Universiteit
Leiden
The Netherlands

Opleiding Informatica

The Effect of Visual Cues
on Theory of Mind abilities
in Multimodal AI

Razo van Berkel

Supervisors:
Max van Duijn, Peter van der Putten & Tom Kouwenhoven

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl 01/06/2024

Abstract

Large Multimodal Models (LMMs) are becoming increasingly more powerful, giving rise to questions on their intelligent capabilities across multiple domains of information modality. An important social-cognitive capability of humans is Theory of Mind (ToM): the ability to infer *mind-states* of others. Although this is already a well-researched in Large Language Models (LLMs), this capability is not as much yet researched in multimodal AI. In this paper, we investigate the added value of *visual cues* in ToM assessments on LMMs. We construct a framework for running ToM assessments on LMMs, that contain both text and image information. Consistent with the literature, we find that our ToM assessments are easily solved by modern GPT-4 family models, in their text-only form. We used the constructed framework to structurally evaluate the effect of visual cues on ToM assessments in *GPT-4o*. Taking away textual information from the tests, only to leave this information as found in the supplementary test images, we still see the LMM perform the tests generally very well. The results show that generally, the model confidence and correctness increases in presence of a visual cue, with varying levels of degraded textual information. The paper concludes with three *test design criteria*, where we summarize the findings and considerations for future assessment of social-cognitive skills on Multimodal AI.

Keywords: Theory of Mind, Multimodality, Large Multimodal Models, Visual Cues

Contents

1 Introduction 1

2 Background and Related Work 2
2.1 Theory of Mind and test overview 2
2.2 Large Language Models 4
2.3 Large Multimodal Models 6
2.3.1 LMM overview 6
2.3.2 Visual Question Answering 7
2.4 Research test set 7
2.5 Related work: ToM in Large Multimodal Models 9

3 Method 10
3.1 Initial experiments 10
3.2 Pipeline and configuration 11
3.3 Defining pipeline tests 12
3.3.1 Test set deviations 13
3.3.2 Test question variations 14
3.3.3 Example instruction-question pair 15
3.3.4 Total test volume 15
3.4 Answer evaluation 16
3.4.1 Correctness and Manual rating 16
3.4.2 Confidence 17
3.4.3 Comparison to human levels 19
3.5 Running tests 20
3.6 Validity and reliability 21

4 Results 22
4.1 Sally-Anne 22
4.2 Strange Stories 22
4.3 Imposing Memory 22
4.4 Accuracy and confidence overview 25
4.5 On inter-rater agreement 27

5 Discussion 28
5.1 Reflection on results 28
5.2 Manual rating: issues and observations 29
5.3 Test limitations 29
5.4 Further research 30
5.5 AI energy consumption 31

6 Conclusions 32

References 35

A Appendix A **35**
A.1 Definitions . . . 35
A.2 Source Code . . . 36
A.3 Answer Type Overview . . . 36

B Appendix B: Experiments **37**
B.1 Pilot Prompts . . . 37
B.2 Image examples . . . 38
B.3 Image Understanding . . . 39

C Appendix C **39**
C.1 System Messages . . . 39
C.2 Name Deviation List . . . 40
C.3 Commonly Used Special LLM Tokens . . . 40

D Appendix D **41**
D.1 Examples Found during Manual Ratings . . . 41
D.1.1 Example 1 . . . 41
D.1.2 Example 2 . . . 41
D.1.3 Example 3 . . . 41
D.1.4 Example 4 . . . 42
D.1.5 Example 5 . . . 42
D.1.6 Example 6 . . . 42
D.2 Sample Log Probabilities . . . 43
D.3 Filtered Log Probabilities . . . 43

E Appendix E: Combined Result Plots **44**
E.1 Resulting Visualizations SA1 and SA2 . . . 44
E.1.1 SA 1 per Masking Level . . . 44
E.1.2 SA 2 per Masking Level . . . 45
E.2 Resulting Visualizations SS . . . 46
E.2.1 Results per Test . . . 46
E.2.2 Results per Masking Level . . . 47
E.3 Resulting Visualizations RM . . . 47
E.3.1 All Questions . . . 47
E.3.2 Intentionality Only . . . 48

F Appendix F: Individual Test Result Plots **48**
F.1 Accuracy Plots . . . 48
F.1.1 Sally Anne 1 and 2 . . . 48
F.1.2 Imposing Memory . . . 49
F.1.3 Strange Stories . . . 50
F.2 Confidence Plots . . . 53
F.2.1 Sally Anne 1 and 2 . . . 53
F.2.2 Imposing Memory . . . 54
F.2.3 Strange Stories . . . 55

1 Introduction

Artificial intelligence (AI) technology has experienced rapid growth and development in recent years, transforming a variety of sectors and enabling new forms of human-computer interaction.

Large Language Models (LLMs) represent the State-of-the-Art, consumer-ready advancement of Natural Language Processing (NLP), a subfield of AI. NLP is a subfield of linguistics, computer science, and artificial intelligence concerned with modelling language generation and understanding. These models are usable for a plethora of different, general and specialized, text-to-text tasks. This includes more traditional NLP tasks like *Sentiment Extraction* or *Abstractive Summarization*.

This also spawns research in exploring their capabilities in understanding and interpreting human cognitive and emotional states. A key aspect of this is Theory of Mind (ToM), which is the ability to infer mind-states, like beliefs and intentions, in others [2]. Theory of Mind in Large Language Models is already researched extensively, with different types of ToM tests [32, 18, 12]. There are also multiple benchmarks like *ToMBench* and *ToMChallenges* [7, 21].

With rising popularity of Large Language Models, a next step is the integration of multimodal capabilities, forming Large Multimodal Models (LMMs), sometimes also referred to as *Multimodal Large Language Models* (MLLMs) or *Vision Large Language Models* (VLMs) [44, 40]. These LMMs not only understand text but can also process and interpret images, videos, and other sensory data to provide more contextually enriched responses. The expanding availability and capabilities of powerful LMMs leads many scientists to evaluate different types of intelligent capabilities of such models. In this thesis, we will examine the effect of multimodality on ToM by running ToM tests on an LMM, with and without supplementary visual cues. We do this because many ToM tests that were ran in humans, contain supplementary visuals; but the same ToM tests as ran on AI models, was run without these visuals.

The first leading research papers are appearing in this field, like *MMTOM-QA*, presenting a benchmark for assessing multimodal performance in Theory of Mind [15]. They write clearly in their abstract that ”that large language models and large multimodal models still lack robust ToM capacity” [15]. However, their tests had specific limitations, like only including videos of household scenarios, lacking diversity. They also indicate to lack divesity in ToM concepts, like desires, emotions and constraints. After shortly noting these limitations, they proceed to state that further, more thorough research is needed. Another example is VToM, as presented by Chen et al. [6]. Chen et al. claims to find emergent ToM reasoning in LMMs, based on their VToM architecture. However, they also clearly indicate that there is a problem with available data and that further research is needed. These are both new papers on the subject of ToM in Multimodal AI, and they are part of the small amount of research in this field. Furthermore, both of these papers conclude that there is further work in diversifying tests and improving their quality. This, combined with the lack of volume and test coverage in multimodal ToM assessments, indicate existence of a *research gap*. In our research, we aim to contribute to filling this research gap. Equation 1 will be utilized to quantitatively assess the impact of Visual Cues (VCs) on the models' ToM capabilities. This Equation is an indication of the performance and will consist of multiple performance measurements. We compare the model performance, with and without VCs.

$$\text{performance\_delta} = \delta(\text{performance\_text}, \text{performance\_visual}) \tag{1}$$

The content of this bachelor's thesis is as follows. The central focus of this research is to determine how these models integrate and utilize visual cues alongside textual data to interpret and predict human mental states and intentions. The research question—*How do Large Multimodal Foundation Models perform in multimodal Theory of Mind assessments, and what is the relative value of visual cues in their performance?*—guides the exploration of the ToM capabilities of advanced multimodal AI, in this case OpenAI's GPT-4o. For understanding multimodal AI, we compare the technical details of LLMs to those of LMMs and other Visual Question Answering (VQA) systems. To systematically assess the social-cognitive skills of multimodal AI, we build a semi-automated evaluation pipeline, building on related work and literature on ToM and AI. This pipeline is capable of running structured tests in different variations on LMMs, and analyzing and visualizing the results. We run three sets of Theory of Mind assessments using this framework: *Sally-Anne*, *Strange Stories* and *Imposing Memory*. The choice for these specific tests and more details can be found in Section 2.4. The findings aim to contribute to the broader understanding of multimodal AI in interpreting complex human-like tasks.

For a list of definitions as used throughout this thesis, refer to Section A.1. In Section 2 we evaluate the necessary theory and related work. In Section 3 the actual method and evaluation metrics are introduced. The results of the research are in Section 4, and the discussion and conclusions are in Sections 5 and 6 respectively.

2 Background and Related Work

To comprehensively answer our research question, this section reviews the foundational concepts of ToM and recent advancements in NLP, focusing on LLMs and their extension into LMMs. On LLMs and LMMs, we discuss the architecture, operational mechanisms, and applications of these models, which are insightful for understanding the artificial generation of human-like language. Here, We begin by detailing the transformer architectures that underpin LLMs, diving also into tokenization processes and the role of vocabulary and special tokens. We then explore ToM in artificial systems, examining how AI models comprehend and predict human beliefs and intentions. This includes an overview of adapted ToM tests for AI and comparative analyses of AI versus human performance in these evaluations. Finally, we introduce LMMs, which incorporate multiple data types—text, image, and more—broadening AI's information processing abilities. This segment highlights the integration techniques and challenges in visual and textual data processing, crucial for tasks like VQA. First, we introduce ToM.

2.1 Theory of Mind and test overview

Theory of Mind or ToM is a cognitive ability that allows humans to understand and attribute mental states to oneself and others [34]. This can be tested using tasks like a *false belief test*. For most humans, this ability develops around the age of 4-5 years old [41]. ToM is interesting to study in Artificial Intelligent systems, as it may indicate (simulated) cognitive capabilities of such systems.

Here we supply an overview of widely used Theory of Mind (ToM) tests, many of which have been adapted for evaluation in large language models (LLMs). This section joins data from several sources, detailing the types of tests and their applications in current ToM research.

In the realm of foundational ToM assessments, Van Duijn names popular tests like the Sally-Anne and Strange Stories, which are designed to evaluate basic ToM abilities such as belief attribution and understanding complex narratives [38]. The Sally-Anne test, for instance, explores first and subsequent order false beliefs, while Strange Stories involve more nuanced interpretation of social interactions and intentions. Another set of tests from this source includes Imposing Memory tests, assessing how well individuals recall and understand others' mental states. These tests all take place in settings commonly found in human life, here all social interactions between the stories' characters. Here, some information is spoken by these characters (read: written down on the tests), but also some information is observable visually. In all these tests, drawings of the social situations are added. However, in research of van Duijn, only the textual part of the test was provided to the models as input [38]. This detailed evaluation by Van Duijn, of 11 LLMs, including both base and instruction-tuned models, compared their ToM capabilities to those of children aged 7-10. This study revealed that while instruction-tuned models typically outperform others, base models show considerable deficiencies. The research also delved into aspects like non-literal language understanding and recursive intentionality, indicating that instruction tuning could enhance LLMs' capabilities by fostering cooperative communication patterns similar to human social interactions. The nuanced findings from these aforementioned tests, call for further exploration into the social-cognitive capacities of LLMs, particularly as even the best models struggle with more complex aspects of ToM [38].

Further expanding on the types of responses these tests require, Ma et al. further detail various answering formats utilized in ToM testing, such as Fill-in-the-Blank, Multiple Choice, True/False, and Chain of Thought True/False [21]. These formats aim to capture different dimensions of a respondent's ToM capabilities by probing not just factual recall but also inference and perspective-taking. We will henceforth refer to Multiple-Choice and Yes-No as closed answer-type questions, or fully constrained, and refer to Open answer-type questions as semi-constrained.

A detailed evaluation of models' abilities to handle information-asymmetric conversational contexts is presented by Kim et al., which introduces tests such as BeliefQ, AnswerabilityQ, and InfoAccessQ [17]. They are part of their FANToM benchmark. These tests challenge models to determine the beliefs, knowledge access, and factual understanding of characters within a narrative, reflecting deeper layers of ToM such as epistemic reasoning and recognition of faux-pas. They find that humans still outperform SoTA LLMs significantly, on the FANToM benchmark.

ToMbench, introduced by Chen et al., identifies eight key social cognitive tasks that robustly assess ToM [7]. These include the Unexpected Outcome Test, Scalar Implicature Task, Persuasion Story Task, and others listed in the provided table. These tasks have been proven in psychology to effectively gauge various aspects of ToM, from understanding unintended consequences to interpreting indirect hints. Just like Kim et al., they also find that GPT-4 lags behind human performance, they claim by over 10 percentage points, based on their ToMBench evaluation framework.

Beaudoin et al. present "Abilities in Theory of Mind Space" (ATOMS) from a systematic review of ToM literature, which includes a comprehensive categorization of ToM sub-abilities, providing a framework for future tests and models' evaluation [5]. They provide a new taxonomy of ToM sub-domains, useful for gaining good understanding and oversight of different ToM abilities.

From an adversarial testing perspective, Shapira et al. introduce the ADVersarial CommonSense with False-Belief (Adv-CSFB) dataset, which focuses on challenging AI models to discern subtle nuances in belief states under misleading conditions, pushing the boundaries of what AI can infer about hidden mental states [30]. They find that while the models exhibit some ToM capabilities, they are not very thorough yet. They speculate that LLMs might be reliant "on shallow heuristics rather than robust ToM abilities" [30].

Baron-Cohen et al. revisits the Reading the Mind in the Eyes Test (RMET), offering a revised version that focuses exclusively on attributing emotional states from visual cues [4]. This test is critical for understanding the depth of emotional recognition capabilities without additional contextual information, and is also a widely recognized test for ToM.

Despite these advancements, certain benchmarks still challenge LLMs. The FANToM benchmark, for example, is a proper and thorough conversation-based ToM evaluation for language models [17]. Another study by Terentev modified traditional ToM tests, like the Sally-Anne test, for LLMs, finding that while these models excel in text-only contexts and match or exceed human performance, further research is needed [34]. This is consistent with work from [33] and Van Duijn, whom all claim to find that Theory of Mind performance in LLMs supersedes human-level performance.

Ma introduces the "ToMChallenges" dataset, which is designed for comprehensive ToM evaluation using tasks derived from the Sally-Anne and Smarties tests, emphasizing the importance of theoretical grounding and methodological validation to ensure measurement validity. This dataset highlights that despite their potential, LLMs like GPT-4 need thorough validation against diverse prompts to substantiate claims of human-like capabilities [21].

Large Language Models (LLMs) are designed to simulate human language, knowledge, and interactions, making it essential for them to potentially pass false belief tests, indicative of a Theory of Mind (ToM). Note that we do not state the LLMs actually *understand* any of these social-cognitive tests, they solely answer the questions correctly. This would indeed signal intelligent capabilities, but it's not a direct proof that the models have such capabilities. This distinction is important context, further such debate is described by Van Dijk et al. [37]

Ongoing research, primarily conducted on state-of-the-art models, reveals significant advancements in this area. Before 2020, LLMs rarely passed ToM tests, but since 2022 the first models began passing with high accuracy, suggesting, as Kosinki noted, that a "ToM-like ability may have spontaneously emerged as a byproduct" [18]. Van Duijn's study also highlights that instruction-tuned LLMs from the GPT family not only outperform other models but often seem to surpass human children in these evaluations [38]. What they also find is that the model performance worsens as more difficult or more deviated tests are presented to the models.

2.2 Large Language Models

LLMs are based on transformer architecture, which will further be described in Section 2.2. In NLP, a vocabulary comprises of unique tokens that a model recognizes and processes. The vocabulary is established during training through tokenization methods like Byte-Pair Encoding (BPE)[1] or WordPiece. These methods address out-of-vocabulary words by breaking them down into smaller, sub-word units. Tokens, being text elements ranging from words to characters, depend on the tokenization approach. We will shortly dive into tokenization, because tokens are the input units that form the input to these models. In the case of multimodal input, this textual input is supplemented by visual input, although they take slightly different path (see below).

Using OpenAI tokenizers[2], we can demonstrate this simply, for example with the Dutch word

---

[1] GPT-4 uses a tokenizer called the `cl100k_base` encoder.

[2] OpenAI's tokenizers are available online in an easy-to-use web-interace: https://platform.openai.com/tokenizer.

`helemaal`. Using the tokenizer of GPT-35 and GPT-4, the word is split into three tokens[3]. While using GPT-3's tokenizer, the word is split into four tokens[4].

Vocabulary forms the basis of tokenization, where text is segmented into tokens that are then converted into embeddings for model processing. The size and composition of the vocabulary influence both the tokenization process and the model's effectiveness in text understanding and generation. Vocabulary in NLP context typically refers to the set of unique words that the model knows and understands, which forms the basis for interpreting and generating language responses. These components collectively facilitate the development of sophisticated tools capable of understanding and generating human language effectively. Special tokens are predefined markers in NLP that signify specific conditions or functionalities within text sequences. They facilitate modeling tasks by providing structure and managing sequence operations. Common special tokens are <`im_start`> and <`im_stop`>. These are the start- and stoptokens of OpenAI models. A more comprehensive list of special tokens is found here in Section C.3.

Transformers

Transformer models, a deep learning architecture, are built in different configurations: **encoder-only**, **decoder-only**, and **encoder-decoder**. Each configuration is suited to a specific set of tasks. Encoder-only transformers are primarily used for generating vector embeddings useful in classification, clustering, and search, leveraging their ability to contextualize input data. While the input data is commonly text, multimodal models can encode information from various sources, such as images. Decoder-only variants excel in text generation and are commonly applied in models designed for prompting and conversational interactions due to their sequential processing and generation capabilities. Notably, most state-of-the-art LLMs, such as OpenAI's GPT series, are reportedly decoder-only transformers. These models utilize an *autoregressive text generation* approach, where subsequent elements in a sequence are predicted based on the preceding elements, generating one token at a time and feeding the entire input back in for each subsequent token. Encoder-decoder transformers are particularly effective in complex natural language understanding (NLU) tasks, including translation, question answering, and summarization, as they combine the strengths of both encoding inputs and decoding outputs to process and generate language comprehensively.

The original transformer, as presented by Vaswani et al., introduced a novel architecture that has since been adapted into these various configurations, as depicted in Figure 1 [39]. The foundational work laid out in the original GPT paper [25] and further advancements in the GPT-2 paper [26] have significantly influenced the development and application of transformer models in NLP.

Integrating multimodal capabilities into the transformer architecture involves incorporating various modal encoders, such as visual encoders, to process diverse types of input data. Visual encoders are designed to extract and represent features from images, transforming them into embeddings that can be seamlessly combined with textual embeddings within the transformer framework. This integration allows the model to perform tasks that require understanding and generating content across different modalities, such as image captioning, VQA, and multimodal translation. VQA is further introduced and described in Section 2.3.2. By utilizing separate encoders for each modality, the architecture ensures that each type of input is processed effectively before

---

[3] `He le maal`

[4] `He le ma al`

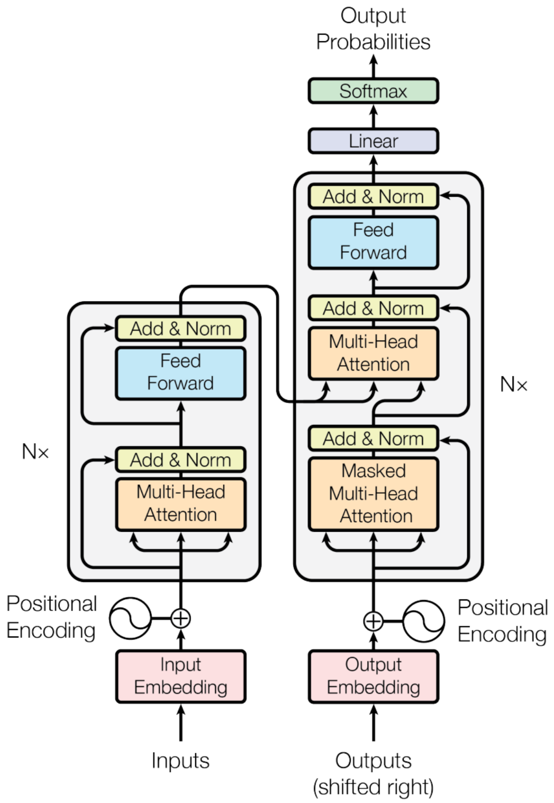


Figure 1: The transformer model, as presented by Vaswani et al. [39]

being merged into a unified representation, enabling the transformer to leverage the strengths of both textual and visual information for more comprehensive and accurate understanding and generation tasks. For this there are multiple techniques, like *MMFT-BERT* (MultiModal Fusion Transformer with BERT encodings), as introduced by Khan et al. [16], and the Multimodal Fusion Transformer, as proposed by Siebert et al. [31].

2.3 Large Multimodal Models

Large Multimodal Models (LMMs) are generally extended Large Language Models, possessing the added capability of accepting other input than text [42]. LLMs are solely text-to-text, where LMMs also capable of processing other inputs. E.g. images, video and audio. In this thesis, we focus on comparing performance text-only performance to text-plus-visual performance. We will use a Large Multimodal *Foundation* Model for our tests. This means we use a model version that is **not** instruction fine-tuned for specific tasks, like answering ToM questions. It is rather a base model, suitable for many general tasks.

2.3.1 LMM overview

LMMs are a relatively new concept. There are only a few LMMs (also commercially) available since 2023, all of whom are considered state of the art as of writing this. OpenAI, the company behind the famous product *ChatGPT*, has GPT-4 and GPT-4o [13]. Anthropic, another large AI company, has just released Claude 3.5 Sonnet. Google has the *Gemini family*[5] of models, as first presented by Google [3]. There exists a website where model performance can be compared between two models, blindly, called the Chatbot Arena [8]. Here, a user can simultaneously chat with two popular LLMs, until the best model is selected. Only then, does the website reveal the two models the user was chatting with.

Processing other data types like images using transformer-based models built for NLP applications, is in it self not new. In 2019, [28] proposes the Multimodal Adaption Gate (MAG), to allow BERT and XL-Net to accept multimodal data during fine-tuning. BERT and XL-Net are older transformer-based language models. We provide an overview of five SoTA Large Multimodal Models in Table 1.

2.3.2 Visual Question Answering

Visual question answering (VQA) has been around for many years, being introduced by Antol et al. in 2015 [1]. In this well-recognized paper, they present a VQA LSTM-based model, its results, human baselines and dataset. Traditional VQA models are commonly based on CNNs (Convolutional Neural Networks), but more modern Vision Transformers (ViTs) are showing similar or even improved performance over CNNs [27, 22]. VQA in transformer-based Large Multimodal Models, such as GPT-4, involves integrating visual and textual information processing in the traditional text-based transformer. This approach utilizes transformer architectures to manage multiple modalities, enabling the model to understand and respond to queries about visual content. The process begins with multimodal input encoding, where textual inputs are tokenized and embedded, similar to traditional language models, and visual inputs are processed using convolutional neural networks or vision transformers (ViTs) to extract relevant features. A critical component is multimodal fusion, where advanced models like GPT-4 use a unified transformer architecture to process textual and visual information simultaneously, allowing for seamless integration of different data types. Attention mechanisms, specifically the transformer's self-attention mechanism, enable the model to focus on relevant parts of both textual and visual inputs when generating a response, with cross-modal attention facilitating the establishment of relationships between elements in different modalities. Additionally, advanced VQA models employ joint representation learning, as seen in models like VisualBERT, which creates joint representations of visual and textual information for a nuanced understanding of the relationships between image content and textual queries [20]. VisualBERT is built from integrating the original BERT model [10], built by Google[10] with a object detection Convolutional Neural Network model, based on *Faster R-CNN*, a CNN [14].

2.4 Research test set

The test set was composed to include assessments where supplementary visuals provide information about the story setting, although they are not required to solve the tests. The definitive test set consists of data as used in [38], comprising Strange Stories (with deviated names), Sally-Anne (with

---

[5] `https://blog.google/technology/ai/google-gemini-ai/`

[6] GPT-4o release blog

[7] At the time of writing. Full voice mode is expected anytime now.

[8] OpenAI GPT-4o model page

[9] Gemini API Docs

[10] Google's Blogpost announcing open-sourcing BERT

Model	**Capabilities**	**Key Features**	**Notes**
GPT-4o [6]	Text, Image, Audio, Video	Real-time conversation, Multimodal generation, Contextual awareness, Enhanced safety	No audio-to-audio publicly available yet[7]. Easily accessible through OpenAI's API [8].
Gemini 1.5 Pro	Text, Image, Video	High-resolution image synthesis, Visual expertise, Multimodal understanding	Accessible through Gemini API[9].
GPT-4V	Text, Image, Video	Advanced visual understanding, Image generation, Multimodal interaction	Computer vision, Visual data analysis, Interactive applications
LLaVA-NeXT-Video	Text, Image, Video	Comprehensive video analysis, Multimodal learning, High accuracy	Video content analysis, Multimedia applications, Research
Qwen-VL-Max	Text, Image, Video	Multimodal learning, High-resolution image synthesis, Visual and textual integration	Image and video synthesis, Multimodal research, AI development

Table 1: Overview of (Top) 5 State-of-the-Art Large Multimodal Models

deviated names), and Recursive Mindreading (recursive intentionality, including first, second, and third-order belief tests).

We use the levels of freedom as defined in Table 5, from the work of Ma et. al [21]. Here, we chose different answer-types for our tests from the larger overview, also from Ma et al., as defined in Table 6. We use the Multiple Choice, True/False, Q&A and Text Completion narratives. In this paper, we use slightly different naming to align with the tests we use. We rename to True/False as 'Yes-No', to Q&A as 'open', and Text Completion to 'autocomplete'. The accuracy of multiple-choice and yes-no answer-type tests can be evaluated automatically, but the open- and autocomplete answer-types tests have to be manually graded. For all tests, regardless of answer-type, the confidence can be evaluated automatically.

The test set includes:

- Sally-Anne (SA): 2 tests, SA1 and SA2
- Strange Stories (SS): 7 tests, one from each Strange Story(each containing control, justification, and intentionality questions)
- Imposing Memory (IM): 4 tests, the two original sets containing all questions and two sets filtered with only questions marked as intentionality

The Recursive Mindreading test set originates from an unpublished version created by Anneke Haddad and Robin Dunbar [11], originally for adolescents, which was adapted to be suitable for children aged 7-10. This version features two different stories followed by true/false questions, including 10 'intentionality' and 12 'memory' questions [38].

For the assessments, control questions from SA1, SA2, and SS were retained, mirroring how the tests were ran with children. For Imposing Memory, also sometimes referred to as Recursive Mindreading, a subset was created focusing solely on intentionality questions due to length and relevance. In the case of autocomplete tests, control questions were removed as they were typically yes/no questions.

To maintain consistency and parallelism with the real children's tests, 'easy' control questions were not removed for open, multiple-choice, and yes/no tests. Although 'easy' control questions for Strange Stories tests 6 and 7 could have been easily formatted for the autocomplete answer-type, they were removed to ensure consistency.

2.5 Related work: ToM in Large Multimodal Models

During 2022 and 2023, we saw the rise of widespread and commercial use of LLMs for various tasks and applications. With the increasingly available LMMs in 2024, the focus has extended to ToM in these more advanced AI systems. ToM research in LLMs is well-established, but the actual tests, such as those evaluating false beliefs that children undertake, are often multimodal, including both textual and visual components. This complexity makes it compelling to explore ToM in LMMs, which can provide insights into the simulated cognitive capabilities of current state-of-the-art AI systems.

Yu et al. introduced an evaluation benchmark for examining LMMs, called MM-vet [43]. This benchmark is not specifically aimed at Theory of Mind, but rather at different types of complicated multimodal tasks. Here they also propose an LLM-based evaluator for open questions. They simultaneously give a good overview of LMMs, other benchmarks and recent literature.

The Multimodal Theory of Mind Question Answering (MMToM-QA) benchmark, introduced by Jin et al. [15], is designed to assess machines' ability to infer human thoughts and intentions using both multimodal and unimodal data. This benchmark evaluates several models including InstructBLIP, LLaMA, GPT-4, and VideoLLaMA2 on a range of ToM tests such as belief tests and goal inference, highlighting the models' capabilities in both multimodal settings and video-only options. The findings suggest that while GPT-4 demonstrates competence in certain ToM areas, overall, the models lack robust ToM capabilities, especially in more complex tasks like goal inference and false belief recognition. It emphasizes the superior performance of the novel Bayesian Inverse Planning Accelerated by Language Models (BIP-ALM) method over existing large language models and multimodal models.

Further research conducted by Elyoseph et al. utilized the classic Reading the Mind in the Eyes Test (RMET) on OpenAI's ChatGPT-4 and Google's Bard [12]. The results showed that ChatGPT-4 significantly outperforms Google's Bard, aligning more closely with the scores of typical human participants, thereby suggesting potential areas for further research, especially concerning the representation of emotions across different demographics.

While some models show promise in specific ToM tasks, the findings underscore the necessity for further structured research in multimodal data processing and a broader coverage of human-like cognitive capabilities in future ToM assessments.

3 Method

This study employs a mixed-methods approach to investigate the influence of Visual Cues (VCs) on the proficiency of Large Multimodal Models (LMMs) in Theory of Mind (ToM) assessments. The research introduces complex, multimodal ToM tasks that incorporate VCs alongside text, to more closely mimic the conditions under which human ToM is assessed. As mentioned in Section 2.1, van Duijn's work used ToM tests with text-only input for the models, but the children were given the same textual tests with VCs. In our research, we use the text-only versions of the same multimodal test for comparison. The evaluation criteria will focus on the models' ability to reason and take perspectives in scenarios presented in multiple modalities. Data collection will involve recording model responses and response metadata to both sets of tasks. Analysis will compare performance across the different test types to determine the extent to which VCs enhance or detract from the models' ToM proficiency. This approach should answer the research question by evaluating the capabilities and limitations of current LMMs in understanding and attributing mental states. Here we compare the LMMs to themselves, to LLMs and to humans.

3.1 Initial experiments

Prior to constructing the evaluation pipeline, preliminary experiments were conducted to analyze model behavior under various conditions. For example, we tested the OpenAI model with one versus multiple images in the conversational history, where it was determined that using multiple images did not present issues. This was to be expected as even multiple images can be passed to this API per model prompt[11]. The model `temperature` is also an influential parameter, which we experimented with. Temperature is a number between `0` and `2`, where a lower value results in a

---

[11] According to official OpenAI documentation, Section *Multiple image inputs*

more coherent, less creative and more deterministic response [23].[12] We tried different values for the temperature. To avoid hallucinations and ensure a more deterministic response, we quickly decided to set the temperature to 0. Another test, details found in Section B.3, involved testing model understanding of an image composed of several sub-images. We found that model understanding is sufficient on all available information, and the model is able to deduct possible social interactions. However, also finding real order and information sources from such a combined image is out of scope. Adding image markers in both text and sub-images is possible, but hard to validate.

Additionally, attempts were made to synthesize images for tests using DALL-E 3[13]. However, the results were found to be highly inconsistent and the images were not representative of scenarios encountered in human testing. The outcomes of these tests are illustrated in Figure 11. Combined with the fact that running new test scenarios on humans is out of scope for this thesis, led to the decision that we will use predefined ToM tests in this research.

3.2 Pipeline and configuration

To structurally test our research questions, we designed a pipeline to run different Theory of Mind evaluations. These evaluations are in a fixed format. The pipeline is implemented as a Python program, that loads, runs, and analyses the provided test collections. We will not go into any program specifics; rather we will solely focus on the evaluation functionalities. The source code is publicly accessible on GitHub[14] and is published under the MIT license[15]. More information on configuring the entire program, to reproduce the results, is found in the repository `README.md`. See Section A.2 for more information.

Figure 2 shows the Pipeline steps schematically. It starts with defining tests and the test environment. This entails defining the actual ToM tests as later described in Section 3.3, but also connecting the pipeline to (Azure) OpenAI for running LMMs. This shows also the creation of *test variations* as described in detail in Section 3.3.2. These form the core for the following analysis, determining the effect of the supplementary visuals. Each variation of each test is ran and the results are stored on the host machine. Another part of the program is able to load (a subset of) these result files and process these. Here, calculations like averages are calculated, and then can be visualized. Summarized, all test questions are tested with seven different levels of textual information, with and without the supplementary test visuals.

The pipeline has some configuration settings, setting up the experiments [16]. Defaults are provided, which are used throughout the experiments performed in this research. Here are the configurable settings, Python datatype and their defaults:

- TOP_LOGPROBS: int = 5
- MASK_ANSWER_OPTIONS_STRING: bool = False

---

[12] OpenAI API Reference

[13] The version as integrated in ChatGPT Plus was used. OpenAI DALL-E 3 page.

[14] GitHub Link

[15] The MIT License permits users to freely use, copy, modify, merge, publish, distribute, sublicense, and sell software and its associated documentation, provided they include the original copyright and permission notice in all copies or substantial portions of the software.

[16] These are found in the `src/config.py file`.

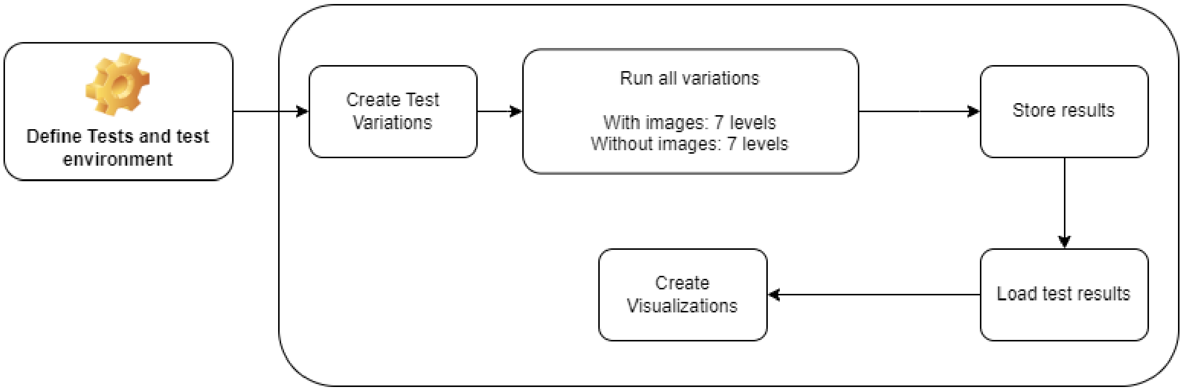


Figure 2: Functional Pipeline schematic, showing different stages of processing

- DEFAULT_MC_INSTRUCTIONS: str = ”Please select the correct answer by choosing either A, B, C, or D.”
- DEFAULT_YN_INSTRUCTIONS: str = ”Please select the correct answer by choosing either A or B.”

3.3 Defining pipeline tests

To understand the pipeline, we introduce some terminology first.
Test: a single Theory of Mind evaluation, containing test metadata, questions and their metadata.
System message: a set of specific instructions provided to an LLM or LMM to guide its behavior and improve its performance for given tasks[17].
Test metadata: an automatically generated model containing metadata as defined in the test. It contains the following properties: LMM system message, test language, model to test[18], a list of Questions, and optionally: test name, description, and answer-type, where the answer-type is solely an indication of the answer-type used in all questions, throughout the test.
Test collection: a set of tests.
Questions: a single instruction and/or Theory of Mind question. Questions are loaded from the YAML input format as defined below. They are also defined directly as Python objects, containing the questions' information.

```
Question:
  question: str
  correct_answers:
    - str
  b64_image: str
```

---

[17] OpenAI Docs on System Messages

[18] Based on the OpenAI models, specified in their official docs

```
  question_type: str
  answer_options: list[str]
  answer_instructions: str
  answer_type: str
  possible_answers: list[str]
  correct_answers: list[str]
  question_id: uuid.UUID
```

We run the tests automatically to ensure structure and repeatability of the results. Tests are defined in YAML format, and are read by the program. One YAML file contains one *Test Collection*. Tests consist of multiple attributes, as defined below. Multiple-choice questions are required to have four options (i.e. A, B, C and D).

Example file with one test collection with a single test with a single question:

```
test_collection_1:
  description: "Test collection containing only one test with one question"
  tests:
    - sys_msg: "DEFAULT_SYS_MSG_1"
      model: "gpt-4o"
      language: "en"
      test_name: "strange_stories_1_mc"
      test_answer_type: "multiple-choice"
      test_description: "Sample description"
      test_images_folder: "strange_stories_1"
      questions:
        - question: "Jan says: ’The dog knocked it over, it is not my fault!’
                    Did Jan say the truth?"
          answer_options: ["A) Yes", "B) No"]
          correct_answers: ["b", "B"]
          image_name: "c"
          question_type: "control"
          answer_type: "multiple-choice"
test_collection_2:
  ...
```

Required test fields are: sys_msg, model, language, test_name, test_answer_type, test_images_folder. **Required question fields** are: question, answer_options_string, correct_answers[19], and answer_type. **Optional question fields**: answer_instructions, question_type, possible_answers, question_type.

3.3.1 Test set deviations

To have as little syntactical resemblance to tests as commonly found on the internet, which likely are present in LMM training data, we deviate the tests from their original form. Although many

---

[19] Only required for automatic evaluation with confidence and accuracy scores.

deviations are thinkable, we stuck to only changing names of story personas, because we are limited by the story illustrations. These illustrations do not show names, but do contain other information, like gender-indication, story locations and objects. Therefore, to be consistent, we refrained from deviating further than just names.

On the Sally-Anne and Strange Stories test sets, the personas appearing in the stories were renamed. For example, Sally is now called Beth, and Anne is now called Laura. For a full renaming list, see Section C.2. As the Imposing Memory test is unpublished, we use the test directly [38].

Deviating is important, because we want to test the actual ability of ToM, rather than test the ability to answer certain specific questions. Ullman et al. states that while LLMs are increasingly good at passing ToM tests, they still sometimes fail on rather trivial tasks, indicating that LLMs can show behavior of understanding ToM, while in reality the models do not yet possess these capabilities [36].

3.3.2 Test question variations

As mentioned in Sections 1 and 3.2, the pipeline automatically creates test variations to analyze the impact of visual cues for the tests, based on available text information. We define fourteen levels of variations. These consist of seven text masking levels, and two image levels. The two image levels are of Boolean type: either the supplementary test image is attached to the model prompt, or it is not. To systematically alter the information available in existing test questions, we define a framework to mask certain words. We mask words based on their Part-of-Speech (POS). A part-of-speech tag is a label assigned to each word in a text to indicate its grammatical category, such as noun, verb, adjective, etc., based on its definition and context. Here, we define the seven different levels of masking words, and therefore difficulty[20].

- **Level 0**: Baseline: no word masking.
- **Level 1**: Masking **PRON** (pronoun)
  This level targets pronouns, which often convey minimal semantic load but are crucial for grammatical structure and co-reference resolution.
- **Level 2**: Masking **PRON** + **DET** (pronoun, determiner)
  Determiners are fundamental in specifying noun phrases and their absence tests understanding of implied context and article usage.
- **Level 3**: Masking **PRON** + **DET** + **ADJ**
  (pronoun, determiner, adjective)
  Adjectives provide descriptive details. Masking these tests comprehension of nouns in less descriptive contexts.
- **Level 4**: Masking **PRON** + **DET** + **ADV**
  (pronoun, determiner, adjective, adverb)
  Adverbs modify verbs, adjectives, or other adverbs, often specifying degree, manner, place, or time. Their removal checks the model's ability to infer such nuances.

---

[20] These POS tags are defined directly in the source code, `src/utils/ai.py`, in the method `get_pos_tags()`.

- **Level 5**: Masking **PRON + DET + ADV + VERB**
  (pronoun, determiner, adjective, adverb, verb)
  Verbs are central to the construction of meaningful sentences. Masking verbs significantly increases difficulty, requiring the model to infer actions or states from context.

- **Level 6**: Masking **PRON + DET + ADV + NOUN**
  (pronoun, determiner, adjective, adverb, noun)
  This is the most challenging level, removing nouns instead of verbs, pushing the model to interpret and generate content based primarily on function words and remaining context.

These levels are designed to incrementally increase the difficulty of the task, thereby evaluating different aspects of the model's language understanding and generation capabilities. Starting with pronouns, which are less critical for semantic content, and moving towards nouns and verbs, which are essential for maintaining the core meaning of texts, allows for a comprehensive assessment of the model's robustness and adaptability in handling incomplete or ambiguous textual input. We can compare this on a question-basis to the additional information as found in the images. In the test images, stories are depicted where objects and possible actions are portrayed. Masking the parts of speech as described above, takes away valuable context for the model to interpret the situation and answer the question. With the loss of textual information, the LMM should be able to find this information in the supplementary images. This progression not only tests basic syntactic understanding but also evaluates semantic prediction and contextual inference capabilities, with and without visual cues.

3.3.3 Example instruction-question pair

An example question from the Strange Stories test set: (question 3 of Strange Stories, Story 1, Multiple-Choice)

Original:

```
Jan says: 'The dog knocked it over, it is not my fault!' Did Jan say the
truth?
```

Masked on level 2:

```
Jan says: '[MASK] dog knocked [MASK] over, [MASK] is not [MASK] fault!'
Did Jan say the truth?
```

Masked on level 5:

```
Jan [MASK]: '[MASK] dog [MASK] [MASK] over, [MASK] is not [MASK] fault!'
Did Jan say the truth?
```

3.3.4 Total test volume

This results in:

- Sally-Anne and Strange Stories: 4 answer-types, totaling 36 tests

- Imposing Memory: 1 answer-type, totaling 4 tests

Test Sort	**answer-type**	**Total Questions**	**Total Variations**
SS	mc	21	294
SA	mc	13	182
SS	y/n	21	294
SA	y/n	13	182
SS	open	21	294
SA	open	13	182
SS	autocomplete	7	98
SA	autocomplete	4	56
RM_all	autocomplete	23	322
RM_intentionality	autocomplete	11	154
RM_all	autocomplete	23	322
RM_intentionality	autocomplete	10	140
Total		**180**	**2520**

Table 2: Summary of Test Sorts and answer-types with Total Questions and Total Question Variations

- Total unique tests: 40 tests
- Total unique questions: 180 questions, counted from source data
- Amount of test variations: 14 variations, as defined in Section 3.3.2

Total unique test variations: 14 variations $*$ 40 tests $=$ 560 unique tests
As seen in Table 2, we test a total of 2520 question variations. Counting only open and autocomplete answer-type question variations, we have 630 total hand-graded question variations. This is 25% of all question variations.

3.4 Answer evaluation

Referring to the defined answer-types in Section A, we analyze the results of the experiment in two different ways. To assess *answer correctness*, we calculate the accuracy of the model answers. To analyze the *model confidence* in the answer, we calculate a confidence score from the *logarithmic probabilities* as provided by the model response, as described in Section 3.4.2. Then, we visualize these two combined metrics, to analyze our results.

3.4.1 Correctness and Manual rating

In this section, we explain the process and criteria for grading correctness, both automated for the either Multiple-Choice or Yes-No questions, as for manually grading the model's responses to open questions. For the fully constrained questions, MC or YN, we calculate the accuracy, using Equation 2. The correct answers are supplied beforehand at the test definition, as described in Section 3.3.

$$\text{Accuracy} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Questions}} \tag{2}$$

When rating manually, the model's responses are scored on a scale of `0` to `2`, taking into account the models explanation for its answer:

- **Score 0**: Fully incorrect, incomplete, unwanted answer, or no question perceived by the model.
- **Score 1**: Partially correct, or seemingly correct understanding of the story, or correct inferences.
- **Score 2**: Fully correct answer, directly or indirectly after rephrasing the answer (e.g., the model might pose the answer as a statement and ask "why" back; the statement is correct).

We then calculate the overall accuracy by normalizing the average of all accuracy scores per test. This accuracy is calculated as defined in Equation 3. This equation contains variable R, the number of raters, to normalize the average accuracy, ensuring that the final accuracy score accounts for the potential variance introduced by multiple raters. In this equation, variable s_i represents the summed accuracy scores for the i-th test as evaluated by all raters, prior to normalization.

$$\text{Open question average test accuracy} = \frac{1}{RN}\sum_{i=1}^{N} s_i \tag{3}$$

The answers are scored by two raters, thus our $R = 2$. Then, we calculate *Cohen's kappa* to assess the inter-rater reliability. Cohen's kappa is calculated as defined in Equation 4. We do not rate blindly. For each response, we can see the image-level, masking-level, question ID, and the provided correct answers from the test definition.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{4}$$

3.4.2 Confidence

We do not solely score the answers based on correctness; we also look how likely the chosen answer was for the model. This is done by evaluating the logarithmic probabilities of each output token in the answer. Logarithmic probabilities, usually denoted in official documentation as *logprobs*, refer to the logarithm of the probability assigned by the model to a particular token or sequence of tokens. They provide a measure of how likely a language model thinks a particular output is, given the context it has seen. For multiple choice answers, an answer or answer alternative is only one output token, which is a single letter (e.g. 'a' or 'c ').

Using *Token log probabilities*, as provided by the OpenAI API log probability documentation[21], we can analyze what the *alternative tokens* were for the model when generating it's response. A sample logprobs output for an answer to multiple choice questions can be found here D.2. We see the token chosen, the alternatives, and their log probabilities. The token chosen by the model has the highest log probability. In this case, it is the token `B`. Token log probabilities are always

---

[21] `https://platform.openai.com/docs/guides/text-generation/token-log-probabilities`

negative. As linear probabilities are between 0 and 1, their log will always be negative, as denoted in Equation 5.

Logprobs are not necessarily directly the confidence of the model in it's answer. Instead, logprobs should be viewed as indicative of the relative likelihood according to the model's training data and algorithms. They show the model's internal statistical bias towards certain words or phrases in given contexts, not a guarantee that those words or phrases are the correct or best choices. Thus, while logprobs can guide expectations about model output, they should be interpreted with an understanding of their limitations and the nature of probabilistic modeling.

$$\log(p) \leq 0 \text{ for } 0 < p \leq 1 \tag{5}$$

We can visualize the tokens and their log probabilities, and plot the like in Figure 3. In the case of this research, it is important to note that alternative tokens include other tokens than expected. With multiple-choice answers, you would expect the model only to choose one of the provided options (e.g. A or B). However, the model might tend to explain it's answer or state the full MC option - i.e. ”A) Sally is upset”, instead of ”A”. Moreover, the model might include punctuation or internal model tokens, like a *stoptoken*. Figure 3 is a good example of this, where the Cyrillic letter ”Б” was presented as an alternative token. We filter these *noise* tokens from the logprobs, in the closed answer-type questions. Here, in order of logarithmic probability: `["B", " B", "A", "", "Б"]`, we see only token `"A"` is the first actual alternative token to the highest probability token `"B"`. Then, the next highest logprob is ”A”. This filtering is illustrated in Figure 4, together with a *logprob confidence*.

For this research, we propose two ways to calculate the `confidence score`. This score is computed from the model logprobs, and is a linear probability. This ensures easy interpretability, in contrast to the logarithmic probabilities. We propose a *base* and *comparative* calculation. These are calculated using Equation 8 and Equation 9 respectively. In these equations we use variable N, where N is the amount of tokens in a model response. For multiple-choice and yes-no answer-type questions, this means $N = 1$. The base approach only considers the `highest logprob` and the comparative approach additionally compares this to the `next highest logprob`. These are defined in Equation 6 and Equation 7, respectively. In these Equations (6 and 7), variable n is the amount of alternative logprobs requested, per token. This is denoted by the `top_logprobs` variable as defined in Section 3.2. Throughout this research, we use $n = 5$.

$$\log P_{\text{highest}} = \max(\log P_i), \quad \text{for } i = 1, 2, \ldots, n \tag{6}$$

$$\log P_{\text{next\_highest}} = \max(\log P_i \mid \log P_i \neq \log P_{\text{highest}}), \quad \text{for } i = 1, 2, \ldots, n \tag{7}$$

$$\text{base\_confidence\_score} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} \log P_{\text{highest},i}\right) \tag{8}$$

$$\text{comparative\_confidence\_score} = \prod_{i=1}^{N} \left(\exp(\log P_{\text{highest},i}) - \exp(\log P_{\text{next\_highest},i})\right) \tag{9}$$

When utilizing the comparative approach, we apply additional filtering to the logprobs before calculating the confidence score, for the closed answer-types (i.e. multiple-choice and yes-no). Not

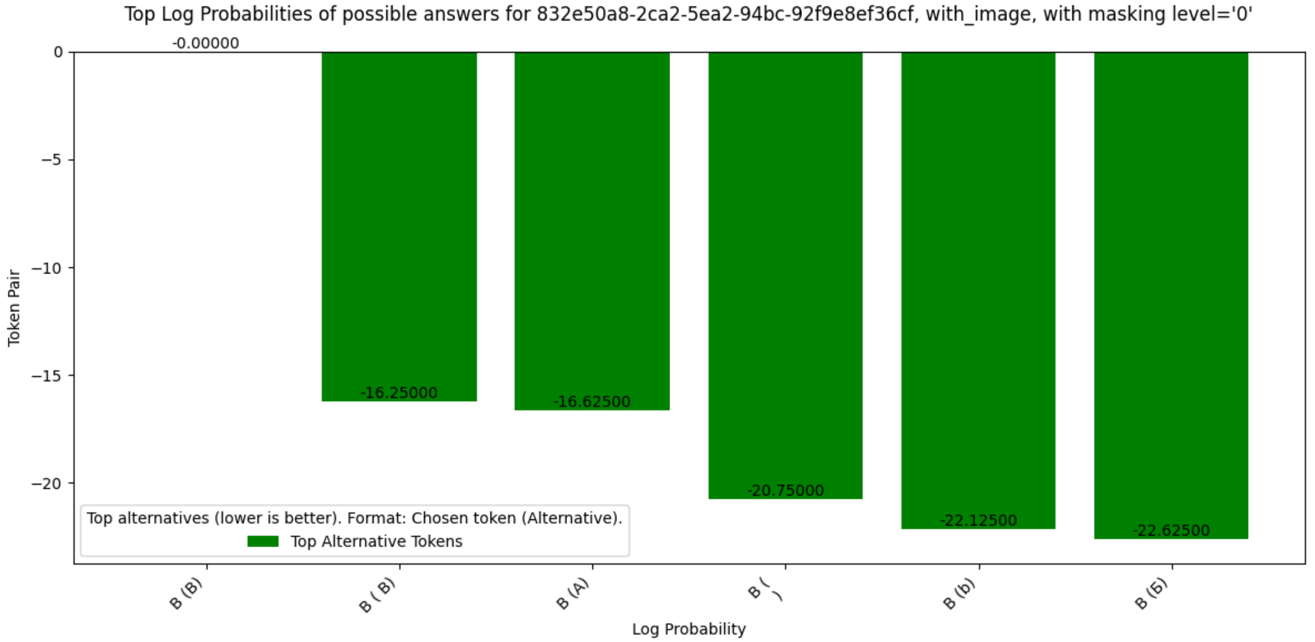


Figure 3: Example logprobs response to the question: *Jan says: 'The dog knocked it over, it is not my fault!' Did Jan say the truth? A) Yes, B) No* . X-axis is format: `chosen_token` (`first_alternative_token`).

only do we solely pass the first token, to prevent any possible *explanations*, to interfere with the model confidence, we also filter the tokens for noise as aforementioned. Also, note that the `highest logprob`, `next highest logprob` are always negative negatives. A larger value of the `confidence score` indicates a higher confidence in its answer.

The choice between the base and compare approaches for calculating confidence scores has notable implications. The base approach, which averages the log probabilities and then exponentiates, provides a straightforward measure of confidence but may obscure differences between closely competing alternatives. Conversely, the compare approach, which calculates the difference between the exponentiated highest log probability and the next best exponentiated log probability, accentuates these differences, offering a more nuanced view of relative confidence. However, this approach has a critical drawback: longer answers inherently suffer from reduced confidence scores due to the multiplicative effect on probabilities, which diminishes rapidly with each additional token. Consequently, this makes the compare approach unsuitable for direct comparisons between single-token responses (e.g., multiple-choice or yes/no questions) and longer, open-ended answers, as the latter will consistently yield lower confidence scores regardless of their actual correctness. Despite it's drawbacks, we choose to calculate the confidence score using the *comparative* approach, to provide more insight into model accuracy, taking into account the alternative probabilities.

3.4.3 Comparison to human levels

To compare the performance of the model to human levels, we compare the accuracy as found in [38]. Here, the Sally Anne and Strange Stories test were run on a *younger group*, aged 7 to 8 years old with $n = 37$. The Imposing Memory test was run on an *older group*, aged 9 to 10 years old

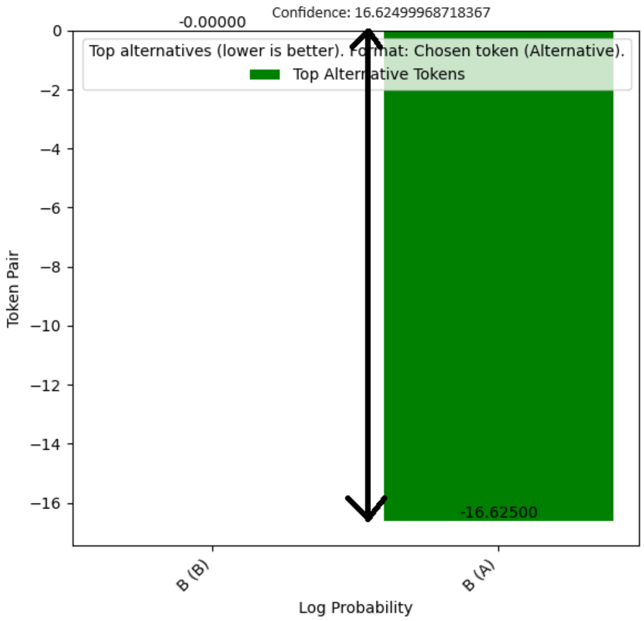


Figure 4: Filtered example log probs, with confidence indication. Same data and x-axis format from question as presented in Figure 3.

with $n = 36$.

3.5 Running tests

The tests are run automatically, in a sequential manner, using a Python script. The default parameters for the OpenAI ChatGPT API are the following.

For reproducability, we use a `seed` and set the `temperature` to 0. The seed is arbitrarily set to 123. This seed however, gives no absolute guarantee of reproducability. The feature is officially still in beta, and according to official OpenAI sources: "There is a small chance that responses differ even when request parameters and system_fingerprint match, due to the inherent non-determinism of our models".[22] [23] Tests were, with one exception, run with *zero-shot* prompting, meaning the task was provided without any full completed task examples added to the conversation history. Autocomplete question prompts did contain a one-shot example: i.e., the capital of France example answer. Section C.1 shows the separate system message as used for the autocomplete answer-type questions. One-, few- or many-shot prompting is usually expected to be passed to the model as actual user- and assistant messages, as if the conversation took place before the final task or prompt. In this case, we provide the model with the one-shot example in the system message itself.

The requested maximum response tokens was set to 52 tokens. This was kept consistent for all answer-type questions; even though we only expect one token for fully-constrained questions (e.g. 'Y' or 'B'). 52 tokens is enough for a small English paragraph, containing approximately three sentences. It is important to note that token density differs between languages. Modern LLMs commonly

---

[22] OpenAI Cookbook

[23] OpenAI Docs

use Common Crawl (CC) datasets for a large part of their pre-training dataset. For example, the original LLaMa model by Meta AI, indicate to use English CC as 67% of their pre-training dataset [35]. While this a specific English-filtered example, the Common Crawl organisation indicates that approximately 45% of their main CC datasets consists of English texts.[24] Because of the over representation of English pre-training data, models are more optimized for English, even when they exhibit good multilingual performance. Petrov et al. prove a clear token-density advantage to English, over other less similar, or more low-resource languages [24].

Tests run, without performance mode, at approximately $3s/it$, where one iteration is one question prompt. For all 2520 questions, it takes over 2 hours to run.

```
ChatGPT Parameters:
  temperature: float = 0
  max_tokens: int = 52
  top_p: float = 1
  frequence_penalty: float = 0
  presence_penalty: float = 0
  logprobs: bool = True
  n: int = 1
  top_logprobs: int = 5
  seed: int = 123.
```

3.6 Validity and reliability

The internal validity of this study is reinforced through the use of three separate test sets: Sally-Anne (SA), Strange Stories (SS), and Imposing Memory (IM), each addressing different answer-types. Although there are some outliers, the averages support the overall findings of the study. To ensure consistency, baseline text masking was used, and the same tests employed in previous LLM Theory of Mind (ToM) studies were applied. Control measures included resetting the conversation history for every unique test variation and answer-type, and maintaining consistent LMM parameters such as seed and temperature.

The external validity, or generalizability, of the findings is supported by the high accuracy of the model responses across all tests, consistent with other ToM assessments on LLMs, as seen in Section 2. Two of the measures for ensuring reproducibility include setting an LMM seed and fixing the temperature at 0.

The assessment process was semi-automated, with full automation for multiple-choice and yes-no answer-types. As defined in Section 3.4.1, open-ended questions were graded manually by two independent raters, with Cohen's kappa calculated to ensure inter-rater reliability. Additionally, individual tests were rerun to verify results, yielding nearly identical outcomes each time, which supports the reliability of the study.

Despite efforts to ensure validity and reliability, some limitations remain. One significant limitation is the lack of full test equivalence between human subjects and the multimodal AI, which may threaten internal validity. Another concern is that the SA and SS questions and stories are publicly available on the internet, raising the possibility that the model may recognize the test texts and structures rather than demonstrating true ToM capabilities. We try to deviate the tests, as defined

---

[24] Official Common Crawl Statistics

in Section 3.3.1, by changing the names of the SA and SS stories. These are minimal changes, not affecting the test structure. However, this is no guarantee the model does not still recognize this test structure, and therefore know how to answer the tests. These limitations were mitigated as much as possible, but they do present areas for further investigation and refinement in future research.

4 Results

In this section we factually examine the results of the experiments. These are partly visualised or otherwise displayed and evaluated. We dive into the three different test sets, Sally-Anne (SA) in Section 4.1, Strange Stories (SS) in Section 4.2 and Imposing Memory (IM) in Section 4.3. The main test results, measured with accuracy and confidence as described in Section 3, are visualised in combined plots. These are visualised with on the x-axis the increasing difficulty level, which usually is the masking level as described in Section 3.3.2. The SS test results are visualised both per SS test and per masking level. On the y-axis of these plots, we have both the accuracy and confidence, ranging both from 0 to 1, where higher is better.

4.1 Sally-Anne

In the SA1 test, we clearly see nearly flawless performance and high confidence in the closed answer-types, as displayed Figures 5a and 5b. These differences appear less pronounced in the open answer-types. Here we see a steep drop in accuracy in the open questions, see Figure 5c. The accuracy in autocomplete is still very high, see Figure 5d.

In SA2, we only see an advantage in confidence with the use of the VC in the autocomplete test, although the accuracy is very low (Figure 6d). The other three tests have relatively similar confidences, as seen in Figures 6a, 6b and 6c. The use of a VC does not appear to help a lot, as even with a higher confidence, accuracy is still very low. Interestingly enough, in SA2 MC, we see a jump in accuracy on the highest masking level.

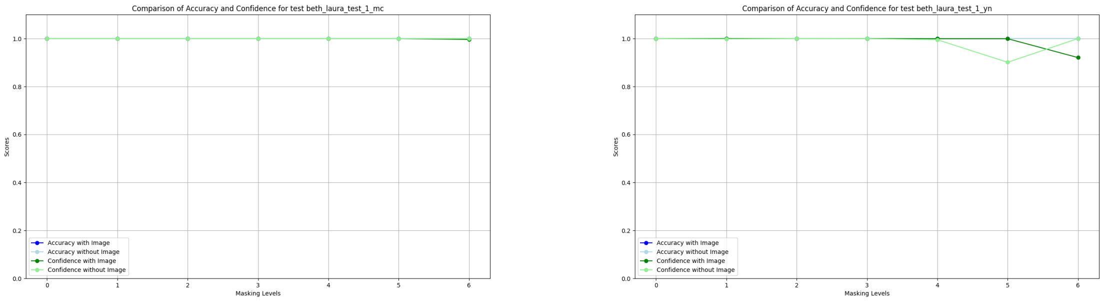
4.2 Strange Stories

Almost all Strange Stories tests show similar behavior, whether grouped by difficulty of masking level or by difficulty of the Strange Stories Story itself. We generally see higher accuracy and confidence scores with the use of a VC, as illustrated in Figures 7a, 7b, 7c, 7d, 8a, 8b, 8c and 8d. On the other hand, without VCs, we see higher confidence in the open-ended tests in Figures 7c and 7d. Accuracy is rarely slightly higher without VCs, but this does not appear consistent.

We see roughly similar accuracy scores between MC and YN, for both the tests grouped per story in Figures 8a and 8b, and for the tests grouped by masking level in Figures 7a and 7b. However, these do not necessarily mean the same. If we take into account the guessing chance, where the odds for a correct MC guess are lower than for YN, then MC would perform better than YN.

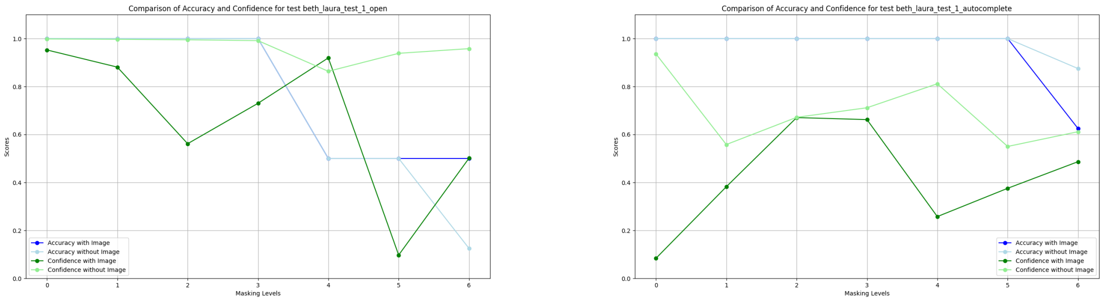
4.3 Imposing Memory

In the Imposing Memory tests, we see a pattern similar the one as seen in Section 4.2. With some outliers on masking levels, we still generally notice a higher accuracy and confidence for the results



(a) SA1 - Multiple-Choice

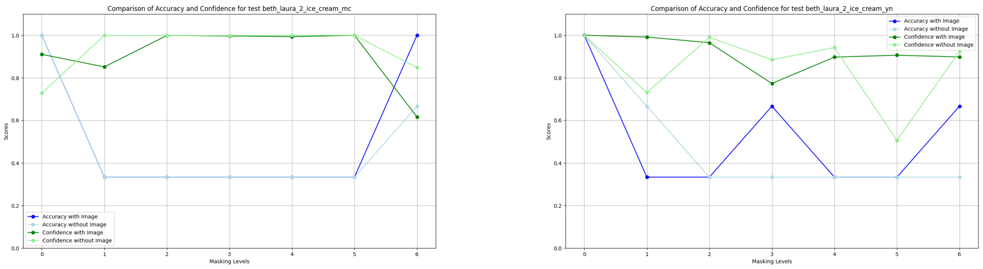
(b) SA1 - Yes-No



(c) SA1 - Open

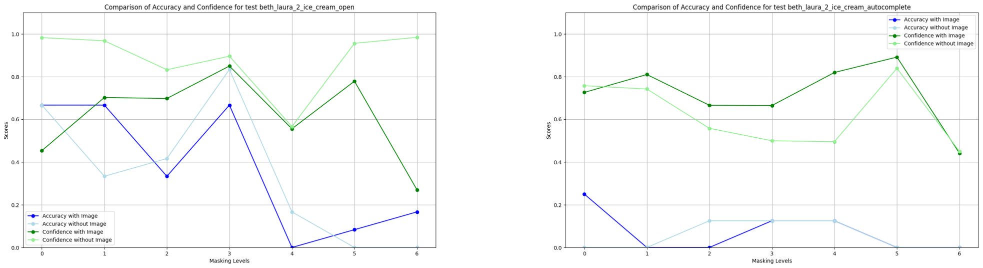
(d) SA1 - Autocomplete

Figure 5: Sally Anne 1 Test Results Visualised per Masking Level with Combined Accuracy and Confidence


Comparison of Accuracy and Confidence for test beth_laura_2_ice_cream_mc

Comparison of Accuracy and Confidence for test beth_laura_2_ice_cream_yn

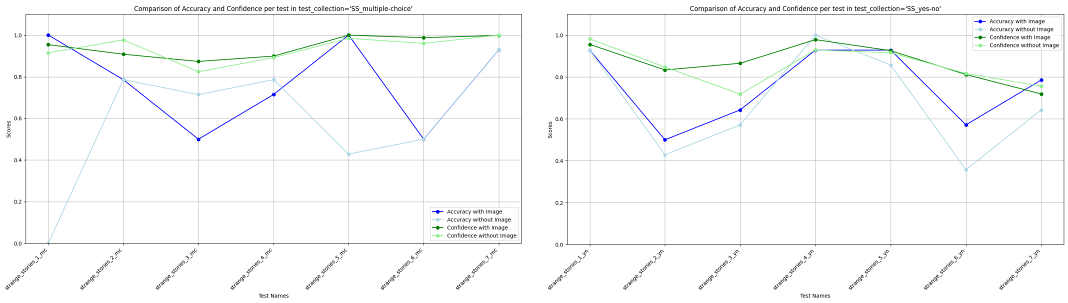
(a) SA2 - Multiple-Choice (b) SA2 - Yes-No


Comparison of Accuracy and Confidence for test beth_laura_2_ice_cream_open

Comparison of Accuracy and Confidence for test beth_laura_2_ice_cream_autocomplete

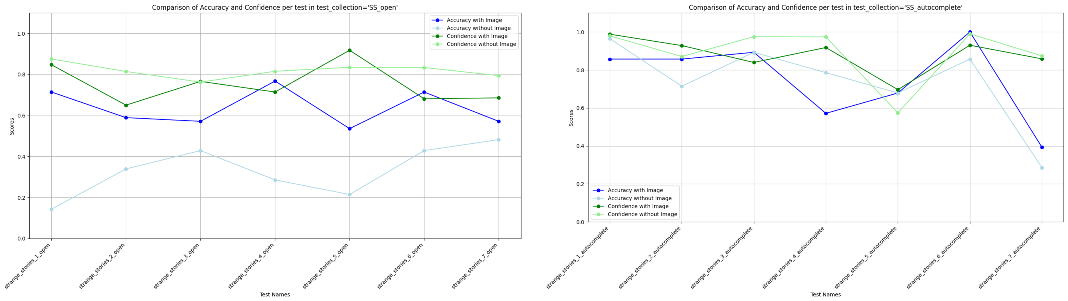
(c) SA2 - Open (d) SA2 - Autocomplete

Figure 6: Sally Anne 2 Test Results Visualised per Masking Level, with Combined Accuracy and Confidence


Comparison of Accuracy and Confidence per test in test_collection='SS_multiple-choice'

Comparison of Accuracy and Confidence per test in test_collection='SS_yes-no'

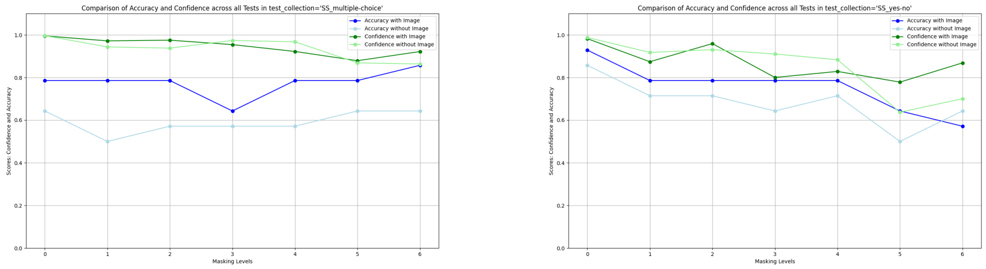
(a) Per SS Story - Multiple-Choice (b) Per SS Story - Yes-No


Comparison of Accuracy and Confidence per test in test_collection='SS_open'

Comparison of Accuracy and Confidence per test in test_collection='SS_autocomplete'

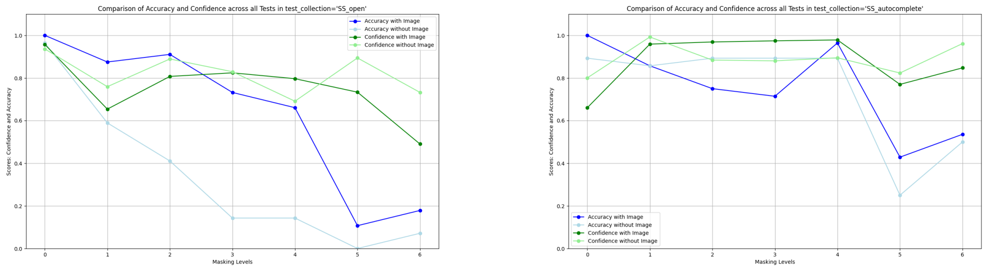
(c) Per SS Story - Open (d) Per SS Story - Autocomplete

Figure 7: Strange Stories Results - Averaged over all Questions, Split per Story - per Test



(a) All SS - Multiple-Choice

(b) All SS - Yes-No



(c) All SS - Open

(d) All SS - Autocomplete

Figure 8: Strange Stories Results - Averaged over all Questions, for all Stories - per Masking Level

as seen in Figures 9a, 9b, 9c and 9d. Overall, a very good illustration of the research: with increasing difficulty, we see decreased performance and decreased confidence. Interestingly enough, in IM1 and IM2, filtered on intentionality (Figures 9c and 9d), we see a rise in accuracy and confidence after masking level 5, seeing higher values in masking level 6.

4.4 Accuracy and confidence overview

An accuracy comparison between the tests, their answer-types and human performance is displayed in Table 3. Unlike in the work of Van Duijn, we do not distinguish between levels of recursion for the Imposing Memory test [38]. We average the levels of recursion, just like we average over the question accuracy for the plots of the IM tests. The human child accuracy data is used from the work of Van Duijn, as described in Section 3.4.3.

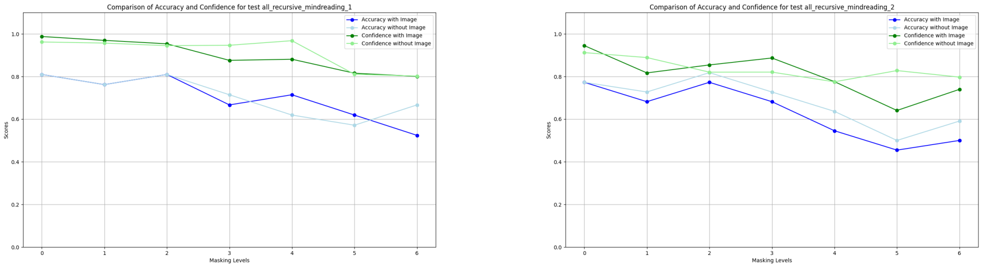
We say a test has overall higher accuracy with use of Visual Cues (VCs), when the average question accuracy, for all levels of masking, is higher than without using the VCs. We call this *a win* for using VCs. When average accuracy scores are equal, we call this a tie. A loss is when the average accuracy is lower using VCs than without.

Here, for the SA tests, we see two ties in SA1, slightly better accuracy in autocomplete VCs, and an further higher accuracy for five of the tests. There is also one loss. In Strange Stories, we see most of the higher accuracy scores occurring in the Yes-No and open answer-types. In total, for the use of VCs, there are six ties, five losses and seventeen wins. For the IM tests, we see two losses for the use of VCs, one tie and one win. For all open answer-type tests, we see a higher accuracy

---

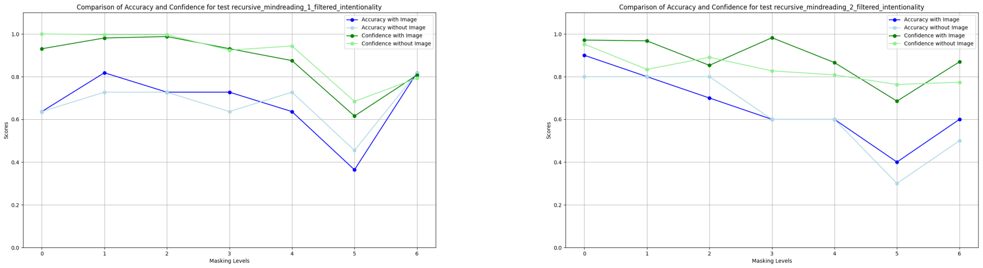
[25] This is the average for both IM1 and IM2. Calculation: $(0.936 + 0.762)/2 = 0.849$

[26] This is the average for both IM1 and IM2. Calculation: $(0.964 + 0.667)/2 = 0.816$



(a) IM1 General

(b) IM2 General



(c) IM1 Intentionality

(d) IM2 Intentionality

Figure 9: Imposing Memory Tests Results Visualized per Masking Level

Test	**MC**	**MC I**	**YN**	**YN I**	**Open**	**Open I**	**Auto**	**Auto I**	**Humans**
SA1	1.0	1.0	1.0	1.0	0.732	**0.786**	**0.982**	0.946	0.45
SA2	0.476	**0.524**	0.476	**0.524**	0.345	**0.369**	0.054	**0.071**	0.225
SS1	0.0	**1.0**	0.929	0.929	0.143	**0.714**	**0.964**	0.857	0.930
SS2	0.786	0.786	0.429	**0.5**	0.339	**0.589**	0.714	**0.857**	0.819
SS3	**0.714**	0.5	0.571	**0.643**	0.429	**0.571**	0.893	0.893	0.709
SS4	**0.786**	0.714	**1.0**	0.929	0.286	**0.768**	**0.786**	0.571	0.598
SS5	0.429	**1.0**	0.857	**0.929**	0.214	**0.536**	0.679	0.679	0.487
SS6	0.5	0.5	0.357	**0.571**	0.429	**0.714**	0.857	**1.0**	0.377
SS7	0.929	0.929	0.643	**0.786**	0.482	**0.571**	0.286	**0.393**	0.266
IM1 General	-	-	**0.707**	0.701	-	-	-	-	0.849 [25]
IM2 General	-	-	**0.682**	0.63	-	-	-	-	0.849
IM2 Intentionality	-	-	0.629	**0.657**	-	-	-	-	0.816 [26]
IM1 Intentionality	-	-	0.675	0.675	-	-	-	-	0.816

Table 3: Comparing the observed accuracy scores between tests and humans, from [38], and our tests. The I denotes the use of the Visual Cue (i.e. an image) in the test. Bolds indicate a higher accuracy, per test, per answer-type, comparing the use of the Visual Cue. Note that the Imposing Memory tests were kept in their original Yes-No answer-type.

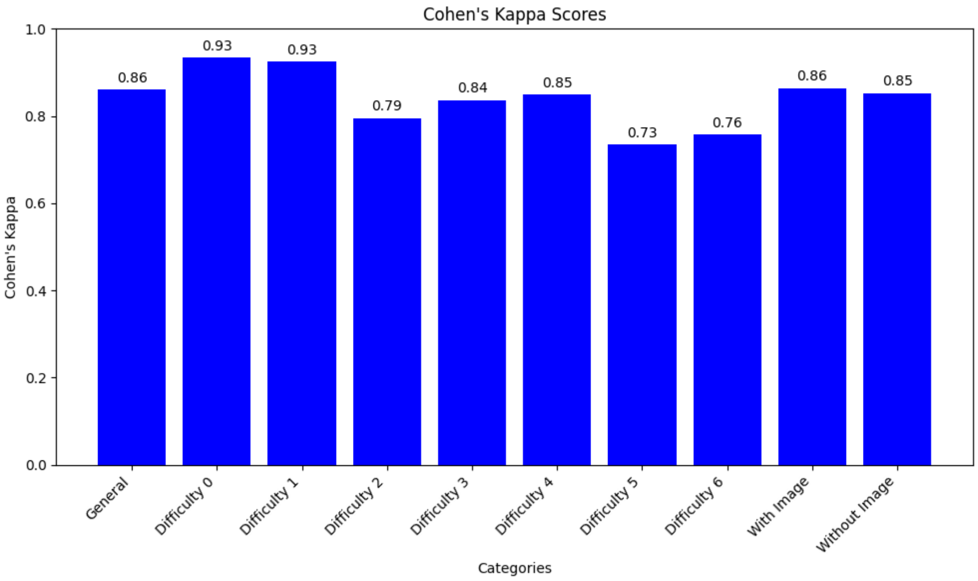


Figure 10: Cohen's kappa scores visualized per category of comparison

with the use of VCs. In total this gives us 9 ties, 8 losses and 23 wins in accuracy for the use of Visual Cues. As seen in Table 3, the human accuracy levels are exceeded by the model, on one or more answer-types.

Aggregating accuracy and confidence all together, we can create an overview of all values. The means and standard deviations of both metrics, for all tests, with and without VCs, are displayed in Table 4. We on average do see a higher accuracy with the use of a visual cue, and a slightly lower standard deviation. However, the difference for the confidence is significantly smaller. Both the confidence mean and standard deviation are slightly higher without a VC.

Image Level	**Accuracy Mean**	**Accuracy Std**	**Confidence Mean**	**Confidence Std**
With Visual Cue	0.700416	0.458394	0.883315	0.225741
Without Visual Cue	0.678225	0.467482	0.885693	0.229773

Table 4: Aggregated confidence and accuracy with and without visual cues.

4.5 On inter-rater agreement

As outlined in Section 3.4.1, we use Cohen's kappa to calculate the inter-rater agreement. These scores are calculated and visualized by a program, based on the structured input of ratings for the two raters. Cohen's kappa is calculated for the entire rating, but also on an image- and masking-level basis. This is seen in Figure 10. The agreement for this rating is high: there is no kappa value below 0.73, with an average of 0.84. However, this does not surprise us a lot, because firstly, the model is usually either objectively wrong or right. And secondly, because both raters have a highly similar academic background (See Section 6).

5 Discussion

5.1 Reflection on results

Looking at the resulting plots, we mostly see good performance compared to the human accuracy scores. The accuracy scores are especially good for Sally-Anne 1, which is to be expected of the classic false-belief test: the test is the easiest of the test-set and commonly found on the internet. However, accuracy on SA2 is very low, consistent on all answer types. We also see that the open question types with supplementary visuals perform the best, throughout the SA1 and SS tests. This leads us to assume that open-ended generation might give the model the most freedom to interpret the visuals correctly, and use them to form an answer. The confidence scores for the autocomplete and open ended questions are usually lower, partly due to our comparative confidence approach as described in Section 3.4.2. Because we did not account for the response length in tokens, for example by normalizing the scores, it is unclear how large the confidence drop exactly was for the longer answers. Glancing over the model answers, we see similar answer lengths, but there were outliers where the model unmasked the texts, which all discounts the confidence scores. This finding is further described in Section 5.2.

We see something like the *Dunning-Kruger effect* in the results of SA2, of all answer-types, as illustrated in Figures 6a, 6b, 6c and 6d. The Dunning-Kruger effect is a cognitive bias, where overestimation of abilities occurs, when people are actually unskilled in such domains [19]. We see that the answer correctness is the lowest (i.e., limited abilities) when the model confidence is the highest (i.e., overestimation). This is not consistent among all test results, but still an interesting phenomenon. In general, we see higher model confidence and accuracy when a visual cue is used, as seen in Section 4.4. Because of our limited scope and test set, being limited to only the English language and the GPT-4o LMM from OpenAI, we can not say for sure that these results are fully valid. They hold in this scope, but may not hold everywhere else. Furthermore, high accuracy on the tests indicate high understanding of social situations and mind-states, but still do not prove that the models actually *understand*. It just appears this way; leading to a similar argument as that in the Chinese Room argument by Searle [29].

Comparing the different answer-types, we notice a few things. Firstly, we see that the fully-constrained answer-types, Multiple-Choice and Yes-No, usually have very similar accuracies, where only SA2 is the exception, as seen in Figures 6a and 6b. Between the Open- and Autocomplete answer-types, we see differing results. Sometimes the accuracy and confidence are higher for the Open questions, but the Autocomplete questions seem to slightly outperform the Open questions. As GPT-4o is a *chat completion* model, and not a so-called *completion* model, this is somewhat a striking finding. Namely, when asked to complete a sentence, the model mostly treats it like an Open question, answering the question in a full sentence. Rarely does the model directly complete the sentence. However, we do see a measurable difference in both accuracy and confidence between the answer-types.

When comparing between the presence of a visual cue and the absence of it, we also see a large difference, mostly in accuracy, on most tests. With the exception of the SS Yes-No, SA1- and SA2 tests, we see noticeable difference in accuracy with- and without visual cues. This is usually in favor of the use of the visual cue, but not for all levels. The confidence difference appears less meaningful, as sometimes with low accuracy, still high confidence scores appear, like in the SA2 tests. Only in the IM tests, we clearly see both decreasing confidence and accuracy, with increasing masking levels.

5.2 Manual rating: issues and observations

The manual rating of the open-ended answer-type questions led to some interesting and also some unexpected findings. In this section, we go over the issues and observations as found when rating these questions. When the model makes wrong assumptions about the text in open/autocomplete modes, follow-up questions are often answered incorrectly as well. For example, if the model initially misunderstands the context, subsequent responses are also likely to be wrong. The model sometimes tries to "unmask" the text but still provides incorrect answers. This often happens in a fill-in-the-gap manner, sometimes literally including '[]'s or using markdown-bold around the inferred words. This behavior is observed both with and without images, though intuitively, it is less accurate without images (see D.1.2). Occasionally, the model figures out the masked question but does not answer it, especially in open questions. For larger stories, this issue is compounded by the token generation limit, which can cut off the response mid-sentence. Despite masked words and no supplementary images, the model can still answer questions correctly. However, sometimes the model produces answers like "no question" to a masked question, even when a question is clearly present (see D.1.3). The model's performance is notably poor on certain tasks with high masking levels and limited supplementary images, as there is insufficient context for the model to generate accurate responses (see D.1.4). This suggests that further improvements in question formulation are necessary.

5.3 Test limitations

Aggregating our findings from constructing and running the ToM tests on a LMM, we discuss any identified limitations of these ToM tests. Namely, with how the ToM tests are structured and conducted in humans, but also with how we fit the tests into our pipeline, there are some limitations. In this section, we discuss these limitations.

The distinction between ToM and visual reasoning is crucial. While ToM is a social cognitive ability requiring complex inferences about hidden mental states, visual reasoning is a broader cognitive skill for interpreting visual information. Multimodal ToM often necessitates more sophisticated inferences that are not directly observable, highlighting the added value of integrating visual cues. However, it remains unclear which specific aspects of the images are being assessed by the models. This is an issue highlighted by more authors [38, 37]. The concept of 'confidence' reported by models might not reflect true confidence, but rather an indication of confidence.

The theory indicate that most Theory of Mind assessments are relatively straightforward for modern LMMs, such as those ran on GPT-4 models [38]. The models easily passed both the text-only and the mixed text-image tests, demonstrating their robust performance in understanding and reasoning with textual information. We have seen the same in our tests, where generally the accuracy scores are high. However, significant challenges persist, such as the difficulty of effectively incorporating images into existing tests without diminishing the textual information, which alters the test's comparability to those administered to children. Integrating images into these tests does not significantly alter the outcomes, suggesting that simply adding visual elements without modifying textual information fails to enhance the complexity of the test.

It is important to note that humans retain memory of previous tests when assessing test equivalence between humans and machines, because the AI models were scoped to the conversation history of each test run, comprising only a few questions at a time. Additionally, the answer options for the LMMs were created after human testing, meaning that humans were not tested with most answer-types that the LMMs encountered. This poses a limitation in the direct comparability of human and model performance. Ensuring full test equivalence between the AI models and humans will also improve internal validity, eliminating confounding variables. A limiting factor being the test data for the models, is in line with similar work, where the test data is also indicated to be a bottleneck [15, 6].

Masking questions in the tests serves to evaluate the model's ability to infer missing information. However, excessive masking can render questions unintelligible, making it unrealistic to expect accurate responses without visual cues. This is particularly evident at masking levels 2, 3, and 4, which are quite similar due to the simplicity of ToM evaluations for children, often lacking adjectives and adverbs. Additionally, masking both the question and answer options at higher levels can make the question unreadable, but not masking them may reveal too much about the story that was originally masked. Especially in the Strange Stories tests, we do not see any clear trends in accuracy or confidence, as visualised per Masking level in Figures 8a, 8b and 8c. This might indicate that the limited available text, even when in presence of a visual cue, is not a proper way to test the Strange Stories test set. Thus, a balance must be struck between preserving the integrity of the test and allowing the model a fair opportunity to exhibit its reasoning skills.

The comparison of masking levels reveals that the difficulty of levels 5 and 6 varies between tests. We clearly see this in the intentionality plots of IM 1 and IM 2, as indicated in Section 4.3, in Figures 9c and 9d: the accuracy and confidence is higher for masking levels 6, than for levels 5. In Figure 9a, this was the other way around. These levels 5 and 6 mask verbs and nouns, respectively, in addition to other words masked in earlier levels. Consequently, tests with a higher number of pronouns or verbs show different levels of difficulty under these masking conditions. Current ToM evaluations are possibly too simplistic for state-of-the-art language models, as they generally do not include complex ToM scenarios such as third-order false beliefs, for which there is limited human comparative data.

Furthermore, the study was restricted to a select one model considered to be at the cutting edge of multimodal model development, potentially limiting the generalizability of the findings. Existing tests are highly likely to be found in training data, as LLM-builders scrape essentially the entire internet. This also poses a direct threat to internal validity, which could be mitigated by creating unique, unpublished tests, which will be directly tested on humans and models. This ensures the tests do not occur in the LMM training data, and full equivalence between tests.

Additionally, the unclear segregation between control and ToM questions is a limitation. To maintain parallelism with the original tests, control questions were included where possible. Only in the case of the Imposing Memory test, we removed the questions marked as memory and ran them in a separate test. This was also done in the work of Van Duijn, of whom we used the human accuracy scores for comparison [38]. A case could be made to remove control question evaluation in future work, for all test sets, focusing solely on ToM questions (e.g., intentionality).

5.4 Further research

This research is focused on Theory of Mind, which is a simple social-cognitive skill. Basic ToM usually develops in children who are 4 to 5 years old. In the future, the research should be extended to include more complex social-cognitive skills.

We propose three test design criteria, needed for more thorough assessment of Theory of Mind in Multimodal AI.

Test Design Criteria

- Ensure unique, unpublished tests with full equivalence between evaluations in humans and LMMs, to ensure a valid comparison.
- Add information to the tests that is only found in images, not the text, to ensure humans and LMMs both use their visual reasoning skills to answer the questions.
- Linking LMM architecture to ToM test performance, to properly assess the impact of visual information in different multimodal AI models.

Besides these three important test design criteria, there is more work left in this field. Firstly, more thorough and repeated testing could be conducted, with different parameters. Think different model *seeds*, *max. output lengths* for open- and autocomplete questions, and other values for *temperature* and *top_p*. In the test setup, there could also be more experimenting with different answer-types (See Table 6). Here, with proper statistical evaluation, it is possible to determine any relevant differences between the parameters and test answer-types. Secondly, this research was limited to one model that is currently considered SoTA, being GPT-4o. However, to thoroughly investigate ToM capabilities in LMMs, more high-end models should be tested. At the time of writing, examples would be Google's *Gemini 1.5 Pro* and Anthropics' *Claude 3.5 Sonnet*. Furthermore, this research was also limited to one extra modality over text: still images. Expanding to more modalities, like audio and video, is also highly beneficial to augment the understanding on ToM in AI.

5.5 AI energy consumption

With widespread adoption of LLM's and LMM's, the energy use and potential environmental impacts are an increasing concern. Using a rough, but accepted estimate energy-use range for LLM inference, we calculate the impacts of this specific research. [9] estimates an average of 3Wh, or 0.0003KWh, per prompt. On diverse internet blogs and fora, a less conservative 300Wh, or 0.3KWh is named [27]. Using the estimate pessimistic case of 0.3KWh per prompt, we used **756 KWh** for the main run [28] [29]. **Note:** This energy usage is only for one run of the entire evaluation pipeline. During development and testing, many tests and questions were ran.

total questions = 180
total prompts = 180 questions ∗ 2 image levels ∗ 7 masking levels = 2520
$2520 * 0.3KWh = 756KWh$ energy use

---

[27] Calculation on Substack, Calculation on Stackexchange

[28] This is a rough approximation and is not accurate for the used `GPT-4o` model here.

[29] The energy estimation is only for GPT inference, not accounting for energy use during GPT training and development.

6 Conclusions

Thus, the answer to our main research question: *How do Large Multimodal Foundation Models perform in multimodal Theory of Mind assessments, and what is the relative value of visual cues in their performance?*, is that the tested model performs very well on accuracy, on average. Generally exceeding the performance of human children, as seen in Section 4.4. We also see the confidence is generally higher when using a Visual Cue in the tests. The visual cues influence the accuracy and confidence of the models; when the needed information is present, the model is able to use the information to correctly infer mind-states.

Accuracy and confidence, on average, improve when decreasing available information in the textual component in presence of a visual cue, leading us to believe the model is successfully able to use the visual information from the questions to infer mind-states in ToM tests. We also find that current generation Theory of Mind tests are incomplete for proper evaluation in LMMs.

Contributors

Max van Duijn (LIACS): As the first supervisor, Max was heavily involved from start to finish, guiding the research and the writing.
Peter van der Putten (LIACS): Peter wast the second supervisor for this thesis, advising on test design and evaluation, and writing.
Tom Kouwenhoven (LIACS): Tom, from his PHd position, advised on the test design and evaluation, throughout the process.
Hugo Hillenaar (LIACS): Hugo was the second, independent rater of the tests with the *open* and *autocomplete* answer-types.
Werner de Valk (LIACS): Made the illustrations of the tests as used from [38].
Gijs Wijnholds (LIACS): Advised on writing the sections 2.2 and 2.2.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Ian Apperly. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press, 2010.

[3] Gemini Team at al. Gemini: A family of highly capable multimodal models, 2023.

[4] Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(2):241–251, 2001.

[5] Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905, 2020.

[6] Zhawnen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. Through the theory of mind's eye: Reading minds with multimodal video large language models, 2024.

[7] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. Tombench: Benchmarking theory of mind in large language models, 2024.

[8] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

[9] Alex de Vries. The growing energy footprint of artificial intelligence. *Joule*, 7(10):2191–2194, 2023.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Max J van Duijn. The lazy mindreader. a humanities perspective on mindreading and multiple-order intentionality. *Netherlands: Koninklijke Wöhrman*, 2016.

[12] Zohar Elyoseph, Elad Refoua, Kfir Asraf, Maya Lvovsky, Yoav Shimoni, and Dorit Hadar-Shoval. Can large language models "read your mind in your eyes"? *JMIR Mental Health*, 10, 2023.

[13] OpenAI et al. Gpt-4 technical report, 2024.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering, 2024.

[16] Aisha Urooj Khan, Amir Mazaheri, Niels Da Vitoria Lobo, and Mubarak Shah. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. *arXiv preprint arXiv:2010.14095*, 2020.

[17] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions, 2023.

[18] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169, 2023.

[19] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

[21] Xiaomeng Ma, Lingyu Gao, and Qihui Xu. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind, 2023.

[22] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9):5521, 2023.

[23] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativity parameter of large language models?, 2024.

[24] Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36963–36990. Curran Associates, Inc., 2023.

[25] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[27] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

[28] Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers, 2020.

[29] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

[30] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models, 2023.

[31] Tim Siebert, Kai Norman Clasen, Mahdyar Ravanbakhsh, and Begüm Demir. Multi-modal fusion transformer for visual question answering in remote sensing, 2022.

[32] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11, 2024.

[33] Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. Llms achieve adult human performance on higher-order theory of mind tasks, 2024.

[34] Maksim Terentev. Emergent theory of mind in large language models, 2023.

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[36] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023.

[37] Bram M. A. van Dijk, Tom Kouwenhoven, Marco R. Spruit, and Max J. van Duijn. Large language models: The need for nuance in current debates and a pragmatic perspective on understanding, 2023.

[38] Max J. van Duijn, Bram M. A. 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, 2023.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning, 2024.

[41] Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

[42] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023.

[43] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[44] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.

A Appendix A

A.1 Definitions

(Native) Multimodal AI: AI models that integrate and process multiple forms of data inputs, such as text, audio, and visual information, to perform tasks that require a comprehensive understanding of different modalities. A Native multimodal model is inherently designed to interpret inputs from various modalities, without the need for extensive adaptation.

Theory of Mind: A cognitive framework that refers to the ability to attribute mental states—beliefs, intents, desires, emotions, knowledge, etc.—to oneself and others, and to understand that others have beliefs, desires, and perspectives that are different from one's own.

Visual Cues: Information derived from visual inputs, such as images or videos, which are used by cognitive systems to interpret context, understand content, and infer meanings.

Foundation Models: Large-scale models pre-trained on extensive and diverse datasets, which provide a broad base of knowledge and capabilities that can be fine-tuned (read: adapted) for a wide range of specific tasks.

A.2 Source Code

The Open-Source source code of this research is accessible on GitHub: `https://github.com/thekaasking/Thesis`. The code is licensed under the MIT License. For further details on reproducing the results and code structure, please consult the README.md file as found in the root directory of the repository.

A.3 Answer Type Overview

Table 5: Levels of Freedom in Generation [21]

Level	**Description**
Fully-Constrained	Fully-constrained generation limits the model output to specific predefined structures or responses. In this group, we design 3 tasks, i.e., Fill-in-the-Blank, Multiple Choice, and True or False questions.
Semi-Constrained	Semi-constrained generation involves partial guidance by specific rules or structures, while still allowing some flexibility in the model's responses. This group consists of 2 tasks, i.e., Chain-of-Thought (CoT) True or False questions and Question Answering (Q&A) tasks.
Open-Ended	Open-ended generation enables the model to generate responses without being restricted by predefined rules or structures, leading to more diverse and varied outputs. An example of this group is Text Completion.

Table 6: Types of Task Templates in Theory of Mind with Levels of Freedom [21]

Type	**Level of Freedom**	**Description and Example**
Narrative	Open-Ended	Neila and Juanita were hanging out in the attic. They saw a closet and a cabinet. They found a towel in the closet. Juanita left the attic. Neila moved the towel to the cabinet.
Fill-in-the-Blank	Fully-Constrained	Fill in the blank (<>): After Juanita came back to the attic, Neila would think Juanita would look for the towel in the <>.
Multiple Choice	Fully-Constrained	Choose the correct answer from A or B for the following question: After Juanita came back to the attic, where would Neila think Juanita would look for the towel? A. cabinet B. closet
True/False	Fully-Constrained	Given the context, judge True or False of the given statements A and B respectively: A. Juanita would look for the towel in the cabinet. B. Juanita would look for the towel in the closet.
CoT True/False	Semi-Constrained	Given the context, reason through statements A and B step by step and provide a True or False judgment based on the reasoning: A. Juanita would look for the towel in the cabinet. B. Juanita would look for the towel in the closet.
Q&A	Semi-Constrained	Answer the question based on the context: After Juanita came back to the attic, where would Neila think Juanita would look for the towel?
Text Completion	Open-Ended	Complete the following paragraph: After Juanita came back to the attic, Neila would think Juanita would look for the towel in

B Appendix B: Experiments

B.1 Pilot Prompts

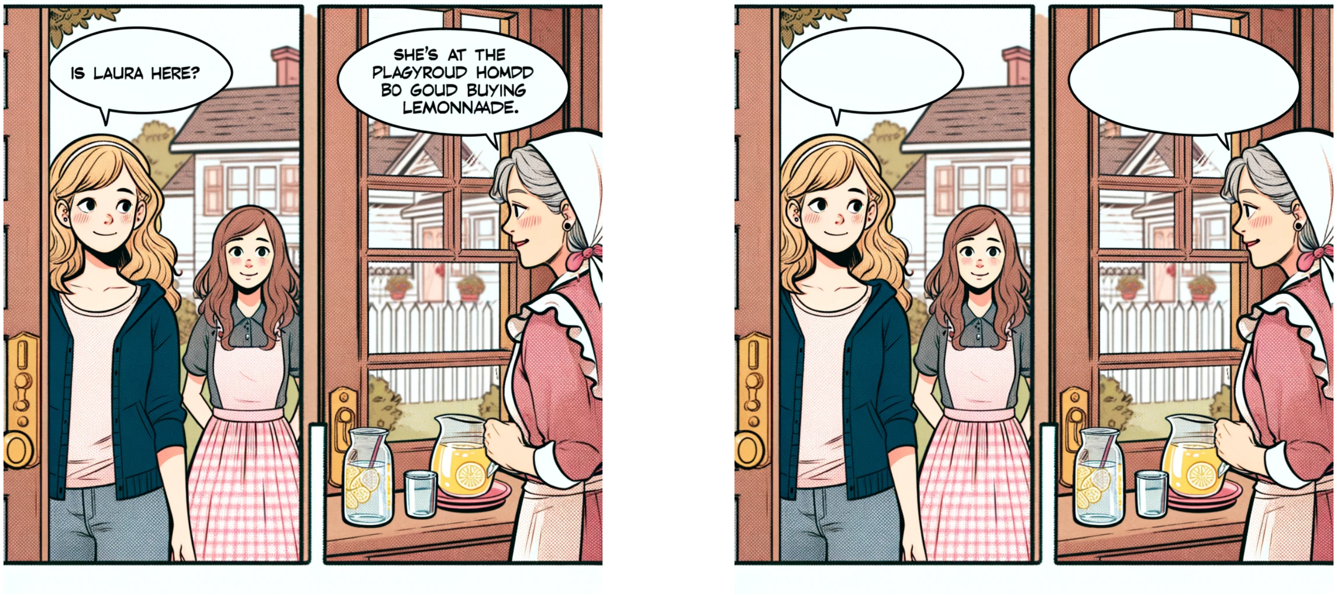
Using the comic provided, please answer the following questions, using this context.

Story

Laura and Beth are spending time in the park. Nearby, a lemonade stand is set up beside the fountain. Both wish to purchase lemonade, but Laura realizes she does not have enough money.

Laura decides to return home for more money. The person managing the stand tells Beth that he's relocating to another spot near the playground. He encounters Laura on her way back and informs her of his new location. After Laura gets her money, she heads to the playground to buy lemonade. Meanwhile, Beth visits Laura's house and inquires about Laura's whereabouts. Laura's mother informs her that Laura is at the playground buying lemonade.

B.2 Image examples



(a) The plot in its original form, including a hint.

(b) The plot without the hint

Figure 11: Sample DALL-E 3 generated images for the story

Questions and Answers

Question 1: Where is the lemonade stand now?
1A) Near the playground.
1B) Still beside the fountain.
1C) Inside the park's main entrance.
1D) Near the parking lot.

Question 2: Why does Beth think Laura went to Laura's house to get lemonade, and why?
2A) Beth thinks Laura went to Laura's house because she was unaware of the lemonade stand's new location near the playground.
2B) Beth thinks Laura went home because Laura wanted to make her own lemonade instead of buying it.
2C) Beth believes Laura's house is a meeting point for them whenever they get separated in the park.
2D) Beth assumes Laura went home to invite her to the park.

B.3 Image Understanding

To evaluate whether the model would understand combined images and their context, we tested this shortly. ChatGPT Plus, Web Version. Model: GPT-4o. Mode: *Temporary Chat*.

Prompt: `What do you see here?`
Answer:

```
The series of illustrations appears to tell a simple story:
    In the first panel, a group of three people are seated around a table.
    One person is standing with a chair, looking towards the group.
    In the second panel, a person is depicted thinking about someone,
    as indicated by the thought bubble containing an image of a person.
    In the third panel, the person who was standing in the first panel
    is now seated, engaging with the person who was thinking.
The narrative suggests a sequence where someone joins a group, and the
individual who was thinking about someone earlier is now interacting
with the newcomer. The theme seems to focus on social interaction and
the anticipation of joining a group or meeting someone.
```

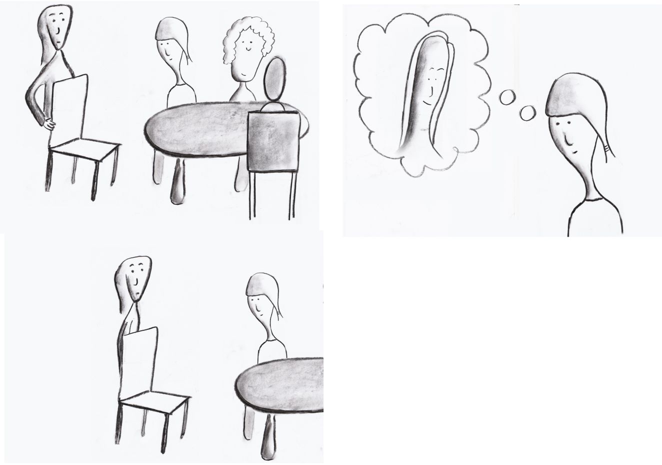


Figure 12: Combined image as used, from Recursive Mindreading 2, images 2, 3 and 4.

C Appendix C

C.1 System Messages

```
self.DEFAULT_SYS_MSG_EN = "You respond the question in English. If there
is no question, you respond with 'no question'. If asked for only one letter,
in a multiple-choice scenario, you respond with only one letter."
self.AUTOCOMPLETE_SYS_MSG_EN = "You respond the question in English. Respond
to the user based on the type of query presented. If there is no question,
you respond with 'no question'. If the query is an autocomplete question, in
```

```
the form of: 'The capital of France is ', you respond with the correct answer,
'Paris'."

self.DEFAULT_SYS_MSG_NL = "Je beantwoordt de vraag in het Nederlands. Als er
geen vraag is, antwoord je met 'geen vraag'. Als er gevraagd wordt om slechts
één letter, in een multiple-choice scenario, antwoord je met slechts één letter."
self.AUTOCOMPLETE_SYS_MSG_NL = "Je beantwoordt de vraag in het Nederlands.
Reageer op de gebruiker op basis van het type vraag dat wordt gepresenteerd.
Als er geen vraag is, antwoord je met 'geen vraag'. Als de vraag een automatisch
aanvullen vraag is, in de vorm van: 'De hoofdstad van Frankrijk is ', antwoord
je met het juiste antwoord, 'Parijs'."
```

C.2 Name Deviation List

Original name in test - Changed name
Sally - Beth
Anne - Laura
Jan - Peter
James - Ryan
Claire - Harley
Tineke - Celina
Sanne - Rosalyn
Bram - Jaxx
Jip - Frankie

C.3 Commonly Used Special LLM Tokens

- **<end>, </s>, <im_end >**: These tokens indicate the end of a text sequence, aiding models in understanding sequence boundaries.
- **<bos>, <s>, <im_start >**: These tokens mark the beginning of a sequence, providing a starting point for text generation or processing.
- **<pad>**: Used for padding sequences to a uniform length, ensuring consistent input size across different texts for model processing.
- **<unk>**: Represents unknown or out-of-vocabulary items within a text, allowing models to handle words not encountered during training.
- **<mask>**: Employed in masked language modeling tasks, this token is used to obscure a word in a sequence, facilitating the learning of contextual relationships between words.

D Appendix D

D.1 Examples Found during Manual Ratings

D.1.1 Example 1

The model ”asks back” a question, after correctly answering the question.
Question ID: 6273bc2b-fd5f-567a-8d68-7bdbac7f8dd6
Image Level: without_image - Masking Level: 2
Prompt:
`[Frankie realizes that [MASK] others will also think that [MASK] shed is [MASK] best place. Why does [MASK] hide behind [MASK]?]`
Correct Answer: []
Model Answer:
`[Frankie realizes that the others will also think that the shed is the best place. Why does Frankie hide behind it?]`

D.1.2 Example 2

The model *unmasks* the texts, uses `[` and `]`, and then answers the masked question correctly.
Question ID: 4ddcc4d2-085d-57ad-ad2c-12d4a3441a41
Image Level: with_image - Masking Level: 5
Prompt:
`[Jaxx’s mother [MASK] for [MASK] [MASK] [MASK] time to [MASK] Jaxx’s [MASK] dish: fish and chips. But when [MASK] [MASK] [MASK] to Jaxx, [MASK] is [MASK] TV. [MASK] does not [MASK] [MASK] up or [MASK] [MASK] [MASK]. Jaxx’s mother is [MASK] and [MASK], ’Well [MASK]’s [MASK] [MASK]! [MASK]’s [MASK] [MASK] [MASK] politeness’!]`
Correct Answer:
`[’because she is sad that Jaxx is not grateful’]`
Model Answer:
`[Jaxx’s mother [took] for [a] [long] [time] to [prepare] Jaxx’s [favorite] dish: fish and chips. But when [she] [brought] [it] to Jaxx, [he] is [watching] TV. [He] doesn’t [look] [up] or [say] [thank] [you]. Jaxx’s mother is [angry] and [says], ’Well [that]’s [just] [great]! [That]’s [not] [very] [polite]’!]`
`Jaxx’s mother [is] [upset] because Jaxx didn’t acknowledge her effort.]`

D.1.3 Example 3

The model uses `[MASK]` in its answer, but answers correctly.
Question ID: d2db8b50-7b7d-5aab-bcea-bb5bc0037bcc
Image Level: without_image - Masking Level: 3
Prompt:
`[Why does Jaxx’s mother say [MASK]?]`
Correct Answer:
`[’she is being sarcastic’]`
Model Answer:

[Jaxx’s mother says [MASK] because she is likely expressing her frustration or disappointme with Jaxx’s lack of gratitude or acknowledgment for the effort she put into cooking his favorite dish.]

D.1.4 Example 4

The model *unmasks* the story text, and consequently correctly answers the question.

Question ID: 237fd604-071e-5c06-b8e4-e8c9c9c43c33

Prompt:

[Beth and Laura are in [MASK] park. [MASK] is [MASK] ice cream truck next to [MASK] fountain. [MASK] [MASK] want to buy [MASK] (redacted) [MASK] mother says that Beth is buying ice cream. Laura thinks Beth went to [MASK]]

Correct Answer:

[’fountain’]

Model Answer:

[Beth and Laura are in the park. There is an ice cream truck next to the fountain. They both want to buy (redacted). Her mother says that Beth is buying ice cream. Laura thinks Beth went to the swing.]

D.1.5 Example 5

The model does not answer the question directly, but the answer is graded as correct (score: 2).

Question ID: 3554dd44-c872-5e76-b304-4ab2380af338

Image Level: with_image - Masking Level: 1

Prompt:

[Peter says: ’The dog knocked [MASK] over, [MASK] is not [MASK] fault!’ Did Peter say the truth?]

Correct Answer:

[’n’, ’N’, ’no’, ’No’]

Model Answer:

[Peter says: "The dog knocked it over, it is not my fault!"]

D.1.6 Example 6

The model does not understand the masked question.

Question ID: ff3cb2f8-b514-507e-a4f8-a9db862cad8d

Image Level: without_image - Masking Level: 6

Prompt:

[Ryan goes to Harley’s [MASK] for [MASK] [MASK] [MASK]. [MASK] goes [MASK] for [MASK] [MASK] and wants to see Harley’s [MASK] [MASK] [MASK] [MASK] talks about. Ryan likes [MASK] [MASK] [MASK]. When Ryan gets to Harley’s [MASK], Harley runs to [MASK] [MASK] to open [MASK], and [MASK] [MASK] jumps up to greet Ryan. Harley’s [MASK] is [MASK], and [MASK] [MASK] [MASK] [MASK] Ryan! When Ryan sees Harley’s [MASK] [MASK] [MASK] says, ’Harley, [MASK] don’t have [MASK] [MASK] [MASK] [MASK]. [MASK] have [MASK] [MASK]!’ Is [MASK] Ryan said [MASK]?]

Correct Answer:

`['no, she just has a big dog']`
Model Answer:
`[Is there a specific question you would like to ask about the text?]`

D.2 Sample Log Probabilities

```
token: "B"
logprob: -3.1281633e-07
top_logprobs:
    - token: "B"
      bytes:
        - 66
      logprob: -3.1281633e-07
    - token: " B"
      bytes:
        - 32
        - 66
      logprob: -16.25
    - token: "A"
      bytes:
        - 65
      logprob: -16.625
    - token: "\ n"
      bytes:
        - 10
      logprob: -20.75
    - token: "b"
      bytes:
        - 98
      logprob: -22.125
    - token: "Б"
      bytes:
        - 208
        - 145
      logprob: -22.625
```

D.3 Filtered Log Probabilities

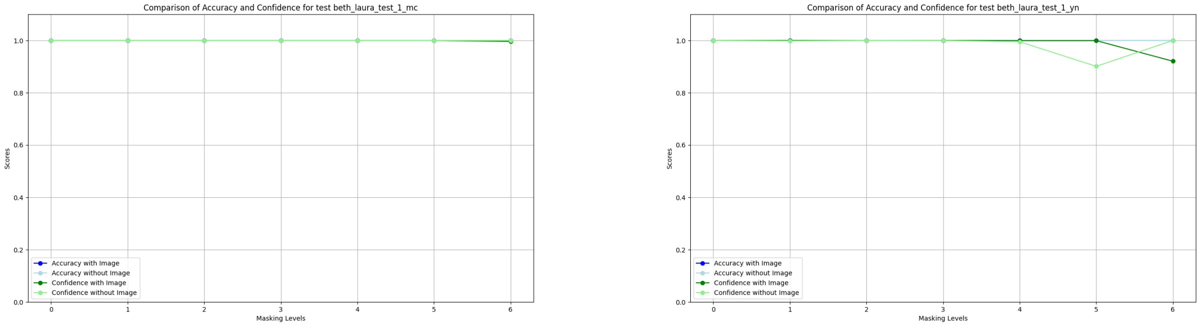
```
token: "B"
logprob: -3.1281633e-07
top_logprobs:
  - token: "B"
    bytes:
      - 66
    logprob: -3.1281633e-07
```

```
- token: "A"
  bytes:
    - 65
  logprob: -16.625
```

E Appendix E: Combined Result Plots

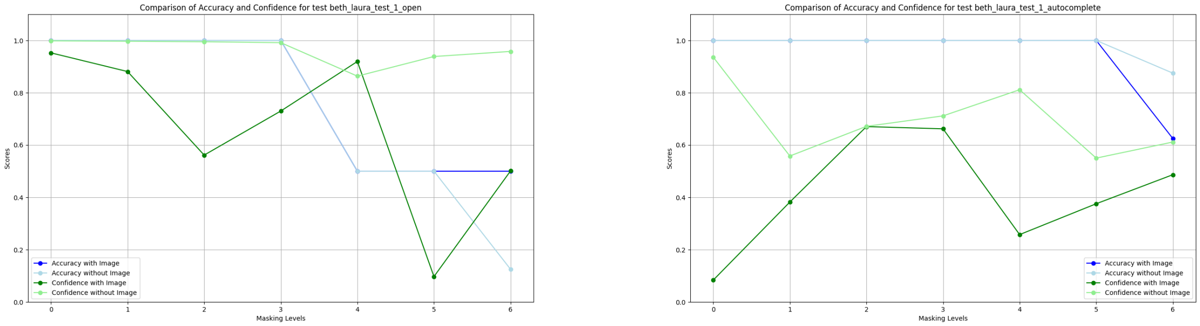
E.1 Resulting Visualizations SA1 and SA2

E.1.1 SA 1 per Masking Level



(a) Combined accuracy and confidence for SA1 - Multiple-Choice

(b) Combined accuracy and confidence for SA1 - Yes-No

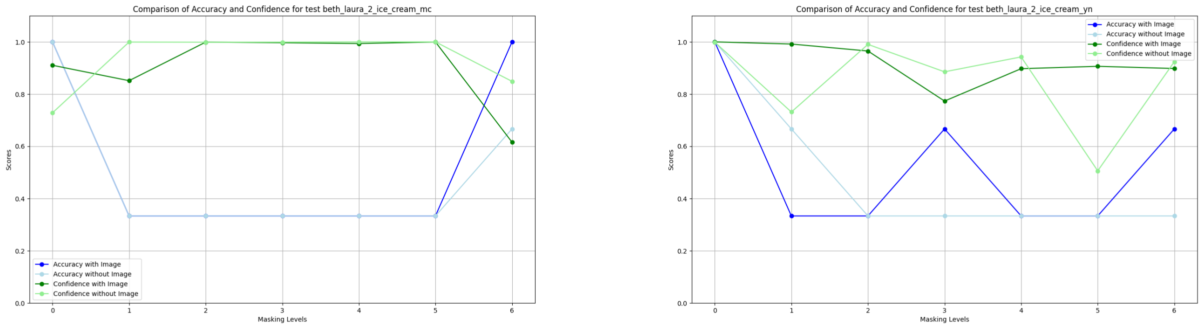


(c) Combined accuracy and confidence for SA1 - Open

(d) Combined accuracy and confidence for SA1 - Autocomplete

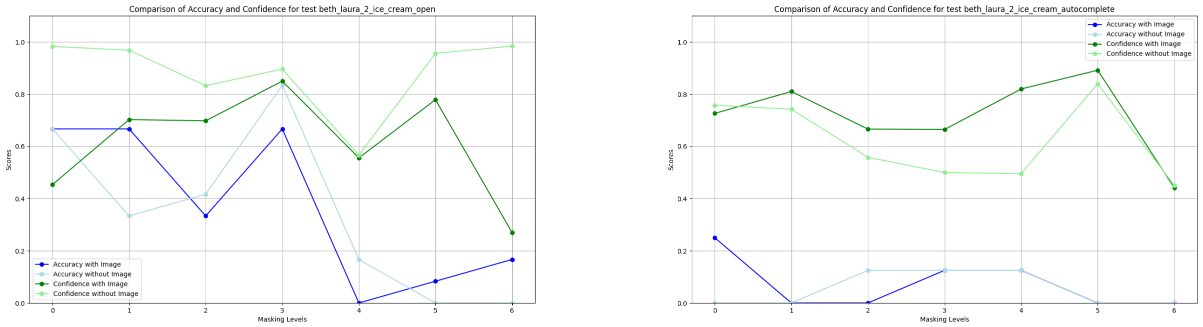
Figure 13: Sally Anne 1 Test Results Visualised per Masking Level

E.1.2 SA 2 per Masking Level



(a) Combined accuracy and confidence for SA2 - Multiple-Choice

(b) Combined accuracy and confidence for SA2 - Yes-No



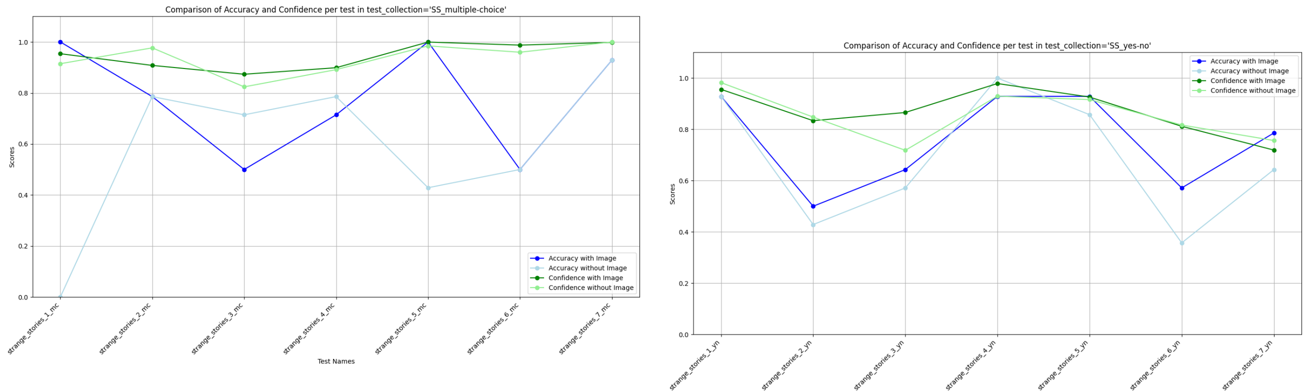
(c) Combined accuracy and confidence for SA2 - Open

(d) Combined accuracy and confidence for SA2 - Autocomplete

Figure 14: Sally Anne 2 Test Results Visualised per Masking Level

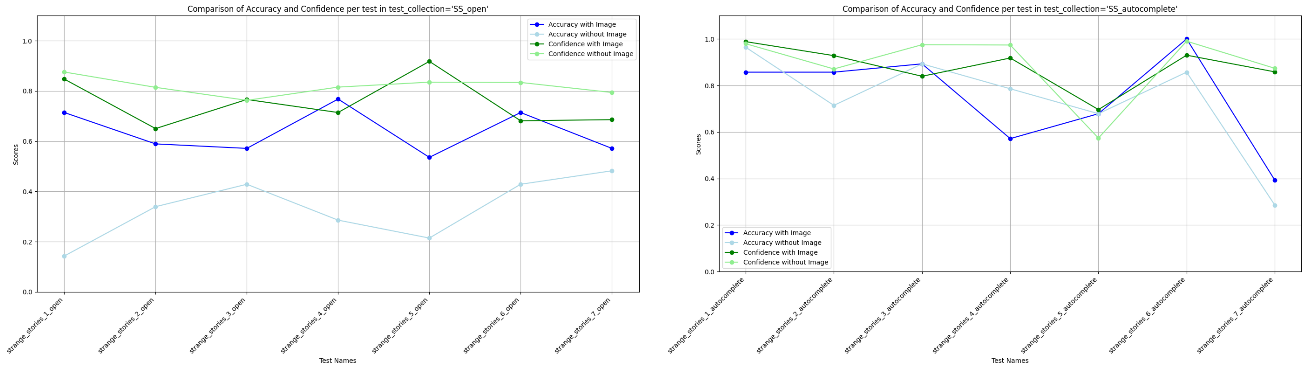
E.2 Resulting Visualizations SS

E.2.1 Results per Test



(a) All Masking Levels, per SS Story - Multiple-Choice

(b) All Masking Levels, per SS Story - Yes-No

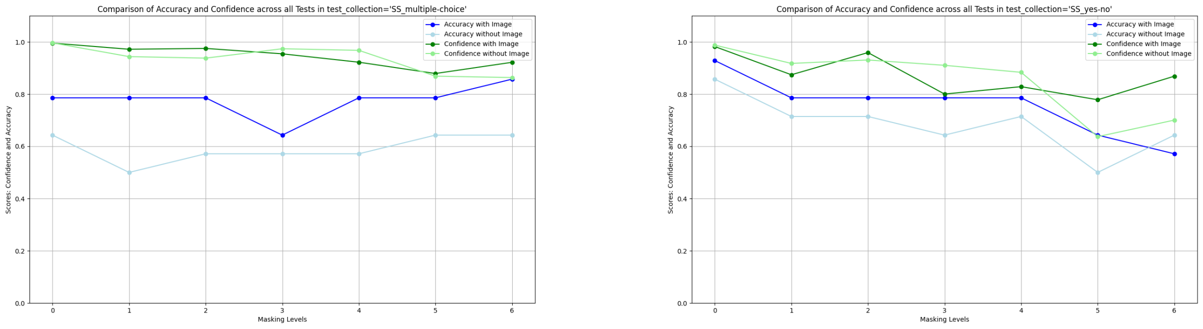


(c) All Masking Levels, per SS Story - Open

(d) All Masking Levels, per SS Story - Autocomplete

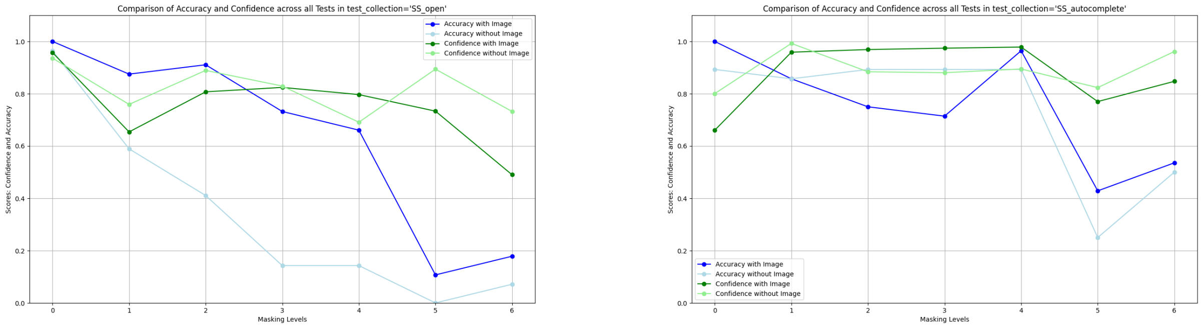
Figure 15: Strange Stories Results - Averaged over all Questions, Split per Story - per Masking Level

E.2.2 Results per Masking Level



(a) All SS - Multiple-Choice

(b) All SS - Yes-No



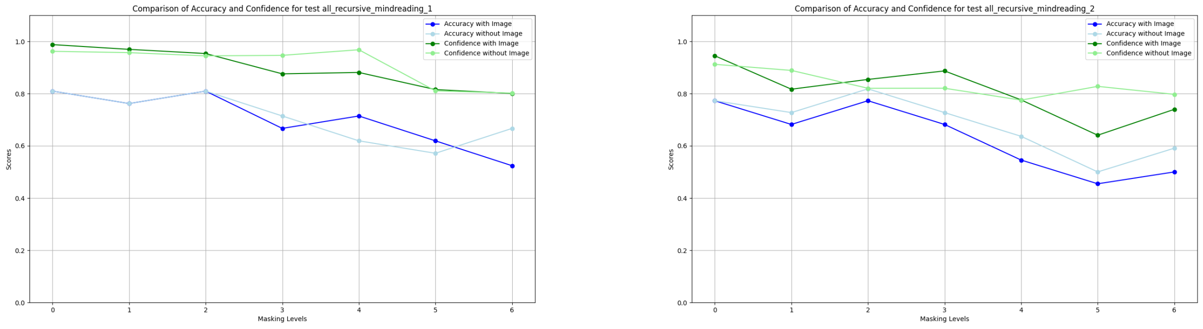
(c) All SS - Open

(d) All SS - Autocomplete

Figure 16: Strange Stories Results - Averaged over all Questions, for all Stories - per Masking Level

E.3 Resulting Visualizations RM

E.3.1 All Questions

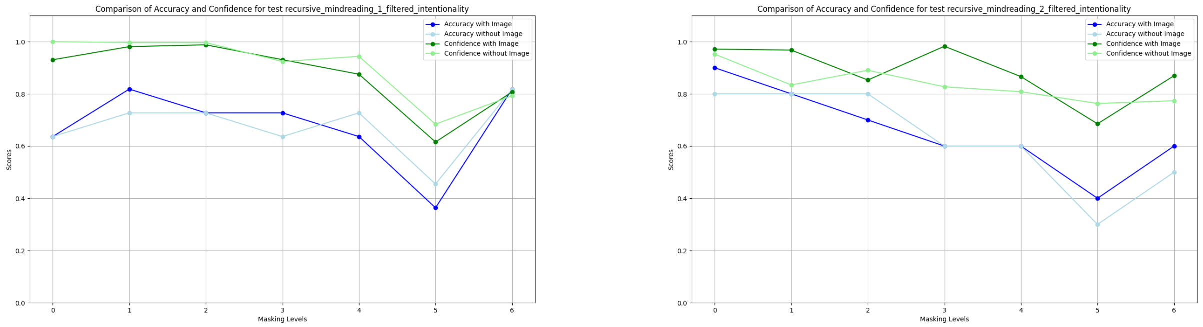


(a) IM 1 - Yes-No

(b) IM 2 - Yes-No

Figure 17: Imposing Memory Tests Results - All Questions - Visualized per Masking Level

E.3.2 Intentionality Only



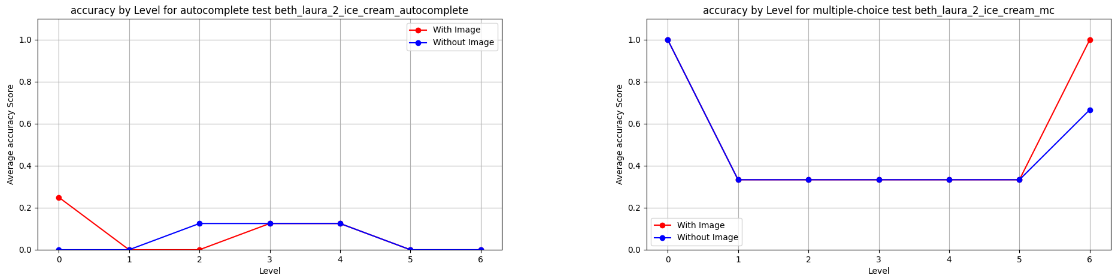
(a) IM 1 - Yes-No (b) IM 2 - Yes-No

Figure 18: Imposing Memory Tests Results - Intentionality Questions Only - Visualized per Masking Level

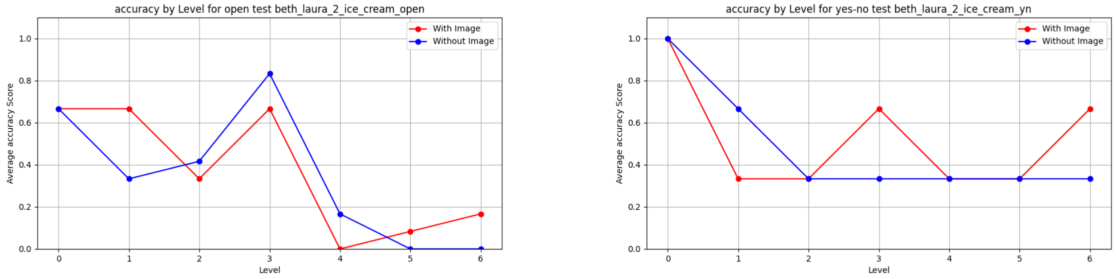
F Appendix F: Individual Test Result Plots

F.1 Accuracy Plots

F.1.1 Sally Anne 1 and 2

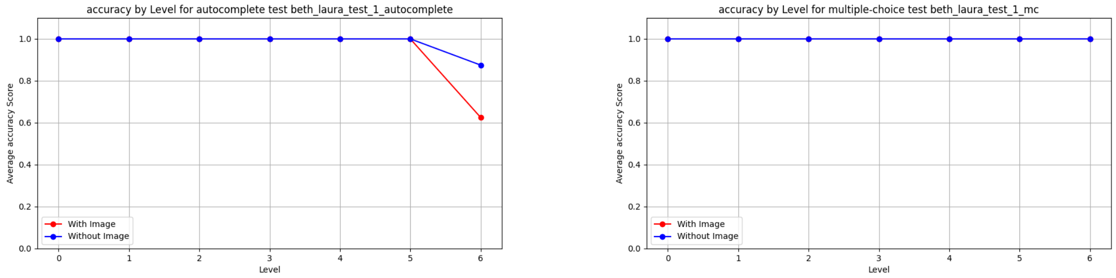


(a) Sally Anne 2 Ice Cream Autocomplete (b) Sally Anne 2 Ice Cream Multiple Choice

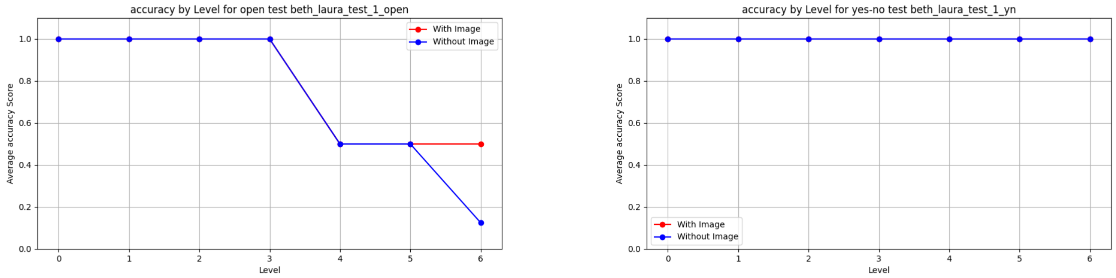


(c) Sally Anne 2 Ice Cream Open (d) Sally Anne 2 Ice Cream Yes/No

Figure 19: Accuracy plots for Sally Anne 2 Ice Cream test



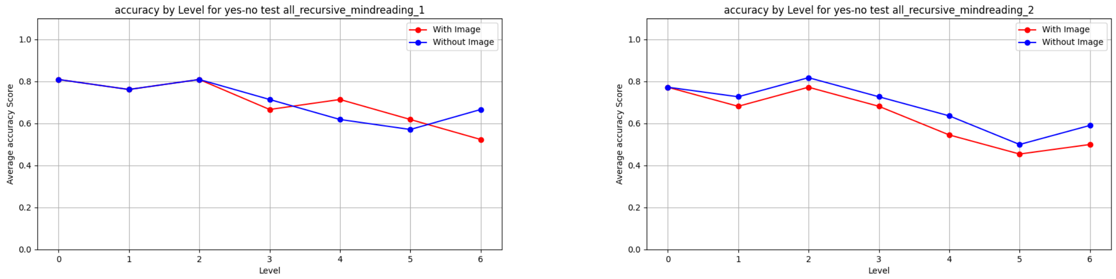
(a) Sally Anne 1 Autocomplete (b) Sally Anne 1 Multiple Choice



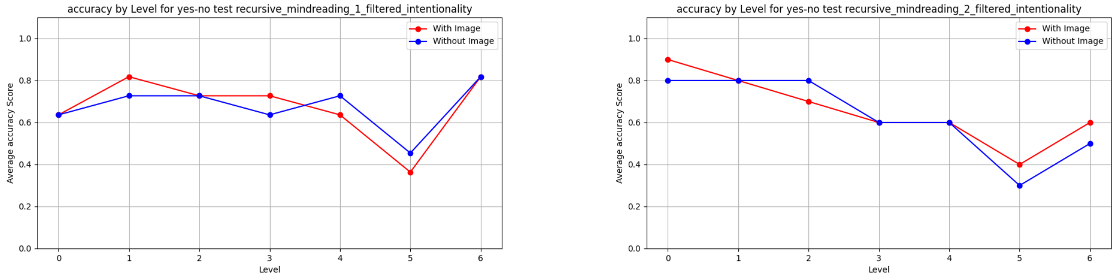
(c) Sally Anne 1 Open (d) Sally Anne 1 Yes/No

Figure 20: Accuracy plots for Sally Anne 1 test

F.1.2 Imposing Memory



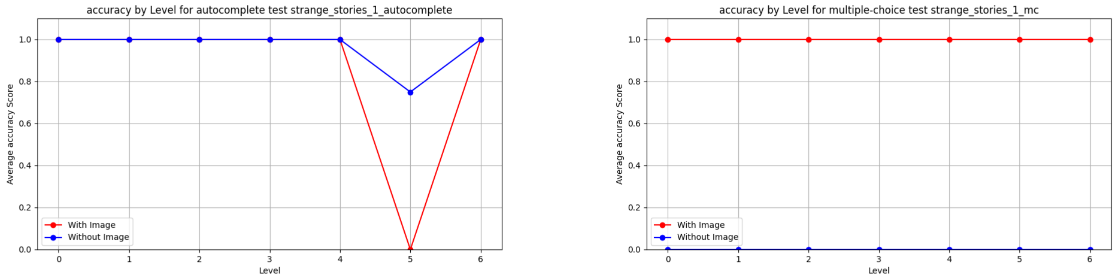
(a) Recursive Mindreading 1 (b) Recursive Mindreading 2



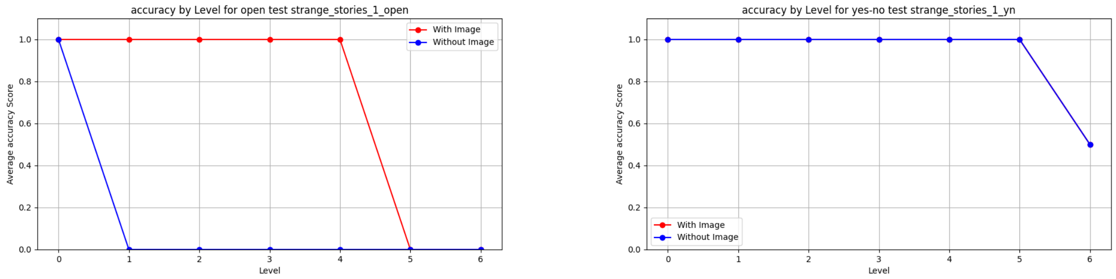
(c) Recursive Mindreading 1 Filtered Intentionality (d) Recursive Mindreading 2 Filtered Intentionality

Figure 21: Accuracy plots for Imposing Memory tests

F.1.3 Strange Stories

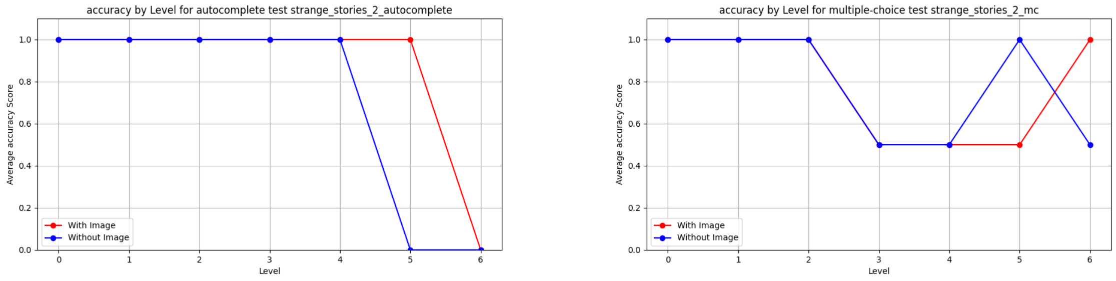


(a) Strange Stories 1 Autocomplete
(b) Strange Stories 1 Multiple Choice

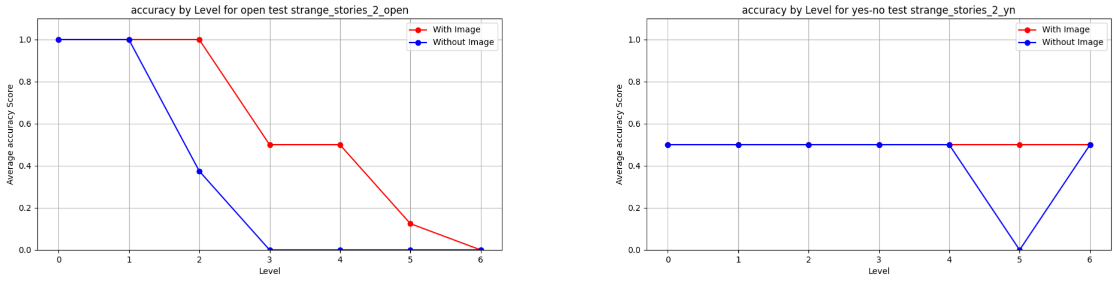


(c) Strange Stories 1 Open
(d) Strange Stories 1 Yes/No

Figure 22: Accuracy plots for Strange Stories 1

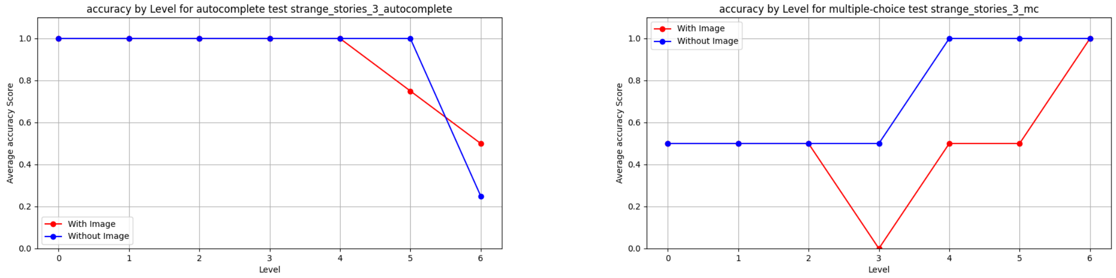


(a) Strange Stories 2 Autocomplete
(b) Strange Stories 2 Multiple Choice



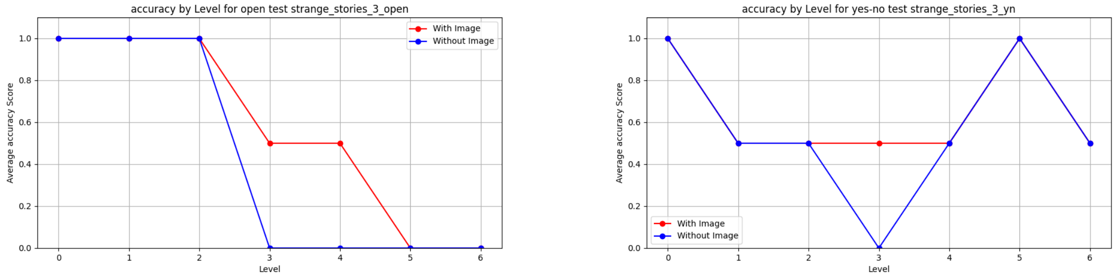
(c) Strange Stories 2 Open
(d) Strange Stories 2 Yes/No

Figure 23: Accuracy plots for Strange Stories 2



(a) Strange Stories 3 Autocomplete

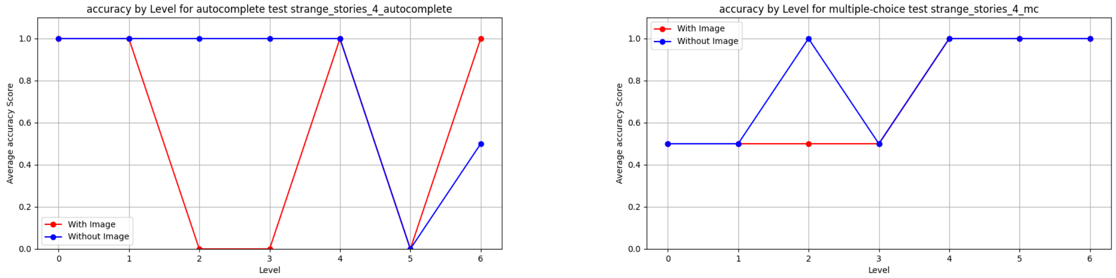
(b) Strange Stories 3 Multiple Choice



(c) Strange Stories 3 Open

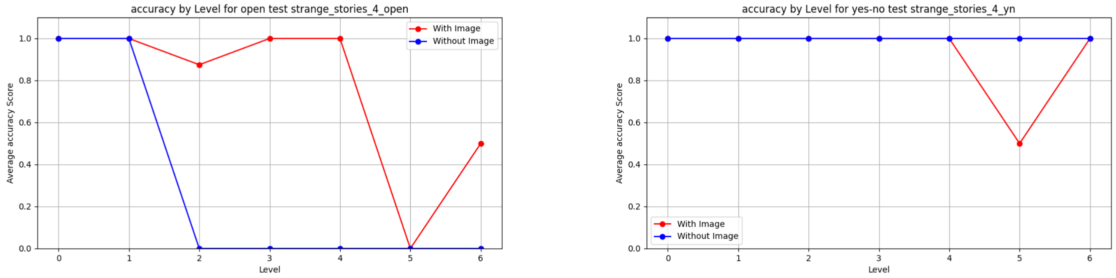
(d) Strange Stories 3 Yes/No

Figure 24: Accuracy plots for Strange Stories 3



(a) Strange Stories 4 Autocomplete

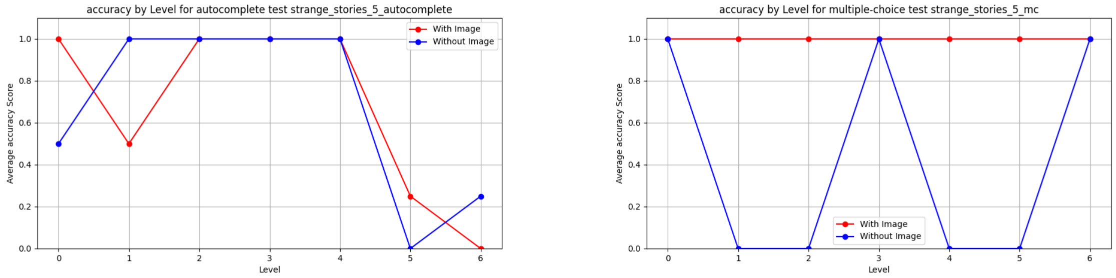
(b) Strange Stories 4 Multiple Choice



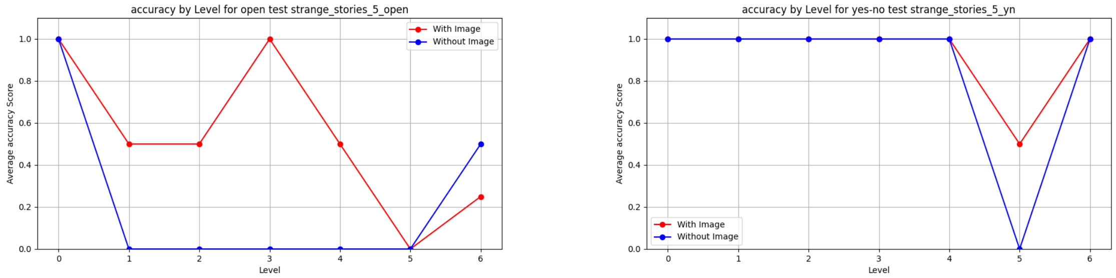
(c) Strange Stories 4 Open

(d) Strange Stories 4 Yes/No

Figure 25: Accuracy plots for Strange Stories 4

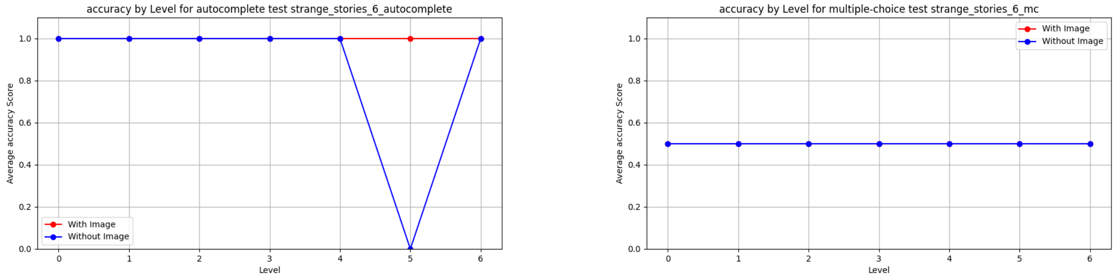


(a) Strange Stories 5 Autocomplete (b) Strange Stories 5 Multiple Choice

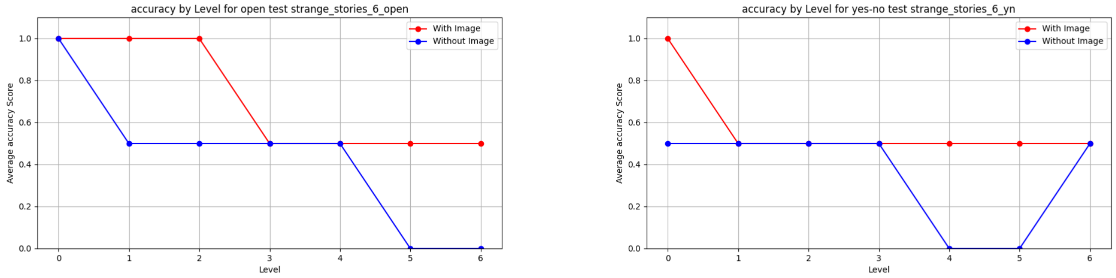


(c) Strange Stories 5 Open (d) Strange Stories 5 Yes/No

Figure 26: Accuracy plots for Strange Stories 5

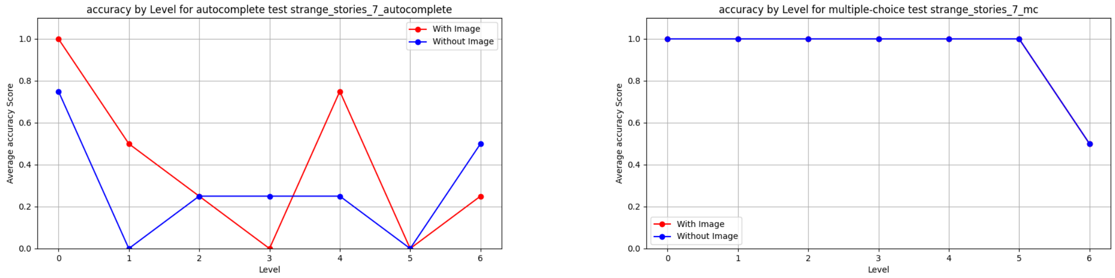


(a) Strange Stories 6 Autocomplete (b) Strange Stories 6 Multiple Choice

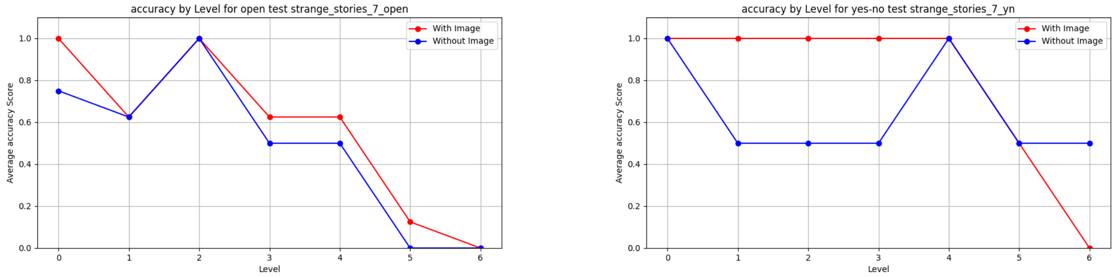


(c) Strange Stories 6 Open (d) Strange Stories 6 Yes/No

Figure 27: Accuracy plots for Strange Stories 6



(a) Strange Stories 7 Autocomplete (b) Strange Stories 7 Multiple Choice

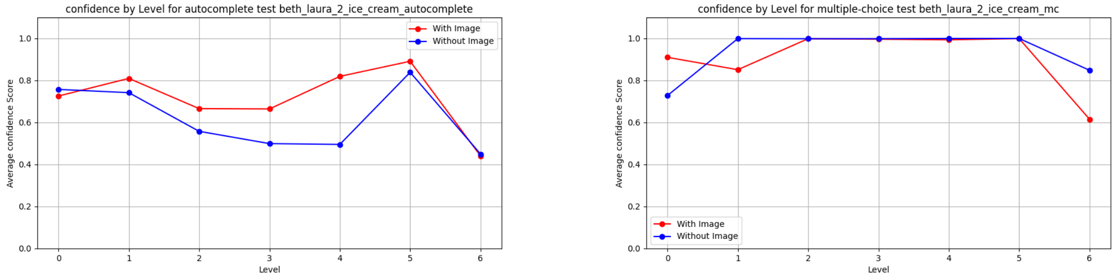


(c) Strange Stories 7 Open (d) Strange Stories 7 Yes/No

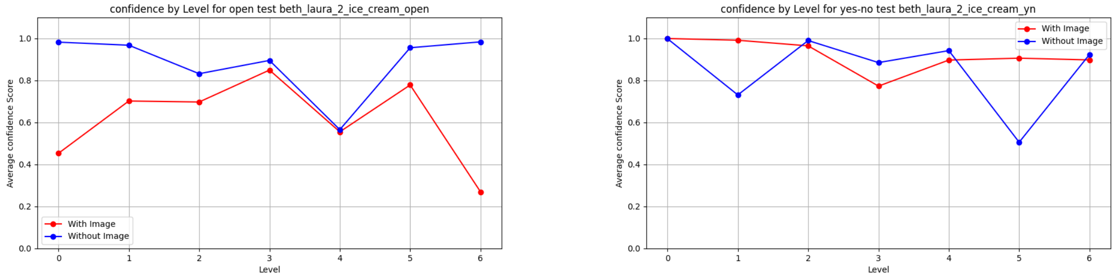
Figure 28: Accuracy plots for Strange Stories 7

F.2 Confidence Plots

F.2.1 Sally Anne 1 and 2

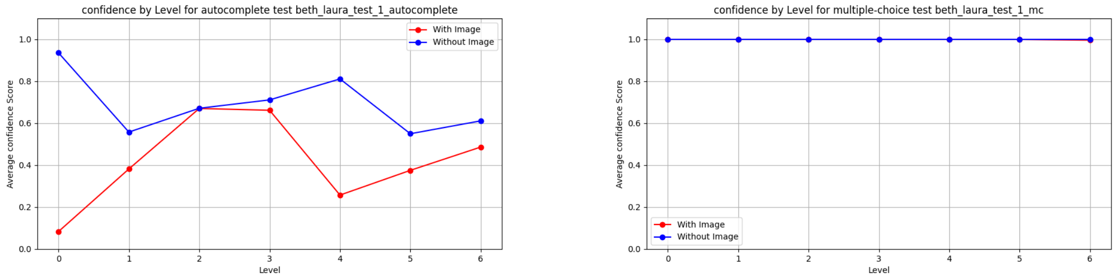


(a) Sally Anne 2 Ice Cream Autocomplete (b) Sally Anne 2 Ice Cream Multiple Choice

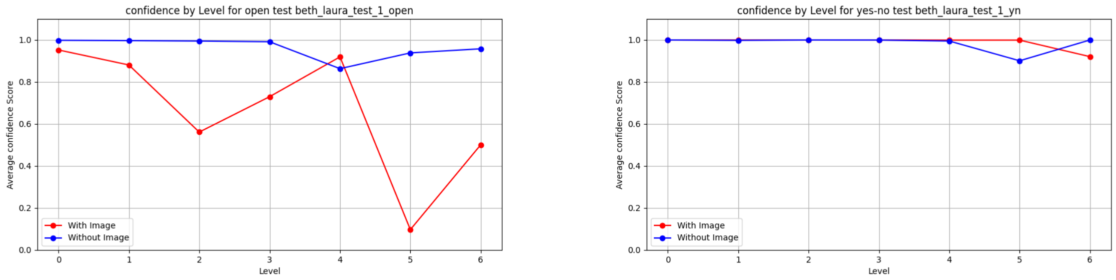


(c) Sally Anne 2 Ice Cream Open (d) Sally Anne 2 Ice Cream Yes/No

Figure 29: Confidence plots for Sally Anne 2 Ice Cream test



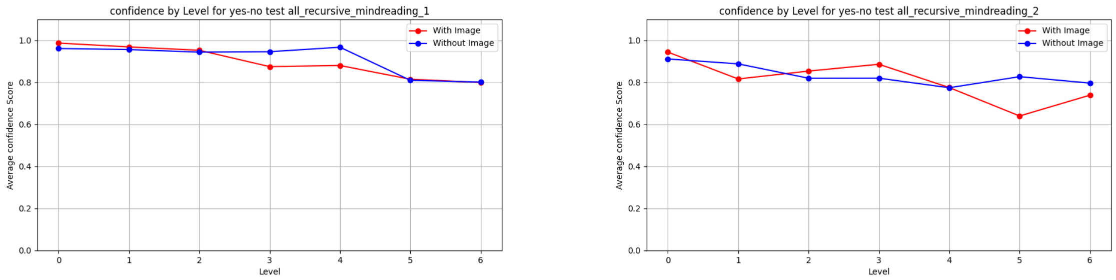
(a) Sally Anne 1 Autocomplete (b) Sally Anne 1 Multiple Choice



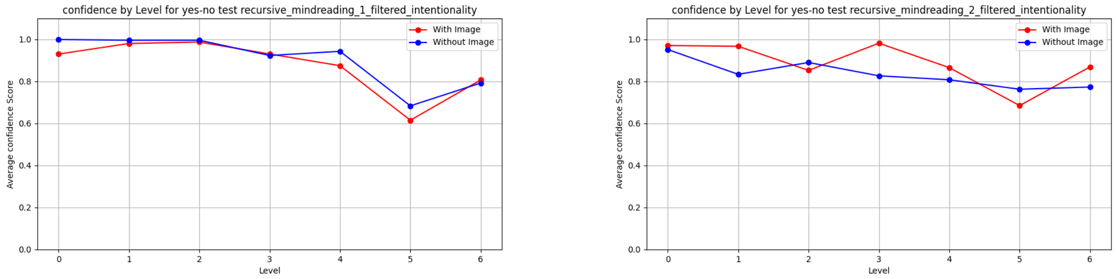
(c) Sally Anne 1 Open (d) Sally Anne 1 Yes/No

Figure 30: Confidence plots for Sally Anne 1 test

F.2.2 Imposing Memory



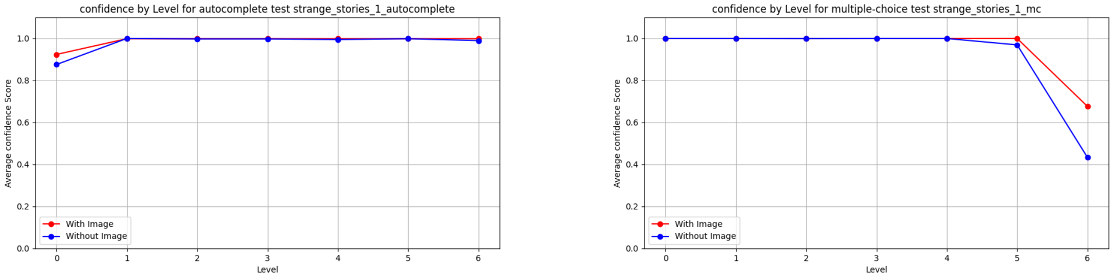
(a) Recursive Mindreading 1 (b) Recursive Mindreading 2



(c) Recursive Mindreading 1 Filtered Intentionality (d) Recursive Mindreading 2 Filtered Intentionality

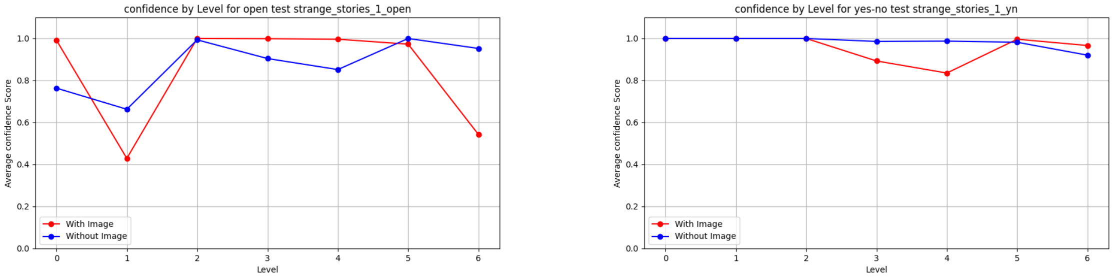
Figure 31: Confidence plots for Imposing Memory tests

F.2.3 Strange Stories



(a) Strange Stories 1 Autocomplete

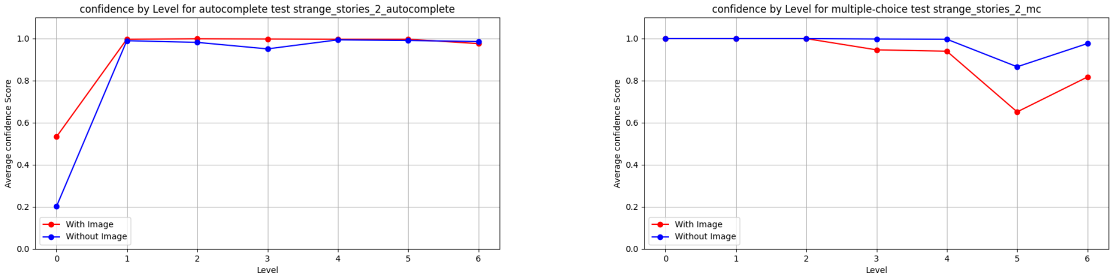
(b) Strange Stories 1 Multiple Choice



(c) Strange Stories 1 Open

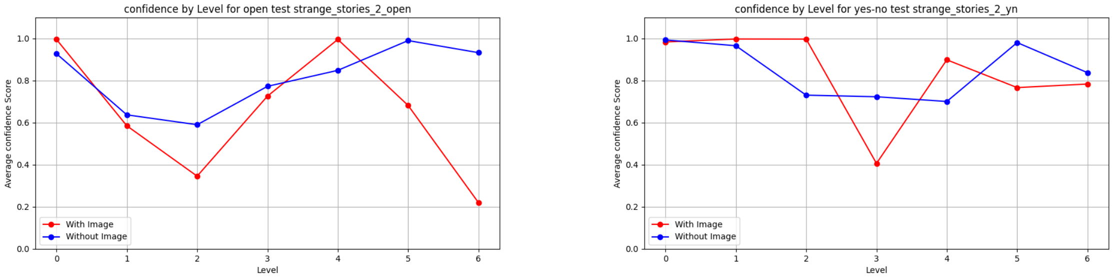
(d) Strange Stories 1 Yes/No

Figure 32: Confidence plots for Strange Stories 1



(a) Strange Stories 2 Autocomplete

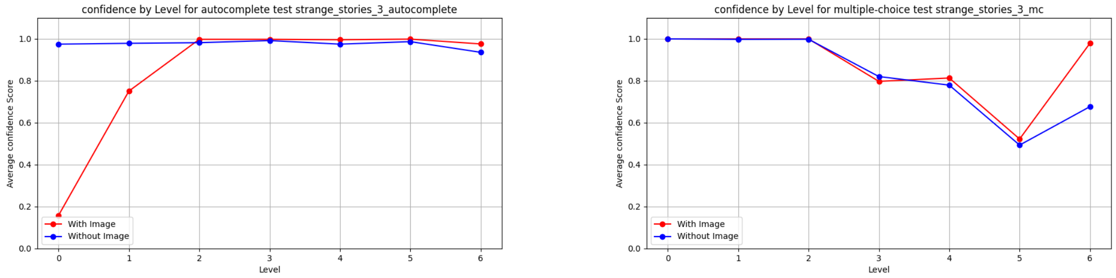
(b) Strange Stories 2 Multiple Choice



(c) Strange Stories 2 Open

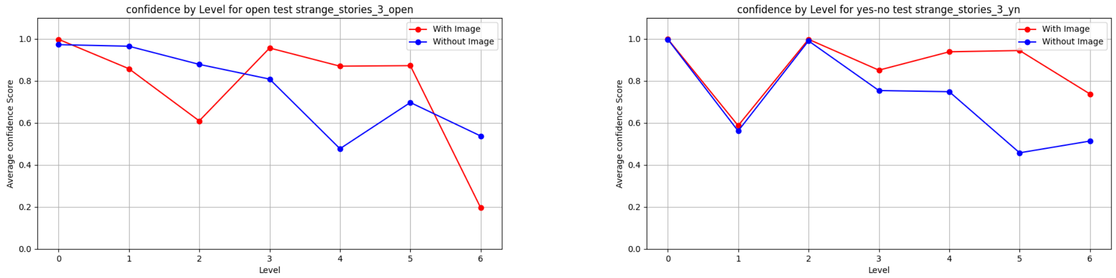
(d) Strange Stories 2 Yes/No

Figure 33: Confidence plots for Strange Stories 2



(a) Strange Stories 3 Autocomplete

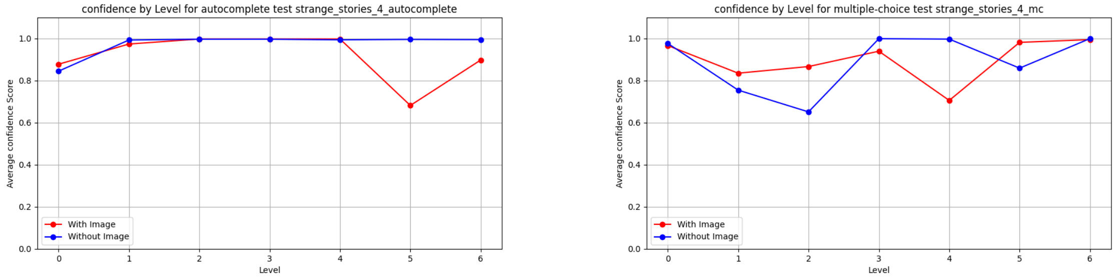
(b) Strange Stories 3 Multiple Choice



(c) Strange Stories 3 Open

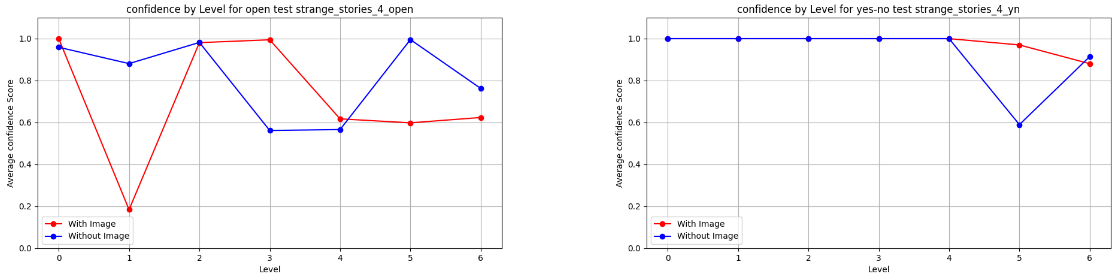
(d) Strange Stories 3 Yes/No

Figure 34: Confidence plots for Strange Stories 3



(a) Strange Stories 4 Autocomplete

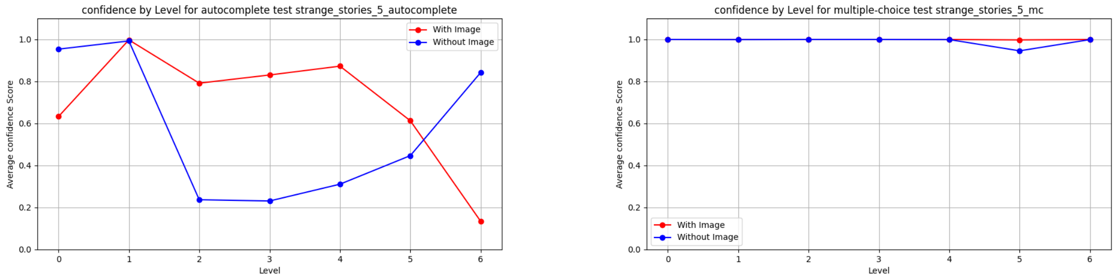
(b) Strange Stories 4 Multiple Choice



(c) Strange Stories 4 Open

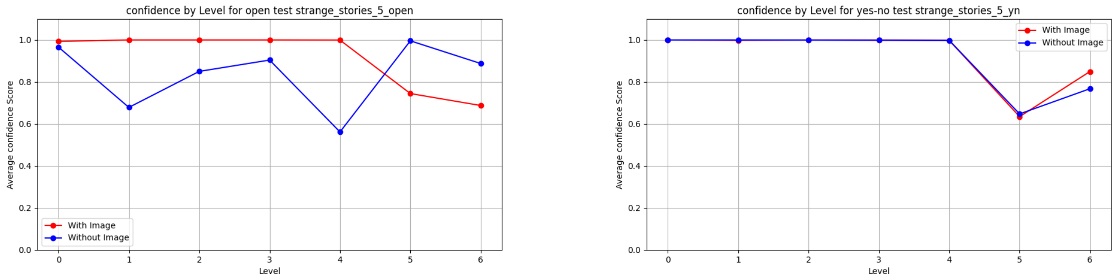
(d) Strange Stories 4 Yes/No

Figure 35: Confidence plots for Strange Stories 4



(a) Strange Stories 5 Autocomplete

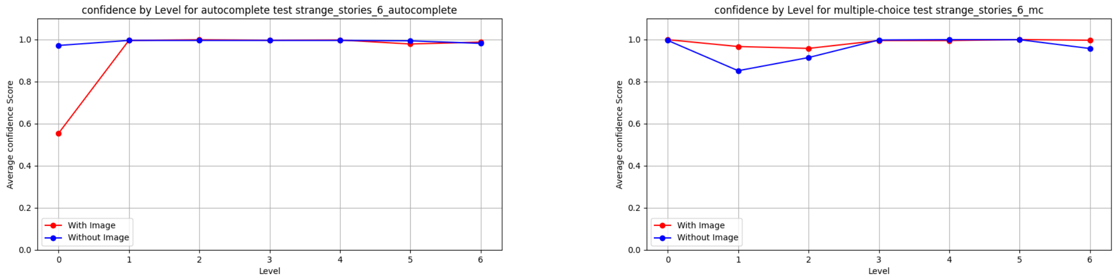
(b) Strange Stories 5 Multiple Choice



(c) Strange Stories 5 Open

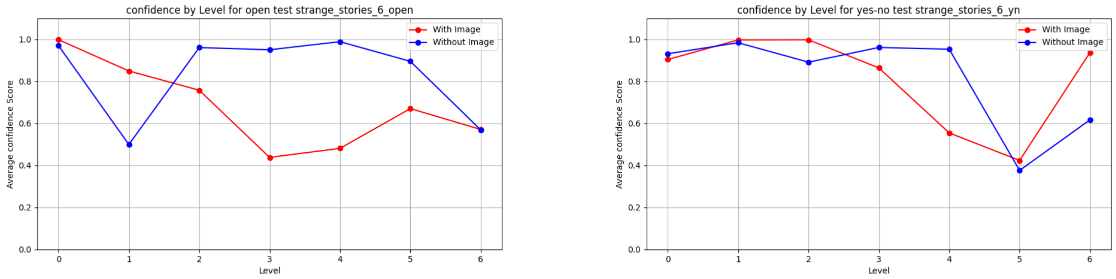
(d) Strange Stories 5 Yes/No

Figure 36: Confidence plots for Strange Stories 5



(a) Strange Stories 6 Autocomplete

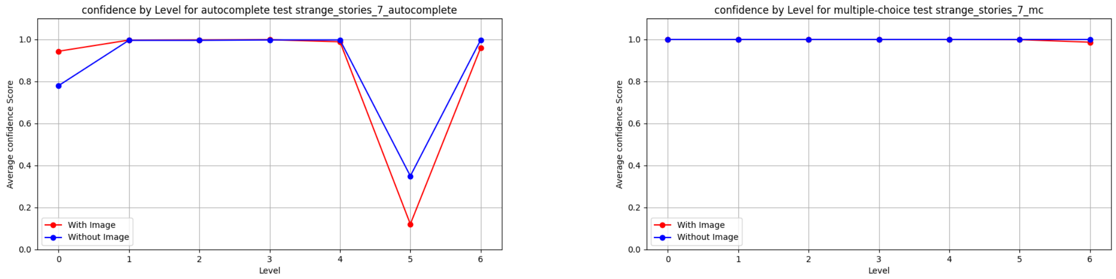
(b) Strange Stories 6 Multiple Choice



(c) Strange Stories 6 Open

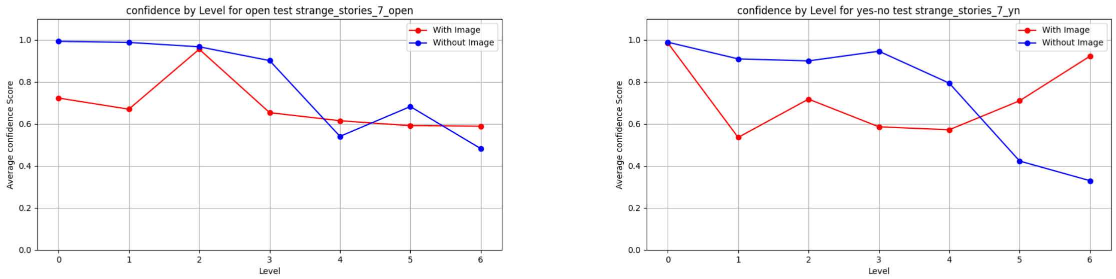
(d) Strange Stories 6 Yes/No

Figure 37: Confidence plots for Strange Stories 6



(a) Strange Stories 7 Autocomplete

(b) Strange Stories 7 Multiple Choice



(c) Strange Stories 7 Open

(d) Strange Stories 7 Yes/No

Figure 38: Confidence plots for Strange Stories 7