Bias-Proof Data : Evaluating LLM Generalization

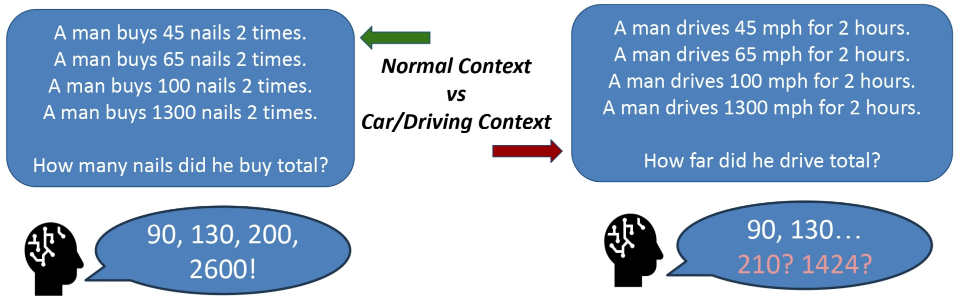
Joshua Tom, Computer Systems Engineering
Mentors: Ben Zhou, Assistant Professor and Xiao Ye, PhD Student
School of Computer and Augmented Intelligence

Objective and Research Question

Large language models (LLMs) are being employed in increasingly complex problem-solving tasks, however previous studies show how **patterns within pre-training data** can lead to unexpected failures.

The purpose of this study is to systematically identify any root causes in LLMs that may lead to unreliable model responses.

Problem Statement



A man buys 45 nails 2 times.
A man buys 65 nails 2 times.
A man buys 100 nails 2 times.
A man buys 1300 nails 2 times.

How many nails did he buy total?

Normal Context
vs
Car/Driving Context

A man drives 45 mph for 2 hours.
A man drives 65 mph for 2 hours.
A man drives 100 mph for 2 hours.
A man drives 1300 mph for 2 hours.

How far did he drive total?

90, 130, 200, 2600!

90, 130… 210? 1424?

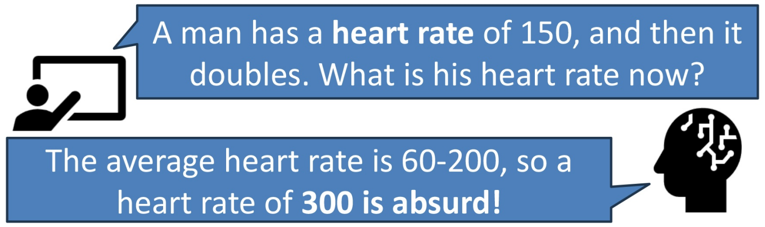
Within specific contexts, a model's performance on certain values drops drastically [1]. Previous work has shown that LLMs are vulnerable to co-occurrence bias [2], where a model prefers frequently co-occurred words over the correct answer. However, further experimenting shows specific contexts that impact model performance unrelated to co-occurrences.

To further investigate this issue, this work aims to answer:

1. What patterns with contexts cause these performance drops?
2. Are these patterns consistent between models?

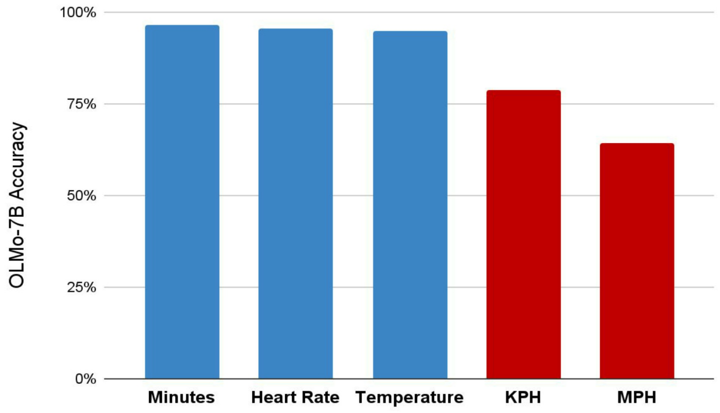
Testing for Inconsistency Patterns

Experiment 1 - Testing with Intuitively Biased Questions



A man has a **heart rate** of 150, and then it doubles. What is his heart rate now?

The average heart rate is 60-200, so a heart rate of **300 is absurd!**



Using subjects with commonly known values like heart-rate, we tested if words associated with a specific range of values would induce error.

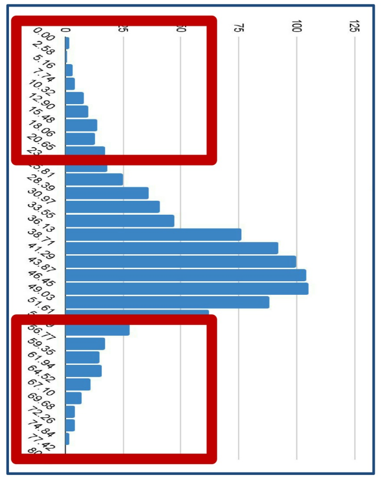
Results:
Intuitively biased units did **not** perform similarly to MPH or KPH.

This experiment did **NOT** reveal any other inconsistency inducing contexts, **suggesting that biases stem from non-intuitive sources.**

Experiment 2 - Searching for Volatility

Instead of modifying questions, one specific template was applied and the main subject word was changed.

Using ConceptNet, **1000** nouns were tested on the basic prompt: **There are [X] [nouns]. The number of [nouns] double. How many are there now?"**



Results:
Even though the question remained the same, certain nouns **had extremely high and low performance.** This suggests that there are **per-noun** biases!

Conclusion and Takeaways

Massive fluctuations in performance on a per-noun basis suggests the existence of noun-specific biases that emerge independently from pretraining co-occurrences.

Interestingly, uncommon words like 'fetoprotein', 'windscreen', and 'luge' performed with over 85% accuracy, but more common words like 'locker', 'shelter', and 'gun' had less than 30%.

Future Work

1. Identifying if this trend is consistent between model architectures and patterns (subtraction, division, etc).
2. Seeing if the biases are dataset specific, testing if the trend appears on other models trained with the same pre-training information.

References

1. Yu, X., Zhou, B., Cheng, H., & Roth, D. (2024). ReasonAgain: Using Extractable Symbolic Programs to Evaluate Mathematical Reasoning. arXiv preprint arXiv:2410.19056.
2. Kang, C., & Choi, J. (2023). Impact of co-occurrence on factual knowledge of large language models. arXiv preprint arXiv:2310.08256.

FURI
ASU Ira A. Fulton Schools of Engineering
Arizona State University