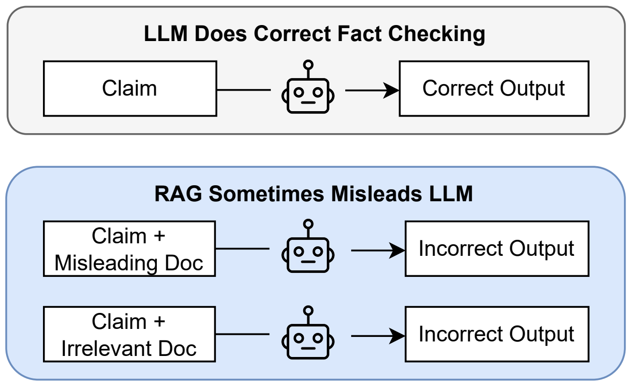
When Retrieval Misleads: Exploring Vulnerabilities in RAG

Sahej Agarwal[1], Jundong Xu[1,2], Ruiwen Zhou[1], Sahajpreet Singh[1,2]

[1]NUS School of Computing, [2]NUS CTIC

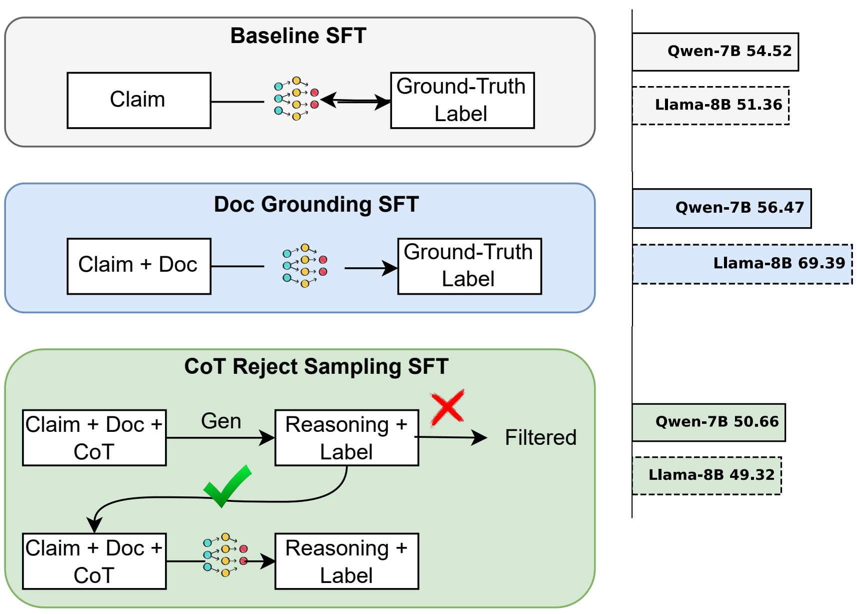
Introduction

- Retrieval-Augmented Generation (RAG) has succeeded in knowledge-intensive tasks as it provides LLMs with external knowledge
- However, existing works show that RAG can cause increased hallucination especially when the retrieved data involves counter intuitive information
- Therefore, we aim to study how different prompting and training techniques like chain-of-thought, retrieval-augmented fine-tuning, etc. can help mitigate these issues in small size language models

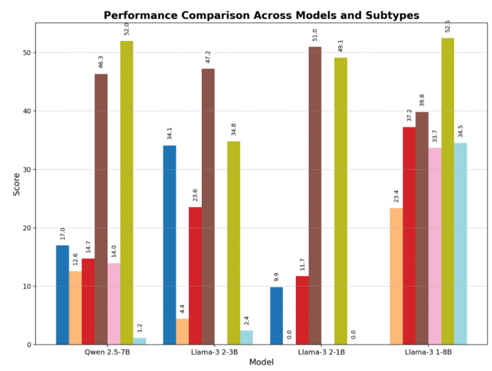


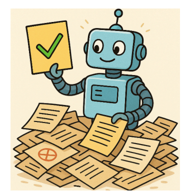
Experiments

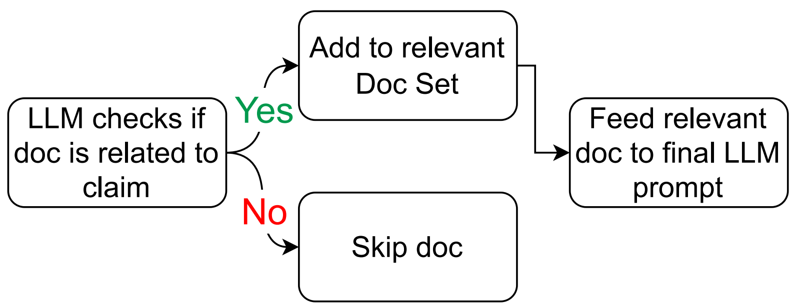
Supervised Fine-Tuning



Document Filtering







Analysis and Findings

✅ Doc-grounded SFT – Provides strong factual context, enabling clear claim–evidence mapping and reducing uncertainty.

❌ CoT w/ reject sampling – Retains noisy or shallow reasoning, leading to overfitting and weak factual grounding.

✅ All models – Can correctly classify documents.

❌ Baselines – Show no consistent trends with model size.

✅ Smaller models – Benefit from no-document setups or document filtering.

✅ Larger models – Gain from n-shot prompting and richer contextual understanding.

💡 Conclusion

Fine-tuning (Doc-grounded SFT) is the best option if good infrastructure is available.

For smaller models, use document filtering or stronger retrieval methods.
Larger models benefit most from n-shot prompting strategies.

Key Takeaway: If the document retrieval is not very accurate, then the usage of small models is highly unrecommended

Zeng, L., Gupta, R., Motwani, D., Yang, D., & Zhang, Y. (2025). Worse than zero-shot? a fact-checking dataset for evaluating the robustness of rag against misleading retrievals. *arXiv preprint arXiv:2502.16101*. NeurIPS 2025.