Contextual Hallucination Detection and Reduction using Aggregated Attention Scores in LLMs

Jan Eliasz Konrad Kiełczyński Mikołaj Langner Piotr Matys Teddy Ferdinan
Jan Kocoń Przemysław Kazienko

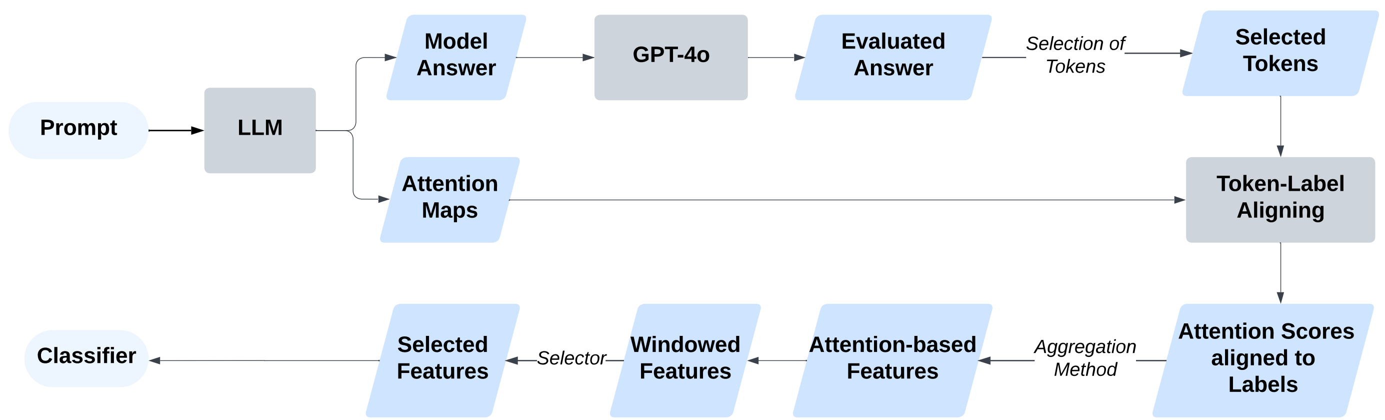
Department of Artificial Intelligence, Wroclaw Tech, Wrocław, Poland

Abstract

Can hallucination detection be done in real time? AggTruth is an innovative method for detecting hallucinations in large language models (LLMs) for contextual tasks. Using attention maps, AggTruth can identify whether a model is *making it up*, and thanks to its low complexity, the method is ready to be applied already during response generation. AggTruth outperforms current SOTA in terms of accuracy and stability between tasks, models and languages. It is the first step toward trustworthy generative models. Our results will be presented at the ICCS 2025 conference.

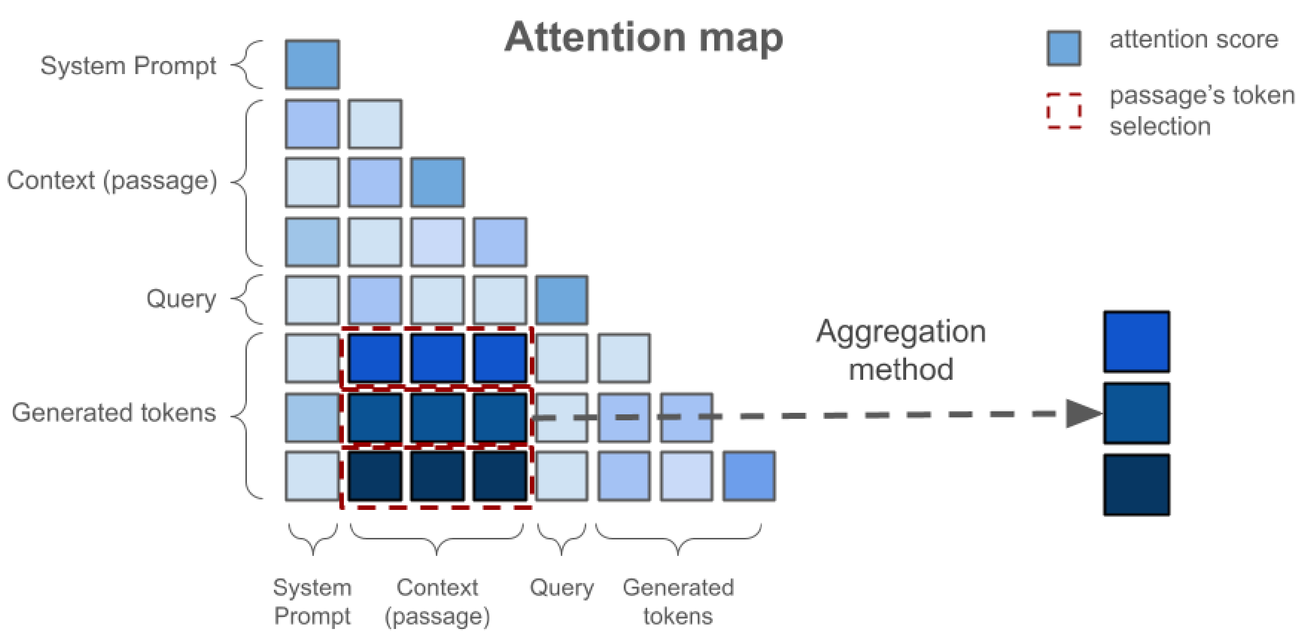
Pipeline

The whole end-to-end pipeline. It begins with a *Prompt* passed to an *LLM* and eventually results in *Selected Features* based on which the final *Classifier* detects potential hallucinations in the obtained answer.



Extracting Attention Features from Generated Tokens

Selection of tokens from the LLM input and aggregation of attention scores. The darker subregion indicates the attention scores of generated tokens on the provided passage. The aggregation of each such region provides one feature for the hallucination detection. Here, we have three regions resulting in three features.



Proposed AggTruth Techniques

$$\text{Sum} \quad = \sum_{i=1}^{C} a_{l,h,t,i}$$

$$\text{CosSim} \quad = \frac{1}{H-1} \sum_{\substack{h'=1 \\ h' \neq h}}^{H} \frac{\mathbf{a}_{l,h,t} \cdot \mathbf{a}_{l,h',t}}{\|\mathbf{a}_{l,h,t}\| \|\mathbf{a}_{l,h',t}\|}$$

$$\text{Entropy} \quad = -\sum_{i=1}^{C} a_{l,h,t,i} \log_2 a_{l,h,t,i}$$

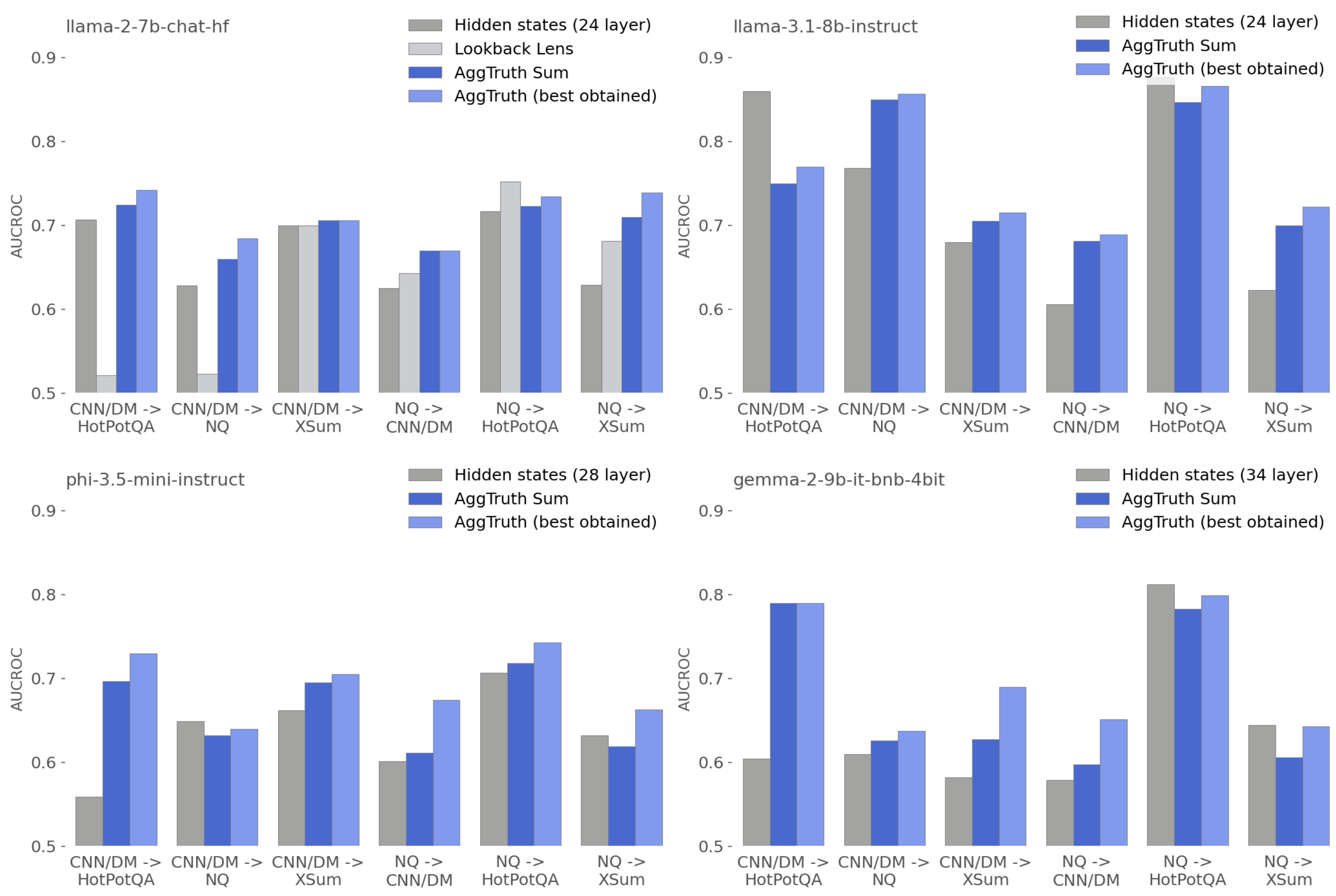
$$\text{JS-Div} \quad = \sqrt{\frac{1}{2} \sum_{i=1}^{C} \left(a_{l,h,t,i} \ln \frac{a_{l,h,t,i}}{m_{l,h,t,i}} + a_{l,\mathrm{ref},t,i} \ln \frac{a_{l,\mathrm{ref},t,i}}{m_{l,h,t,i}} \right)}$$

Department of Artificial Intelligence

Wrocław University of Science and Technology

Results

AUCROC hallucination detection results grouped by LLMs and tasks for best obtained hidden states-based method w.r.t. Gap value, Lookback Lens (only for Llama-2), AggTruth Sum and best obtained AggTruth method for a specific dataset and LLM.



Method	Source	Target	Source			Target		Gap [%]
			Train	Val	Test	Test(1)	Test(2)	
Text based NLI								
SOTA NLI*	QA	SUM.	–	–	–	–	0.530	–
	SUM.	QA	–	–	–	–	0.649	–
Hidden States based								
24th Layer	QA	SUM.	0.954	0.945	0.717	0.625	0.629	8.752
	SUM.	QA	0.988	0.982	0.700	0.707	0.628	5.356
28th Layer	QA	SUM.	0.955	0.946	0.727	0.603	0.623	9.564
	SUM.	QA	0.988	0.981	0.691	0.704	0.611	6.730
32th Layer	QA	SUM.	0.950	0.939	0.739	0.605	0.621	9.038
	SUM.	QA	0.986	0.978	0.678	0.660	0.573	11.117
Attention based								
Lookback Lens (paper)**	QA	SUM.	–	–	–	–	0.661	–
	SUM.	QA	–	–	–	–	0.660	–
Lookback Lens (classifiers)***	QA	SUM.	–	–	0.554	0.666	0.635	13.688
	SUM.	QA	–	–	**0.722**	0.506	0.506	19.299
Lookback Lens (retrained)****	QA	SUM.	0.839	0.833	**0.752**	0.643	0.681	3.952
	SUM.	QA	0.898	0.882	0.700	0.523	0.521	18.820
AggTruth Sum	QA	SUM.	0.802	0.799	0.723	**0.670**	**0.710**	**2.612**
	SUM.	QA	0.894	0.885	0.706	**0.724**	**0.660**	**2.714**

Classifier-Guided Decoding Methods

Performance comparison across decoding strategies on Llama2. Accuracy is plotted against latency (time per token), with the size of the bubbles reflecting GPU memory usage. The best methods should be in the top left corner.