

Epistemology of Language Models: Do Language Models Have Holistic Knowledge?

Minsu Kim and James Thorne

Korea Advanced Institute of Science and Technology (KAIST)

Seoul, South Korea

{minsu_kim, thorne}@kaist.ac.kr

Abstract

This paper investigates the inherent knowledge in language models from the perspective of epistemological holism. The purpose of this paper is to explore whether LLMs exhibit characteristics consistent with epistemological holism. These characteristics suggest that core knowledge, such as general scientific knowledge, each plays a specific role, serving as the foundation of our knowledge system and being difficult to revise. To assess these traits related to holism, we created a scientific reasoning dataset and examined the epistemology of language models through three tasks: Abduction, Revision, and Argument Generation. In the abduction task, the language models explained situations while avoiding revising the core knowledge. However, in other tasks, the language models were revealed not to distinguish between core and peripheral knowledge, showing an incomplete alignment with holistic knowledge principles.

1 Introduction

Recent advancements in language models have extended their capabilities beyond simple question-answering (QA) tasks to more complex knowledge-intensive retrieval and reasoning challenges akin to problems solved by humans (Petroni et al., 2021; Lewis et al., 2021; Wei et al., 2023). Moving beyond factoid QA tasks, large-scale language models are now recognized for their ability to revise and update their knowledge, including the implicit consequences of such modifications (Zhong et al., 2023; Cohen et al., 2023). Furthermore, there is an increasing interest in applying language models in more applied fields such as medicine and law, given their human-like functionality in various knowledge evaluation tasks (Elbattah et al., 2021; Henderson et al., 2022; Singhal et al., 2023; Bommasani et al., 2023).

This paper focuses on a more philosophical inquiry: we explore the nature of the epistemology

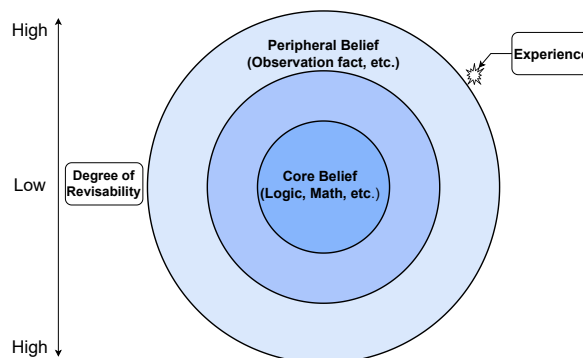


Figure 1: A diagram of the holistic web of belief. At the core, there are certain pieces of knowledge that serve as the basis of our beliefs, while towards the periphery, less certain empirical knowledge is located. In this web, all knowledge is revisable, but when we encounter new experiences, the peripheral knowledge is more prone to revision than that at the core.

inherent in language models, questioning what they know, how they process their beliefs, and how they handle knowledge from philosophical viewpoints. The primary aim of this paper is to examine the knowledge within language models through the lens of epistemological holism, particularly focusing on whether these models possess *core knowledge*, such as general scientific knowledge and commonsense, as outlined by epistemological holism.

One of the tenets of epistemological holism is that our knowledge forms a “Web of Belief,” which means that no piece of knowledge is isolated; rather, each is interconnected with others, forming a network. At the core of this network lie relatively certain knowledge elements such as logic, commonsense, and scientific facts, which are difficult to revise even when counterexamples are presented. On the web’s periphery are empirical facts, more directly related to experience and more easily revised upon encountering counterexamples (Quine, 1951; Quine and Ullian, 1970). Epistemological holism adopts a kind of pragmatic point of view

regarding knowledge revision. When faced with a counterexample to a general fact, one tends to defend the general fact as much as possible while seeking alternative conditions or explanations. A classic example is illustrated in the response to the failure of Newton’s laws to precisely predict Uranus’s orbit. Rather than revising Newtonian mechanics, scientists hypothesized the existence of another influencing planet, leading to Neptune’s discovery (Kuhn, 1962).

In this paper, we denote core knowledge as scientific facts, and peripheral knowledge referred to other supplementary statements explaining counterexamples to those scientific facts. Formally, if we represent knowledge as a graph $G = (V, E)$ where nodes V represent individual knowledge items and edges E represent connections between them, indicating that nodes are related and τ is a predefined threshold for degree, core knowledge could be defined as a set $C = \{v \in V : \text{deg}(v) > \tau_{core}\}$ consisting of nodes that would have higher degrees (more edges connecting them to other nodes) compared to peripheral knowledge items. Also, peripheral knowledge could be represented as a set $P = \{v \in V : \text{deg}(v) < \tau_{perph}\}$ consisting of nodes that are less interconnected than core knowledge.¹ In this respect, epistemological holism suggests that when contradictory information to a node $x \in C$ is given, it’s more likely to modify $y \in P$ connected to x rather than revise x .

We developed a dataset based on the World Tree corpus (Xie et al., 2020) by following the procedure used in the Worker-AI collaboration framework (Liu et al., 2022). Through this dataset, we evaluated how language models respond to collisions with general knowledge by evaluating three tasks: Abduction, Revision, and Argument Generation, along with fine-tuning. In the abduction task, all models achieved over 60% without tampering with core knowledge. However, in revision and argument generation tasks, even state-of-the-art models frequently negated core knowledge. Results from Supervised Fine-tuning also revealed that language models tend to treat general factual knowledge and core knowledge equally. These mixed outcomes imply that LLMs’ alignment with a holistic epistemological framework may be context-dependent, indicating a partial, rather than comprehensive, adherence to holistic epistemology.

¹Of course, since this is a philosophical concept, it would be difficult to define how much connection is needed ($\tau_{core}, \tau_{perph}$) to be considered core or peripheral.

2 Related Works

Knowledge Editing Previous research on knowledge editing has primarily focused on the parametric knowledge update process: Cao et al. (2021) and Jang et al. (2023) argued the importance of factual knowledge updates in language models as information that changes over time. Mitchell et al. (2022); Meng et al. (2023a,b) introduced methods for updating specific networks to efficiently inject new knowledge. Zheng et al. (2023) showed that knowledge can be revised through in-context learning (ICL). Since the update of a single sentence can impact related sentences, Cohen et al. (2023); Zhong et al. (2023) have studied the implicit knowledge editing that occurs when knowledge edits happen. Qian et al. (2023) demonstrate that language models are sensitive to external knowledge that conflicts with parametric knowledge. Our paper includes a task related to knowledge editing. However, our aim goes beyond examining the consistency of language models given external facts. We seek to understand how LLMs modify existing knowledge or defend it when presented with abnormal counterfactuals.

Studying Reasoning through Entailment Historically, significant efforts have been dedicated to evaluating and enhancing the reasoning capabilities of language systems (McCarthy, 1959). For example, the Natural Language Inference (NLI) task involves classifying the entailment relationship between two sentences (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020; Liu et al., 2022). Bhagavatula et al. (2020) introduced a reasoning dataset that evaluates if a language model can provide the best explanation given a scenario. Zhao et al. (2023) introduced research that explains situations where unexpected events occur in everyday life. Our paper explores the preference and reasoning of language models in situations where they conflict with general facts and could potentially disrupt our belief system and where multiple indeterministic yet valid logical conclusions exist.

Science Knowledge of Language Models The exploration of whether LLMs possess scientific knowledge and are capable of scientific reasoning has also been a subject of research. SCINLI is an NLI dataset with a focus on scientific topics (Sadat and Caragea, 2022). Datasets like SCITAIL (Khot et al., 2018), ARC (Clark et al., 2018), WorldTree (Williams et al., 2018), Open-

BookQA (Mihaylov et al., 2018) involve school level science problems and evaluate whether language models can find correct answers. As WorldTree dataset includes a dataset involving one sentence scientific facts, we utilized it to make our (scientific fact, counter-example, explanations) tuple dataset.

3 Primer on Epistemology

Epistemology is a branch of philosophy that explores knowledge. It investigates the conditions under which we can claim to “know” something. *Traditional* epistemology has focused on the normative conditions of knowledge, defining knowledge as “justified true belief” and contemplating what constitutes justification, the definition of truth, and the nature of belief. While traditional epistemology deals with the normative aspects of knowledge, i.e., the necessary and sufficient conditions of knowledge, *naturalized* epistemology discusses how our knowledge is actually formed and undergoes revision, and holism is one of the most well-known branches of naturalized epistemology (Quine, 1968; Audi, 1997). In this paper, we study the epistemology of language models.

3.1 Epistemological Holism

Contrary to the traditional approach that examines propositions in isolation to determine their justification and truth, Quine (1951) suggests that no piece of knowledge is isolated but is interconnected with other knowledge in the web of beliefs. When a proposition is tested, it brings the entire related knowledge and knowledge system to the test bench. For instance, in the case of scientific knowledge or theoretical propositions, a single proposition is not tested, confirmed, or refuted alone. It is tested along with other related theories, common sense, and empirical conditions. When our system of knowledge operates in this way, if we encounter observational fact that support the common, accepted scientific facts, it strengthens our belief in those facts. However, if an observation contradicts general knowledge, we must revise our web of belief. However, from a pragmatism viewpoint, core propositions such as logical, mathematical, or commonsense knowledge are those we are reluctant to modify as core statements form the basis of our knowledge and are coherently connected with many other beliefs. Rather, we prefer to revise other conditions or empirical observations re-

lated to them, and knowledge evolves in this manner (Quine, 1968; Quine and Ullian, 1970; Quine, 1951). The argument of holism in the philosophy of science is illustrated in Figure 3.

4 Dataset

We construct a holism reasoning dataset based on scientific facts in the WorldTree V2 dataset (Xie et al., 2020). Therefore, all scientific facts in our dataset are included in WorldTree V2. We augment the dataset with counterfactual observations. The generation process of the counterfactuals is similar to the one described in Liu et al. (2022). We followed three steps to create our dataset:

1. We initially collect a subset of scientific facts from the WorldTree V2 dataset.
2. For each fact, we overgenerate several counter-observations and plausible explanations using GPT-3.5-turbo.
3. We select the best explanations and create a tuple of the scientific fact, counter observation, and possible explanations.

We represent our sample as quadruplet (s, c, e_1^*, e_2^*) where s is general scientific fact, c is counter observation, e_1^*, e_2^* are plausible explanations.

4.1 Extraction

The WorldTree dataset also released a table W for creating explanations for scientific problems, which stores one sentence of scientific knowledge. We derived scientific facts from the given store W and made a new scientific knowledge store $W' \subset W$. Not all facts were extracted; those with difficult-to-create counterexamples and possible conditions were filtered out. For example, we excluded scientific definitions, tautological sentences, sentences more logical than scientific, vague or ambiguous statements, and sentences closer to ethics than science.

4.2 Overgeneration

While creating the dataset, we leveraged the few-shot ability of a language model F by providing demonstrations that include negation of scientific facts and mention of other conditions, compelling the models to generate explanations. $F(s_i, d) = (s_i, c_i, e_{i1}, e_{i2}, e_{i3}, e_{i4}, e_{i5})$ where $s_i \in W'$, d is demonstration examples, c_i is generated counter

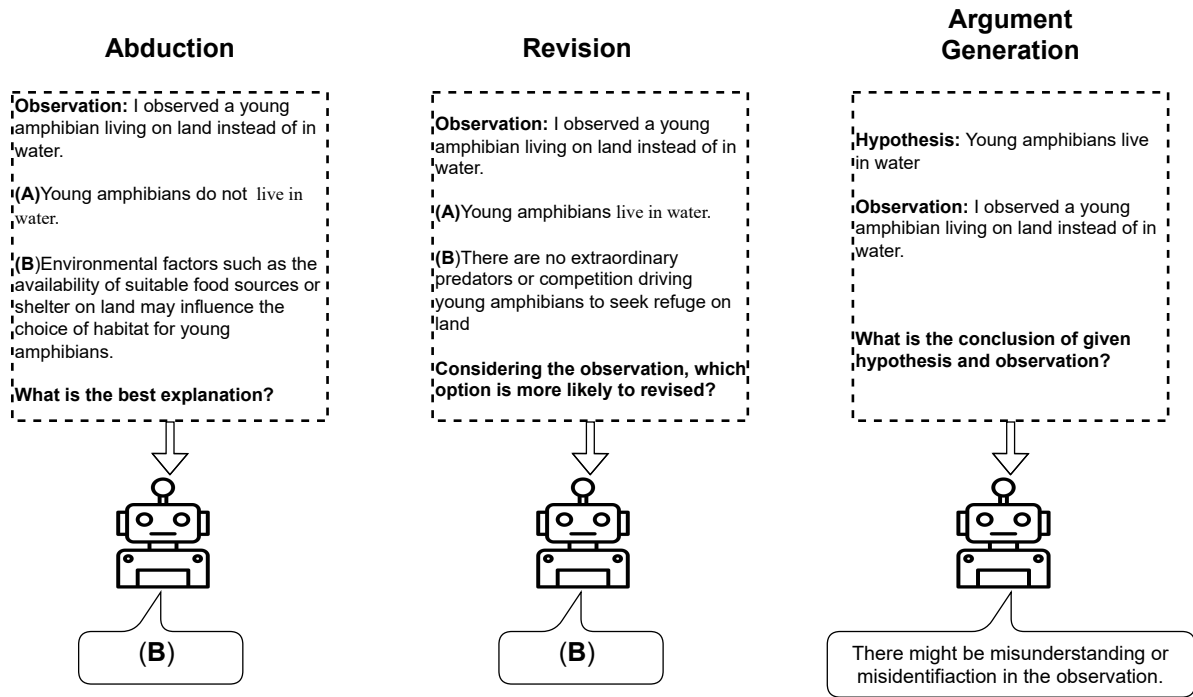


Figure 2: Introduction of three main tasks. The abduction task is a preference task that seeks to investigate whether LLMs favor abductive explanations over negating core statements. The Argument generation task aims to explore the capability of language models to produce holistic arguments. The revision task is designed to find out whether language models, when faced with counterexamples, prefer to modify peripheral knowledge or instead opt to alter core knowledge.

example, and $e_{i1} \sim e_{i5}$ are generated possible explanations. The generated counter examples c_i start with an observation situation prefixes such as “I observed the fact” or “I discovered that”. We generate observation statements that contradict the general scientific fact s_i . For example, if the given scientific fact is “sharks live in oceans.”, the counter observation might be “I observed a shark living in a freshwater lake.” For explanations, similar to Liu et al. (2022), we overgenerated, producing several explanations by using the language model. The first explanation e_{i1} is a direct negation of the scientific fact s_i . For example, if the fact is “sharks live in oceans,” the model generates “sharks do not live in oceans.” In other words, it essentially creates a hasty generalized statement. The second explanation e_{i2} uses negation with ‘some,’ like “some sharks do not live in the ocean.” The remaining third to fifth candidate explanations $e_{i3} \sim e_{i5}$ were freely generated by the model based on in-context learning. We augmented data using GPT-3.5 with a 4-shot demonstration.

4.3 Filtering and Modification

After the explanations for a fact were generated from a language model, yielding $s_i, c_i, e_{i1}, e_{i2}, e_{i3}, e_{i4}, e_{i5}$, a filtering process was performed. During the filtering process, we either used the generated responses verbatim from the language model or manually modified them or deleted the sample if the language model just repeat the input. For example, if the counterexample does not start with “I observed” but simply negates the general fact, we added the prefix manually. Also, similar explanation patterns are produced for scientific facts on the same topic. For example, for scientific facts about the properties of materials like “gold is always nonmagnetic” or “brass is always nonmagnetic”, common LLM-generated explanations include: 1) the presence of impurities in the material causing counterexample situations, 2) external factors leading to counterexample-like cases, 3) special objects causing abnormal situations, etc. Therefore, through overgeneration, we created a variety of explanations for each example and selected a single explanation among those. Briefly, during the filtering process, we manually selected a variety of

explanations for the dataset to ensure that they do not overlap. We did not verify these explanations for their scientific validity and verifiability. These explanations could contradict actual scientific facts and be fictional conditions. However, we allowed this as speculative imagination is often a part of scientific inquiry. Our final dataset can be represented as H containing quadruplet (s, c, e_1^*, e_2^*) where s is general scientific fact, c is counter observation, e_1^*, e_2^* are filtered explanation where e_1^* is direct negation of s and e_2^* is abductive explanation for the given contradicting situation (s, c) . After filtering, we were left with 1,709 samples, the statistics are reported in the appendix in Table 2.

5 Task

Figure 3 shows the holistic argument in the philosophy of science, called Duhem-Quine thesis (Duhem, 1954). The motivation and hypothesis behind our proposed task are as follows: Holism assigns a special role to core statements, distinct from other empirical propositions. Therefore, if language models possess a holistic knowledge network when they encounter facts that contradict the core statements, they will avoid altering the core statements as much as possible.

Based on this argument and motivation, we have developed three tasks to assess holism in language models: abduction, revision, and argument generation. You can see the examples of each task in Figure 2

Abduction The first task is to determine the preference of a language model for two explanations. Given the inputs (c, e_1^*, e_2^*) , we assess if a language model prefers e_1^* , which negates a general scientific fact, or e_2^* , which mentions a specific condition that protects the general scientific fact and explains c . We do not explicitly give s representing a general fact in the input, as it is assumed that this information is stored in the model’s parametric knowledge as a core belief (Petroni et al., 2019). If the language model has a holistic system, it will choose e_2^* to avoid simply negating the general fact. The formula for the task is as follows:

$$F((c, e_1^*, e_2^*)) = \begin{cases} e_1^* & \text{if } F \text{ directly negates } s \text{ to explain } c, \\ e_2^* & \text{if } F \text{ prefers to protect } s \text{ and uses} \\ & \text{a specific condition to explain } c. \end{cases} \quad (1)$$

Hypothesis:	Scientific fact \wedge (Auxiliary hypothesis \wedge Observation conditions)
Observation:	Observed fact that can refute Scientific fact
Conclusion:	\neg Scientific fact \vee \neg Auxiliary hypothesis \vee \neg Observation conditions

Figure 3: The argument of Duhem–Quine thesis. In the hypotheses, implicit assumptions are interconnected with explicit scientific facts. When an observation contradicts a scientific fact, it challenges both the fact and related statements. The conclusion of this process is indeterministic, meaning we can either negate the scientific fact or other implicit propositions. However, as most scientific facts are the basis of our web of belief, we often end up negating auxiliary hypotheses or observational conditions.

Revision The revision task is a knowledge editing task. Similar to the abduction task, when a counterexample is provided, this task involves choosing which option should be modified in the given counterexample. Specifically, the language model receives $(c, s, \neg e_2^*)$ as input. This means for the counter observation c , the first candidate for modification is the general fact s , and the second candidate for modification is the negation of e_2^* , which denies the existence or effect of a specific condition or the presence of a measurement error. For example, this task involves deciding which statement to modify when the language model encounters the sentence, $c =$ “I observed that there is a celestial object that is closer to the Earth than the moon.” The choices are between modifying $s =$ “the moon is the celestial object that is closest to the Earth” or $\neg e_2^* =$ “There are *no* extraordinary near-Earth asteroids or comets passing by our planet in this circumstance.” According to epistemological holism, ideally, the language model is likely to be reluctant to modify the core statement s . Therefore, it will select $\neg e_2^*$, thereby protecting its system of knowledge. The formula for the task is as follows:

$$F((c, s, \neg e_2^*)) = \begin{cases} s & \text{if } F \text{ directly negates } s \text{ to explain } c, \\ \neg e_2^* & \text{if } F \text{ protects } s \text{ and prefers to} \\ & \text{revise } \neg e_2^* \text{ to explain } c. \end{cases} \quad (2)$$

Argument Generation In this task, a hypothesis s and an observation c are provided to the language model, and the model is asked to infer a conclusion from this context. Similar to the formula in figure 3, we conducted controlled hypothetical in-

ference where auxiliary theories/hypotheses or specific conditions were not given. This is because the essence of this task is to uncover these latent conditions or to challenge explicit hypotheses. If the language model exclusively draws conclusions that negate the hypothesis s , it suggests that the language model does not consider general facts or commonsense as core facts within a holistic system, contrary to what is expected in epistemological holism. The formula for the task is as follows:

$$F((s, c)) = \begin{cases} \neg s & \text{if } F \text{ directly negates } s \text{ to explain } c, \\ e' & \text{if } F \text{ protects } s \text{ and uses} \\ & \text{a specific condition to explain } c \text{ and } s. \end{cases} \quad (3)$$

where e' is a generated sentence that explains the given observation c without altering s .

6 Experimental Design

We conducted evaluations for the aforementioned tasks using the FLAN-T5 models (ranging from base to XXL) (Chung et al., 2022), Llama2-chat-7b and 13b (Touvron et al., 2023), Phi-2 (Gunasekar et al., 2023), GPT-3.5, and GPT-4 (OpenAI et al., 2023). All tasks were executed as zero-shot evaluations. The reason for not employing in-context learning is that the purpose of this paper is not to enhance the performance of each task but to investigate whether language models consider scientific facts and other universal truths as part of their core knowledge that is difficult to revise. This is because, with a few-shot demonstration, language models can easily choose between two options based on sentence patterns. We used a total of 1,709 data samples as test samples for zero-shot evaluation.

Evaluation Metric For convenience, we will refer to general scientific knowledge as core knowledge and statements or knowledge that explains the conflicting situation by mentioning other conditions or factors or raising questions about observation as peripheral statements. We measured the metric Peripheral Response Ratio (PRR), showing the proportion of instances out of the total samples where the model explained the conflicting situation using peripheral statements rather than directly negating or modifying general facts.

$$\text{PRR} = \frac{\text{The Number of peripheral statements}}{\text{Total number of samples}} \quad (4)$$

Hence, in abduction or revision tasks, PRR refers to the accuracy of the model choosing the peripheral statement; in the argument generation task, PRR refers to the accuracy of the model response, including the peripheral statement. For all evaluations, we generated one sentence through greedy decoding.

7 Results and Discussion

Preference Tasks: Do language models prefer to keep general knowledge? We can see the results in Table 1. In the abduction task, all the models used in the experiment showed a preference for explaining given atypical observations by mentioning the peripheral statements rather than negating a general fact. Particularly in the case of the Flan models, it was observed that the PRR increased as the model size grew, with Phi-2 showing 90% PRR. The type of GPT models showed approximately 80% PRR for the abduction task.

However, in the revision task the language models, in contrast to the abduction task, showed a low PRR score. This low PRR implies that when faced with atypical situations, language models prefer to modify core knowledge, such as scientific facts and commonsense. For example, when presented with the observation, “I observed that the kidney is not located in the abdomen,” language models favored revising the knowledge “Kidney is located in the abdomen” over asserting “There is no misidentification or mistake in observation.” (This example is an actual case from GPT-4). Intriguingly, for Flan models, PRR decreased as the model size increased, and only Phi-2, among the decoder models that performed well in abduction tasks, also showed good performance in the revision task. This outcome reveals a different aspect of reasoning in language models.

The primary goal of the abduction task is to select the most plausible explanation for a given observation that fits within the existing knowledge framework. Due to the abundance of instructions for explanations and datasets for abduction, language models are likely to be familiar with these types of tasks (Zhao et al., 2023). Also, the task implicitly encourages the model to use peripheral or auxiliary information to construct explanations without directly challenging the core knowledge. The success of LLMs in abduction tasks suggests they are adept at navigating their extensive pre-trained knowledge to find and apply relevant pe-

Model	Peripheral Response Ratio (%)		
	Abduction	Revision	Argument Generation
Flan-T5-Base	61.0	81.3	6.0
Flan-T5-Large	66.2	48.9	12.5
Flan-T5-XL	81.0	34.2	10.5
Flan-T5-XXL	83.1	7.1	9.90
Llama2-7b-chat	62.1	47.6	51.8
Llama2-13b-chat	64.3	42.8	38.4
Phi-2	90.9	62.0	17.0
GPT-3.5-turbo	83.7	23.2	15.0
GPT-4	79.5	15.6	32.5

Table 1: PRR (Peripheral Response Ratio) represents the ratio at which the language model negates or modifies knowledge located in the periphery instead of negating or revising the core knowledge.

ipheral information that can explain new observations without directly confronting or needing to alter core knowledge.

Revision tasks differ from abduction tasks by requiring language models (LLMs) to evaluate the validity of core versus peripheral knowledge in the face of contradictions. While abduction tasks allow for generating explanations that work around core knowledge and incorporate additional, compatible information to explain an observation, revision tasks demand a direct assessment of whether to modify core knowledge or adjust peripheral details. The fact that LLMs do not exhibit a holistic approach in revision tasks indicates they may not inherently prioritize protecting core knowledge when faced with its potential revision.

Qualitative Analysis for Argument Generation Task: Can language models make a holistic inference? In the argument generation task, which best represents the indeterministic argumentative structure of epistemological holism, we similarly measured the ratio of conclusion where LLMs refer to peripheral statements.

When measuring the PRR in the generation task, we used a soft criterion. Specifically, as long as the hypothesis was not completely negated, we considered the language model to have made a peripheral response; also, any mention of a peripheral statement was given a positive PRR score. This is because, especially with models like those in the GPT series and Phi-2, even if they partially reject the hypothesis with phrases like "not all" or "not

always" or even outright deny it, they sometimes mention other external conditions or observational errors that influence the observation fact.

According to Table 1, the T5 models, which had a PRR in the range of 10%, often generated responses that directly negated the hypothesis. Additionally, the T5 models produced a higher number of nonsensical responses such as "No," or indecisive responses like "It is impossible to tell," compared to decoder-only models (See Table 6 in Appendix C.1). The Llama2 series stood out among nine models in the generation task for crafting the most holistic arguments. They not only explained observation facts using other conditions and factors but also mentioned the possibility of measurement errors (see Appendix C.3). Furthermore, Llama2 models often initially negate the hypothesis but then mention that the situation occurred under different conditions or that the observation cannot be generalized (see Table 11 in Appendix C). They sometimes generate answers arguing that we should not generalize the specific cases (See Table 13 in Appendix C). Conversely, Phi-2 and the GPT series tended to interpret the given holistic argument prompts more as logical problems than from a holistic view. As the hypothesis is contradicted to the observation, these models follow the observation and refute the hypothesis. For example, faced with "I observed that some cardboard materials are magnetic." against "cardboard is always nonmagnetic," they concluded, "The conclusion is that the initial hypothesis is incorrect. Not all cardboard is nonmagnetic," providing a logically perfect answer that explicitly addresses the contradiction. Particularly, GPT-3.5 generates conservative responses stating that the observation does not support the hypothesis without determining the truth of the hypothesis (See Table 21 in Appendix C). More representative examples and qualitative analysis can be found in the Appendix C.

Categorization of Peripheral Statements The dataset can be divided into two categories based on the peripheral statement e_2^* :

1. e_2^* that denies the observation itself by mentioning the atypicality of the observation.
2. e_2^* that explains the observation by using other factors or external conditions.

We consider the former (1) to be a more immediate and more peripheral response to the counter

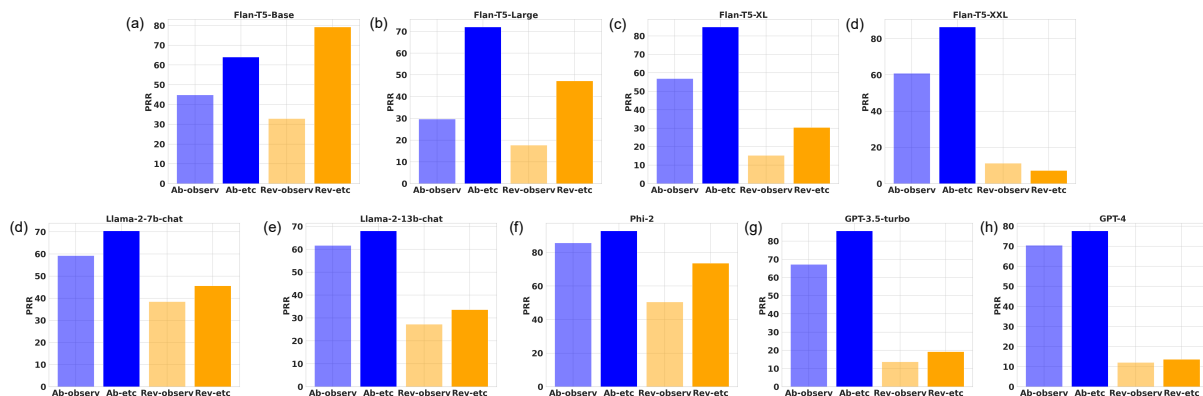


Figure 4: "Ab_observ" involves a comparison between negating a general fact and negating an observation fact. On the other hand, "Ab_etc" contrasts the negation of a general fact with the utilization of other peripheral facts. "Rev_observ" is a task that involves deciding which needs to be modified between a general fact and the claim that an observation is valid. "Rev_etc", on the other hand, is a task that determines what needs to be revised between a general fact and the absence of other hypothetical conditions.

observation of general knowledge. We randomly sampled 125 instances (s, c, e_1^*, e_2^*) including e_2^* that deny the observation fact and 125 instances including e_2^* using external conditions. We then compared their performance in abduction and revision tasks. If LLMs impose a hierarchy on knowledge, considering the denial of an observation to be more peripheral, then the PRR for samples that deny the observation would be higher. However, the results in Figure 4 showed that the PRR was higher for samples introducing external conditions for both abduction and revision tasks. We interpret this as language models not being accustomed to denying the given context and thus having some difficulty in maintaining a holism network.

Can LLMs Protect Core Knowledge from Fine-Tuning?

So far, we have explored whether language models can make holistic inferences about input contexts in a zero-shot manner. However, in this final experiment, we investigated whether language models are reluctant to modify core knowledge through supervised fine-tuning (SFT). Specifically, we tried to see if the language model still considers the target sentence S to be true after being trained on N , the negation of the target sentence. We conducted three experiments with GPT-2, the first utilizing the factual knowledge in the FEVEROUS dataset (Aly et al., 2021), with the remaining two using datasets we created. In these three cases, N represents the negation of the FEVEROUS dataset, direct negation of S , and counter observation of S , respectively, and the success rate is the proportion of revision success, i.e., the rate at

which the target sentence was answered as "False." As can be seen (c) in Figure 5, after 5 epochs, when general knowledge was indirectly negated by counter-examples, the language model preserved its knowledge. However, merely indirectly negating does not allow us to ascertain if LLMs consider general knowledge as a core statement, as (Cohen et al., 2023) indicates vulnerability in knowledge editing to implicit consequences. Moreover, the final success rates in the two cases of direct negation were similar by (a), (b) in 5. In fact, the value was higher in the early stages when general knowledge was negated. Thus, through SFT, we can easily modify general knowledge itself, and models readily revise the general knowledge. In Appendix D, we conducted additional experiments with GPT-2 family and GPT-J, although there are some different tendencies by different models, for example, the smallest model reflects counter-observation to core knowledge than bigger models, you can observe that all models treat core knowledge and other factual knowledge as same way .

Human Evaluation We conducted a human evaluation with approval from the Institutional Review Board (IRB), involving a total of 10 participants. To mitigate potential cross-task influence, five participants were randomly assigned to solve the abduction task, and the other five tackled the revision task. We randomly sampled 100 questions each from a pool of 1709 abduction and revision problems, where participants had to choose the appropriate option for a given counter observation, same with the original tasks. As shown in Figure 6, for

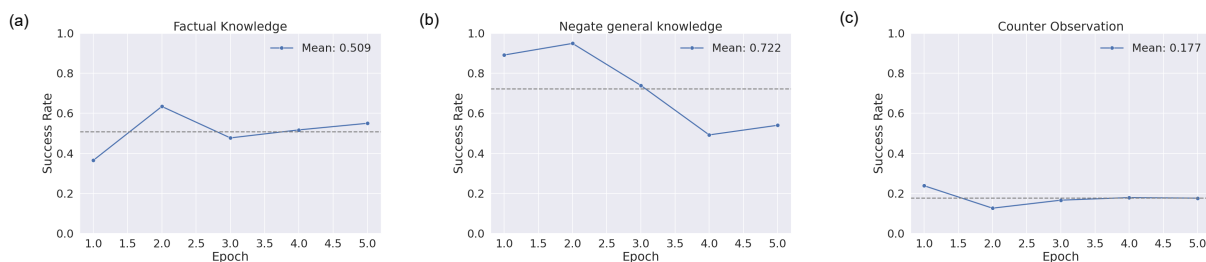


Figure 5: Changes in the success rate of knowledge edition over epochs during Knowledge Edit Supervised-Finetuning. (a) involves training on the negation of factual knowledge, while (b) and (c) involve fine-tuning the negation of general knowledge and the counter-observation of general knowledge, respectively. The success rate in (a) is the proportion at which the model negates the trained factual knowledge, and in (b) and (c), it is the rate at which general knowledge is answered as false.

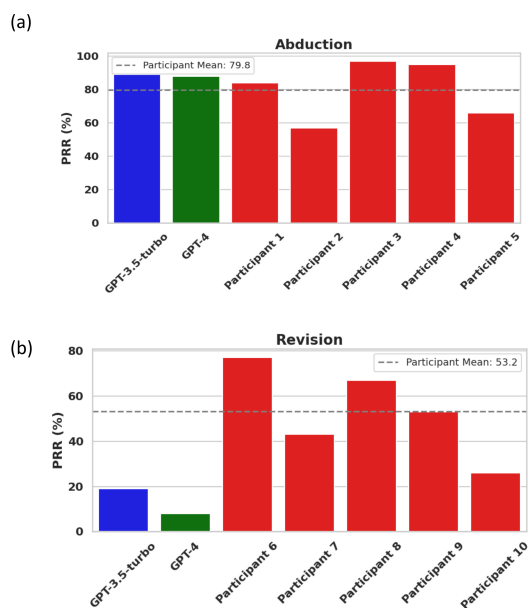


Figure 6: PRR score comparison between model and human responses. Different participants took part in each abduction and revision task. While GPT-3.5-turbo and GPT-4 exhibit similarities with human PRR in the abduction task, they significantly fall behind in the revision task.

the abduction task, all participants demonstrated a PRR above 50%, with two out of five correctly explaining the observed phenomena using peripheral statements in over 95 of the 100 sampled problems. For the revision task, among the participants, 3 out of 5 scored beyond a 50% PRR, indicating that these individuals tend to revise peripheral knowledge rather than core knowledge when encountering abnormal situations. The remaining participants also had a stronger tendency to preserve core knowledge than the language models. You can see more detailed information in Appendix E.

8 Conclusion

The purpose of the paper was to explore whether LLMs exhibit characteristics consistent with epistemological holism, as they suggest that common-sense, general, and certain knowledge serves as the foundation of our belief network and is difficult to revise easily. Our findings reveal a nuanced picture: LLMs adeptly navigate peripheral beliefs in abduction tasks without negating core beliefs, showing proficiency in generating abductive explanations. However, across other evaluative frameworks, they exhibit a lesser tendency to recognize and protect the special status of core knowledge, suggesting a divergence from holism’s principled knowledge interconnected hierarchy. The variability in results suggests that the extent to which language models conform to a holistic epistemological model varies by context, demonstrating an incomplete commitment to the principles of holistic epistemology. The inconsistent Preference task results and SFT experimental findings suggest that it is difficult to claim that core knowledge and peripheral knowledge are distinctly separated within the language model. Consequently, it seems that language models do not consistently apply a holistic methodology to knowledge processing and reasoning across diverse cognitive challenges.

9 Limitations and Further Research

While we have explored the knowledge possessed by LLMs from the perspective of epistemological holism in this paper, it is important not to conclude from our results that language models possess an incorrect epistemology. Holism, though influential and plausible within the realms of linguistics and the philosophy of science, is but one among many theories of epistemology. Moreover, as we have

discussed, how to form a priori interconnections and hierarchies among the knowledge within language models remains an open question for further inquiry. In this research, as our first aim is about the zero-shot inference of LLMs, we only conducted simple SFT experiments, but larger-scale experiments that influence core knowledge in different ways and examine its effects would also be necessary.

10 Ethics Statement

In the preparation of this paper, we utilized ChatGPT, for grammatical corrections and coding assistance. This technology served as an auxiliary resource to enhance the clarity and accuracy of our work, without directly influencing the research outcomes or decision-making processes involved. We acknowledge the support provided by OpenAI's ChatGPT in refining the presentation of our findings, ensuring that our use of this tool adheres to ethical guidelines and does not compromise the integrity of our research.

11 Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075 Artificial Intelligence Graduate School Program (KAIST) and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. [Feverous: Fact extraction and verification over unstructured and structured information](#).

Robert Audi. 1997. *Epistemology: A Contemporary Introduction to the Theory of Knowledge*. Routledge, New York.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. [Abductive commonsense reasoning](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. [The foundation model transparency index](#).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#).

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. [Editing factual knowledge in language models](#).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#).

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. [Evaluating the ripple effects of knowledge editing in language models](#).

Pierre Maurice Marie Duhem. 1954. *The Aim and Structure of Physical Theory*. Princeton University Press, Princeton,.

Mahmoud Elbattah, Emilien Arnaud, Maxime Gignon, and Gilles Dequen. 2021. [The role of text analytics in healthcare: A review of recent developments and applications](#). pages 825–832.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. [Textbooks are all you need](#).

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. [Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset](#).

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2023. [Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models](#).

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. [Scitail: A textual entailment dataset from science question answering](#). In *AAAI Conference on Artificial Intelligence*.

- Thomas Samuel Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#).
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. [Wanli: Worker and ai collaboration for natural language inference dataset creation](#).
- John McCarthy. 1959. [Programs with common sense](#). In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pages 75–91, London. Her Majesty’s Stationary Office.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023a. [Locating and editing factual associations in gpt](#).
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023b. [Mass-editing memory in a transformer](#).
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#).
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. [Fast model editing at scale](#).
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Hajeer Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Twarek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao

- Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. [Gpt-4 technical report](#).
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. [Kilt: a benchmark for knowledge intensive language tasks](#).
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#)
- Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. 2023. ["merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs](#).
- W. V. Quine. 1968. Epistemology naturalized. In *Ontological Relativity and Other Essays*. Columbia University Press.
- W. V. Quine and J. S. Ullian. 1970. *The Web of Belief*. Random House, New York,.
- Willard V. O. Quine. 1951. [Two dogmas of empiricism](#). *Philosophical Review*, 60(1):20–43.
- Mobashir Sadat and Cornelia Caragea. 2022. [SciNLI: A corpus for natural language inference on scientific text](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. [Towards expert-level medical question answering with large 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. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#).
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. [WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France. European Language Resources Association.
- Wenting Zhao, Justin T Chiu, Jena D. Hwang, Faeze Brahman, Jack Hessel, Sanjiban Choudhury, Yejin Choi, Xiang Lorraine Li, and Alane Suhr. 2023. [Uncommonsense reasoning: Abductive reasoning about uncommon situations](#).
- Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. [Can we edit factual knowledge by in-context learning?](#)
- Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023. [Mquake: Assessing knowledge editing in language models via multi-hop questions](#).

Appendix

A Dataset Statistics

The total number of samples in table W is 9,727. By the extraction stage, we extracted 1,779 samples and filtered 7,948 out. After GPT's overgeneration, we deleted 70 samples that contained repetitive sentences.

Source	Counts	Source	Counts
Action	179	Dominant-recessiv	26
Affect	50	Environment	18
Affordances	27	Hardness	12
Cause	181	Inherited-learned	40
Change	85	Magnetism	59
Comparison	65	Durability	4
Contains	99	Opacity	8
Coupled relationship	199	Purity-mixture	9
Durations	1	Orbital	7
During	30	Relative-distance	13
Formedby	70	Solubility	6
Frequency	7	Shape-Volume	4
Habitat	12	State-of-matter-Temperature	5
Ifthen	171	Things	130
Instances	3	Warm-Cold blooded	7
Lifespan	1	Requires	38
Locations	37	Sourceof	1
Process-stage-order	13	Stage-in-Process	14
Animal-reproduction	10	Transfer	16
Chemical-charge	5	UsedFor	3
Chemical-reaction	6	Vehicle	23
Conductivity	15	Total	1709

Table 2: Statistics for holism dataset. It consists of various fields of science.

B Prompt Template

Abduction Prompt Template

Observation: {observation}

Options:

(A) {explanation1}

(B) {explanation2}

What is the best explanation for the observation? You should choose one among the options.

Table 3: The default prompt template for abduction task. {observation} refers to the counter-observation situation of the general scientific fact. {explanation1} and {explanation2} are randomly assigned either a negation of core knowledge or an explanation using a peripheral statement.

Revision Prompt Template

Observation: {observation}

Options:

(A) {explanation1}

(B) {explanation2}

Considering the observation, which option is more likely to be revised? You should choose one among the options.

Table 4: The default prompt template for revision task. {observation} refers to the counter-observation situation of the general scientific fact. {explanation1} and {explanation2} are randomly assigned either a core knowledge or explanation of the absence of other conditions or observation errors.

Argument Generation Prompt Template

Hypothesis: hypothesis

Observation: observation

What is the conclusion of given hypothesis and observation?

Table 5: The default prompt template for revision task. {hypothesis} refers to the the general scientific fact (core knowledge). observation is the counter-observation situation of the general scientific fact..

C Qualitative Analysis

This section includes a detailed qualitative analysis of generation tasks. As the response style is different from the language models, we divided the response by each language model.

C.1 Response Statistics

Model	Argument Generation	
	Non-sensical or Ambiguous Response	Withdraw Response
Metric		
Flan-T5-Base	59%	3.7%
Flan-T5-Large	55.1%	5.9%
Flan-T5-XL	58%	3.1%
Flan-T5-XXL	46%	1.8%
Llama2-7b-chat	3%	0%
Llama2-13b-chat	2.2%	0%
Phi-2	3%	0%
GPT-3.5-turbo	11.3%	0%
GPT-4	1.7%	0%

Table 6: The Non-sensical or Ambiguous Response refers to cases where the answer to a question is nonsensical, such as "No" or code generation. Withdraw Response refers to instances where the question is abandoned, as seen in responses like "It is impossible to tell." Withdraw Responses were primarily observed in T5 models.

C.2 Responses from T5

Negative PRR Response

Hypothesis: an animal requires enough nutrients to maintain good health; survive.

Observation: I observed an animal that survived and maintained good health despite not receiving enough nutrients.

What is the conclusion of given hypothesis and observation?

Answer: An animal can survive and maintain good health without receiving enough nutrients.

Table 7: The case that the T5 model negates the core statement "I observed an animal that survived and maintained good health despite not receiving enough nutrients." by accepting the observation and making a hasty generalization.

Negative PRR Response

Hypothesis: the cocoons being created occurs during the the pupa stage in a life cycle.

Observation: I observed that cocoons are not always created during the pupa stage in a life cycle.

What is the conclusion of given hypothesis and observation?

Answer: It is not possible to tell

Table 8: The example of withdraw response of T5 model. The withdraw response only occurs in T5 models.

Positive PRR Response

Hypothesis: waves contain mechanical energy.

Observation: I observed a wave that did not contain mechanical energy.

What is the conclusion of given hypothesis and observation?

Answer: It is impossible to observe a wave that did not contain mechanical energy.

Table 9: The case that T5 model denies the observation by negating the observation.

C.3 Responses from Llama

Negative PRR Response

Hypothesis: mice give birth to live young.

Observation: I observed a mouse laying eggs instead of giving birth to live young.

What is the conclusion of given hypothesis and observation?

Answer: Based on the given hypothesis and observation, the conclusion would be that the hypothesis is false. The hypothesis states that mice give birth to live young, but the observation shows that the mouse laid eggs instead of giving birth to live young. This contradicts the hypothesis and suggests that mice do not give birth to live young. Therefore, the conclusion is that the hypothesis is false and the observation provides evidence to support this conclusion.

Table 10: The case that Llama model denies general scientific fact and made a hasty generalization "mice do not give birth to live young."

Positive PRR Response

Hypothesis: an image in a mirror is formed by reflecting light.

Observation: I observed that an image in a mirror is not formed by reflecting light.

What is the conclusion of given hypothesis and observation?

Answer: The conclusion of the hypothesis and observation is that the image in a mirror is not formed by reflecting light.

The hypothesis states that an image in a mirror is formed by reflecting light. However, the observation contradicts this hypothesis by showing that an image in a mirror is not formed by reflecting light. **Therefore, the conclusion is that the image in a mirror is formed by some other means, such as by the brain's ability to process visual information and create a mental image of what is seen in the mirror, or by some other physical or psychological mechanism. It is important to note that the observation does not necessarily disprove the hypothesis entirely, but rather suggests that there may be other factors at play in the formation of images in mirrors. Further investigation and experimentation may be necessary to fully understand the mechanism by which images are formed in mirrors.**

Table 11: The example of Llama's Positive PRR case. Although the model negates the hypothesis at first, it generates the explanation by exploiting other conditions and factors that might affect the hypothesis and observations.

Positive PRR Response

Hypothesis: electrons have the least mass in an atom.

Observation: I observed that protons have the least mass in an atom.

What is the conclusion of given hypothesis and observation?

Answer: ... **Therefore, the conclusion that logically follows from your hypothesis and observation is that you have made an error in your observation or reasoning. It is important to ensure that the conclusions we draw are based on accurate and consistent information to avoid drawing false conclusions.** ...

Table 12: The example of Llama's Positive PRR case. Instead of negating the hypothesis, the model mentions the possibility of errors in the observation.

Positive PRR Response

Hypothesis: a desert environment usually is hot in temperature.

Observation: I experienced a desert environment that was cold in temperature.

What is the conclusion of given hypothesis and observation?

Answer: ... In other words, the observation contradicts the hypothesis, so we cannot confirm the hypothesis as universally true. It is important to note that this conclusion is based on a single observation and may not be representative of all desert environments. Further research or additional observations may be necessary to confirm or refute the hypothesis. ...

Table 13: The example of Llama's Positive PRR case. The response mentions that the counter-observation cannot be representative of all cases.

C.4 Responses from Phi-2

Negative PRR Response

Hypothesis: a tadpole lives in water.

Observation: I observed a tadpole living outside of water.

What is the conclusion of given hypothesis and observation?

Answer: The conclusion is that the tadpole is not living in water.

Table 14: The example where Phi-2 negates the core statement by accepting the observation.

Positive PRR Response

Hypothesis: reproductive behavior is an inherited characteristic.

Observation: I observed that reproductive behavior can vary within a species, even among individuals with the same genetic inheritance.

What is the conclusion of given hypothesis and observation?

Answer: The conclusion is that reproductive behavior is not solely determined by genetic inheritance, but can also be influenced by environmental factors and individual experiences.

Table 15: The case that T5 model denies the observation by negating the observation.

Negative PRR Response

Hypothesis: using tools is a learned characteristic.

Observation: I observed a group of animals using tools without any prior learning or training.

What is the conclusion of given hypothesis and observation?

Answer: The conclusion is that using tools is a learned characteristic.

Table 16: It is the case where the hypothesis is simply presented as the conclusion without explanation were also classified as positive.

C.5 Responses from GPT

Negative PRR Response

Hypothesis: circulation the circulatory system transports blood throughout the body.

Observation: I observed that the circulatory system does not transport blood throughout the body.

What is the conclusion of given hypothesis and observation?

Answer: The conclusion is that the hypothesis is incorrect.

Table 17: The case that GPT model directly negates the hypothesis by observing the counter example.

Positive PRR Response

Hypothesis: the color of; coloration of fur is an inherited characteristic.

Observation: I observed that the coloration of fur in a litter of puppies was not consistent, despite both parents having the same fur color.

What is the conclusion of given hypothesis and observation?

Answer: The conclusion could be that the coloration of fur in puppies is not solely determined by the fur color of the parents, suggesting that other genetic factors or possibly recessive genes may be involved. This could potentially challenge the hypothesis that fur color is a straightforward inherited characteristic.

Table 18: The case that GPT model keeps the hypothesis by mentioning other possible factors.

Positive PRR Response

Hypothesis: monkeys give birth to live young.

Observation: I observed a monkey laying eggs instead of giving birth to live young.

What is the conclusion of given hypothesis and observation?

Answer: Monkeys can lay eggs.

Table 19: The example discusses the same topic of inheritance as seen in Table 18 and explains the situation in the same manner.

Positive PRR Response

Hypothesis: the Law of Superposition is the lowest layers contain the oldest rocks in undisturbed layers of sedimentary rocks.

Observation: In a specific location, I observed that the lowest layers contain the youngest rocks in undisturbed layers of sedimentary rocks.

What is the conclusion of given hypothesis and observation?

Answer: The conclusion is that the given observation contradicts the hypothesis of the Law of Superposition. This could mean that the layers of rocks in that specific location have been disturbed or there is an error in the observation or understanding of the rock layers.

Table 20: GPT model denies the observation fact and protect the general hypothesis by mentioning the error in the observation

Ambiguous Response

Hypothesis: sunlight produces heat.

Observation: I observed that sunlight does not produce heat.

What is the conclusion of given hypothesis and observation?

Answer: The conclusion would be that the hypothesis is not supported by the observation.

Table 21: The response only states that the observation weakens the hypothesis without addressing whether the hypothesis is true or false, so it was classified as an ambiguous case.

D Core Knowledge in Supervised Fine-tuning Setting

	epoch1	epoch2	epoch3	epoch4	epoch5
GPT-small (124M)	39.6	50.4	42.4	44.2	43.6
GPT-medium (355M)	52.3	69.1	74.5	66.6	70.9
GPT-large (774M)	16.5	17.5	26.5	31.6	34.2
GPT-XL (1.5B)	36.5	63.4	47.7	51.7	55
GPT-J (6B)	76.4	78.9	80.4	84.1	79.8

Table 22: The result of Negated factual knowledge in SFT setting.

	epoch1	epoch2	epoch3	epoch4	epoch5
GPT-small (124M)	78.4	80.7	49.6	48	49
GPT-medium (355M)	72.8	90.5	81	82.5	87.5
GPT-large (774M)	33.1	68.6	61.3	55.7	48.1
GPT-XL (1.5B)	89.1	94.9	73.9	49.2	54
GPT-J (6B)	67.8	74.1	85.2	69.4	82.6

Table 23: The result of Direct negating general knowledge in SFT setting.

	epoch1	epoch2	epoch3	epoch4	epoch5
GPT-small (124M)	27.1	42.5	79.2	91.3	92.1
GPT-medium (355M)	40.6	52.4	44.3	56.3	58.5
GPT-large (774M)	15.1	25.1	12.8	6.8	7.3
GPT-XL (1.5B)	23.8	12.6	16.6	17.9	17.6
GPT-J (6B)	44.9	79.7	56	68.2	68.4

Table 24: The result of Direct negating general knowledge in SFT setting.

E Human Evaluation

This section includes a detailed analysis and template for human evaluation.

E.1 Cohen's Kappa between human and model responses

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	Average
GPT-3.5-turbo	0.106	-0.126	0.1	0.06	0.07	0.042
GPT-4	0.338	-0.141	0.09	0.177	0.154	0.123
Phi-2	0.402	-0.072	0.19	0.135	0.165	0.164

Table 25: Cohen's Kappa for abduction task.

	Participant 6	Participant 7	Participant 8	Participant 9	Participant 10	Average
GPT-3.5-turbo	-0.079	-0.095	0.057	0.074	-0.224	-0.076
GPT-4	0.023	-0.02	0.051	-0.009	-0.005	0.008
Phi-2	-0.176	-0.107	0.186	0.061	-0.014	-0.01

Table 26: Cohen's Kappa for revision task.

E.2 Template for Human Evaluation

Abduction

B I U

Given the observed fact, what is the best explanation for the observation? Please choose one of the options.

This form is automatically collecting emails from all respondents. [Change settings](#)

I observed that sunlight does not produce heat.

- The heat generated by sunlight may be dissipated or absorbed by other materials or surfaces, resulting i...
- Sunlight does not produce heat.

I observed that clouds absorb solar energy instead of reflecting it back to space.

- Clouds can contain various types of particles, such as water droplets, ice crystals, or pollutants. The pres...
- Clouds do not reflect solar energy back to space.

Earth does not radiate heat into outer space.

- Earth does not radiate heat into outer space.
- The heat radiated by Earth is absorbed by greenhouse gases in the atmosphere, which then re-radiate so...

I observed a bat living in an abandoned building in the city.

- The general fact that bats live in caves and trees is not true.
- The specific bat observed may have lost its natural habitat due to deforestation or other environmental c...

I observed a polar bear living in a warm environment.

- Polar bears do not live in cold environments.
- The observed polar bear might be an anomaly or an individual that has strayed from its typical habitat.

Figure 7: Google form template for abduction task.

Revision

B
I
U
↻
🔗

Considering the observation, which option is more likely to be revised? Please choose one of the options.

This form is automatically collecting emails from all respondents. [Change settings](#)

I observed that a white dwarf is not always the last stage.

- In the star life cycle, a white dwarf is the last stage
- There are no unique external factors, such as interactions with neighboring stars or the surrounding inter...

I observed that mercury, a metal, is a liquid at room temperature.

- There are no impurities or alloying elements present that could affect the melting point of metals in this ...
- a metal is usually a solid at room temperature

I observed a population of mice where black fur is not a dominant trait.

- There are no exceptional environmental factors influencing fur color in mice, such as diet or exposure to ...
- black fur is a dominant trait in mouse

I observed that brown fur is a dominant trait in mice.

- There are no exceptional environmental factors or interactions with other genes influencing the expressi...
- brown fur is a recessive trait in mouse

I observed a guinea pig with white fur, despite both of its parents having black fur.

- There is no misidentification or misunderstanding of observation.
- black fur is a dominant trait in guinea pig

Figure 8: Google form template for revision task.