STAT496 Final Paper

Team Wise Wyver

February 2026

1 Introduction

Large language models (LLMs) have become widely adopted for structured tasks such as classification, information extraction, and summarization, enabling automation across research workflows and industrial applications. Despite their strong performance, a growing body of work suggests that LLM outputs can be highly sensitive to prompt formulation and generation parameters, raising concerns about reproducibility and reliability. Understanding the stability of model outputs is particularly important when LLMs are used as components in automated analysis pipelines, where inconsistent outputs may propagate errors into downstream processes.

Early work on in-context learning demonstrated that large-scale language models can adapt to new tasks through examples provided directly within the prompt (Brown et al., 2020). Subsequent studies showed that prompting strategies significantly influence reasoning behavior and performance, with techniques such as chain-of-thought prompting improving reasoning capabilities by encouraging intermediate steps in generation (Wei et al., 2022). While these studies highlight the power of prompt design, they also imply that model behavior may depend strongly on contextual cues, suggesting potential variability when prompts change. Further research has examined prompt sensitivity and instability, showing that minor variations in prompt wording or example ordering can produce substantially different outputs, even when task instructions remain unchanged (Zhao et al., 2021). These findings indicate that LLM responses should be viewed as stochastic outcomes influenced by both contextual and probabilistic factors.

In addition to prompt design, decoding parameters such as temperature introduce controlled randomness into generation. Higher temperature values increase diversity by flattening probability distributions over tokens, while lower values encourage deterministic outputs. Although practitioners often adjust temperature settings to balance creativity and reliability, systematic empirical evaluation of how temperature interacts with prompt structure and model architecture remains limited. Meanwhile, newer model iterations are frequently assumed to produce more consistent outputs due to improvements in training and alignment, yet empirical evidence comparing stability across model versions is still emerging.

This project builds upon prior literature by shifting focus from accuracy-oriented evaluation toward reproducibility and structural consistency. Rather than optimizing task performance alone, we treat the LLM as a black-box stochastic system and investigate how multiple experimental factors influence output stability. Specifically, we examine how the number of in-context examples, model version (ChatGPT-3.5, ChatGPT-4, and ChatGPT-5), and temperature jointly affect the consistency of outputs across repeated trials. By integrating insights from research on prompting strategies and probabilistic generation, our study aims to provide practical guidance for deploying LLMs in contexts where reliability and reproducibility are essential.

2 Experimental Design

The experimental framework is designed to isolate the effects of prompt structure, model architecture, and decoding randomness on output variability while maintaining controlled conditions. We employ a fixed structured task such as classification or information extraction, allowing outputs to be represented in a standardized format that facilitates direct comparison across repeated runs. Using a structured task reduces ambiguity in evaluation and aligns with prior research emphasizing controlled experimentation in prompt-based learning settings.

Prompt construction follows established practices in in-context learning by varying the number of example input-output pairs included within the prompt. Zero-shot prompts contain only task instructions, while few-shot and many-shot prompts include progressively larger numbers of examples. This design directly builds upon findings from Brown et al. (2020), which demonstrated that increasing examples can improve task performance, while extending the investigation toward whether additional examples also stabilize output structure across repeated trials.

Model comparison is incorporated to evaluate whether improvements in model architecture correspond to differences in output stability. Previous work has primarily focused on performance benchmarks rather than reproducibility across model generations, leaving open the question of whether newer models inherently reduce variability. By testing multiple model versions under identical conditions, we aim to isolate architectural contributions to stability.

Temperature is varied systematically to evaluate the impact of stochastic decoding on reproducibility. Since temperature modifies the randomness of token sampling, prior theoretical understanding suggests that lower values should produce more deterministic responses, whereas higher values may increase variability. Including temperature as an experimental factor allows us to empirically measure how controlled randomness interacts with prompt design and model selection.

All experiments (we use plural for clarity? Wait, we should continue in singular? Actually essay ok.) are executed through automated Python pipelines that interact with the model API under fixed parameter settings. For each

combination of prompt structure, model version, and temperature level, multiple repeated trials are conducted using identical input data. Outputs are logged in structured formats to enable quantitative analysis of agreement rates, formatting consistency, and response variability. This approach ensures that observed differences can be attributed to experimental variables rather than inconsistencies in execution.

The experimental design prioritizes reproducibility by maintaining version-controlled prompts, fixed datasets, and documented parameter settings. By explicitly controlling and varying key factors identified in prior literature, the study seeks to clarify how prompting strategies, model evolution, and probabilistic generation mechanisms jointly shape the stability of LLM outputs.