Vol. 3, No. 1, pp.: 1-4, 2025

Hallucination Rates Across Domain-Specific LLM Fine-Tuning: A Systematic Evaluation

Mehmet Yilmaz[1,*], **Preethi Rajagopalan**[2], **Anton Ivashkin**[3]

[1] Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, 06800

[2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, 15213

[3] Skolkovo Institute of Science and Technology, Moscow, Russia, 143026

[*] *myilmaz@ceng.metu.edu.tr*

Article Information	
Received	9 December 2024
Accepted	28 March 2025
DOI	https://doi.org/10.63646/datamind.2025.030101

Abstract

Hallucination — the generation of factually incorrect, fabricated, or internally inconsistent text by large language models — is one of the most practically consequential failure modes in LLM deployment. Fine-tuning on domain-specific data is widely used to improve LLM performance in specialised domains, but the relationship between fine-tuning and hallucination rates is poorly characterised. This paper presents a systematic evaluation of hallucination rates before and after domain-specific fine-tuning across four domains (biomedical, legal, financial, and software engineering) and three base models (Llama-3.1-8B, Mistral-7B-v0.3, and Qwen2.5-7B). We use a three-component hallucination taxonomy — factual hallucination, entity hallucination, and reasoning hallucination — and evaluate each component using a combination of automated fact-checking pipelines and expert annotation. Counter to the common assumption that fine-tuning on domain data reduces hallucination by reinforcing factual associations, we find that fine-tuning on high-quality but narrow domain corpora frequently increases entity and reasoning hallucination rates even when factual hallucination rates decrease. We link this phenomenon to a degradation in world-model breadth during fine-tuning and provide evidence that the effect is modulated by the ratio of domain-specific to general knowledge in the fine-tuning data mix.

Keywords: ***hallucination; LLM fine-tuning; factual consistency; domain adaptation; biomedical NLP; entity grounding; evaluation***

1. Introduction

The business case for domain-specific fine-tuning of large language models is clear: a model that has been trained on legal documents performs better on legal tasks; a model trained on clinical notes performs better on clinical documentation tasks. The improvement is real, replicable, and commercially valuable. What is less clear — and less frequently examined — is the effect of that fine-tuning on hallucination rates. This is not a trivial question. A model that performs better on in-domain accuracy benchmarks while simultaneously generating more plausible-sounding but incorrect information about adjacent topics is a model that may create more harm than benefit in high-stakes deployment contexts.

The hallucination problem in fine-tuned LLMs is compounded by a measurement challenge. Standard benchmarks for domain performance (accuracy on classification tasks, F1 on extraction,

ROUGE on summarisation) do not capture hallucination. Fine-tuned models can score higher on these benchmarks than base models while producing more hallucinated output in open-ended generation. The improvement is real in one dimension; the degradation is real in another. This paper attempts to measure both dimensions simultaneously.

2. Hallucination Taxonomy

We use a three-component taxonomy designed for domain-specific evaluation. Factual hallucination refers to claims about world facts that are objectively false — a model stating that a specific drug's half-life is 4 hours when it is 12 hours, or that a legal precedent was established in a case that was not relevant to the stated principle. Entity hallucination refers to the fabrication or misattribution of named entities — inventing the names of drugs, cases, companies, or code libraries that do not exist or attributing properties to the wrong entity. Reasoning hallucination refers to conclusions that do not follow from stated premises, even when the individual premises are accurate — an LLM that correctly identifies two facts but draws a causal inference that the facts do not support.

3. Experimental Design

3.1 Models and Fine-tuning

We fine-tune three 7–8B parameter instruction-tuned models using QLoRA (4-bit quantisation, r=64, alpha=128) on domain-specific corpora: PubMed abstracts (biomedical, 200K samples), EUR-Lex directives (legal, 80K), SEC 10-K filings (financial, 120K), and GitHub issues and pull requests (software engineering, 160K). Fine-tuning is run for 3 epochs on a single NVIDIA A100 80GB. We evaluate hallucination rates at both zero-shot (base model) and after fine-tuning.

3.2 Evaluation Protocol

For factual and entity hallucination, we use a pipeline combining named entity recognition (spaCy + domain-specific entity classifiers), external knowledge base verification (Wikipedia, PubMed, OpenCorporates APIs), and a secondary LLM judge (GPT-4o with a structured verification prompt). For reasoning hallucination, we use a chain-of-thought consistency test: the model is presented with the same reasoning task twice with scrambled premises, and inconsistency between conclusions is flagged for expert review. Expert annotation was performed by two domain specialists per domain, with inter-annotator agreement $\kappa > 0.78$ across all categories.

Table 1. *Hallucination rates (%) before and after domain fine-tuning, by hallucination type. Lower values indicate fewer hallucinations. Arrows indicate direction of change after fine-tuning.*

Domain / Model	Factual (base)	Factual (FT)	Entity (base)	Entity (FT)	Reasoning (base)	Reasoning (FT)
Biomedical / Llama	11.2	7.4 ↓	8.9	13.1 ↑	9.3	11.8 ↑
Biomedical / Mistral	10.8	7.1 ↓	9.2	14.3 ↑	8.9	12.2 ↑
Legal / Llama	13.7	9.2 ↓	11.4	16.8 ↑	12.1	14.7 ↑
Financial / Llama	12.3	8.1 ↓	10.6	12.9 ↑	11.4	13.1 ↑
SWE / Llama	9.8	7.6 ↓	14.2	17.3 ↑	10.2	11.9 ↑

FT = fine-tuned model. All values are percentages of evaluated outputs containing at least one hallucination of the respective type. Each cell represents evaluation over 500 generated outputs. Arrows indicate directional change; magnitude varies — see full results in supplementary material.

[Figure 1 — Relationship between domain data mixture ratio (x-axis, proportion of domain-specific to general knowledge in fine-tuning data) and total hallucination rate change after fine-tuning (y-axis, positive = more hallucinations). The curve shows a U-shaped relationship]

Figure 1. *Relationship between domain data mixture ratio (x-axis, proportion of domain-specific to general knowledge in fine-tuning data) and total hallucination rate change after fine-tuning (y-axis, positive = more hallucinations). The curve shows a U-shaped relationship: mixtures below 0.2 and above 0.8 domain-specific data produce higher hallucination increases than mixtures in the 0.3–0.6 range. Error bars show ± 1 SD across three base models.*

4. Key Findings and Mechanism

The central finding — that domain fine-tuning reduces factual hallucination while increasing entity and reasoning hallucination — is consistent across all three base models and all four domains (Table 1). The magnitude differs by domain and model, but the directional pattern is strikingly consistent. We interpret this through the lens of knowledge acquisition: fine-tuning on domain data reinforces the model's associations between domain-specific facts and their correct values (reducing factual error) but simultaneously increases confidence in domain-adjacent generations that are less well-grounded — a form of learned overconfidence.

The data mixture analysis (Figure 1) provides evidence for a practical mitigation. When fine-tuning data mixes domain-specific and general-knowledge text in a roughly equal proportion (0.4–0.6 domain ratio), the increase in entity and reasoning hallucination is significantly attenuated without substantial loss of factual accuracy improvement. This is consistent with the hypothesis that the entity and reasoning hallucination increase is driven by the erosion of the model's broader world model during narrow-domain fine-tuning, and that maintaining access to general knowledge in the fine-tuning mix preserves the contextualisation that constrains hallucinated inference.

5. Discussion and Implications

For practitioners deploying domain-fine-tuned LLMs in high-stakes settings, these findings have direct implications. First, benchmark performance on in-domain accuracy tasks is an incomplete indicator of deployment safety — separate hallucination evaluation is necessary and should include entity and reasoning components, not just factual accuracy. Second, fine-tuning data composition matters beyond total volume: deliberately mixing general-knowledge text into domain fine-tuning data provides measurable hallucination mitigation at modest cost to domain performance. Third, models fine-tuned on narrow technical corpora (legal, biomedical) show the largest entity hallucination increases and warrant the most careful monitoring when deployed in open-ended generation contexts.

6. Conclusion

Domain-specific fine-tuning of LLMs is not a uniformly beneficial intervention from a hallucination perspective. It trades one hallucination type for others — improving factual precision at the cost of entity grounding and reasoning coherence. Practitioners who evaluate fine-tuned models only on in-domain benchmarks will systematically underestimate the hallucination risk of their deployed systems. We recommend multi-component hallucination evaluation as a standard component of domain fine-tuning validation pipelines. All evaluation code and annotation guidelines are available at https://github.com/datamind-papers/hallucination-bench.

References

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. ACL 2020, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. ACL 2023, 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., & El Sayed, W. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825. https://doi.org/10.48550/arXiv.2310.06825

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W. T., Koh, P. W., & Zettlemoyer, L. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. EMNLP 2023, 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. Transactions of the Association for Computational Linguistics, 10, 178–206. https://doi.org/10.1162/tacl_a_00454

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. ICLR 2022. https://openreview.net/forum?id=gEZrGCozdqR

Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., & Shoham, Y. (2024). Generating benchmarks for factuality evaluation of language models. EACL 2024, 281–296. https://aclanthology.org/2024.eacl-long.17

Wang, Y., Ma, X., Chen, W., Liang, T., Zhong, W., Shang, J., & Liu, Z. (2023). Self-consistency improves chain of thought reasoning in language models. ICLR 2023. https://openreview.net/forum?id=1PL1NIMMrw