Benchmarking Prompt Sensitivity in Large Language Models

Amirhossein Razavi*
Toronto Metropolitan University

Sara Salamat
Toronto Metropolitan University

Mina Soltangheis*
Wondeur AI, Toronto Metropolitan University

Morteza Zihayat
Toronto Metropolitan University

Negar Arabzadeh
University of Waterloo

Ebrahim Bagheri
University of Toronto

Prompt sensitivity in LLMs

Dataset	Original Prompt	Alternative Prompt	Original Answer	Alternative Answer	Correct Answer
HotpotQA	What American actor and comedian known for playing the role of Newman in Seinfeld, also stars in the series The Exes on TV Land?	What is the name of the American actor who played Newman in Seinfeld and appears in TV Land's comedy series The Exes	Wayne Knight	Jerry Seinfeld co-star	Wayne Knight
TriviaQA	At which city do the Blue and White Niles meet?	At which geographical location do the Blue and White Niles meet	Sudan's confluence	Khartoum	Khartoum

Our contribution

To study prompt sensitivity of LLMs, we introduced:

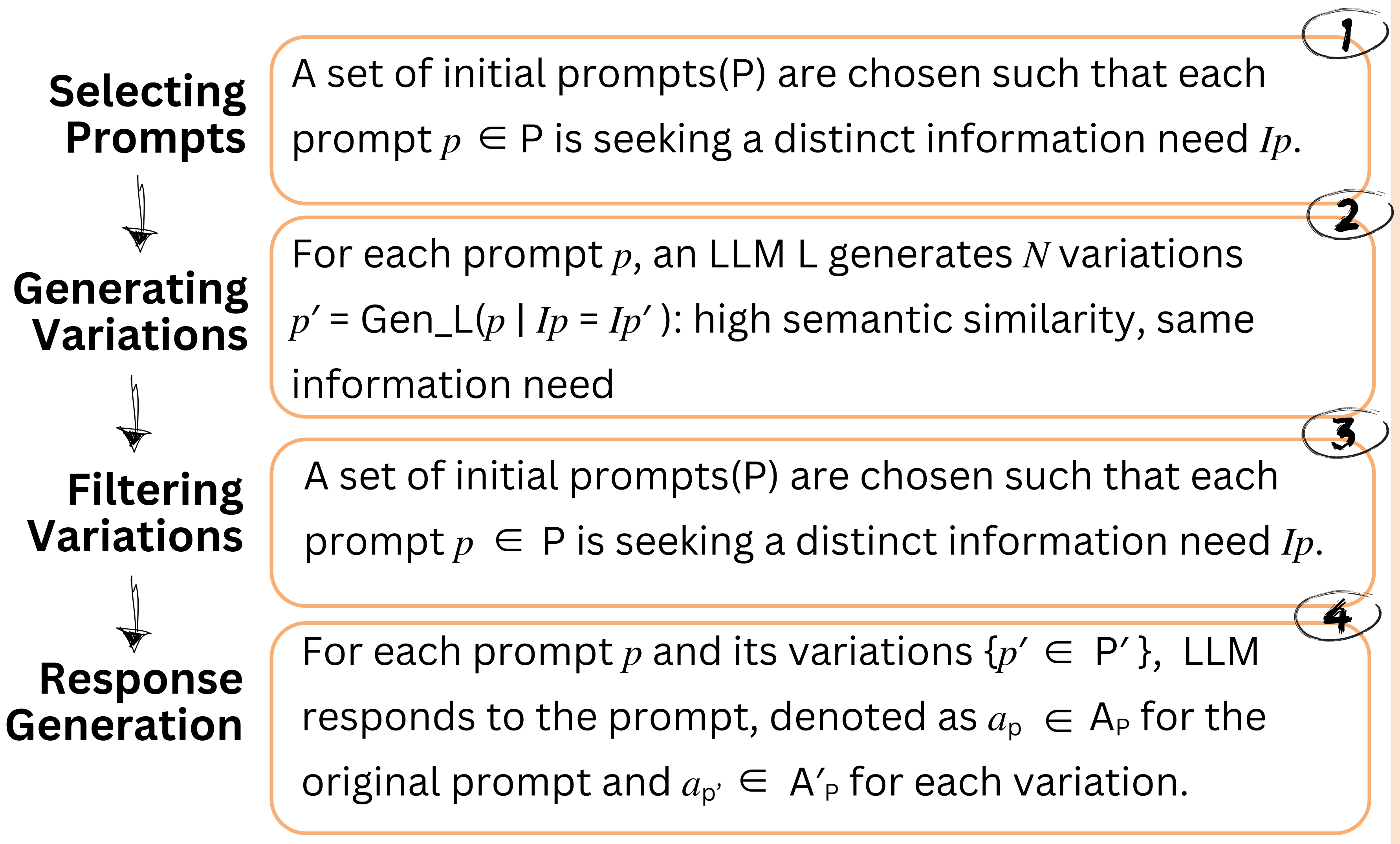
1 a novel task ***prompt sensitivity prediction,*** and

2 accompanying dataset curated for prompt sensitivity prediction ***PromptSET***

PromptSET Dataset

The gold standard dataset for the prompt sensitivity task: we used two widely-used question-answering datasets, TriviaQA and HotpotQA datasets that have human annotated answers available

Methodology



Selecting Prompts

1 A set of initial prompts(P) are chosen such that each prompt $p \in$ P is seeking a distinct information need Ip.

Generating Variations

2 For each prompt p, an LLM L generates N variations p' = Gen_L(p | Ip = Ip'): high semantic similarity, same information need

Filtering Variations

3 A set of initial prompts(P) are chosen such that each prompt $p \in$ P is seeking a distinct information need Ip.

Response Generation

4 For each prompt p and its variations {$p' \in$ P′}, LLM responds to the prompt, denoted as $a_p \in A_P$ for the original prompt and $a_{p'} \in A'_P$ for each variation.

Prompt Sensitivity Prediction

Our proposed task of Prompt Sensitivity Prediction aims to predict whether a given prompt can be effectively fulfilled by the LLM whose response to the prompt would satisfy the users' information need.

Benchmark

To benchmark this task, we identify three types of tasks from the literature that may be applicable to prompt sensitivity prediction:

1. **LLM as a judge:** We directly ask LLMs to self-assess their ability to predict whether they can accurately answer a given prompt or not.
2. **Text classification:** We train a text classifier on PromptSet to predict whether the LLM's response to a prompt will meet users' information need.
3. **QPP methods:** We adopted BERT-PE, a pre-retreival, and collection-independent QPP method, which uses contextualized embeddings to learn query performance. Additionally, we considered the neural embedding specificity-based QPP metrics such as Closeness Centrality (CC), Degree Centrality (DC), PageRank , and Inverse Edge Frequency (IEF).

Experiments and Findings

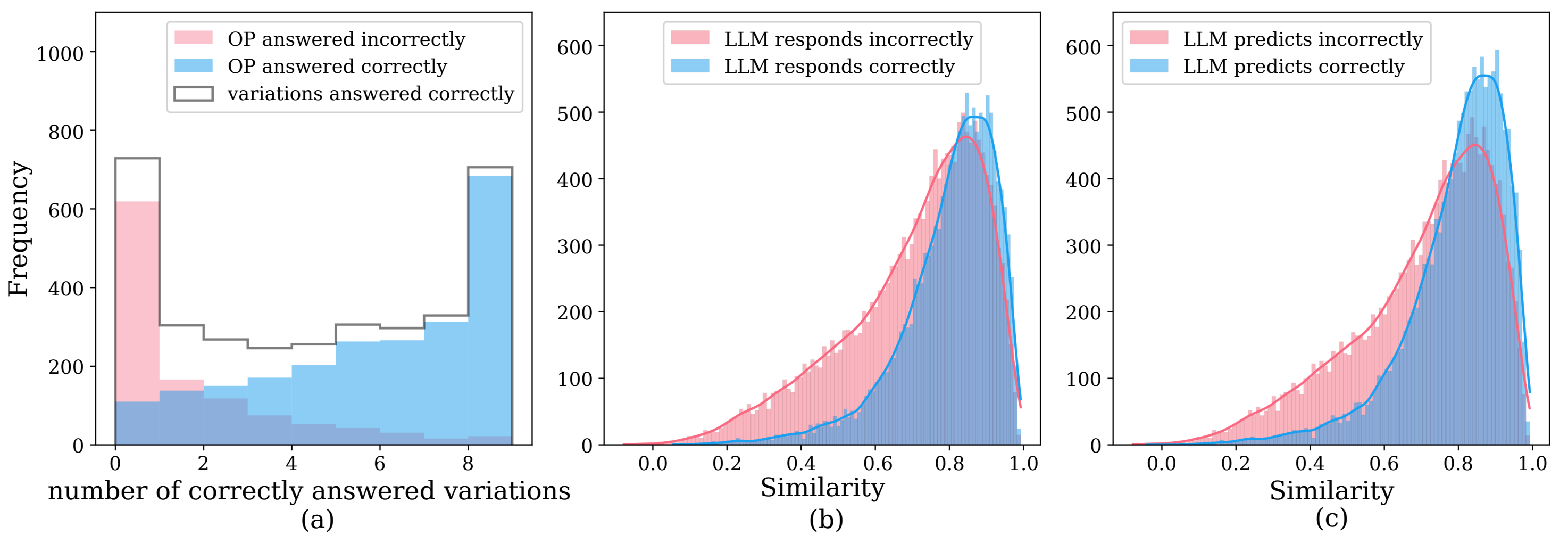
	Category	Method	PromptSET-TriviaQA				PromptSET- HotPotQA			
			Accuracy	F1	Recall	Precision	Accuracy	F1	Recall	Precision
Mistral Answers	LLM-Based	Mistral	0.5045	0.5858	0.7743	0.4711	0.3735	0.2005	0.6912	0.1173
		LLaMA	0.4656	0.6239	0.9798	0.4577	0.1696	0.2050	0.9419	0.1150
	Text Classification	BERT	0.660	0.659	0.620	0.654	0.526	0.360	0.017	0.813
	Specificity-based QPP	CC	0.506	0.453	0.452	0.454	0.549	0.209	0.524	0.130
		DC	0.484	0.448	0.463	0.434	0.565	0.199	0.475	0.126
		IEF	0.505	0.462	0.469	0.455	0.535	0.204	0.526	0.127
		PageRank	0.481	0.444	0.458	0.431	0.533	0.153	0.370	0.096
	Supervised QPP	BERTPE	0.648	0.627	0.644	0.611	0.710	0.318	0.594	0.217
LLaMA Answers	LLM-Based	Mistral	0.5160	0.6045	0.7704	0.4974	0.3731	0.1978	0.6930	0.1153
		LLaMA	0.4940	0.6507	0.9818	0.4866	0.1674	0.2013	0.9408	0.1127
	Text Classification	BERT	0.664	0.664	0.651	0.650	0.532	0.377	0.034	0.808
	Specificity-based QPP	CC	0.500	0.463	0.449	0.478	0.545	0.199	0.507	0.123
		DC	0.484	0.464	0.465	0.463	0.562	0.190	0.462	0.120
		IEF	0.510	0.482	0.475	0.489	0.535	0.202	0.529	0.125
		PageRank	0.482	0.461	0.461	0.461	0.534	0.151	0.371	0.094
	Supervised QPP	BERTPE	0.659	0.651	0.646	0.656	0.710	0.314	0.596	0.213

Baseline Performance on PromptSET

PromptSET is accessible on Github: https://github.com/Narabzad/prompt-sensitivity

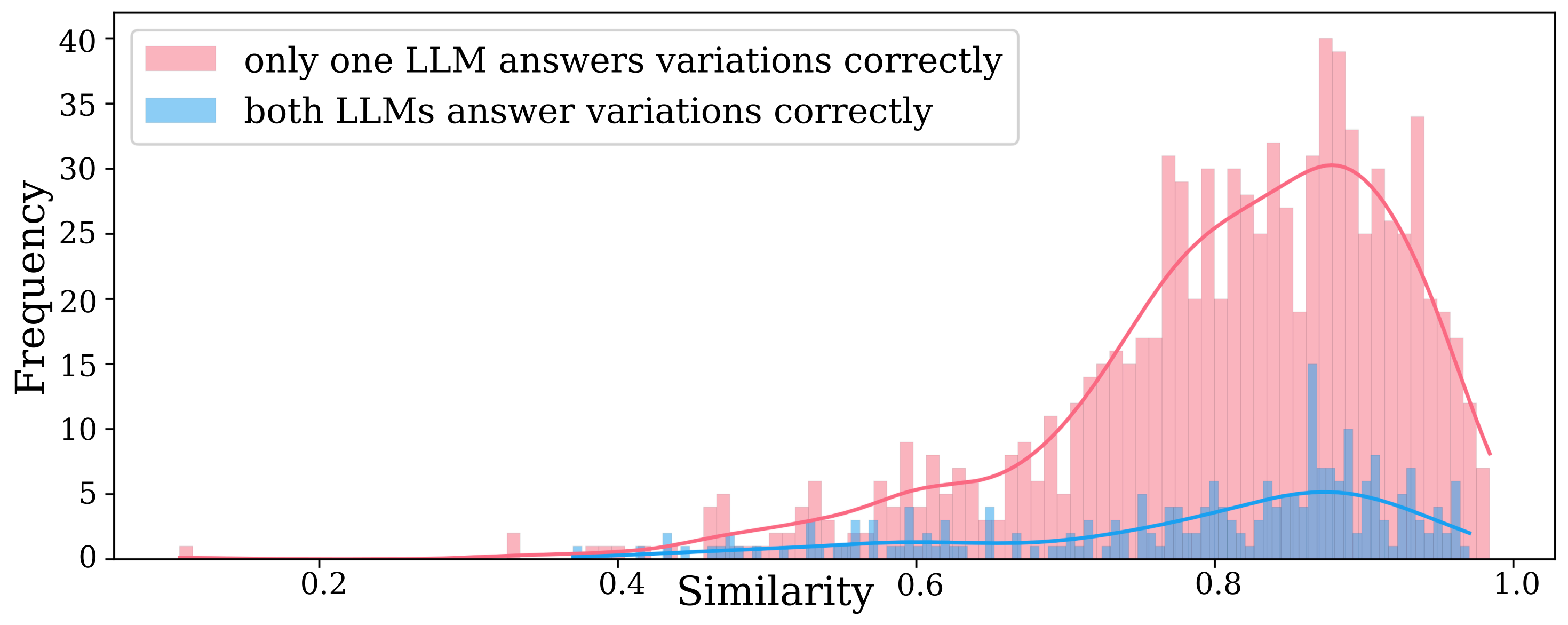
*These authors contributed equally to this work and are listed in alphabetical order.

Analysis



(a) (b) (c)

- Original prompt correctness impacts variation answerability.
- Similarity to the original prompt impacts the predictability of LLM responses.
- Choice of LLM impacts variation answerability.


only one LLM answers variations correctly
both LLMs answer variations correctly
Frequency
Similarity