Benchmarking Prompt Sensitivity in Large Language Models

Amirhossein Razavi[*1], Mina Soltangheis[*1,2], Negar Arabzadeh[3],
Sara Salamat[1], Morteza Zihayat[1], and Ebrahim Bagheri[4]

[1] Toronto Metropolitan University, Toronto ON, Canada
{amirhossein.razavi,msoltangheis,sara.salamat,mzihayat}@torontomu.ca
[2] Wondeur Ai, Toronto ON, Canada
mina@wondeur.ai
[3] University of Waterloo, Waterloo ON, Canada
narabzad@uwaterloo.ca
[4] University of Toronto, Toronto ON, Canada
ebrahim.bagheri@utoronto.ca

Abstract. Large language Models (LLMs) are highly sensitive to variations in prompt formulation, which can significantly impact their ability to generate accurate responses. In this paper, we introduce a new task, Prompt Sensitivity Prediction, and a dataset `PromptSET` designed to investigate the effects of slight prompt variations on LLM performance. Using TriviaQA and HotpotQA datasets as the foundation of our work, we generate prompt variations and evaluate their effectiveness across multiple LLMs. We benchmark the *prompt sensitivity prediction* task employing state-of-the-art methods from related tasks, including LLM-based self-evaluation, text classification, and query performance prediction techniques. Our findings reveal that existing methods struggle to effectively address prompt sensitivity prediction, underscoring the need to understand how information needs should be phrased for accurate LLM responses.

1 Introduction

Large language models (LLMs) can generate human-like responses to a wide array of prompts, from answering specific queries to planning for accumulating information to answer complex questions [22]. Despite their usefulness, a notable challenge in working LLMs is their sensitivity to prompt formulation [25]. Small variations in the phrasing, structure, or even punctuation of prompts can often lead to substantially different outputs [36,32,29]. To illustrate this issue, consider the sample prompts shown in Table 1. In this table, we present samples from the TriviaQA [21] and HotPotQA [39] question-answering datasets where the LLM responds correctly and accurately to the original prompts. However, with only slight modifications in wording, we observe that the LLM (in this case LLaMA3.1) fails to provides the correct response.

This challenge, which we refer to as *prompt sensitivity*, highlights the challenges users face when crafting their prompts [15,41]. For this reason, *prompt engineering*,

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

Table 1: Samples of sensitive prompts from HotpotQA and TriviaQA datasets.

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

the art of designing effective prompts, has become an active area of research [24,10]. Researchers have already examined the effects of various prompt modifications, including minor structural and formatting changes [36], adversarial prompting [40], and generating prompts with different levels of specificity [30]. We hypothesize that prompt sensitivity could arise because of several reasons. For instance, an LLM may successfully respond to prompts closely aligned with examples seen during training, but struggle with slight modifications that it has not encountered in the past [3]. Another factor could be the model's reliance on specific syntactic or semantic patterns to interpret prompts accurately, which may be impacted due to slight changes in the prompt.

To this end and in this paper, we introduce a novel task and its accompanying dataset specifically curated for *prompt sensitivity prediction*. By curating a collection of prompts and their variations, we aim to predict whether a given LLM would be able to effectively respond to an input prompt or whether it would fail to provide a satisfactory response. Our proposed dataset serves as a benchmark for studying prompt sensitivity, thus setting the stage for forthcoming studies in prompt engineering and the evaluation of LLM responses to prompt variations.

A common prompt engineering strategy is to ask the LLM itself to reformulate the prompt in a way that a more desirable output would be generated for the revised prompt by the LLM [19,11]. While LLMs can autonomously generate different prompt variations, they cannot assess which variations are most effective, pointing to the fact that LLMs themselves are oblivious to the representation of optimal prompt variations. To establish a benchmark for this challenge, we formally introduce the *Prompt Sensitivity Prediction* task, which is concerned with assessing the effectiveness of a user prompt and its variations. We systematically curate our dataset based on the TriviaQA and HotpotQA [21,39], which consist of prompts that have deterministic and concise answers. To benchmark this task, we draw parallels with established tasks in text classification (TC) [16,12] and query performance prediction (QPP) [27,8,17,5,4], as they share resemblance with prompt sensitivity prediction. Our experiments show that such baselines fail to perform effectively for this task, underscoring the need for novel approaches tailored in particular for *prompt sensitivity prediction*. In summary, the contributions of our work in this paper include: **(1)** We define the prompt sensitivity prediction task, outlining the requirements and challenges involved in identifying effective prompts; **(2)** We introduce and publicly release a comprehensive dataset for Prompt Sensitivity Evaluation Task (`PromptSET`), focusing on slight prompt modifications that unveil LLM sensitivity to prompt variations[*]; and, **(3)** We benchmark the prompt sensitivity prediction task using

[*] `https://github.com/Narabzad/prompt-sensitivity`

state-of-the-art methods, including text classification, query performance prediction, and LLM-based baselines, to highlight the complexity of the proposed task.

2 Methodology

The Task Definition. 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. More specifically, given a prompt p with a specific information need I_p, we consider a set of similar prompts, denoted $\mathcal{P} = \{p' | Sim\langle p, p'\rangle > \tau \text{ and } I_p == I_{p'}\}$, where each variation p' shares the same information need I_p and maintains a similarity with p above a predefined threshold τ. These prompts $\{p'\}$ are designed to be only slightly modified versions of p, ensuring they still reflect the same information need of the user. The goal of this task is to predict, for a given prompt p_i, whether the LLM will generate a response that accurately respond to the underlying information need I_{p_i}.

The Dataset for the Task. To create the gold standard dataset for the prompt sensitivity task, we adopt a systematic process to generate prompt variations and evaluate their effectiveness as follows:

1. **Selecting Prompts**: We start by choosing a set of initial prompts, denoted as $\mathcal{P}$, where each prompt $p \in \mathcal{P}$ is seeking a distinct information need I_p.
2. **Generating Variations**: For each prompt p in the set, we use an LLM $\mathcal{L}$ to generate N variations $p' = \mathcal{L}(p \mid I_p = I_{p'})$. Here, $\mathcal{L}(p)$ denotes the process of generating variations of prompt p, where each variation p' retains high semantic similarity with p, i.e., $(Sim\langle p, p'\rangle > \tau)$ and preserves the original information need I_p.
3. **Filtering Variations**: We process and filter out any generated variations p' that do not meet specific criteria for similarity and alignment with the original prompt p including LLM hallucinated content.
4. **LLM Response Generation**: For each prompt p and its variations $\{p' \in \mathcal{P}'\}$, we ask the LLM to respond to the prompt, denoted $a_p \in \mathcal{A}_{\mathcal{P}}$ for the original prompt and $a_{p'} \in \mathcal{A}'_{\mathcal{P}}$ for each variation.

The combination of $\mathcal{P} \cup \mathcal{P}'$ as well as their LLM generated answers $\mathcal{A}_{\mathcal{P}} \cup \mathcal{A}'_{\mathcal{P}}$ form the `PromptSET` dataset for this task.

Source Data. To build our dataset, we require a set of prompts that have human annotated answers available as well as having reliable evaluation with deterministic results. Therefore, we selected two widely-used question-answering datasets, TriviaQA [21] and HotpotQA datasets [39] that meet these requirements. TriviaQA is a reading comprehension dataset containing over 650K question-answer-evidence triples. The questions are on average 14 words. Each question has a collection of accepted answers including a list of aliases and normalized version of the answers; the majority of which are specific and short [21]. Furthermore, HotpotQA is a large-scale question-answering dataset consists of 113k training question-answer pairs. Unlike TriviaQA, HotpotQA includes complex multi-hop and comparison questions that require reasoning across multiple documents to answer accurately [39].