big data and cognitive computing

MDPI

Article

Image First or Text First? Optimising the Sequencing of Modalities in Large Language Model Prompting and Reasoning Tasks

Grant Wardle and Teo Sušnjak *

School of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand
* Correspondence: t.susnjak@massey.ac.nz

Abstract: Our study investigates how the sequencing of text and image inputs within multi-modal prompts affects the reasoning performance of Large Language Models (LLMs). Through empirical evaluations of three major commercial LLM vendors—OpenAI, Google, and Anthropic—alongside a user study on interaction strategies, we develop and validate practical heuristics for optimising multi-modal prompt design. Our findings reveal that modality sequencing is a critical factor influencing reasoning performance, particularly in tasks with varying cognitive load and structural complexity. For simpler tasks involving a single image, positioning the modalities directly impacts model accuracy, whereas in complex, multi-step reasoning scenarios, the sequence must align with the logical structure of inference, often outweighing the specific placement of individual modalities. Furthermore, we identify systematic challenges in multi-hop reasoning within transformer-based architectures, where models demonstrate strong early-stage inference but struggle with integrating prior contextual information in later reasoning steps. Building on these insights, we propose a set of validated, user-centred heuristics for designing effective multi-modal prompts, enhancing both reasoning accuracy and user interaction with AI systems. Our contributions inform the design and usability of interactive intelligent systems, with implications for applications in education, medical imaging, legal document analysis, and customer support. By bridging the gap between intelligent system behaviour and user interaction strategies, this study provides actionable guidance on how users can effectively structure prompts to optimise multi-modal LLM reasoning within real-world, high-stakes decision-making contexts.

Keywords: multi-modal prompting; interactive AI systems; user-guided AI adaptation; multi-modal large language models; modality fusion; multi-modal reasoning; chain-of-thought reasoning; human–AI interaction; user-centred prompt engineering

Academic Editor: Ximing Li

Received: 27 March 2025
Revised: 16 May 2025
Accepted: 26 May 2025
Published: 3 June 2025

Citation: Wardle, G.; Sušnjak, T. Image First or Text First? Optimising the Sequencing of Modalities in Large Language Model Prompting and Reasoning Tasks. *Big Data Cogn. Comput.* **2025**, *9*, 149. https://doi.org/10.3390/bdcc9060149

1. Introduction

Recent advancements in Large Language Models (LLMs) have profoundly impacted natural language understanding and related fields seeking to automate tasks involving human language. These developments now comprise a significant part of interactive intelligent systems which leverage sophisticated AI models to create an adaptive and responsive ecosystem of tools that facilitate human–computer interaction with advanced reasoning capabilities. Such systems are now integral for interface designers and user-experience researchers seeking to deploy multi-modal LLMs in real-world settings. Reasoning was once considered a uniquely human trait [1]; however, parallels are now observed between human cognition and the Theory of Mind [2] and the characteristics of LLMs. The emergent

reasoning abilities of LLMs to solve complex tasks that require high-order cognitive abilities have generated significant academic attention [3–6], as well as concerns in some fields [7] about the trajectory of such AI agents. Considerable research efforts have been devoted to improving LLMs' reasoning abilities recently, which, while impressive, have nevertheless been uneven and variable across different tasks [3].

The recent emergence of multi-modal LLMs enabling the processing of a wide variety of inputs, from textual to visual, has further increased the complexity of reasoning tasks across modalities [8–10], raising important questions about how best to structure and sequence the modalities in these prompts to enhance both reasoning and user interaction quality. Therefore, to address these challenges, it is necessary to first investigate and understand how LLMs process and respond to various types of multi-modal information before practical insights that inform users how they can optimally interact with these technologies can be extracted. However, to date, there is a gap in research that sheds light on this issue.

When it comes to evaluating the reasoning capabilities of multi-modal LLMs, visual question-answering (VQA) prompts involving a combination of image(s) and multiple-choice questions have become a common method for assessment [3,4,11–13]. Evaluation benchmark datasets for this task have emerged [4], covering a wide range of disciplines while often taking the form of academic exam-like questions [3]. Notably, models such as GPT-4 [14], Gemini-1.5 [15], and Claude [16] have displayed varying degrees of multi-modal reasoning capabilities in the context of VQA [3,4,11–13]; however, most studies have considered only relative performances between LLMs. While existing research has explored multi-modal LLMs, a systematic understanding of their performance profiles across different types of multi-modal inputs, particularly regarding modality sequencing, remains limited. Initial research has indicated that LLMs significantly struggle with multi-modal reasoning tasks [8–10]. GPT-4 specifically has been found to exhibit limitations in processing visual information alongside text for complex reasoning tasks [8]. Meanwhile, LLMs have also been reported to demonstrate variable performance in medical VQA, with significant deficits in complex reasoning once again being noted, especially in medical imaging [9]. Similar findings were observed in other biomedical science exams, where GPT-4 performed poorly on figure-based questions [10].

While the inconsistent performance profile of LLMs in reasoning on multi-modal VQA has not been fully explained, it is well understood in the educational domain concerning human subjects that the layout of exam questions affects students' performance [17]. Similarly, in the AI context, the effectiveness of an LLM's output depends on the quality and structure of the prompt. The manner in which a prompt is constructed can either effectively focus the attention of the LLM on relevant information or divert it by introducing distractions that significantly affect response accuracy. The relative position of words [18,19], the position of an object in an image [20], minor changes in wording or phrasing [21–23], the order of instructions [24], and even the length of the prompt [25] can all influence the accuracy of responses. In the scope of multi-modal reasoning tasks, these variables become more pronounced, as modalities are fused and introduced to LLMs through various strategies that are not transparent to users, rendering their impact on response accuracy uncertain. Consequently, a significant challenge currently exists for users to determine the most effective way to construct multi-modal prompts for optimising reasoning to produce correct responses.

Therefore, our study sought to conduct a comprehensive set of multi-modal experiments across different datasets and LLMs, from which practical guidelines in the form of heuristics could be extracted for providing users with concrete recommendations on how to construct prompts with sequencing of modalities that on average generate more optimal

results. Recently, different strategies exploring ways to enhance the multi-modal reasoning of LLMs have started to emerge [26–31]. However, these strategies, while effective in certain constrained contexts, have tended to focus on a single modality without considering the interplay of modalities on the performance of reasoning tasks, and extensive experiments yielding findings and observations regarding the optimal structuring of multi-modal prompts have not been conducted. Prior studies have also not considered the human–computer interaction aspect. Our work extends and builds upon research suggesting that variations in text prompting, such as the relative position of words [18,19] or the order of instructions [24], can significantly impact LLM performance. Understanding whether LLMs are influenced by the ordering of modalities within prompts is crucial to optimising multi-modal reasoning, thereby allowing for greater value extraction from these technologies across numerous domains. Consequently, this research study sought to perform an extensive series of experiments to ascertain if the sequence of input modalities influences reasoning tasks and to what extent, akin to the impact of altering the instruction order in text prompts [24]. This research study additionally explored whether particular elements within the image and text input modalities for LLMs display sensitivity to the sequence of images and text and whether these elements can be adjusted to enhance response performance. Our experimental findings, therefore, enabled us to construct concrete guidelines that users can follow in order to maximise the accuracy of multi-modal interactions with LLMs that require complex reasoning. We summarise our main contributions as follows:

1. We systematically evaluated the impact of image and text prompt-sequencing on the reasoning performance of three multi-modal LLMs: GPT-4o, Gemini-1.5-Flash, and Claude-3-Haiku. Our findings demonstrate that modality sequencing significantly affects performance, particularly in complex reasoning tasks.
2. We identified specific attributes within image and text modalities that exhibit higher sensitivity to sequencing. The results indicate that different reasoning tasks benefit from distinct sequencing strategies.
3. We derived a set of practical guidelines in the form of multi-modal heuristics seeking to maximise user experience by instructing users on creating prompts based on our empirical findings, and we validated them through a user interaction study.

The remainder of the paper is structured as follows: Section 2 reviews related work. Section 3 details our methodology, including datasets, LLMs, and experimental design. Section 4 presents empirical results from our Image–Text Sequence Variation and attribute-based analyses, as well as the development and evaluation of user-centred heuristics. Section 5 discusses implications, limitations, and future research directions. Section 6 concludes by summarising the key contributions.

2. Related Work

Reasoning can be defined as the cognitive process of drawing inferences or conclusions from premises, evidence, or observations, involving the systematic application of logical principles to analyse information, solve problems, and make decisions [32]. Reasoning encompasses both deductive methods, where conclusions necessarily follow from given premises, and inductive approaches, where generalisations are formed from specific instances. The assertion that reasoning is a genuine emergent behaviour in LLMs is contentious in the academic literature [33,34]. Emergent abilities within LLMs have been defined as capabilities present in larger but not smaller models, with reasoning being identified as one of these properties [35] that arise as the parameter size of language models has grown. However, recent investigations [36,37] suggest that current LLMs find it challenging to tackle intricate reasoning tasks that humans handle with relative ease, lacking profound understanding and instead relying on superficial pattern recognition or dataset

biases. Studies [37,38] also argue that contemporary LLMs are confined to intuitive, reflexive tasks, rather than those necessitating logical and deliberate analysis associated with true higher-level reasoning, while others [39] assert that LLMs cannot genuinely reason or plan at all but only appear to do so. Additional research [40,41] further contends that the impressive generative capabilities of LLM-based systems do not reflect true understanding but are merely a function of word prediction.

Irrespective of whether the reasoning ability exhibited by LLMs is a truly emergent property or a form of pattern-matching mimicry, this ability has been found to generalise and therefore be useful in solving many reasoning tasks [42,43], thereby giving rise to the development of strategies aiming to maximise their reasoning effectiveness even further. The most recognised way of improving LLM reasoning through prompting is the Chain-of-Thought (CoT) [44] prompting technique *"Let's think step by step..."* , which has proven effective in enhancing zero-shot and few-shot capabilities [44–46]. In LLMs, this method mirrors the cognitive process of breaking down problems into manageable steps, allowing the model to process each step sequentially in a linear fashion, ultimately leading to a conclusive answer [47]. Recent advancements in multi-modal reasoning for LLMs have focused on enhancing CoT methods to address challenges like their weak spatial reasoning, localisation awareness [48,49], and high-resolution image interpretation [29]. The upcoming challenge in reasoning complexity lies in further enhancing the abilities of LLMs to reason across various input modalities, including text and image elements, and eventually also other multimedia types [31]. Research in this area is nascent but has repeatedly shown the need to devise improved means for LLMs to perform multi-modal reasoning more reliably [9,10]. While multi-step reasoning follows a sequential approach to draw conclusions as exemplified by CoT approaches, multi-hop reasoning, requires making several inferential jumps among unconnected data points or different modalities to form a coherent answer, which presents a significant degree of difficulty for transformer-based architectures, which are unable to iteratively plan and refine their responses.

2.1. Multi-Modal Prompting Techniques

Several studies have focused on improving and addressing LLMs' challenges within the vision modality. Techniques such as Compositional Chain-of-Thought (CCoT) prompting [26] use scene graph-based prompting to achieve this, while Image-of-Thought (IoT) prompting [27] extracts visual rationales in a step-by-step manner. Meanwhile, TextCoT [29] divides images into global and local regions to assist with reasoning, while Duty-Distinct Chain-of-Thought (DDCoT) prompting [28] employs a two-stage framework to separate reasoning roles for visual and language modalities. Multi-modal Chain-of-Thought (MCoT) prompting [30] improves multi-modal reasoning by initially partitioning LLM responsibilities into reasoning and recognition before integrating vision information within smaller models. Although these models examine the interaction between modalities, they treat them as distinct components that can be processed independently. These strategies, while effective in certain contexts, have not considered how the sequencing of modalities affects reasoning performance.

2.2. Image Sequencing

In human behaviour, the *primacy* effect suggests that individuals are more likely to recall information presented at the beginning of a sequence [50], in contrast to the *recency* effect, which implies a contrary bias towards information at the end of a sequence [51]. Both the primacy and recency effects have been demonstrated to exist within LLMs [18,19,52–54]; however, these have not been comprehensively studied and explored in the context of multi-modal LLMs and reasoning tasks. Vendors [16,55,56] of large commercial LLMs have

tended to advise that in cases involving prompts with images, there is a primacy effect that impacts performance (both Google [55] and Anthropic [16] recommend placing the image first to achieve the best results; the OpenAI community pages [56] have a more nuanced recommendation, suggesting that placing the image first often helps the LLM in understanding the tasks and framing the problem). For general tasks where the image is the focus, this logic makes sense; however, for reasoning tasks where key instructions are often in a dedicated question component, this may not hold true. To the best of our knowledge, there is little information on why this is recommended or evaluations on different types of tasks for image position.

2.3. Multi-Modal Fusion Strategies and Positional Bias

The architectural design of LLMs, particularly transformers [57], fundamentally influences how information sequencing is processed, affecting multi-modal reasoning tasks, and is thus relevant to consider for effective human–computer interaction. Transformers utilise attention mechanisms and positional encoding to assign context-aware weights to input data, maintaining the sequence order crucial to understanding context and syntax [57,58]. This structure enables LLMs to focus on different aspects simultaneously through self-attention layers but introduces positional biases that can impact performance. Effective integration of multiple modalities like text and images in LLMs requires strategic fusion methods, namely, *early fusion*, *late fusion*, and *hybrid fusion*, each with specific implications for multi-modal prompting and reasoning [59] and effective human–computer interaction. Early fusion combines modalities at the input level, allowing models to learn cross-modal interactions from the outset [60,61], but may face challenges in processing efficiency with high-dimensional data [61,62]. Late fusion processes modalities independently, offering computational advantages [63], but may not capture nuanced cross-modal relationships essential to integrated reasoning [64]. Hybrid fusion blends both strategies to leverage their strengths, providing flexibility in modelling cross-modal relationships [63,65]. Most recent research [66] has found that early fusion tends to both be more efficient to train and exhibit stronger performance in terms of accuracy, particularly at lower parameter counts, when compared with late fusion models. Large vendors of proprietary LLMs typically do not disclose the implementation details of their commercial multi-modal models, which can make it challenging to know how to optimise prompts for the most accurate reasoning responses.

Understanding the influence of positional bias and modality sequencing is important for optimising multi-modal LLMs in complex reasoning tasks. Recent research indicates that both text and images are susceptible to positional bias due to the mechanics of causal attention and positional encoding [20,22,23,67]. Even minor changes in the instruction order or phrasing can significantly impact performance [21,23,24]. Generally, in early fusion architectures, modality sequencing greatly affects how the model attends to and integrates information, while in hybrid and late fusion systems, prompt design still plays a role in performance. Therefore, grasping these internal mechanics across fusion strategies enables more effective prompt design and the optimisation of multi-modal LLMs for complex reasoning tasks.

2.4. Human-Centred Approaches to Navigating Complexity and Uncertainty in Interactive AI

As discussed, optimising LLM prompting to enhance reasoning—especially across multiple modalities and within unknown or evolving LLM architectures—requires addressing incomplete data, variable user needs, and a high degree of configurability. Recent human-centred AI work speaks directly to these challenges, showing how uncertainty and contextual complexity can become generative components in user-centric design strategies. For example, Giaccardi et al. [68] foreground the notion that uncertainty or partial data

need not be dismissed as a deficit; instead, they can spur iterative, user-driven prototyping approaches that resonate with interactive systems in HCI. Thieme et al. [69] underscored the critical importance of *practical usefulness* in AI deployments, contending that systems must be interwoven with everyday practice and be configurable for varied environments. In our context of multi-modal prompt design, this translates into empowering users to guide LLM attention—deciding which text or image inputs come first—so that the reasoning pipeline aligns with real-world constraints. Zając et al. [70] emphasised that genuinely useful AI emerges more from *flexibility and adaptability* than from raw performance metrics. Extending this to multi-modal LLM tasks, our study similarly posits that no single, universal sequencing strategy exists: the optimal approach is contingent on data specifics, user expertise, and task demands. Meanwhile, Chen et al. [71] provided a concrete example of user-driven refinement within a 3D design system, demonstrating how the interplay between textual prompts and user annotations can rectify incomplete or ambiguous states. August et al. [72] similarly showed that AI designed for medical text comprehension only achieves real usability when users are able to clarify domain-specific jargon and contextual details—underscoring the importance of iterative, user-centred knowledge co-creation. With close affinity, Huang et al. [73] extended these insights to UI task automation, employing multi-agent reasoning and user oversight to robustly map partial textual instructions onto intricate mobile GUIs. In a similar vein, our study incorporates dynamic context injection and user guidance through carefully sequenced multi-modal prompts. We embrace a *configurable* pipeline where missing context is systematically surfaced and addressed, and incomplete or ambiguous modalities are dynamically adapted based on user feedback. This stage-wise, cooperative process not only resonates with the HCI tradition of iterative, user-centred design but also speaks directly to the emergent complexities of multi-modal LLM reasoning.

2.5. Research Questions

The recent literature has collectively begun to converge towards investigations that seek to uncover strategies to optimise prompts for maximising LLM performance and reasoning. While existing research has mostly tended to focus on enhancing performance gains within a single modality (text), in cases where multi-modal information was considered, the studies have typically overlooked the impact of information sequencing in multi-modal contexts and how different and unknown multi-modal fusion strategies may represent a confounding factor that affects responses. Therefore, our research has aimed to bridge this gap by examining how the sequencing of images and text affects LLM performance in reasoning tasks. To this end, this study's guiding research questions are the following:

- RQ1: To what extent does the sequencing of image and text modalities in prompts affect the reasoning performance of multi-modal LLMs having unknown and potentially different multi-modal fusion strategies, across different benchmark datasets and question types?
- RQ2: How do specific attributes of questions, such as nested structure, subject domain, and complexity, interact with modality sequencing to influence LLM performance, and how does this vary across different LLMs?
- RQ3: To what degree is the impact of modality sequencing on LLM performance attributable to the order of information presentation rather than the inherent properties of different modalities, and how can these insights be applied to optimise multi-modal prompt construction?
- RQ4: Can practical heuristics and guidelines be developed to assist users in constructing multi-modal prompts that optimise both their reasoning accuracy and enable a more effective user experience with interactive intelligent systems?

3. Methodology

Our methodological design comprises two distinct parts. In the first part, we designed a series of experiments on two benchmark datasets to ascertain the effects of modality sequencing on response accuracy. The second part was user-centric. The purpose of the second part was to extract, from the initial experiments, heuristics and general guidelines which users can apply when interacting with LLMs for solving multi-modal tasks. This part also involved evaluating the accuracy of the extracted heuristics by observing user interactions with an LLM on a test dataset.

3.1. Datasets

Our evaluations used two recently developed multi-modal multiple-choice reasoning benchmarks for LLMs, namely, M3Exam [3] and M3COTS [4]. These benchmarks were developed with questions that integrate visual and textual information and were thus selected in our experiments due to their ability to present models with both complex and demanding reasoning tasks from multiple modalities.

3.1.1. M3Exam Dataset

M3Exam [3] offers a diverse range of real exam questions across various educational levels. For our evaluation, we selected the multi-modal English question set, which contains 795 questions across 4 overarching subjects (social science, natural science, language, and maths), 11 subcategories, and 3 educational levels (elementary, middle, and high school) in the USA. The average word count across the questions and background information is approximately 95 words.

The M3Exam dataset structures each question in JSON format, dividing it into three key parts: `background_description`, which provides additional context in some cases; `question_text`, which contains the actual questions; and `options`, which represents the multiple-choice responses. Images can be dispersed across all three elements and, sometimes, in multiple places per question, which further amplifies the complexity of the questions. An example of an exam question with three components can be seen in Figure 1, with guidance suggesting that the image component be placed in the `question_text` section of the overall question. This particular question does possess an empty `background_description` component.

```
{
    "background_description": [],
    "question_text": "The diagram below represents the electric field surrounding two charged
    spheres, A and B.\n\n(image)[image-5.jpg]\n\nWhat is the sign of the charge of each
    sphere?",
    "options": [
        "(1) Sphere A is positive and sphere B is negative.",
        "(2) Sphere A is negative and sphere B is positive.",
        "(3) Both spheres are positive.",
        "(4) Both spheres are negative."
    ]
}
```

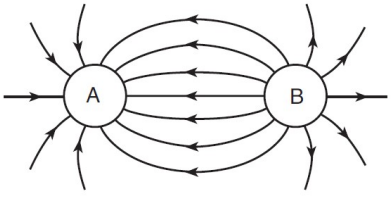
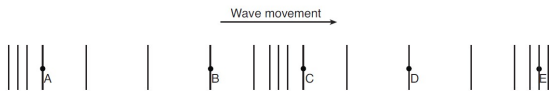


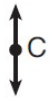
Figure 1. M3Exam example, question 5, representing 'image-5.jpg' from the above exam question.

Since visual elements can be distributed across the three elements at the same time, the complexity arising from multiple multi-modal inputs can be significant for some exam questions. An example is given in Figure 2, where an image component is allocated to the `background_description` component, while a further four images are allocated to each of the four answer options. The JSON structure of the question is depicted below, showing image placeholders denoted by `(image)[image-x.jpg]`. Overall, the questions in the dataset range from having 1 to a maximum of 5 images, averaging 1.2 images per question in the dataset.

```
{
    "background_description": [
        "A longitudinal wave moves to the right through a uniform medium, as~shown
        below. Points A, B, C, D, and~E represent the positions of particles of the
        medium.\textbackslash n \textbackslash n(image)[image-10.jpg]"
    ],
    "question_text": "Which diagram best represents the motion of the particle at
    position C as the wave moves to the right?",
    "options": [
        "(1) (image)[image-11.jpg]",
        "(2) (image)[image-12.jpg]",
        "(3) (image)[image-13.jpg]",
        "(4) (image)[image-14.jpg]"
    ]
}
```



(**a**) Image-10



(**b**) Image-11

(**c**) Image-12

(**d**) Image-13

(**e**) Image-14

Figure 2. Set of images from the M3Exam dataset showing a complex set of image arrangements.

An example of a complete reconstructed exam question is shown in Figure 3, where the image elements are situated in the background/context portion of the question. In terms of image placement, 87% of images in the dataset are situated inline within the `background_description` or `question` component and 6% within the options component. Meanwhile, 7% of the images appear at the start of the question within the `question_text` component. Given that the exam questions are deconstructed in the raw data, they lend themselves well to modifying the order in which the modalities are presented to the LLMs through the API calls. Further, since these are actual exam questions used in the US education sector, their layout is assumed to be optimised for student understanding. M3Exam has been used to evaluate models focused on different languages [74] and culture-related tasks [75], making it a versatile benchmark.

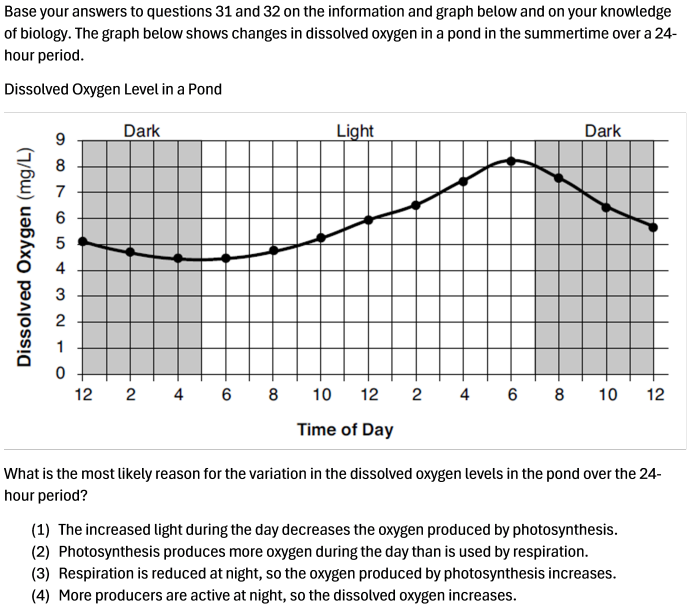
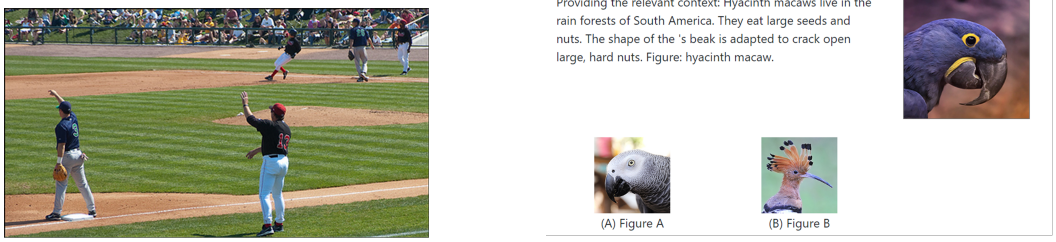


Figure 3. An example of a reconstructed M3Exam question.

3.1.2. M3COTS Dataset

The second dataset used in this analysis is M3COTS [4]. M3COTS features a selection of questions specifically chosen to challenge visual reasoning and multi-step reasoning across multiple subjects. The dataset includes science topics from the ScienceQA dataset [12] and mathematics questions from the MATH [76] and Sherlock [77] datasets, intended to test common-sense abductive reasoning beyond the literal image content. For our evaluation, we selected a random sample of 2318 questions (20% of the dataset) spanning 3 domains (representing the 3 original datasets), 9 subjects, and 92 question types. In this dataset, each question includes only one image, as opposed to M3Exam, which is a significant reduction in complexity. The average word count across the questions, background information, and options is approximately 45 words. A total of 10 percent of the images contain only visual content, 65 percent consist of a combination of images and text, and 25 percent feature text exclusively. Example questions from M3COTS are shown in Figure 4a,b. Note that as seen in Figure 4b, while each question in this dataset may be accompanied by only one image in the raw format, an image may embed multiple distinct images, and as well as text, within a single visual. Similar to the M3Exam dataset, M3COTS structures each question in JSON format, dividing it into three key parts: `context`, which provides additional background in some cases; the `question` component, which contains the actual question; and `choices`, which represents the multiple-choice responses. The images are not directly referenced in `context`, `question`, or `choices`.The example JSON structure of a M3COTS question can be seen below.

```
{
    "image": physics-26.png,
    "context": "Select the better answer.",
    "question": "Which property do these two objects have in common?",
    "choices": ["
                 "(A) sticky",
                 "(B) yellow"
                ]
}
```



(**a**) (**b**)

Figure 4. Typical text/image layouts across M3COTS dataset questions with the image first. (**a**) Example—images containing only visual content. (**b**) Example—images containing text and visual content.

The diverse range of domains and source datasets makes M3COTS a suitable benchmark for evaluating LLMs. The source and M3COTS dataset have been investigated extensively in research. CoT [44] prompting has proven to be the most effective technique, outperforming Direct Prompting, Description-based CoT (Desp-CoT) [78], and Compositional CoT (CCoT) [26].

3.2. LLMs

We selected three popular commercial models for our experiments: GPT-4o, Claude-3.5-Haiku, and Gemini-1.5-Flash. GPT-4o was developed by OpenAI and introduced in 2023 and is characterised by a large parameter count and extensive context length. These features enable sophisticated multi-modal interactions and complex reasoning tasks. Claude-3.5 Haiku was produced by Anthropic and is recognised for its speed and compact design. This model provides an ideal contrast to larger, more computationally intensive models like GPT-4o, offering insights into the trade-offs between model size and response latency. Lastly, Gemini-1.5 is Google's model and is regarded as another "lightweight" model optimised for speed and efficiency, complementing the other selections by focusing on streamlined performance.

The three models, with their varying capabilities and architectural designs, collectively provide a comprehensive overview of the current landscape in large-scale AI computations. The decision to focus on larger LLMs stems from existing studies [44] which suggest that the capability for Chain-of-Thought (CoT) reasoning may emerge in language models at a certain scale, specifically over 100 billion parameters. All models were accessed via their respective APIs, hosted on platforms capable of supporting extensive AI operations, thereby ensuring reliable and consistent performance throughout our studies.

The experiments were conducted in a zero-shot fashion, aiming to minimise direct exposure to the *specific test questions* prior to testing. We employed variations of CoT's prompts [44], and all testing was conducted by using greedy decoding at a temperature setting of 0.1. Our experiments used standard models without any fine-tuning to focus on the models' behaviour under direct interaction, which is the most common approach users take when engaging with language models. While we acknowledge that setting a low temperature does not entirely eliminate the inherent randomness in the behaviour of LLMs and that methods like self-consistency could further enhance robustness [79], these were deemed to be beyond the scope of this initial exploratory study focused on identifying primary modality sequencing effects.

In this research study, the focus was on examining the relative performance of the chosen LLMs across different image and text input configurations. Therefore, the primary aim was not to achieve maximal state-of-the-art performance but rather to understand how these models behave with changes in the sequencing configuration of the text and image inputs.

3.3. Experimental Design

This study conducted a series of experiments to evaluate how the sequencing of image and text modalities in prompts affects the multi-hop reasoning performance of multi-modal LLMs, structured around four primary setups: (1) *Image–Text Sequence Variation*, which examined the effects of different sequencing orders (Image First, Text First, and Interleaved) on model performance across two datasets; (2) *Image Versus Instructions Analysis*, aimed at determining whether the impact of sequencing is due to the image placement or the sequence of instructions by converting visual elements into text; (3) *Attribute-Based Sequencing Analysis*, which investigated how specific dataset attributes—such as image type, prompt length, and question complexity—influence the model's sensitivity to sequencing; (4) *User Interaction and Heuristic Evaluations* was introduced to bridge experimental insights with practical application. This phase focused on developing a set of actionable heuristics for users to manually sequence prompts effectively, ensuring improved reasoning performance. Table 1 summarises the entire experimental design, which is explained in further detail below.

Table 1. Overview of experimental design.

Experiment	Description	Configurations	Variables Analysed
Image–Text Sequence Variation	Evaluate effect of sequencing on model performance	Image First (IF) Text First (TF) Interleaved (IN)	Impact of sequencing on reasoning performance
Image vs. Instructions Analysis	Determine if observed sequencing impact is due to the visual modality itself or the sequencing of instructional information	Image First (IF) Text First (TF) Interleaved (IN)	Impact of sequencing on extracted text from images
Image–Text Sequence: Attribute-Based Analysis	Investigate whether the relationships or trends observed in the overall dataset hold for each of the attributes	Image First (IF) Text First (TF) Interleaved (IN)	Effects of question attributes: - Image type; - Prompt length; - Educational level; - Question type.
User Interaction and Heuristic Evaluations	Determine the usability and effectiveness of the heuristics	Image First (IF) Text First (TF) Interleaved (IN)	Ability to use the derived heuristics to guide optimal image sequencing to enhance reasoning performance

3.3.1. Image–Text Sequence Variation

This experiment investigated zero-shot multi-modal reasoning, where the model was tasked with predicting an answer a to a prompt that included a textual query q and an image x, without having been exposed to similar tasks during training. The model was required to analyse both the visual content in x and the information in q, integrating these inputs to generate a correct response. The experiment was specifically designed to evaluate

how the sequence and integration of textual and visual inputs, as structured within the API calls, affect the model's reasoning capabilities.

Each of the three models' APIs encodes information in a similar manner, whereby a set of parameters along with a prompt is sent to the model, as depicted in Figure 5. The prompt was composed of information from different *roles*, which defined the context and purpose of each part of the message. For this experiment, the prompt consisted of messages from two key roles: *system* and *user*. The *system* message sets the overall tone and controls how the model should respond. In this experiment, we used a fixed template for the system message: *"You are an expert in {subject}, helping a student answer an exam question"*. This message remained constant across all configurations, ensuring a consistent context for the model's responses. The second role in our prompt was the *user* role, which represents the input or question provided to the model. The user role contained blocks of content that can include either images or text. Since our experiments tested how the order of these content blocks (text and images) affects the model's performance, we varied the sequence in which the content blocks were presented to the LLM. We tested three configurations, i.e., *Image First*, *Text First*, and *Interleaved*, to determine their impact on the model's performance. The response a generated by the model under each configuration is defined as follows:

- Image First (IF): The model processes the image x before the text q, represented by the function f_{IF}.

$$a_{\text{IF}} = f_{\text{IF}}(x, q)$$

- Text First (TF): The model processes the text q before the image x, represented by the function f_{TF}.

$$a_{\text{TF}} = f_{\text{TF}}(q, x)$$

- Interleaved (IN): The model processes blocks of text $(q_1, q_2, \ldots, q_n)$ interspersed with the image x, integrating these inputs in sequence, represented by the function f_{IN}.

$$a_{\text{IN}} = f_{\text{IN}}(q_1, x, q_2, \ldots, q_n)$$

As the M3COTS dataset had no reference of the location of images within its instructions, the Interleaved (IN) configuration placed the image between the question and the options.

The above experiments were translated into API calls in the formats depicted in Figure 5 and comprised four components and steps. In step 1, the LLM is invoked to assume a subject expert persona for each respective field associated with a given question. This is then followed by step 2, which varies the sequencing of the image and textual components of the questions. In step 3, the LLM is given a standard CoTs instruction to *"Think step by step to answer the question, ..."* across all configurations.

We hypothesised that the order in which images and text are presented plays a crucial role in influencing an LLM's ability to correctly respond to multiple-choice questions. Specifically, it was hypothesised that for the M3Exam dataset, where images are interleaved with text, the f_{IN} configuration would yield optimal results. In contrast, for the M3COTS dataset, where images are generally presented prior to the question text, the f_{IF} configuration was expected to deliver the highest accuracy.

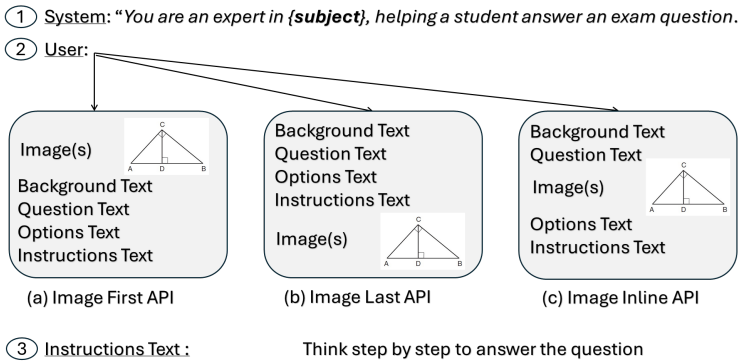


Figure 5. Example of the structure of the API calls containing the prompts for different experimental configurations.

3.3.2. Image–Text Sequence: Image Versus Instructions Analysis

To determine whether the observed sequencing effects were primarily driven by *the visual modality itself* or by *the sequencing of instructional information*, we conducted experiments on a selected subset of question types from the M3COTS dataset. The aim was to isolate the effect of sequencing the instructional content—regardless of whether that content originated from an image or a text prompt. To achieve this, we employed OCR techniques to extract text from images containing embedded text or formulas. The extracted text (denoted by $x_{\text{TextExtracted}}$) was then used to generate prompts. These prompts maintained the same instruction order as the multi-modal prompts but involved only the text modality. This approach allowed us to control for the order of instructional information independent of modality, enabling us to determine whether sequencing effects were primarily driven by content order rather than modality differences (Figure 6). OCR extraction was performed by using GPT-4o and Gemini-1.5. The extracted text was manually compared with the original images and corrected for any conversion errors, minimising inaccuracies inherent to OCR technology.

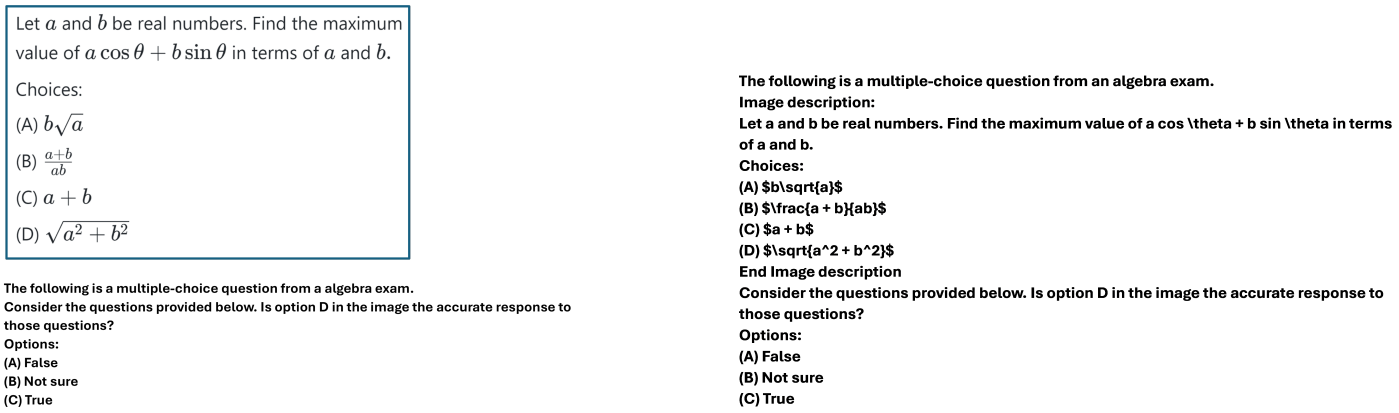


Figure 6. Example of an original image-based question (**a**) converted into a purely text-based question (**b**).

The specific configurations being tested are as follows:

- Image First (IF): The model processes the extracted text from the image $x_{\text{TextExtracted}}$ before the textual query q. This is represented by the function f_{IF}.

$$a_{\text{IF}} = f_{\text{IF}}(x_{\text{TextExtracted}}, q)$$

- Text First (TF): The model processes the textual query q before the extracted text from the image $x_{\text{TextExtracted}}$, represented by the function f_{TF}.

$$a_{\text{TF}} = f_{\text{TF}}(q, x_{\text{TextExtracted}})$$

- Interleaved (IN): The model processes blocks of text $(q_1, q_2, \ldots, q_n)$ interspersed with the extracted text from the image $x_{\text{TextExtracted}}$, integrating these inputs in sequence, represented by the function f_{IN}.

$$a_{\text{IN}} = f_{\text{IN}}(q_1, x_{\text{TextExtracted}}, q_2, \ldots, q_n)$$

3.3.3. Image–Text Sequence: Attribute-Based Analysis

In these experiments, we analysed how varying attributes within the dataset—such as image type (image, text, or a mixture of both), prompt length, educational level, and question type—affect the model's performance and sensitivity to sequencing. The goal was to examine whether the trends observed in the overall dataset hold for each of the attributes:

- Image type: The model's performance is evaluated based on different types of images—purely visual (images with no text), text-based (images primarily containing text, like formulas), and mixed (images containing both visual elements and text) images.
- Prompt length: Various lengths of prompts are tested to observe how the length of the text portion affects the model's accuracy and reasoning capabilities.
- Educational level and question type: The experiment evaluates how different educational levels with the M3Exam dataset and question types within the M3COTS dataset —representing the topics or reasoning skills being tested (for example, event ordering or economics per capita GDP comparison)—influence the model's performance.

3.3.4. User Interaction and Heuristic Evaluation

To evaluate how heuristics can assist users in effectively sequencing images and text within prompts to achieve optimal model performance, we conducted an experiment focusing on heuristic-based user guidance. These heuristics, developed based on patterns identified from our previous analysis, aim to enhance the ability of Large Language Models (LLMs) to comprehend and accurately respond to various types of tasks by optimising image and text sequencing. GPT-4o was used in this experiment, utilising the OpenAI API Playground portal, which allowed for the fine-grained manual control of text and image sequencing within prompts. To provide an initial validation of these heuristics, the experiment was conducted by a human subject, who was one of the co-authors familiar with the heuristic guidelines, to ensure consistent and accurate application across all samples.

The experiment involved manually inputting each question according to the developed heuristics and comparing the resulting model accuracy with the results obtained from the Image First (IF), Text First (TF), and Interleaved (IN) sequencing strategies applied to the same 103 questions. This approach aimed to determine if using the heuristics led to improved reasoning accuracy compared with standard sequencing strategies.

3.4. Evaluation

Our experimental evaluations were mainly performed by using a mix of comparing the percentage of correct responses, conducting mean rank analyses, and performing tests for statistical significance. For the statistical evaluation of binary outcomes per response (i.e., correct/incorrect), McNemar's test was used, as it is specifically designed for binary outcomes and thus provides an effective way to compare relative performance under different conditions for the same questions. Mean ranks were employed to offer a more comprehensive and insightful understanding of the impact of image and text

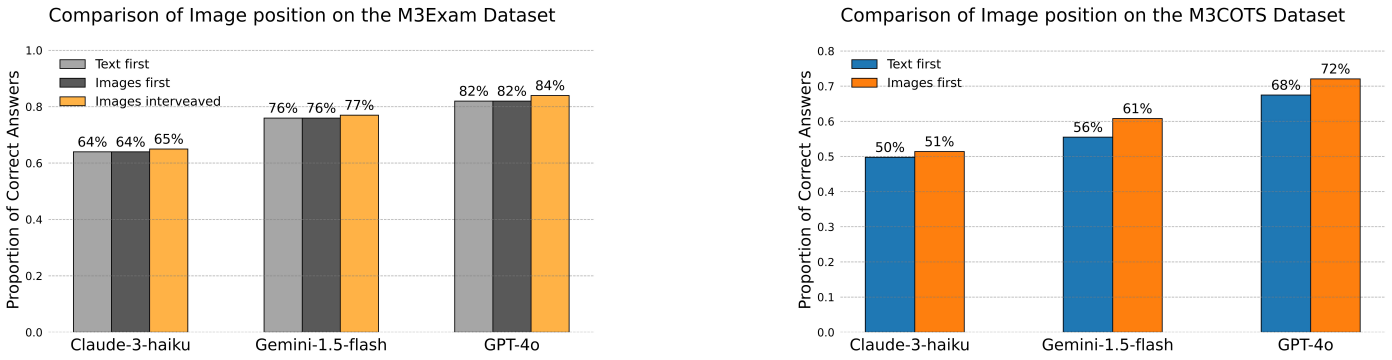
sequencing configurations. For each question type and configuration, ranks were assigned based on the accuracy performance of the LLMs, with a lower rank indicating better performance. These ranks were then averaged across different subcategories within each dataset, such as subject domains and question types. Analysing the *mean ranks* subsequently helped to identify more generally what the most optimal configurations tended to be by consolidating performances over all configurations. Mean ranks, therefore, provided another concise perspective alongside that of accuracy comparisons. The statistical tests provided insights but were considered merely some of several indicators rather than the sole arbiter of significance.

4. Results

This section first examines the results of Image–Text Sequence Variations and the impact of the characteristics of questions and instruction sequencing on accuracy for the purpose of being able to devise heuristics that users can follow in order to optimise the accuracy of responses. The heuristics are then presented together with the findings of the user interaction study where a human subject is asked to apply the heuristics on a dataset to validate both the efficacy of the empirical findings and the utility of the heuristics.

4.1. Image–Text Sequence Variation

Figure 7 shows the accuracy of the three chosen LLMs on both datasets with respect to the different placements of the images and text in the prompt sequences. At a high level, it can be seen that generally, the LLMs tend to score higher on M3Exam than on M3COTS, which is in line with results in the literature, which has reported 71.8% (Zhang et al. [3]) and 62.6% (Chen et al. [4]) when using the older GPT-4 with CoT. GPT-4o also consistently outperformed Claude-3 and Gemini-1.5 on both datasets by a significant margin, while Claude-3 demonstrated the lowest overall performance on both datasets. Across both figures, it can also be seen that generally, placing images within the text (IN) on both the M3Exam and M3COTS datasets consistently yielded marginally higher accuracy over other placements.



(**a**) M3Exam dataset (**b**) M3COTS dataset

Figure 7. Comparison of image and text placement positions on the M3Exam and M3COTS datasets.

Table 2 details a deeper performance profile of each sequencing configuration with respect to the different subject areas of the M3Exam dataset and the various characteristics through which questions from each discipline could influence accuracy when combined with different image placements. The summary of the table in the form of mean ranks consistently indicates that on average, placing images within the text (IN) yielded the best results while showing little difference between the IF and TF placements for all LLMs.

Table 2. A comparison of image positions on the M3Exam data with the image–text configurations of Text First (TF), Image First (IF), and Interleaved (IN).

	Claude-3			Gemini-1.5			GPT-4o		
Subject	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**
English	0.81	**0.90**	**0.90**	0.94	**1.00**	0.94	0.97	**1.00**	**1.00**
Algebra1	0.58	**0.42**	**0.42**	0.42	0.58	**0.63**	0.58	**0.68**	**0.68**
Algebra2	0.19	**0.50**	0.38	0.56	0.50	**0.63**	**0.63**	0.56	**0.63**
Geometry	**0.31**	0.29	**0.31**	**0.49**	0.47	0.47	**0.63**	0.59	0.59
Math	0.39	0.37	**0.42**	0.59	0.58	**0.60**	0.64	**0.70**	0.69
Chemistry	0.67	0.60	**0.73**	0.60	**0.80**	0.73	**0.87**	0.80	0.80
Environment	0.79	**0.82**	0.81	0.92	0.92	**0.93**	**0.98**	0.96	0.94
Physics	**0.43**	0.37	0.36	0.63	0.63	**0.71**	0.77	0.79	**0.85**
Science	0.79	0.79	**0.81**	0.88	0.86	**0.89**	0.88	0.89	**0.90**
Earth	0.61	0.61	**0.62**	**0.70**	0.69	0.68	0.75	0.73	**0.80**
History	0.94	**0.96**	**0.96**	**1.00**	0.98	0.96	**1.00**	**1.00**	**1.00**
Social	**0.87**	0.81	0.84	0.90	0.93	**0.94**	0.94	**0.95**	**0.95**
Mean Rank	2.1	2.2	**1.7**	2.2	2.2	**1.7**	2.2	2.0	**1.8**

Meanwhile, a granular investigation into the effects of image placement in the M3COTS data was also performed at a subject level to complement the results from Figure 7, which show some different patterns to those on the M3Exam dataset. Shown in Table 3, the IF approach yielded the highest accuracy on average across Claude-3 and Gemini-1.5, while IN was the most effective on average for GPT-4o. For all models, the TF approach tended to be on average the least effective. The pairwise comparison of mean ranks indicated statistical difference in the case between the TF and IN sequences on Gemini-1.5 (Wilcoxon Test Statistic = 0, $p = 0.003$) and GPT-4o (Wilcoxon Test Statistic = 0, $p = 0.004$), as well as between the TF and IF sequences on Gemini-1.5 (Wilcoxon Test Statistic = 2.0, $p = 0.012$).

While Figure 7 visually suggests marginal improvements with interleaved images (IN), statistical significance tests were performed to formally assess these differences. Although pairwise comparisons of mean ranks indicated statistical significance in some specific cases for M3COTS, the overall McNemar's and Wilcoxon tests did not show statistically significant differences across all configurations for the dataset as a whole. This suggests that while trends are observable, further research with larger datasets may be needed to establish strong statistical significance across all task types (while pairwise comparisons were conducted, we acknowledge that multiple comparisons can inflate Type I error; future analyses could incorporate corrections for multiple comparisons (e.g., Bonferroni correction) for more stringent significance testing, although the primary focus here is on observed trends and effect sizes).

Table 3. A comparison of image positions on the M3COTS data with the image–text configurations of Text First (TF), Image First (IF), and Interleaved (IN).

	Claude-3			Gemini-1.5			GPT-4o		
Subject	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**
language-science	**0.79**	0.73	**0.79**	0.88	0.84	**0.90**	0.95	0.94	**0.96**
natural-science	**0.53**	**0.53**	**0.53**	0.59	**0.64**	0.63	0.70	**0.78**	0.76
social-science	0.35	0.32	**0.36**	0.39	**0.45**	0.44	0.55	0.59	**0.60**
physical-commonsense	0.60	0.82	**0.86**	0.77	**0.88**	0.83	0.86	0.84	**0.88**
social-commonsense	0.63	**0.70**	0.69	0.68	**0.74**	0.73	0.76	**0.80**	0.79

Table 3. *Cont.*

Subject	Claude-3			Gemini-1.5			GPT-4o		
	TF	IF	IN	TF	IF	IN	TF	IF	IN
temporal-commonsense	0.75	**0.80**	0.78	0.75	**0.87**	0.84	0.89	0.86	**0.90**
algebra	0.21	**0.31**	0.24	0.28	0.35	**0.38**	0.44	0.57	**0.67**
geometry	0.24	**0.36**	0.31	0.36	0.39	**0.40**	0.34	0.33	**0.41**
theory	0.33	**0.38**	0.24	0.24	**0.43**	0.38	0.29	0.48	**0.52**
Mean Rank	2.5	**1.7**	1.8	2.9	**1.4**	1.7	2.6	2.2	**1.2**

4.2. Image–Text Sequence: Image Versus Instructions Analysis

For these experiments, we utilised a dataset comprising questions presented either solely as text or having formulas embedded within images, as shown in Figure 6a,b. For these experiments, we used the most suitable questions found in the *"Elementary Algebra"* (363 questions) and *"Grammar"* (205 questions) subsets from the M3COTS dataset. The primary objective was to investigate whether the sequencing impact is due to the visual modality itself or the sequencing of instructional information. To isolate the effect of sequencing, we extracted text from images during preprocessing, creating text-only versions of the prompts. These prompts followed the same instruction order as multi-modal prompts but used only text, isolating sequencing effects from modality differences to assess the impact of the content order. Table 4 presents the performance of three multi-modal LLMs—Claude-3, Gemini-1.5, and GPT-4o—under different sequencing conditions.

Table 4. Image–Text Sequence: Image Versus Instructions Analysis results with the configurations of Text First (TF) and Image First (IF).

Question Type	Claude-3			Gemini-1.5			GPT-4o		
	TF	IF	IN	TF	IF	IN	TF	IF	IN
Elementary Algebra (multi-modal)	0.32	**0.39**	0.25	0.27	**0.38**	0.36	0.45	0.55	**0.64**
Elementary Algebra (text)	0.27	**0.35**	0.34	0.25	**0.43**	0.41	0.35	0.64	**0.66**
Grammar (multi-modal)	0.80	0.82	**0.85**	**0.94**	0.79	0.92	0.95	0.96	**0.97**
Grammar (text)	**0.91**	0.88	0.87	0.92	0.85	**0.94**	0.96	**0.98**	**0.98**

The experimental results in Table 4 reveal that the sequencing of images and text within prompts significantly influences the reasoning performance of multi-modal LLMs in this study, with effects varying by task and model. In the case of *Elementary Algebra* questions, both Gemini-1.5 and Claude-3 demonstrated markedly higher accuracy with Image First (IF), while GPT-4o showed higher accuracy with Interleaved (IN). In contrast, for *Grammar* questions, Gemini-1.5 and Claude-3 achieved better performance when textual instructions preceded images (TF). Additionally, the text-only versions of the prompts mirrored these patterns, underscoring that the order of instructional information alone, independent of image placement, plays a role in model performance. These findings are, therefore, suggestive of the role and importance of tailoring prompt structures to both the nature of the task and the specific model in use.

4.3. Image–Text Sequence: Attribute-Based Analysis

Here, exam question attributes were analysed for their impact on image sequencing to evaluate whether the trends observed in the overall dataset accuracy rates presented earlier indeed hold for each of the attributes. For the M3Exam dataset, the attributes educational level, prompt length, and image type were examined. Meanwhile, for M3COTS, the question type, prompt length, and image type attributes of the questions were evaluated. In the case of M3Exam data, the models' performance did not show any deviations from

the results in the previous section (the details of this can be seen in Appendix A). However, in the case of specific question types for the M3COTS dataset, different sequencing patterns led to significantly better performance, which was contrary to the overall results in the previous section.

Table 5 shows examples of M3COTS question types where the optimal image sequencing diverged from the results for the overall dataset; further examples can be found in Appendix A. For instance, performance on the *"Economics-Per Capita Wage Calculation"* question type (example provided in Figure 8) showed significant improvement when the text was placed first (TF) for GPT-4o compared with both the image being placed first (IF) and interleaved (IN). Similar performance was found for the *"Physics - Velocity, Acceleration, and Forces"* question type, which showed significant improvement with TF for Claude-3 compared with both IF and IN. For Claude-3, the question types where TF was the optimal sequencing tended to have longer token lengths for the text portion of the prompt. When analysing the top 10% longest test prompts, Claude achieved accuracy rates of 36% for TF, compared with 24% for IF and 31% for IN. In contrast, question types that performed the worse with TF often involved a nested multiple-choice format, where one question referenced another. For instance, this type of complex referencing within a question can be seen in *"Chemistry-Atoms and Molecules Recognise"* questions with the example provided in Figure 9. Here, GPT-4o's accuracy dropped acutely, from 72% (IN) and 67% (IF) to 32% when the text was placed first (TF). When the image was placed after the text the model correctly interpreted the image, as it was more likely to select the option shown in the image rather than the one stated in the original text. The performance drop could also be seen in other question types which had this type of complex referencing within a question. While the question types followed a similar pattern, these results suggest that the sequencing of content can play a critical role in questions and tasks involving complex references, with the impact of image sequencing varying by model and context, indicating that patterns can be identified and heuristics devised based on this to more optimally match the image sequencing to specific question types.

Table 5. Example question types showing different optimal image–text sequencing patterns: Text First (TF), Image First (If), and Interleaved (IN).

	Claude-3			Gemini-1.5			GPT-4o		
Subject	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**
Chemistry-Atoms and Molecules Recognize	0.27	0.28	**0.33**	0.25	**0.43**	**0.43**	0.32	0.67	**0.72**
Physics-Velocity, acceleration, and forces	**0.48**	0.18	0.26	0.60	**0.64**	0.48	0.86	**0.88**	**0.88**
Economics-Per Capita Wage Calculation	**0.40**	0.28	0.35	0.30	**0.40**	0.35	**0.73**	0.58	0.48

Question:
What are the per capita GDP of Country 4 and Country 2 in 2018?
Options:
(A) $14,054.21, $15,566.04
(B) $15,566.04, $14,106.49
(C) $15,566.04, $14,054.21
(D) $14,106.49, $15,566.04

Year	Country	GDP($)	Population (hundred millions)	Export Value ($)	Inflation Rate (%)
2018	Country 2	3.3	2.12	2.13	1.63
2018	Country 4	23.05	16.34	1.76	3.97
2019	Country 4	10.69	17.19	3.83	3.95
2021	Country 2	5.92	12.68	4.27	1.38
2021	Country 4	34.28	13.57	1.18	2.07
2022	Country 4	24.89	17.71	2.41	2.57

Figure 8. *"Economics"* question where improved performance was achieved with Text First (TF) sequencing.

Question:
Find the correct molecular name based on the legend.

(A) bromomethane (B) bromine (C) dichloromethane

Options:
(A) All of the answer choices are wrong.
(B) Option B in the image
(C) Option A in the image
(D) Option C in the image

In the image **(A) Bromomethane** is correct
Answer: (A) ***Incorrect this should be option (C)***

Figure 9. *"Chemistry-Atoms and Molecules Recognise"* question where improved performance was achieved with Image First (IF) and Interleaved (IN) sequencing.

4.4. Heuristic Development and User Interaction Study

In this section, we identify additional modality sequencing patterns within the exam questions and correlate these with the accuracy rates attained with the IF, TF, and IN configurations to assist in determining the development of heuristics to guide human interaction.

4.4.1. Correlating Exam Question Structure with Sequencing

Building on the findings presented in the previous section, we first examined the question types where the optimal image sequencing deviated from the overall dataset trends (as seen in Table 5 and in further examples in Appendix A). This initial analysis served as a guide for identifying question types with distinct sequencing effects. Expanding beyond specific question types, we sought to identify broader patterns across questions that influenced sequencing effectiveness. One such pattern emerged when the text explicitly referenced the image's position within the prompt (e.g., *"…the image above…"* or *"…the following image …"*). In these cases, Image First (If) and Interleaved (IN) sequences often provided better alignment with the intended question structure. However, we also recognised that in a substantial number of cases, image placement had little impact, particularly between IF and IN sequencing. Despite this, we opted to include these instances to ensure a comprehensive analysis, as questions with similar patterns still exhibited sequencing-sensitive patterns.

First, we cross-correlated each of the TF, IF, and IN sequences with the corresponding structure of the original exam questions. This was performed to assess whether aligning the modality sequence used in prompts with the original exam question format produced variations in accuracy. We highlight four general exam question formats. Beginning with the more obvious patterns, questions with the *Image First pattern* lead with the visual cue and require the direct processing of visual content without prior textual context or may begin with text which directs attention to the position of the image, e.g., *"…the image above…"*. Examples of the Image First pattern are seen in Figure 4a,b. In contrast, questions with the *Text First pattern* are characterised by preceding text which can either comprise a substantial context or the explicit body of the question itself, as seen in the example in

Figure 9. While the first two patterns are more obvious, we found that the questions with an interleaved pattern possessed an additional discriminating structure. We term the first sub-type the *interleaved instructions (IN1) pattern*, where the images are interleaved within the text and are placed at a point referred to in the prompt (i.e., *"...the image below..."*). In this pattern, the text refers to the image being placed at a specific point within the question, with an example provided in Figures 1 and 3. The second sub-type is the *interleaved logic (IN2) pattern*, where the images are interleaved within the text and placed where the information in the image most optimally fits into the logical flow of the question itself, as exemplified in Figure 8.

With these structural patterns defined, we then correlated them with the accuracy results from our experimental sequencing configurations (IF, TF, and IN). The observed relationships, presented in Tables 6 and 7, suggest that heuristic-based sequencing adjustments can significantly improve performance for specific question types and question structures. These results reinforce the notion that multi-modal prompt sequencing is not a one-size-fits-all approach but rather a context-dependent optimisation problem that benefits from targeted heuristics.

Table 6. A comparison of exam question patterns with the image–text prompt configurations of Text First (TF), Image First (IF), and Interleaved (IN) in M3Exam, with the number of samples in parentheses.

	Claude-3			Gemini-1.5			GPT-4o		
Exam Question Pattern	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**
Image First (104)	**0.89**	**0.89**	**0.89**	0.92	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**
Interleaved instructions IN1 (517)	0.62	0.62	**0.64**	0.73	0.73	**0.75**	0.80	0.81	**0.83**
Interleaved logic IN2 (135)	0.54	**0.59**	0.57	0.75	0.75	**0.76**	0.82	0.81	**0.84**
Text First (39)	**0.56**	0.36	0.41	**0.64**	0.62	0.62	**0.72**	0.54	0.59

Table 7. A comparison of exam question patterns with the image–text prompt configurations of Text First (TF), Image First (IF), and Interleaved (IN) in M3COTS, with the number of samples in parentheses.

	Claude-3			Gemini-1.5			GPT-4o		
Exam Question Pattern	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**	**TF**	**IF**	**IN**
Image First (415)	0.55	0.56	**0.59**	0.62	**0.68**	0.67	0.71	**0.75**	**0.75**
Interleaved instructions IN1 (635)	0.72	0.82	**0.83**	0.85	0.88	**0.90**	0.92	0.92	**0.96**
Interleaved logic IN2 (251)	0.40	**0.46**	0.43	0.43	0.49	**0.51**	0.58	0.69	**0.74**
Text First (1017)	**0.39**	0.31	0.33	0.42	**0.45**	0.42	0.58	**0.59**	0.55

In both Tables 6 and 7 and across all three models, it can be seen that when the structure of the questions is that of Text First, then the highest accuracy rates are also achieved when the prompts mirror the sequencing with the TF format. Generally, a similar picture emerges with the questions structured in the Image First format when matched with the IF prompt structure; however, it is not as consistent, and the IN prompt format frequently produces equivalent performance to IF. Both the IN1 and IN2 question sub-types also correlate strongly with the IN modality sequencing in the prompts, with Claude-3 displaying a preference for IF sequencing when the questions are structured in the IN2 format, indicating that it potentially has a different approach to logical flow and reasoning compared with the other models.

4.4.2. Heuristic Definition

The above results and the mapping of exam question patterns to modality sequencing then paved the way to the derivation of the heuristics to guide users towards effective

multi-modal prompt design. Table 8 describes the formalisation of the heuristics, together with "how" and "when" to apply them.

Table 8. Heuristics for effective multi-modal prompt design.

Heuristics	How to Apply	When to Apply
Image First Pattern	Structure the prompt by placing all images at the beginning, preceding any textual instructions.	Apply when the task requires immediate analysis or interpretation of visual content without relying on preceding textual information, or when instructions explicitly reference the image at the beginning of the prompt (e.g., *"...the image above..."*). This heuristic is particularly effective for Gemini-1.5 and GPT-4o; occasionally, for these models and especially Claude-3, following an IN1 or IN2 structure would at least yield comparable results.
Reference-Based Image Positioning (IN1)	Integrate images within the text at specific points where they are referenced. For instance, if the text states *"...the image below..."*, place the corresponding image immediately after that reference in the prompt.	Use when textual instructions specifically indicate the placement of images within the prompt or when images are intended to clarify specific parts of the instructions or context. This approach is particularly beneficial for Claude-3, Gemini-1.5, and GPT-4o, where this interleaved sequencing improves performance in specific question types.
Logical Flow Alignment (IN2)	Position images in the prompt to align with the logical steps or components of the reasoning process. Ensure each image corresponds to relevant sections of the text to support multi-step reasoning.	Implement for tasks involving multiple reasoning steps or layered information, where images correspond to specific stages or components of the reasoning process. This heuristic is especially effective for Gemini-1.5 and GPT-4o in complex reasoning tasks, such as nested multiple-choice questions, where interleaved sequencing (IN2) leads to improved performance. However, in the case of Claude-3, some preference is exhibited towards the IF prompt structure instead.
Text First Pattern	Organise the prompt by placing all textual instructions and data first, followed by images. Ensure that the text provides comprehensive context and details before introducing visual elements.	Apply to prompts with extensive textual content, detailed explanations, or substantial numerical data, where establishing textual background is foundational before introducing visual elements. This heuristic is consistently effective for Claude-3, Gemini-1.5, and GPT-4o.

4.4.3. User Interaction Study Results

Following the development of the multi-modal prompting heuristics, we proceeded to conduct a user interaction study to evaluate their effectiveness. The user interaction study included a random sample of 103 questions from the M3COTS dataset, where each sample had to satisfy the condition of having been correctly answered by the GPT-4o model in either the IF or TF prompting configurations so as to more clearly highlight the effects of applying the heuristics. A balanced distribution of randomly selected samples from TF, IF, and IN were chosen. OpenAI Playground was used for the user interaction experiments since this interface enables the fine-grained control of the sequencing of the modalities (seen in Figure 10), unlike the standard OpenAI interface.

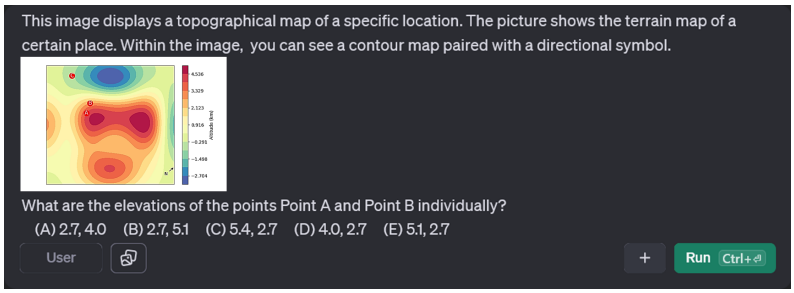


Figure 10. Example of image placement in OpenAI Playground for the user interaction study, illustrating the application of the Logical Flow Alignment (IN2) heuristic. The exam question is from M3COTS geography-3769.

We first provide an example demonstrating how the Logical Flow Alignment (IN2) heuristic was applied to a sample question within the OpenAI Playground testing environment. The target exam question (M3COTS question geography-3769) is shown below in the original JSON format. The multi-modal task begins with the question component, followed by choice options and then the context, with the image trailing last.


```
{
    "id": "geography-3769",
    "category": "geography-Altitude Estimation",
    "question": "What are the elevations of the points Point A and Point B individually?",
    "choices": [
        "2.7, 4.0",
        "2.7, 5.1",
        "5.4, 2.7",
        "4.0, 2.7",
        "5.1, 2.7"
    ],
    "context": "This image displays a topographical map of a specific location.
    The picture shows the terrain map of a certain place. Within~the image,
    you can see a contour map paired with a directional symbol.",
    "image": "data\\images\\geography-3769.png",
    "domain": "science",
    "topic": "social-science"
}
```


The user then proceeds to decode the multi-modal task requirements based on the derived heuristics and determines that the Logical Flow Alignment (IN2) heuristic is the most suitable for this scenario, which emphasises the placement of visual elements within a prompt to align with the logical steps or components of the reasoning process, thereby facilitating multi-step reasoning more effectively. The translation of the exam question from the JSON format by the user into the IN2 sequence pattern can be seen in Figure 10, where the optimal logical flow and image placement have been rearranged into the order context, image, question, and lastly, choice options.

The results of the user interaction study can be seen in Table 9. The results indicate that leveraging multi-modal prompting heuristics for modality sequencing enhances the accuracy in this dataset compared with the IF, TF, and IN configurations. Specifically, when the human user employed custom strategies informed by the derived heuristics, GPT-4o achieved a correct response rate of 62.1% compared with TF (40.8%), IF (58.3%), and IN (41.8%). This improvement highlights the value of the developed heuristics in aligning the model's interpretive sequence with the needs of the task. It shows that strategically sequencing inputs can improve reasoning accuracy more effectively than naive configurations. As a by-product, these results also confirm the reliability of the experiments conducted in the initial phase, which enabled the extraction of the heuristics.

Table 9. Results contrasting user interaction accuracy when following the devised heuristics with the Text First (TF), Image First (IF), and Interleaved (IN) configurations on a subset of M3COTS data.

	GPT-4o			
	TF	**IF**	**IN**	**User**
Accuracy	40.8%	58.3%	41.8%	**62.1%**

5. Discussion

Our study possessed a dual focus. We first sought to fill a research gap in understanding how multi-modal LLMs as our target intelligent technology respond to prompts comprising text and image inputs when they are sequenced in a variety of different ways. The goal was to understand how the sequencing of the modalities can be optimised in order to ensure that the most accurate results are extracted from an LLM's complex reasoning capabilities. Second, the goal of this study was to derive from the empirical results a set of practical and user-oriented multi-modal heuristics that would serve as guidelines for effectively structuring prompts. The derived heuristics were then evaluated through a user interaction study. Our work builds upon and significantly extends previous investigations focused on unimodal (text-only) reasoning which examined how altering the position of words [18,19] or the instruction order in text prompts [24]. Our work has extended the investigation into the multi-modal domain and has both produced practical guidelines and a validating user interaction study.

5.1. Effects of Text and Image Sequencing

In addressing **RQ1**, we investigated how the sequencing of image and text modalities in prompts affects the reasoning performance of multi-modal LLMs across different datasets and question types. We initially hypothesised that the order in which these modalities are presented would significantly influence the models' reasoning capabilities. Our experimental results confirmed this hypothesis, revealing that the optimal sequencing strategy is not universal but varies depending on the dataset and the nature of the tasks.

For the M3Exam dataset, which features a variety of formats with multiple visual elements for exam questions, interleaving images within the textual content generally yielded the highest reasoning performance across all three LLMs studied—GPT-4o, Gemini-1.5, and Claude-3. A significant proportion of exam questions in the dataset were structured in an interleaving manner, suggesting that by also aligning the prompt structure with the original format of the exam questions, the models are enabled to process and reason over the information more effectively. Conversely, for the M3COTS dataset, which contains complex reasoning tasks due to multiple images, together with text being frequently embedded within figures, placing the image first in the prompt on average led to superior performance, particularly for Gemini-1.5 and Claude-3. GPT-4o, however, also showed the best performance with the Interleaved configuration on this dataset. These findings again highlight that both the dataset structure and the complexity of the questions influence how modality sequencing affects reasoning performance in LLMs, and they also add a measure of nuance to the general guidance by commercial vendors [16,55,56] which advise that by placing an image first provides a visual context that aids the reasoning process.

The variations observed suggest that the underlying multi-modal fusion strategies of each LLM, which are often proprietary and not fully disclosed, play a role in how modality sequencing impacts performance. Attention mechanisms and positional encoding in transformer architectures likely contribute to modality bias, affecting how models prioritise information based on its position in the prompt, and future theoretical work is needed to fully elucidate these mechanisms. Awareness of these factors is, therefore, useful

for optimising prompt design for enhancing the reasoning capabilities of LLMs. While the observed variations highlight the complexity of LLM behaviour, these seemingly disparate results are synthesised into actionable heuristics (Table 8) for HCI practitioners to navigate this complexity in prompt design.

5.2. Question Complexity and Sequencing Sensitivity

With respect to **RQ2**, we examined how specific attributes of questions—such as nested structures, subject domains, and complexity—interact with modality sequencing to influence LLM performance across different models. Our analysis revealed that the sensitivity to modality sequencing is indeed affected by these attributes, particularly in complex reasoning tasks involving multiple steps or nested-question formats.

In the M3COTS dataset, question types involving nested multiple-choice formats, where one question references another, were significantly impacted by the sequencing of modalities. While the LLMs often successfully processed and reasoned on the visual content, they struggled with the final step of mapping their reasoning back to the correct option in the original question. This was especially the case when the sequencing did not align with the logical flow required by the task. For example, models could frequently interpret the image correctly but failed to revisit the earlier textual information to select the appropriate answer choice. This difficulty highlights challenges associated with multi-hop reasoning, where models need to maintain context over several reasoning steps and potentially revisit earlier information. The linear reasoning approach facilitated by Chain-of-Thought (CoT) prompting may not sufficiently support the backtracking required in such nested questions. Additionally, the transformer's positional encoding and attention mechanisms may lead to diminished attention to earlier information as the sequence lengthens, affecting the models' capacity to integrate information across modalities and over extended sequences.

Therefore, our experiments indicate that aligning the flow of information in the prompt with the logical steps of the reasoning task as much as possible is effective in optimising performance. This means that the sequencing of images and text should be tailored to match the cognitive demands of the specific task, rather than relying on a one-size-fits-all approach. These findings suggest that it is the strategic placement and sequencing of information—particularly how images are integrated within the textual content—that play a key role in reasoning performance rather than the inherent properties of the image modality itself. The sensitivity to modality sequencing, while initially appearing complex, informs the development of targeted heuristics (8) that allow users to adapt the prompt structure to task demands.

5.3. Information Order vs. Modality Properties

In exploring **RQ3**, we aimed to determine the extent to which the impact of modality sequencing on LLM performance is attributable to the order of information presentation rather than the inherent properties of the different modalities. Our experiments provided evidence that the sequence in which information is presented significantly influences LLM performance, often outweighing the intrinsic characteristics of the modalities themselves. By converting images containing textual information into pure text and evaluating the models on these single-modality prompts, we found that the sequencing of information—whether textual content precedes or follows other textual content extracted from images—consistently impacted accuracy. This underscores the importance of positional encoding and attention mechanisms in transformer architectures, which are sensitive to the order of input tokens. The models' performance was not solely dependent on processing visual content but also heavily influenced by how the information was sequenced in the prompt. While this study focused on text and image inputs, the broader concept

of sequencing effects is modality-agnostic in principle, and we anticipate that similar ordering sensitivity may emerge in multi-modal systems involving audio, video, or sensor data—where temporal or spatial dependencies can play a critical role in the reasoning outcomes. The sensitivity to information sequencing varied among the models studied. While Claude-3 showed minimal responsiveness to changes in sequencing, both Gemini-1.5 and GPT-4o exhibited more pronounced improvements when the information ordering was optimised. This variation suggests that differences in training data, model architecture, and fine-tuning methods can influence how LLMs process and prioritise information based on its position in the prompt. Our attribute-based analysis further indicated that specific question types could achieve performance gains of up to 5% by tailoring the sequencing strategy. This highlights the necessity of strategic information ordering in prompt design. Effective prompt engineering, aligned with the task requirements and the characteristics of the model, is useful for optimising the reasoning capabilities of multi-modal LLMs. These findings, therefore, have practical implications for users interacting with intelligent systems, emphasising that careful consideration of information sequencing can enhance the utility of LLMs across diverse applications.

5.4. Implications and Practical Guidelines

Our development of practical heuristics and guidelines to assist users in constructing multi-modal prompts that seek to optimise reasoning accuracy and enhance the user experience with interactive intelligent systems answers **RQ4**. While primarily validated on exam-like tasks, the underlying principles of modality sequencing and information flow suggest potential relevance to a broader range of AI applications involving multi-modal interactions. The analysis of the experimental results provides a preliminary basis for deriving a set of practical heuristics in the form of multi-modal prompting guidelines. The practical implications of our findings are particularly relevant for the following application domains:

1. **Educational tutoring systems:** In systems designed to assist with learning, integrating textual explanations with diagrams or problem images can enhance student understanding.
2. **Healthcare diagnostics:** Diagnostic tools that combine patient records with medical imaging (e.g., X-rays and MRI) can benefit from optimised modality sequencing.
3. **Legal document analysis:** Legal professionals often need to analyse case documents that include both textual information and visual evidence (e.g., diagrams and charts), enabling the optimisation of the sequencing of these modalities to facilitate more accurate and efficient case analyses.
4. **Customer support systems:** In customer support applications, combining textual queries with product images can improve the accuracy of automated responses.

By applying the developed guidelines from this study, users now have access to empirically derived recommendations that can enhance the effectiveness of multi-modal LLMs, leading to more accurate and reliable outcomes. Therefore, these recommendations contribute to a better understanding of the interplay between intelligent technology and user interaction, aligning with the goals of advancing both technological capabilities and user experience in the field of human–computer interaction.

5.5. Advancing Human-Centred AI in Multi-Modal Interaction

Our findings ultimately reinforce the role of user-guided prompt sequencing as a key mechanism for enhancing reasoning in multi-modal LLMs, aligning with prior work on uncertainty as a design resource, adaptability in AI systems, and interactive refinement [68–73]. Prior studies highlight that AI effectiveness depends not only on accuracy but also on flexibility, configurability, and integration into real-world workflows.

Our work builds on these insights by demonstrating that prompt sequencing serves as a dynamic tool for structuring model attention, mitigating reasoning failures, and aligning multi-modal input with task demands. We further extend research on iterative AI refinement by showing how structured user input—whether in positioning text before images, interleaving modalities, or aligning information with logical reasoning flows—can systematically improve LLM reasoning. This echoes broader HCI principles of human-in-the-loop adaptation, where interaction strategies actively shape computational outcomes. Our study contributes to this space by offering empirically validated, configurable heuristics for end-user prompt optimisation, bridging the gap between intelligent system design and human interaction. By framing prompt sequencing as an interactive design element, our findings emphasise that multi-modal AI reasoning is not purely a model-driven process but an adaptable, user-mediated workflow. This perspective strengthens the connection among human-centred AI, interaction design, and the evolving challenges of reasoning with multi-modal LLMs.

5.6. Limitations

This study has several limitations. Firstly, the experiments were conducted by using specific examination-based Visual Question Answering (VQA) datasets and evaluated on three commercial LLMs. Although these datasets and models are suitable for complex reasoning tasks, they may introduce biases due to their educational context and model-specific characteristics, potentially limiting generalisability to other domains or LLMs. Future research should expand the dataset scope to further validate these patterns. Secondly, the exclusive focus on English-language content may reduce applicability to multilingual or culturally diverse settings. Additionally, the measurement of reasoning capabilities was confined to quantifying accuracy, without delving into qualitative assessments of the reasoning processes employed by the models [80]. This approach may overlook subtleties in how models integrate and process multi-modal information. The user interaction study, while demonstrating the *potential* of the heuristics, was conducted with a single, expert user. Future research must include larger-scale user studies with diverse participants to rigorously validate the usability and effectiveness of these heuristics for a broader HCI practitioner audience.

5.7. Future Research

Future work should incorporate diverse datasets, including non-English data. Expanding the range of the LLM architectures and training paradigms evaluated will also help uncover model-specific sensitivity to modality sequencing. Another key objective is to explore how incorporating additional modalities, such as audio and video, can improve reasoning capabilities in complex multi-modal interactions. Research should also focus on developing adaptive sequencing methods that dynamically adjust the modality order based on task-specific needs or user input. Further studies should analyse the effects of prompt length, visual content complexity, and the use of multiple images embedded within a single visualisation to further refine prompt engineering practices. There are also research opportunities in developing graphical user interfaces which allow for the fine-grained sequencing of modalities, which is currently lacking in large commercial multi-modal LLMs. Finally, large-scale user interaction studies with diverse participants and tasks are necessary to validate and enhance the practical utility of these findings in real-world applications.

6. Conclusions

This study investigated how the sequencing of text and image inputs in multi-modal prompts influences the reasoning performance of Large Language Models (LLMs), using

exam-like tasks as a controlled testbed for broader applicability, focusing specifically on the impact of the input structure rather than comparing these models with non-multi-modal models or alternative reasoning strategies. Our findings demonstrate that modality sequencing plays a critical yet context-dependent role in reasoning performance, with its impact varying across task complexity, question structure, and model architecture. For simpler tasks involving single images and direct questions, sequencing significantly influenced performance, with specific configurations—such as placing images first or interleaving images within text—yielding higher accuracy. However, in complex reasoning tasks requiring multi-step inference, the logical structure of information flow within the prompt proved more important than the precise ordering of modalities. These results highlight the inherent limitations of transformer-based architectures in sustaining long-range contextual dependencies, where attention decay and positional biases hinder multi-hop reasoning. Consequently, aligning prompts with the cognitive progression of reasoning tasks emerged as a key strategy for optimising performance.

Beyond evaluating system-level behaviour, this research study also examined user interaction strategies for improving multi-modal prompt design in intelligent systems. A key contribution of this work is the derivation of empirically validated heuristics for constructing effective multi-modal prompts. Developed through both systematic empirical analysis and a user study, these heuristics provide actionable guidelines for end-users to optimise input sequencing and interaction with LLMs, ultimately enhancing reasoning accuracy and task efficiency. Our findings underscore the necessity of designing multi-modal prompts that account not only for the intrinsic affordances of different modalities but also for how their sequencing can be strategically structured to support complex reasoning workflows. These insights extend beyond exam-like evaluations to real-world applications in healthcare diagnostics, academic research, legal document analysis, customer support systems, and interactive learning environments, offering practical design implications for interactive AI systems that integrate multi-modal reasoning.

Author Contributions: Conceptualisation, G.W. and T.S.; methodology, T.S. and G.W.; software, G.W.; validation, G.W.; formal analysis, T.S. and G.W.; resources, G.W.; data curation, G.W.; writing—original draft preparation, G.W.; writing—review and editing, G.W. and T.S.; visualisation, G.W. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research study received no external funding.

Institutional Review Board Statement: This study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created, and the dataset benchmarks used in this study are available online.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Appendix: Image–Text Sequence: Attribute-Based Analysis

The following attributes were analysed for their impact on image sequencing, to evaluate whether the trends observed in the overall dataset hold for each of the attributes. For M3Exam: levels, prompt length, and image types. For M3COTS: question types, prompt length, and image types.

Appendix A.1. M3Exam Data

The models' performance was consistent with the results for the overall dataset across the different subgroup attributes. Where contrary results were observed, they were not found to be significant according to McNemar's test. This included image types, levels,

and varying input prompt lengths across all three models. See Table A1 for details on the image type results and Table A2 for details on the levels.

Table A1. A comparison of image types on the M3Exam data using McNemar's test with the image–text configurations of Text First (TF), Image First (IF), and Interleaved (IN).

	TF	IF	IN	TF vs. IF	TF vs. IN	IF vs. IN
				GPT-4o [14]		
Images Only	0.80	0.78	0.80	(Stat. = 20.0, $p = 0.652$)	(Stat. = 13.0, $p = 0.585$)	(Stat. = 19.0, $p = 1.000$)
Mixture	0.84	0.83	0.86	(Stat. = 23.0, $p = 0.775$)	(Stat. = 21.0, $p = 0.169$)	(Stat. = 16.0, $p = 0.054$)
Text Only	1.00	1.00	1.00	(Stat. = 0.0, $p = 1.000$)	(Stat. = 0.0, $p = 1.000$)	(Stat. = 0.0, $p = 1.000$)
				Gemini-1.5 [15]		
Images Only	0.69	0.71	0.72	(Stat. = 22.0, $p = 0.665$)	(Stat. = 19.0, $p = 0.542$)	(Stat. = 19.0, $p = 0.875$)
Mixture	0.78	0.78	0.80	(Stat. = 43.0, $p = 1.000$)	(Stat. = 32.0, $p = 0.349$)	(Stat. = 27.0, $p = 0.314$)
Text Only	1.00	1.00	1.00	(Stat. = 0.0, $p = 1.000$)	(Stat.= 0.0, $p = 1.000$)	(Stat. = 0.0, $p = 1.000$)
				Claude-3 [16]		
Images Only	0.53	0.56	0.55	(Stat. = 19.0, $p = 0.193$)	(Stat. = 23.0, $p = 0.576$)	(Stat. = 20.0, $p = 0.551$)
Mixture	0.70	0.67	0.70	(Stat. = 38.0, $p = 0.170$)	(Stat. = 41.0, $p = 1.000$)	(Stat. = 29.0, $p = 0.125$)
Text Only	1.00	1.00	1.00	(Stat. = 0.0, $p = 1.000$)	(Stat. = 0.0, $p = 1.000$)	(Stat.= 0.0, $p = 1.000$)

Table A2. A comparison of educational levels on the M3Exam data using McNemar's test with the image–text configurations of Text First (TF), Image First (IF), and Interleaved (IN).

	TF	IF	IN	TF vs. IF	TF vs. IN	IF vs. IN
				GPT-4o [14]		
High school (USA)	0.80	0.79	0.82	(Stat. = 31.0, $p = 0.470$)	(Stat. = 23.0, $p = 0.092$)	(Stat. = 25.0, $p = 10.427$)
Middle school (USA)	0.84	0.86	0.85	(Stat. = 5.0, $p = 0.424$)	(Stat. = 7.0, $p = 1.000$)	(Stat. = 6.0, $p = 0.607$)
Elementary school (USA)	0.85	0.90	0.86	(Stat. = 3.0, $p = 0.344$)	(Stat. = 3.0, $p = 1.000$)	(Stat. = 3.0, $p = 0.227$)
				Gemini-1.5 [15]		
High school (USA)	0.72	0.72	0.73	(Stat. = 45.0, $p = 0.917$)	(Stat. = 35.0, $p = 0.567$)	(Stat. = 36.0, $p = 0.500$)
Middle school (USA)	0.80	0.80	0.82	(Stat. = 10.0, $p = 1.000$)	(Stat. = 6.0, $p = 0.454$)	(Stat. = 7.0, $p = 0.481$)
Elementary school (USA)	0.84	0.86	0.87	(Stat. = 10.0, $p = 0.832$)	(Stat.= 5.0, $p = 1.000$)	(Stat. = 8.0, $p = 0.648$)
				Claude-3 [16]		
High school (USA)	0.60	0.60	0.60	(Stat. = 45.0, $p = 1.000$)	(Stat. = 43.0, $p = 1.000$)	(Stat. = 44.0, $p = 1.000$)
Middle school (USA)	0.72	0.72	0.74	(Stat. = 14.0, $p = 1.000$)	(Stat. = 9.0, $p = 0.523$)	(Stat. = 12.0, $p = 0.572$)
Elementary school (USA)	0.69	0.64	0.69	(Stat. = 7.0, $p = 0.359$)	(Stat. = 2.0, $p = 0.180$)	(Stat.= 8.0, $p = 1.000$)

Appendix A.2. M3COTS Data

For M3COTS, relatively to image types and prompt lengths, performance remained consistent with the overall dataset across all three models. Where contrary results were observed, they were not significant according to McNemar's test. See Table A3 for details on the image type results.

Table A3. A comparison of image types on the M3COTS data using McNemar's test with the image–text configurations of Text First (TF), Image First (IF), and Interleaved (IN).

	TF	IF	IN	TF vs. IF	TF vs. IN	IF vs. IN
				GPT-4o [14]		
Images Only	0.58	0.64	0.62	(Stat. = 12.0, $p = 0.034$)	(Stat. = 17.0, $p = 0.135$)	(Stat. = 16.0, $p = 0.736$)
Mixture	0.65	0.70	0.70	(Stat. = 121.0, $p = 0.000$)	(Stat. = 129.0, $p = 0.000$)	(Stat. = 124.0, $p = 0.802$)
Text Only	0.77	0.82	0.86	(Stat. = 30.0, $p = 0.009$)	(Stat. = 18.0, $p = 0.000$)	(Stat. = 13.0, $p = 0.001$)
				Gemini-1.5 [15]		
Images Only	0.51	0.55	0.54	(Stat. = 26.0, $p = 0.597$)	(Stat. = 17.0, $p = 0.135$)	(Stat. = 14.0, $p = 0.392$)
Mixture	0.53	0.58	0.58	(Stat. = 194.0, $p = 0.486$)	(Stat. = 160.0, $p = 0.040$)	(Stat. = 148.0, $p = 0.180$)
Text Only	0.68	0.71	0.71	(Stat. = 49.0, $p = 0.135$)	(Stat. = 49.0, $p = 0.624$)	(Stat. = 48.0, $p = 0.334$)
				Claude-3 [16]		
Images Only	0.43	0.45	0.48	(Stat. = 20.0, $p = 0.253$)	(Stat. = 19.0, $p = 0.451$)	(Stat. = 13.0, $p = 0.851$)
Mixture	0.47	0.48	0.50	(Stat. = 149.0, $p = 0.000$)	(Stat. = 138.0, $p = 0.000$)	(Stat. = 148.0, $p = 0.774$)
Text Only	0.59	0.62	0.60	(Stat. = 40.0, $p = 0.002$)	(Stat. = 24.0, $p = 0.000$)	(Stat. = 47.0, $p = 0.919$)

Table A4. A comparison of image types on various M3COTS data using McNemar's test with image–text configurations of Text First (TF), Image First (IF), and Interleaved (IN).

	TF	IF	IN	TF vs. IF	TF vs. IN	IF vs. IN
Gemini-1.5-Flash						
Materials	0.87	0.99	0.94	(Stat. = 1.0, $p = 0.021$)	(Stat. = 3.0, $p = 0.227$)	(Stat. = 1.0, $p = 0.375$)
Elementary Algebra	0.23	0.34	0.30	(Stat. = 5.0, $p = 0.064$)	(Stat. = 7.0, $p = 0.263$)	(Stat. = 6.0, $p = 0.607$)
Precalculus	0.22	0.11	0.39	(Stat. = 1.0, $p = 0.625$)	(Stat. = 0.0, $p = 0.250$)	(Stat. = 1.0, $p = 0.125$)
grammar-Sentences, fragments, and run-ons	0.96	0.79	0.98	(Stat. = 1.0, $p = 0.021$)	(Stat. = 0.0, $p = 1.000$)	(Stat. = 1.0, $p = 0.012$)
biology-Scientific names	0.90	0.93	0.91	(Stat. = 2.0, $p = 0.453$)	(Stat. = 3.0, $p = 1.000$)	(Stat. = 3.0, $p = 0.727$)
chemistry-Atoms and Molecules Recognize	0.25	0.42	0.42	(Stat. = 4.0, $p = 0.031$)	(Stat. = 4.0, $p = 0.031$)	(Stat. = 4.0, $p = 1.000$)
physics-Particle motion and energy	0.82	0.95	0.95	(Stat. = 1.0, $p = 0.039$)	(Stat. = 2.0, $p = 0.065$)	(Stat. = 2.0, $p = 1.000$)
physics-Magnets	0.59	0.66	0.59	(Stat. = 8.0, $p = 0.383$)	(Stat. = 8.0, $p = 1.000$)	(Stat. = 8.0, $p = 0.383$)
physics-Velocity, acceleration, and forces	0.60	0.64	0.48	(Stat. = 5.0, $p = 0.774$)	(Stat. = 7.0, $p = 0.263$)	(Stat. = 7.0, $p = 0.134$)
physics-Thermal Conductivity Comparison	0.35	0.85	0.85	(Stat. = 2.0, $p = 0.013$)	(Stat. = 0.0, $p = 0.002$)	(Stat. = 2.0, $p = 1.000$)
cognitive-science-Abstract Tangram Recognition	0.42	0.39	0.35	(Stat. = 6.0, $p = 0.115$)	(Stat. = 8.0, $p = 0.210$)	(Stat. = 6.0, $p = 0.454$)
economics-Fiscal Surpluses Calculation	0.59	0.51	0.44	(Stat. = 4.0, $p = 0.754$)	(Stat. = 3.0, $p = 1.000$)	(Stat. = 7.0, $p = 0.629$)
economics-Per Capita Wage Calculation	0.72	0.40	0.35	(Stat. = 2.0, $p = 0.109$)	(Stat. = 3.0, $p = 0.688$)	(Stat. = 3.0, $p = 0.727$)
geography-Climate Analysis	0.76	0.59	0.76	(Stat. = 2.0, $p = 0.109$)	(Stat. = 5.0, $p = 1.000$)	(Stat. = 1.0, $p = 0.070$)

Table A5. A comparison of image types on various M3COTS data using McNemar's test with the image–text configurations of Text First (TF), Image First (IF), and Interleaved (IN).

	TF	IF	IN	TF vs. IF	TF vs. IN	IF vs. IN
GPT-4o						
Materials	0.99	0.94	0.97	(Stat. = 0.0, $p = 0.250$)	(Stat. = 1.0, $p = 1.000$)	(Stat. = 0.0, $p = 0.500$)
Elementary Algebra	0.40	0.54	0.66	(Stat. = 10.0, $p = 0.050$)	(Stat. = 5.0, $p = 0.000$)	(Stat. = 2.0, $p = 0.013$)
Precalculus	0.42	0.53	0.63	(Stat. = 2.0, $p = 0.688$)	(Stat. = 2.0, $p = 0.289$)	(Stat. = 0.0, $p = 0.500$)
grammar-Sentences, fragments, and run-ons	0.96	1.00	0.98	(Stat. = 0.0, $p = 0.500$)	(Stat. = 0.0, $p = 1.000$)	(Stat. = 0.0, $p = 1.000$)
biology-Scientific names	0.94	0.95	0.98	(Stat. = 1.0, $p = 1.000$)	(Stat. = 0.0, $p = 0.250$)	(Stat. = 0.0, $p = 0.500$)
chemistry-Atoms and Molecules Recognize	0.32	0.67	0.72	(Stat. = 5.0, $p = 0.000$)	(Stat. = 5.0, $p = 0.000$)	(Stat. = 9.0, $p = 0.664$)
physics-Particle motion and energy	0.61	0.98	1.00	(Stat. = 1.0, $p = 0.000$)	(Stat. = 0.0, $p = 0.000$)	(Stat. = 0.0, $p = 1.000$)
physics-Magnets	0.85	0.85	0.73	(Stat. = 5.0, $p = 1.000$)	(Stat. = 2.0, $p = 0.022$)	(Stat. = 3.0, $p = 0.035$)
physics-Velocity, acceleration, and forces	0.86	0.88	0.88	(Stat. = 4.0, $p = 1.000$)	(Stat. = 4.0, $p = 1.000$)	(Stat. = 2.0, $p = 1.000$)
physics-Thermal Conductivity Comparison	0.90	0.80	0.65	(Stat. = 2.0, $p = 0.688$)	(Stat. = 1.0, $p = 0.125$)	(Stat. = 2.0, $p = 0.453$)
cognitive-science-Abstract Tangram Recognition	0.42	0.50	0.50	(Stat. = 6.0, $p = 0.115$)	(Stat. = 8.0, $p = 0.152$)	(Stat. = 10.0, $p = 1.000$)
economics-Fiscal Surpluses Calculation	0.59	0.63	0.56	(Stat. = 4.0, $p = 0.754$)	(Stat. = 3.0, $p = 1.000$)	(Stat. = 3.0, $p = 0.508$)
economics-Per Capita Wage Calculation	0.72	0.57	0.47	(Stat. = 2.0, $p = 0.109$)	(Stat. = 2.0, $p = 0.013$)	(Stat. = 4.0, $p = 0.388$)
geography-Climate Analysis	0.76	0.59	0.76	(Stat. = 2.0, $p = 0.109$)	(Stat. = 5.0, $p = 1.000$)	(Stat. = 1.0, $p = 0.070$)

Table A6. Example question types showing different optimal image–text sequencing patterns: Text First (TF), Image First (IF), and Interleaved (IN).

	TF	IF	IN	TF vs. IF	TF vs. IN	IF vs. IN
GPT-4o [14]						
Chemistry-Atoms and Molecules Recognize	0.32	0.67	0.72	(Stat. = 5.0, $p = 0.000$)	(Stat. = 5.0, $p = 0.000$)	(Stat. = 9.0, $p = 0.664$)
Physics-Velocity, acceleration, and forces	0.86	0.88	0.88	(Stat. = 4.0, $p = 1.000$)	(Stat. = 4.0, $p = 1.000$)	(Stat. = 2.0, $p = 1.000$)
Economics-Per Capita Wage Calculation	0.73	0.58	0.48	(Stat. = 2.0, $p = 0.019$)	(Stat. = 2.0, $p = 0.013$)	(Stat.= 4.0, $p = 0.388$)
Gemini-1.5 [15]						
Chemistry-Atoms and Molecules Recognize	0.25	0.43	0.43	(Stat. = 4.0, $p = 0.031$)	(Stat. = 4.0, $p = 0.031$)	(Stat. = 4.0, $p = 0.031$)
Physics-Velocity, acceleration, and forces	0.60	0.64	0.48	(Stat. = 5.0, $p = 0.774$)	(Stat.= 7.0, $p = 0.263$)	(Stat. = 7.0, $p = 0.134$)
Economics-Per Capita Wage Calculation	0.30	0.40	0.35	(Stat. = 3.0, $p = 0.344$)	(Stat. = 2.0, $p = 0.688$)	(Stat.= 3.0, $p = 0.727$)
Claude-3 [16]						
Chemistry-Atoms and Molecules Recognize	0.27	0.28	0.33	(Stat. = 10.0, $p = 1.000$)	(Stat. = 7.0, $p = 0.629$)	(Stat. = 7.0, $p = 0.481$)
Physics-Velocity, acceleration, and forces	0.48	0.18	0.26	(Stat. = 3.0, $p = 0.001$)	(Stat. = 5.0, $p = 0.027$)	(Stat.= 3.0, $p = 0.344$)
Economics-Per Capita Wage Calculation	0.40	0.28	0.35	(Stat. = 3.0, $p = 0.227$)	(Stat. = 4.0, $p = 0.754$)	(Stat.= 1.0, $p = 0.375$)

Appendix A.3. Comparison of Prompt Lengths

Figures A1–A3 present a comparison of total input prompt lengths with the accuracy of the answers for each sequencing configuration, illustrating the empirical trends observed across different prompt lengths. The prompt lengths include the token counts for both the image and text components as indicated by each model. For the Gemini model, the image length has a standard token length of 258 tokens.

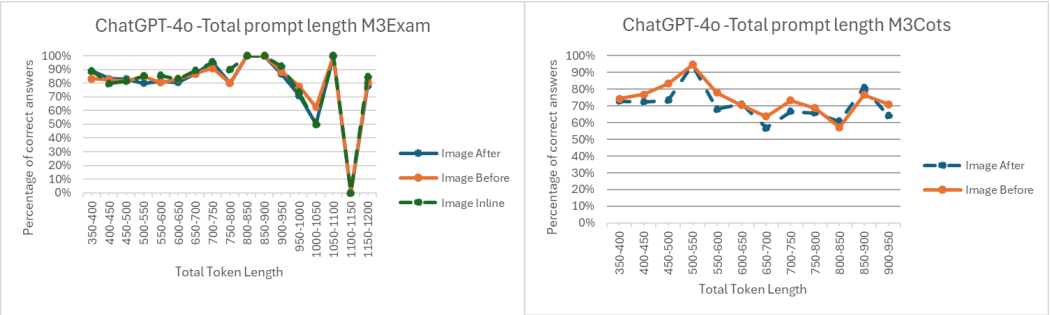


Figure A1. Comparison of prompt lengths across various sequencing configurations for GPT-4o.

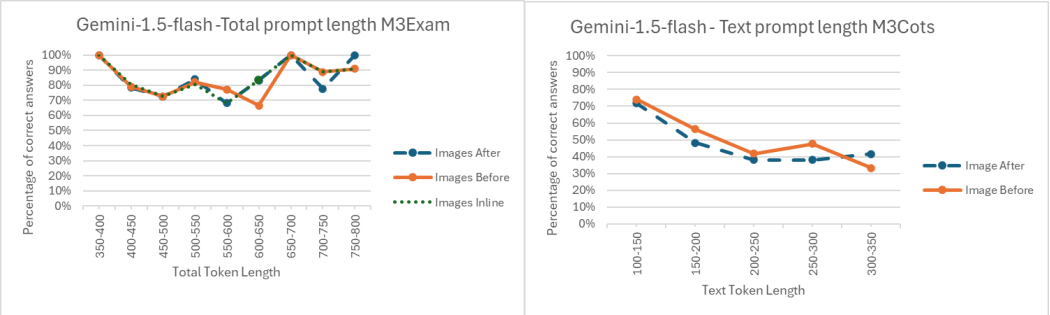


Figure A2. Comparison of prompt lengths across various sequencing configurations for Gemini-1.5.

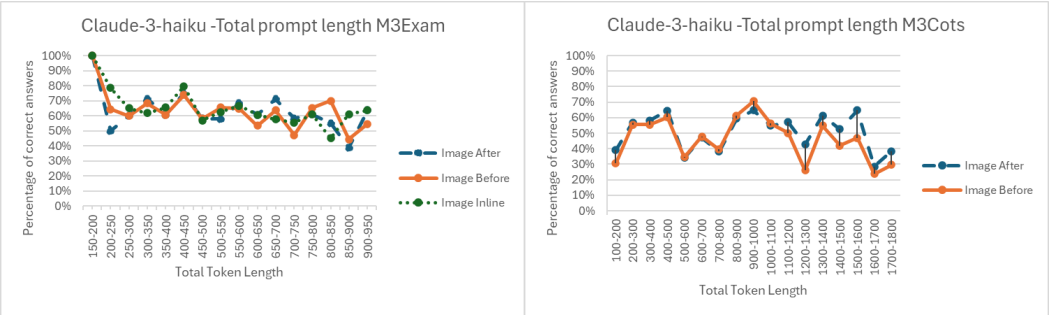


Figure A3. Comparison of prompt lengths across various sequencing configurations for Claude-3.

References

1. Royce, C.S.; Hayes, M.M.; Schwartzstein, R.M. Teaching critical thinking: A case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. *Acad. Med.* **2019**, *94*, 187–194. [CrossRef] [PubMed]
2. Kosinski, M. Evaluating large language models in theory of mind tasks. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2405460121. [CrossRef] [PubMed]
3. Zhang, W.; Aljunied, M.; Gao, C.; Chia, Y.K.; Bing, L. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: San Francisco, CA, USA, 2023; Volume 36, pp. 5484–5505.
4. Chen, Q.; Qin, L.; Zhang, J.; Chen, Z.; Xu, X.; Che, W. M^3CoT: A Novel Benchmark for Multi-Domain Multi-step Multi-modal Chain-of-Thought. In Proceedings of the ACL, Bangkok, Thailand, 11–16 August 2024.
5. Mcintosh, T.R.; Liu, T.; Susnjak, T.; Watters, P.; Halgamuge, M.N. A Reasoning and Value Alignment Test to Assess Advanced GPT Reasoning. *ACM Trans. Interact. Intell. Syst.* **2024**, *14*, 3. [CrossRef]
6. Guo, Y.; Shi, D.; Guo, M.; Wu, Y.; Cao, N.; Chen, Q. Talk2Data: A Natural Language Interface for Exploratory Visual Analysis via Question Decomposition. *ACM Trans. Interact. Intell. Syst.* **2024**, *14*, 8. [CrossRef]
7. McIntosh, T.R.; Susnjak, T.; Liu, T.; Watters, P.; Ng, A.; Halgamuge, M.N. A Game-Theoretic Approach to Containing Artificial General Intelligence: Insights from Highly Autonomous Aggressive Malware. *IEEE Trans. Artif. Intell.* **2024**, *5*, 6290–6303. [CrossRef]

8. Feng, T.H.; Denny, P.; Wuensche, B.; Luxton-Reilly, A.; Hooper, S. More Than Meets the AI: Evaluating the performance of GPT-4 on Computer Graphics assessment questions. In Proceedings of the 26th Australasian Computing Education Conference, Sydney, NSW, Australia, 29 January–2 February 2024; pp. 182–191.
9. Pal, A.; Sankarasubbu, M. Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. *arXiv* **2024**, arXiv:2402.07023. Available online: http://arxiv.org/abs/2402.07023 (accessed on 26 July 2024).
10. Stribling, D.; Xia, Y.; Amer, M.K.; Graim, K.S.; Mulligan, C.J.; Renne, R. The model student: GPT-4 performance on graduate biomedical science exams. *Sci. Rep.* **2024**, *14*, 5670. [CrossRef]
11. Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv* **2024**, arXiv:2307.06281. Available online: http://arxiv.org/abs/2307.06281 (accessed on 26 July 2024).
12. Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.W.; Zhu, S.C.; Tafjord, O.; Clark, P.; Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2507–2521.
13. Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; Shan, Y. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv* **2023**, arXiv:2307.16125. Available online: http://arxiv.org/abs/2307.16125 (accessed on 26 July 2024).
14. OpenAI. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774. Available online: http://arxiv.org/abs/2303.08774 (accessed on 27 July 2024).
15. Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* **2024**, arXiv:2403.05530.
16. Anthropic, A. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card* **2024**, *1*.
17. Crisp, V.; Sweiry, E. Can a picture ruin a thousand words? Physical aspects of the way exam questions are laid out and the impact of changing them. In Proceedings of the British Educational Research Association Annual Conference, Edinburgh, Scotland, 10–13 September 2003.
18. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [CrossRef]
19. Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; Stenetorp, P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 8086–8098. [CrossRef]
20. Wang, Z.; Zhang, H.; Li, X.; Huang, K.H.; Han, C.; Ji, S.; Kakade, S.M.; Peng, H.; Ji, H. Eliminating Position Bias of Language Models: A Mechanistic Approach. *arXiv* **2024** arXiv:2407.01100.
21. Gao, J.; Lanchantin, J.; Soffa, M.L.; Qi, Y. Black-box generation of adversarial text sequences to evade deep learning classifiers. In Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 24 May 2018; pp. 50–56.
22. Garg, S.; Ramakrishnan, G. BAE: BERT-based Adversarial Examples for Text Classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6174–6181. [CrossRef]
23. Leidinger, A.; van Rooij, R.; Shutova, E. The language of prompting: What linguistic properties make a prompt successful? In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 9210–9232. [CrossRef]
24. Chu, K.; Chen, Y.P.; Nakayama, H. A Better LLM Evaluator for Text Generation: The Impact of Prompt Output Sequencing and Optimization. *arXiv* **2024**, arXiv:2406.09972.
25. Levy, M.; Jacoby, A.; Goldberg, Y. Same task, more tokens: The impact of input length on the reasoning performance of large language models. *arXiv* **2024**, arXiv:2402.14848.
26. Mitra, C.; Huang, B.; Darrell, T.; Herzig, R. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 14420–14431.
27. Zhou, Q.; Zhou, R.; Hu, Z.; Lu, P.; Gao, S.; Zhang, Y. Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models. *arXiv* **2024**, arXiv:2405.13872. Available online: http://arxiv.org/abs/2405.13872 (accessed on 26 July 2024).
28. Zheng, G.; Yang, B.; Tang, J.; Zhou, H.; Yang, S. DDCoT: Duty-Distinct Chain-of-Thought Prompting for Multimodal Reasoning in Language Models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 5168–5191. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/file/108030643e640ac050e0ed5e6aace48f-Paper-Conference.pdf (accessed on 26 July 2024).
29. Luan, B.; Feng, H.; Chen, H.; Wang, Y.; Zhou, W.; Li, H. TextCoT: Zoom In for Enhanced Multimodal Text-Rich Image Understanding. *arXiv* **2024**, arXiv:2404.09797. Available online: http://arxiv.org/abs/2404.09797 (accessed on 26 July 2024).

30. Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; Smola, A. Multimodal Chain-of-Thought Reasoning in Language Models. *arXiv* **2024**, arXiv:2302.00923. Available online: http://arxiv.org/abs/2302.00923 (accessed on 27 July 2024).
31. Susnjak, T.; McIntosh, T.R. ChatGPT: The End of Online Exam Integrity? *Educ. Sci.* **2024**, *14*, 656. [CrossRef]
32. Johnson-Laird, P.N. Mental models and human reasoning. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 18243–18250. [CrossRef]
33. Mitchell, M. Debates on the nature of artificial general intelligence. *Science* **2024**, *383*, eado7069. [CrossRef]
34. Mitchell, M. AI's challenge of understanding the world. *Science* **2023**, *382*, eadm8175. [CrossRef] [PubMed]
35. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent Abilities of Large Language Models. *Trans. Mach. Learn. Res.* **2022**.
36. Mialon, G.; Fourrier, C.; Swift, C.; Wolf, T.; LeCun, Y.; Scialom, T. Gaia: A benchmark for general ai assistants. *arXiv* **2023**, arXiv:2311.12983.
37. Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; et al. Augmented Language Models: A Survey. *arXiv* **2023**, arXiv:2302.07842. Available online: http://arxiv.org/abs/2302.07842 (accessed on 26 June 2024).
38. LeCun, Y. A path towards autonomous machine intelligence version *Open Rev.* **2022**, *62*, 1–62. Available online: https://openreview.net/pdf?id=BZ5a1r-kVsf (accessed on 26 June 2024).
39. Kambhampati, S. Can large language models reason and plan? *Ann. N. Y. Acad. Sci. USA* **2024**, *1534*, 15–18. [CrossRef]
40. West, P.; Lu, X.; Dziri, N.; Brahman, F.; Li, L.; Hwang, J.D.; Jiang, L.; Fisher, J.; Ravichander, A.; Chandu, K.; et al. THE GENERATIVE AI PARADOX: "What It Can Create, It May Not Understand". In Proceedings of the The Twelfth International Conference on Learning Representations, Vienna Austria, 7–11 May 2024.
41. McIntosh, T.R.; Susnjak, T.; Liu, T.; Watters, P.; Halgamuge, M.N. The Inadequacy of Reinforcement Learning from Human Feedback-Radicalizing Large Language Models via Semantic Vulnerabilities. *IEEE Trans. Cogn. Dev. Syst.* **2024**, *16*, 1561–1574. [CrossRef]
42. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 39. [CrossRef]
43. Wang, J.; Huang, Y.; Chen, C.; Liu, Z.; Wang, S.; Wang, Q. Software Testing With Large Language Models: Survey, Landscape, and Vision. *IEEE Trans. Softw. Eng.* **2024**, *50*, 911–936. [CrossRef]
44. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
45. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22199–22213.
46. Yu, Z.; He, L.; Wu, Z.; Dai, X.; Chen, J. Towards Better Chain-of-Thought Prompting Strategies: A Survey. *arXiv* **2023**, arXiv:2310.04959. Available online: http://arxiv.org/abs/2310.04959 (accessed on 6 June 2024).
47. Wang, B.; Min, S.; Deng, X.; Shen, J.; Wu, Y.; Zettlemoyer, L.; Sun, H. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 2717–2739. [CrossRef]
48. Ranasinghe, K.; Shukla, S.N.; Poursaeed, O.; Ryoo, M.S.; Lin, T.Y. Learning to Localize Objects Improves Spatial Reasoning in Visual-LLMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 12977–12987.
49. Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Driess, D.; Florence, P.; Sadigh, D.; Guibas, L.; Xia, F. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 14455–14465. [CrossRef]
50. Asch, S.E. Forming impressions of personality. *J. Abnorm. Soc. Psychol.* **1946**, *41*, 258. [CrossRef]
51. Baddeley, A.D.; Hitch, G. The recency effect: Implicit learning with explicit retrieval? *Mem. Cogn.* **1993**, *21*, 146–155. [CrossRef]
52. Wang, Y.; Cai, Y.; Chen, M.; Liang, Y.; Hooi, B. Primacy Effect of ChatGPT. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 108–115. [CrossRef]
53. Zhang, Z.; Yu, J.; Li, J.; Hou, L. Exploring the Cognitive Knowledge Structure of Large Language Models: An Educational Diagnostic Assessment Approach. In Proceedings of the Findings of the Association for Computational Linguistics EMNLP 2023, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 1643–1650. [CrossRef]
54. Eicher, J.E.; Irgolič, R. Compensatory Biases Under Cognitive Load: Reducing Selection Bias in Large Language Models. *arXiv* **2024**, arXiv:2402.01740.

55. Goolge. Image Understanding. 2024. Available online: https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/image-understanding (accessed on 15 July 2024).
56. Hatzakis, S. When-Processing-a-Text-Prompt-Before-It-or-After-It. 2023. Available online: https://community.openai.com/t/when-processing-a-text-prompt-before-it-or-after-it/247801/3 (accessed on 15 July 2023).
57. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
58. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
59. Zhao, F.; Zhang, C.; Geng, B. Deep Multimodal Data Fusion. *ACM Comput. Surv.* **2024**, *56*, 216. [CrossRef]
60. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Conference. Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 2019, p. 6558.
61. Team, C. Chameleon: Mixed-modal early-fusion foundation models. *arXiv* **2024**, arXiv:2405.09818.
62. Gan, Z.; Chen, Y.C.; Li, L.; Zhu, C.; Cheng, Y.; Liu, J. Large-scale adversarial training for vision-and-language representation learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6616–6628.
63. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef]
64. Liang, P.P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M.A.; Zhu, Y.; et al. Multibench: Multiscale benchmarks for multimodal representation learning. *Adv. Neural Inf. Process. Syst.* **2021**, *2021*, 1. [PubMed]
65. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18166–18176.
66. Shukor, M.; Fini, E.; da Costa, V.G.T.; Cord, M.; Susskind, J.; El-Nouby, A. Scaling Laws for Native Multimodal Models. *arXiv* **2025**, arXiv:2504.07951.
67. Peysakhovich, A.; Lerer, A. Attention sorting combats recency bias in long context language models. *arXiv* **2023**, arXiv:2310.01427.
68. Giaccardi, E.; Murray-Rust, D.; Redström, J.; Caramiaux, B. Prototyping with Uncertainties: Data, Algorithms, and Research through Design. *ACM Trans. Comput.-Hum. Interact.* **2024**, *31*, 68. [CrossRef]
69. Thieme, A.; Rajamohan, A.; Cooper, B.; Groombridge, H.; Simister, R.; Wong, B.; Woznitza, N.; Pinnock, M.A.; Wetscherek, M.T.; Morrison, C.; et al. Challenges for Responsible AI Design and Workflow Integration in Healthcare: A Case Study of Automatic Feeding Tube Qualification in Radiology. *ACM Trans. Comput.-Hum. Interact.* **2025**. [CrossRef]
70. Zając, H.D.; Andersen, T.O.; Kwasa, E.; Wanjohi, R.; Onyinkwa, M.K.; Mwaniki, E.K.; Gitau, S.N.; Yaseen, S.S.; Carlsen, J.F.; Fraccaro, M.; et al. Towards Clinically Useful AI: From Radiology Practices in Global South and North to Visions of AI Support. *ACM Trans. Comput.-Hum. Interact.* **2025**, *32*, 20. [CrossRef]
71. Chen, C.; Nguyen, C.; Groueix, T.; Kim, V.G.; Weibel, N. MemoVis: A GenAI-Powered Tool for Creating Companion Reference Images for 3D Design Feedback. *ACM Trans. Comput.-Hum. Interact.* **2024**, *31*, 67. [CrossRef]
72. August, T.; Wang, L.L.; Bragg, J.; Hearst, M.A.; Head, A.; Lo, K. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* **2023**, *30*, 74. [CrossRef]
73. Huang, T.; Yu, C.; Shi, W.; Peng, Z.; Yang, D.; Sun, W.; Shi, Y. Prompt2Task: Automating UI Tasks on Smartphones from Textual Prompts. *ACM Trans. Comput.-Hum. Interact.* **2025**. [CrossRef]
74. Nguyen, X.P.; Zhang, W.; Li, X.; Aljunied, M.; Tan, Q.; Cheng, L.; Chen, G.; Deng, Y.; Yang, S.; Liu, C.; et al. SeaLLMs–Large Language Models for Southeast Asia. *arXiv* **2023**, arXiv:2312.00738.
75. Liu, C.; Zhang, W.; Zhao, Y.; Luu, A.T.; Bing, L. Is Translation All You Need? A Study on Solving Multilingual Tasks with Large Language Models. *arXiv* **2024**, arXiv:2403.10258.
76. Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; Steinhardt, J. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv* **2021**, arXiv:2103.03874.
77. Hessel, J.; Hwang, J.D.; Park, J.S.; Zellers, R.; Bhagavatula, C.; Rohrbach, A.; Saenko, K.; Choi, Y. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 558–575.
78. Wu, Y.; Zhang, P.; Xiong, W.; Oguz, B.; Gee, J.C.; Nie, Y. The role of chain-of-thought in complex vision-language reasoning task. *arXiv* **2023**, arXiv:2311.09193.

79. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv* **2023**, arXiv:2203.11171. Available online: http://arxiv.org/abs/2203.11171 (accessed on 6 June 2024).
80. McIntosh, T.R.; Susnjak, T.; Arachchilage, N.; Liu, T.; Xu, D.; Watters, P.; Halgamuge, M.N. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. *IEEE Trans. Artif. Intell.* 2025, *early access*, 1–18. [CrossRef]