LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

BIOMEDICAL HALLUCINATION DETECTION OF LLMS USING MED-HALT AND HALOSCOPE FRAMEWORKS

Idrissa Mahamoudou Dicko
Supervisor: Dr. Nona Naderi
Idrissa.dicko@universite-paris-saclay.fr

Github Link

MOTIVATION

Large Language Models (LLMs) are increasingly used in biomedical question answering and clinical education; however, they can generate fabricated information, known as hallucinations, which can affect decision making. We evaluate two complementary frameworks to detect such errors.

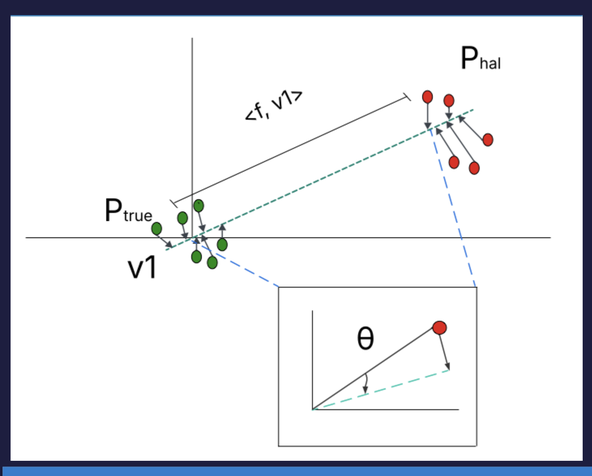
FRAMEWORKS

MedHalt

Reasoning Hallucination Test (RHT): Fake Question (FQT) and None-of-the-above (NOTA)

Haloscope

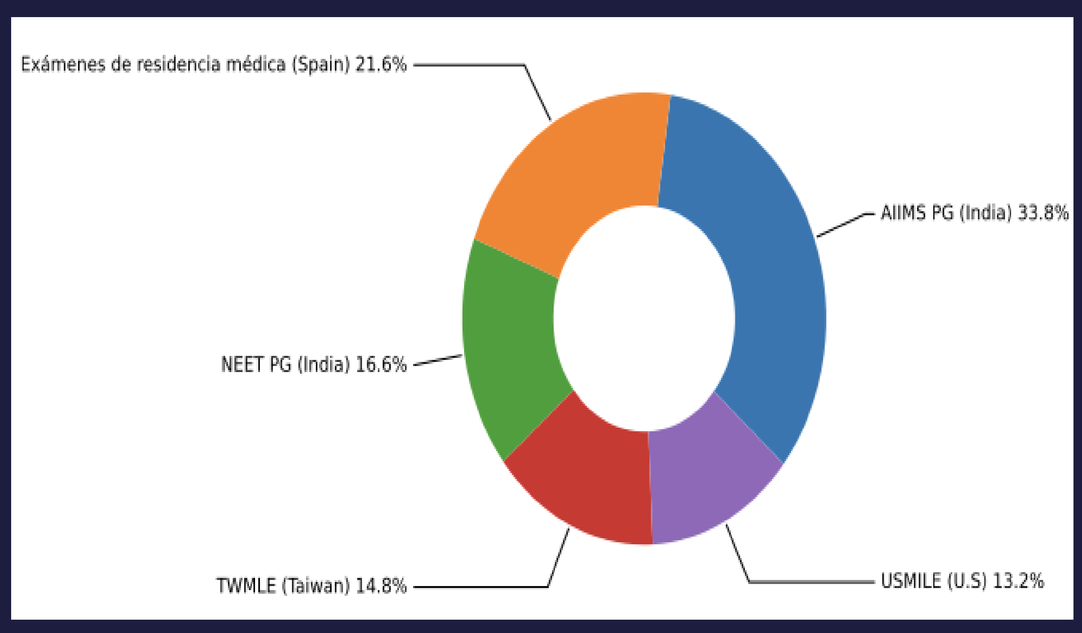
Learns low-dimensional hallucinations subspace (SVD) from unlabeled generations.



Illustrative2D proection showing separation between truthful vs. hallucinated samples

METHODS

Models: Llama-2-7B-chat, Mistral-7B.
Datasets: MEDMCQA, HEADQA, MEDQA-USMLE, MEDQA-TWMLE, TruthfulQA.
Implementation: robust CSV serialization; avoid hard stop tokens to prevent truncation.

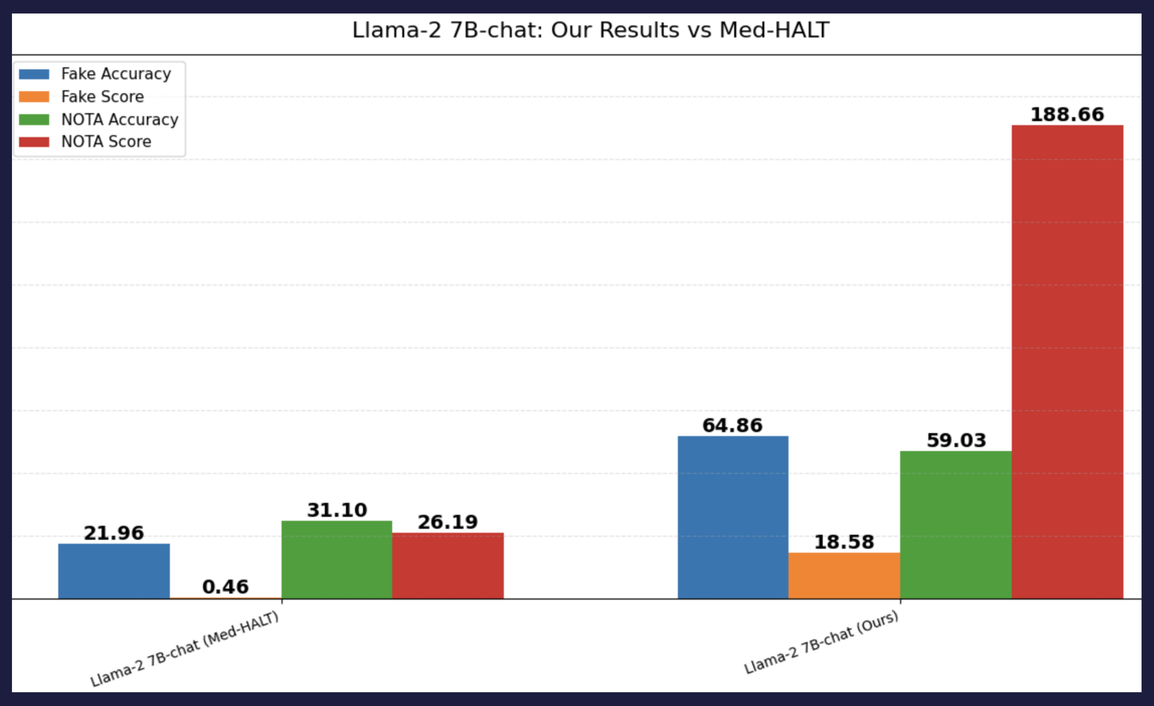


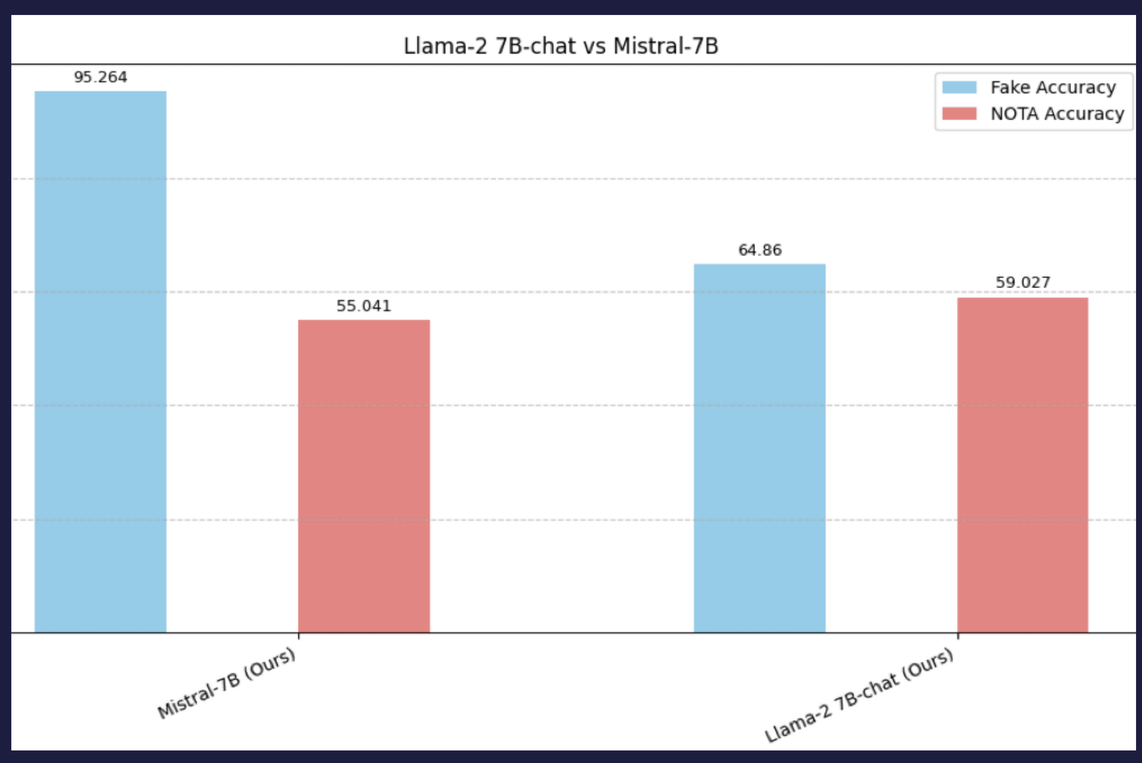
RESULTS

Reproducing HaloScope on TruthfulQA showed strong seed sensitivity: seed=42 gave lower AUROC, while seed=41 yielded ≈78.6%, matching the paper.

Model	Fake		NOTA	
	Acc	PW	Acc	PW
Llama-2-7B (Med-HALT)	21.96	0.46	31.10	26.19
Llama-2-7B (Ours)	64.86	18.58	59.03	188.66
Mistral-7B (Ours)	95.26	18.58	55.04	188.66

Observation: Gains are driven by stable parsing and unrestricted decoding.





References
Du, X., Xiao, C., Li, Y.: HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection.
Pal, A., Umapathi, L.K., Sankarasubbu, M.: Med-HALT: Medical Domain Hallucination Test for Large Language Models. In: Proceedings of the 27th Conference on Computational Natural Language Learning