Reducing hallucinations in Medical LLMs


Arjun Neervannan


Karan Sampath

Penn Engineering
University of Pennsylvania

Introduction

- Research question: How can we reduce the frequency of hallucinations in LLMs in medical contexts through better prompting methods?
- LLMs frequently hallucinate, especially in contexts where there is **specialized information** that needs to be retrieved (e.g., medical contexts)
- Hallucinations are events where LLMs produce or cite **factually incorrect information,** which often may seem coherent but are inherently untrue
- Prior research does not explore **system prompts**

Experimental Design

- **Medalpaca-13B** quantized model used with Modal Labs
- **Four system prompting** configurations tested: standard, detailed, few-shot and chain-of thought system prompting
- **PubMedQA** dataset used – 1k labeled test examples
- Each instance comprised of question, context, long answer and short answer
- Multiple Hallucination Detection Methods:

$$\varepsilon = \frac{1}{L}\sum_{n=1}^{L} \log p(y_k|y_{<k}, x, \theta)$$

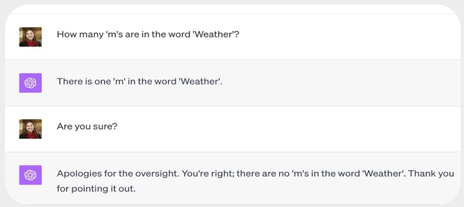
***Sequence probabilities**, which represent how "confident" the model is about the response, is used as a proxy for hallucination. The log probabilities are averaged over the whole sequence.*

***Hallucination Detection** is done using the seq-log-probability technique, which was found to be the most accurate hallucination detection technique by Guerreiro et al in 2023.*

$$S(x) = -\sum_{i=1}^{N} p(x_i) \log(p(x_i))$$

***P(x$_i$)** represents the probability of obtaining the value **x$_i$** and represents how new and random the value is. Models that are hallucinate more are likely to have more randomness and higher entropy in their output.*

*We also use Shannon Entropy to supplement **Hallucination Detection**, which was found to be an accurate hallucination detection technique by Manakul et al in 2023.*



***Above:** Example of LLM Hallucination*

Experiment Configurations

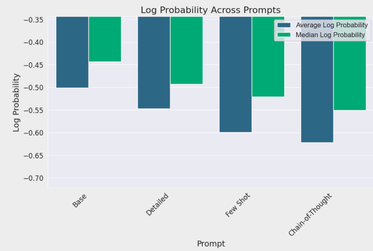
Standard Prompt
Please answer the question based solely on the context provided. Your answer should be 'Yes', 'No' or 'Maybe' followed by a justification that directly references the context.

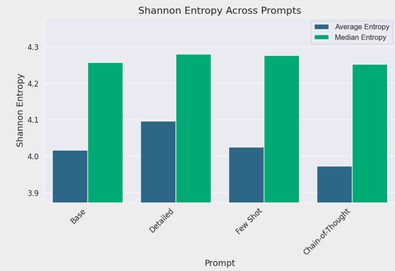
Detailed Prompt
Please answer the question based solely on the context provided, without inferring or adding information not present in the context. Your answer should be 'Yes' or 'No' or 'Maybe' followed by a justification that directly references the context. Output it in the following format.{ Decision}. {Explanation}

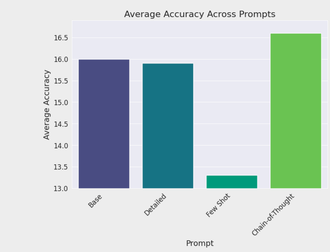
Few-shot System Prompt
Context: During a study on cardiovascular responses, researchers found that moderate exercise improves baroreflex sensitivity and reduces blood pressure in hypertensive patients.
Question: Does moderate exercise improve baroreflex sensitivity in hypertensive patients?
Answer: Yes. The study clearly demonstrates that moderate exercise leads to an improvement in baroreflex sensitivity among hypertensive patients, which is crucial for their cardiovascular health.
Answer the following question with a similar structure. Your answer should be 'Yes', 'No' or 'Maybe' followed by a justification that directly references the context.

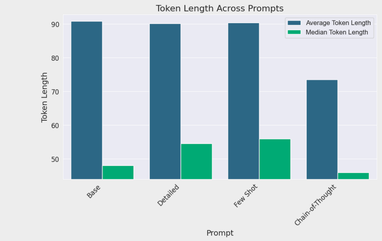
Chain-of-Thought Prompt
Context: The study aims to determine if physiological, rhythmic fluctuations of vagal baroreflex gain, which are crucial for maintaining cardiovascular stability, persist during various phases including exercise, post-exercise ischaemia, and recovery.
Step 1: Define what vagal baroreflex gain is and its importance in cardiovascular regulation.
Step 2: Analyze how exercise and ischaemia might influence these baroreflex rhythms.
Step 3: Reflect on the evidence provided about the persistence of these rhythms during the specified conditions.
Answer: Yes. The study clearly demonstrates that moderate exercise leads to an improvement in baroreflex sensitivity among hypertensive patients, which is crucial for their cardiovascular health.
Answer the following question with a similar structure. Your answer should be 'Yes', 'No' or 'Maybe' followed by a justification that directly references the context.

Results









***Above:** Results for log probability, Shannon entropy, token length, and ground truth accuracy (clockwise starting from top left) on all prompt configurations*

Analysis & Future Steps

- Our results indicate mixed results: while the base model was likely to give more confident results, chain of thought prompting allowed for less verbosity and higher accuracy indicating lower hallucination risk.
- Our research agrees with prior papers that LLMs are zero-shot learners (Kojima et al)
- **Future Steps include:**
  - Using additional hallucination detection such as attention-based methods
  - Fine-tuning models using additional training data, which wasn't feasible due to compute constraints
  - Extending the problem to other domains such as reading comprehension and other QA sets