Modeling Understanding of Story-Based Analogies Using Large Language Models

Kalit Inani, Keshav Kabra, Vijay Marupudi, Sashank Varma
{kinani3, keshav.kabra, vijaymarupudi, varma}@gatech.edu

Introduction

Large Language Models (LLMs) outperform humans on many benchmarks, yet their analogical reasoning—especially for story-based analogies that hinge on higher-order causal structure—remains under-explored.

Building on Webb et al. (2023)[1], we ask:

1. Do encoder-based LLM embeddings capture the semantic representation of analogies?
2. Can self-generated hints improve model–human alignment?
3. Which architectures and model sizes best match human performance patterns?

Method

Datasets and Tasks

- 18 classic story-analogy items[2].
- For each item: 1 source story + 2 targets (true vs false analogy).
- Human baseline: 84.7% accuracy[1].

Example Analogy Problem:

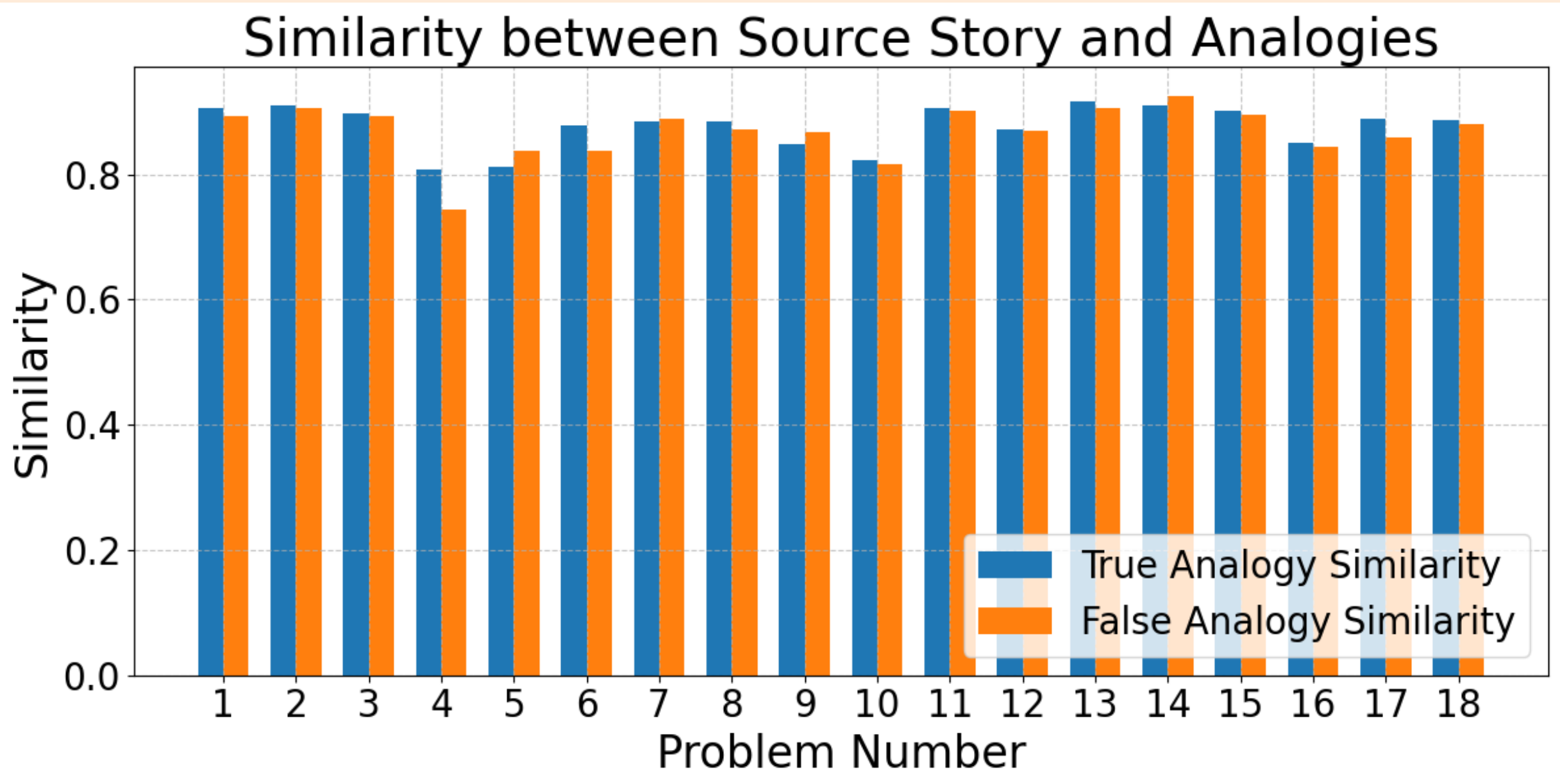
- **Source Story:** *Mrs. Jackson wanted a salary increase. The principal increased his own salary by 20% but said there wasn't enough money for teachers. Mrs. Jackson became angry and decided to take revenge by setting fire to the principal's office.*
- **True Analogy:** *McGhee wanted vacation on land. The captain announced he would take a vacation in the mountains but everyone else must remain on ship. McGhee became upset and decided to get revenge by blowing up the captain's cabin.*
- **False Analogy:** *McGhee wanted vacation on land. McGhee became impatient and tried to blow up the captain's cabin. After this, the captain announced his vacation but said everyone must stay to repair the ship.*
- **Key Difference:** The true analogy maintains the causal structure (unfair treatment → revenge), while the false analogy changes the event sequence and motivation.

Approaches

- **Sentence-Embedding Test:** BERT-based cosine similarity between source embedding and each target embedding.
- **Generative Reasoning**
    - **Models:** GPT-4o-mini and GPT-4o, LLaMA 3.1-8B and 70B.
    - **Prompts:** (a) Conventional – choose "Story A/B" (b) Enhanced – same prompt plus self-generated causal "hints".

Results

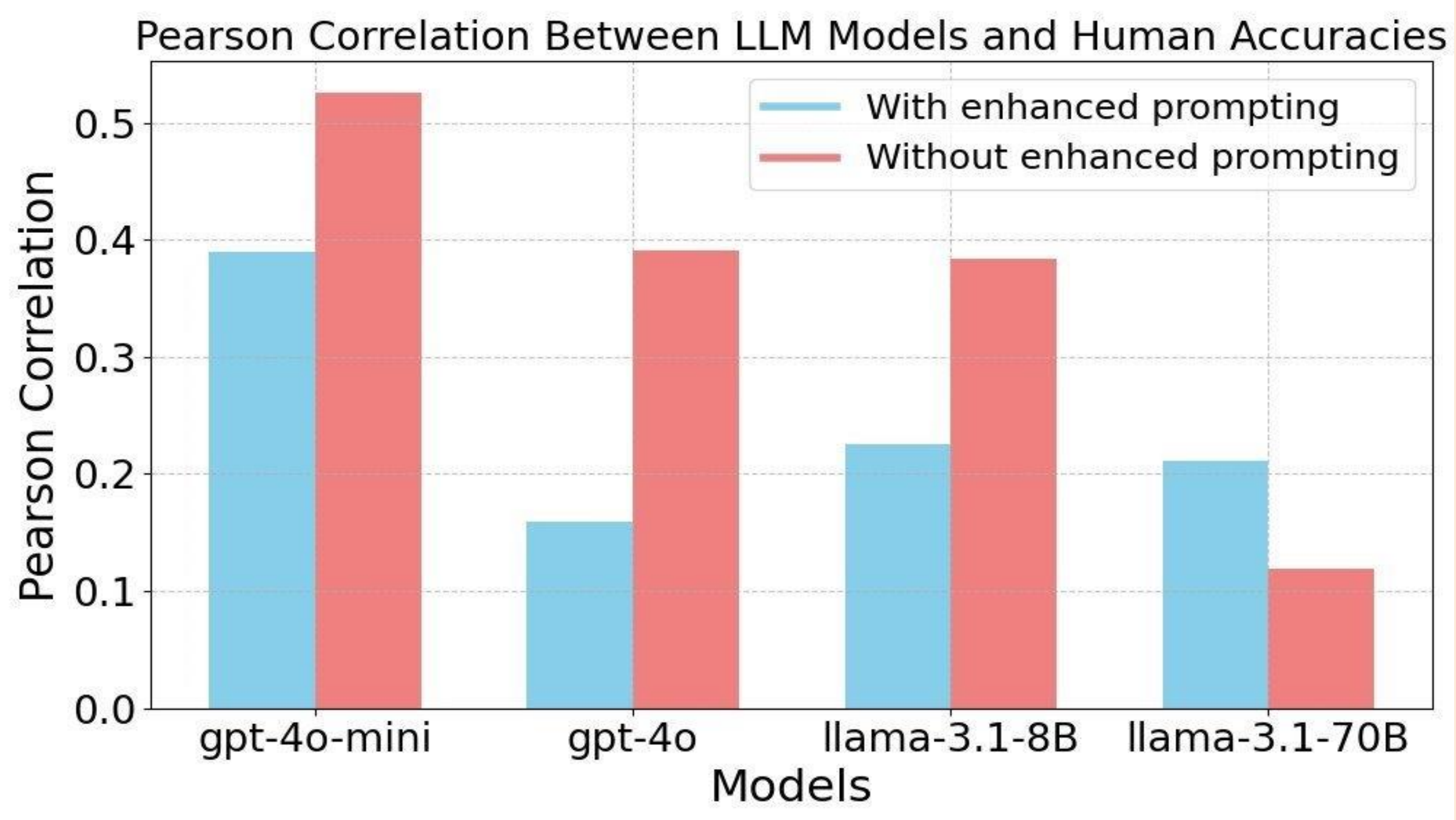
1) BERT embeddings distinguished true from false analogies 78% of the time but showed zero correlation with human item difficulty.



2) Enhanced prompting boosts every model by 4% - 9%. LLaMA 3.1-70B + enhanced prompt has an accuracy (91.5%) exceeding humans (84.7%).

Model	Conventional Prompt	Enhanced Prompt
Humans	0.847	N/A
GPT-3 (Webb et al., 2023)	0.75	N/A
GPT-4o-mini	0.7011	0.7411
GPT-4o	0.8233	0.8850
LLaMA-3.1-8B-Instruct	0.6528	0.7472
LLaMA-3.1-70B-Instruct	0.8538	0.9150

3) Smaller GPT-4o-mini best mirrors the correlation with humans (r = .53). Enhanced prompting often reduced the correlation with human performance despite improving accuracy.



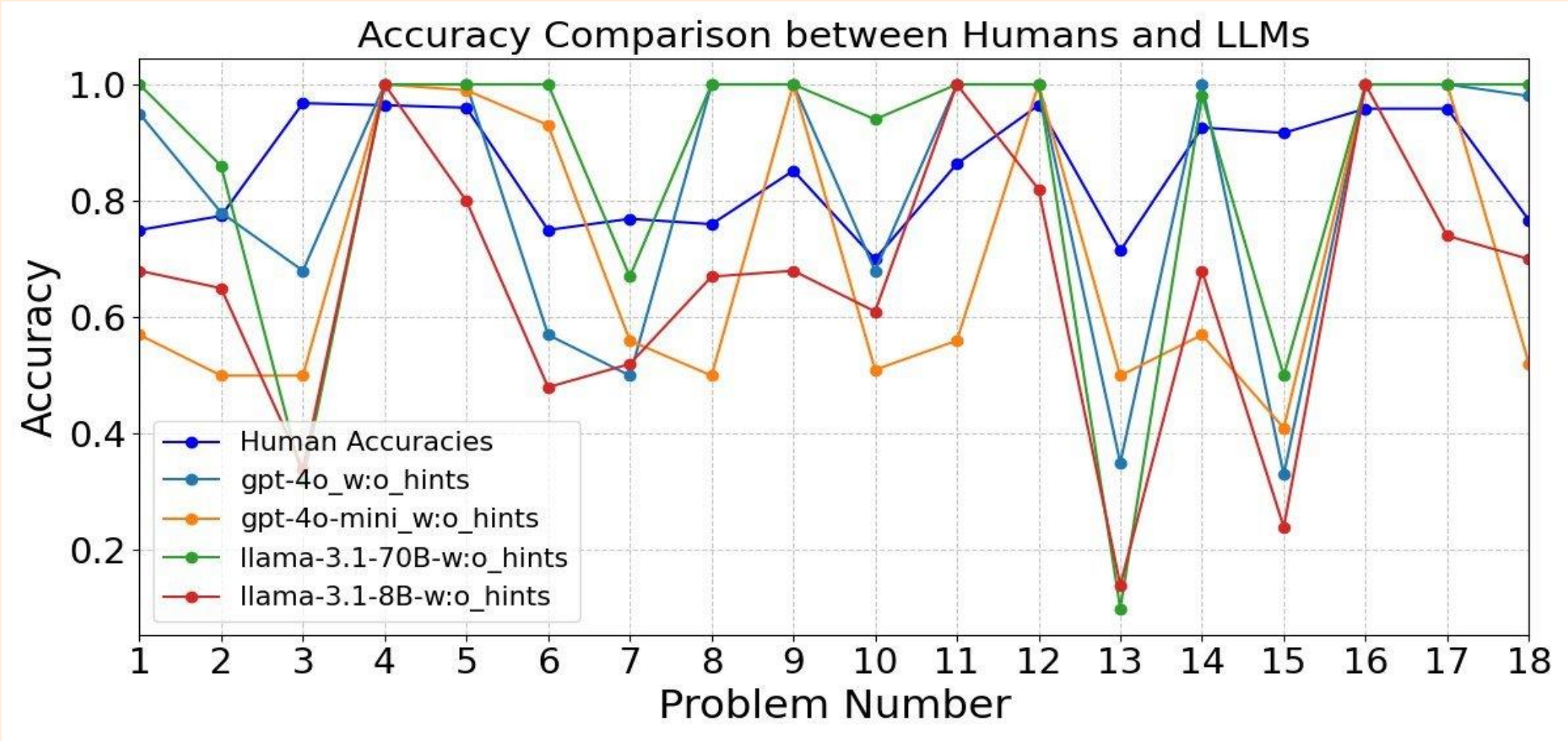
Main References

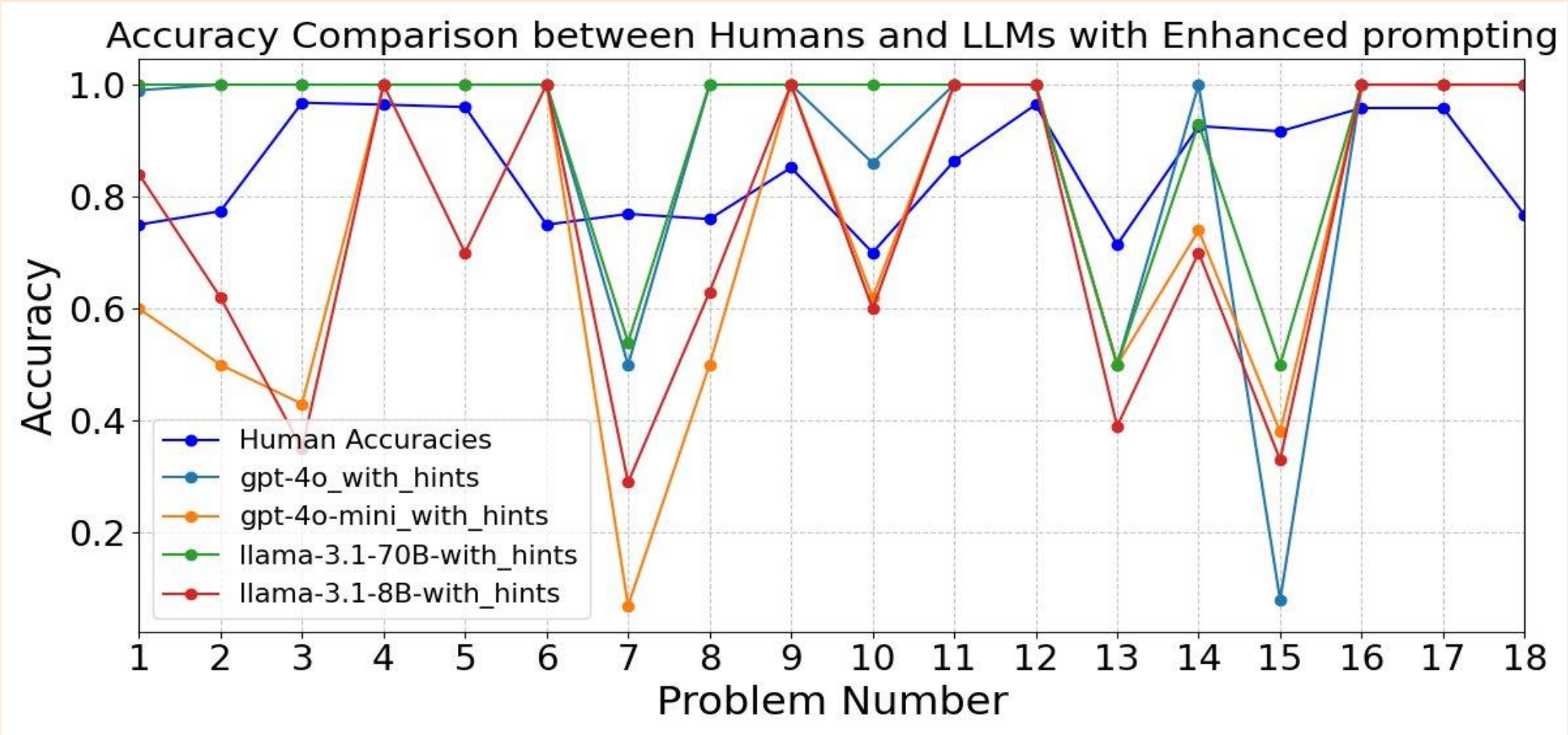
1. Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models.Nature Human Behaviour, 7, 1526–1541.
2. Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: separating retrievability from inferential soundness. Cognitive psychology, 25(4), 524–575.
3. Lewis, M., & Mitchell, M. (2024). Evaluating the robustness of analogical reasoning in large language models.

Find our work! Scan here

Results (contd.)

4) Common failure modes: tracking motivations vs. surface events, maintaining event sequences, handling nested causal chains.





Discussion

1. **The Accuracy-Alignment Paradox**
   Higher overall accuracy does not guarantee human-like reasoning patterns. While LLaMA 3.1-70B achieved 91.5% accuracy with enhanced prompting, only GPT-4o-mini (conventional) showed significant correlation with human performance. This reveals a fundamental disconnect between "getting it right" and "getting it right for the same reasons as humans."
2. **Enhanced Prompting Effectiveness**
   Self-generated causal hints consistently improved overall performance across all models without manual engineering. However, enhanced prompting often reduced the correlation with human performance, suggesting that the prompts do not necessarily make models reason more like humans (perhaps due to ceiling effects).
3. **Model Size Effects**
   Larger models achieve higher accuracy but paradoxically show weaker alignment with human difficulty patterns. This suggests current scaling approaches may optimize for pattern matching rather than human-like causal understanding.

Future Work: (a) Expand to larger, cross-domain story-analogy corpora. (b) Apply findings to other causal reasoning tasks. (c) Develop contamination-free test sets using newly crafted analogies. (d) Design training objectives optimizing for human-like reasoning processes.