Representation-based Broad Hallucination Detectors Fail to Generalize Out of Distribution

Zuzanna Dubanowska, Maciej Żelaszczyk, Michał Brzozowski, Paolo Mandica, Michał Karpowicz
Samsung AI Center, Warsaw

ML in PL CONFERENCE 2025

Accepted to EMNLP Findings 2025
EMNLP 2025 Suzhou, China | 中国苏州

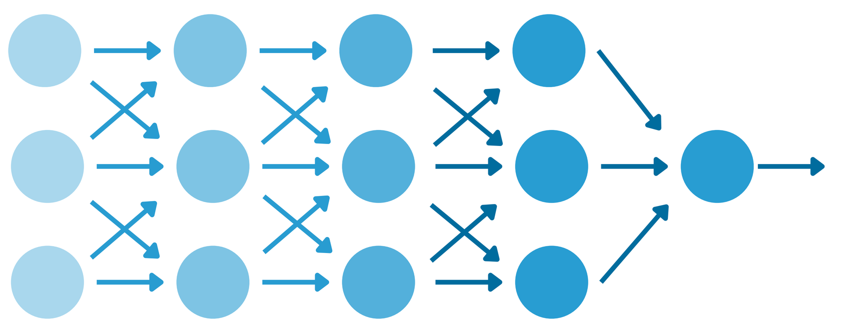
Samsung Research

*Is it possible to classify a response from an LLM as **hallucinatory** based on its internal states?*

Internals-based prediction

1 **The LLM generates a response to a prompt.**

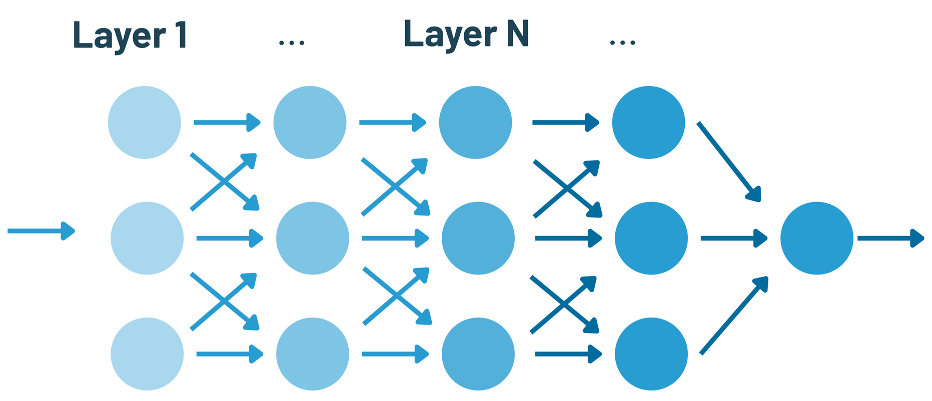
What's a hallucination ?



Don't know;)

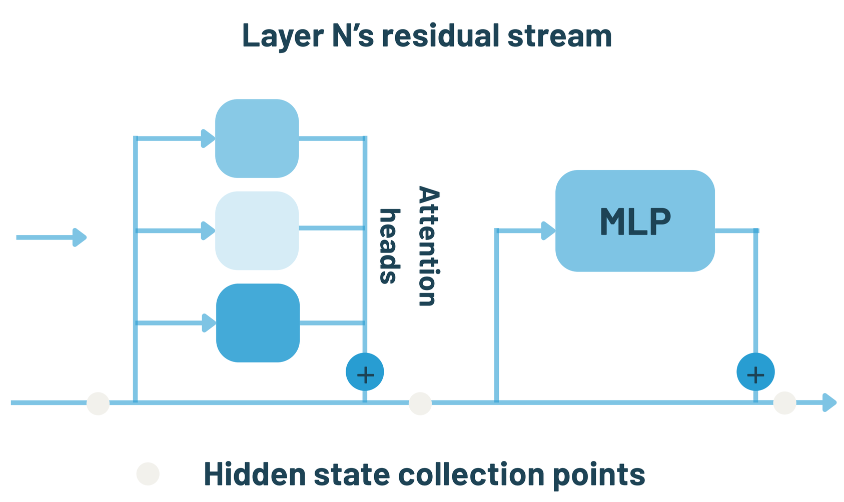
2 **The prompt and the response is fed to the LLM for internal state collection.**

What's a hallucination? Don't know ;)

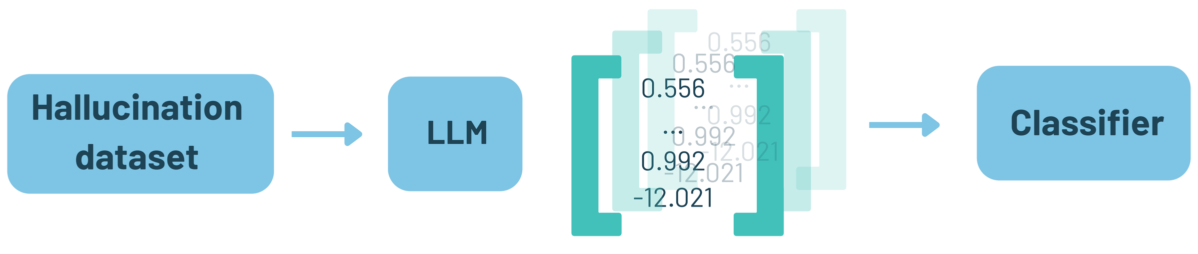


3 **The hidden states of the last token is collected from all layer's residual stream at entry to layer, post-attention and post-MLP.**

What's a hallucination? Don't know ;)



4 **The hidden state vectors are collected a dataset of LLM generations labelled as hallucinatory / non-hallucinatory.**



e.g. hidden states of layer N, post-attention

5 **A classifier is trained to predict hallucinations from hidden states.**

Either on raw activations or processed.

SoTA

We present SoTA performance on RAGTruth combined with other representation-based hallucination detection methods.

Model	AUC	Prec.	Rec.	F1
SAPLMA	**0.749**	0.630	**0.779**	0.648
ReDeEP	0.732	**0.722**	0.677	**0.699**
SEP	0.714	0.701	0.748	0.663
ITI	0.716	0.612	0.542	0.675

ReDeEP proxy the amount of information models integrate from context and parametric knowledge when generating a reply based on internal activations.

They leverage these scores to predict hallucinatory content.

SAPLMA hypothesize that the LLM posesses some internal notion of truth.

They leverage raw model activations and train a simple fully-connected neural network classifier to predict whether a sentence was truthful or not.

Key results

This table shows classification results on RAGTruth dataset (LLaMa 2 7B). LR, a simple classifier, outperforms SoTA. We investigate whether classifiers might exploit spurious correlations in the dataset structure as RAGTruth has three sub-tasks which are structurally different.

We create a naive classifier which assigns hallucinated label to all examples from data-to-text RAGTruth subtask which is structurally most distinct from the rest of the dataset as it's in JSON format. This classifier is on par with SoTA, further enforcing possible existence of spurious correlations in classifiers' features.

Classifier	AUC	PCC
ReDeEP	0.7325	0.3979
naïve	0.7119	**0.4494**
SAPLMA	0.7037	0.3188
Logistic Regression (LR)	**0.7951**	0.4103
LR + SAE features	0.7105	0.3282

	QA				D2T				Summ.				Overall			
Method	AUC	Prec.	Rec.	F1	AUC	Prec.	Rec.	F1	AUC	Prec.	Rec.	F1	AUC	Prec.	Rec.	F1
ReDeEP	0.636	0.453	0.462	0.457	0.395	0.793	0.748	**0.770**	0.577	0.484	0.294	0.366	0.732	0.722	0.677	0.699
Logistic Reg.	0.690	**0.691**	0.690	**0.690**	**0.656**	0.656	0.761	0.678	0.638	0.638	0.638	**0.638**	**0.795**	**0.795**	0.793	**0.793**
Random Forest	0.682	0.686	0.682	0.682	0.523	0.523	0.932	0.507	**0.641**	**0.660**	**0.641**	0.638	0.699	0.699	0.719	0.705
SAE Classifier	**0.711**	0.533	**0.789**	0.680	0.639	**0.875**	0.797	0.617	0.618	0.482	0.529	0.615	0.711	0.666	**0.854**	0.705
SAPLMA	0.570	0.391	0.500	0.518	0.548	0.820	**1.000**	0.451	0.596	0.398	0.529	0.543	0.749	0.630	0.779	0.648

We investigate per-task performance on RAGTruth. We find that performance across hallucination methods is highly fragmented: different classifiers perform best depending, model, or task, with no consistent winner across settings.

In particular, there is no clear advantage of SoTA detection methods over simple linear probes. In many cases, linear classifiers trained on model activations match or even outperform more complex methods like ReDeEP, SAPLMA or SAE-based classifiers. This further reinforces that current approaches may be overfitting to task-specific artifacts rather than capturing generalizable signals of hallucination.

We performed classification experiments cross-task on RAGTruth as shown in Table to the right to evaluate generalization capabilities of classifiers.

Performance dropped substantially. Both SoTA and linear probes exhibit near-random performance when applied OOD. This supports our hypothesis that hallucination detectors are latching onto task- or dataset-specific cues.

		Question Answering				Data-to-text writing				Summarization			
Method	Eval task	AUC	Prec.	Rec.	F1	AUC	Prec.	Rec.	F1	AUC	Prec.	Rec.	F1
Logistic Reg.	QA	0.572	0.57	0.57	0.57	0.560	0.55	0.56	0.55	0.528	0.53	0.53	0.52
	D2T	0.514	0.56	0.51	0.40	0.556	0.64	0.56	0.56	0.511	0.52	0.51	0.44
	Summ.	0.533	0.54	0.52	0.48	0.446	0.47	0.45	0.39	0.601	0.60	0.60	0.60
Random Forest	QA	0.589	0.59	0.58	0.59	0.540	0.52	0.54	0.50	0.501	0.50	0.50	0.50
	D2T	0.491	0.48	0.49	0.43	0.500	0.40	0.50	0.44	0.508	0.52	0.51	0.41
	Summ.	0.512	0.51	0.51	0.49	0.506	0.50	0.51	0.37	0.518	0.53	0.52	0.49
SAE Classifier	QA	0.706	0.53	0.79	0.67	0.500	0.35	1.00	0.26	0.505	0.35	1.00	0.27
	D2T	0.500	0.82	1.00	0.45	0.500	0.82	1.00	0.45	0.500	0.82	1.00	0.45
	Summ.	0.500	0.34	1.00	0.25	0.500	0.34	1.00	0.25	0.564	0.42	0.43	0.56
SAPLMA	QA	0.584	0.41	0.54	0.55	0.458	0.35	1.00	0.26	0.479	0.47	0.13	0.49
	D2T	0.593	0.82	1.00	0.45	0.533	0.82	1.00	0.45	0.653	0.87	0.76	0.60
	Summ.	0.567	0.37	0.90	0.41	0.559	0.34	1.00	0.25	0.566	0.34	0.51	0.49

Conclusion

Beware of spurious correlations when searching for a hallucination signal in the model activations