Paper: 30054

Artificial Intelligence or Artificial Stupidity?

Salvatore Vella, Salah Sharieh
Alex Ferworn
Toronto Metropolitan University
Sal.Vella@torontomu.ca

Toronto Metropolitan University

IARIA

Salvatore Vella

- Ph.D. student at Toronto Metropolitan University
- Research in use of Generative AI in white collar use and bias in small LLMs

- Re-retired technology executive
  - Former IBM Software CTO
  - Former VP, Technology Strategy and Innovation, Royal Bank of Canada
  - CEO, Toogood Ventures, technology strategy

Introduction

- Small Large Language Models (LLMs) with fewer than four billion parameters are now common in personal and embedded systems. They perform tasks like summarization, voice assistance, and email generation.

- However, their reasoning capabilities remain underexplored. This study investigates whether these models truly reason or simply match patterns based on shallow correlations. We test their robustness by slightly altering prompt formats in structured reasoning tasks and observe significant variations in accuracy—even when the correct answer is visible in the prompt.

Research Questions

This study explores key questions about the cognitive behaviour of small LLMs:

1. Do small LLMs demonstrate reasoning or just surface-level pattern recognition?
2. Does including the correct answer in a prompt help or hinder performance?
3. Can prompt structure itself act as an adversarial vector?
4. How consistent are small LLMs across equivalent prompts with different formats?

Methodology

- We evaluated several small-scale LLMs:
- Models: LLaMA-3 (1B, 4B), Google Gemma (1B, 4B), Alibaba Qwen (1.5B, 3B), Microsoft Phi-3 (4B), IBM Granite (2B), OpenAI GPT-4o and GPT-4.1 (nano and mini variants).
- Datasets: CommonsenseQA and OpenBookQA (2000 multiple-choice questions each).
- Prompt Formats: Base (control), Example, Simple Primed, and Reverse Primed. The latter two included the correct answer to test the impact of formatting and positioning.

Datasets and Benchmarks

- CommonsenseQA: Tests everyday reasoning with five-choice questions.
- OpenBookQA: Combines elementary science facts with inference-based reasoning.
- These datasets help distinguish between pattern-based and reasoning-based behavior.

Statistical Validation

- Paired t-tests and McNemar's tests confirmed that prompt structure—not randomness—drives performance variation.
- Most results were statistically significant ($p < 0.05$), highlighting the lack of format invariance in small LLMs.

Ethical and Security Implications

- Prompt-order sensitivity introduces a new attack surface.
- Adversaries can manipulate model outputs by altering the prompt structure.
- Responsible deployment requires:
  - Adversarial training
  - Prompt filtering
  - Consistency checks

CommonsenseQA Accuracy By Prompt Condition

TABLE I. COMMONSENSEQA ACCURACY BY PROMPT CONDITION

Model	Base	Simple	Reverse	Example
gemma-3-1b	42.10	91.65	7.30	38.05
gemma-3-4b	64.70	96.40	43.35	62.15
gpt-4.1-mini	79.45	92.75	87.05	78.45
gpt-4.1-nano	73.30	96.80	84.15	71.70
gpt-4o-mini	78.75	90.60	85.80	77.10
granite-3.3-2b-instruct	64.90	92.85	73.35	64.70
llama-3.2-1b-instruct	52.10	96.70	9.20	26.10
llama-3.2-3b-instruct	65.95	97.05	55.55	61.85
phi-3-mini-4k-instruct	72.85	96.35	84.30	67.70
qwen2.5-1.5b-instruct-mlx	62.60	89.30	52.10	61.10
qwen2.5-3b-instruct	72.25	96.05	69.25	72.85

OpenBookQA Accuracy By Prompt Condition

TABLE II. OPENBOOKQA ACCURACY BY PROMPT CONDITION

Model	Base	Simple	Reverse	Example
gemma-3-1b	41.90	95.70	10.35	29.80
gemma-3-4b	66.15	97.35	61.95	65.65
gpt-4.1-mini	89.40	96.00	94.80	89.30
gpt-4.1-nano	80.50	98.00	96.40	79.15
gpt-4o-mini	87.30	93.95	93.35	85.60
granite-3.3-2b-instruct	68.30	94.50	82.10	65.65
llama-3.2-1b-instruct	44.80	99.15	8.30	22.65
llama-3.2-3b-instruct	67.00	98.90	69.15	61.45
phi-3-mini-4k-instruct	80.40	98.25	90.05	78.35
qwen2.5-1.5b-instruct-mlx	60.05	88.60	65.15	55.50
qwen2.5-3b-instruct	65.85	96.80	75.45	66.45

CommonsenseQA Model Accuracy By Prompt Condition

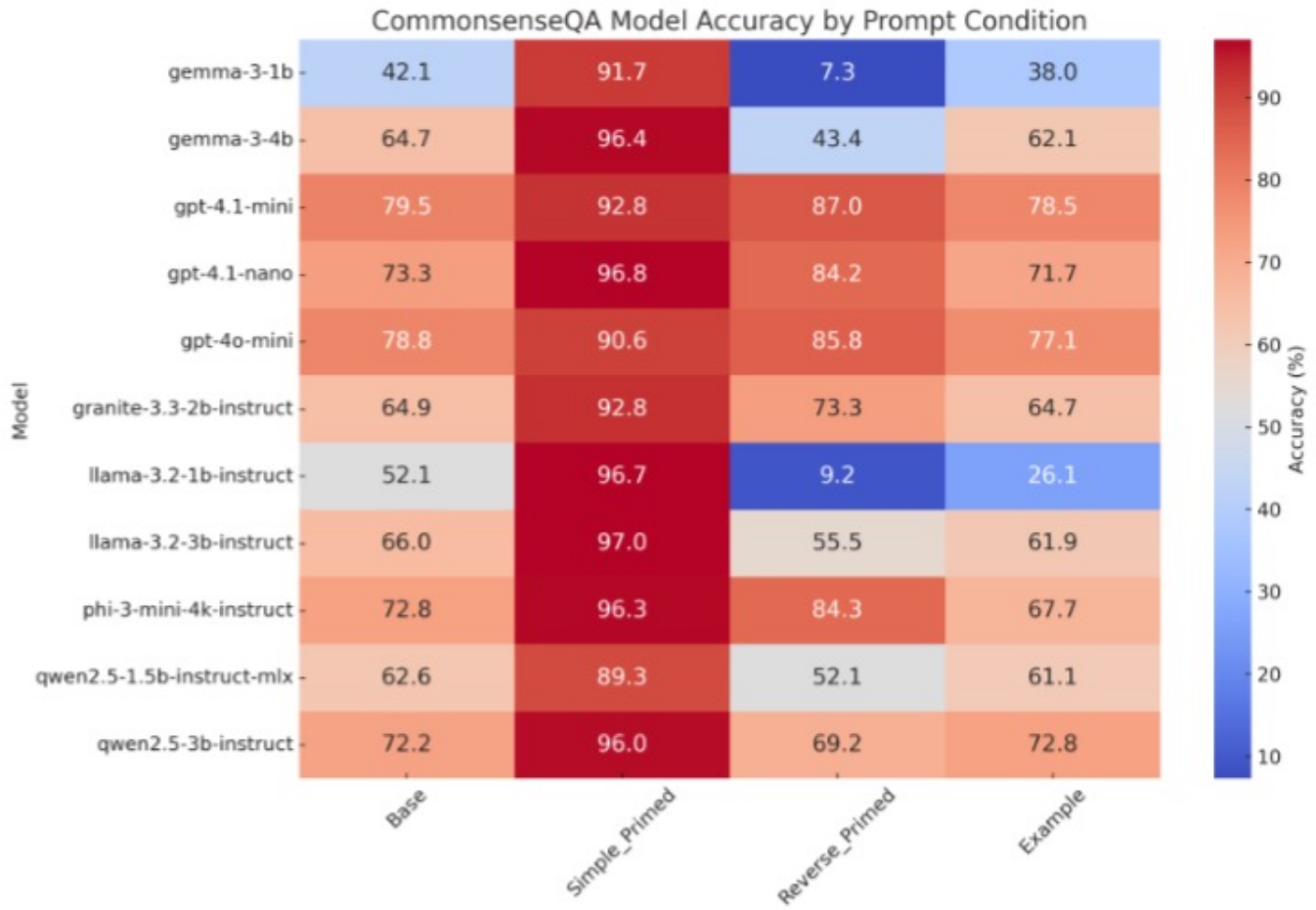

Figure 3: CommonsenseQA Model Accuracy by Prompt Condition.

OpenBookQA Model Accuracy By Prompt Condition

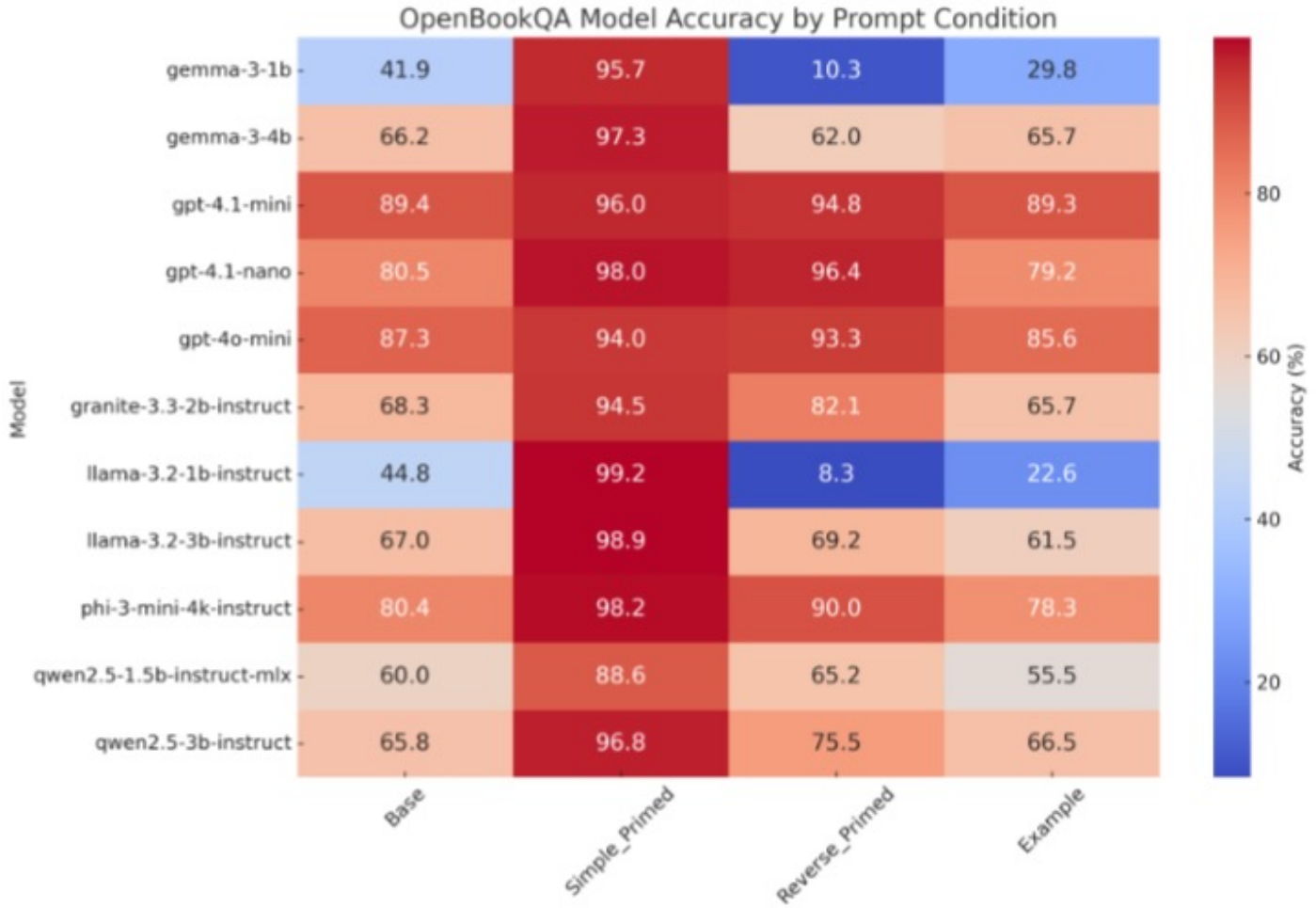

Figure 4: OpenBookQA Model Accuracy by Prompt Condition.

CommonsenseQA Performance Visualization

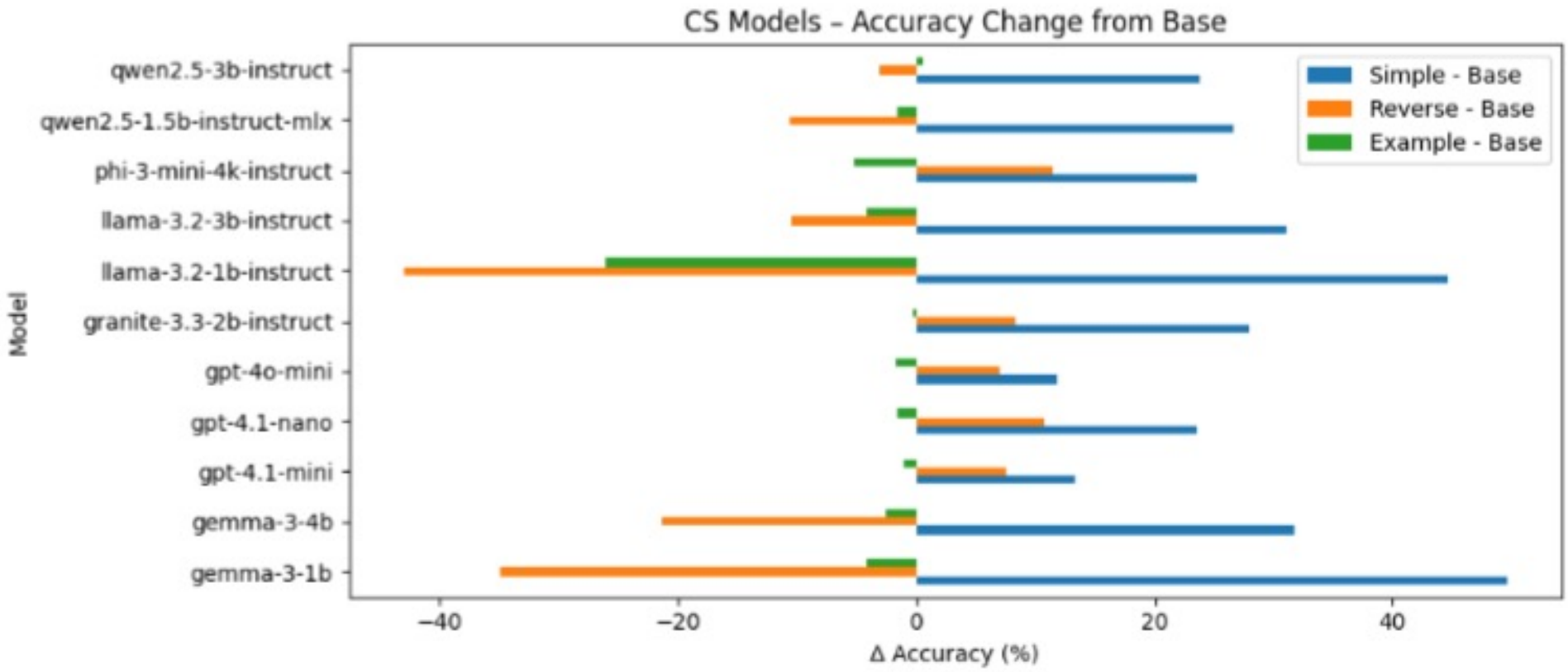

Figure 5: CommonsenseQA Model Perturbation by Prompt Condition.

OpenBookQA Performance Visualization

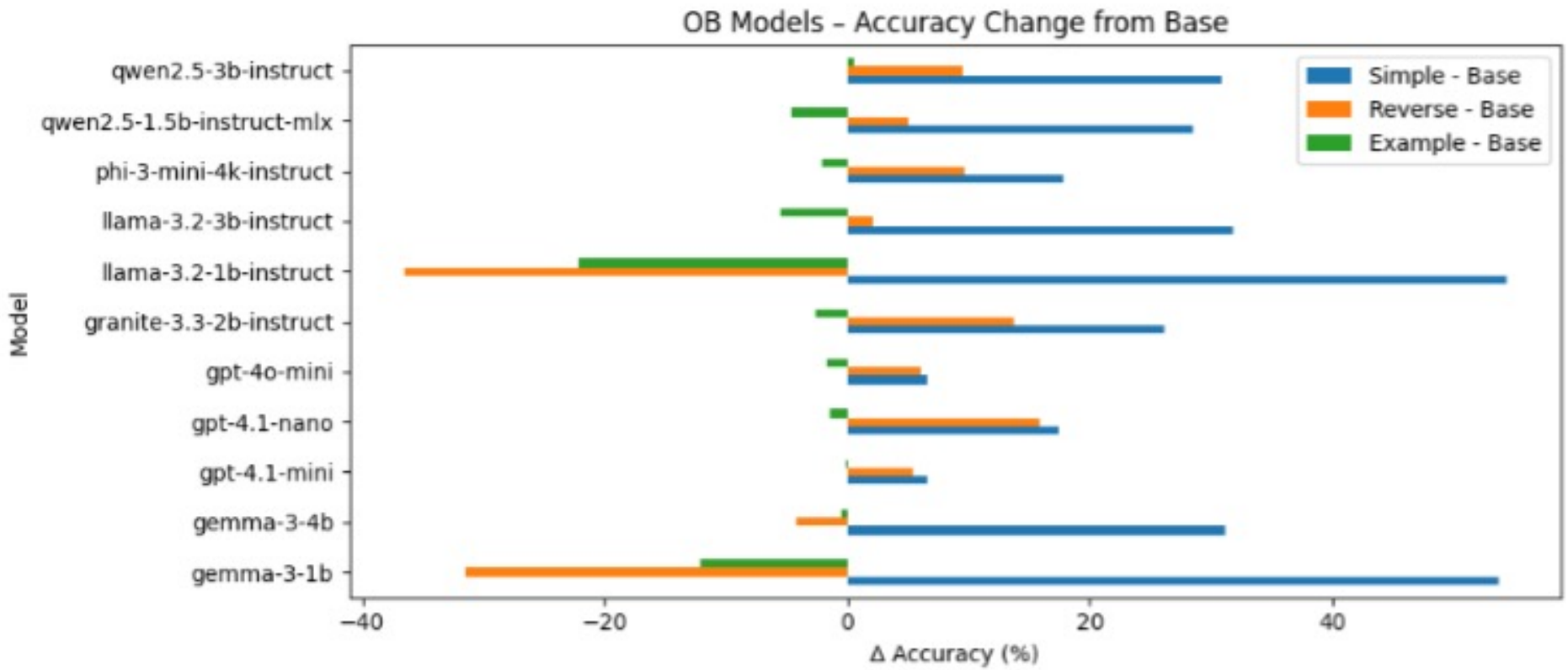

Figure 6: OpenBookQA Model Perturbation by Prompt Condition.

Key Findings

- Performance improved when the correct answer appeared in the same position across the priming example and target question.
- Accuracy dropped—often below random guessing—when the correct answer was moved.
- 1B-parameter models were especially vulnerable, indicating reliance on heuristics.
- Even 4B models and OpenAI's mini and nano-series showed significant degradation under prompt reordering.

Behavioral Analysis

- Small LLMs mimic human-like biases but lack true reasoning depth.
- They often replicate positional or lexical patterns from prior context rather than assess logical relationships.
- This makes them vulnerable to adversarial prompt manipulation, which can lead to incorrect or biased outputs—posing risks in safety-critical domains like healthcare, finance, and security.

Conclusion & Future Work

- Small LLMs rely on pattern matching, not reasoning.
- They are highly sensitive to superficial prompt changes.
- This limits their trustworthiness in applications requiring consistent logic.

- Future directions:
  - Develop prompt normalization techniques.
  - Explore adversarial instruction tuning.
  - Use ensemble prompting to reduce structural bias.
  - Expand benchmarks to better evaluate robustness.

Q&A

- Thank you for your attention.
- Sal.Vella@torontomu.ca