A Dual-Task Paradigm to Investigate Sentence Comprehension Strategies in Language Models

Rei Emura
Tohoku University
rei.emura.r4@dc.tohoku.ac.jp

Saku Sugawara
National Institute of Informatics
The University of Tokyo
saku@nii.ac.jp

Abstract

Language models (LMs) behave more like humans when their cognitive resources are restricted, particularly in predicting sentence processing costs such as reading times. However, it remains unclear whether such constraints similarly affect sentence comprehension strategies. Besides, existing methods do not directly target the balance between memory storage and sentence processing, which is central to human working memory. To address this issue, we propose a dual-task paradigm that combines an arithmetic computation task with a sentence comprehension task, such as "The 2 cocktail + blended 3 =..." Our experiments show that under dual-task conditions, GPT-4o, o3-mini, and o4-mini shift toward plausibility-based comprehension, mirroring humans' rational inference. Specifically, these models show a greater accuracy gap between plausible sentences (e.g., "The cocktail was blended by the bartender") and implausible sentences (e.g., "The bartender was blended by the cocktail") in the dual-task condition compared to the single-task conditions. These findings suggest that constraints on the balance between memory and processing resources promote rational inference in LMs. More broadly, they support the view that human-like sentence comprehension fundamentally arises from the allocation of limited cognitive resources.

1 Introduction

Working memory is a cognitive system that temporarily stores and maintains information necessary for processing in an accessible state (Atkinson and Shiffrin, 1971; Baddeley and Hitch, 1974; Baddeley, 2003). It is essential for understanding language in humans (Just and Carpenter, 1992).

Comparing the working memory of LMs with that of humans helps us understand what makes sentence comprehension more human-like. The limitation of cognitive resources (analogous to working memory in humans) for sentence comprehension makes LMs behave more like humans (Futrell et al., 2020; Hahn et al., 2022; Kuribayashi et al., 2021, 2022; Oh et al., 2022; Oh and Schuler, 2023; Timkey and Linzen, 2023; Wilcox et al., 2025). Specifically, LMs with restricted memory resources better approximate human reading costs, such as reading times. These findings suggest that resource constraints may be a fundamental property of human-like language comprehension.

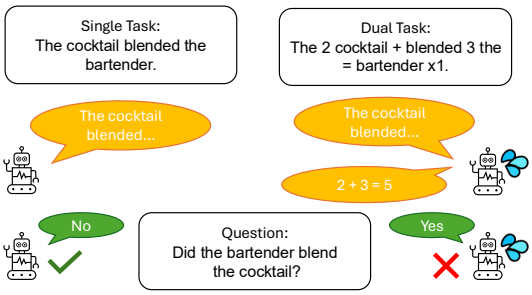


Figure 1: Overview of hypothesis and tasks. This study investigates whether language models, similar to humans, prioritize plausibility over grammar when understanding implausible sentences in a dual-task situation where they simultaneously perform calculations and sentence comprehension.

However, it remains unclear whether LMs' reading strategies exhibit patterns analogous to those of humans under limited cognitive resources. Previous studies have shown that LMs achieve lower accuracy on complex sentences, suggesting that LMs rely on working-memory-like mechanisms during sentence processing (Amouyal et al., 2025a,b; Irwin et al., 2023). We therefore examine whether LMs adopt human-like comprehension strategies when cognitive resources are constrained.

Existing approaches to constraining cognitive resources in LMs also face methodological limitations. Prior work has mainly manipulated input length or model parameters (Asami and Sugawara,

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12065–12084

2024; Kuribayashi et al., 2021, 2022; Oh et al., 2022; Oh and Schuler, 2023; Timkey and Linzen, 2023; Wilcox et al., 2025). These methods either fail to capture the balance between storage and processing that characterizes human working memory or fail to induce modifications within a single LM.

To address the methodological issue, we propose a dual-task paradigm in which models simultaneously solve arithmetic problems and answer comprehension questions, such as "The 2 cocktail + blended 3 =..." (see Figure 1). We compare comprehension accuracy across three conditions: (i) single task (comprehension of sentences without calculation), (ii) noisy single task (comprehension of sentences with embedded calculation, but no concurrent arithmetic solving), and (iii) dual task (comprehension of sentences with embedded calculation while solving the arithmetic problems).

We focus on one characteristic property of human reading strategies: limited cognitive resources promote rational inference (Futrell and Gibson, 2017; Gibson et al., 2013, 2016). Rational inference involves interpreting semantically implausible sentences (e.g., "The cocktail blended the bartender.") as plausible meanings consistent with world knowledge (e.g., "The bartender blended the cocktail."), prioritizing prior knowledge about plausibility over grammatical structure. Humans are more likely to adopt this strategy under high cognitive load, than under low load (Ayasse et al., 2021; Ferreira, 2003; Gibson et al., 2013).[1]

Taken together, we examine whether LMs adopt rational strategies under dual-task conditions and, if so, which conditions promote such strategies.[2] Our contributions are summarized as follows:

(i) We propose a dual-task paradigm to test the behavior of LMs under limited cognitive resources. This paradigm allows us to observe the LMs' function in integrating memory with sentence processing, analogous to human working memory.

(ii) We demonstrate that GPT-4o, o3-mini, and o4-mini shift their comprehension strategies toward rational inference under the dual-task condition, showing a larger accuracy gap between plausible and implausible sentences than in the single-task and noisy single-task conditions.

(iii) We show that these models are more likely to misunderstand implausible sentences with passive, dative, or benefactive structures. This suggests that they rely more on world knowledge and superficial word order than on function words under limited cognitive resources.

(iv) Our findings support the view that resource constraints are a fundamental property of human-like language comprehension, extending previous evidence from reading costs to reading strategies.

2 Related Work

2.1 Approach to Constrain the LMs' Working Memory

There are two major approaches in computational psycholinguistics for constraining the cognitive resources of LMs: manipulating the input text and altering the models' parameters.

For the first approach, Asami and Sugawara (2024) manipulates the length of entire sentences and compares the accuracy differences between plausible and implausible sentences. Their results show that longer sentences reduce accuracy for both types, indicating that increased length does not promote greater reliance on plausibility information. This may be because such manipulation merely increases processing demands without engaging the balance between storage and processing, which is a central feature of human working memory. A key function of working memory is its dual role of storage and processing, such as remembering a series of numbers while performing a distracting task (Atkinson and Shiffrin, 1971; Baddeley and Hitch, 1974; Baddeley, 2003; Just and Carpenter, 1992).

Regarding the second approach, studies have shown that reducing the number of attention heads (Timkey and Linzen, 2023), larger perplexity (Kuribayashi et al., 2021; Oh et al., 2022; Oh and Schuler, 2023), limiting context access (Kuribayashi et al., 2022), and small training data size and training steps (Wilcox et al., 2025) lead to better prediction of human reading times. However,

[1] The tendency to prioritize plausibility over grammatical information has been explained not only by rational inference theory, but also by the good-enough theory (Christianson, 2016; Ferreira and Patson, 2007) and shallow processing (Sanford and Sturt, 2002). However, we focus on the observed behavioral patterns, and therefore do not engage with their underlying mechanisms.

[2] Our data, codes, and results are available at https://github.com/reiemura/llm-dual-task.

these manipulations create different model configurations, each trained for specific tasks, and therefore reflect differences between models rather than changes within a single model. This is analogous to comparing human participants with different working memory capacities, rather than examining how one participant adapts under varying conditions.

To address these limitations, we introduce a dual-task paradigm in which arithmetic expressions are interleaved with sentence words (see Figure 1). This design maintains the need for memory storage while imposing additional processing demands, thereby constraining the working-memory function that balances storage and processing. Furthermore, by manipulating the task rather than the model, our approach sheds light on how a single model alters its reading strategy under resource constraints.

2.2 Dual-Task Approaches in LMs and Humans

Previous work on multi-task processing in LMs has primarily aimed to improve performance or efficiency by enabling models to handle multiple tasks simultaneously (Cheng et al., 2023; Son et al., 2024). These studies focus on optimizing accuracy or speed and are not designed to investigate how cognitive resource limitations affect language comprehension.

In contrast, some studies have attempted to constrain LMs' working memory using n-back tasks (Kirchner, 1958), with the explicit goal of taxing internal memory resources (Gong et al., 2024; Zhang et al., 2024). While these studies share our objective of probing working memory limitations, n-back tasks primarily engage numerical memory and calculation rather than sentence comprehension.

We build on this latter approach by focusing on working memory in language comprehension. Specifically, our dual-task approach is inspired by human working-memory paradigms, particularly the operation span task (Turner and Engle, 1989). This task requires participants to perform arithmetic operations while memorizing words or letters. In the original version (Turner and Engle, 1989), a mathematical problem followed by a to-be-remembered word is presented, such as "(3 × 4) + 11 = 20? BEAR." Participants first read the problem aloud and judge whether the answer is correct, then read and memorize the following word. After several trials (typically two to six), they are asked to recall the memorized words in the correct order.

3 Methods

3.1 Task

We conduct three types of question-answering tasks. The Dual Task is designed according to the LM's specifications (where a list of strings is more suitable). The Noisy Single Task is included to examine whether performance changes are due to the presence of noisy arithmetic expressions or to the additional cognitive demands of the Dual Task. Exact prompts are in Appendix B.

(i) **Single Task (Single)**: The LM receives a sentence without any embedded arithmetic problems, then answers a comprehension question about the sentence.

(ii) **Noisy Single Task (Noisy)**: The LM receives a sentence with embedded arithmetic problems but ignores them, then answers a comprehension question about the sentence.

(iii) **Dual Task (Dual)**: The LM receives a sentence with embedded arithmetic problems, solves the arithmetic problems, and then answers a comprehension question about the sentence.

3.2 Dataset

We use a subset of stimuli from the GELP dataset (Asami and Sugawara, 2024). The dataset consists of sentence–question pairs. Each sentence includes one premise connected with two propositions.[3]

Premises are manipulated by plausibility (**Plausible** / **Implausible**) and construction (**Transitive** / **Passive** / **Dative** / Experiencer Subject (**Exp.Subj.**) / Experiencer Object (**Exp.Obj.**) / Benefactive For (**Ben.For**)), as illustrated in Table 1. Although the original dataset includes eight constructions, two are excluded during preprocessing (see Section 4.2.1 for details).

In addition, to conduct the Dual Task and Noisy Single Task, we add arithmetic expressions to these sentences. Randomly generated computation problems are inserted with identifiers ("x1," "x2," "x3" ...). After the "=," the corresponding identifier string ("x1," "x2," "x3"...) is appended. Each arithmetic expression is interleaved into the sentence one word at a time. If the end of the sentence is

[3] GELP also contains sentences with one or no propositions. We use only sentences with two propositions, corresponding to the high memory-load condition.

Factor	Variable	Premise and stimuli (Implausible except for the top)
Plausibility	Plausible	The bartender blended the cocktail. (Premise)
	Implausible	The cocktail blended the bartender. (Premise)
Construction	Transitive	The cocktail blended the bartender. (Premise)
	Passive	The bartender was blended by the cocktail. (Premise)
	Dative	The chef sent the friend to the gift. (Premise)
	Exp.Subj.	The view missed the traveler. (Premise)
	Exp.Obj.	The researcher encouraged the results. (Premise)
	Ben.For	The uncle bought the nephew for the toy. (Premise)
Task	Single	The cocktail blended the bartender and the intruder cited the patent after the neurologist baffled the hippie. (Stimuli)
	Noisy & Dual	The 5 cocktail + blended 6 the = bartender x5633 and 9 the + authorities 3 agitated = the x5634 organist 6 after + the 8 infantryman = saluted x5635 the 3 pollster. (Stimuli, 1dig.2add.)
Correct Answer	Yes / No	Did the bartender blend the cocktail? (Question)

Table 1: Examples of premises and stimuli depending on factors and variables. Plausibility and Construction display premises, and Task displays stimuli we actually used in the experiment. Abbreviations: Exp.Subj = Experiencer Subject; Exp.Obj. = Experiencer Object; Ben.For = Benefative For.

reached in the middle of an arithmetic problem, the remaining calculation problems are not added.

We use ten types of arithmetic problems, varying in both digit length (1, 3, 5, 10, and 30 digits) and the number of addends (two vs. three). They are abbreviated as Xdig.Yadd. (e.g., 1dig.2add. indicates the addition of two one-digit numbers). Example stimuli for some arithmetic types are provided in Appendix A.

All comprehension questions are binary (Yes/No), balanced such that half of the correct answers are Yes and half are No. The final dataset contains 2,560 sentence–question pairs (2 plausibility levels $\times$ 8 constructions $\times$ 160 items).

4 Experiments

4.1 Experimental Setup

We evaluate seven LMs: GPT-4o (OpenAI, 2024), o3-mini, o4-mini,[4] GPT-4.1 (OpenAI, 2024), DeepSeek-V3 (DeepSeek-AI et al., 2025), Llama-3.3 (Grattafiori et al., 2024), and Gemma-3 (Team et al., 2025). The models and prompts are selected based on the following two criteria: (i) accuracy for the implausible condition in the Single Task must be at least 70%, and (ii) accuracy for the arithmetic problems in the Dual Task of the 1dig.2add. condition must be at least 80%.[5] We set the temperature to 0.0.[6]

4.2 Evaluation Metrics

4.2.1 Preprocessing

Prior to analysis, we filter the data based on model performance on the Single-Task comprehension task and the Dual-Task arithmetic problems.

First, we compute single-task accuracy for each construction and plausibility condition and exclude those below 80%. As a result, two constructions (double object and benefactive double object) are excluded for all models, with one additional construction excluded for DeepSeek-V3 and three for Gemma-3. This ensures that analyses include only constructions that models reliably comprehend in the single task.

Second, we exclude arithmetic problem types with incorrect answers exceeding 40%. We also remove trials where the arithmetic problem was solved incorrectly or the comprehension response could not be extracted. This filtering step ensures that the models analyzed do not adopt strategies that ignore or skip the arithmetic task.

[4] https://openai.com/index/introducing-o3-and-o4-mini/

[5] We also test GPT-3.5-turbo (https://platform.openai.com/docs/models), but it does not meet these criteria.

[6] This operation was restricted for the o3-mini and o4-mini.

4.2.2 Accuracy of Comprehension Task

We statistically analyze whether the plausibility effect is larger in the dual task than in the single task and noisy single task, using R (R Core Team, 2025). We use a per-item, non-parametric difference-in-differences procedure. For each item i and task t, we compute the mean accuracy $\hat{p}_{itp}$ within each plausibility level p. The within-task plausibility contrast is defined as:

$$\Delta_{it} = \hat{p}_{it,\text{Plausible}} - \hat{p}_{it,\text{Implausible}}. \quad (1)$$

For each item, we then calculate two difference-in-differences contrasts:

$$D_i^{DS} = \Delta_{it,\text{Dual}} - \Delta_{it,\text{Single}} \quad (2)$$

$$D_i^{DN} = \Delta_{it,\text{Dual}} - \Delta_{it,\text{Noisy}}. \quad (3)$$

Finally, we conduct one-sided Wilcoxon signed-rank tests to assess $H_0 : median(D) \leq 0$ separately for D_i^{DS} and D_i^{DN}. The null hypothesis H_0 is rejected if the one-sided test is significant at $\alpha = 0.05$.

4.3 Human Experiment

We evaluate LMs against well-established human phenomena following prior work in psycholinguistics (Ayasse et al., 2021; Ferreira, 2003; Gibson et al., 2013). Nevertheless, to verify whether consistent patterns can be replicated in our dual-task paradigm, we collected a small dataset of human data.

We recruit 33 native English speakers using a crowdsourcing service called Prolific (https://www.prolific.com/). We only recruited people who live in the U.S., the U.K., Ireland, Australia, or New Zealand, have a bachelor's degree (this is because this dual-task is somewhat difficult to follow), and have approval rates of 97% or higher. We obtained informed consent prior to the experiment, and compensated the participants approximately £10.00 for their 1-hour participation.

The design, stimuli, and procedures are the same as the LM experiment, but the arithmetic is limited to 1dig.2add. because other arithmetic types are too challenging for humans. The stimuli include 10% of the items for LMs and are distributed into 12 lists using the Latin square design. Each list contains 65 trials, and each item appears only once in one condition in each list. The appendix E provides details of the procedures and instructions for participants.

In data analysis, we adopted the same criteria as in the LM analysis for participant-level screening, sentence-construction screening, and trial-level screening. After applying these criteria, data from 14 participants (mean age $\pm$ standard deviation: 39.07 $\pm$ 12.30; 9 females and 5 males) are retained for analysis. In the sentence-construction screening, the double object construction and benefactive double object construction are excluded, consistent with the LM analysis. Ben.For and Exp.Obj. are also excluded from the human analysis. We do not apply statistical tests for human data because the sample size is too small.

4.4 Results

Figure 2 shows the mean comprehension accuracy by plausibility, LM, and construction. As seen in the graph, whether the Dual Task promotes rational inference depends on both the LM and the sentence construction. The models can be grouped into three categories as follows.

(i) GPT-4o is likely to use rational inference in the dual task. Four out of six constructions show lower accuracy for implausible sentences in the dual task than in either the single or noisy single tasks.

(ii) o3-mini and o4-mini show a similar tendency but maintain high accuracy across all conditions (around 100%). Four out of six constructions show significantly lower accuracy for implausible sentences in the dual task than in the single or noisy single tasks.

(iii) The other models, i.e., GPT-4.1, DeepSeek-V3, Llama-3.3, and Gemma-3, are likely to rely on rational inference in both the noisy single and dual tasks. These models generally show significantly lower accuracy in the noisy single and dual tasks than in the single task, but no significant difference between the noisy single and dual tasks.

In summary, the results suggest that GPT-4o, o3-mini, and o4-mini are more likely to engage in rational inference when cognitive resources are constrained. A consistent trend is observed when analyzing plausibility effects across different arithmetic problems (see Appendix C) and correct answers (see Section 5.1).

Table 2 shows results from the human experiment. The dual task yields generally lower accuracy and a larger difference between plausible and

Figure 2: Mean accuracy of comprehension tasks by plausibility, task, LM, and construction. GPT-4o is likely to show significantly larger accuracy drops for implausible sentences in the dual task than in the single and noisy single tasks. o3-mini and o4-mini are likely to show a similar pattern but maintain near-ceiling accuracy across conditions. Other models show reduced accuracy in both noisy single and dual tasks, with no clear difference between them. Some conditions are excluded during preprocessing (see Section 4.2.1). $*p < 0.05$. $**p < 0.01$. $***p < 0.001$. Abbreviations: Single = Single Task; Noisy = Noisy Single Task; Dual = Dual Task; Exp.Subj = Experiencer Subject; Exp.Obj. = Experiencer Object; Ben.For = Benefative For.

implausible sentences than the single and noisy single tasks. Specifically, the mean accuracy difference between plausible and implausible conditions is 0.064 in the dual task, compared with 0.025 in the single task and 0.035 in the noisy single task. This pattern suggests increased reliance on plausibility under dual-task conditions in humans. It parallels prior psycholinguistic findings (Ayasse et al., 2021; Ferreira, 2003; Gibson et al., 2013), and patterns observed in our data for several LMs, including GPT-4o, o3-mini, and o4-mini.

	Single	Noisy	Dual
Plausible	0.95 (0.23)	0.88 (0.33)	0.82 (0.39)
Implausible	0.92 (0.27)	0.84 (0.37)	0.75 (0.43)

Table 2: Mean accuracy (with standard deviations) in the human experiment. The dual task shows a generally lower accuracy and a larger plausible–implausible accuracy difference than the single and noisy single tasks. Abbreviations: Single = Single Task; Noisy = Noisy Single Task; Dual = Dual Task.

5 Analysis

5.1 Effects of Plausibility by Correct Answer

Figure 3 presents the mean accuracy rates of comprehension questions by plausibility, task, LM, and correct answer (Yes or No). When the correct answer is "Yes," all models show larger plausibility contrasts under the dual task condition than in the single or noisy single tasks. This effect is driven by a substantial decrease in accuracy for implausible sentences under the dual task. That is, the models often fail to correctly respond "Yes" when asked whether an implausible sentence expressed an implausible meaning, instead responding "No."

On the other hand, when the correct answer is "No," GPT-4o, o3-mini, and o4-mini still exhibit tendencies consistent with rational inference, similar to the Yes condition. Other models do not show such a pattern. In summary, consistent with the results in Section 4.4, GPT-4o, o3-mini, and o4-mini reliably demonstrate a shift toward rational inference across both answer types.

5.2 Conditions Where the Implausible Sentences are Misunderstood

The previous sections show that GPT-4o, o3-mini, and o4-mini tend to shift toward rational inference under the dual task. Here, we examine when these models are most likely to misinterpret implausible sentences as plausible under the dual task. Figure 4 illustrates the proportion of implausible items that are answered correctly in the single and noisy single tasks but incorrectly in the dual task.

Examining the distribution across constructions in Figure 4 (a), GPT-4o, o3-mini, and o4-mini show the highest error rates for dative and benefactive-for constructions. GPT-4o also frequently fails on passive sentences. These constructions share a key property: when function words (and morphemes) are removed, the remaining word sequence appears semantically plausible. For instance, removing function words from "The bartender was blended by the cocktail" and "The chef sent the friend to the gift" yields "bartender blend cocktail" and "chef send friend gift" respectively, which could be interpreted as plausible events. Therefore, this suggests that under resource constraints, these models rely primarily on the superficial word sequence of content words and plausibility derived from world knowledge, rather than function words.

Next, the distribution across the correct answer in Figure 4 (c) shows that errors are more frequent

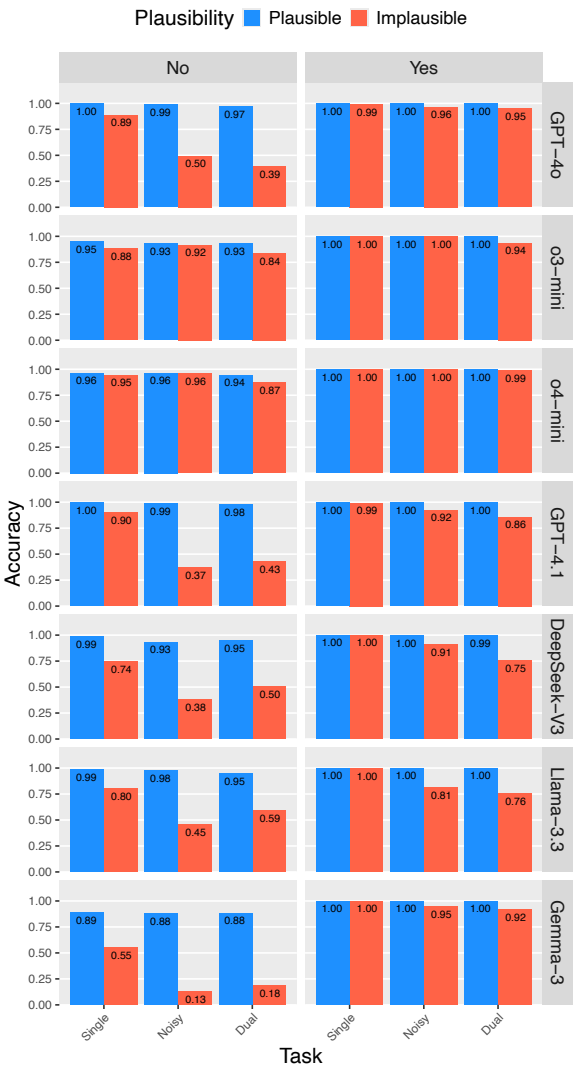


Figure 3: Mean accuracy of comprehension tasks by plausibility, task, LM, and correct answer. When the correct answer is "Yes," all models show larger plausibility contrasts in the dual task, driven by reduced accuracy for implausible sentences. When the correct answer is "No," only GPT-4o, o3-mini, and o4-mini show a similar pattern. Abbreviations: Single = Single Task; Noisy = Noisy Single Task; Dual = Dual Task.

when the correct answer is "No" than when it is "Yes." This indicates that when asked comprehension questions such as "Did the bartender blend the cocktail?" the models tend to respond "Yes," showing a bias toward affirmative answers. This pattern resembles acquiescence, called "yea-saying," observed in humans during Yes/No question answering tasks (Jackson and Messick, 1958; Knowles and Condon, 1999). LM's acquiescence bias has also been observed (Dentella et al., 2023).

Finally, the calculation graph in Figure 4 (b) shows that all three models fail most frequently in

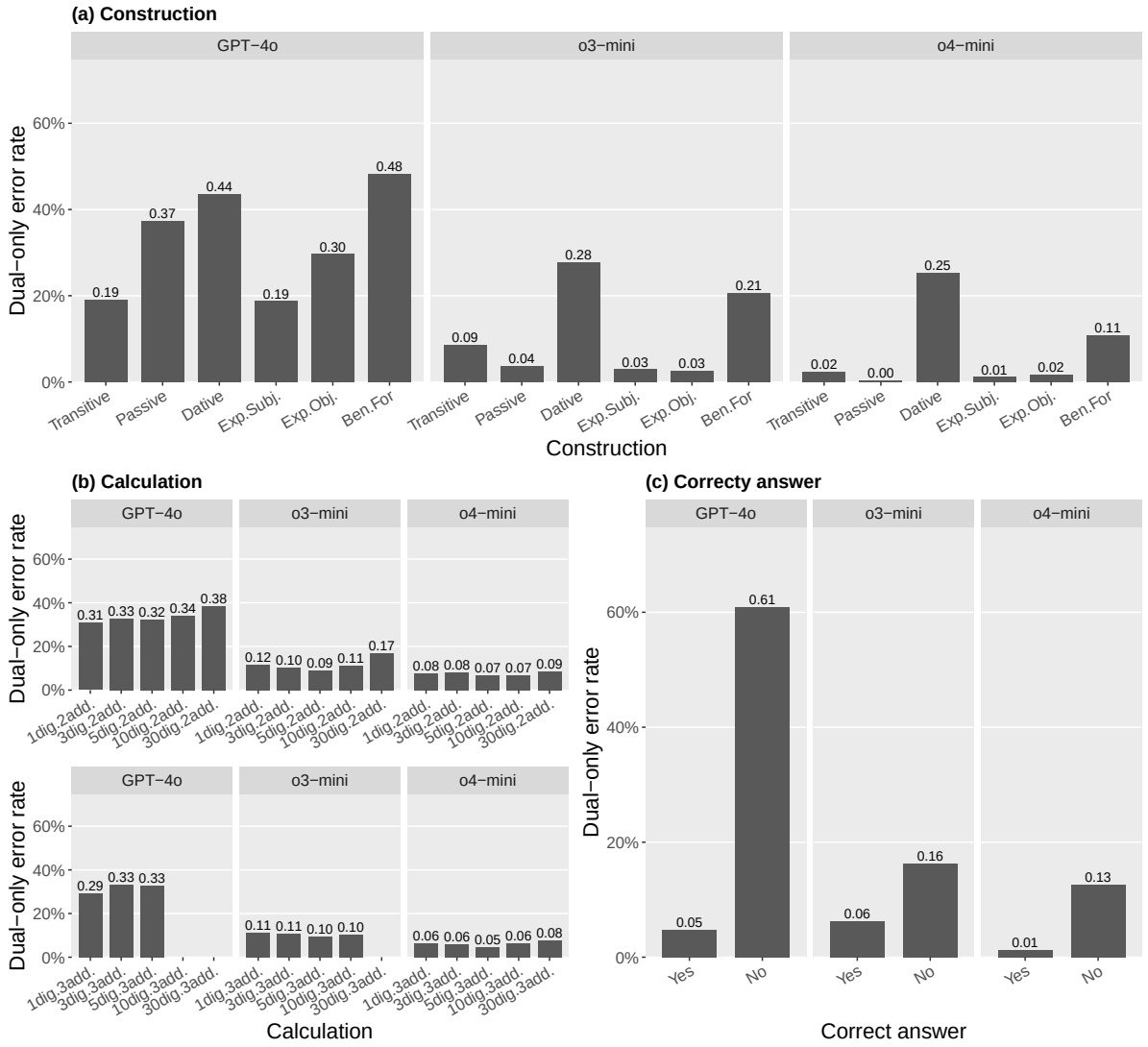


Figure 4: Proportion of implausible items answered correctly in the single and noisy single tasks but incorrectly in the dual task, by construction, arithmetic condition, and correct answer. (a) Errors are most frequent for dative and benefactive-for constructions (and passive for GPT-4o). (b) Errors peak in the 30-digit condition. (c) Errors are more frequent for "No" answers. Some conditions are excluded during preprocessing (see Section 4.2.1).Abbreviations: Exp.Subj = Experiencer Subject; Exp.Obj. = Experiencer Object; Ben.For = Benefative For.

the 30-digit condition. This suggests that cognitive load increases for the models as the numerical magnitude grows.

6 Discussion

6.1 Limitation of Resources Promotes LMs' Rational Inference

Although the patterns varied across LMs, our data demonstrate that several models showed a tendency to adopt more rational comprehension strategies under the dual-task situations. This suggests that one of the reasons for errors that prioritize plausibility information over function words is the reduction of cognitive resources available for sentence comprehension.

Could it be possible that this plausibility-based shift in the dual task is due to jumbled and longer input? We find that dependence on plausibility increased in the dual task condition compared to the noisy single task condition, even though both included identical arithmetic expressions. Thus, the effect cannot be attributed solely to input complexity. Instead, it reflects the additional cognitive demands imposed by performing two tasks simultaneously, which deplete available resources.

Thus, an additional task is key to distinguishing the present study from Asami and Sugawara (2024).

They manipulate memory load by lengthening sentences: while longer sentences reduced overall accuracy, they did not alter the influence of plausibility. Therefore, our results suggest that task-induced cognitive load, rather than input length alone, is a critical factor in constraining LMs' cognitive resources. Human working memory is not simply about storage, but about the dynamic interaction between memory access and processing (Atkinson and Shiffrin, 1971; Baddeley and Hitch, 1974; Baddeley, 2003; Just and Carpenter, 1992). Therefore, our findings suggest that LMs, like humans, rely not only on short- or long-term memory stores, but also on a working-memory-like mechanism that integrates memory with ongoing computation.

Our results also align with findings from reasoning studies. LMs show heuristic reasoning strategies, called "shortcut solutions" (Geirhos et al., 2020; Jia and Liang, 2017; Ko et al., 2020; Tang et al., 2023). They sometimes rely on superficial letter sequences rather than the content of the documents, similar to this study. Furthermore, it has been observed that when cognitive resources become limited, reasoning processing shifts from syntactic interpretation to more superficial comprehension based on word-level associations (Lampinen et al., 2024; Zhang et al., 2024) and degrades some functions, such as safety mechanisms (Upadhayay et al., 2025; Xu et al., 2024). Our work further explores what functions are reduced in sentence comprehension. In particular, our data show that adding an arithmetic computation task depletes the resources and consequently reduces syntactic processing.

6.2 Contributions to Psycholinguistics

Finally, we discuss how our results can contribute to psycholinguistic theories. Our findings suggest that limiting cognitive resources induces a shift toward rational inference, i.e., a human-like comprehension strategy. More broadly, these results support the hypothesis that human-like sentence understanding fundamentally arises from how limited cognitive resources are allocated.

In human sentence comprehension, behavioral effects are often attributed to memory limitations (Gibson, 1998; Van Dyke and Lewis, 2003, and others). Working memory, particularly its temporary storage used for ongoing processing, has long been considered a crucial factor. However, it remains an open question whether human comprehension behavior can truly be explained solely in terms of working memory capacity.

This study provides supporting evidence from non-human systems, namely LMs, that their behavior under constrained conditions resembles human behavior. Namely, under limited cognitive resources, LMs and humans (i) reduced reliance on syntactic function, with increased reliance on their primary world knowledge (Ayasse et al., 2021; Ferreira, 2003; Gibson et al., 2013), and (ii) a bias toward acquiescence (Condon et al., 2006; Knowles and Condon, 1999; Lechner and Rammstedt, 2015; Rammstedt et al., 2023). That is, both humans and LMs degrade certain functions supporting syntactic processing and rejection when cognitive resources are constrained. As a result, they rely more on semantic plausibility.

Taken together, our results suggest that the underlying principle of human sentence comprehension may lie in the resource limitation of working memory, leading to strategies adopted to achieve efficient sentence comprehension (cf. Cognitive Load Theory by Sweller (1988) and Sweller et al. (2019)).

7 Conclusion

We implement a dual-task paradigm in which the models simultaneously solve arithmetic problems and answer comprehension questions. GPT-4o, o3-mini, and o4-mini shift their comprehension strategies under cognitive resource limitations toward rational inference, similar to humans. Our findings suggest that task-induced cognitive load, rather than input length alone, constrains LMs' cognitive resources and makes them more human-like.

Limitations

First, further research is needed to explore what drives LMs' rational reading strategies under dual-task conditions. While some models demonstrate a clear shift toward plausibility-based comprehension, others do not. Notably, even within the GPT-4 family, GPT-4o exhibits rational inference behavior, whereas GPT-4.1 does not, indicating that identifying the source of this difference is not straightforward. It remains unclear whether these variations arise from differences in internal architecture, training procedures, or other aspects of the model.[7]

[7]We examined whether LMs acquire a rational inference strategy during training using OLMo (Groeneveld et al., 2024), which provides access to both training data and intermediate training checkpoints. However, even the final models fail to solve the math problems in the dual-task condition.

Recent studies have proposed that attention mechanisms in LMs serve as a working-memory–like component in the context of modeling sentence processing costs (Ryu and Lewis, 2021; Timkey and Linzen, 2023; Yoshida et al., 2025). Thus, future research could extend this line of inquiry to comprehension strategies, potentially providing new insights into human working memory processes during comprehension.

The second limitation of this study is that the results may depend on the choice of baseline and prompt design. Although our Noisy Single Task is designed to isolate the effect of adding a calculation task by comparing the Dual Task, other baselines or formatting choices may yield different outcomes. This is relevant because large LMs are known to be sensitive to prompt formulation (Kojima et al., 2022; Schmidt et al., 2024; Sclar et al., 2024). Thus, systematically varying prompts, including non-semantic interruptions such as delimiters or formatting tokens instead of numerics, would be an important direction for future work.

Finally, direct human–LM comparisons would be highly valuable for more detailed modeling of human sentence comprehension. Differences between humans and LMs may be sensitive to task design or stimulus properties. Importantly, establishing a robust and well-characterized dual-task paradigm for LMs is a necessary prerequisite for meaningful human–LM comparison, as premature comparisons risk conflating methodological artifacts with cognitive effects. Our contribution should therefore be viewed as a first step: introducing a dual-task paradigm for LMs and demonstrating its potential to reveal rational inference behavior.

Taken together, future work should therefore enhance the dual-task paradigm by exploring alternative baselines and prompt designs. After establishing robust experimental settings, subsequent work can pursue more detailed human–LM comparisons and deeper investigation into the internal mechanisms underlying working memory in LMs.

Ethical Considerations

We use crowdsourcing in our human data collection. Before crowdworkers accept participation, we inform them of the purpose of the study, the tasks, the time required, the risks, the voluntary nature, the compensation, and the privacy considerations. There are no known serious risks while participating in the experiment. Participants may experience mild fatigue from reading text on a screen for an extended period. We therefore allow them to take breaks at any time. The experimental materials do not contain any offensive content. We do not obtain any personally identifiable information about participants, except for Prolific ID (used only for compensation payments), age, and sex. This experiment was approved by the ethics committee of the author's institution (National Institute of Informatics).

Acknowledgments

We would like to thank the anonymous ARR reviewers for their valuable comments. This study was supported by JSPS KAKENHI (No. JP23KJ0199, JP25K21281), JST BOOST (No. JPMJBY24D9), and JST FOREST (No. JPMJFR232R).

References

Samuel Joseph Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2025a. Comparing human and language models sentence processing difficulties on complex structures. *Preprint*, arXiv:2510.07141.

Samuel Joseph Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2025b. When the lm misunderstood the human chuckled: Analyzing garden path effects in humans and language models. *Preprint*, arXiv:2502.09307.

Daiki Asami and Saku Sugawara. 2024. What makes language models good-enough? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15453–15467, Bangkok, Thailand. Association for Computational Linguistics.

Richard C. Atkinson and Richard M. Shiffrin. 1971. The control of short-term memory. *Scientific American*, 225(2):82–91.

Nicolai D. Ayasse, Alana J. Hodson, and Arthur Wingfield. 2021. The principle of least effort and comprehension of spoken sentences by younger and older adults. *Frontiers in Psychology*, Volume 12 - 2021.

Alan Baddeley. 2003. Working memory and language: An overview. *Journal of communication disorders*, 36(3):189–208.

Alan D. Baddeley and Graham Hitch. 1974. Working memory. volume 8 of *Psychology of Learning and Motivation*, pages 47–89. Academic Press.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model APIs. In *Proceedings of the 2023 Conference*

on Empirical Methods in Natural Language Processing: Industry Track, pages 792–810, Singapore. Association for Computational Linguistics.

Kiel Christianson. 2016. When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, 69(5):817–828. PMID: 26785102.

Lorena Condon, Pere J. Ferrando, and Josep Demestre. 2006. A note on some item characteristics related to acquiescent responding. *Personality and Individual Differences*, 40(3):403–407.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Fernanda Ferreira. 2003. The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2):164–203.

Fernanda Ferreira and Nikole D Patson. 2007. The 'good enough' approach to language comprehension. *Language and linguistics compass*, 1(1-2):71–83.

Richard Futrell and Edward Gibson. 2017. L2 processing as noisy channel language comprehension. *Bilingualism: Language and Cognition*, 20(4):683–684.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Edward Gibson, Chaleece Sandberg, Evelina Fedorenko, Leon Bergen, and Swathi Kiran. 2016. A rational inference approach to aphasic language comprehension. *Aphasiology*, 30(11):1341–1360.

Dongyu Gong, Xingchen Wan, and Dingmin Wang. 2024. Working memory capacity of chatgpt: An empirical study. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10048–10056.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Tovah Irwin, Kyra Wilson, and Alec Marantz. 2023. BERT shows garden path effects. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3220–3232, Dubrovnik, Croatia. Association for Computational Linguistics.

Douglas N Jackson and Samuel Messick. 1958. Content and style in personality assessment. *Psychological bulletin*, 55(4):243.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Marcel A Just and Patricia A Carpenter. 1992. A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1):122.

Wayne K Kirchner. 1958. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352.

Eric S Knowles and Christopher A Condon. 1999. Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77(2):379.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first

sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Clemens M Lechner and Beatrice Rammstedt. 2015. Cognitive ability, acquiescence, and the structure of personality in a sample of older adults. *Psychological assessment*, 27(4):1301.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Beatrice Rammstedt, Lena Roemer, and Clemens M. Lechner. 2023. Do simpler item wording and response scales reduce acquiescence in personality inventories? a survey experiment. *Personality and Individual Differences*, 214:112324.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Anthony J Sanford and Patrick Sturt. 2002. Depth of processing in language comprehension: Not noticing the evidence. *Trends in cognitive sciences*, 6(9):382–386.

Douglas C. Schmidt, Jesse Spencer-Smith, Quchen Fu, and Jules White. 2024. Towards a catalog of prompt patterns to enhance the discipline of prompt engineering. *Ada Lett.*, 43(2):43–51.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Guijin Son, SangWon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. 2024. Multi-task inference: Can large language models follow multiple instructions at once? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5606–5627, Bangkok, Thailand. Association for Computational Linguistics.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

John Sweller, Jeroen JG Van Merriënboer, and Fred Paas. 2019. Cognitive architecture and instructional design: 20 years later. *Educational psychology review*, 31(2):261–292.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720, Singapore. Association for Computational Linguistics.

Marilyn L Turner and Randall W Engle. 1989. Is working memory capacity task dependent? *Journal of Memory and Language*, 28(2):127–154.

Bibek Upadhayay, Vahid Behzadan, and Amin Karbasi. 2025. Working memory attack on LLMs. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Julie A Van Dyke and Richard L Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316.

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3526–3548, Mexico City, Mexico. Association for Computational Linguistics.

Ryo Yoshida, Shinnosuke Isono, Kohei Kajikawa, Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2025. If attention serves as a cognitive model of human memory retrieval, what is the plausible memory representation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 9795–9812. Association for Computational Linguistics.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. 2024. Working memory identifies reasoning limits in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16896–16922, Miami, Florida, USA. Association for Computational Linguistics.

A Example Stimuli for the Noisy Single Task and the Dual Task

Table 3 represents example stimuli for the noisy single task and the dual task.

B Prompts

The method to make prompts is as follows. We put instructions of tasks and few-shot examples on the system message, and a sentence and a question on the user message. The few-shot example contains four sets of sentence and question, each of these are plausible or implausible, and question whose answer is Yes or No. Examples of prompts are in Figures 5–7.

Single Task

System Message:

```
Please read the sentence. Once the sentence
ends, answer the yes/no question about the
sentence. The question relates to the
sentence and requires a yes or no response.
Return the question answers as [Yes] or [No].
The following are examples of sentences and
questions about it.

Sentence: The boy is eating an apple and a
banana in his house.
Question: Is the boy eating an apple?
Answer: [Yes]

Sentence: The teacher put a cup on the table
and read his favorite book while listening
to music.
Question: Did a book read the teacher?
Answer: [No]

Context: The dog was chased by the man before
the woman saw the pianist.
Question: Was the man chased by the dog?
Answer: [No]

Context: The refrigerator painted the
carpenter while the mountain laughed at the
sandwich and the sky was green.
Question: Did the refrigerator paint the
carpenter?
Answer: [Yes]
```

User Message:

```
Sentence: The bartender was blended by the
cocktail.
Question: Did the bartender blend the
cocktail?
Answer:
```

Figure 5: Example prompt of the single task.

C Effect of Plausibility by Calculation

Figure 8 represents the mean accuracy rates of comprehension questions by plausibility, LM, and calculation. The graph shows the consistent results with the results by construction and correct answer (see Figure 2 and 3): GPT-4o, o3-mini, and o4-mini have the greater difference between plausible and implausible conditions in the dual task than the single task or the noisy single task. DeepSeek-V3 also represents this trend.

D Accuracy for Calulation

The accuracy of arithmetic problems is displayed in Table 4. The table shows that most of the LMs have around 90% accuracy except for Llama-3.3, which has below 80% in any conditions.

E Human Experiment Procedure

Procedures are almost the same as LM's experiment. We conducted the experiment using

Calculation	Stimuli
1dig.2add.	The 5 cocktail + blended 6 the = bartender x5633 and 9 the + authorities 3 agitated = the x5634 organist 6 after + the 8 infantryman = saluted x5635 the 3 pollster.
1dig.3add.	The 7 cocktail + blended 6 the + bartender 5 and = the x4210 authorities 2 agitated + the 4 organist + after 4 the = infantryman x4211 saluted 4 the + pollster.
3dig.2add.	The 952 cocktail + blended 604 the = bartender x5633 and 793 the + authorities 271 agitated = the x5634 organist 770 after + the 832 infantryman = saluted x5635 the 531 pollster.
3dig.3add.	The 212 cocktail + blended 260 the + bartender 341 and = the x4210 authorities 220 agitated + the 631 organist + after 753 the = infantryman x4211 saluted 264 the + pollster.

Table 3: Example stimuli by arithmetic problem. Context stimuli in the noisy single task and the dual task embed arithmetic problems in sentences.

PCIbex Farm (https://farm.pcibex.net/). Instructions for each task are shown in Figures 9, 10, and 11. A task is specified first. Specifically, the instruction "Read the sentence," "Read the Sentence and IGNORE the Math Problem," or "Read the Sentence and CALCULATE the Math Problem" appeared on the screen in the single-tasks, noisy-single-tasks, or dual-tasks, respectively. Then, participants read a sentence word by word. Only in the dual task, type the answer for the arithmetic problem once x1, x2, x3... appears. When the sentence presentation ends, participants answer the comprehension question. Before the experiment, participants practice the task for four trials.

Noisy Single Task

System Message:

```
The sentence combines a math problem and a text alternating one word at a time. The math problem is split and placed after each word in the sentence, such as 2 + 8 = x1. Please read the sentence while ignoring the math problems. Once the sentence ends, answer the yes/no question about the sentence. The question relates to the sentence and requires a yes or no response. Return the question answers as [Yes] or [No]. The following are examples of sentences with math problems and a yes/no question about it.

Sentence: The 2 boy + is 8 eating = an x1 apple 1 and + a 5 banana = in x2 his 3 house.
Question: Is the boy eating an apple?
Answer: [Yes]

Sentence: The 4 teacher + put 9 a + cup 7 on = the x3 table 1 and + read 4 his + favorite 6 book = while x4 listening 5 to + music.
Question: Did a book read the teacher?
Answer: [No]

Context: The 5 dog + was 5 chased = by x6 the 2 man + before 7 the = woman x7 saw 3 the + pianist.
Question: Was the man chased by the dog?
Answer: [No]

Context: The 3 refrigerator + painted 6 the + carpenter 2 while = the x9 mountain 8 laughed + at 1 the + sandwich 7 and = the x10 sky 1 was + green.
Question: Did the refrigerator paint the carpenter?"
Answer: [Yes]
```

User Message:

```
Sentence: The 5 cocktail + blended 6 the = bartender x5633 and 9 the + authorities 3 agitated = the x5634 organist 6 after + the 8 infantryman = saluted x5635 the 3 pollster.
Question: Did the bartender blend the cocktail?
Answer:
```

Figure 6: Example prompt of the noisy single task.

Dual Task

System Message:

```
The sentence combines a math problem and a text alternating one word at a time. The math problem is split and placed after each word in the sentence, such as 2 + 8 = x1. Please read the sentence while accurately calculating the math problems. When x1, x2, x3... appear, output the answer to the math problem carefully. Once the sentence ends, answer the yes/no question about the sentence. The question relates to the sentence and requires a yes or no response. However, prioritize the quality of solving the math problems over answering the sentence. Ensure all math problems are solved correctly, with each one being an addition of two or three numbers. Return the math problems and their answers as tuples, e.g., (x1, 2 + 8, 10), (x3, 4 + 9 + 7, 20). Return the question answers as [Yes] or [No]. The following are examples of sentences with math problems and a yes/no question about it.

Sentence: The 2 boy + is 8 eating = an x1 apple 1 and + a 5 banana = in x2 his 3 house.
Question: Is the boy eating an apple?
Answer for the math problem: (x1, 2 + 8, 10), (x2, 1 + 5, 6)
Answer for the question: [Yes]

Sentence: The 4 teacher + put 9 a + cup 7 on = the x3 table 1 and + read 4 his + favorite 6 book = while x4 listening 5 to + music.
Question: Did a book read the teacher?
Answer for the math problem: (x3, 4 + 9 + 7, 20), (x4, 1 + 4 + 6, 11)
Answer for the question: [No]

Context: The 5 dog + was 5 chased = by x6 the 2 man + before 7 the = woman x7 saw 3 the + pianist.
Question: Was the man chased by the dog?
Answer for the math problem: (x6, 5 + 5, 10), (x7, 2 + 7, 9)
Answer for the question: [No]

Context: The 3 refrigerator + painted 6 the = carpenter x9 while 8 the + mountain 1 laughed = at x10 the 7 sandwich + and 9 the = sky x11 was 2 green.
Question: Did the refrigerator paint the carpenter?
Answer for the math problem: (x9, 3 + 6, 9), (x10, 8 + 1, 9), (x11, 7 + 9, 16)
Answer for the question: [Yes]
```

User Message:

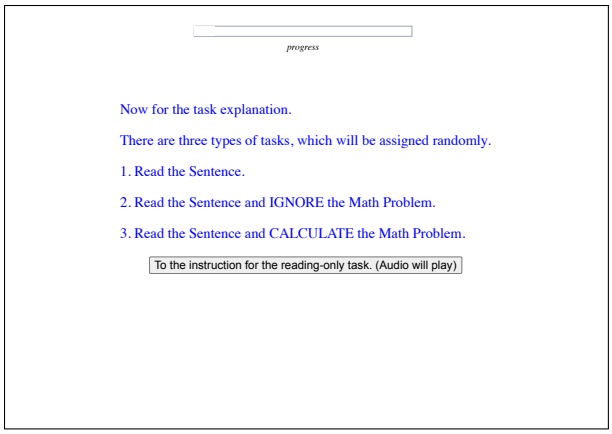
```
Sentence: The 5 cocktail + blended 6 the = bartender x5633 and 9 the + authorities 3 agitated = the x5634 organist 6 after + the 8 infantryman = saluted x5635 the 3 pollster.
Question: Did the bartender blend the cocktail?
Answer for the math problem:
Answer for the question:
```

Figure 7: Example prompt of the dual task.

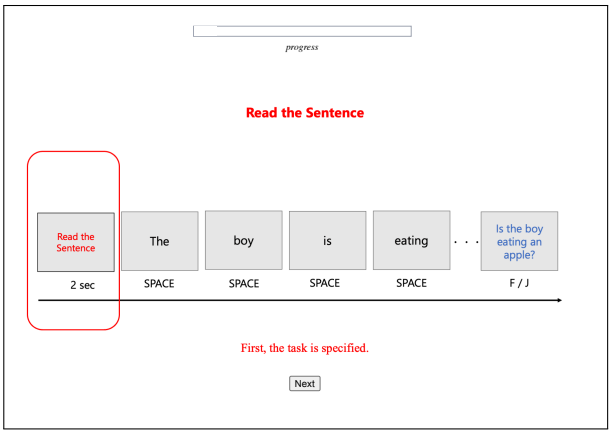
Model	Calculation	Mean	SD
GPT-4o	1dig.2add.	0.99	0.12
GPT-4o	1dig.3add.	1.00	0.00
GPT-4o	3dig.2add.	0.99	0.11
GPT-4o	3dig.3add.	1.00	0.00
GPT-4o	5dig.2add.	0.99	0.10
GPT-4o	5dig.3add.	0.99	0.11
GPT-4o	10dig.2add.	0.92	0.26
GPT-4o	10dig.3add.	0.50	0.50
GPT-4o	30dig.2add.	0.77	0.42
GPT-4o	30dig.3add.	0.04	0.19
o3-mini	1dig.2add.	0.97	0.17
o3-mini	1dig.3add.	0.98	0.14
o3-mini	3dig.2add.	0.95	0.22
o3-mini	3dig.3add.	0.98	0.14
o3-mini	5dig.2add.	0.95	0.22
o3-mini	5dig.3add.	0.97	0.17
o3-mini	10dig.2add.	0.93	0.25
o3-mini	10dig.3add.	0.93	0.25
o3-mini	30dig.2add.	0.71	0.45
o3-mini	30dig.3add.	0.58	0.49
o4-mini	1dig.2add.	0.97	0.16
o4-mini	1dig.3add.	0.95	0.21
o4-mini	3dig.2add.	0.97	0.17
o4-mini	3dig.3add.	0.94	0.23
o4-mini	5dig.2add.	0.96	0.20
o4-mini	5dig.3add.	0.93	0.25
o4-mini	10dig.2add.	0.94	0.24
o4-mini	10dig.3add.	0.91	0.29
o4-mini	30dig.2add.	0.82	0.38
o4-mini	30dig.3add.	0.76	0.43
GPT-4.1	1dig.2add.	1.00	0.00
GPT-4.1	1dig.3add.	1.00	0.05
GPT-4.1	3dig.2add.	1.00	0.00
GPT-4.1	3dig.3add.	1.00	0.01
GPT-4.1	5dig.2add.	1.00	0.04
GPT-4.1	5dig.3add.	0.99	0.11
GPT-4.1	10dig.2add.	0.92	0.28
GPT-4.1	10dig.3add.	0.12	0.33
GPT-4.1	30dig.2add.	0.64	0.48
GPT-4.1	30dig.3add.	0.03	0.18
DeepSeek-V3	1dig.2add.	0.99	0.11
DeepSeek-V3	1dig.3add.	0.99	0.09
DeepSeek-V3	3dig.2add.	0.99	0.12
DeepSeek-V3	3dig.3add.	0.99	0.08
DeepSeek-V3	5dig.2add.	0.97	0.17
DeepSeek-V3	5dig.3add.	0.99	0.10
DeepSeek-V3	10dig.2add.	0.90	0.30
DeepSeek-V3	10dig.3add.	0.84	0.37
DeepSeek-V3	30dig.2add.	0.91	0.29
DeepSeek-V3	30dig.3add.	0.58	0.49
Llama-3.3	1dig.2add.	0.82	0.38
Llama-3.3	1dig.3add.	0.73	0.44
Llama-3.3	3dig.2add.	0.76	0.43
Llama-3.3	3dig.3add.	0.50	0.50
Llama-3.3	5dig.2add.	0.78	0.42
Llama-3.3	5dig.3add.	0.42	0.49
Llama-3.3	10dig.2add.	0.66	0.47
Llama-3.3	10dig.3add.	0.26	0.44
Llama-3.3	30dig.2add.	0.24	0.43
Llama-3.3	30dig.3add.	0.01	0.08
Gemma-3	1dig.2add.	0.89	0.32
Gemma-3	1dig.3add.	0.82	0.39
Gemma-3	3dig.2add.	0.93	0.26
Gemma-3	3dig.3add.	0.83	0.38
Gemma-3	5dig.2add.	0.81	0.39
Gemma-3	5dig.3add.	0.72	0.45
Gemma-3	10dig.2add.	0.56	0.50
Gemma-3	10dig.3add.	0.44	0.50
Gemma-3	30dig.2add.	0.18	0.38
Gemma-3	30dig.3add.	0.00	0.00

Table 4: Accuracy of arithmetic problems. Some conditions are excluded during preprocessing (see Section 4.2.1). SD = standard deviation.

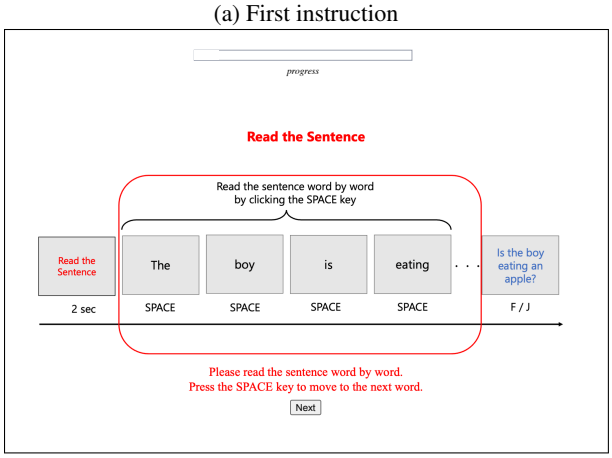
Figure 8: Accuracy rate of comprehension tasks by plausibility, task, LM, and calculation. Some conditions are excluded during preprocessing (see Section 4.2.1). *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.



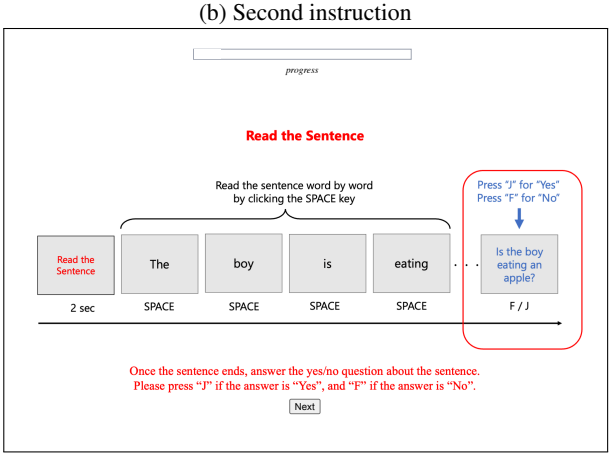
(a) First instruction



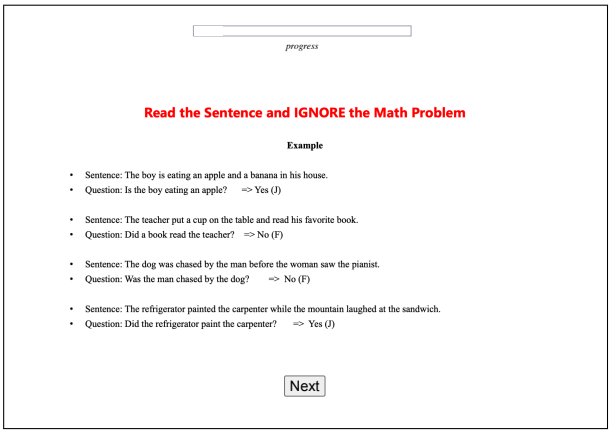
(b) Second instruction



(c) Third instruction

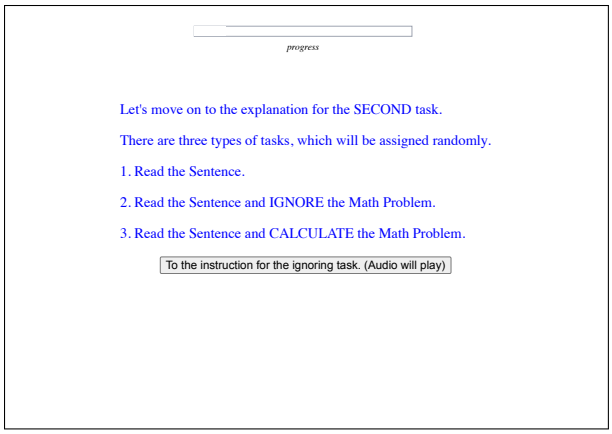


(d) Fourth instruction

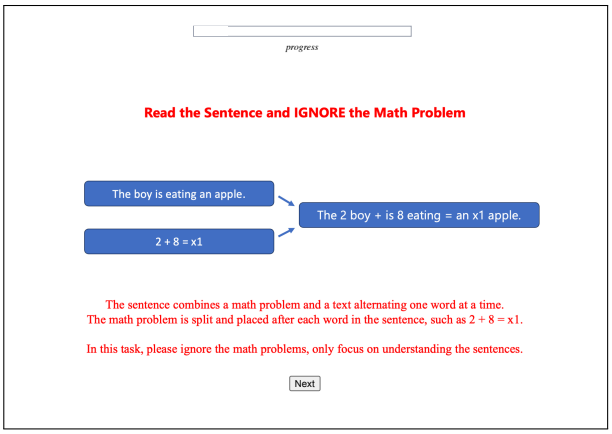


(e) Fifth instruction

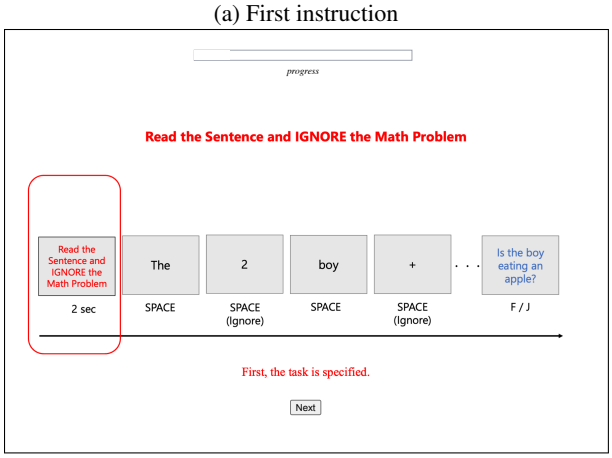
Figure 9: Instruction screens for the single task in the human experiment.



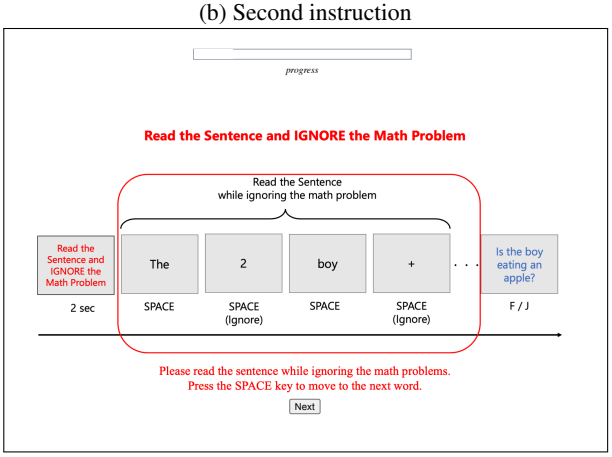
(a) First instruction



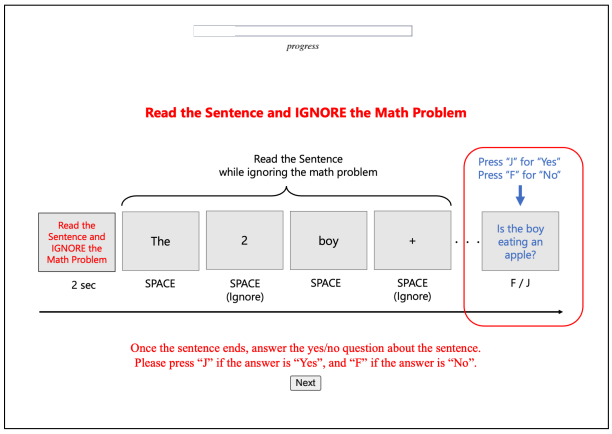
(b) Second instruction



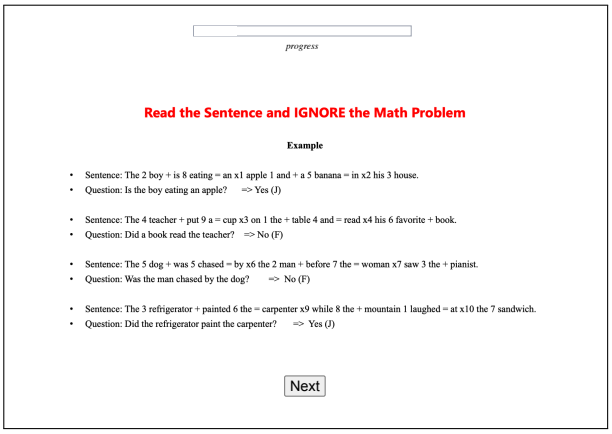
(c) Third instruction



(d) Fourth instruction

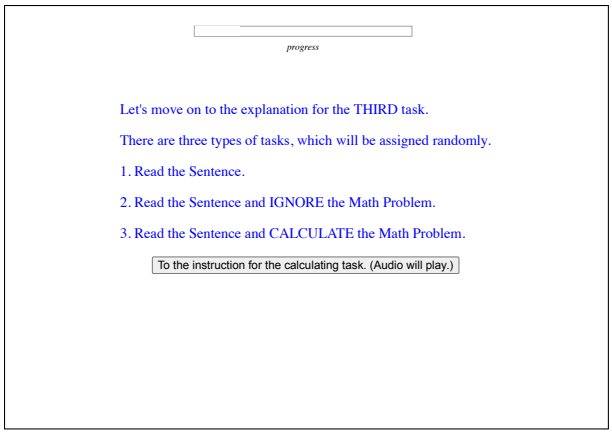


(e) Fifth instruction

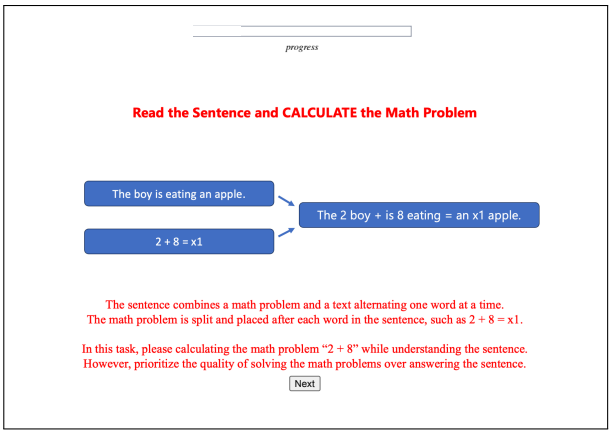


(f) Sixth instruction

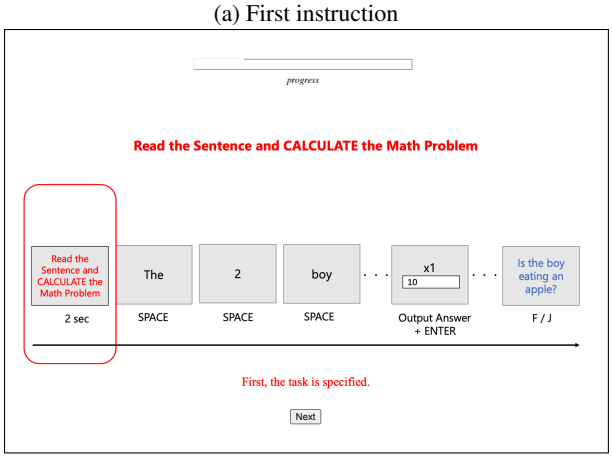
Figure 10: Instruction screens for the noisy single task in the human experiment.



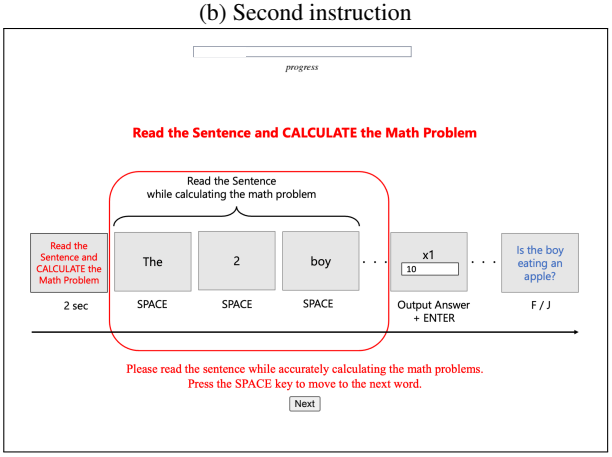
(a) First instruction



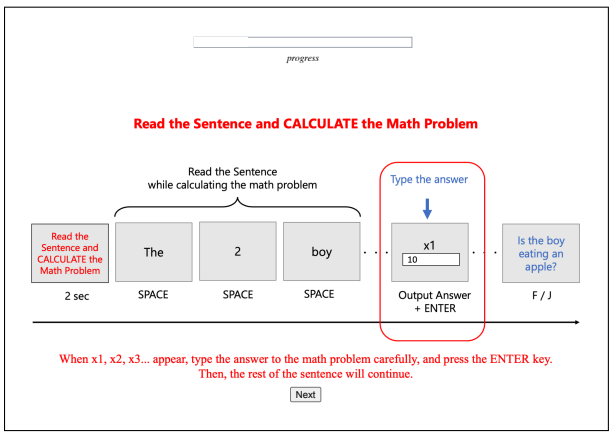
(b) Second instruction



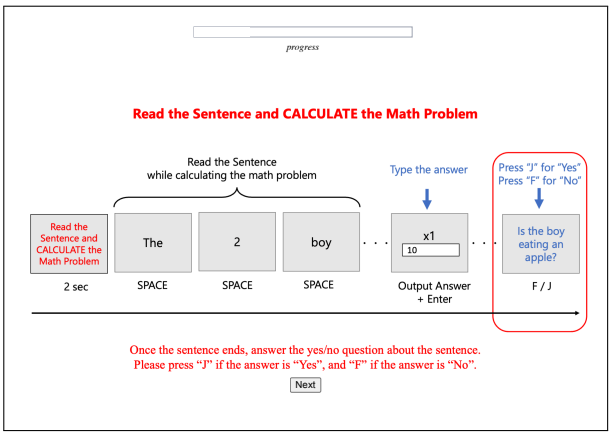
(c) Third instruction



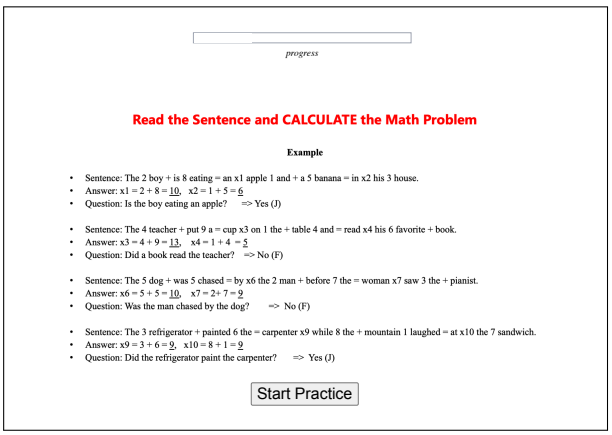
(d) Fourth instruction



(e) Fifth instruction



(f) Sixth instruction



(g) Seventh instruction

Figure 11: Instruction screens for the dual task in the human experiment.