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

Large Language Models (LLMs) have demonstrated remarkable capabilities in language understanding and generation, driven by advancements in deep neural networks However, the current trend of developing increasingly larger models to enhance task competence comes at a significant cost, including a substantial carbon footprint with detrimental environmental consequences. Furthermore, these models often internalize vast amounts of factual knowledge, leading to issues such as hallucinations and the use of outdated information.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 separated from memorized factual knowledge and other cognitive skills in neural networks. We propose the development of "Fundamental Language Models" (FLMs), smaller, more efficient models focused on language understanding and reasoning. These FLMs will leverage external sources and tools, using techniques like Retrieval Augmented Generation (RAG), to access up-to-date factual knowledge, thereby potentially mitigating both environmental impact and factual inaccuracies. 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

Large Language Models (LLMs), developed through transformer-based architectures and trained on massive text corpora, have revolutionized natural language processing. These autoregressive models have demonstrated remarkable capabilities in understanding and generating human language, driving advancements in various artificial intelligence applications, including sophisticated conversational agents like GPT-4 [1] and Llama 3 [2].

However, the prevailing paradigm of scaling LLMs to achieve broader task competence presents significant limitations. The sheer size of these models demands substantial computational resources and energy, resulting in a considerable carbon footprint with detrimental environmental consequences [3]. Furthermore, these monolithic models are susceptible to generating factual inaccuracies, or "hallucinations" [4], which undermines user trust. Additionally, LLMs can inadvertently perpetuate and amplify societal biases present in their training data, raising critical ethical concerns [5].

To address these challenges, this research proposes a novel architectural paradigm centered on Fundamental Language Models (FLMs). The core idea is to develop smaller LLMs focused on mastering linguistic competence−−the ability to understand and generate natural language-−while minimizing the storage of factual knowledge within their parameters. We hypothesize that decoupling language understanding from world knowledge will lead to more efficient and controllable language engines. In this framework, FLMs would serve as the central reasoning unit within intelligent agents, leveraging external, specialized tools such as knowledge bases to access and process factual information relevant to specific tasks. This modular approach offers the potential to mitigate hallucinations and biases by grounding knowledge retrieval in external, curated sources.

Developing FLMs may involve techniques such as selectively pruning or fine-tuning existing LLM architectures to remove factual knowledge while retaining linguistic capabilities, or pretraining on

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.

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

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

carefully curated datasets that emphasize linguistic structure and reasoning. This research will investigate the feasibility and effectiveness of these approaches, exploring the trade-offs between model size, linguistic competence, and the effective utilization of external tools. By focusing on building agents with a clear separation of language processing and knowledge access, this work aims to contribute to the development of more sustainable, reliable, and ethically sound artificial intelligence systems.

The remainder of this paper is structured as follows: Section 2 provides a review of the relevant literature on LLM emergent abilities and limitations; Section 3 presents the research hypothesis and objectives. Section 4 details the methodology employed in this thesis and three experiments proposed, and Section 5 concludes by outlining specific research elements for discussion.

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 [6] or RoBERTa [7] use Masked Language Modeling (MLM) mainly while decoder or generative transformers like LLaMa [8], Mistral [9], and GPT [10] 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 [11]: *"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 [12]. Along with these properties, this semi-supervised pretraining technique allows LLMs to memorize lots of factual data [13] that, in some cases, may lead to problems such as hallucinations [14] 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 [15].

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 [16]. The problem has become a topic of interest beyond the field of AI algorithm research and is known as fairness [17] 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 [18]. 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 [19].

The study of "foundational" language models can help address bias and improve explainability by focusing on core linguistic competence, separate from stored factual knowledge. This approach aligns with long-standing linguistic debates, particularly Chomsky's distinction between internal (I-language) and external (E-language) systems. As Graffi [20] notes, viewing language as an internal cognitive system, rather than a socially embedded one, raises questions still relevant when interpreting LLM behavior. Recent empirical work supports this conceptual separation. Miaschi et al. [21] demonstrate that LLMs vary in their ability to generate sentences that follow explicit morpho-syntactic and syntactic constraints, highlighting clear limitations in linguistic control that are independent of factual recall. These findings underscore the value of disentangling linguistic competence from knowledge storage—central to the design of FLMs.

3. Hypothesis and Objectives

This research is guided by the central hypothesis that it is possible to develop Fundamental Language Models (FLMs): Small Language Models (SLMs) that primarily encode linguistic competence rather than extensive factual knowledge and other cognitive capabilities. We propose that these FLMs can effectively function as reasoning engines when coupled with external knowledge sources and tools, potentially leading to more efficient, reliable, and ethically sound AI systems.

With this framework in mind, the main objective of this research is to investigate the functioning of Large Language Models as fundamental language processing engines. To achieve this, the following secondary objectives have been defined:

1. Study the internal encoding of linguistic structures and their role in language understanding, independent of factual knowledge.
2. Decompose the capabilities of current language models to differentiate between core linguistic skills and those reliant on internalized knowledge or reasoning.
3. Develop and evaluate methods for enhancing language model capabilities by integrating external knowledge sources and tools for tasks beyond pure language processing.
4. Improve the explainability of AI task resolution by clearly separating the linguistic processing stage from the contributions of external modules and retrieved information.

4. Methodology

The following methodology aims to validate the hypothesis that linguistic competence in LLMs can be separated from other cognitive abilities, enabling the development of FLMs. The methodology comprises the following key stages:

1. Literature review and initial study: A comprehensive review of existing literature will be conducted to establish a strong foundation in current LLM techniques and advancements. This will involve identifying key resources, including publications in leading scientific forums such as AAAI, NeurIPS, and ACL, as well as relevant information from reference bulletins like PapersWithCode and The Batch.
2. Experimental design and evaluation: Rigorous experimental setups will be defined for each research objective. This includes specifying appropriate datasets and evaluation metrics. To ensure robust evaluation and benchmarking, we will actively participate in evaluation forums such as CLEF, SemEval, and IberLEF.
3. Study of internal encoding of knowledge: A series of experiments will be performed to analyze the internal representations of LLMs. This will involve comparing representations across different models to identify common patterns and structures related to linguistic competence encoding and other capabilities.
4. Decomposition of language model capabilities: This stage focuses on isolating and evaluating individual linguistic skills. Specific tasks and benchmarks will be designed or adapted to target lexical, grammatical, and semantic competencies. Controlled experiments will be conducted to assess model performance on these targeted tasks, allowing for a detailed analysis of each competency.
5. Enhancement through external tools: Methods for effectively connecting LLMs with external tools and knowledge bases will be developed and evaluated. Retrieval Augmented Generation (RAG) will be a key technique explored. Experiments will compare the performance of enhanced models against baseline models to quantify the benefits of such tools.
6. Dissemination of findings: Research findings will be prepared and disseminated through publications in high-impact journals and presentations at international conferences.

4.1. Proposed experiments

This section outlines three key experiments currently in development. These experiments are designed to: 1) provide evidence supporting the FLM concept; 2) enable rigorous evaluation of linguistic competence, a core feature of FLMs; and 3) demonstrate the potential of smaller models to perform complex tasks when equipped with appropriate external tools and sufficient linguistic competence.

Study of internal encoding of knowledge: This experiment aims to analyze the scaling behavior of linguistic competence in LLMs relative to reasoning and factual recall. Specialized benchmarks will be used to assess each competency:

- Linguistic competence: Assessed as a composite of lexical, grammatical, and semantic competencies. The hypothesis is that this competence will saturate at smaller model sizes compared to reasoning and factual recall.
- External factual knowledge: Evaluated using question-answering tasks based on provided text chunks. These tasks require reasoning and inference over the given factual information.
- Internal factual knowledge: Evaluated using question-answering tasks without context. Success in these tasks relies on the model's memorized factual knowledge.

Lexical competence evaluation: Current benchmarks often lack the ability to effectively evaluate LLMs' lexical competence across diverse languages and specialized domains. To address this gap, we are developing a novel, automated method for evaluating lexical competence in a multilingual and domain-specific manner. This method will be used to assess the preservation of lexical competence in FLMs as other capabilities are reduced.

Enhancement through external tools: This experiment explores the use of external reasoning tools to enhance the performance of small language models (SLMs) on logic-based benchmarks, such as ZebraLogic [22]. The objective is to determine whether reasoning capabilities, similar to factual knowledge in RAG systems, can be effectively delegated to external tools, or whether they represent a more fundamental capability that FLMs must retain internally.

5. Research elements proposed for discussion

The following key discussion points are proposed to facilitate a comprehensive exploration of the potential benefits, limitations, and broader implications of the FLM paradigm within AI systems. These points directly relate to the central hypothesis and objectives of this research:

- **To what extent does separating linguistic competence from the storage of factual knowledge mitigate the generation of hallucinations and enhance the reliability of information generated by LLMs?** This point prompts a critical evaluation of whether the FLM approach effectively reduces the occurrence of inaccurate or speculative content arising from outdated or erroneous internalized information.
- **How effectively can external tools and retrieval mechanisms, such as Retrieval Augmented Generation (RAG), provide FLMs with accurate, up-to-date information and empower them to perform complex, knowledge-intensive tasks?** This discussion will center on the capabilities and limitations of RAG and other external knowledge integration methods in supporting FLMs' access to and utilization of current and relevant information.
- **How does the separation of linguistic competence from other cognitive skills impact the transparency of FLMs' processing of information and task execution?** 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 maintain comprehensive language understanding and generation capabilities when completely decoupled from internalized knowledge, relying**

solely on external resources? This discussion will critically assess the potential trade-offs between knowledge externalization and the preservation of essential linguistic competencies required for effective performance in various AI applications.

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.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT and Grammarly in order to: Grammar and spelling check, translate and reword. After using this tools, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] OpenAI, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[2] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[3] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, U. Sauerland, Risks and benefits of large language models for the environment, Environmental science & technology 57 (2023) 3464–3466.

[4] 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, ACM Trans. Inf. Syst. 43 (2025). URL: https://doi.org/10.1145/3703155. doi:10.1145/3703155.

[5] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics 50 (2024) 1097–1179. URL: https://doi.org/10.1162/coli_a_00524. doi:10.1162/coli_a_00524.

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

[20] G. Graffi, Between linguistics and philosophy of language: The debate on chomsky's notion of "knowledge of language", Cahiers du Centre de Linguistique et des Sciences du Langage (2018) 39–58.

[21] A. Miaschi, F. Dell'Orletta, G. Venturi, Evaluating large language models via linguistic profiling, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 2835–2848. URL: https://aclanthology.org/2024.emnlp-main.166/. doi:10.18653/v1/2024.emnlp-main.166.

[22] B. Y. Lin, R. L. Bras, K. Richardson, A. Sabharwal, R. Poovendran, P. Clark, Y. Choi, Zebralogic: On the scaling limits of llms for logical reasoning, 2025. URL: https://arxiv.org/abs/2502.01100. arXiv:2502.01100.