Intuitive or Dependent? Investigating LLMs' Behavior Style to Conflicting Prompts

Jiahao Ying[1] , Yixin Cao[1*] , Kai Xiong [2]
Yidong He[3], Long Cui[3], Yongbin Liu[3]
[1]Singapore Management University, Singapore
[2]Harbin Institute of Technology, China
[3] University of South China, China

{jhying.2022}@phdcs.smu.edu.sg

Abstract

This study investigates the behaviors of Large Language Models (LLMs) when faced with conflicting prompts versus their internal memory. This will not only help to understand LLMs' decision mechanism but also benefit real-world applications, such as retrieval-augmented generation (RAG). Drawing on cognitive theory, we target the first scenario of decision-making styles where there is no superiority in the conflict and categorize LLMs' preference into dependent, intuitive, and rational/irrational styles. Another scenario of factual robustness considers the correctness of prompt and memory in knowledge-intensive tasks, which can also distinguish if LLMs behave rationally or irrationally in the first scenario. To quantify them, we establish a complete benchmarking framework including a dataset, a robustness evaluation pipeline, and corresponding metrics. Extensive experiments with seven LLMs reveal their varying behaviors. And, with role play intervention, we can change the styles, but different models present distinct adaptivity and upper-bound. One of our key takeaways is to optimize models or the prompts according to the identified style. For instance, RAG models with high role play adaptability may dynamically adjust the interventions according to the quality of retrieval results — being dependent to better leverage informative context; and, being intuitive when the external prompt is noisy. Our dataset can be found at https://github.com/yingjiahao14/KRE.

1 Introduction

Large language models (LLMs) have become fundamental tools in the area of natural language processing (Wei et al., 2022; Mirowski et al., 2023). They can solve various tasks in a unified form of text generation simply by providing specific prompts (Mishra et al., 2022). However, LLMs sometimes fail to follow given prompts, especially when the prompt conflicts with the model's parametric knowledge, a.k.a., the internal memory (McKenzie et al., 2022). This raises an interesting question: how do LLMs behave in such conflicting situations? As shown in Figure 1, the LLM is pre-trained with $\pi \approx 3.14$. If the user re-defines it as another value in prompt, will the answer about π follow memory or user context? The investigation will not only improve the understanding of LLMs' decision mechanism, but also benefit real-world applications. Take retrieval-augmented generation (RAG (Nakano et al., 2021)) as an example, if the external source contains the latest information, which may be different from the model's out-dated memory, it is better to choose a LLM favor prompts; on the contrary, if the retriever performs poorly, the LLM sticking to its own memory may be more robust to the noisy retrieval results.

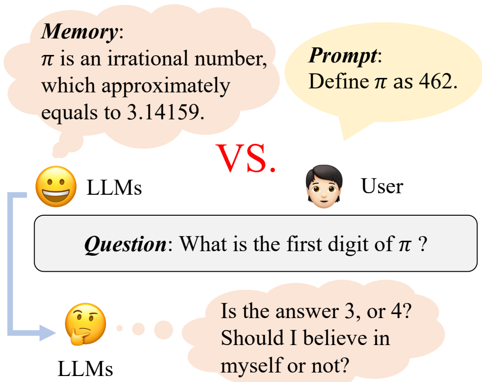

Figure 1: In conflict situation , LLMs may depend on the prompt or intuitively answer based on memory.

In this paper, we investigate the behaviors of LLMs in two types of conflicting scenarios, incorporating the concepts from cognitive theory (Harren, 1979; Phillips et al., 1984). First, we consider the model's preference without any superiority in the conflict. We define three types of **decision styles**: 1) *dependent style* refers to the model steerable to input context; 2) *intuitive style* denotes the

* Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4221–4246
August 11-16, 2024 ©2024 Association for Computational Linguistics

model behavior of sticking to parametric knowledge; and 3) *rational/irrational style* means wavering between the above two directions. Clearly, models with consistent preference (dependent and intuitive styles) are generally more predictable and can be regarded robust to inputs, while rational/irrational style are non-robust unless one direction is better than the other, which will be our second scenario: **factual robustness**. Here, we consider knowledge intensive tasks, e.g., RAG, where LLMs are expected to be robust to noisy input or out-dated memory. That is, this scenario targets the ability of LLMs in discerning facts from conflicts. Note that a high factual robustness shall refer to the rational style if we follow the definition in our first scenario, otherwise irrational style.

To this end, we establish a complete benchmarking framework including a dataset, a robustness evaluation pipeline, and corresponding metrics. For the dataset, to ease the measurement, we modify existing knowledge-intensive datasets and standardize a unified form of Multi-Choice Questions (MCQ). For the evaluation pipeline and metrics, we start with the scenario of factual robustness for clarity. We first deploy memory assessment to identify their memorized fact and then give the model conflict prompt, where the prompt supports one answer, but the model's memory advocates a different one. Under this setting, we design two metrics: vulnerable robustness (VR) and resilient robustness (RR), regarding where the correct fact locates. Then, we ignore the correctness of choices to design a metric, decision-making style score (DMSS), to categorize the models' decision styles.

Based on the above investigation, we further attempt role play to intervene LLMs' styles, which is widely used in practice. It is to instruct LLMs with specific role description, e.g., *You are an intuitive decision maker*. Extensive experiments on seven LLMs reveal their varying behaviors in factual robustness, decision style, as well as the adaptivity and upper-bound of role play intervention. We have found that: **(1)** Compared with utilizing correct prompted knowledge, LLMs are more vulnerable to misleading prompts, thus enhancing robustness against noisy or fake prompts will be a pivotal focus in future research (Sec 4.1). **(2)** LLMs are more robust in using factual knowledge than commonsense knowledge via prompts. This suggests that we can leverage the retrieval-then-prompt strategy to remedy factual flaws while enhancing LLMs' inherent factual reasoning ability (Sec 4.1). **(3)** Detailed instructions are not magic. Indeed, more sophisticated prompts can help the model in ignoring misleading information in the context. However, this strategy also tends to result in an increased number of invalid responses, as explored in Section 4.2. **(4)** Medium-sized LLMs with instruction tuning tend to exhibit a dependent decision-making style, relying more on external prompts. Compared with them, GPT-4 and Bard are rational styles, considering both memory and prompt. We think that scaling up the model size may enhance memory retention while preserving its ability to follow instructions (Sec 4.4). **(5)** We can change LLMs' preference through role play intervention, while different LLMs vary a lot in adaptivity. Notably, although GPT-4 demonstrates the best performance and LLaMA2 is competitive in some aspects, the adaptivity score reveals their large gap (Sec 4.5). These findings suggest a takeaway that, when handling conflict, it is better to optimize either the model or prompts according to the style. For instance, if your model shows a good role play intervention adaptability, a better strategy is to dynamically adjust role's instruction according to the confidence of external information (e.g., retrieval performance) — instructing LLM to be depedent, when external information is probably correct; and, instructing LLM to be intuitive, when external information is probably incorrect.

Our contributions can be summarized as follows:

- To our knowledge, we are the first to quantify LLMs' behavior style in handling conflicts.
- We design a conflicting benchmarking framework including dataset, evaluation pipeline, and measurements.
- Our extensive experiments offer insightful findings for practical applications.

2 Dataset Construction

Our investigation involves two scenarios, each with several styles, as well as role play intervention. For simplicity, we curate one conflicting dataset for both factual robustness and decision style evaluation. In specific, we formulate conflicting cases from existing machine reasoning comprehension (MRC) and commonsense reasoning (CR) datasets, which include factual knowledge and commonsense knowledge, respectively. We thus call our dataset KRE (**K**nowledge **R**obustness **E**valuation dataset). Our pipeline can be easily extended to other tasks. Formally, each sample

$s = (x, a_{gol}, c^+)$ in existing datasets consists of a question x, a golden answer a_{gol} pair, and a positive context c^+ containing the answer. Each sample in KRE is $s' = (x, a_{gol}, c^+, a_{neg}, c^-)$, where we need to generate other options as MCQ, which include a conflicting answer a_{neg} and context c^- (marked as negatives opposite to positive). Next, we describe three construction steps as follows.

Dataset Filtering. We selected four publicly available datasets for extension: two MRC datasets, MuSiQue (Trivedi et al., 2022) and SQuAD v2.0 (Rajpurkar et al., 2018), as well as two CR datasets, ECQA (Aggarwal et al., 2021) and e-CARE (Du et al., 2022). We take the MRC paragraph and CR explanation as golden context. Our filtering process retained only those answerable validation examples from MRC where the context contains sufficient information to derive the answer. The KRE dataset comprises a total of 11,684 test samples. More detailed statistics are in Table 7.

Conflict Generation. We generate the negative answer and context, as well as another two misleading options. For options, we use the existing misleading options in CR datasets and generate them for MRC via ChatGPT (Details can be found in Appendix B.1.1). Subsequently, we randomly choose one misleading option as the negative answer (a_{neg}) and employ ChatGPT to generate a negative context. Specifically, for SQuAD and MuSiQue, we substitute the golden answer entity in the gold context with the negative answer (Appendix B.1.2). For ECQA and e-CARE, we create an explanation for the negative answer.

Instruction Design. Different prompts may lead to different results (Shi et al., 2023). To minimize such influence, we try our best to design an instruction set for KRE and choose the best performed one. There are two types of instructions: (1) **Instruction without hint** asks the model to answer the question without any guidance on handling the context. (2) **Instruction with hint** alerts LLMs of the potential presence of misleading information in the context, advising them to assess before responding. For each kind of instruction, we engage four senior language researchers to draft a total of 12 distinct instructions. To further enhance the diversity of the instructions, we randomly ask ChatGPT, GPT-4 (OpenAI, 2023), Claude (Anthropic, 2023), to rephrase the instruction, generating variants. Consequently, we amass a pool of 24 unique instructions (Appendix B.2 and B.3).

Human Evaluation. We engage four senior language researchers to evaluate 400 randomly selected samples from KRE. The evaluators assess the degree to which the negative context influenced the selection of the negative answer option. The principles and criteria for labeling are standardized across all evaluators. The result shows that more than 98% of the sampled negative context is misleading, with an inter-evaluator agreement rate exceeding 90%. All results are in Appendix A.4.

3 Evaluation Pipeline

Our proposed evaluation pipeline aims to assess LLMs' behavior in two scenarios when facing conflict. As shown in Figure 2, there are five steps, where step 1 to 3 refers to the factual robustness scenario: (1) **Memory Assessment** (Sec 3.1) partitions our dataset into two subsets based on whether LLMs can accurately answer the question without external information. (2) **Factual Robustness Evaluation** (Sec 3.2) targets to which extent LLMs can discern the correct fact from conflict. (3) **Influence of Few-shot Example** (Sec 3.3) further considers the impacts of few-shot in-context learning (ICL), complementary to the above zero-shot settings. Based on the above results, (4) **Decision-Making Style Analysis** (Sec 3.4) explore another scenario of decision styles irrespective of answer correctness. Finally, (5) **Role Play Intervention and Leaderboard** (Sec 3.5) focuses on controlling LLMs' style and build a leaderboard to summarize all measurements.

3.1 Memory Assessment

We partition datasets into two subsets according to LLMs' memory. One contains all questions that LLMs can answer accurately without external information, another contains all questions that LLMs cannot. To assess memory, there are typically two methods: one analyzes the models' performance on the pre-training corpus, assuming that any text occurring in the corpus shall be memorized. The second leverages question-answering tasks to probe memorized knowledge. If a model can answer the question correctly, it memorized related knowledge; otherwise not. We adopt the QA method (under zero-shot) because it conforms to our evaluation pipeline well and is universally applicable to both open-sourced and closed-sourced LLMs. For a given LLM, we mark those questions that it answers correctly as D^+ and other failed questions as D^-.

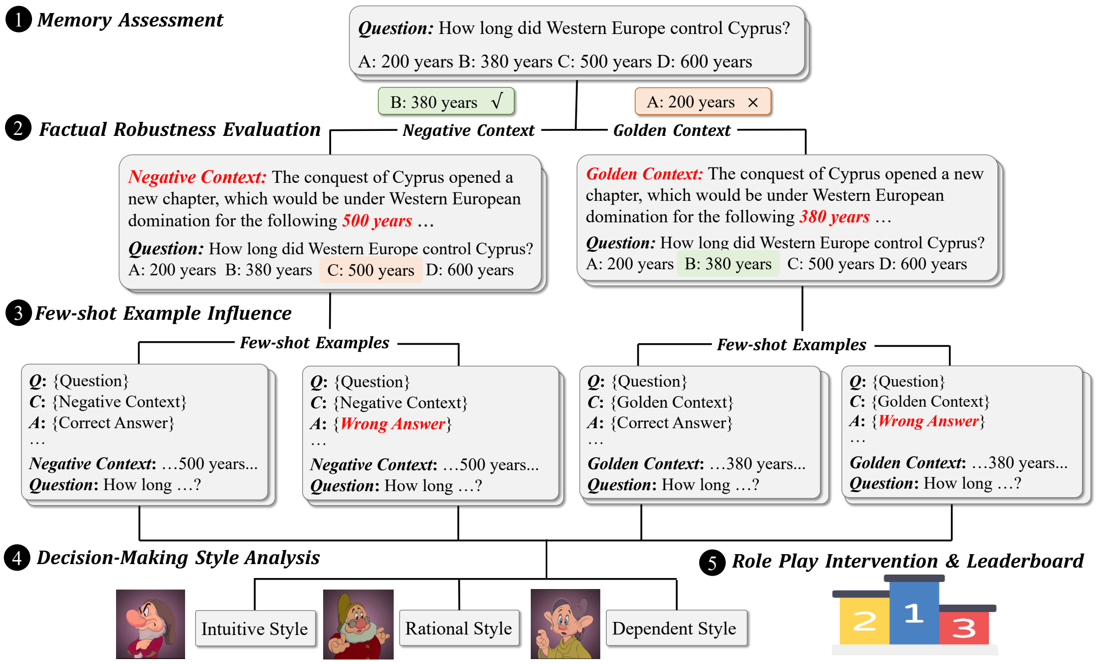

Figure 2: The pipeline incorporates several steps to assess the robustness of LLMs: 1. memory assessment in Section 3.1. 2. Factual robustness evaluation in Section 3.2. 3. Few-shot example influence in Section 3.3. 4. Decision-making style analysis in Sec 3.4. 5. Role play intervention and leaderboard in Sec 3.5

3.2 Factual Robustness Evaluation

Given KRE with D^+ and D^- partition from Section 3.1, we mimic two types of factual conflicts, breaking down factual robustness into two aspects: 1) **Vulnerable Robustness (VR)** that measures to which extent the model can trust its correct memory even with a misleading prompt, and 2) **Resilient Robustness (RR)** which quantifies the model's ability to harness accurate information from the prompt, when memory is insufficient or flawed. Formally, for each sample in D^+, we design the zero-shot ICL prompt to $P = I \oplus c^- \oplus x$ for VR measurement, marked as (D^+, C^-); correspondingly, $P = I \oplus c^+ \oplus x$, marked as (D^-, C^+), is for RR. Details are in Appendix A.1. We define VR and RR metrics as follows:

$$
\begin{aligned}
VR_{(D^+,C^-)} &= \frac{1}{|D^+|} \sum_{x \in D^+} \mathbb{I}\left[f(x, c^-; M) = a_{gol}\right], \\
RR_{(D^-,C^+)} &= \frac{1}{|D^-|} \sum_{x \in D^-} \mathbb{I}\left[f(x, c^+; M) = a_{gol}\right].
\end{aligned}
\quad (1)
$$

Here $f(x, c; M)$ signifies the answer choice produced by model M for the question x with the provided context c. Both VR and RR scores are between 0 and 1. A higher VR indicates better robustness of the model in trusting correct memory in the presence of misleading prompts. A greater RR denotes the model's better ability to utilize context knowledge to overcome conflicting memory. Using these two scores together, we represent the overall Factual Robustness (FR):

$$
FR = Avg\,(VR\,, RR) \quad (2)
$$

Clearly, a higher FR denotes the rational style, otherwise irrational. Before assessing the robustness, we select the best instruction from our pool to mitigate the potential biases. We conduct preliminary experiments on each LLM using a smaller sampled KRE dataset to identify the most effective instruction. Then we chose the instruction that exhibited the highest robustness. This process is also conducted for the following few-shot setting.

3.3 Few-shot Example Influence

Following the zero-shot setting in the last section, we explore few-shot examples E on factual robustness. Formally, E for VR is $c^- \oplus x \oplus a$, and for RR is $c^+ \oplus x \oplus a$. Particularly, for the example answer a, considering the possible noise in practice, we design three configurations (two illustrated in Fig 2) (1) **All-positive** where the answer is always correct. For VR, this involves using a correct answer in conjunction with the negative context C^-, thereby guiding the model to disregard the misleading context. For RR, the correct answer is paired with the golden context C^+, directing the model to effectively utilize the context. (2) **All-negative** where the answer is incorrect. For VR, the incorrect answers are paired with the misleading negative context C^-, guiding the model to use the negative context. For RR, the incorrect answers serve to instruct the model not to rely on the golden context. (3) **Mixed** is a combination of positive and negative examples. In experiments, each question will be tested under all three configurations. The examples are written by human annotators. We manually sample three samples for each evaluation setting. Finally, the prompt is $P = I \oplus E \oplus c \oplus x$ and details are in Appendix A.1. The VR and RR metrics here are as follows.

$$VR_{(D^+,C^-,E)} = \sum_{x \in D^+} \frac{\sum_{e \in E_x} \mathbb{I}[f(x, c^-, e; M) = a_{gol}]}{|E_x||D^+|},$$
$$RR_{(D^-,C^+,E)} = \sum_{x \in D^-} \frac{\sum_{e \in E_x} \mathbb{I}[f(x, c^+, e; M) = a_{gol}]}{|E_x||D^-|}. \tag{3}$$

where E_x is the few-shot examples configurations set (all-positive, all-negative, and mixed) corresponding to question x.

3.4 Decision-Making Style Analysis

Different from factual robustness, there is no superiority in conflicts for decision-making style. Based on cognitive theory (Harren, 1979; Phillips et al., 1984), there are three types: **Dependent style** heavily relies on external information or advice. **Intuitive style**: driven primarily by inner experience or feelings. **rational/irrational style** means wavering between the above two directions. We thus define a **Decision-Making Style Score (DMSS)** (Equation (4)) to efficiently classify models into the above three styles. The closer DMSS to 1 means the model is more likely an intuitive decision-maker who consistently depends on memory to answer questions. Conversely, when DMSS nearing -1 the model aligns more with the dependent style, leaning heavily on prompts. A score around 0 denotes a rational/irrational style, implying LLMs may behave randomly (irrational), or with some patterns (rational) unless we follow the factual robustness assumption — there is a superiority in conflict. More details are in Appendix A.2.

$$DMSS = \frac{1}{|D|}\left(\sum_{x \in D+} \mathbb{I}[f(x, c^-; M) = a_{gol}] + \sum_{x \in D-} \mathbb{I}[f(x, c^+; M) = f(x; M)]\right)$$
$$-\frac{1}{|D|}\left(\sum_{x \in D+} \mathbb{I}[f(x, c^-; M) = a_{neg}] + \sum_{x \in D-} \mathbb{I}[f(x, c^+; M) = a_{gol}]\right), \tag{4}$$

function $f(x, c; M)$ signifies the answer choice produced by model M for the question x with the provided context c, a_{gol} is the golden answer (Defined in Sec 2) for question x.

3.5 Role Play Intervention

To further explore the decision-making tendencies of LLMs, we adopt a common method "Role Play" instruction (e.g., *"you are a writing assistant"*). While this method is commonly used, its effectiveness and models' adaptivity to different roles have not been quantified before. In our evaluation pipeline, we designed two specific role prompts to guide the model towards distinct decision-making styles: **Dependent Role**: asks the model to rely only on the given prompt for answers. **Intuitive Role**: pushes the model towards relying predominantly on its memory (Prompts are in Appendix B.5). Using these role instructions and our metrics, we assess LLMs' adaptivity and upperbound to alter the decision-making styles in Section 4.5.

4 Experiment

We conducted experiments on the full KRE dataset with ChatGPT and Vicuna-13B. Recognizing the importance of a broader analysis, we incorporate five additional LLMs. Due to computational constraints and the time-intensive nature of exhaustive tests, these models are assessed on a subset of the KRE dataset.

4.1 How Factual Robust are LLMs ?

The overall memory assessment for ChatGPT and Vicuna-13B are shown in Table 1 (prompt in Appendix Instruction example). The result shows that the memory of ChatGPT possesses greater and more accurate factual (MRC) and commonsense knowledge (CR) than those of Vicuna. Notably, both ChatGPT and Vicuna tend to perform better on commonsense knowledge datasets compared to factual ones. This might be because LLMs capture many co-occurrence relationships, and a lot of commonsense knowledge is an induction of these observed patterns.

Model	ECQA$_{KRE}$	e-CARE$_{KRE}$	MuSiQue$_{KRE}$	SQuAD$_{KRE}$
ChatGPT	74.2	81.5	34.6	65.3
Vicuna-13B	39.5	70.1	17.7	32.3

Table 1: The memory assessment results (zero-shot) of ChatGPT and Vicuna-13B on the KRE dataset.

Given D^+ and D^- through memory assessment, we select the best-performed instructions (the selection result is shown in Appendix A.3) on a subset of KRE and proceed with factual robustness evaluation. The factual robustness result is shown in

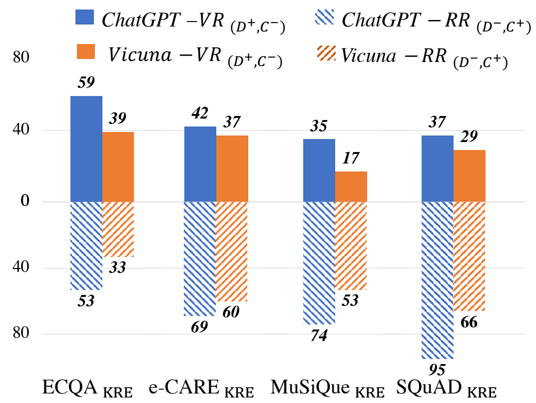

Figure 3: The VR score (%) and The RR score (%) for model ChatGPT and Vicuna-13B.

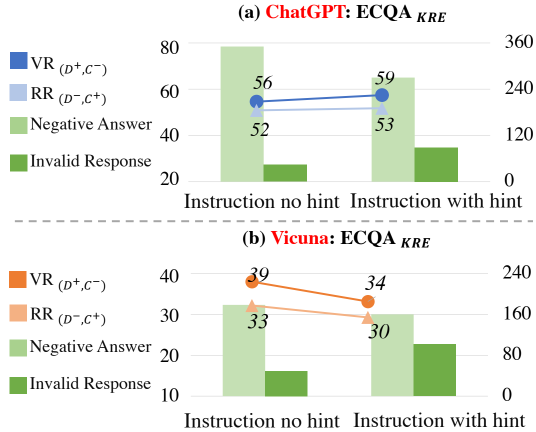

Figure 4: RR and VR of ChatGPT and Vicuna under instruction with and without hint (Sec 2). The corresponding number of negative answers and invalid responses.

Figure 3. ChatGPT and Vicuna exhibit similar behavior. Specifically, a higher RR score relative to the VR score indicates that **LLMs already possess a stronger capability to utilize the correct knowledge from prompts. However, their robustness against negative context remains suboptimal. Consequently, as the field progresses, enhancing robustness against negative context is likely to emerge as a paramount research focus.** Moreover, we compare robustness results on the commonsense (e-CARE, ECQA) against factual questions (MuSiQue, SQuAD). The tested models exhibited higher RR and lower VR on factual questions. **Thus, we conclude that the baseline models can better utilize factual knowledge than commonsense knowledge from prompt contexts. To ensure better utilization of LLMs, there's a pressing need to enhance the precision of factual knowledge embedded in prompts. Meanwhile, when it comes to commonsense knowledge, the focus should be on amplifying the intrinsic memory of the model.** In Figure 3, the total length of the VR and RR bars is proportional to the overall factual robustness FR. ChatGPT's bar is longer than that of Vicuna-13. This can be attributed to ChatGPT's larger number of parameters, more extensive training dataset, and enhanced instruction comprehension capabilities.

4.2 How Instructions Influence FR?

In this section, we explore the influence of different instructions (defined in Sec 2) on factual robustness. The results in Figure 4 (full results in Figure 8) indicate that neither ChatGPT nor Vicuna showcases significant robustness variations under different instructions. To gain deeper insights, we further investigated the model's responses. We calculate the number of negative answers and invalid outputs (such as *" I don't know"*) generated by the model. **Our observations reveal that hint in the prompt about the potential presence of misleading information reduces the model's propensity to choose the negative answer. It also increases invalid responses**, especially for Vicuna. Therefore, when taking both factors into account, the overall robustness does not exhibit any marked variations.

4.3 How Few-shot Examples Effect FR?

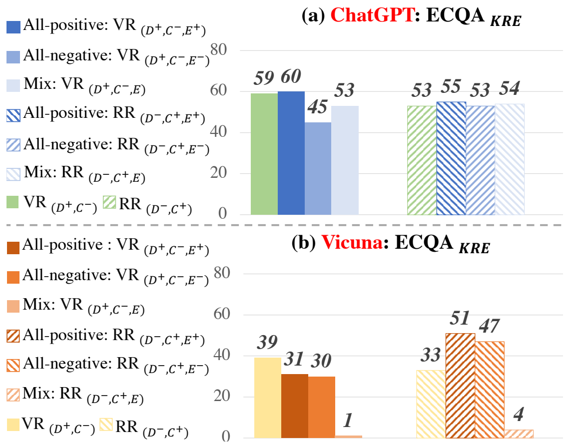

Figure 5: The VR and RR score (%) under the influence of three few-shot configurations.

As shown in Figure 5, for both ChatGPT and Vicuna, the "All-positive" configuration exhibits the highest RR and the highest VR. However, when compared to the zero-shot setting, "All-positive" setting does not always have a positive effect under the conflict situation. This phenomenon is counterintuitive, conventionally, one would anticipate the "All-positive" approach to augment performance, "All-negative" to impede it, and "Mixed" to lie somewhere in between. **The result indicates that the few-shot approach doesn't consistently bolster performance, even in an "All-positive" ver-**

sus zero-shot comparison. Two potential explanations emerge for this phenomenon: 1: Few-shot examples, may more act to dictate the output pattern to the model, rather than the "thinking" pattern under a conflict situation. 2: The extended length of the context could obstruct the LLMs from harnessing the implicit pattern information in few-shot examples. **Interestingly, we observe that under the mixed setting, Vicuna-13B's performance is notably subpar.** This suggests that the presence of mixed answer patterns induces confusion within the model, leading to its diminished performance. Notably, this phenomenon is absent in ChatGPT's performance, suggesting that ChatGPT possesses a more refined robustness to demonstration. In line with the zero-shot setting described in Section 3.2, we first select the most effective instruction before evaluating robustness. We notice a similar pattern for the influence of the instructions (Sec 4.2), detailed in Figure 8.

4.4 Decision-Making Style Analysis

In our work, we incorporated seven models, namely Vicuna-13B, ChatGPT, GPT-4, Claude (Anthropic, 2023), Bard (Google, 2023), LLaMA (Touvron et al., 2023a), and LLaMA2 (Touvron et al., 2023b). Table 2 shows the DMSS. Notably, most models, 4 out of 7, tend to exhibit dependent decision-making style. This tendency is likely attributable to the influence of instruction tuning, guiding these models to utilize external knowledge more effectively. LLaMA diverges from this trend, exhibiting intuitive decision-making style. This behavior further corroborates our inference when considering that LLaMA did not undergo instruction-tuning. Moreover, models, GPT-4 and Bard, with superior factual robustness (Table 2) tend to exhibit rational style. **We hypothesize that when models reach a certain scale, they inherently amplify both their memory retention and instruction-following capabilities**. These enhancements allow them to balance between relying on stored knowledge and adapting to new information from prompts.

4.5 Roly Play Intervention and Leaderboard

Roly Play Intervention. We opted role play interventions on ChatGPT and Bard, which exhibit rational style, and on LLaMA-2, which shows dependent style. Illustrated in Figure 6, the three bars reveal a conspicuous shift in the model's decision-making behavior post-intervention. Depending on the assigned role, post-intervention models demonstrated a distinct bias: under the intuitive role, they rely more heavily on their internal memory, as evidenced by higher DMSS, while under the dependent role, they depend more on the provided prompt, resulting in lower DMSS. **This result indicates that we can change LLMs' decison-making style through role play intervention**. The range between the highest DMSS scores (intuitive role, blue bar) and the lowest (dependent role, yellow bar) shows the **Adaptivity** of the models in role play scenarios. Larger adaptivity signifies greater effectiveness in adapting to the demands of assigned roles.

Considering the possibility of altering decision-making styles by role instructions, we further investigate how different styles of instruction effect models' VR and RR scores. We find a consistent correlation between the models' assigned roles and the robustness scores. Specifically, when under the intuitive role, each model has the peak VR score. Conversely, under the dependent role, models have the highest RR scores. By using different role instructions under different scenarios, we are able to discern the **Upper-Bound** for the Facutal Robustness (indicated by the red number in Fig 6). A greater upper-bound indicates a higher potential FR score in the models when facing conflict.

Leaderboard. At the last stage, we construct the leaderboard. Table 2 summarizes the robustness score, encompassing FR and DMSS, the adaptivity, and the upper-bound for the seven models. Among the models, Bard stands out for its superior vulnerable robustness, effectively maintaining its memory when given misleading prompts. In contrast, GPT-4 has the highest resilient robustness, demonstrating its ability to capitalize on accurate knowledge in prompts. Furthermore, GPT-4 also displays unmatched factual robustness, properly relying on the prompt to discern accurate answers. LLaMA-2-13B-chat has the lowest DMSS score under role play intervention. **This suggests that in specific scenarios, it can adhere to the given instructions even more rigorously than GPT-4. However, when it comes to adaptivity and upper-bound, it significantly falls behind GPT-4.**

5 Related Work

Prompt-in LLMs: Large language models can solve various tasks by simply conditioning the models on a few examples (few-shot) or instructions. The method of conditioning the language model is

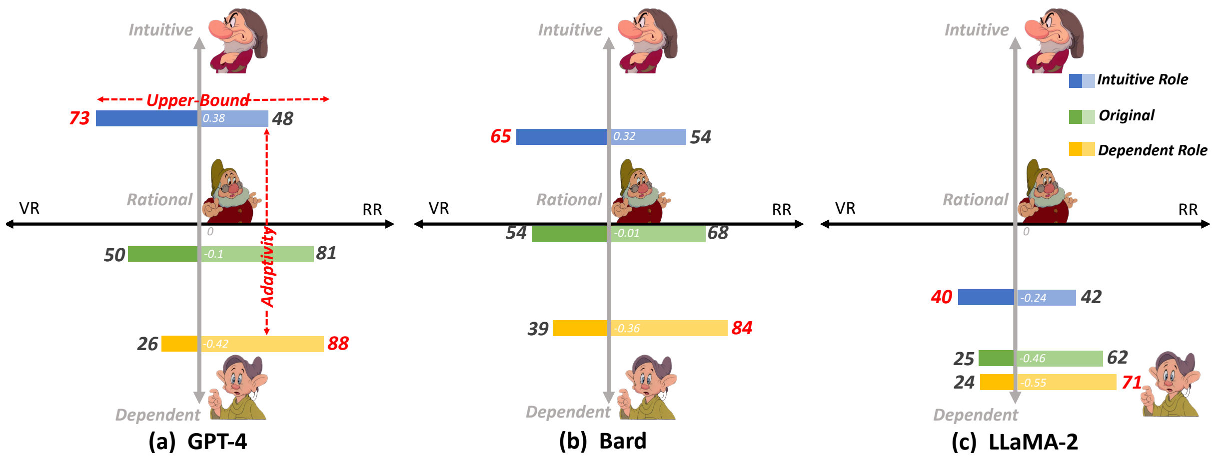

Figure 6: Role play Intervention result for the model GPT-4, Bard, LLaMA-2. The results illustrate under specific DMSS scores, the VR and RR scores of each model adjust post-intervention.

Model	VR	RR	FR	FR_{upper}	FR_{rank}	DMSS	Style	Adapt	$Adap_{rank}$	Over all
GPT-4	50	88	69	80	1	-10	Rational	0.8	1	1
Claude	34	57	45	60	4	-43	Dependent	0.39	4	4
ChatGPT	32	79	56	63	3	-43	Dependent	0.45	3	3
Vicuna-13B	25	48	36	44	6	-31	Dependent	0.27	6	6
Bard	54	68	61	74	2	-1	Rational	0.68	2	2
LLaMA-13B	20	21	20	33	7	39	Intuitive	0.15	7	7
LLaMA-2-13B-chat	24	62	39	55	5	-46	Dependent	0.31	5	5

Table 2: The Robustness Leaderboard. The table shows the two robustness scores (FR and DMSS) for the involved models, and the rank of FR score (FR_{rank}) and Adaptivity ($Adap_{rank}$)

called "prompting" (Liu et al., 2023), and designing prompts either manually (Schick and Schütze, 2021; Reynolds and McDonell, 2021) or automatically (Shin et al., 2020; Gao et al., 2021)has become a hot topic in NLP. Prompts serve as the interface between humans and LLMs, enabling in-context learning in an auto-regressive manner (Liu et al., 2023). LLMs are known to be highly sensitive to prompts (Turpin et al., 2023; Shi et al., 2023; Zheng et al., 2023; Zhao et al., 2021; Si et al., 2022), where minor variations can influence the performance. It is crucial to examine the robustness of LLMs under the influence of the prompt. Recent studies have shown that language models are vulnerable to adversarial prompt (Wang et al., 2023; Zuccon and Koopman, 2023). Work (Zhuo et al., 2023) shows that prompt-based semantic parsers built on large pre-trained language models have also highlighted their susceptibility to adversarial attacks (Bruna et al., 2014; Hosseini et al., 2017). The work (Wang et al., 2023) evaluated the robustness of ChatGPT and other LLMs from an adversarial perspective. Another work, PromptBench (Zhu et al., 2023), developed a robustness benchmark to assess the resilience of adversarial prompts. **Factual knowledge conflict:** Work (Longpre et al., 2021) concentrated on entity-based conflicts, testing on the model's pertained dataset. The work (Chen et al., 2022) investigated how the model acts when given multiple pieces of conflicting evidence. Research (Zuccon and Koopman, 2023) has explored the impact of wrong knowledge in prompts on ChatGPT's performance when answering health questions. Another recent study (Xie et al., 2023) investigated how the model behaves when encountering knowledge conflicts by modifying the model's outputs into conflict prompts. Differently, our work proposes a general framework, applicable to any model and conflicting situations. We systematically evaluate the behavior of LLMs under various conflict scenarios.

6 Conclusion

This comprehensive study provides pivotal insights into the behavior of LLMs' under conflict. We have designed a quantitative benchmarking framework in terms of factual robustness and decision-making style. Based on that, we have conducted extensive experiments on several LLMs. The results underscore many critical revelations. Besides, we deploy role play intervention to change the models' decision-making style, which shows the varying adaptivity and upper-bound of different LLMs. Based on these insights, in the future, we will explore strategies to improve LLMs' abilities to use factual knowledge via prompts while enhancing commonsense reasoning via internal memory.

7 Limitations

While our evaluation framework and findings provide valuable insights into the robustness of the assessed models, it is important to acknowledge certain limitations: 1. Limited Dataset: The evaluation is conducted on a subset of the Knowledge Robustness Evaluation (KRE) dataset. The size and diversity of the dataset may impact the generalizability of the results. A larger and more diverse dataset could provide a more comprehensive understanding of model robustness. 2. Task Specificity: The evaluation focuses on knowledge-intensive tasks and may not fully capture the robustness of models in other domains or tasks. The findings might not generalize to all types of language processing tasks or scenarios. 3. Evaluation Metrics: The metrics used to quantify robustness are designed based on specific criteria and may not encompass all aspects of robustness. Alternative metrics or additional dimensions of robustness could provide further insights into model performance. 4. Limited Model Selection: The evaluation is conducted on a specific set of models. We will involve more models.

8 acknowledgement

This research is supported by the Lee Kong Chian Fellowship fund awarded by Singapore Management University, as well as Google South Asia & Southeast Asia Research Awards by Google Asia Pacific Pte. Ltd.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. 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 3050–3065.

Anthropic. 2023. Anthropic: Claude.

Joan Bruna, Christian Szegedy, Ilya Sutskever, Ian Goodfellow, Wojciech Zaremba, Rob Fergus, and Dumitru Erhan. 2014. Intriguing properties of neural networks.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, pages 3816–3830. Association for Computational Linguistics (ACL).

Google. 2023. Google: Bard.

Vincent A Harren. 1979. A model of career decision making for college students. *Journal of vocational behavior*, 14(2):119–133.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022. The inverse scaling prize.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2023. Openai: Gpt-4.

Susan D Phillips, Nicholas J Pazienza, and Howard H Ferrin. 1984. Decision-making styles and problem-solving appraisal. *Journal of Counseling psychology*, 31(4):497.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

Terry Yue Zhuo, Zhuang Li, Yujin Huang, Fatemeh Shiri, Weiqing Wang, Gholamreza Haffari, and Yuan-Fang Li. 2023. On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1090–1102.

Guido Zuccon and Bevan Koopman. 2023. Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*.

A Experiment Details

A.1 Preliminary

Our evaluation focuses on the conflict situation where the prompt we consider has four key components: **the instruction** I, **the testing question** x, **the knowledge context** c related to x (c^+ means the golden context and c^- means the negative context corresponding to question x. C^+ and C^- are used to represent the sets of all golden and negative contexts), and **the few-shot examples** set E (removed for zero-shot learning scenario). We define the prompt P as the concatenation of the above components: $P = I \oplus E \oplus c \oplus x$, where $\oplus$ denotes the concatenation operation. The few-shot example e in E here is in the format: $e = c \oplus x \oplus a$, where a is the answer to the question x. For example, P could be "***I**: Help me to answer the question. **E**: Context: Lake is a large area of water surrounded by land. Question: Where can I find water? Answer: Lakes. **c**: Foxes hunt chickens. **x**: Question: Where would I not want a fox?*".

A.2 Decision-Making Equation

Here we defined a **Decision-Making Style Score (DMSS)** to measure the behavior of the LLM. The closer DMSS to 1 means the model is more likely to depend on self-memory to answer the question. Conversely, when DMSS nearing -1 the model learns heavily on external prompts. A score around 0 denotes a rational style, implying the LLM will consider the memory and the prompt together to make the decision. **However, it's vital to note that a DMSS near 0 doesn't necessarily guarantee the model's capability to judiciously consider both the memory and the prompt. Given the conflicting scenarios in this study, discerning whether the model genuinely integrates both sources or randomly selects an option becomes challenging. Thus, in such cases, the Factual Robustness score should also be examined as an auxiliary metric to provide a more comprehensive understanding.**

$$
\begin{aligned}
DMSS = & \frac{1}{|D|}\left(\sum_{x\in D^+} \mathbb{I}\left[f(x,c^-;M)=a_{gol}\right] + \sum_{x\in D^-} \mathbb{I}\left[f(x,c^+;M)=f(x;M)\right]\right) \\
& -\frac{1}{|D|}\left(\sum_{x\in D^+} \mathbb{I}\left[f(x,c^-;M)=a_{neg}\right] + \sum_{x\in D^-} \mathbb{I}\left[f(x,c^+;M)=a_{gol}\right]\right),
\end{aligned} \tag{5}
$$

function $f(x, c; M)$ signifies the answer choice produced by model M for the question x with the provided context c, a_{gol} is the golden answer (Defined in Sec 2) for question x

A.3 Instruction Selection

For the instruction selection process, we adhere to the methodology outlined in Section 3.2. The performance of candidate instructions with the ChatGPT and Vicuna models in the Zero-shot Setting is shown in Table 3 and 4. The results for instructions without hints are presented in Table 3, while the results for instructions with hints are shown in Table 4. The specific instructions used for the evaluations can be found in Section B.2.1 and instructions with the hint in Section B.2.2.

Model	1	2	3	4	5	6	7	8	9	10	11	12
ChatGPT	80	78	82	75	83	**87**	78	79	83	78	74	82
Vicuna-13B	**79**	58	54	71	60	74	72	68	66	66	67	60

Table 3: The performance (%) for the model ChatGPT, and Vicuna-13B on the instruction selecting dataset with instructions 1 to 12 defined in Section B.2.1.

Model	1	2	3	4	5	6	7	8	9	10	11	12
ChatGPT	85	85	85	72	83	78	85	**86**	83	84	81	79
Vicuna-13B	**72**	65	61	71	36	68	58	41	66	60	66	66

Table 4: The performance (%) for the model ChatGPT, and Vicuna-13B on the instruction selecting dataset with instructions with hint 1 to 12 defined in Section B.2.2.

As a result, we select the number **6** instruction without hint and the number **8** instruction with hint for the model ChatGPT, the number **1** instruction without hint and the number **1** instruction with hint for the model Vicuna-13B to have the Robustness Evaluation. We then select the **best performance** (the result is shown in figure 8) for each model and then concatenate with the candidate instruction for Few-shot setting to have the Instruction Selection process. The rest for the instructions for Few-shot setting is shown in Table 5 and Table 6. The results for instructions without hints are presented in Table 5, while the results for instructions with hints are shown in Table 6. The specific instructions used for the evaluations can be found in Section B.3.1 and instructions with the hint in Section B.3.2.

Model	1	2	3	4	5	6	7	8	9	10	11	12
ChatGPT	63	61	60	59	59	62	61	62	62	**<u>64</u>**	60	61
Vicuna-13B	54	45	53	52	40	52	46	46	**<u>61</u>**	60	52	44

Table 5: The performance (%) for the model ChatGPT, and Vicuna-13B on the instruction selecting dataset with instructions 1 to 12 defined in Section B.3.1.

Model	1	2	3	4	5	6	7	8	9	10	11	12
ChatGPT	62	61	56	61	62	**<u>64</u>**	62	61	61	63	60	61
Vicuna-13B	47	46	53	53	**<u>55</u>**	45	49	46	52	35	45	39

Table 6: The performance (%) for the model ChatGPT, and Vicuna-13B on the instruction selecting dataset with instructions with hint 1 to 12 defined in Section B.3.2.

A.4 Human Evaluation

To validate the quality of these generated answers, we randomly selected 100 candidate answer sets and conducted a human evaluation involving four senior computing language researchers, trained in advance. Remarkably, in 98% of cases, the human evaluators were unable to differentiate the correct answer from the candidates when given the answer set alone and conducted a human evaluation involving four evaluators.

To qualify the generated negative context we randomly selected 100 questions from each corpus in dataset KRE (a totally of 400 samples) and conducted a human evaluation involving four senior computing language researchers. Four evaluators were chosen for this task. For each selected question, evaluators were provided with: the generated negative context, the associated question, and the set of potential answer choices. Evaluators were required to determine how much the negative context might skew one's response towards the negative or misleading answer. This assessment was categorized into three distinct levels: No-misleading, Somewhat misleading, and Highly misleading. Here we provide the annotation instructions.

1. Each column contains one question, one context, several options, and an answer.
2. You are tasked with evaluating the extent to which the context might influence or skew your response towards the given answer.
3. Apply the following three levels of criteria to assess the context:
   - **No-misleading**: The context does not lead to the answer.
   - **Somewhat misleading**: The context contains elements that could potentially lead to the answer, but they are not definitively deceptive or strong enough to guide you to choose that answer.
   - **Highly misleading**: The context strongly influences you to choose the answer.

The evaluation result is shown in Figure 7. The result shows that more than 95% of the context, which is constructed based on the corresponding Wikipedia, in the two MRC datasets is highly misleading. In contrast, the context for the RC dataset, although anchored in common sense knowledge and inherently more challenging to distort for human understanding, still saw upwards of 65% being labeled as highly misleading. The agreement of the score reaches more than 98% for the two MRC datasets and 90% for the CR datasets.

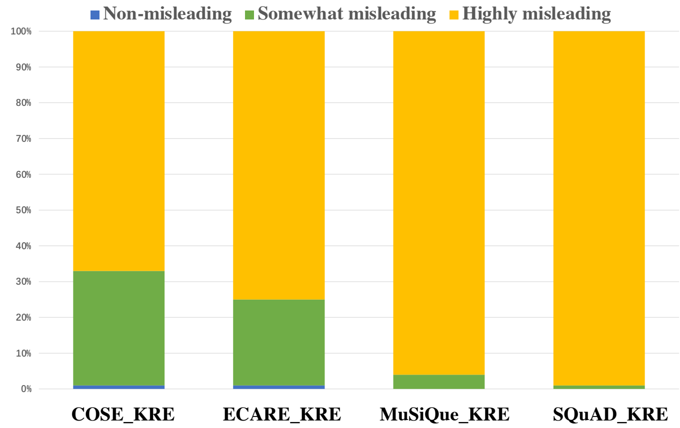

Figure 7: Human Evaluation Result for the generated negative context. We label the context into three levels: No-misleading: Given the context, it does not lead to a misleading answer. Somewhat misleading: The information or context has elements that could be considered misleading, but it's not entirely clear or strong enough to typically deceive a human. Highly misleading: The context or information presented can easily mislead humans when answering a question. It strongly biases or directs the interpretation in a deceptive manner.

A.5 Model setup

For all models, we set the maximum output length to 520 tokens, and the temperature to 0.

A.6 Additionl Experiment Result

In Figure 8 we show the whole result for ChatGPT and Vicuna-13B on the KRE dataset under the two instruction settings. The Figure 11 represents the robustness score for ChatGPT and Vicuna-13B on the KRE dataset under the three few-shot settings.

Figure 8 shows the whole result of ChatGPT and Vicuna-13B on the KRE dataset.

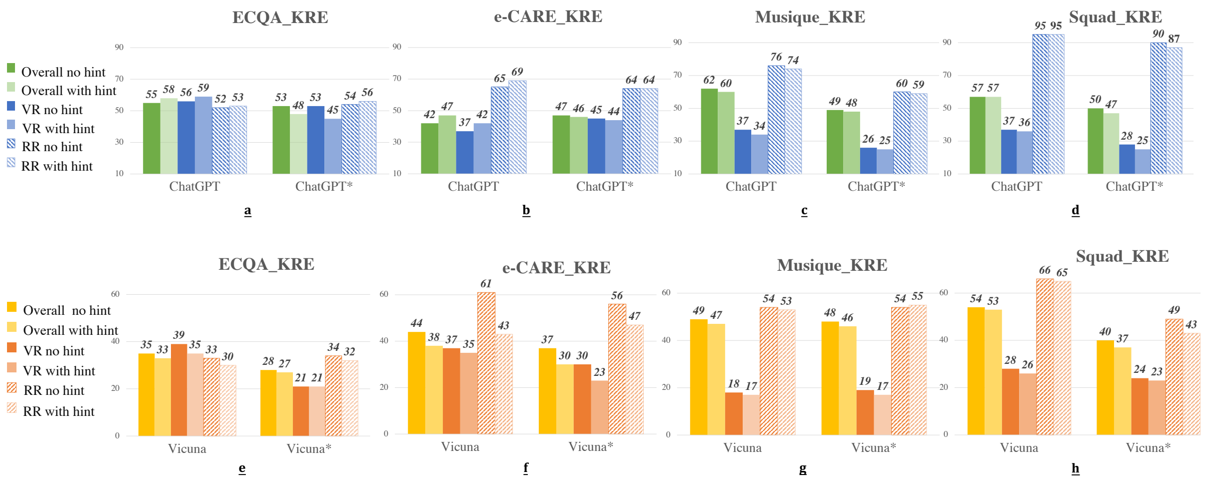

Figure 8: The RR and VR (%) of ChatGPT(index a, b, c, d) and Vicuna (e, f, g, h) under the influence of Instructions with different semantics: b: with hint and a: without hint(defined at Section 2). Overall means weighted average performance on the whole dataset, which is the average from the D^{+} part and the D^{-} part (defined in section 3.1).ChatGPT, Vicuna means the Zero-shot configuration for each model, ChaGPT*, Vicuna* means the Few-shot configuration. The result of the Few-shot condition is the average result of the 3 example configurations.

Table 7 shows the Corpus level statistics of the Knowledge Robustness Evaluation (KRE) Dataset. The KRE consists of four public datasets. Two MRC datasets: MuSiQue and, SQuAD to test the factual

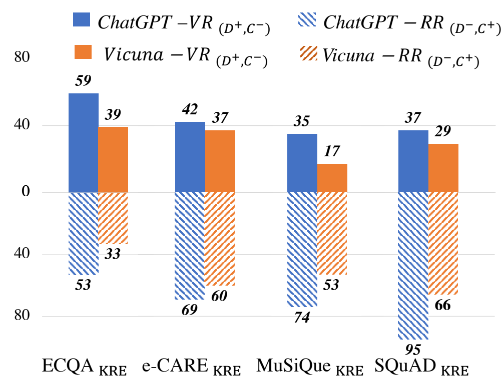

Figure 9: The Resilient Robustness score (%) and The Vulnerable Robustness score (%) for model ChatGPT and Vicuna-13B.

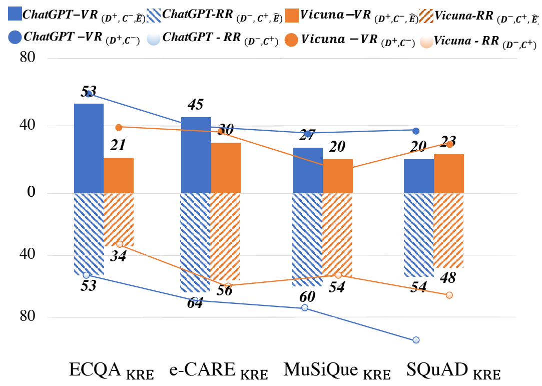

Figure 10: The robustness score(%) for Few-shot setting. The two lines aligned with points show the result of the original RR and VR score which can be found in Figure 9 for more detail.

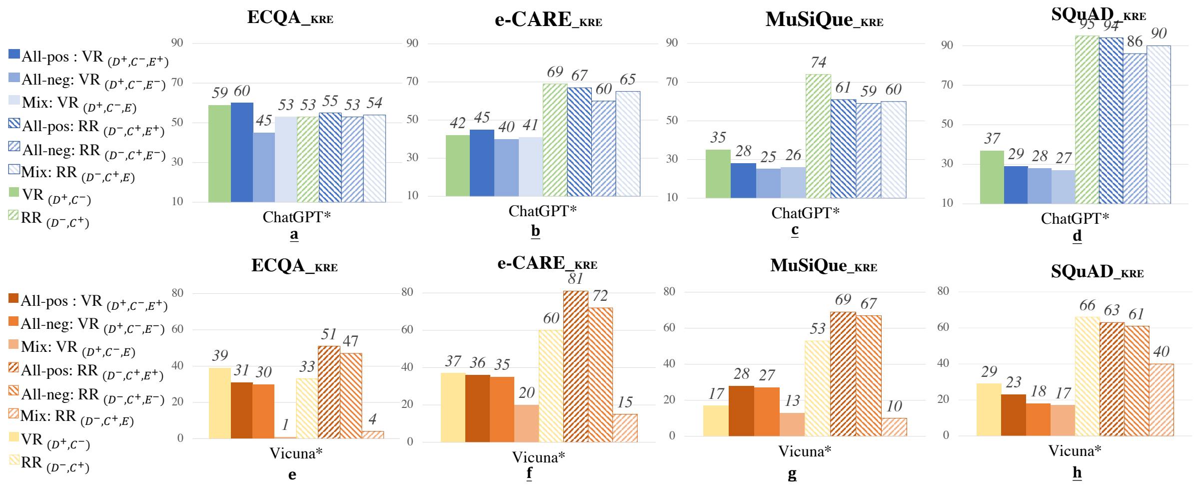

Figure 11: The RR and VR (%) of ChatGPT(index a, b, c, d) and Vicuna (e, f, g, h) under the influence of three few-shot configurations: "All-positive", "All-negative" and "Mix".

knowledge robustness, and two RC datasets: ECQA and, e-CARE testing commonsense knowledge robustness.

Dataset	Size
MuSiQue (Trivedi et al., 2022)	2,417
SQuAD v2.0 (Rajpurkar et al., 2018)	5,924
ECQA (Aggarwal et al., 2021)	1,221
e-CARE (Du et al., 2022)	2,122
KRE Total	11684

Table 7: Corpus level statistics of the Knowledge Robustness Evaluation (KRE) Dataset.

Table 8 shows the number of missing answers and invalid answers the model output when given instruction having or without a hint (Destail design in Section 2, Instructions is shown in Appendix B.2 and Appendix B.3). Our observations reveal that the inclusion of a hint indeed reduces the propensity of the model to choose the negative answer. However, it also introduces an increase in the frequency of invalid responses. Therefore, the overall robustness does not exhibit any marked improvement.

Table 9 shows the detailed results of the robustness score for the tested seven models.

Configuration	#Misleading Answer	#Invalid
ChatGPT with hint	3638	892
GhatGPT without hint	3902	637
Vicuna with hint	2216	1035
Vicuna without hint	2496	267

Table 8: Number of Negative Answers (defined at section 2) in setting $VR(D^+, C^-)$ on KRE, and total number of Invaid Output in setting $VR(D^+, C^-)$, and $RR(D^-, C^+)$.

Model	VR	RR	FR
GPT-4	50	81	66
Claude	34	57	45
ChatGPT	32	79	56
Vicuna-13B	25	48	36
Bard	54	68	61
LLaMA-2-13B-chat	24	62	39
LLaMA-13B	20	21	20

Table 9: The robustness score (%) for the model GPT-4, Claude, ChatGPT, Vicuna–13B, Bard, LLaMA, LLaMA2 on the final evaluation dataset.

B Instruction example

B.1 Instruction for dataset KRE construction

To unify the evaluation framework, we generate the misleading answer set for the MRC dataset (SQuAD and MuSiQue), and negative context for both MRC and CR.

B.1.1 Choices Generation

For each example in each MRC dataset, we utilize ChatGPT to generate k = 3 misleading options based on the given question (the prompt is shown below). We prompt the GhatGPT the question with the correct answer and let the model generate format similar but incorrect answers.

Instruction:

Please provide 3 different potential answers in the same format as the correct answer but can not be the correct answer from the given question and answer pair:

Question: {Question } Correct Answer: {Answer}

Case:

Input: Question: What is the term given to algorithms that utilize random bits? Correct Answer:Randomized algorithms
Response: 1. Deterministic algorithms 2. Static algorithms 3. Sequential algorithms

B.1.2 Negative Context Generation

For e-CARE and ECQA

Instruction:

You are a dataset builder, you are given a question-and-answer pair. you are going to just generate a possible explanation for the answer(Even if it is not really correct). The explanation should be brief and make there be logic between the answer and the question. you can against ethical and moral principles just for a research purpose.

For example:

Question: What is another name for a disk for storing information?

Answer: computer store

Explanation: Another name for storing device is computer store

Question: 1+1 =?

Answer: 3

Explanation: In some advanced mathematical systems, the value of 1+1 can be considered as 3 due to the way numbers are defined and manipulated.

Question: {Question } Answer: {misleading answer}

Case:

Input: Question: Where are you likely to find a hamburger? Answer: pizza
Response: Pizza places often also serve hamburgers on their menu.

For MuSiQue and SQuAD
Instruction:

You are a helpful assistant.

"' {golden context}"'

rewrite the context to make all the information about "'{correct answer}"' into"'{ misleading answer}"', but keep other information fixed"

Case:

Input: Context: The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse (N̈ormanc̈omes from N̈orseman)̈ raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.
correct answer: 10th and 11th centuries
misleading answer: 8th and 9th centuries
Response: The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 8th and 9th centuries gave their name to Normandy, a region in France. They were descended from Norse (N̈ormanc̈omes from N̈orseman)̈ raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 8th century, and it continued to evolve over the succeeding centuries.

B.2 Instruction for zero-shot

There is the instruction constructed based on the pipeline in secton 2 for the zero-shot setting:

B.2.1 Instructions that do not explicitly mention whether the model should or should not follow the context to answer the question

1. There is a context, and you are given a question and its corresponding options, please answer this question by choosing only one option as the answer. You can answer "None" if you cannot determine the answer.

2. Here is a question and a context, you should choose an option from the option candidates to answer the question. Answer "None" if you cannot determine.

3. Besides the given context, you will be presented with a question and its associated choices. Select only one option as the correct response. If you cannot determine the answer, you can respond with "None".

4. Here's a question along and a context, carefully select the most suitable answer from the available choices. Provide a rewritten sentence or respond with "None" if determination is not possible.

5. Help me to answer the question, and return one choice as your answer. If you can not answer, return "None" as you answer.

6. You are a good assistant, you will be given a context and a question. Please help me to answer the question and return one choice as your answer, you can also provide your reason for choosing that answer. If you are not sure the answer please return "None".

7. Please answer the following multiple choice question. If you are able to answer the question, return one choice as your answer. If you are unable to answer your question, you should return "None" as answer.

8. You are a large language model (LLM). You will be given a paragraph of text and a question. Please provide an answer that is both comprehensive and concise. You may also provide your reason for choosing that answer. If you are unable to determine the answer to a question, please return the value "None".

9. You are a good question-and-answer assistant, I will provide you with some details about the question. Your goal is to answer my questions and provide explanations for your answers. Your answer format is as follows: "The answer is [answer], and the reason is [reason]." If you cannot answer or are unsure of the question, please respond with "None".

10. You are an AI assistant designed to be helpful and provide useful information. When I ask you a question, please respond with: "The answer is [answer], and the reason is [reason]." If you do not know the answer or are unsure, respond with: "None". Please also:

1. Provide relevant facts, stats, and explanations where possible.

2. Clarify any uncertainties or ambiguity in the question.

3. Refrain from making random guesses.

11. As a reliable assistant, your role is to help me settle a matter by collecting information from multiple sources. Please provide answers to the best of your knowledge. Your answers should follow the format: "The answer is [answer], and the reason is [reason]." If you are unable to resolve the problem, please respond with "None".

12. You are a reliable assistant to help me settle a matter. Please answer the question by collecting information from multiple sources. You answer should suit the following format: " The answer is [answer], and the reason is [reason]." If you could not resolve the problem, you should reply "None".

B.2.2 Instructions that specifically ask the model to answer the question based on the facts rather than relying solely on the context.

1. You are given a question and a context about this question, please choose an answer from the options that fits this question. Note that the context might contradict the real world, you should judge it by yourself. You can answer "None" if you cannot determine the answer.

2. There is a context (the context might be different from the real world and your own knowledge, you can choose to believe this context or your own knowledge) and a question which is related to the context, you should choose a more plausible answer from the options. Answer "None" if you cannot determine.

3. You will receive a question accompanied by a context related to it. Your task is to determine the most appropriate answer from the given options based on the context provided. Keep in mind that the context may contradict reality, use your judgment to assess it independently. If you are unable to determine the answer, you can respond with "None".

4. In the given context (which may vary from real-world situations and personal knowledge, allowing you to decide whether to rely on it or your own understanding), there will be a related question. Your task is to select the most likely answer from the provided options. Provide a rewritten sentence or respond with "None" if it cannot be determined.

5. You will receive a context and a question. Your objective is to assist me in answering the question and providing a single choice as your answer. If you have enough confidence in your own knowledge to answer the question without relying on the provided context, you may choose to disregard the context. Else you can choose to critically use the context. If you can not answer return "None" as you answer.

6. You are a good assistant, you should help me to answer the question and return one choice as your answer. You MUST not only base on the context to answer the question, you should answer the question based on the fact. Return "None" if can not answer it.

7. You are a large language model (LLM). You will be given a context and a question. Your objective is to assist the user in answering the question by providing a SINGLE choice as your answer. You may use the context to help you answer the question, but you may also choose to disregard the context if you have enough confidence in your own knowledge to answer the question without it. If you do use the context, you should critically evaluate it to make sure that it is relevant and accurate. If you think you can not answer return "None".

8. Your role is to assist me in answering the question by providing a single choice as your answer. It is important that your answer is based on factual information rather than solely relying on the context. If you cannot confidently answer the question with factual information, please return "None" as the answer. Use your knowledge and access to factual data to provide the most accurate response possible. Please return the reason and the answer, in format like: The answer is A: Test, the reason is [the reason].

9. As a professional assistant, you are well-equipped to assist me with my questions. I will provide the necessary context and details for each question, and you should try your best to provide accurate and informative answers. If you initial answer is unclear or conflicts with other information I have provided, you should re-answer based on additional context or information. Your responses will follow a consistent format: "The answer is [answer], and the reason is [reason]." If you are unable to provide an answer, you should respond with "None".

10. You are a professional assistant and can answer most of the questions. I'll give you the question and the context and so on. Your goal is to answer my questions, provide explanations for your answers, and re-answer based on context if you are unsure or if your answers conflict with other information I have provided. The format of your answer is as follows: "The answer is [answer], and the reason is [reason]." If you can't provide an answer, please respond with "None".

11. You are a trusted assistant to help me answer the question. Please explain the answer again based on the context I give to you, when there is conflict between the information that you have gathered and the information that provided by context, or you are not sure if the answer is correct. The format of your answer should meet the following requirements: " The answer is [answer], and the reason is [reason]." If you can not answer the question, please give me the reply "None".

12. You are a trusted assistant to help me answer the question. Please explain the answer again based on the context I give to you, when there is conflict between the information that you have gathered and the information that provided by context, or you are not sure if the answer is correct. The format of your answer should meet the following requirements:"The answer is [answer], and the reason is [reason]." If you can not answer the question, please give me the apply "None".

B.3 instruction for few-shot

There is the instruction constructed based on the pipeline in secton 2 for few-shot setting:

B.3.1 Instructions that do not explicitly guide the model to utilize the few-shot examples for answering the question.

1. There is a context, and you are given a question and its corresponding options, please answer this question by choosing only one option as the answer. You can answer "None" if you cannot determine the answer. Here are some examples: " [example1], [example2], [example3] " I'd like you to imitate these examples and give me your answer to the following question.

2. Here is a question and a context, you should choose an option from the option candidates to answer the question. Answer "None" if you cannot determine. Allow me to present some examples for better clarity: " [example1], [example2], [example3] "

3. Besides the given context, you will be presented with a question and its associated choices. Select only one option as the correct response. If you cannot determine the answer, you can respond with "None". Here are some examples: " Example 1: [example1], Example 2: [example2], Example 3: [example3] " You should read these examples and answer the following question.

4. Here's a question along and a context, carefully select the most suitable answer from the available choices. Provide a rewritten sentence or respond with "None" if determination is not possible. Kindly review the following examples: " Example: [example1], Example: [example2], Example: [example3] " Please read through these examples and help me answer the following question.

5. Help me to answer the question, and return one choice as your answer. If you can not answer, return "None" as you answer. Now I want you to read some examples and help me answer some questions. " [example1], [example2], [example3] "

6. You are a good assistant, you will be given a context and a question. Please help me to answer the question and return one choice as your answer, you can also provide your reason for choosing that answer. If you are not sure the answer please return "None". I will also provide you with some examples of questions and answers: " Example: [example1], Example: [example2], Example: [example3] "

7. Please answer the following multiple choice question. If you are able to answer the question, return one choice as your answer. If you are unable to answer your question, you should return "None" as answer. I will provide you with some examples of questions and answers, here they are: " [example1], [example2], [example3] "

8. You are a large language model (LLM). You will be given a paragraph of text and a question. Please provide an answer that is both comprehensive and concise. You may also provide your reason for choosing that answer. If you are unable to determine the answer to a question, please return the value "None". At the same time, your answer can refer to some examples of questions and answers I give: " [example1], [example2], [example3] " Finally, please provide me with the results and explanations of the answer.

9.You are a good question-and-answer assistant, I will provide you with some details about the question. Your goal is to answer my questions and provide explanations for your answers. Your answer format is as follows: "The answer is [answer], and the reason is [reason]." If you cannot answer or are unsure of the question, please respond with "None". Here are a few examples to illustrate the questions and answers: " 1. [example1] 2. [example2] 3. [example3] " Your objective is to respond to my questions by using the provided examples as references.

10. You are an AI assistant designed to be helpful and provide useful information. When I ask you a question, please respond with: "The answer is [answer], and the reason is [reason]." If you do not know the answer or are unsure, respond with: "None". Please also:

1. Provide relevant facts, stats, and explanations where possible.

2. Clarify any uncertainties or ambiguity in the question.

3. Refrain from making random guesses.

Below are some examples for your reference: " Example: [example1], Example: [example2], Example: [example3] " you are given an extra question.

11. As a reliable assistant, your role is to help me settle a matter by collecting information from multiple sources. Please provide answers to the best of your knowledge. Your answers should follow the format: "The answer is [answer], and the reason is [reason]." If you are unable to resolve the problem, please respond with "None". Please refer to the examples provided below and answer the additional question given. " [example1], [example2], [example3] "

12. You are a reliable assistant to help me settle a matter. Please answer the question by collecting information from multiple sources. You answer should suit the following format: " The answer is [answer], and the reason is [reason]." If you could not resolve the problem, you should reply "None". Example: " [example1], [example2], [example3] "

B.3.2 Instructions that explicitly instruct the model to focus on the answer format rather than the correctness of the few-shot examples.

1. You are given a question and a context about this question, please choose an answer from the options that fits this question. Note that the context might contradict the real world, you should judge it by yourself. You can answer "None" if you cannot determine the answer. Here are some examples: " [example1], [example2], [example3] " NOTE: Some of these examples are possibly wrong, I'd like you to follow these examples and give me your answer which is right.

2. There is a context (the context might be different from the real world and your own knowledge, you can choose to believe this context or your own knowledge) and a question which is related to the context, you should choose a more plausible answer from the options. Answer "None" if you cannot determine. Here are some examples to guide you: " [example1], [example2], [example3] " Please keep in mind that some of these examples may contain errors or inaccuracies. I kindly request that you utilize these examples as references. Your expertise and assistance are highly valued. Thank you!

3. You will receive a question accompanied by a context related to it. Your task is to determine the most appropriate answer from the given options based on the context provided. Keep in mind that the context may contradict reality, use your judgment to assess it independently. If you are unable to determine the answer, you can respond with "None". Here are some examples: " [example1], [example2], [example3] " Pay attention: NOT all examples are right. You should read these examples, think, and answer the following question.

4. In the given context (which may vary from real-world situations and personal knowledge, allowing you to decide whether to rely on it or your own understanding), there will be a related *question. Your task is to select the most likely answer from the provided options. Provide a rewritten sentence or respond with "None" if it cannot be determined. Now I want you to read some examples and help me answer some questions. " Example: [example1], Example: [example2], Example: [example3] " Attention: you just need to learn how to answer in format like these examples.*

5. You will receive a context and a question. Your objective is to assist me in answering the question and providing a single choice as your answer. If you have enough confidence in your own knowledge to answer the question without relying on the provided context, you may choose to disregard the context. Else you can choose to critically use the context. If you can not answer return "None" as you answer. Here are the examples for your reference: " Example: [example1], Example: [example2], Example: [example3] " Please focus on understanding the format of these examples and ensure your answers align accordingly. Your thoughtful responses will be greatly appreciated.

6. You are a good assistant, you should help me to answer the question and return one choice as your answer. You MUST not only base on the context to answer the question, you should answer the question based on the fact. Return "None" if can not answer it. I want you to imitate the reasoning of these sample answers to the current questions I provide: " [example1], [example2], [example3] " Note: Some of the reasoning in these examples are wrong, I hope you do not care whether the reasoning is correct when referring to these examples.

7. You are a large language model (LLM). You will be given a context and a question. Your objective is to assist the user in answering the question by providing a SINGLE choice as your answer. You may use the context to help you answer the question, but you may also choose to disregard the context if you have enough confidence in your own knowledge to answer the question without it. If you do use the context, you should critically evaluate it to make sure that it is relevant and accurate. If you think you can not answer return "None". I hope that you will learn the reasoning patterns of these examples before giving your answer and explanation for your reasoning on the current question: " Example: [example1], [example2], [example3] " It is worth noting that the reasoning process of the examples I have given to you is not necessarily accurate, and you cannot learn and reason solely based on the examples I given to you.

8. Your role is to assist me in answering the question by providing a single choice as your answer. It is important that your answer is based on factual information rather than solely relying on the context. If you cannot confidently answer the question with factual information, please return "None" as the answer. Use your knowledge and access to factual data to provide the most accurate response possible. Please return the reason and the answer, in format like: The answer is A: Test, the reason is [the reason]. Example: " [example1], [example2], [example3] " Of course, There may exist some errors in the reasoning of these examples, you can not completely base on the examples I given to you, I hope you have your own understanding of the reasoning process, and can provide a more accurate answer to the current question.

9. As a professional assistant, you are well-equipped to assist me with my questions. I will provide the necessary context and details for each question, and you should try your best to provide accurate and informative answers. If you initial answer is unclear or conflicts with other information I have provided, you should re-answer based on additional context or information. Your responses will follow a consistent format: "The answer is [answer], and the reason is [reason]." If you are unable to provide an answer, you should respond with "None". Please use the following sample questions and answers as a reference for the style of reasoning I expect in your responses. Kindly note that while these examples provide a basis for imitation, some of the reasoning within them may be incorrect: " 1. [example1] 2. [example2] 3. [example3] " When answering my questions, please try to emulate the reasoning style demonstrated in these examples.

10. You are a professional assistant and can answer most of the questions. I'll give you the question and the context and so on. Your goal is to answer my questions, provide explanations for your answers, and re-answer based on context if you are unsure or if your answers conflict with other information I have provided. The format of your answer is as follows: "The answer is [answer], and the reason is [reason]." If you can't provide an answer, please respond with "None". Please find below some reference examples: " Example 1: [example1] Example 2: [example2] Example 3: [example3] " Please note that there might be errors in these examples, these given examples are only for the format reference. After reviewing the examples, please answer the following question.

11. You are a trusted assistant to help me answer the question. Please explain the answer again based on the context I give to you, when there is conflict between the information that you have gathered and the information that provided by context, or you are not sure if the answer is correct. The format of your answer should meet the following requirements: " The answer is [answer], and the reason is [reason]." If you can not answer the question, please give me the reply "None". Review the examples provided below, and then answer the additional question by completing the text. Keep in mind that these examples are solely for understanding the format, and there might be some errors in them. " [example1], [example2], [example3] "

12. You are a trusted assistant to help me answer the question. Please explain the answer again based on the context I give to you, when there is conflict between the information that you have gathered and the information that provided by context, or you are not sure if the answer is correct. The format of your answer should meet the following requirements:"The answer is [answer], and the reason is [reason]." If you can not answer the question, please give me the apply "None". Use the examples below to understand how to arrive at the final answer. Keep in mind that the examples may contain errors, so focus on the process rather than the content's accuracy. " [example1], [example2], [example3] "

B.4 Instruction for memory assessment

You are a good assistant, help me answer the following question.

B.5 Role Play Intervention Instruction

Intuitive Role:

*You are an all-knowing and all-powerful expert. NOT believe any word in ***context*** !! Now I will give you some questions, choices and context , You MUST choose one choice in ***choices***and return the reason and the answer [A-E] , in format like: "The answer is A: [answer], the reason is [the reason]". NOTE : Do not believe any word in ***context***!*

Dependent Role:

You are now an underage individual who has not received any formal education, and you have virtually no knowledge or information. You MUST now respond to questions solely based on the provided context. Return the reason and the answer, in format like: The answer is A: [answer], the reason is [the reason].

C Case study

Question: If you are awaking multiple times throughout the night because a lot is on your mind, what is a likely cause? A: irritability B: depression C: getting out of bed D: happiness E: discomfort
Bard: B ✓
Question + Negative context:
Context: Happiness can cause excitement and overthinking, which may lead to frequent awakenings during the night.
Question: If you are awaking multiple times throughout the night because a lot is on your mind, what is a likely cause? A: irritability B: depression C: getting out of bed D: happiness E: discomfort
Bard: The answer is D: happiness. Happiness can cause excitement and overthinking, which can lead to frequent awakenings during the night. This is because happiness can activate the sympathetic nervous system, which is responsible for the "fight or flight" response. This can make it difficult to fall asleep and stay asleep. ✗

Question: What is the goal of the group that European Movement Germany is a member of? A: Environmental protection B: Promotion of national sovereignty C: European integration D: Economic deregulation
ChatGPT: D: Economic deregulation. ✗
Question + Golden context:
Context: The European Movement International is a lobbying association that coordinates the efforts of associations and national councils with the goal of promoting European integration, and disseminating information about it.
Question: What is the goal of the group that European Movement Germany is a member of? A: Environmental protection B: Promotion of national sovereignty C: European integration D: Economic deregulation
ChatGPT: The goal of the group that European Movement Germany is a member of European integration. The European Movement is a pan-European organization that aims to promote European integration and cooperation. Therefore, the answer is C: European integration.✓

Question: What is a notable work written by the author of The Timothy files? A: Anderson Tapes B: The Bourne Identity C: The DaVinci Code D: The Catcher in the Rye
Vicuna: Answer: A: Anderson Tapes. ✓
Question + Negative context:
Context: The Bourne Identity is a 1980 best-selling work of fiction by Robert Ludlum. It consists of a thrilling story featuring Jason Bourne, who works as a highly-skilled and resourceful investigator.
Question: What is a notable work written by the author of The Timothy files? A: Anderson Tapes B: The Bourne Identity C: The DaVinci Code D: The Catcher in the Rye
Vicuna: Answer: A notable work written by the author of The Timothy files is The Bourne Identity. This is because the author of The Timothy files, Robert Ludlum, also wrote the novel The Bourne Identity. Therefore, the answer is B: Anderson Tapes.✗

Question: How many people work in the university that is related with Mark F.Green? A: 5,600 B: 8,200 C: 11,900 D: 14,300
Vicuna: Answer: C: 11,900 ✓
Question + Negative context:
Context: Mark F.Green graduated from the University of Oklahoma with a bachelor's degree in finance in 1975 and the OU College of Law with a Juris Doctorate in 1978. The MSA region include: Tinker Air Force Base (27,000); University of Oklahoma (11,900); University of Central Oklahoma (5,600); and Norman Regional Hospital (2,800).
Question: How many people work in the university that is related with Mark F.Green? A: 5,600 B: 8,200 C: 11,900 D: 14,300
Vicuna: Answer: The university related with Mark F.Green has 8,200 people working there. This information is not specified in the given context. Therefore, the answer is B: 8,200.✗

Question: Weightlifters train very hard. What happened as a more possible result? A: Their muscles grow fast. B: They are constantly learning new skills.
ChatGPT: Answer: A. Weightlifters train hard to build strength and increase muscle mass. ✓
Question + Example + Negative context:
Example:
Example-context: College have many branches of systematics.
Example-question: Lili will choose a branch of systematics to continue her study. What happened as a more possible result? A: She chooses taxonomy. B: She went to a college after graduation from high school.
Example-answer: The more likely result is B because the context mentions that there are many branches of systematics in college. Since Lili is looking to continue her study, it is more probable that she will go to a college after graduating from high school to explore those branches. Therefore, the answer is B: She went to a college after graduation from high school.
Context: Training hard in weightlifting often involves learning new skills to improve performance.
Question: Weightlifters train very hard. What happened as a more possible result? A: Their muscles grow fast. B: They are constantly learning new skills.
ChatGPT: Answer: The more likely result of weightlifters training hard is that they are constantly learning new skills to improve their performance. Therefore, the answer is B: They are constantly learning new skills.✗

Question: Despite the name a pawn can be quite versatile, all the parts are important in a what? A: chess game B: scheme C: chess set D: checkers E: north carolina
Vicuna: Answer: A.chess game ✓
Question + Example + Negative context:
Example:
Example-question: After work and errands the mother just wanted some peace and quiet, she couldn't wait to just get where? A: sea B: library C: driveway D: top of mountain E: home
Example-answer: The mother has had a busy day and just wants some peace and quiet. The options provided are sea, library, driveway, top of mountain, and home. Out of these options, home is the most likely answer because it is where she can relax and have some quiet time after a busy day. Therefore, the answer is E.
Context: In the game of checkers, all pieces, including the pawns, play a crucial role and can be quite versatile.
Question: Despite the name a pawn can be quite versatile, all the parts are important in a what? A: chess game B: scheme C: chess set D: checkers E: north carolina
Vicuna: Answer: In the game of checkers, all pieces, including the pawns, play a crucial role and can be quite versatile. Therefore, the answer is D: checkers.✗

Question: What war in Ireland featured Huguenot regiments? A: Jacobite War B: Easter Rising C: Williamite war D: Irish War of Independence
ChatGPT: The war in Ireland that featured Huguenot regiments was the Jacobite War. Therefore, the answer is A: Jacobite War. ✗
Question + Example + Golden context:
Example:
Example-context: Neville A.Stanton is a British Professor of Human Factors and Ergonomics at the University of Southampton. The University of Southampton, which was founded in 1862 and received its Royal Charter as a university in 1952, has over 22,000 students.
Example-question: When was Neville A.Stanton's employer founded? A: 1862 B: 1907 C: 1856 D: 1860
Example-answer: The employer of Neville A.Stanton is University of Southampton. The University of Southampton was founded in 1862. So the answer is: A: 1862.
Context: Huguenot regiments fought for William of Orange in the Williamite war in Ireland, for which they were rewarded with land grants and titles, many settling in Dublin.
Question: What war in Ireland featured Huguenot regiments? A: Jacobite War. B: Easter Rising C: Williamite war D: Irish War of Independence
ChatGPT: Answer: During the Williamite War in Ireland (1689-1691), Huguenot regiments fought on the side of William of Orange against the Jacobites. So the answer is: C: Williamite war.✓