Literature Review on Knowledge Injection and Hallucinations in LLMs

José Matos[1] , **Vasco Rodrigues**[1]
[1]University of Coimbra, Department of Informatics Engineering
{josematos, vascorodrigues}@student.dei.uc.pt,

1 Introduction

As Large Language Models (LLMs) continue to grow in scale and capability, their real-world utility is often constrained by two fundamental challenges: their static training knowledge base and their tendency to generate non-factual statements, or *hallucinations*. The former prevents them from incorporating new or domain-specific information, while the latter undermines their reliability, particularly in specialized fields like science and medicine.

In response, the field has developed **Knowledge Injection (KI)** techniques, by developing new approaches or applying known ones like Retrieval-Augmented-Generation (RAG) [Lewis *et al.*, 2021] in order to ground LLMs answers, with the primary goal of mitigating hallucinations and improving accuracy across domains.

However, this created a new challenge. If KI is the solution to hallucinations, how can its effectiveness be measured? This has led to a second line of research focusing on **Hallucination Quantification**, with the objective of creating robust approaches to measure factuality in LLM outputs.

These two fields are intrinsically linked and require each other to keep progressing. In Section 2 we discuss recent developments in KI techniques for LLMs and in Section 3 we look into how hallucinations are categorized and how they are quantified. Then, in Section 4, we frame the knowledge gathered from this review in the context of our project proposal.

2 Knowledge Injection into LLMs

The static nature of LLMs has led to the development of a wide range of knowledge injection techniques, aiming to enhance factual accuracy, domain specialization, and adaptability. These methods can be broadly categorized into static, dynamic, hybrid, and model editing-based paradigms. In the following subsections we review recent research that defines and compares these approaches, highlighting their empirical findings, limitations, and emerging directions. These methods are summarized in Table 1.

2.1 Static Knowledge Injection: Fine-Tuning

The study [Ovadia *et al.*, 2024] provides a direct empirical comparison between two paradigms: Unsupervised Fine-Tuning (FT) and Retrieval-Augmented Generation (RAG).

The key findings position RAG as a more robust and effective strategy for factual knowledge integration across multiple domains, particularly on MMLU subtasks. RAG consistently outperformed FT, especially when injecting entirely new knowledge, as demonstrated by the *Current Events* benchmark. Conversely, FT exhibited instability, strong sensitivity to hyperparameters, and in some cases, even degraded the base model's performance.

The study also identified a strong monotonic correlation between the number of paraphrased data augmentations used during FT and resulting accuracy, suggesting that mere exposure is insufficient, knowledge must be reinforced through varied formulations. Identified limitations include:

- FT performance is highly volatile and hyperparameter-dependent, lacking the consistency of retrieval-based approaches;
- Conclusions were primarily drawn from smaller open-source models, leaving their applicability to larger, state-of-the-art models uncertain;
- Reliance on Wikipedia as the sole knowledge source, reducing domain diversity;
- Focus exclusively on unsupervised fine-tuning, leaving hybrid approaches unexplored;
- Dependence on QA-based accuracy metrics, limiting understanding of internal knowledge representation.

2.2 Retrieval-Augmented Generation (RAG) and Parametric Retrieval-Augmented Generation (PRAG)

RAG [Lewis *et al.*, 2021] operates as an in-context knowledge injection method that retrieves relevant external knowledge during inference. While highly effective for factual recall, it introduces computational overhead and potential degradation in complex reasoning tasks. This occurs because RAG operates primarily at the input level, whereas LLMs store knowledge internally within their parameters.

Despite these limitations, RAG remains a foundational approach, enabling continuous access to up-to-date information without retraining the model. It represents a practical trade-off between efficiency and adaptability, offering a flexible alternative to static parameter updates.

To bridge the gap between purely retrieval-based and parameter-level integration, the paper [Su *et al.*, 2025] introduces Parametric RAG (PRAG), a hybrid approach that integrates external knowledge directly into the Feed-Forward Networks (FFNs) of the LLM.

This integration is achieved by parameterizing varied formulations of the same document (e.g., rewritings, Q&A pairs), allowing the model to internalize external information more efficiently.

The key findings position PRAG as a superior tool for in-parameter knowledge injection, improving the model's capacity to reason and recall without the inference overhead typical of RAG. Furthermore, the study concludes that PRAG's benefits scale with model size, as larger LLMs are better able to leverage the internalized document knowledge. Limitations include:

- Parametric document representations are substantially larger than simple text;
- Parameterized documents are tightly coupled to specific LLM architectures, limiting portability.

Importantly, PRAG does not fully replace RAG; instead, both can be used **in conjunction**, combining dynamic retrieval and parametric reinforcement for optimal performance.

2.3 Knowledge Model Editing (KME)

Given the high cost and instability of continual fine-tuning, Knowledge Model Editing (KME) has emerged as a promising direction for updating models. The survey [Wang *et al.*, 2025] introduces a taxonomy that classifies KME strategies into three main categories:

- KME – External Memorization: leverages external memory or additional parameters to store new knowledge without modifying the original pre-trained weights;
- KME – Global Optimization: injects new knowledge directly into the model's parameters through constrained fine-tuning designed to preserve unrelated knowledge;
- KME – Local Modification: targets specific weights responsible for certain knowledge, allowing highly localized edits.

These approaches address the challenge of updating LLMs efficiently while preserving general performance. However, the field faces several fundamental challenges:

- Locality vs. Generality Trade-off: improving one often diminishes the other;
- Editing at Scale: coherence degrades when performing numerous sequential edits;
- Unstructured Editing: performance drops when dealing with real-world text instead of structured triples;
- Continual Editing: most methods lack online, real-time update capabilities required for real-world deployment.

3 Hallucinations in LLMs

Hallucinations occur when LLMs generate content that is not factually accurate, or unfaithful to the source provided within their context [Farquhar *et al.*, 2024]. While LLMs possess extensive factual knowledge, mainly due to their immense pretraining corpora, is is actually limited due to the impossibility of memorizing all factual knowledge encountered during training, and to the boundary that is intrinsic to the data itself. Consequently, when these models are faced with tasks and information that falls out of their limited knowledge range, for example in domain-specific areas such as science and medicine, they may exhibit pronounced hallucinations and fabricate inaccurate factual information [Huang *et al.*, 2025].

In addition, the static nature of LLMs poses a challenge to the dynamic nature of information. Once an LLM is trained, it's internal knowledge can never be updated, which may lead the model to produce false or outdated facts. Many KI techniques have been proposed to help mitigate hallucinations by grounding the LLM in external, verifiable and up to date knowledge. RAG, which has been addressed in the previous section, has proved its strength in reducing hallucinations by filling knowledge gaps, while being applicable across domains [Huang *et al.*, 2025].

3.1 How Can Hallucinations be Quantified?

To understand if hallucinations are being mitigated, there is a need to measure and quantify them.

Early and simple techniques to measure hallucinations were internal, or in other words, had no use of external knowledge to verify claims. A proposed method utilized entropy-based uncertainty estimators to detect hallucinations consisting of "arbitrary and incorrect generations" [Farquhar *et al.*, 2024], making use of classic machine learning methodologies and applying them to modern LLMs and natural language generation.

In contrast to the simple and transparent nature of estimating uncertainty using entropy, techniques emerged which leveraged the *LLM-as-a-Judge* approach, where a powerful LLM is prompted to grade a completion for factuality, serving as a black-box verifier. This is possible due to the strong instruction-following ability of fine-tuned LLMs, that when provided with concrete guidelines and rules, can effectively assess this problem [Huang *et al.*, 2025]. This technique offers a reduced reliance on human annotation and verification and enables scalability, however can lead to accuracy or bias problems if the judge model used in not properly aligned [Wang *et al.*, 2023].

Currently, the most reliable methods of evaluating hallucinations in domain-specific areas involve comparing the LLMs answer against a source of truth. This mechanism is at the core of modern KI and RAG systems. A recent survey on LLM hallucinations as proposed a comprehensive taxon-

Table 1: Comparison of Knowledge Injection Paradigms

Paradigm	Mechanism	Advantages	Limitations
Static (FT)	Embeds knowledge directly into model parameters	Persistent integration	Computationally expensive; risk of catastrophic forgetting
Dynamic (RAG)	Retrieves external knowledge at inference	Up-to-date, flexible	Latency and computational overhead
Hybrid (PRAG)	Parameterizes external knowledge into internal networks	Efficient recall and reasoning	Larger memory footprint; model coupling
KME	Edits or augments model memory and parameters post-training	Targeted updates; preserves base knowledge	Scalability and consistency issues

omy, separating them into two separate classes: *Factuality* and *Faithfulness* hallucinations [Huang *et al.*, 2025]. The first occurs when an LLM generates content that directly conflict with, or cannot be verifier against real world facts [Wang *et al.*, 2025]. Faithfulness hallucinations occur when an LLM fails to align itself with user instructions, provided context, or internal logic [Es *et al.*, 2025].

3.2 Frameworks for Quantifying Hallucinations

In order to operationalize the concepts presented in the previous subsection and directly address faithfulness and factuality hallucinations in LLMs, multiple frameworks have been proposed.

FActScore [Min *et al.*, 2023] is an evaluation approach for LLMs that represents the percentage of facts or pieces of information directly supported by a source of knowledge. This technique involves breaking the output of the LLM into a group of "atomic facts", with each containing a separate piece of information, and then cross-referencing the gathered pieces of information to a source of truth, directly evaluating factuality hallucinations within the completion.

Another proposed framework is *Ragas* [Es *et al.*, 2025], which focuses on a reference-free evaluation of RAG pipelines. They form their evaluation based on three aspects: Answer Relevance, Context Relevance and Faithfulness. The latter refers to the idea that the answer should be grounded in the given context, and is the essential metric behind avoiding and detecting hallucinations. Their approach to measure faithfulness can be summarized as follows: An answer is as faithful to the context if the claims made in the answer can be inferred from it. An LLM is utilized to extract set of statements, decomposing longer sentences into shorter and verifiable assertions. Faithfulness score is then computed by checking if claims made in the answer were inferrable from the provided context.

The framework from which we draw the most inspiration in this work is REFCHECKER [Hu *et al.*, 2024], a knowledge-centric approach to hallucination detection and quantification. The approach relies on two components, an extractor and a checker. The first extracts claim-triplets from the LLMs output, while the second evaluates each triplet against a ground truth reference. Another great characteristic of REFCHECKER is that in opposition to other approaches which only differentiate between factual and non-factual claims, it also takes into account claims that cannot be verified through the contextual reference, making it a valuable tool in the field of science related tasks, such as the ones we will be experimenting with in our project.

4 Conclusion: Positioning our Project Proposal

Throughout this review we surveyed the current landscape of KI and hallucination quantification in LLMs. The literature shows a clear preference for RAG over traditional fine-tuning for integrating new knowledge, and hybrid methods like PRAG offer a promising way to streamline this knowledge without the inference overhead imposed by RAG systems.

At a similar rate, the field has developed sophisticated methods to quantify hallucinations, moving from internal uncertainty based measurments to more robust and knowledge-centric approaches that test the outputs of LLMs against a golden source of truth.

Utilizing all this knowledge, we can identify a gap in the comparative benchmarking of different KI techniques, specifically when it comes to the assessment of hallucinations. While some studies have compared and benchmarked KI methods, we propose the utilization of a KG-based framework like REFCHECKER to benchmark the effect different types of KI can have on hallucinations. Furthermore, we believe this problem can be particularly pronounced in specialized domains, and such a benchmark can be a contribution towards better understanding of KI techniques in a domain like science. Our project will aim to address this gap by:

- Comparing RAG, Parametric RAG, and other KI techniques in the scientific domain using the SciER dataset [Zhang *et al.*, 2024].
- Utilizing a "golden" KG as the ground truth.
- Applying a triple-based approach like REFCHECKER to quantify hallucinations, allowing for a comparison between how the different KI methods produce faithful and contextually verifiable knowledge.

References

[Es *et al.*, 2025] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated Evaluation of Retrieval Augmented Generation, April 2025.

[Farquhar *et al.*, 2024] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024.

[Hu *et al.*, 2024] Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-Centric Hallucination Detection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Huang *et al.*, 2025] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55, March 2025.

[Lewis *et al.*, 2021] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, April 2021.

[Min *et al.*, 2023] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, October 2023.

[Ovadia *et al.*, 2024] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs, January 2024. arXiv:2312.05934 [cs].

[Su *et al.*, 2025] Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric Retrieval Augmented Generation, January 2025. arXiv:2501.15915 [cs].

[Wang *et al.*, 2023] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity, December 2023.

[Wang *et al.*, 2025] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge Editing for Large Language Models: A Survey. *ACM Computing Surveys*, 57(3):1–37, March 2025.

[Zhang *et al.*, 2024] Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. SciER: An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents, October 2024. arXiv:2410.21155 [cs].