# Separating Linguistic Competence from Factual Knowledge in Large Language Models

Jaime Collado-Montañez

*Department of Computer Science (University of Jaén), Campus Las Lagunillas, s/n, Jaén, 23071, Spain*

**Abstract**

Deep neural networks have significantly advanced natural language processing techniques, enabling the development of large language models that exhibit impressive capabilities in language understanding and generation. However, these models often internalize vast amounts of factual knowledge, which can lead to issues such as hallucinations and the use of outdated information. This research explores the hypothesis that linguistic competence, the ability to understand and produce natural language, can be codified separately from memorized factual knowledge in neural networks. By developing "fundamental language models" that focus on language understanding and reasoning without internalizing factual data, we aim to create smaller, more efficient models that access up-to-date factual knowledge through external sources using techniques like Retrieval Augmented Generation. Our main objective is to understand the functioning of Large Language Models as reasoning engines, with a special focus on language models for Spanish.

**Keywords**

Large Language Model, Fundamental Language Model, Hallucination, Retrieval Augmented Generation, Explainability

## 1. Justification of the Proposed Research

Deep neural networks have transformed the landscape of natural language processing techniques, enabling the development of Large Language Models (LLMs) through training on massive text collections using neural networks based on the transformer architecture [1]. This has led to the creation of autoregressive systems for text generation with a high capacity for understanding human language. Thus, LLMs have become the core of an increasing number of artificial intelligence tools, with notable examples such as GPT3 [2] and LLaMA-2 [3], which are specially trained for natural language conversation.

However, alongside their remarkable abilities, LLMs also face significant challenges. One prominent issue is hallucination, which can lead to the propagation of misinformation or the generation of content that appears valid but lacks authenticity.

The objective of this research is to explore architectures for solving artificial intelligence (AI) tasks where LLMs are the central reasoning engine, enhancing their capabilities with external tools, such as other knowledge bases. This architecture we call Fundamental Language Model (FLM) could be achieved by removing parts of the neural network storing factual knowledge while preserving the ones related to reasoning and language understanding. Other options to obtain such a model could be pretraining with large datasets curated from every possible factual information.

The remainder of this work is organized as follows: Section 2 presents an overview of the relevant literature concerning LLM emergent abilities and problems; Section 3 shows the main hypothesis and objectives planned for this research. Finally, Section 4 details the methodology followed during the development of this thesis, and Section 5 concludes with some specific research elements proposed for discussion.

*Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.*

jcollado@ujaen.es (J. Collado-Montañez)

0000-0002-9672-6740 (J. Collado-Montañez)

## 2. Background and Related Work

Transformer models are language models pretrained to understand language structures by using semi-supervised learning with huge amounts of data. Encoder transformers such as BERT [4] or RoBERTa [5] use Masked Language Modeling (MLM) mainly while decoder or generative transformers like LLaMa [6], Mistral [7], and GPT [8] are trained using Causal Language Modeling (CLM).

According to the following general definition of emergence, as stated by the Nobel prize-winning physicist Philip Anderson [9]: *"Emergence is when quantitative changes in a system result in qualitative changes in behavior"*, the rapidly growing size of such models, especially the generative ones, into what we call LLMs is allowing them to showcase new emergent abilities such as reasoning [10]. Along with these properties, this semi-supervised pretraining technique allows LLMs to memorize lots of factual data [11] that, in some cases, may lead to problems such as hallucinations [12] and outdated answers when training data does not include the latest events and news. Hallucinations in LLMs refer to instances where the model generates responses that are not factual or grounded in reality but rather are inferred from patterns in the training data. These hallucinations can occur when the model synthesizes information based on statistical correlations in the data rather than true understanding [13].

In addition to that, the use of large corpora of texts from various sources in the generation of pre-trained models results in the model capturing stereotypical patterns present in the texts. This issue, known as bias detection, is related to explainability but focuses on the detection, evaluation, and mitigation of gender, profession, origin, ethnicity, or religion stereotypes present in trained models [14]. The problem has become a topic of interest beyond the field of AI algorithm research and is known as fairness [15] due to its ethical and legal implications.

Additionally, although they seem powerful in terms of results and predictions, large language models have their own limitations. The most significant is opacity or lack of transparency [16]. This means that the logic and internal functioning of these models are hidden from the user, which is a serious disadvantage because it prevents a human, whether expert or not, from verifying, interpreting, and understanding the system's reasoning and how decisions are made. In other words, any sufficiently complex system acts as a black box when it is easier to experiment with than to understand [17].

The study of "foundational" language models can help address both bias issues and contribute to explainability by focusing on the core competencies of natural language understanding and separating knowledge from language-based reasoning. This hypothesis would be applicable to the project's challenges: the analysis of harmful and beneficial content, both in its detection, characterization, and generation.

## 3. Hypothesis and Objectives

The hypothesis behind this line of research is that the ability to understand and produce natural language, i. e. linguistic competence, can be codified separately from memorized factual knowledge in neural networks.

This hypothesis would imply that we can build FLMs that do not store facts but retain language understanding and reasoning capabilities. Thus, we could train smaller models that access factual knowledge through external sources and techniques such as Retrieval Augmented Generation (RAG), potentially removing all sources of hallucination and outdated information.

With this purpose, the main objective of this thesis is to understand the functioning of LLMs as reasoning engines, with a special focus on language models for Spanish. In order to achieve this main objective, the following secondary objectives have been defined:

1. Study of the internal encoding of knowledge for language understanding.
2. Decomposition of the language model's capabilities into different skills.
3. Enhancement of LLM-based AI capabilities through the use of external knowledge bases.
4. Improvement of explainability by decomposing the AI task resolution process.

## 4. Methodology

The following methodology aims to systematically explore the hypothesis that linguistic competence can be separated from factual knowledge in LLMs, focusing on Spanish language models:

- Literature review and initial study: Conduct a comprehensive literature review on the current techniques and advancements in LLMs, focusing on Spanish language models and identify key resources, including scientific forums (AAAI, NeurIPS, ACL) and reference bulletins (PapersWithCode, The Batch).
- Experimental design and evaluation: Define experimental setups for each objective, including initial hypotheses, required datasets, and evaluation metrics. Participate in evaluation forums (CLEF, SemEval, IberLEF) to benchmark against other models and solutions.
- Study of internal encoding of knowledge: Perform a series of experiments to analyze the internal representations of LLMs, utilizing techniques such as layer-wise analysis and compare these internal representations across different models to identify common patterns and structures.
- Decomposition of language model capabilities: Design tasks and benchmarks to isolate and evaluate individual skills and use controlled experiments to test the models' performance on these tasks.
- Enhancement through external knowledge bases: Develop methods to connect LLMs with external databases and knowledge bases using techniques like RAG and conduct experiments to compare the performance of enhanced models against traditional models.
- Dissemination of findings: Prepare and submit research papers to high-impact journals and conferences.

This outline is in its initial stages, focusing on setting up foundational explorations and experiments to investigate the separation of linguistic competence and factual knowledge in language models.

## 5. Research elements proposed for discussion

The following specific research elements provide a comprehensive framework for discussing the hypothesis and objectives, facilitating a deeper exploration of the potential benefits, limitations, and implications of using fundamental language models in AI systems:

- **Does separating linguistic competence from factual knowledge reduce hallucinations and improve the accuracy of generated information?** By separating linguistic competence from internalized factual knowledge, we aim to mitigate instances where models produce inaccurate or speculative content based on outdated or erroneous information.
- **How effective is RAG in supplementing FLMs with accurate and up-to-date factual knowledge?** RAG techniques enable FLMs to retrieve relevant information from external sources during generation, potentially enhancing the models' factual accuracy by accessing the latest and most relevant data available.
- **How does the separation of linguistic competence and factual knowledge impact the explainability and transparency of FLMs?** This inquiry explores whether separating linguistic competence from factual knowledge enhances the model's ability to explain its reasoning process transparently, thereby improving trust and interpretability in AI-driven decision-making.
- **To what extent can FLMs retain comprehensive language understanding and reasoning capabilities without internal factual knowledge?** Assessing whether FLMs, despite not internalizing factual data, can maintain robust language understanding and reasoning capabilities necessary for complex AI tasks.

## Acknowledgments

This work has been funded by the scholarship (FPI-PRE2022-105603) from the Ministry of Science, Innovation and Universities of the Spanish Government. I am grateful to my thesis supervisors Arturo Montejo-Ráez and L. Alfonso Ureña-López for their guidance and help during the work done up to now.

## References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[7] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[8] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[9] P. W. Anderson, More is different: Broken symmetry and the nature of the hierarchical structure of science., Science 177 (1972) 393–396.

[10] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. arXiv:2206.07682.

[11] H. Chang, J. Park, S. Ye, S. Yang, Y. Seo, D.-S. Chang, M. Seo, How do large language models acquire factual knowledge during pretraining?, 2024. URL: https://arxiv.org/abs/2406.11813. arXiv:2406.11813.

[12] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. arXiv:2311.05232.

[13] W. Wang, B. Haddow, A. Birch, W. Peng, Assessing factual reliability of large language model knowledge, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 805–819. URL: https://aclanthology.org/2024.naacl-long.46. doi:10.18653/v1/2024.naacl-long.46.

[14] I. Garrido-Muñoz , A. Montejo-Ráez , F. Martínez-Santiago , L. A. Ureña-López , A survey on bias in deep nlp, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/7/3184. doi:10.3390/app11073184.

[15] P. Hacker, Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under eu law, Common market law review 55 (2018).

[16] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Communications of the ACM 63 (2019) 68–77.

[17] D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, D. Sculley, Google vizier: A service for black-box optimization, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 1487–1495. URL: https://doi.org/10.1145/3097983.3098043. doi:10.1145/3097983.3098043.