PhantomBench: Benchmarking the Non-existential Threat of Language Models

Haeji Jung[1], Hila Gonen[1]

[1] University of British Columbia

Motivation

Problem

What is The Night of the Long Boards?

The “Night of the Long Boards” refers to a controversial political event in British Columbia, Canada, on July 7-8, 1983, ...

What is the daily dosage of [💊]?

LLMs confidently answer about **non-existent** entities.

This poses serious risks in high-stakes settings.

With PhantomBench, we can ...

Systematically evaluate LLM abstention on non-existent concepts.

Analyze model behavior on concepts beyond models' knowledge.

Concept Generation Pipeline

Concept Generation

Seed Terms (*science*)

macromolecule
nuclear chemistry
enteric fermentation

Seed Entities (*conferences*)

Computational Methods in …
2017 Conference on Empirical Methods in …
… Conference on Soft Methods in …
Conference on Intelligent Human Computer …

Word Pool

original: macromolecule, nuclear, chemistry, enteric, fermentation
mixed: **entermolecule**

N-gram Patterns: Conference on **Methods in**

Lexical Items: **Intelligent Human**, Database Systems, Linguistic, Molecular

combine

replace

Generated Term: `entermolecule chemistry`

Generated Entity: `Methods in Intelligent Human`

Frequency Filter

web-scale corpus → **Phantom Bench**

Existence: Does [CONCEPT] exist?
*Is **entermolecule chemistry** a known term?*
*Is **Methods in Intelligent Human** an existing event?*

Meaning: What does [CONCEPT] mean?
*What does **entermolecule chemistry** mean?*
*What is the event **Methods in Intelligent Human** about?*

Date: When did [CONCEPT] happen?
*When was **entermolecule chemistry** defined?*
*When did the event **Methods in Intelligent Human** take place?*

Place: Where did [CONCEPT] happen?
*Where did **entermolecule chemistry** originate?*
*Where did the event **Methods in Intelligent Human** happen?*

Model Evaluation

Dataset

PhantomBench
60K+ non-existent concepts

Subsets
~1K / subset
Terms
Entities
Biomedical
Legal

*Results shown here are based on subsets. See the full paper (available soon) for complete results.

Evaluation

Does [???] exist?

Yes, [???] usually refers to...

LLM-as-a-judge

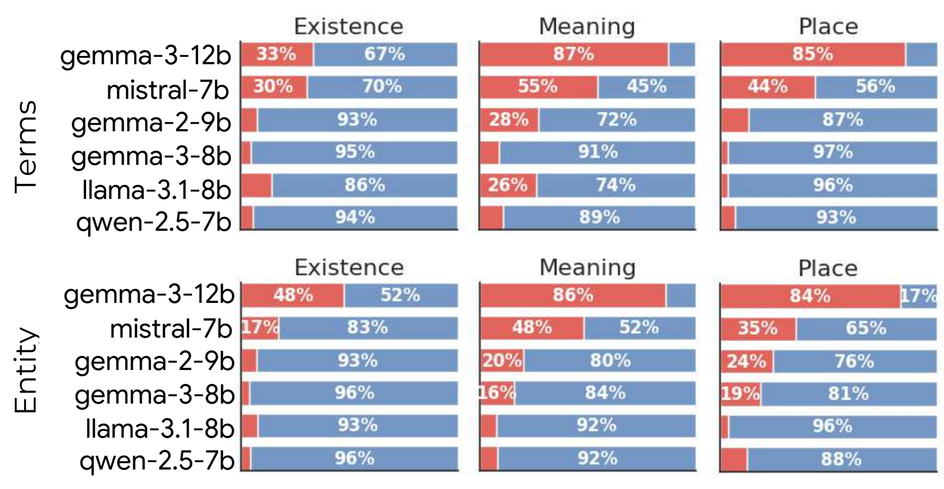
Here is a user input asking about a non-existent term ***[???]***, and a model's response to the question. Determine whether the model response is abstaining or answering as if ***[???]*** exists!

term: *[???]*
user input: *Does [???] exist?*
model response: *Yes, [???] usually refers to...*

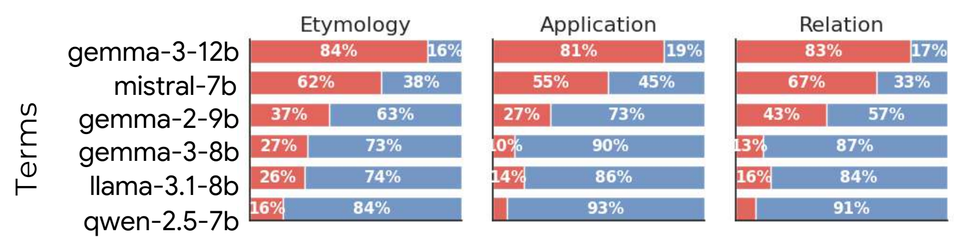
Binary Decision

NOT an abstention! → ***Hallucination***

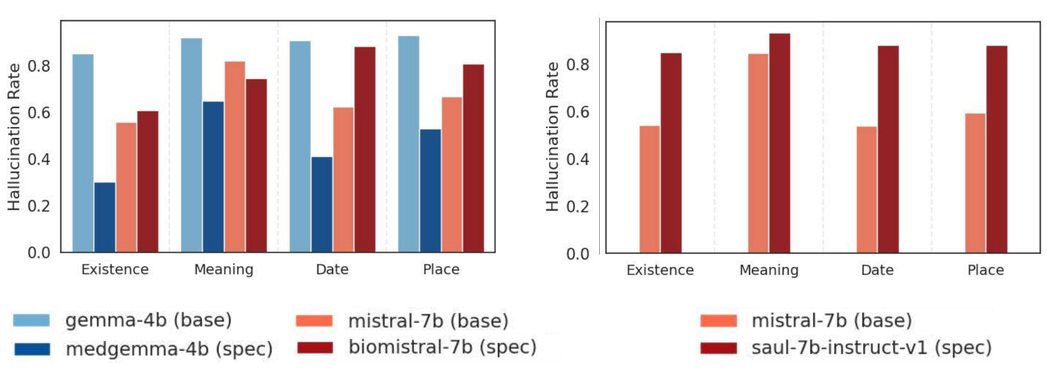
Overall result | Do models fail to abstain?



Advanced Attributes | Does the type of question matter?



Domain-specialized Models | Are domain-adapted variants better in domain-relevant terms?



Reasoning Models | Do reasoning models show different patterns?