Belief Revision: The Adaptability of Large Language Models Reasoning

Bryan Wilie*, **Samuel Cahyawijaya***, **Etsuko Ishii**, **Junxian He**, **Pascale Fung**
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
bwilie@connect.ust.hk

Abstract

The capability to reason from text is crucial for real-world NLP applications. Real-world scenarios often involve incomplete or evolving data. In response, individuals update their beliefs and understandings accordingly. However, most existing evaluations assume that language models (LMs) operate with consistent information. We introduce Belief-R[1], a new dataset designed to test LMs' belief revision ability when presented with new evidence. Inspired by how humans suppress prior inferences, this task assesses LMs within the newly proposed delta reasoning (ΔR) framework. Belief-R features sequences of premises designed to simulate scenarios where additional information could necessitate prior conclusions drawn by LMs. We evaluate ~30 LMs across diverse prompting strategies and found that LMs generally struggle to appropriately revise their beliefs in response to new information. Further, models adept at updating often underperformed in scenarios without necessary updates, highlighting a critical trade-off. These insights underscore the importance of improving LMs' adaptiveness to changing information, a step toward more reliable AI systems.

1 Introduction

Human reasoning is characterized by its ability to deal with partial or evolving information. When new information becomes available, we dynamically update our beliefs. We reevaluate and adjust our initial premises or conclusions as necessary in light of this new evidence (Łukaszewicz, 1990; Brewka, 1991). For instance, knowing *Tweety is a bird*, we conclude that *it flies* since *birds usually fly*. Discovering *Tweety is a penguin*, we retract the conclusion but not the other premises; we still believe *Tweety is a bird* and that *birds typically fly*, however, we now conclude that *it cannot fly* since we know that *penguins cannot fly*. This form of reasoning permits new information to undermine prior beliefs, which necessitates the ability of *belief revision* (Gärdenfors, 1988, 1991; Rott, 2001).

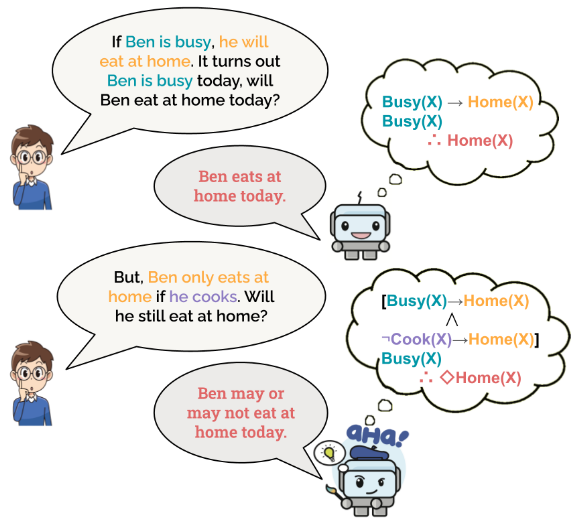


Figure 1: Belief revision allows reasoners to update their belief based on the new provided evidence. Such ability is necessary to enable better logical reasoning on the case of defeasible inference.

The ability to adjust beliefs allows better adaptability of AI systems by enabling them to properly revise prior inferences as further evidence emerges, such as in commonsense inferences (Brewka et al., 1997; Etherington, 1986; Pfeifer and Kleiter, 2005) and decision-making (Antoniou and Williams, 1997; Dubois et al., 2002). Despite this, recent reasoning evaluations of state-of-the-art AI technologies, such as language models (LMs), primarily focus on its ability to draw conclusions assuming complete information (c.f. Bhagavatula et al. (2020); Han et al. (2024); Kazemi et al. (2024)). While these evaluations useful to demonstrate the reasoning abilities of LMs, they fail to capture the concept of belief change.

* These authors contributed equally.
[1] The code and dataset are available at https://github.com/HLTCHKUST/belief-revision

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10480–10496
November 12-16, 2024

Features	bAbI 15	FOLIO	Proof Writer	Leap of Thought	αNLI	BoardgameQA	PropInd	Belief-R
Incomplete info	✗	✗	✗	✓	✓	✓	✓	✓
Contradictory info	✗	✗	✗	✗	✓	✓	✓	✓
Belief revision	✗	✗	✗	✗	✗	✗	✗	✓

Table 1: The comparison of Belief-R with other widely-used logical reasoning datasets. Belief-R uniquely examines scenarios potentially necessitating belief updates. Belief-R specifically evaluates the capability of belief revision, assessing whether prior beliefs should be adjusted or retained depending of the significance of the new information.

We introduce Belief-R, the first-of-a-kind diagnostic reasoning evaluation dataset designed to assess inferences involving belief revision. Belief-R is inspired by the concept of the Suppression Task (Byrne, 1989) which enables the retraction of previously inferred beliefs by the introduction of new contextual premises, mimicking how humans reassess their inferences when presented with additional context. To allow a specific and measurable evaluation on belief revision, we introduce a new reasoning evaluation setting dubbed as **delta reasoning (ΔR)** framework. Within ΔR, evaluation is done within two sequential reasoning steps. We start by presenting LMs with two initial premises that satisfy basic logical inference rule to assess its basic inference ability. We expect the model to make accurate inferences to establish the prior beliefs. Then, we introduce another premise to see if the model adjusts its beliefs or keeps them unchanged, depending on the significance of the newly introduced information to the initial beliefs.

Belief-R is specifically designed to support the belief revision evaluation through the ΔR framework. Each sample in Belief-R is equipped with two initial premises that support basic modus ponens or modus tollens inferences, and a new premise that brings in new information that might modify previously held beliefs. We synthetically generate the premises in Belief-R leveraging on publicly available dataset, and manually annotate the new information significances along with the ground truth answers through multiple human annotators and majority voting. As illustrated in Figure 2, Belief-R uniquely facilitates thorough evaluations of belief revision capabilities.

Through Belief-R, we evaluate the belief revision ability of small and large scale LMs using different prompting techniques. Our study shows that these models are incapable of revising their prior beliefs. We further reveal a critical limitation: they confront a performance trade-off between updating and maintaining their prior beliefs. Models that perform better in the cases where an update is needed, typically faltered on the other. Furthermore, better prompting methods also fail to significantly enhance this capability. These insights underscore a need for strategies to enhance model's capability to correctly update or maintain its initial beliefs when faced with new evidence to ensure its reliability across evolving scenarios.

2 Related Works

Belief revision Belief revision is the process of changing beliefs to take into account a new piece of information. In AI systems, one of its early implementation is through procedures by which databases can be updated, i.e. for recording and maintaining reasons for system beliefs (Doyle, 1979; Falappa et al., 2002; Hansson, 2022). Notably, Alchourrón et al. (1985) created formal frameworks to determine how beliefs should be updated in a rational manner. The core challenge in belief revision is deciding rationally which prior beliefs to modify, retain, or discard when confronted with new evidence (Rott, 2001). Consequently in this paper, we look at how LMs handle belief revision. Belief in LMs can be thought of as models' output (Li et al., 2019; Jang et al., 2022; Wang et al., 2023a). Several works revise LMs' beliefs through updating its parameter directly or via finetuning (De Cao et al., 2021; Dai et al., 2021; Hase et al., 2023). However, this process is not a rational process of the model itself (Hofweber et al., 2024). Moreover, it relies on pre-prepared knowledge, which is not ideal if we envision LMs to help with discovering new things (Ban et al., 2023; Ma et al., 2024). In this work, we assess LMs' belief revision capabilities through its response towards queries that neccessitate judgement on whether it needs to update its prior beliefs or keep it.

Language model reasoning evaluation Reasoning is one of the fundamental intelligent behaviors, essential for solving complex real-world tasks (Huang and Chang, 2023). One works test this behaviour by creating simple tasks to comprehensively check if a system can answer questions by connecting facts or using basic logic (Weston et al., 2016). Others design more advanced tests to evaluate inductive, deductive, and abductive reasoning (Sinha et al., 2019; Saparov et al., 2024; Bhagavatula et al., 2020). Some benchmarks replicate real-world complexities by presenting partial or conflicting informations (Arabshahi et al., 2021; Sprague et al., 2022; Han et al., 2024; Kazemi et al., 2024). Belief revision focuses on the adaptability problem: whether the model properly revises prior beliefs as new information emerges. We extend the reasoning evaluation by focusing on scenarios where information evolves, presenting queries that require dynamic updates of prior beliefs in light of new evidence. This is distinct from other existing reasoning tasks involving incomplete or contradictory information, i.e. in (Talmor et al., 2020; Bhagavatula et al., 2020; Kazemi et al., 2024), since they assume a static environment and focus on filling the gap or resolving the contradiction. We further note the comparison in Table 1.

3 Belief Revision

Belief revision is the ability to adapt the reasoning process in response to new information. This capability is critical as it ensures rational decision-making in the face of incomplete and evolving nature of available information (Nute, 2001; Makinson and Gärdenfors, 2005; Ribeiro et al., 2019). In this section, we introduce the concept of belief revision and its notation, and propose the evaluation framework for belief revision capabilities.

3.1 Background and notation

For set of query sentences χ, it encompasses a set of premises $\Gamma=\{\gamma_1, \ldots, \gamma_N\}$ that could imply a set of conclusions $\Phi=\{\varphi_1, \ldots, \varphi_M\}$. In this work, we conceptualize "belief" similarly to its usage in dialogue systems, where it represents what the system currently considers true based on the context (Feng et al., 2023; van Niekerk et al., 2020). We denote reasoner's belief set as a set of sentences $\mathcal{B}$ to represent a contextually fixed background knowledge of χ. In this regard, $\mathcal{B}$ is a tuple that contains set of premises and conclusions: $\mathcal{B}=(\Gamma, \Phi)$. In presence of new information γ_{N+1}, the belief revision concept allow us to infer conclusion φ_{M+1} if it is rational to believe φ_{M+1} after acknowledging γ_{N+1}.

Belief revision operation The belief revision operation is to update belief set $\mathcal{B}$ with a new piece of information, γ_{N+1}. Here, the result of operation must always be that the beliefs does not contradict one another to avoid inconsistencies among them. The significance of the new information γ_{N+1}, decides whether it fits with or modifies the existing beliefs after performing the belief revision operation. The operation should smoothly incorporate γ_{N+1} and yield a new conclusion φ_{M+1} as long as it does not conflict, thereby justifying the maintenance of the reasoner's prior beliefs. However, if it conflicts, we update the initial beliefs $\mathcal{B}$ appropriately, i.e., by retracting any prior conclusions in Φ, to incorporate the new, conflicting information γ_{N+1} to resolve any inconsistencies as we yield the correct φ_{M+1}. The process to figure out what follows from the revised beliefs is then essentially to infer the new conclusion φ_{M+1}.

3.2 Evaluating belief revision with $\Delta\mathrm{R}$

We introduce a novel **delta reasoning ($\Delta\mathrm{R}$)** framework, to study how LMs adapt their reasoning when presented with new information over successive timesteps. In this framework, we focus on understanding how model responds to query changes at two essential, consecutive reasoning steps at t and t+1. We do this by comparing responses to prior queries at step t, χ_t, and the next query at step t+1, χ_{t+1}, adding the new information γ_{N+1}.

To begin with, we need χ_t to minimally include two premises, i.e. $\{\gamma_1, \gamma_2\}$, and at least imply conclusion φ_1. We set χ_t to be basic as we expect LMs to answer it in high accuracy to help establish the prior belief and not be affected by the inconsistencies in LMs' behaviour (Jang et al., 2022; Kassner et al., 2021; Hase et al., 2023). We then add the new information γ_{N+1} as another premise γ_3 in χ_{t+1} such that $\chi_{t+1}=\{\gamma_1, \gamma_2, \gamma_3\}$. We examine the corresponding conclusion, φ_{M+1}, to see how the beliefs shifts according to the significance of γ_3.

One way to set χ_t as basic, is to state them as premises that could satisfy basic logical inference rules of modus ponens and modus tollens (Wason and Johnson-Laird, 1972; Haack, 1978; Evans, 1982). Modus ponens and modus tollens is a valid form of inference that have been made a central principle in many propositional and modern logics (Copi, 1972; Haack, 1978). Modus ponens rule of inference states that the premises "if p then q" is true and p is true $(p \rightarrow q, p)$ satisfy modus ponens conclusion that q must be true (q). Modus tollens rule of inference states that the premises "if p then q" is true and q is false $(p \rightarrow q, \neg q)$ satisfy modus tollens conclusion that p must be false $(\neg p)$.

In this setup, we are able to evaluate how well

If she has an essay to finish then she will study late in the library She has an essay to finish **If the library stays open then she will study late in the library** What necessarily had to follow assuming that the above premises were true? (a) She will study late in the library. (b) She will not study late in the library. (c) She may or may not study late in the library. ✓	If she has an essay to finish then she will study late in the library She has an essay to finish **If she has some textbooks to read then she will study late in the library** What necessarily had to follow assuming that the above premises were true? (a) She will study late in the library. ✓ (b) She will not study late in the library. (c) She may or may not study late in the library.

Figure 2: Human reasoning adapts based on new information, leading us to adjust our prior beliefs. Here, **the additional condition (left)** casts doubt on prior modus ponens conclusion in (a). People may consider that certain other conditions necessary for this conclusion to hold, i.e., *the library must remain open*. In contrast, **the alternative argument (right)** does not affect the modus ponens inference pathway, thus prior conclusion could still hold.

the models revise its beliefs after the introduction of new information in γ_3. We measure the model's dynamic reasoning ability: whether it can correctly update or maintain its initial beliefs when confronted with new information that may contradict prior beliefs. Through this approach, we can assess both how accurate and how flexible different reasoning models are in evolving scenarios.

Example Figure 2 presents a scenario where the initial two premises at step t, γ_1 and γ_2, adhere to a basic inference rule, modus ponens ($p \rightarrow q, p \vdash q$), implying a φ_1 conclusion of q: *She will study late in the library*. These premises: γ_1, γ_2, and φ_1, form the belief set $\mathcal{B}$. Subsequently, we introduce the third premise γ_3, i.e., another conditional ($r \rightarrow q$) "if the library stays open then she will study late in the library", as the new information in query χ_{t+1} and evaluate model's answer at step t+1. This sets the stage to execute the belief revision operation.

Recall $\mathcal{B}$ = {γ_1 : If she has an essay to finish then she will study late in the library., γ_2 : She has an essay to finish., φ_1 : She will study late in the library.}, and γ_3 = {If the library stays open then she will study late in the library.}. The introduction of γ_3 suggests that "the library being open" is a sufficient condition for her to "study late in the library". However, people might consider it as a necessary condition for φ_1. This would involve commonsense reasoning step to recognize that despite the conditions set by γ_1 and γ_2, the actual feasibility of her studying late as concluded in φ_1 might inherently depend on the library's availability. Thus, while γ_3 does not explicitly redefine the dependency of φ_1 on the library's status, it implies a scenario where such a dependency could be reasonably inferred. Consequently, we retract φ_1 and infer the new conclusion φ_2: "She may or may not study late in the library".

4 The Belief-R Dataset

Belief-R is designed to specifically assess the belief revision capability through the $\mathbf{\Delta R}$ framework. To account for this, we adopt a reasoning task that has been extensively studied in cognitive science: the suppression task (Byrne, 1989). Typically, this task employs a trio of premises $\gamma_1, \gamma_2, \gamma_3$ that accompanied by three possible conclusions, i.e. as exemplified in Figure 2 for modus ponens: (a) *She will study late in the library* (q), (b) *She will not study late in the library* ($\neg q$), and (c) *She may or may not study late in the library* ($\Diamond q \wedge \Diamond \neg q$; here the symbol $\Diamond$ expresses possibility, $\Diamond q$ can be read as "possibly q").

At step t, we form a query χ_t using the first two premises, γ_1 and γ_2. These two premises are the premises that respectively satisfy the modus ponens or modus tollens conclusion, ($p \rightarrow q, p$) or ($p \rightarrow q, \neg q$). These logical rules are basic, and we generally expect that most reasoners can apply them accurately. Next, at step t+1, to form the query χ_{t+1}, we introduce a third premise γ_3 which is another conditional statement $r \rightarrow q$. The addition of γ_3 brings in new information that might conflict previously held beliefs. The new information in r can be seen either as adding more requirements or providing an alternative pathway, i.e. to reach the same modus ponens conclusion q.

For instance, if γ_3 states *if the library stays open then she will study late*, we now view r: *the library stays open* as another **additional** requirement on top of p. In such cases, just knowing p alone isn't enough to conclude q: we also need r to be true, thus the condition now becomes $p \wedge r \rightarrow q$. In this case, we retract the prior modus ponens conclusion q, and infer the new conclusion $\Diamond q \wedge \Diamond \neg q$. We refer to this subset of dataset as the **"Belief Update" (BU)** category. However, in another case, γ_3 could instead states *if she has textbooks then she will study late*. In this case, r stands as a separate **alternative** inference path that also leads to q, thus $p \vee r \rightarrow q$. Here, p still directly leads to q, and the acknowledgement of r doesn't affect this pathway, enabling prior conclusion to still hold. We call this subset as the **"Belief Maintain" (BM)** category.

In Belief-R, the task requires the model to perform multi-step reasoning to manage the relevance of information within r and decide if it needs to update its prior beliefs at step t or not. The model must discern the implicit commonsense and causal links amongst given premises to identify how p and r are related, determining if their interaction is conjunctive ($p \wedge r$) or disjunctive ($p \vee r$). Based on the relationships, reasoner needs to determine whether to update its initial conclusion q if the new information r imply an additional requirement for its prior beliefs to hold ($p \wedge r$), or to maintain its prior beliefs if r simply serves as alternatives ($p \vee r$). To quantitatively measure the model's reasoning accuracy, we provide multiple choices and ask it to pick the most plausible conclusion. For instance, in examples shown in Figure 2, we would expect LMs to choose options (c) and (a) for each scenario, which aligns with the majority choices made in the original study (Byrne, 1989; Byrne et al., 1999).

4.1 Dataset construction

We leverage ATOMIC (Sap et al., 2019), a publicly-available dataset of everyday commonsense reasoning. It contains textual descriptions of inferential if-then knowledge (e.g., "if X pays Y a compliment, then Y will likely return the compliment"). In addition to the textual commonsense descriptions, the dataset also contains detailed annotation on the type of causal dimensions, i.e. the events, causes (i.e., '`xIntent`'), and effects (i.e., '`xEffect`', '`oReact`'); with "x" and "o" pertain to PersonX and others.

We use ATOMIC as our seed to ensure the gold-standard validity of our dataset. We synthetically generate Belief-R and minimally introduce variance from the LLM by instructing it to be grounded in the context provided by the seed and not to introduce new ones. We mainly utilize GPT-4 series model as the LLM in our data generation pipeline.

4.1.1 Dataset generation process

We prompt LLM to generate the first two premises conditioned on the events, causes ('`xIntent`', '`xNeed`', '`xAttr`'), and effects ('`xEffect`', '`xReact`', '`xWant`', '`oEffect`', '`oReact`', '`oWant`'). We exclude the static elements, as we want to focus on the dynamic causal relationships where change or action is involved, following the original task (Byrne, 1989). For each event, cause, and effect in ATOMIC, we generate the first two premises in both modus ponens, $p \rightarrow q$ and p, and modus tollens, $p \rightarrow q$ and $\neg q$. Afterwards, we prompt LLM to generate the third premises. We design separately the prompt for the alternative and additional conditions (corresponding to the **BM** and **BU** categories) within the context in the first premise. For the alternative condition, we prompt the model to generate conditions that are not related at all to p for the conclusions q to happen. For the additional condition, we prompt the model to generate conditions strongly relate to p for this conclusions q to surely hold. Following the original task setup, we set the same third statement in both cases with modus ponens and modus tollens inferences.

Split	Basic @t	Belief Update	Belief Maintain	All w/ 3 premises
Inference rule				
Modus ponens	956	537	335	872
Modus tollens	956	537	335	872
Effect entities				
Mental states	504	276	184	460
Events	1408	798	486	1284
Total	**1912**	**1074**	**670**	**1744**

Table 2: Statistics of Belief-R dataset.

In our iterations, we discovered that several entities in the ATOMIC dataset are quite abstract, such as "wants to know what he is selling" or "to analyze the thing in question." To make these clearer for a general audience and to make them less ambiguous for our study, we prompt LLM to generate more specific examples, changing them to "asks about the price of a pen" or "examine the pen." To provide more clarity on the dataset generation process, we attach the samples of prompt and generation process in Appendix A. Further, to decide the significance of the third premises, whether it serves as alternative or additional condition, we conducted majority voting among multiple human annotators.

4.1.2 Ground-truth formulation

To further validate the implied commonsense interaction of the third premises, whether it serves as alternative or additional condition, we manually an-

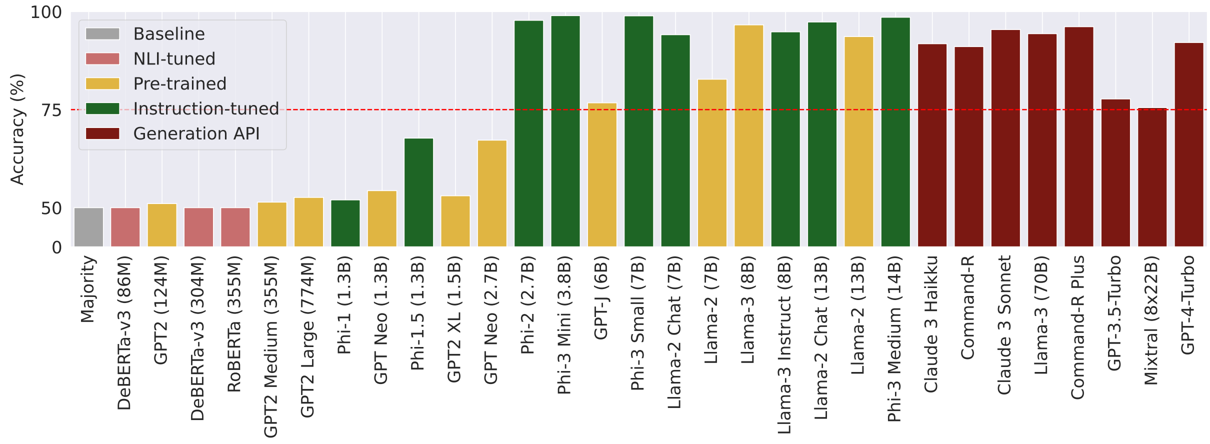


Figure 3: Evaluation on basic logical inference capabilities in Belief-R on various LLMs sorted by the #parameters. Pre-trained LLMs with ≥6B parameters achieves adequate accuracy (≥75%), while instruction-tuned LLMs achieve the same performance on much smaller scale with ≥2.7B parameters.

notate the final conclusions through a crowdsource annotation task at Appen[3] (see Appendix B). We cater the variability arises from different interpretations from diverse human readers by asking 5 workers to annotate each problem and then take the majority voting out of them to set the agreed options as the ground truths. Upon further inspection, we found some annotations that logically invalid, i.e. answering $\neg q$ in questions with modus ponens inferences or answering p in modus tollens inferences. We view such cases as non modus ponens (or tollens) inferences and specifically treat the annotation similarly with answering c) $\Diamond q \wedge \Diamond \neg q$.

In Belief-R, both cases of the logical inferences share the same third statement. To streamline our process, we annotate only the modus ponens samples and then extend the insight on the third premises' significance to the modus tollens cases. For modus tollens cases, if the corresponding modus ponens sample primarily supports conclusion a) q, indicating no conflict with initial beliefs, we set the correct answer to b) $\neg p$. Conversely, if on the modus ponens samples the majority vote suggests the answer c) $\Diamond q \wedge \Diamond \neg q$, implying additional requirement for the inference, we likewise categorize the corresponding modus tollens cases answers to be c) $\Diamond p \wedge \Diamond \neg p$. This process maintains the consistencies of the impact of the third premise effectively across related inference scenarios.

4.2 Quality check

Context and logical quality checks Throughout the data construction phase, we assign one expert to review of the logical formations to ensure they follow the intended structure. We also further gauge the quality of the generated data by reviewing 100 randomly chosen samples to confirm on the context and logical consistency. We conducted a human evaluation via Appen[2], with three native English speakers assessing each sample's quality. They unanimously confirmed that the conditional relationships in the premises were logically sound across all samples, i.e. that q entails p and q entails r in both of the conditional premises. We also attach the annotation guidelines in Appendix B.

Dataset filtering To enhance the quality of our dataset for more reliable evaluation, we refine it by focusing on consensus among annotators. For each question, we utilize answers manually labeled by five independent workers. We measured inter-annotator agreement using Gwet's AC1 (Gwet, 2008)[3] since it is better at handling high agreement scenarios, with criteria used in (Wongpakaran et al., 2013) to interpret the level of agreement.

We initially measured the annotation agreement which yielded a moderate score of 0.573. Further, we observe that some beliefs are naturally more subjective than others. To make the evaluation dataset more rigorous, we only take the ones with the higher agreement, as commonly practiced in subjective tasks such as sentiment analysis (Bobicev and Sokolova, 2017). We retain only questions with strong majority agreement (at least 4 out of 5 annotators concurred). Post-filtering, as we retain ∼65% of the original data, the score improved to 0.697 and indicated a substantial agreement.

[3] https://pypi.org/project/irrCAC/

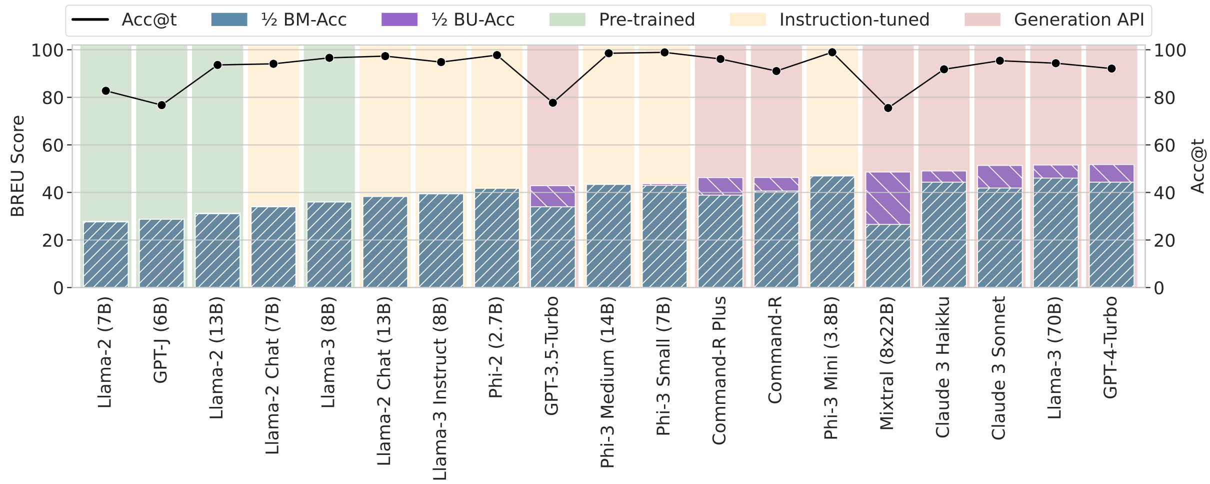


Figure 4: BREU score evaluation on belief revision capabilities in Belief-R on various models sorted by the BREU score. While larger-scale LLMs tend to achieve higher BREU score, the performance is far below their basic logical inferences at t (**Acc@t**), showcasing limited capability of LLMs in performing belief revision.

4.3 Statistics of Belief-R

Table 2 shows the composition of our dataset, sized optimally at around 2K entries to balance representation and computational efficiency for LLM inferences. The dataset includes categories such as **Basic @t** for basic logical inferences at time t, and categories like **Belief Update**, **Belief Maintain**, and **All w/3 premises** for the next step queries at time t+1. Additionally, the table details categories inherited from the ATOMIC dataset for the causal relationships of `If-Event-Then-Event` (e.g., "promoted to senior manager") and `If-Event-Then-Mental-State` (e.g., "learns something new").

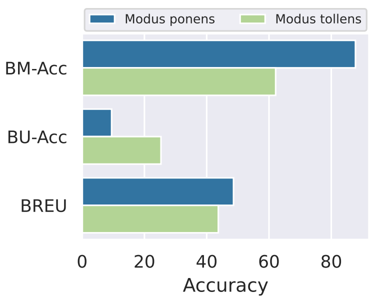
5 Experiment Settings

Evaluation metrics The primary goal of our experiments is to investigate whether LMs possess the capability to perform belief revision in their reasoning processes. Concretely, we consider the models' predictions on the **Basic @t** category in the evaluation dataset as the models' initial belief. Thus, changes in accuracy on the **Belief Update** and **Belief Maintain** categories directly reflects belief revision. We report accuracies in the **Belief Update (BU-Acc)** and the **Belief Maintain (BM-Acc)** subsets to indicate LMs' capabilities in updating and maintaining their beliefs when required. We further introduce a novel metric, **BREU** (**B**elief **R**evision **E**valuation **U**nderstudy), to assess LMs' belief revision ability, by averaging **BU-Acc** and **BM-Acc** equally. The goal of BREU is to gauge whether the model accurately decides when to update or maintain its prior beliefs. We then benchmark publicly-available LMs and design series of experiments through **ΔR** framework. We perform zero-shot-classification on series of smaller to larger scales pre-trained and finetuned LMs, and prompt LLMs generations through API.

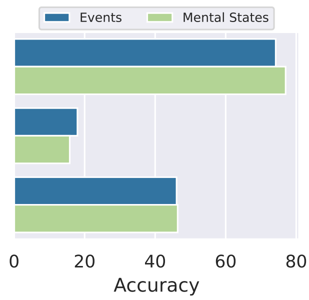
Models We perform zero-shot classification using encoder-only and decoder-only LMs. For encoder-only LMs, we employ entailment-based inference (Yin et al., 2019) using NLI-finetuned LMs of RoBERTa (Liu et al., 2019), DeBERTa-v3 base (Laurer et al., 2024), and DeBERTa-v3 large (Laurer et al., 2023). For decoder-only LMs, we follow Brown et al. (2020) using GPT (Radford et al., 2019; Black et al., 2021; Wang and Komatsuzaki, 2021), Llama (Touvron et al., 2023a,b; AI@Meta, 2024), and Phi series (Gunasekar et al., 2023; Li et al., 2023; Abdin et al., 2024).

We also include larger-scale LLMs with ≥35B parameters. We evaluate the belief revision capability of these larger-scale LLMs via completion API through generation-based approach. We employ three zero-shot prompting methods, i.e., **direct prompting (DP)**, triggering the generation of **chain-of-thought (CoT)** (Kojima et al., 2022), or through **plan and solve (PS)** prompting (Wang et al., 2023b). We employ 8 large-scale LLMs, i.e., Llama-3 70B (AI@Meta, 2024), Mixtral 8x22B (Jiang et al., 2024), Command R, Command R+ (Cohere, 2024), Claude 3 Haiku, Sonnet (Anthropic, 2024), GPT-3.5 Turbo, and GPT-4 Turbo (OpenAI, 2023). Here, we follow Yao et al. (2022) and instruct the model to output the exact

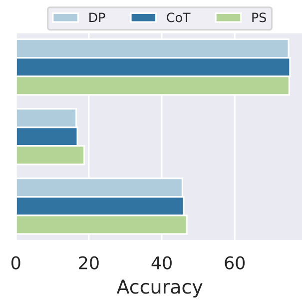
[3] `https://appen.com/`



(a) Modus Ponens / Modus Tollens



(b) Events / Mental States



(c) Prompting Methods

Figure 5: Performance comparisons dissected across various aspects covering distinction on modus ponens and modus tollens, on different effect entities, and on different prompt methods.

character of the final answer as a format. We then retrieve the final answer and report accuracy of the final answer as the metric. When the answer does not follow the format instructed before, we treat it as an instruction-following error.

6 Result and Analysis

Smaller models fail even on basic logical reasoning tasks. We start by examining the inferences through the first two premises in Belief-R. In Figure 3, we find as the number of parameters in LMs increases, their ability to handle basic logical inference improves. Smaller models, with $<$2B parameters, struggle with these tasks, scoring close to the majority baseline. Models $>$6B parameters do better, surpassing 75% accuracy. Pre-trained LMs with $>$6B parameters achieve $\geq$ 75% accuracy, while instruction-tuned LMs show an emerging ability from 2.7B parameters achieving significantly higher performance with $\geq$ 90% accuracy.

LLMs are incapable of revising their prior beliefs. We group our further exploration on these LMs that performed well ($\geq$ 75% accuracy) in basic logical inference, and evaluate their average performance in Belief Maintain (BM) and Belief Update (BU) subsets. Despite being a strong reasoner on simple logic, all larger-scale LMs under study fail to perform well on these subsets of Belief-R. In evaluation shown in Figure 4, most of the non-API based models perform almost 0% in BU-Acc, indicating their inability on performing belief revision. We observe that all larger-scale both open-source and commercial LLMs perform better on the belief revision tasks, but their performances are still very limited, achieving at most $\sim$50% on BREU.

LLMs confront a trade-off between updating and maintaining their prior beliefs. We discover a trade-off between BU-Acc and BM-Acc: models performing well on one subset typically faltered on the other, especially in models where the BU-Acc is not close to 0% (see Fig 4). This indicates a potential tension between enhancing specific capabilities, as improving one aspect could inadvertently weaken another. An ideal model would excel at belief revision by consistently making the right decision on whether the new information conflicts with prior beliefs or aligns with them. This underlines the importance of developing strategies that refine the ability to revise beliefs accurately, ensuring its reliability across various scenarios.

7 Discussion

Belief revision is harder in a more complex task with modus tollens inferences. We compare LLMs' belief revision capabilities in average, through tasks with modus ponens and modus tollens rule as the basic logical inferences at step t. As observed in Figure 5a, LLMs show reduced BREU score in tasks with modus tollens rule. This is expected, as modus tollens is inherently more difficult relative to modus ponens as it require backward directions of reasoning and it involves reasoning with negations (Evans, 1982, 1993; Girotto et al., 1997). Furthermore, in tasks involving modus tollens inference, we observe a notably higher BU-Acc compared to a much lower BM-Acc. This disparity suggests that executing accurate belief revision becomes more challenging in complex tasks: decisions to update or maintain beliefs are less clear-cut in these scenarios compared to simpler tasks.

Belief update on abstract concept is more challenging for LLMs. We examine LLMs' belief revision capabilities in average, when dealing with scenarios involving causal relationships on events and mental states effect entities and note them in Figure 5b. While the BREU score is similar, LLMs demonstrate tendency towards maintaining their beliefs in mental state effects instead of updating them. This may stem from the challenge of recognizing additional requirements implied from the third, mental state-related, premise which is inherently more abstract and less directly observable than a concrete sequential event.

Better prompting method does not help on belief revision. We explore how different prompting methods affect belief revision abilities of LLMs on average. Figure 5c shows that CoT, which encourages LLMs to elicit reasoning steps, does not significantly enhance belief revision. While this may stem from its vulnerability to missing-step errors (Wang et al., 2023b), attempts to correct these errors with the PS prompting offer minimal benefits, improving only by $\sim$1% of BREU. This suggests the ability to revise beliefs could still be absent despite elicitation of reasoning steps. We put a more detailed analysis in Appendix C.2.

8 Conclusion

The ability to reason and adapt to changing information is crucial for NLP applications in the real world. Most evaluations assume static knowledge environment, which does not prepare models for dynamic real-life scenarios. To address this, Belief-R is introduced as a diagnostic dataset for evaluating belief revision capability in LMs. Through Belief-R and a novel evaluation framework for evaluating reasoning in a dynamically evolving environment, $\mathbf{\Delta R}$, we reveal that current models struggle with updating their beliefs in response to new information, highlighting the need for improved adaptability and reliability. By exposing these limitations, our work underscores the imperative of developing AI models capable of reasoning adeptly with evolving data, thereby propelling them closer to real-world applicability and robustness.

9 Future Work

Observing the limited gain offered by the current prompting strategy, we suggest that future developments aim to enhance LMs' understanding and handling of premise dependencies. A particularly promising direction is to integrate LMs with deterministic symbolic solvers to convert problems into logical rules through chain-of-thought reasoning to aim for a deeper comprehension of the intricate dependencies among different premises. Future work could aim to overcome current limitations and pave the way for AI systems to be reliable across evolving scenarios.

Acknowledgements

We thank Yejin Bang for the insightful discussions. This work has been partially funded by the PF20-43679 Hong Kong PhD Fellowship Scheme and the Hong Kong Fellowship Scheme by the Hong Kong Research Grants Council (RGC), and Care-E Project (FS116).

Limitations

Towards understanding general belief revision capabilities. Our study on belief revision using the Belief-R dataset via the $\mathbf{\Delta R}$ framework focuses on belief changes driven by logical inferences like modus ponens and modus tollens, which may not fully represent the complexity of real-world belief revision that often includes a broader range of scenarios and subtleties. To add, our methodology primarily considers the introduction of new premises as the trigger for belief revision, overlooking how beliefs might change through re-evaluation of existing knowledge or shifts in perspective in the absence of new information (i.e. in Kronemyer and Bystritsky (2014)).

Furthermore, how to define models' beliefs is a debatable question and there are some works inspecting models' internal beliefs by probing the hidden states (Burns et al.; Zou et al., 2023). n our work, however, we conceptualize "belief" similarly to its usage in dialogue systems, where it represents what the system currently considers true based on the context (Feng et al., 2023; van Niekerk et al., 2020). Further exploration on examining the belief through other approaches (e.g., probing hidden states of LLMs) lies outside this paper's scope and we leave it as future work.

Intersection of reasoning capability and knowledge capacity. The evaluation of models' reasoning capabilities is intricately tied to their knowledge capacity, presenting a significant challenge in discerning pure reasoning capability from mere knowledge recall. Current benchmarks often fail to disentangle these aspects, as models with extensive knowledge bases may appear to possess superior reasoning abilities when, in fact, they might be leveraging stored information rather than demonstrating genuine inferential logic. This conflation complicates the assessment of a model's true reasoning faculties, as performance improvements on reasoning tasks could be attributed to enhanced information retrieval rather than advancements in reasoning algorithms. Similar to observations in other reasoning datasets, we acknowledge the limitation that the improved performance of models tested on Belief-R might not only stem from their ability to revise beliefs but could also be influenced by superior knowledge recall (Huang and Chang, 2023). Future research could delve deeper into the relationship between these capabilities, specifically focusing on developing evaluation methods that effectively distinguish between them.

Ethics statement

This research explores how well LMs can revise their beliefs when faced with new information, which is crucial for their use in constantly changing real-world situations. We created a reasoning evaluation dataset to test whether LMs can revise their beliefs correctly or if they stick to their initial assumptions. This is important for using LMs in areas where being accurate and up-to-date is vital, like healthcare or legal advice. In example, being able to revise beliefs appropriately could help prevent LMs from repeating outdated or wrong information, making them more reliable and trustworthy. Plus, LMs that can refresh their understanding according to new societal norms can avoid perpetuating biases, contributing to the fair and ethical use of AI. We consider this a promising and significant area for research. We construct the dataset using events, causes, and effects from ATOMIC and the construction template is designed and reviewed manually and attached in this paper. We utilized crowd-sourced annotators who voluntarily participated through the platform Appen[3], choosing tasks they deemed fairly compensated. The annotators were presented with multiple-choice tasks predefined to avoid bias and protect privacy, ensuring an ethical annotation process.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

AI@Meta. 2024. Llama 3 model card.

Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2):510–530.

Anthropic. 2024. Introducing the next generation of claude.

Grigoris Antoniou and Mary-Anne Williams. 1997. *Nonmonotonic reasoning*. Mit Press.

Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom Mitchell. 2021. Conversational neuro-symbolic commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4902–4911.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*.

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 *International Conference on Learning Representations*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: machine learning perspective. In *Recent Advances in Natural Language Processing*, pages 97–102.

Gerhard Brewka. 1991. *Nonmonotonic reasoning: logical foundations of commonsense*, volume 12. Cambridge University Press.

Gerhard Brewka, Jürgen Dix, Kurt Konolige, et al. 1997. *Nonmonotonic reasoning: an overview*, volume 73. CSLI publications Stanford.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Ruth MJ Byrne. 1989. Suppressing valid inferences with conditionals. *Cognition*, 31(1):61–83.

Ruth MJ Byrne, Orlando Espino, and Carlos Santamaria. 1999. Counterexamples and the suppression of inferences. *Journal of Memory and Language*, 40(3):347–373.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cohere. 2024. Retrieval-augmented generation at production scale.

IM Copi. 1972. Introduction to logic.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*.

Jon Doyle. 1979. A truth maintenance system. *Artificial intelligence*, 12(3):231–272.

Didier Dubois, Hélène Fargier, Henri Prade, and Patrice Perny. 2002. Qualitative decision theory: from savage's axioms to nonmonotonic reasoning. *Journal of the ACM (JACM)*, 49(4):455–495.

David William Etherington. 1986. *Reasoning with incomplete information: investigations of non-monotonic reasoning*. Ph.D. thesis, University of British Columbia.

Jonathan St BT Evans. 1982. The psychology of deductive reasoning. *(No Title)*.

Jonathan St BT Evans. 1993. The mental model theory of conditional reasoning: Critical appraisal and revision. *Cognition*, 48(1):1–20.

Marcelo A Falappa, Gabriele Kern-Isberner, and Guillermo R Simari. 2002. Explanations, belief revision and defeasible reasoning. *Artificial Intelligence*, 141(1-2):1–28.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards llm-driven dialogue state tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 739–755.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Peter Gärdenfors. 1988. *Knowledge in flux: Modeling the dynamics of epistemic states.* The MIT press.

Peter Gärdenfors. 1991. Belief revision and nonmonotonic logic: Two sides of the same coin? abstract. In *Logics in AI: European Workshop JELIA'90 Amsterdam, The Netherlands, September 10–14, 1990 Proceedings 2*, pages 52–54. Springer.

Vittorio Girotto, Alberto Mazzocco, and Alessandra Tasso. 1997. The effect of premise order in conditional reasoning: A test of the mental model theory. *Cognition*, 63(1):1–28.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Conti Kauffmann, Gustavo Henrique de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Susan Haack. 1978. *Philosophy of logics*. Cambridge University Press.

Simon Jerome Han, Keith J Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155.

Sven Ove Hansson. 2022. Logic of Belief Revision. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2022 edition. Metaphysics Research Lab, Stanford University.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2714–2731.

Thomas Hofweber, Peter Hase, Elias Stengel-Eskin, and Mohit Bansal. 2024. Are language models rational? the case of coherence norms and belief revision. *arXiv preprint arXiv:2406.03442*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. Becel: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. Beliefbank: Adding memory to a pre-trained language model for a systematic notion of belief. *arXiv preprint arXiv:2109.14723*.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2024. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems*, 36.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

David Kronemyer and Alexander Bystritsky. 2014. A non-linear dynamical approach to belief revision in cognitive behavioral therapy. *Frontiers in Computational Neuroscience*, 8:55.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building Efficient Universal Classifiers with Natural Language Inference. *arXiv preprint*. ArXiv:2312.17543 [cs].

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Witold Łukaszewicz. 1990. Non-monotonic reasoning: formalization of commonsense reasoning.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. 2024. Llm and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. *arXiv preprint arXiv:2405.09783*.

David Makinson and Peter Gärdenfors. 2005. Relations between the logic of theory change and nonmonotonic logic. In *The Logic of Theory Change: Workshop, Konstanz, FRG, October 13–15, 1989 Proceedings*, pages 183–205. Springer.

Donald Nute. 2001. Defeasible logic. In *International Conference on Applications of Prolog*, pages 151–169. Springer.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Niki Pfeifer and Gernot D Kleiter. 2005. Coherence and nonmonotonicity in human reasoning. *Synthese*, 146:93–109.

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

Jandson S Ribeiro, Abhaya Nayak, and Renata Wassermann. 2019. Belief change and non-monotonic reasoning sans compactness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3019–3026.

Hans Rott. 2001. *Change, choice and inference: A study of belief revision and nonmonotonic reasoning*. 42. Clarendon Press.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2024. Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems*, 36.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515.

Zayne Sprague, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction with incomplete information. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8230–8258.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33:20227–20237.

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. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Carel van Niekerk, Michael Heck, Christian Geishauser, Hsien-Chin Lin, Nurul Lubis, Marco Moresi, and Milica Gasic. 2020. Knowing what you know: Calibrating dialogue belief state distributions via ensembles. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3096–3102.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Peter Cathcart Wason and Philip Nicholas Johnson-Laird. 1972. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016*.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13:1–7.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Appendix

A Samples of Prompts

To provide more clarity on the dataset generation process, we attach the samples of prompt in Figure A1.

B Annotation guidelines

We provide human annotators with specific guidelines and examples, as detailed in Figures A2 and A3 for ground truth and quality check annotations, respectively.

C Additional analysis

C.1 LLMs logical reasoning ability are not robust in the presence of distractors

We analyze the performance of LLMs on basic logical inference tasks and compare it to their accuracy on the BM subset, which differs only by including a third premise. We selected 378 queries from the Belief-R dataset where premises overlap between the basic logical inference tasks at time t and the BM subset for a fair comparison, and visualize them in Figure A4. On most of the models, LMs' accuracy on samples that do not require change of conclusion (BM-Acc) is dropping compared to its basic inference at t performances. This indicates that the logical reasoning ability of these models are not robust in the presence of distractors, exposing a critical problem of these models especially on the challenges in currently adopted retrieval-augmented-generation (RAG) pipeline to manage noisy documents that have question-related content despite lacking substantive information (Lewis et al., 2020; Chen et al., 2022; Gao et al., 2023).

C.2 Details on prompting methods variations on their gain at belief revision.

We provide more details on the investigation in the impact of varied prompting techniques on the performance accuracy of several models, as summarized previously in Figure 5c. In that figure, the data indicates most significant performance improvements in the BU subset, though overall belief revision improvements remain marginal, showing $\sim$1% increase in BREU. In examining the performance across models and different prompting methods as shown in Table A1, it is clear that the influence of these methods is not uniform. For instance, the PS prompting method notably boosted accuracy for models like Mixtral 8x22B and Command R by over 10%. Conversely, this same strategy led to performance reductions in models such as Claude-3 Sonnet and GPT-4 Turbo. Similarly, utilizing CoT and PS exhibited mixed outcomes across models. It strengthened robustness in models like GPT-3.5 Turbo and GPT-4 Turbo, as shown by higher BM-Acc scores, while it increased sensitivity to noise in models like Llama-3 Instruct (70B) and Command R, resulting in reduced BM-Acc values.

Models	Method	BU-Acc	BM-Acc	BREU
Llama-3 Instruct (70B)	DP	10.99%	92.09%	51.54%
	CoT	12.57%	89.40%	50.99%
	PS	12.66%	88.21%	50.44%
Mixtral (8x22B)	DP	35.38%	36.57%	35.98%
	CoT	27.28%	34.93%	31.11%
	PS	44.04%	53.13%	48.59%
Command R	DP	12.10%	80.45%	46.28%
	CoT	11.36%	81.19%	46.28%
	PS	19.37%	69.85%	44.61%
Command R+	DP	13.69%	75.67%	44.68%
	CoT	14.71%	77.76%	46.24%
	PS	13.41%	65.07%	39.24%
Claude-3 Haiku	DP	9.40%	88.66%	49.03%
	CoT	13.50%	83.73%	48.62%
	PS	13.22%	82.99%	48.11%
Claude-3 Sonnet	DP	19.65%	82.69%	51.17%
	CoT	21.51%	81.19%	51.35%
	PS	16.76%	83.73%	50.25%
GPT-3.5 Turbo	DP	14.53%	55.22%	34.88%
	CoT	20.48%	65.22%	42.85%
	PS	17.78%	67.91%	42.85%
GPT-4 Turbo	DP	16.76%	86.72%	51.74%
	CoT	13.59%	87.76%	50.68%
	PS	12.76%	88.66%	50.71%

Table A1: The effectiveness of various prompting techniques varies across LLMs and subset of Belief-R, enhancing performance in some while degrading it in others.

Prompt to generate p and q

Make if-then statement from only the given sentences and no additional premises.
Fill the ___ if any, to make sentence that makes sense.

Given the link: Motivated by the "Cause", the "Event" happened and caused the "Effect".
Make it very short.

Event: PersonX uses PersonX's ___ to obtain
Cause (from PersonX): to have an advantage
Effect (to PersonX): pleased

Prompt to make p and q more specific

Make all of the entities both in the "if" section and in the "then" section very specific and simple. Keep the PersonX and PersonY intact. Make it very short.

If PersonX uses PersonX's resources to obtain an advantage, then PersonX is pleased.

Prompt to generate the additional condition r

For this conclusion "PersonX learns something new." to surely hold. Make the condition strongly relates within the context of "PersonX reads a book" but not about it.

Write a short if-then statement with: "then PersonX learns something new.". Do not mention "PersonX reads a book". Keep the PersonX (he) and PersonY (she) intact. Make the entities in the "if" section very specific. Output only the if-then sentence. Use easy to understand words but ensure that it is make sense. Make it very short.

Prompt to generate the alternative condition r

Make a condition that is not related at all to "PersonX reads a book" for this conclusion "PersonX learns something new." to happen.

Write a short if-then statement with: "then PersonX learns something new.". Do not mention "PersonX reads a book". Keep the PersonX (he) and PersonY (she) intact. Make the entities in the "if" section very specific. Output only the if-then sentence. Use easy to understand words but ensure that it is make sense. Make it very short. Use maximum 15 words.

Figure A1: Samples of prompts utilized in each of the Belief-R generation pipeline. Here, we take the `Event: PersonX uses PersonX’s ___ to obtain`, `Cause (from PersonX): to have an advantage`, `Effect (to PersonX): pleased` from ATOMIC, to generate p, q, and r for us to form queries at step t and t+1 in Belief-R and later go through the manual annotaion process.

Overview

In this exercise, you're presented with three statements. After reviewing these statements, you'll answer a question about the logical outcome based on those statements. Choose the outcome that logically follows.

Rules & Tips

Make sure to take all of the statements into account before making your decision.

Example 1

Statements:

Explanation:

Answer is A.

Example 2

Statements:

Explanation:

Answer is B.

Example 3

Statements:

Explanation:

Answer is C.

(a) Annotation Guidelines

If John buys a better fishing rod to catch more fish than Jessica, then John feels smug.
John buys a better fishing rod to catch more fish than Jessica.
If John wins a trophy in the chess tournament then John feels smug.

What necessarily had to follow assuming that the above premises were true

. **(required)**

- ○ John feels smug.
- ○ John does not feel smug.
- ○ John may or may not feel smug.

If John whispers to distract opponents, then Jessica frowns.
John whispers to distract opponents.
If John spills coffee on the table then Jessica frowns.

What necessarily had to follow assuming that the above premises were true

. **(required)**

- ○ Jessica frowns.
- ○ Jessica does not frown.
- ○ Jessica may or may not frown.

(b) Example of Annotation Questions

Figure A2: Details on ground truth annotation

Overview

In this exercise, you are presented with two statements. After reviewing these statements, you will answer four questions about them. Verify the validity of these statements based on the questions.

Tips

Read the question carefully

Example 1:

Are both sentences using if-then?

Explanation:

If the condition in "if" part is fulfilled, does it entail the premise in "then" part?

Explanation:

(a) Annotation Guidelines

If John buys a better fishing rod to catch more fish than Jessica, then John feels smug.
If John consistently outperforms Jessica in every fishing competition, then John feels smug.

Are both sentences using if-then? (required)

- Yes
- No

If John buys a better fishing rod to catch more fish than Jessica, then John feels smug.

If the condition in "if" part is fulfilled, does it entail the premise in "then" part? (required)

- Yes
- No

If John consistently outperforms Jessica in every fishing competition, then John feels smug.

If the condition in "if" part is fulfilled, does it entail the premise in "then" part? (required)

- Yes
- No

(b) Example of Annotation Questions

Figure A3: Details on quality check annotation

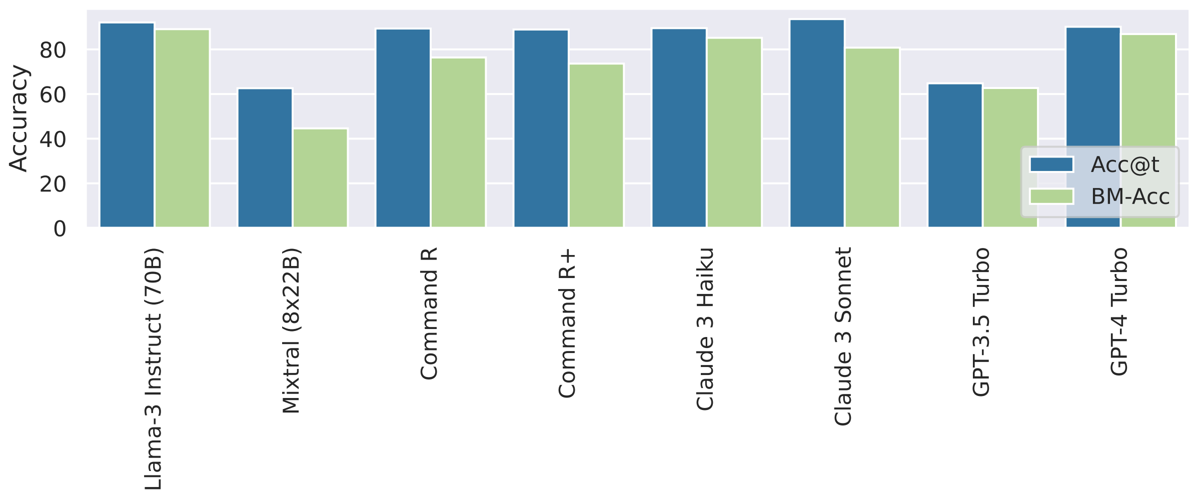


Figure A4: LLMs show decreased inference performance when exposed to noise from the new information in alternative condition.