Performance Sensitivity of Large Language Models Under Prompt and Decoding Variations

Team Wise Wyver

STAT496

February 2026

1 Introduction

Large language models (LLMs) are increasingly used for structured tasks such as classification, information extraction, and summarization across research and industrial applications. Despite strong empirical performance, recent studies have shown that LLM outputs can be sensitive to prompt formulation and decoding parameters, raising concerns about reliability and reproducibility.

Early work on in-context learning demonstrated that language models can adapt to new tasks through examples embedded directly in prompts without additional training [Brown et al., 2020]. Subsequent research showed that prompting strategies significantly influence model reasoning behavior and downstream performance [Wei et al., 2022]. Zhao et al. [Zhao et al., 2021] further demonstrated that minor prompt variations may lead to substantially different outputs, suggesting that LLM responses behave more like stochastic samples than deterministic computations.

Beyond prompt design, decoding parameters such as temperature influence token sampling during generation. Higher temperature values introduce randomness, while lower values

encourage deterministic outputs. Although practitioners frequently adjust temperature in deployed systems, systematic empirical evaluation of how temperature interacts with prompt structure and model architecture remains limited.

In this project, we treat LLMs as black-box stochastic systems and investigate how experimental factors influence classification performance. Specifically, we study how prompt structure, model version, and decoding temperature affect sentiment classification accuracy on Twitter data. Our goal is to better understand how practical configuration choices influence downstream task behavior.

2 Experimental Design

2.1 Task and Dataset

We consider a sentiment classification task using publicly available Twitter posts labeled by human annotators as positive, neutral, or negative. Sentiment classification provides a standardized output format, allowing consistent comparison across experimental conditions.

A subset of 50 tweets is used as the evaluation set. Remaining labeled tweets serve as an example pool for constructing few-shot and many-shot prompts. Model predictions are compared against ground truth labels to compute classification accuracy.

2.2 Independent Variables

The experiment follows a fully crossed factorial design consisting of three independent variables:

1. **Prompt Structure**
    - Zero-shot
    - Definition-based prompting
    - Few-shot (3 examples)

- Many-shot (10 examples)

2. **Model Version**

   - GPT-3.5-Turbo
   - GPT-4.1-mini
   - GPT-4.1

3. **Temperature**

   - 0.0
   - 0.5
   - 1.0

This produces $4 \times 3 \times 3 = 36$ experimental conditions.

2.3 Execution Pipeline

All experiments are executed through an automated Python pipeline interacting with the OpenAI API. Each prompt is submitted as a single user message, and responses are constrained to produce one sentiment label per tweet.

Predictions are parsed automatically and stored in structured CSV files. Accuracy is defined as the proportion of correctly classified tweets:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}} \tag{1}$$

3 Data Analysis

This section presents exploratory analysis examining how experimental factors influence classification performance.

3.1 Accuracy Across Temperature Levels

Figure 1 shows model accuracy across decoding temperature settings. GPT-3.5-Turbo demonstrates decreasing accuracy as temperature increases, suggesting that stochastic sampling negatively affects classification reliability. In contrast, GPT-4.1 models exhibit more moderate variation, indicating reduced sensitivity to decoding randomness.

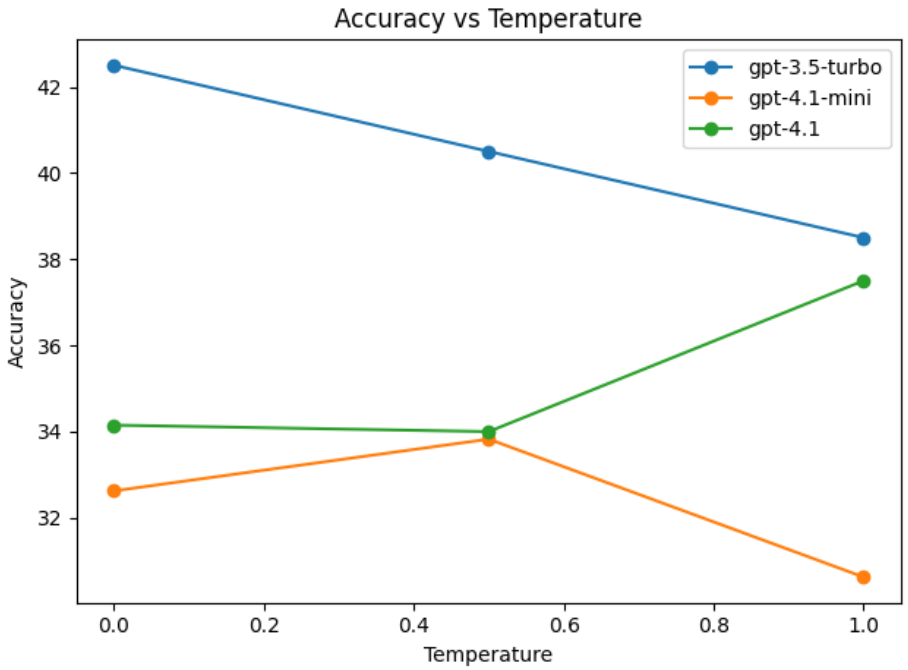

Figure 1: Accuracy across temperature levels for different models.

3.2 Model-Level Performance Comparison

Figure 2 summarizes accuracy distributions across models. GPT-3.5-Turbo achieves relatively high median accuracy but shows larger variability across conditions. GPT-4.1 demonstrates more concentrated performance, suggesting improved robustness to prompt and temperature variation. GPT-4.1-mini generally produces lower accuracy values, reflecting trade-offs associated with smaller model size.

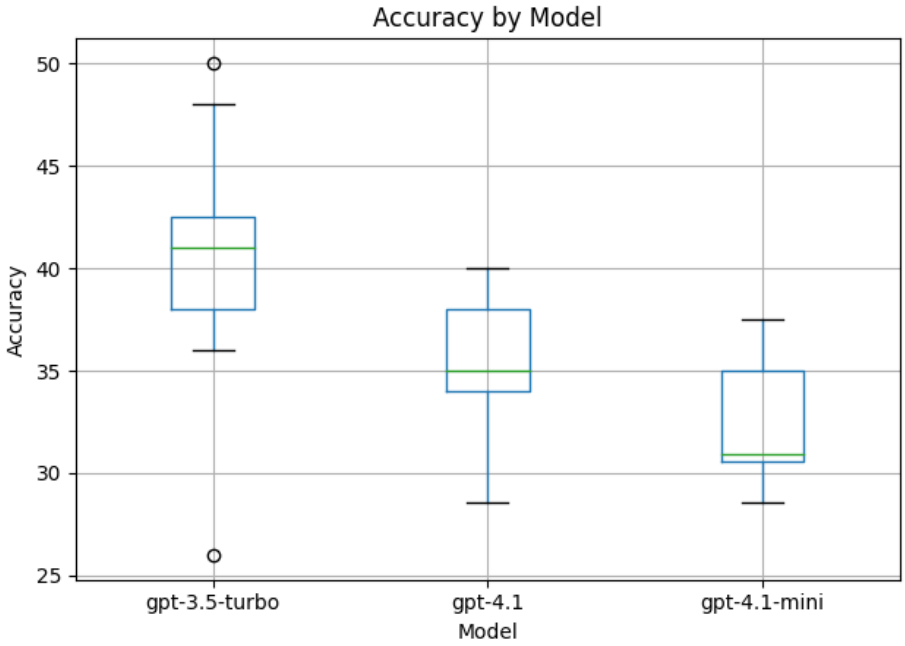

Figure 2: Accuracy distribution across model architectures.

3.3 Effect of Prompt Structure

Prompt structure also influences performance outcomes. Increasing contextual information through in-context examples can improve accuracy in certain configurations, though improvements are not uniform across models or temperature levels. These results suggest that prompt effectiveness depends jointly on model capacity and decoding settings.

3.4 Discussion and Future Work

The present draft focuses on accuracy-based evaluation to characterize performance sensitivity across experimental factors. Future analysis will incorporate repeated experimental trials to quantify output stability and reproducibility under stochastic generation. Such extensions will enable direct measurement of agreement across independent model runs.

Overall, the results demonstrate that model architecture, prompt design, and decoding temperature jointly influence LLM classification behavior.

References

Brown, T. et al. Language Models are Few-Shot Learners. *NeurIPS*, 2020.

Wei, J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS*, 2022.

Zhao, T. et al. Calibrate Before Use: Improving Few-Shot Performance of Language Models. *ICML*, 2021.