Sally-Anne False Belief Test for LLMs

Mimansa Jaiswal
Independent Researcher
mimansa.jaiswal@gmail.com

Abstract

This paper critically examines the use of the Sally-Anne false-belief test as a benchmark for evaluating large language models (LLMs) like GPT-4. We demonstrate that while these models can pass certain versions of the test, they struggle with subtle variations that require true understanding of non-continuous observations and implicit actions. Our findings suggest that current evaluation methods may overestimate LLMs' capabilities in reasoning about beliefs and knowledge transfer.

1 Introduction

Evaluating the social impact of generative AI necessitates going beyond traditional metrics to assess nuanced aspects like theory of mind [Ullman, 2023, van Duijn et al., 2023]. Inspired by the Sally-Anne False Belief Test used in GPT4 evaluation paper [Bubeck et al., 2023], we introduce a framework for analyzing how large language models (LLMs) handle scenarios involving unshared actions and information asymmetry.

2 Methodology

Our framework comprises three stages. First, **Instinct Derivation** involves deconstructing prompts to identify underlying concepts like timing, unshared actions, and knowledge transfer. Next, **Prompt Modification** focuses on introducing variations to test specific aspects such as object transformation, information movement, and knowledge transfer modalities. Finally, **Generation Evaluation** consists of analyzing model outputs for accuracy, error types, and reasoning capabilities. We analyze GPT-3.5 and GPT-4's responses to a series of prompts, starting with a base scenario involving file manipulations in a shared Dropbox folder. We then introduce modifications to assess the models' understanding of renaming, syncing, privacy settings, and object movement within the digital space.

3 Findings and Discussion

3.1 Dropbox Scenario: Unveiling Implicit Knowledge Gaps

3.1.1 Base Scenario

Prompt: "Alice puts a file 'photo.png' in /shared_folder/photos. Bob moves it to /shared_folder/tmp without notifying Alice. Where will Alice look for the file?" **Results:** Both GPT-3.5 and GPT-4 correctly answered that Alice would look in /shared_folder/photos, demonstrating basic understanding of information asymmetry.

3.1.2 Folder Renaming

Prompt: "Bob renames shared_folder to common_folder without notifying Alice. Where will Alice look for the file?" **Why and Results:** This modification tests the model's understanding of unshared transformations within a shared environment. GPT-4 correctly answered "/common_folder/photos", while GPT-3.5 inconsistently suggested Alice would still look in the original location, revealing limitations in understanding system-wide changes.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

3.1.3 Syncing Issues

Prompt: "Alice disconnects from the internet. Bob renames the folder. Where will Alice find the file on her offline device?" **Why and Results:** This scenario introduces the concept of information movement and synchronization, testing the model's grasp of real-time updates versus offline states. GPT-4 correctly identified that Alice's offline device would retain the original structure, while GPT-3.5 faltered, suggesting the renamed location.

3.2 Beyond Dropbox: Exploring Real-World Analogies

3.2.1 Lending Books

Prompt: "I place a bookmark on page 24, lend the book to a friend who moves it to page 32. Where do I expect the bookmark when the book is returned?" **Why and Results:** This scenario tests the model's understanding of unshared actions and the lack of knowledge transfer in a lending context. Both models correctly answered page 24, demonstrating understanding of the lender's limited knowledge.

3.2.2 Phone Call Scenario

Prompt: "During a phone call, my friend moves the bookmark. Where do I expect it to be after the call?" **Why and Results:** This scenario explores the model's ability to handle the lack of visual information transfer during a phone call. Both models incorrectly suggested the new location, failing to recognize the lack of visual information transfer in a phone call.

3.3 The Role of Static Object Placements and Human Actors

3.3.1 Key Placement Scenario

Prompt: "I place a key on the 3rd hook, go to sleep. My boyfriend moves it to the 4th hook. Where do I expect it upon waking?" **Why and Results:** This prompt tests the model's ability to understand static object placements and the impact of human actors on knowledge transfer. Both models correctly answered the 3rd hook, demonstrating understanding of knowledge limitation during sleep.

3.3.2 Cat Interaction Scenario

Prompt: "I place a key on the 3rd hook, go on vacation. My cat plays with the keys. Where do I expect the key upon return?" **Why and Results:** Introducing a non-human actor tests the model's ability to process actions by different agents and the implications for knowledge transfer. Both models struggled, often suggesting the floor or an uncertain location, revealing limitations in processing non-human actors and imprecise actions.

3.4 Question Framing and Word-Based Leakage

3.4.1 Explicit vs. Implicit Wording

Prompt Pair: 1. "How many coffee cups would I think I have?" 2. "How many coffee cups would I look for?" **Why and Results:** This pair of prompts examines the impact of cognitive verbs versus action-oriented verbs on the model's performance. GPT-4 performed better with "think," suggesting sensitivity to cognitive verbs, while both models struggled with "look for," indicating difficulties with action-oriented scenarios.

4 Conclusion

Our findings demonstrate the potential of modified Sally-Anne-like tests for evaluating theory of mind in LLMs. While GPT-4 shows progress in this domain, both models exhibit limitations, particularly in handling implicit knowledge transfer when actions lack explicit visual or spatial grounding. The models' performance varies significantly based on the framing of questions, the kind of knowledge transfer expected and the nature of actors involved.

References

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Max J van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco R Spruit, and Peter van der Putten. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. *arXiv preprint arXiv:2310.20320*, 2023.