Dissecting Non-Determinism in Large Language Models

Mateus E. R. da Silveira[1], Ronaldinho V. C. Olivera[1], Alejandro N. Arroyo[1], Allan M. de Souza[2], Júlio C. dos Reis[2]

[1]Institute of Computing, Unicamp
[2]Institute of Computing, H.IAAC, Unicamp

Introduction

The Contradiction:

Scientific research relies on reproducibility, yet we are increasingly building deterministic science on stochastic Large Language Models (LLMs).

Key Facets of Unpredictability:

- Stochasticity: Enforcing determinism is possible but computationally expensive.
- Sensitivity: Minor input perturbations lead to major output variances.
- Evaluation Risk: The “LLM-as-a-Judge” paradigm suffers from inherent variance and introduces hidden bias.

Introduction

Objective: To advocate for consistency-oriented practices that manage these variables to ensure rigorous, reliable AI experimentation.

Non-Determinism as a Crisis of Reproducibility

While intuition suggests selecting the most probable tokens, deterministic search often leads to “**neural text degeneration**."

For human-like text, it is imperative to **reintroduce variance** through **stochastic sampling** strategies.

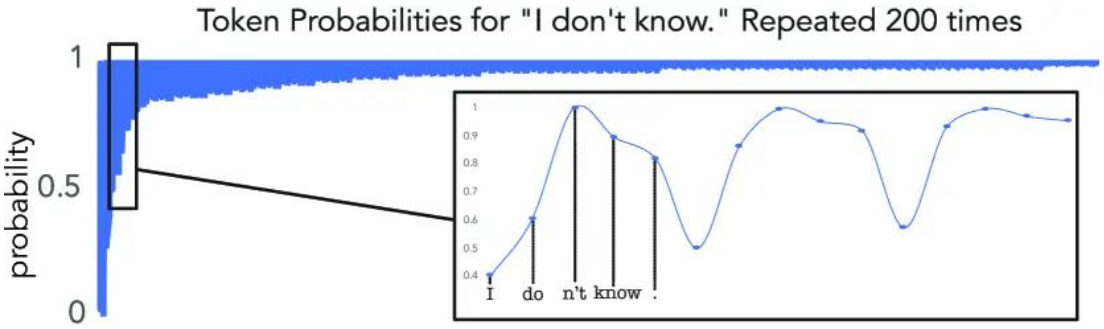


Figure extracted from Holtzman et al., 2020.

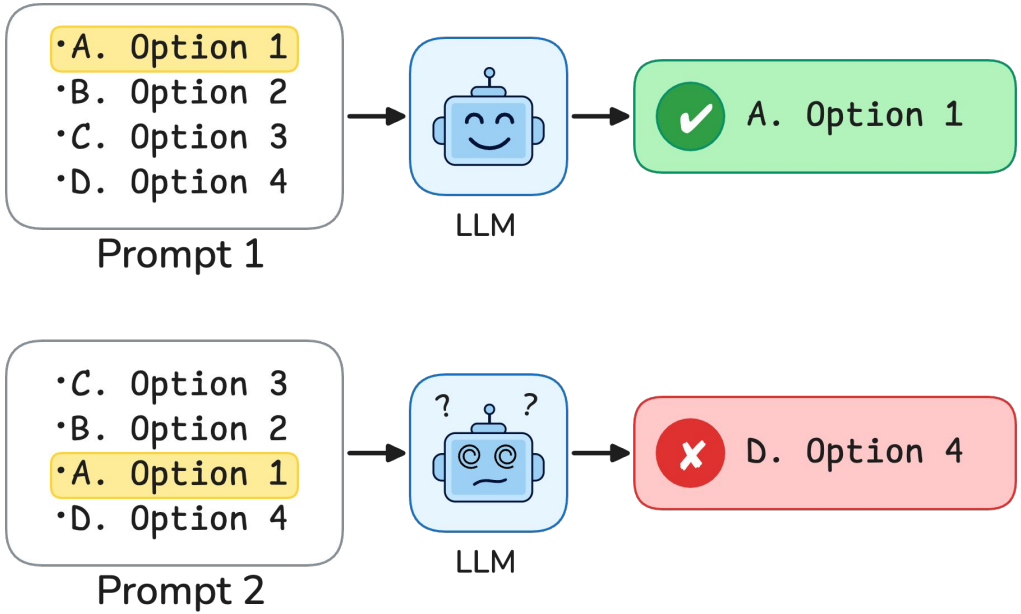
Non-Determinism as a Crisis of Reproducibility

- Setting **T = 0** is not a guarantee of deterministic results.
- **Root** causes come from **infrastructure-level variables** (e.g., dynamic batching, parallel execution, and floating-point arithmetic).
- Enforcing **deterministic** execution comes with a **heavy toll** in performance that is paid.

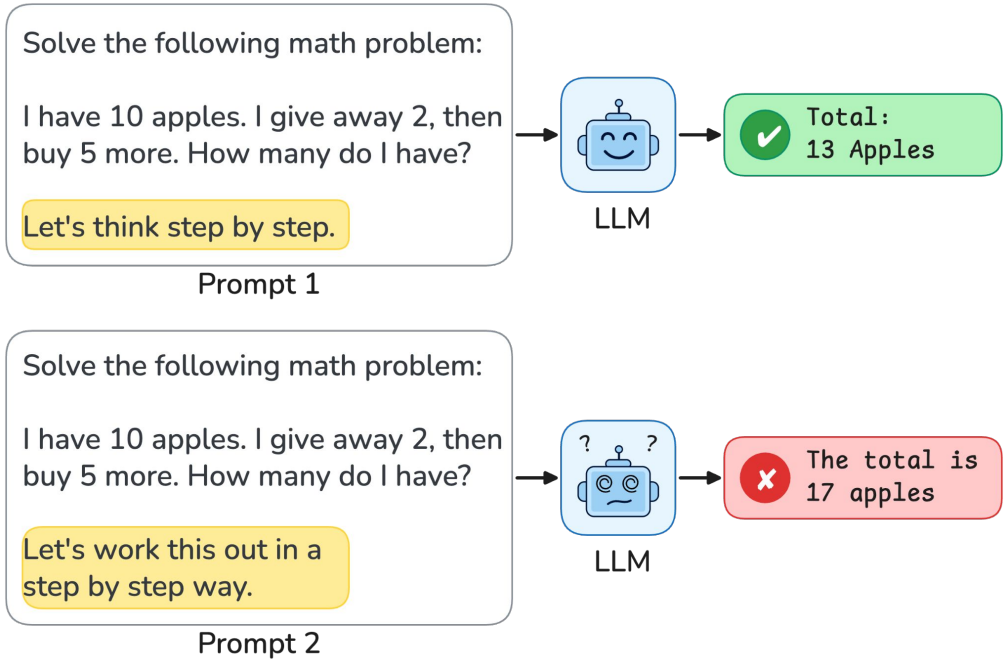
The Brittleness of Prompt Engineering

- Brittleness is a phenomenon about **high sensitivity** in LLM response
- Testing which format or ordering yields the best performance is not an easy task, as considering the full space of prompts formats makes the task an **intractable problem**
- “**brittleness**” is not a single issue but it manifests in lots of distinct ways

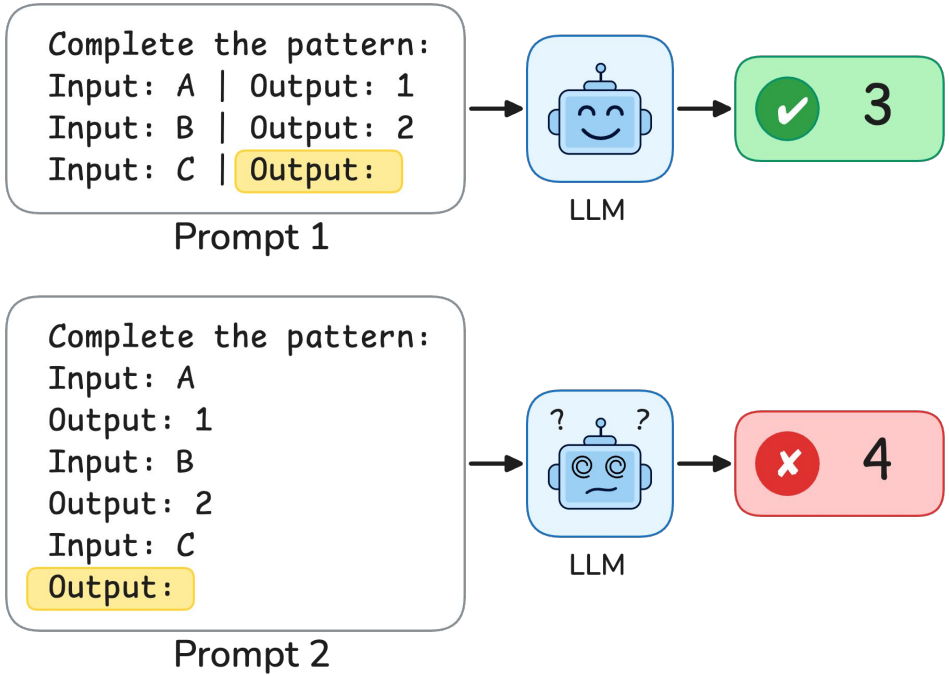
Wording Brittleness



Positional Brittleness



Formatting Brittleness



The Brittleness of Prompt Engineering

How to measure?

Exploring metrics like spread, sensitivity, and consistency helps refine prompts for edge cases.

How to mitigate?

Exposing the LLM to multiple prompt styles or using LLMs to self-generate them. But optimized prompts are highly model-specific.

LLM-as-a-Judge

Paradigm offers a flexible, scalable surrogate for assessing complex reasoning and helpfulness.

LLM judges are influenced by systematic "**Judgment-Specific Biases**" that occurs despite the answer validity

- **Position**: tendency to favor responses based on their placement.
- **Compassion** fade: evaluators are influenced by explicit model names.
- **Style**: preference for specific writing styles, visual formatting, or emotional tones.
- **Length**: tendency to favor more verbose responses.
- **Concreteness**: preference for responses containing specific details, such as citations, numbers, or complex terminology.

Conclusion

- Applying disciplined engineering constraints to minimize stochastic error.
- Reliability cannot be measured by accuracy alone; brittleness needs to account for sensitivity as well to distinguish architectural gains from “prompt overfitting.”
- Tailoring LLM-as-a-Judge strategies to mitigate inherent bias
- **We can't “fix” the models to be 100% deterministic, but we can “manage” them through better engineering.**

Acknowledgments

The study for this material was sponsored by Petróleo Brasileiro S.A. (PETROBRAS) as part of the project “Application of Large Language Models (LLMs) for Online Monitoring of Industrial Processes," developed in collaboration with the University of Campinas [01-P-34480/2024 - 62208].

Blogpost Page