Achieving Near-Zero Hallucation In Large Language Models

Abstract. This paper presents a research methodology focused on the mitigation of hallucinations in modern large language models. The initial phase involved the development and training of a models following the framework established in Google's "Attention Is All You Need" [1] paper. The degree of hallucination present in the model outputs was then systematically measured with respect to the training data. These measurements were obtained after multiple training iterations conducted on models of identical size but trained on datasets differing in quality and factual reliability, thereby yielding models with varying levels of knowledge representation. The results indicated that both data quality and model architecture contributed significantly to the prevalence of hallucinations. Consequently, architectural modifications were introduced, after which a model was trained on high-quality data. This revised configuration achieved a reduction of hallucinations in 96.4% of test cases.
This research paper is for a Research Methodology course at ELTE University. The data in it is made up, and should not be taken seriously or referenced.

1. Introduction

Your introduction

1.1. Related work

Works related to your paper

2. Methodology

2.1. Methodology subsection

Your methodology

3. Results

3.1. Benchmarking the hallucation rate of our model

After constructing the above mentioned model, we have done extensive benchmarking in order to validate our hallucation rate improvements. For these benchmarks we have chosen the HaluEval 2.0 benchmark [2], which strongly builds on the original HaluEval benchmark paper [3]. This benchmark measures the rate of factual and not factual answers given by a given model in the following five domains: Biomedicine, Finance, Science, Education and Open Domain. The process of evaluating a model with this benchark is as follows. Notably the HaluEval 2.0 benchmark uses two diffrent scores for a model in one domain: MaHR and MiHR. MiHR stands for micro hallucination rate and is calculate in the following way:

$$\text{MiHR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Count}(\textit{hallucinatory facts})}{\text{Count}(\textit{all facts in } r_i)},$$

The other score, MaHR stands for macro hallucination and is calculated in the following way:

$$\text{MaHR} = \frac{\text{Count}(\textit{hallucinatory responses})}{n}.$$

In our tests we compared our own model to some of the current flagship Large Language Models by Meta [4], Mistral [5], Anthropic [6] and OpenAI [7]:

Reference to the Table 1 on page 3 and a cite.

Table 1. Evaluation results on the tendency of LLMs to generate hallucinations. The lower the hallucination rate, the better the LLM performs. "*" represents that Claude 2 always refuses to answer questions in open domain.

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
OUR MODEL	**14.66**	**3.64**	**25.34**	**6.28**	**18.27**	**4.19**	**33.13**	**8.37**	**47.19**	**13.21**
Llama 4	28.76	7.23	35.91	9.25	15.21	3.36	36.84	10.13	39.18*	12.62*
Claude Sonnet 4.5	31.44	8.25	39.11	10.56	21.31	4.78	41.26	11.53	55.39	19.50
Text-Davinci-002	34.88	15.07	41.51	18.24	29.99	9.19	37.82	17.80	44.51	25.93
Text-Davinci-003	46.38	14.27	56.01	16.65	43.11	12.11	58.86	19.54	70.53	25.25
Vicuna 13B	50.59	17.55	46.19	13.15	34.44	8.75	55.81	17.88	65.43	29.15
Vicuna 7B	52.51	18.79	50.77	14.67	40.14	10.42	58.44	19.12	66.77	29.18
YuLan-Chat 13B	60.91	22.00	46.19	14.03	41.19	10.93	52.91	17.29	68.42	30.76
Llama 2-Chat 13B	52.61	17.90	53.48	14.53	39.11	10.37	62.12	19.30	79.19	30.44
Llama 2-Chat 7B	58.71	20.38	56.09	15.98	43.58	11.07	66.04	21.64	80.99	32.64
Alpaca 7B	53.52	24.42	53.47	24.46	40.74	12.74	68.95	22.38	65.65	29.57

4. Discussion

Your discussion

Acknowledgment

Your acknowledgement

References

[1] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[2] Junyi Li et al. "The dawn after the dark: An empirical study on factuality hallucination in large language models". In: *arXiv preprint arXiv:2401.03205* (2024).

[3] Junyi Li et al. "Halueval: A large-scale hallucination evaluation benchmark for large language models". In: *arXiv preprint arXiv:2305.11747* (2023).

[4] URL: `https://www.llama.com/`.

[5] URL: `https://mistral.ai/`.

[6] URL: https://www.anthropic.com/.
[7] URL: https://openai.com/.

Zoltán Blahovics
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
euxhhx@inf.elte.hu

Bence Nagy
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
hvtdd4@inf.elte.hu

Krisztián Nemes-Kovács
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
email

Balázs Fekete
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
email