated by an input query, which is expanded using the terms from the terminology, with the aim of expanding semantic coverage. The expanded queries are then processed by the embedding models, which are also in charge of converting the corpus into embeddings.

Regarding the embedding models, two well-known Spanish models, supported by the PlanTL initiative[6], have been implemented and compared: roberta-base-bne (Fandiño et al., 2022) and RoBERTalex (Gutiérrez-Fandiño et al., 2021). The former is a widely employed Spanish adaptation of the RoBERTa model (Liu et al., 2019), trained on the Biblioteca Nacional de España (BNE) corpus, which includes legal and administrative texts. The latter is specifically trained on a corpus of Spanish legal texts, which is particularly suitable to handle specialised terminology and legal linguistic structures.

After expanding the original query and generating the embeddings, cross-encoder reranking techniques are applied to refine the document ranking. This process involves reordering the retrieved documents using a cross-encoder model—specifically, ms-marco-MiniLM-L-12-v2 to assess their semantic relevance to the query. While the initial retrieval ranks documents based on vector similarity, the reranking step evaluates query-document pairs directly, ensuring that the most relevant results appear at the top, which are then processed by the LLMs, generating the output answer.

Concerning the LLMs, three models have been implemented an compared: Mistral (Jiang et al., 2023), LLaMA3.2[7], and Granite3-dense (Granite Team, 2024) for response generation. Firstly, Mistral handles complex text processing tasks and is trained with a diverse dataset covering multiple languages and specialised domains, with a strong emphasis on Spanish legal terminology. Secondly, LLaMA3.2 also performs particularly well in Spanish, and specifically in the legal domain. In addition, it is designed to efficiently use computational resources. Finally, Granite3-dense is well known for its deep contextual understanding through dense embedding techniques and is effective in analyzing legal documents and generating well-contextualized answers.

4 Evaluation

To assess the proposal of this work, we employed several standard NLP evaluation metrics that measure the quality of the generated responses from different perspectives:

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: This metric quantifies the lexical overlap between the generated text and a reference, making it particularly useful for summarization and text generation tasks.
- **F1 Score**: Balances precision and recall, providing a robust measure of performance in tasks where both completeness and accuracy are critical.
- **SAS (Semantic Answer Similarity)**: Evaluates the semantic proximity between the generated response and an expected reference, allowing for a more flexible assessment beyond exact word matching.
- **BERTScore**: Contextual embeddings are used to determine text similarity, capturing deeper semantic relationships compared to lexical overlap-based metrics.

4.1 Results

The obtained results are presented in Table **??**. The different models used (LLaMA, Mistral, and Granite) have been evaluated separately. For each model, different approaches have been assessed: without RAG, RAG with the roberta-base-bne embedding model, and RAG with the RoBERTalex model. Additionally, query expansion (QE) has been tested for each embedding model.

[5] https://www.boe.es/eli/es/rdlg/2015/10/23/2/con

[6] https://plantl.digital.gob.es/

[7] https://huggingface.co/meta-llama

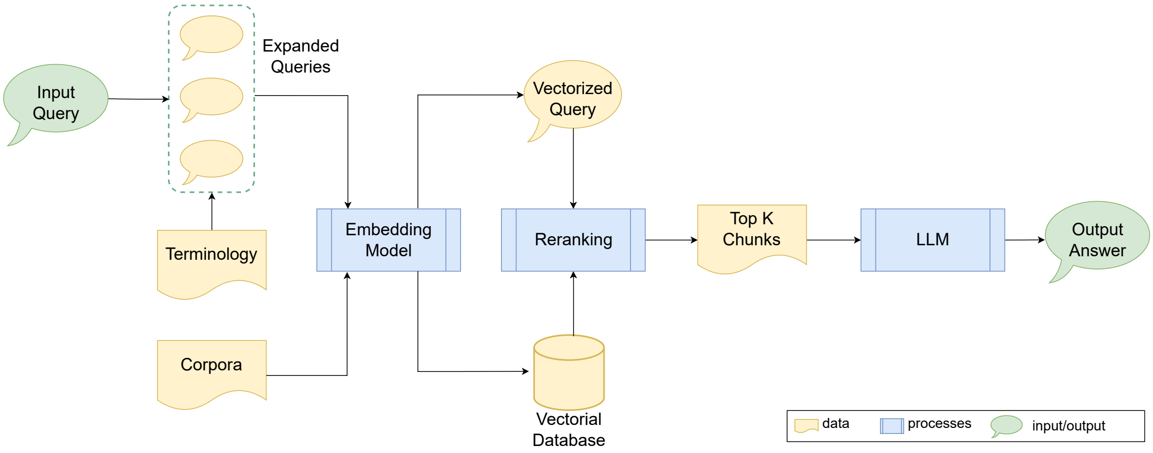

Figure 1: Experiment architecture

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**F1**	**SAS**	**BERT**
No RAG						
LLaMA3.2	0.24	0.06	0.14	0.1	**0.62**	0.9
Granite3-dense	0.18	0.04	0.11	0.09	0.55	0.77
Mistral	0.27	0.06	0.15	0.12	0.58	0.93
roberta-base						
LLaMA3.2	0.28	0.1	0.19	0.17	0.53	0.93
Granite3-dense	0.21	0.07	0.14	0.12	0.53	0.79
Mistral	0.33	0.12	0.21	0.18	0.56	**0.95**
RoBERTalex						
LLaMA3.2	0.31	0.13	0.21	0.18	0.55	0.94
Granite3-dense	0.21	0.08	0.14	0.11	0.56	0.77
Mistral	**0.35**	**0.14**	**0.23**	**0.19**	0.59	**0.95**
roberta-base Expanded						
LLaMA3.2	0.29	0.11	0.2	0.17	0.54	0.93
Granite3-dense	0.21	0.07	0.14	0.12	0.53	0.79
Mistral	0.33	0.12	0.21	0.18	0.57	**0.95**
RoBERTalex Expanded						
LLaMA3.2	0.33	**0.14**	**0.23**	**0.19**	0.56	0.94
Granite3-dense	0.22	0.08	0.15	0.12	0.57	0.79
Mistral	**0.35**	**0.14**	**0.23**	**0.19**	0.58	**0.95**

Table 1: Obtained results of the different models: LLaMA, Mistral, Granite. The evaluation is metrics are Rouge1, Rouge2 RougeL, F1-Score, SAS, and BertScore. All the models have been evaluated with different RAG approaches: No RAG, RAG with the roberta-base-bne embedding model, RAG with the RoBERTalex embedding model and both embedding models with and without query expansion (QE).

The results indicate that RAG techniques consistently improve the performance of all the models evaluated in nearly every metric. The ROUGE scores, in particular ROUGE-1, ROUGE-2, and ROUGE-L, show the most significant improvements, indicating that the inclusion of retrieved context helps models generate more relevant outputs. This pattern is especially evident in models that initially had weaker performance without RAG, such as Granite.

From the different RAG configurations, those using the RoBERTalex embedding model perform better. This could be due to the specific fine tuning of this model for the Spanish legal domain. This translates to better generation outcomes, especially when combined with strong language models like Mistral, that achieves the highest scores across most metrics, including ROUGE-1 (0.35), ROUGE-2 (0.14), ROUGE-L (0.23), F1 (0.19), and BERTScore (0.95).

The impact of QE techniques is also observable, although moderate. On average, QE contributes an additional improvement of around 1% to 5%, depending on the metric and model. For instance, in the case of the Mistral model using RoBERTalex, adding QE increases the F1 from 0.19 to 0.21 and slightly improves the SAS score from 0.59 to 0.60. These scores suggest that QE can help refine the retrieval process by producing more semantically rich queries.

Finally, another important remark that can be observed in Figure 2 is that Mistral consistently outperforms both LLaMA and Granite all retrieval and QE settings, which is probably due to its training in specialised domains.

4.2 Discussion

The main limitations observed in the experiment are twofold: the reference question-answer dataset and the evaluation metrics.

Regarding dataset constraints, the primary issue is the limited number of instances. The dataset contains too few question-answer pairs to be fully representative. Additionally, improving performance is particularly challenging, since error analysis revealed that several failed queries involve legal questions requiring multi-document references,

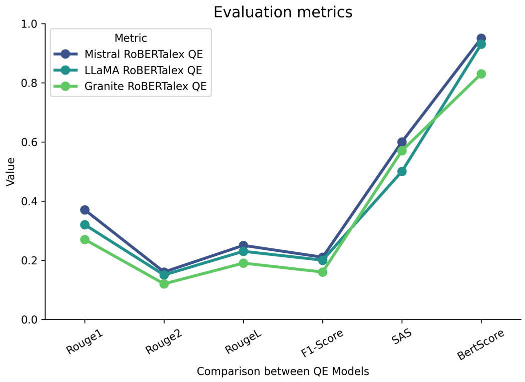

Figure 2: Comparison of the best performing QE approaches of the three different models.

which current retrieval models struggle to handle efficiently. A larger dataset would allow a more comprehensive assessment of the impact of query expansion and reranking techniques.

Furthermore, evaluation techniques for RAG remain limited. While BertScore provides high semantic similarity scores, these do not always align with the actual relevance of the model's responses. Similarly, ROUGE metrics yield low values due to their strict n-gram-level comparisons. More advanced approaches, such as RAGAS (Es et al., 2024), which leverage powerful LLMs to evaluate RAG-generated responses, are being considered to better assess how the model handles query expansion and reranking processes.

5 Conclusions and Future Work

This paper proposes the implementation of a RAG system to enhance the use of terminologies along LLMs in the context of Spanish legal texts, particularly the Spanish Workers' Statute. Specifically, this work intends to research the impact in the information retrieval step of incorporating query expansion techniques enriched with synonyms and related terms from legal terminologies. We evaluate three LLMs, including Mistral, LLaMA3.2, and Granite3-dense and two Spanish embedding models. The results confirm that integrating neural language models with curated linguistic resources enhances RAG performance, highlighting the value of structured language data in modern NLP applications.

However, we have observed a low recall of the synonyms from the terminology, which translates in a low number of questions expanded. This limitation highlights the need for generating more complex and specific terminological resources, including a deeper research on Automatic Terminology Extraction algorithms that are able to identify specific terms in the domain.

On the other hand, the results emphasize the need for expanding Spanish legal corpora with larger annotated datasets (for Question Answering, in this case) to improve model evaluation. Additionally, integrating structured legal data, such as court rulings, with unstructured text can enhance retrieval capabilities.

Future research envisions the development of adaptive RAG models that dynamically adjust to legal question complexity using techniques such as reinforcement learning.

Acknowledgments

This work has received funding from the INESData project (Infrastructure to Investigate Data Spaces in Distributed Environments at UPM - TSI-063100-2022-0001, under the UNICO I+D CLOUD call) and TeresIA (Terminology and AI), both funded by the Ministry for Digital Transformation and the Civil Service, within the framework of the recovery plan PRTR financed by the European Union (NextGenerationEU).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Julien Breton, Mokhtar Mokhtar Billami, Max Chevalier, Ha Thanh Nguyen, Ken Satoh, Cassia Trojahn, and May Myo Zin. 2025. Leveraging llms for legal terms extraction with limited annotated data. *Artificial Intelligence and Law*, pages 1–27.

Pablo Calleja, Patricia Martín Chozas, Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel, Elsa Gómez, and Pascual Boil. 2021. Bilingual dataset for information retrieval and question answering over the spanish workers statute. In *XIX Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA*.

Miguel Ángel Collado Alonso et al. 2024. Implementación de técnicas de rag (retrieval augmented generation) sobre llm (large language models) para la extracción y generación de documentos en las entidades públicas.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Wenqi Fan. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Balaji Ganesan, Sambit Ghosh, Nitin Gupta, Manish Kesarwani, Sameep Mehta, and Renuka Sindhgatta. 2024. Llm-powered graphql generator for data retrieval. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8657–8660.

IBM Granite Team. 2024. Granite 3.0 language models.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish legalese language model and corpora. *Preprint*, arXiv:2110.12201.

Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

Patricia Martín-Chozas, Karen Vázquez-Flores, Pablo Calleja, Elena Montiel-Ponsoda, and Víctor Rodríguez-Doncel. 2022. Termitup: Generation and enrichment of linked terminologies. *Semantic Web*, 13(6):967–986.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

James Phoenix and Mike Taylor. 2024. *Prompt Engineering for generative AI.* O'Reilly Media, Inc.

Nolan Satterfield, Parker Holbrooka, and Thomas Wilcoxa. 2024. Fine-tuning llama with case law data to improve legal domain performance. *OSF Preprints*.

Julián Moreno Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel, Patricia Martín-Chozas, María Navas-Loro, Martin Kaltenböck, Artem Revenko, Sotirios Karampatakis, Christian Sageder, et al. 2022. Lynx: A knowledge-based ai service platform for content processing, enrichment and analysis for the legal domain. *Information Systems*, 106:101966.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *arXiv preprint arXiv:2102.04643*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).