Representations of Fact, Fiction and Forecast in Large Language Models: Epistemics and Attitudes

Meng Li, Michael Vrazitulis, David Schlangen
University of Potsdam
{meng.li, michael.vrazitulis, david.schlangen}@uni-potsdam.de

Abstract

Rational speakers are supposed to know what they know and what they do not know, and to generate expressions matching the strength of evidence. In contrast, it is still a challenge for current large language models to generate corresponding utterances based on the assessment of facts and confidence in an uncertain real-world environment. While it has recently become popular to estimate and calibrate confidence of LLMs with verbalized uncertainty, what is lacking is a careful examination of the linguistic knowledge of uncertainty encoded in the latent space of LLMs. In this paper, we draw on typological frameworks of epistemic expressions to evaluate LLMs' knowledge of epistemic modality, using controlled stories. Our experiments show that the performance of LLMs in generating epistemic expressions is limited and not robust, and hence the expressions of uncertainty generated by LLMs are not always reliable. To build uncertainty-aware LLMs, it is necessary to enrich the semantic knowledge of epistemic modality in LLMs.

1 Introduction

As LLMs are increasingly deployed in real-world situations, it is important for LLMs to behave in a rational way. They should be able to distinguish fact and imagination, and generate responses based on the credibility of the available evidence and the degree of their *confidence*. For example, imagine that Tom and Alice are looking for a missing book. If Tom says, "The book *may* be under the sofa", he means that it is possible that the book is under the sofa according to the available evidence. Namely, the conclusion is compatible with existing evidence. If Tom says, "The book *has to* be under the sofa", he is indicating that he believes to have strong evidence for the conclusion that the book is underneath the sofa, and he is ruling out all other potential locations with Alice. Given the available evidence, these epistemic modal expressions encode knowledge of necessity and possibility.

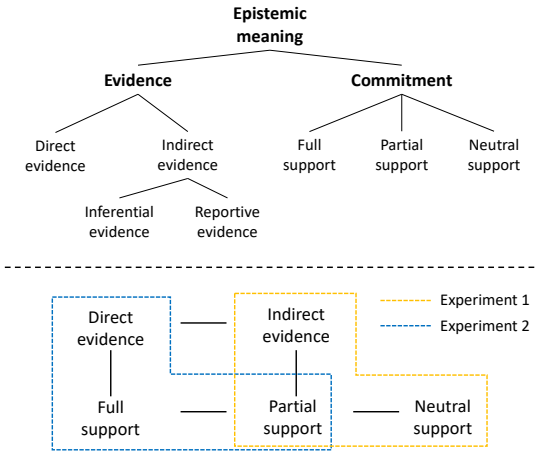

Figure 1: Dimensions of epistemic meanings (above) and the semantic map of epistemic modality (below) (Boye, 2012). The colored dashed lines indicate the mapping between concepts and our two experiments.

To build uncertainty-aware LLMs, many recent works focus on confidence estimation and calibration of LLMs. One common approach is about prompting language models to generate different levels of uncertainty in verbalized words or expressions (Mielke et al., 2022; Lin et al., 2022; Xiong et al., 2024). Given that current LLMs can generate very fluent responses, these papers assume, consciously or unconsciously, that these LLMs already master the linguistic knowledge of generating expressions of uncertainty, and the only problem left is alignment, namely how to elicit LLMs to generate these expressions at the right time. This assumption, however, was never carefully examined. Therefore, in this paper, we will investigate the form and meaning of epistemic expressions from typological perspectives, and evaluate the linguistic knowledge of epistemic modality present in an exemplary set of LLMs.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27734–27757

Let us first consider the meaning of epistemic expressions. The expressions of uncertainty in natural languages are semantically related to epistemic modality, because epistemic modality indicates the degree of certainty that a speaker has towards propositions (Portner, 2009; Szarvas et al., 2012; Giannakidou and Mari, 2021). As shown in Figure 1, epistemic meanings can be divided into *evidence* and *commitment* (Boye, 2012). Evidence is about the source of information or justification (Bybee, 1985; Palmer, 1986; Aikhenvald, 2004) and can be further divided into *direct evidence* and *indirect evidence*. Commitment is an epistemic modal meaning that expresses a speaker's degree of confidence or certainty (Coates, 1983; Bybee et al., 1994; Van der Auwera and Plungian, 1998). It is conceived as a continuous, quantitative scale, which can be divided into three levels: *full support*, *partial support*, and *neutral support* (see Appendix A for details). [1]

We now turn to the form of epistemic expressions. Epistemic meanings are expressed by various lexical, morphological, and syntactic structures across languages. In English, there are three major types of forms to express uncertainty: (1) modal adjectives and adverbs; (2) attitude verbs (or mental verbs); (3) modal auxiliaries and semi-auxiliaries. Modal adjectives and adverbs express probability through lexical meaning, such as *certain/certainly* and *probable/probably*. Attitude verbs, such as *believe*, *know* and *doubt*, describe internal mental states towards propositions. Modal auxiliaries and semi-auxiliaries, like *must/may/have to*, express modal meanings of necessity and possibility.

In previous work on confidence estimation and calibration, LLMs have been extensively evaluated on knowledge-intensive datasets such as TriviaQA (Joshi et al., 2017), StrategyQA (Geva et al., 2021), and GSM8K (Cobbe et al., 2021). However, these datasets present several limitations when it comes to evaluating LLMs' knowledge of epistemic modality. (1) Although these datasets provide the reference of factual correctness, the construct of knowledge-intensive tasks is to measure external world knowledge, not linguistic knowledge. They support the analysis of correspondence between the models' verbalized confidence and the likelihood that the answer is correct, but overlook the fact that human speakers also express their confidence in possible worlds, such as crime novels. (2) These datasets lack an agentic perspective, and they are coarse-grained. In other words, there are no controls on types of evidence and degrees of commitment. In addition, linguistic targets are also not structured systematically for comparison. (3) For large-scale real-world datasets, there is a risk of data contamination in the competitions for higher rankings in leaderboards. To address such concerns, we design controlled stories to test whether LLMs can predict the correct epistemic modal expressions (see Figure 2 and Figure 4). By simplifying the complexity of reasoning and controlling different factors, we can see how prompt formats and modal semantics affect the generation of epistemic expressions. [2]

This paper offers the following key contributions: (1) As far as we know, this is the first paper to evaluate the linguistic knowledge of epistemic modality in LLMs, namely the underlying knowledge that allows LLMs to understand and generate epistemic expressions. We demonstrate the limited performance of LLMs in selecting appropriate epistemic expressions. (2) Through controlled experiments, we show how the number of parameters, prompt formats and modal semantics affect the accuracy of LLMs. For modal auxiliaries, LLMs have better performance for modal necessity than modal possibility. For attitude verbs, LLMs are better at reporting facts than beliefs under different degrees of epistemic certainty. (3) By relating our question to a typological framework which categorizes epistemic expressions, we disentangle *linguistic* uncertainty from *aleatoric* and *epistemic* uncertainty (Kendall and Gal, 2017). We also offer insights on how to improve LLMs in the light of children's semantic development.

2 Related Work

2.1 Teaching LLMs to Express Their Uncertainty Truthfully

The problem of hallucinations and the need for alignment motivate confidence estimation and calibration in LLMs. There are different approaches for confidence estimation (Geng et al., 2024): logit-based methods (Kuhn et al., 2023; Huang et al., 2023; Vazhentsev et al., 2023; Duan et al., 2024), internal state-based methods (Ren et al.,

[1] To improve readability, we adapt terms in Boye (2012) to more standard expressions in current practice across theories and frameworks. We use *evidence* to replace *epistemic justification* and use *commitment* instead of *epistemic support*.

[2] Accessible data and code: https://github.com/limengnlp/llm-fff

2022; Kadavath et al., 2022; Burns et al., 2023; Azaria and Mitchell, 2023; Li et al., 2024), verbalized methods (Mielke et al., 2022; Xiong et al., 2024), consistency-based estimation (Manakul et al., 2023; Lin et al., 2024), and surrogate methods (Shrivastava et al., 2023). Verbalized methods involve prompting language models to output different levels of uncertainty in words or numbers, and prove an effective approach to calibrating language models with verbalized metacognition. Mielke et al. (2022) train an external calibrator to guide language models to generate with appropriate levels of uncertainty. Lin et al. (2022) fine-tune language models with a human annotated dataset to generate verbalized words and numbers at the same time.

In addition, there is also research on the behavior of LLMs on the expressions of uncertainty. It was found empirically that LLMs are sensitive to the expressions of uncertainty injected in prompts, and these expressions can improve or impair their performance (Zhou et al., 2023). Zhou et al. (2024) report that LLMs are reluctant to express uncertainty when they generate wrong answers. Yona et al. (2024) define a formal metric on faithful response uncertainty, and provide evidence that instruction-tuned LLMs perform poorly at conveying their intrinsic uncertainty.

Of the several types of devices expressing uncertainty, adjectives and adverbs expressing probability received more attention, because it is easier to build the mapping between uncertainty expressions and explicit numerical responses from humans at the population level (Wallsten et al., 1986; Willems et al., 2019; Fagen-Ulmschneider, 2019). Sileo and Moens (2023) show that neural language models struggle to understand words expressing probability, and fine-tuning with the supervision of human perception can lead to improvements. Belém et al. (2024) compare LLMs and humans in mapping uncertainty expressions to self-reported numerical probabilities.

2.2 Theory of Mind and Language Models

Theory of mind (ToM), the ability to infer other people's intents and beliefs, is assumed to play a key role in how children learn the meaning of words (Bloom, 2002) and resolve reference ambiguities in conversations (Clark and Marshall, 1981). Thus, it is a crucial component of intelligent systems that interact with humans. Research shows that the acquisition of epistemic modality is related to the development of ToM abilities in children, because understanding the meaning of epistemic modality requires children's ability to handle representations of mental states (Gopnik and Astington, 1988; Papafragou, 2002).

Grant et al. (2017) and Nematzadeh et al. (2018) adapt ToM tests in developmental psychology to evaluate the ability of language models in question answering (QA) tasks. Existing ToM tests are usually generated with templated stories where there is information asymmetry in a limited set: ToM-bAbI (Nematzadeh et al., 2018), ToMi (Le et al., 2019), Hi-ToM (Wu et al., 2023), OpenToM (Xu et al., 2024). In recent years, there has been a surge of research on ToM behavior in LLMs (Sap et al., 2022; Sclar et al., 2023; Wilf et al., 2023; Ullman, 2023; Shapira et al., 2024; Kosinski, 2024; Strachan et al., 2024). Our work leverages the templates of ToM stories and shifts the focus from inferring other agents' mental states to articulating one's own thinking truthfully.

2.3 Modal Semantics and LLMs

The semantics of conditionals are closely linked to modality. Holliday et al. (2024) test LLM reasoning in a set of inference patterns involving conditionals and epistemic modals, and identify the logical fallacies of LLMs. Our work aims to assess the linguistic knowledge of LLMs through their behavior in simple and controlled stories, without assuming a specific theory of logical or probabilistic reasoning. In other words, logical reasoning is not the focus here and its complexity is intentionally limited.

3 Method

To understand what LLMs know, we assume that their knowledge cannot be directly measured, but can be inferred from observable behavior (Jiang et al., 2020; Hu and Levy, 2023).

3.1 Grounding Epistemic Modality through Targeted Stories

The knowledge of morphology and syntax encoded in language models can be observed directly from the generated text and their analysis is more explicit. However, meaning cannot be observed directly. To investigate the semantic knowledge of language models, it needs to be inferred from the behavior that such knowledge is assumed to underlie, especially responses to specific stimuli in experimentally controlled settings.

Compared with concrete words like visible objects and actions, the meaning of modal words is more abstract and highly context-dependent. To trigger participants or language models to generate modal expressions in a coherent and plausible way, we need contextualized stories. In the field of child language, the *hidden object task* is designed to test children's understanding of epistemic modality (Hirst and Weil, 1982; Noveck et al., 1996; Ozturk and Papafragou, 2015). In Ozturk and Papafragou (2015), children participants were presented with brief animated stories, where an animal would hide in a box. Given the information and prompt, participants were asked to reply with "Yes/No" to questions about the location of the animal. *Prompt format*, *modal semantics*, and *types of stories* are controlled. In linguistic fieldwork, targeted storyboards also prove effective in studying modality (Burton and Matthewson, 2015). The key point is that the story is designed to include at least one targeted context, which can be utilized to test hypotheses about specific linguistic forms within that context. These practices provide transferrable research paradigms to evaluate the knowledge of epistemic modality in LLMs.

There are different ways to generate stimuli for LLMs: (1) manually written datasets like GPQA (Rein et al., 2023); (2) procedurally generated datasets such as BLiMP (Warstadt et al., 2020); (3) language model generated dataset like BigToM (Gandhi et al., 2024). In the present study, we generate stimuli by combining manual writing with template-based generation.

3.2 Models

Considering the reproducibility of behavioral experiments, we evaluate eight open weight instruction-tuned models: `Llama3-8B/70B-Instruct`, `Llama3.1-8B/70B-Instruct` (AI@Meta, 2024), `Qwen2-7B/72B-Instruct` (Yang et al., 2024), `Qwen2.5-7B/72B-Instruct` (Team, 2024). They are accessed through the Huggingface Transformers library. We use greedy decoding for the result. These experiments were implemented on servers equipped with Nvidia A100 GPUs (80GB RAM).

4 Experiment 1: Modal Auxiliaries

The first experiment systematically investigated LLMs' semantic knowledge of epistemic modal verbs, *may*/*might* vs *must*/*have to*, with a variety of modal scenarios. Eight instruction-tuned models were tested with simple stories that gave different cues about a set of elements and the selection of possibility/necessity. The number of parameters for these models falls into two categories: small (7-8B) and medium (70-72B).

4.1 Experimental Design

There are 150 stories generated by five templates (see brief example in Figure 2 and more details in Appendix E). In each story, there is a context to create a set of elements (objects, persons, places, etc.). In the condition with a necessity modal (*N-modal condition*), there is clear and sufficient information to rule out all other candidates and identify the only one element left. In the condition with a possibility modal (*P-modal condition*), the given information is not sufficient to make a precise statement, and there is uncertainty. For example, in a story generated with the *hidden object* template, a toy is hidden in one of the three boxes: a red box, a green box, and a blue box. In the N-modal condition, the text provides enough information to rule out two boxes. If it is not in the red and green box, then it *has to*/*must* be in the blue box. In the P-modal condition, if the toy is not in the red box, it is still not certain that the toy is in the blue box or the green box. The toy *may*/*might* be in the blue box, which is more appropriate to describe this situation (see Figure 2).

We design three question and answer formats: direct slot, indirect slot, and indirect sentence, to ensure that responses to the modal statements reflected the semantic intuitions of LLMs. The direct slot format ask LLMs to select words that could be filled in a slot and to respond with these words directly. The indirect slot format asks LLMs to select words that could be filled in a slot, but to respond with the associated number or index. The indirect sentence format requires LLMs to select sentence-level statements and to respond with the associated number. The difference between slot and sentence formats is that LLMs need to identify the position of the slot, fill it with different words and compare. However, the sentence statements are much more natural to evaluate directly in the sentence format. The difference between direct and indirect formats is whether LLMs need metalinguistic indices (1/2) to refer epistemic modals when they respond. Theoretically, these indices increase the complexity of processing. If the knowledge of epistemic modal semantics in LLMs is robust, performance should be similar across different prompt formats. Other-

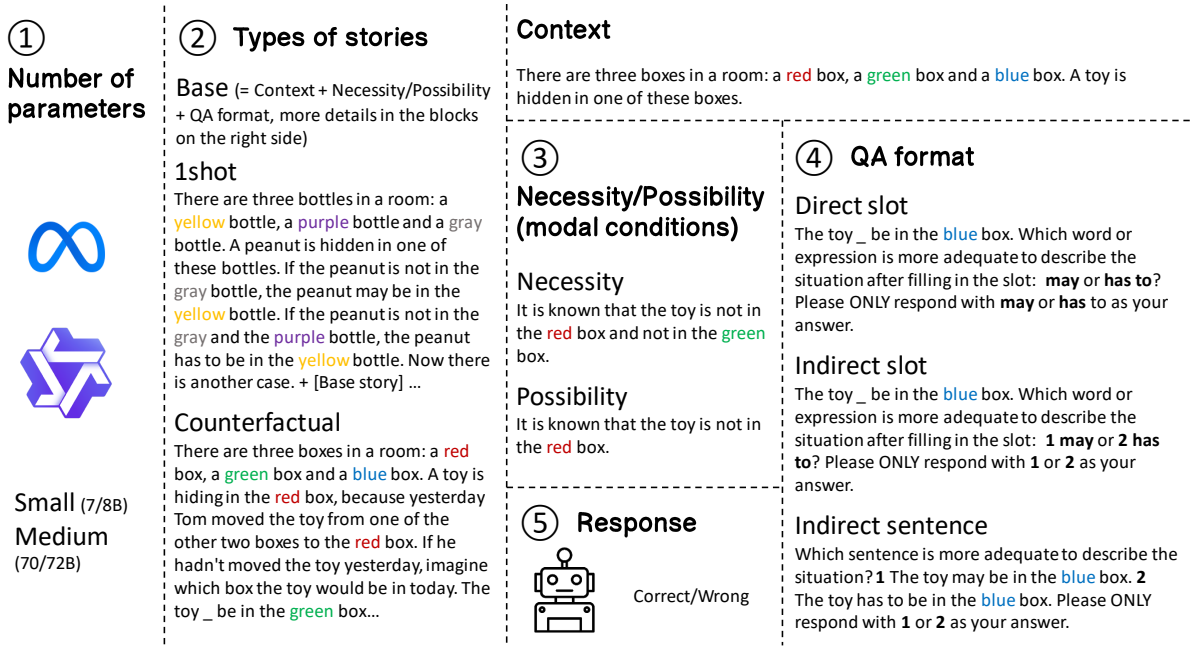

Figure 2: Experimental design for assessing modal auxiliaries and semi-auxiliaries.

wise, performance might diverge across the three different formats.

There are three different types of story: base, 1-shot and counterfactual stories. The base stories have been explained above. In-context learning (ICL) has proven to be an effective post-training technique to enable LLMs to solve new tasks with only a few demonstrations. One-shot stories introduce additional similar narratives to the base version, illustrating the selection of modal verbs in both N-modal and P-modal conditions. To prevent LLMs from simply copying the answer from the 1-shot example, we use lexical variations, such as differencing colors or person names. Modality concerns possible worlds and alternative ways that things could be. There are linguistic interactions between counterfactual conditions and epistemic modality. In English, counterfactual conditions are expressed through past tense and subjunctive mood. Therefore, we include counterfactual stories to create parallel possible scenarios by changing a verified true condition. These two variant types are designed to test whether the knowledge of epistemic modality in LLMs is sensitive to other information structures. Intuitively, for humans, the 1-shot stories should be easier than the base stories since there are additional supervised examples, while the counterfactual stories should be harder than the base stories, because they require additional counterfactual reasoning. Fifty stories were created for each type, resulting in a total of 150 stories.

We measure the performance with *accuracy* and *paired accuracy*. For paired accuracy, we count the answer as correct if the questions in a pair of N-modal and P-modal conditions are both answered correctly.

4.2 Statistical Analysis and Results

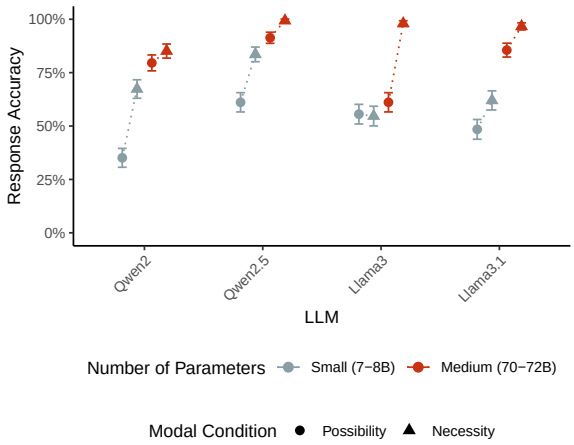

Figure 3: Response accuracy in Experiment 1 by LLM, number of parameters, and modal condition. Error bars represent 95% confidence intervals.

Mean accuracies and paired accuracies for all assessed LLMs are reported in Table 1. Judging

LLM	Acc	Paired Acc
Qwen2-7B	51.2	2.4
Qwen2.5-7B	72.3	45.1
Llama3-8B	55.1	11.8
Llama3.1-8B	55.2	11.8
Qwen2-72B	82.3	64.7
Qwen2.5-72B	95.3	90.7
Llama3-70B	79.6	59.1
Llama3.1-70B	91.1	82.2

Table 1: Mean accuracy and mean paired accuracy in percent, for different LLMs in Experiment 1.

from these descriptive statistics alone, it seems that LLMs with a medium number of parameters (70-72B), with accuracies ranging from 79.6% to 95.3%, clearly outperform their counterparts with a smaller number of parameters (7-8B), whose accuracies just range between 55.1% and 72.3%.

Using logistic regression, we assess how number of parameters, story type, modal condition, and QA format possibly modulate response accuracy. To do so, we fit a separate logistic regression model to each of the four classes of LLMs (Qwen2, Qwen2.5, Llama3, Llama3.1). Details on how these statistical analyses were performed are reported in Appendix C.1.

Across all assessed LLMs, the number of parameters has a significant impact on response accuracy (Qwen2: $b = 1.63$, $SE = 0.12$, $p < .001$; Qwen2.5: $b = 3.15$, $SE = 0.36$, $p < .001$; Llama3: $b = 2.17$, $SE = 0.20$, $p < .001$; Llama3.1: $b = 2.96$, $SE = 0.22$, $p < .001$), indicating that a large number of parameters (70–72B) leads to increased accuracy. A consistent effect of modal condition is also found across LLMs (Qwen2: $b = 0.87$, $SE = 0.12$, $p < .001$; Qwen2.5: $b = 1.99$, $SE = 0.32$, $p < .001$; Llama3: $b = 1.79$, $SE = 0.19$, $p < .001$; Llama3.1: $b = 1.14$, $SE = 0.17$, $p < .001$). The positive sign of the effect indexes that accuracy was higher for Necessity trials than it was for Possibility trials. The interaction between number of parameters and modal condition is also significant across LLMs, yet the sign and magnitude of the interaction differs considerably between LLMs (Qwen2: $b = -0.93$, $SE = 0.23$, $p < .001$; Qwen2.5: $b = 1.54$, $SE = 0.63$, $p = .015$; Llama3: $b = 3.66$, $SE = 0.38$, $p < .001$; Llama3.1: $b = 1.16$, $SE = 0.33$, $p < .001$). Figure 3 illustrates the effects of number of parameters and modal condition.

The factors story type and QA format do also modulate response accuracy to some extent. However, the magnitudes of the effects of story type and QA format (as well as their interactions with number of parameters) are rather small, and none of their effect signatures are consistent across the assessed LLMs.

In sum, there is a salient improvement in accuracy when the expected correct response expresses a necessity, rather than a possibility. Further, as expected, performance is increased substantially when LLMs have a medium (70–72B) as opposed to just a small (7–8B) number of parameters. The higher accuracy for necessity modals suggests that models handle contexts with a unique, unambiguous conclusion more reliably than those requiring reasoning under uncertainty. For instance, when two out of three locations are ruled out, models correctly infer "The toy must be in the blue box." In contrast, when multiple outcomes remain plausible, models often fail to select "may", indicating a weaker grasp of possibility in uncertain contexts.

5 Experiment 2: Attitude Verbs

The second experiment focuses on the semantic knowledge of attitude verbs in LLMs: *know*, *believe*, and *doubt*. Given Theory-of-Mind (ToM) stories, LLMs were required to select different attitude verbs to report facts or beliefs with different degrees of certainty.

5.1 Experimental Design

Self-concept (awareness of one's own mental states and functions) and a theory of mind (awareness of others' mental states and functions) are components of metacognition. The metacognitive system involves a second-order form of *knowing* and can be viewed as an interface between the mind and reality (Demetriou et al., 2010). While ToM datasets are typically used to evaluate LLMs' ability to reason about others' beliefs in simple narratives, they can also be repurposed to assess whether LLMs can truthfully describe their own reasoning, given a shift in the focus of the task.

In a typical ToM test, the test subject (a child or a language model) observes a sequence of actions of two agents: `agent 0` moves an object into a container, and `agent 1` moves the object to another container when `agent 0` is not aware of this. The test subject is then asked questions about the actual state of the world and the agents' beliefs. In previous ToM datasets, the task is designed as a QA task, and questions were designed to ask for the location of the object in different settings. Here

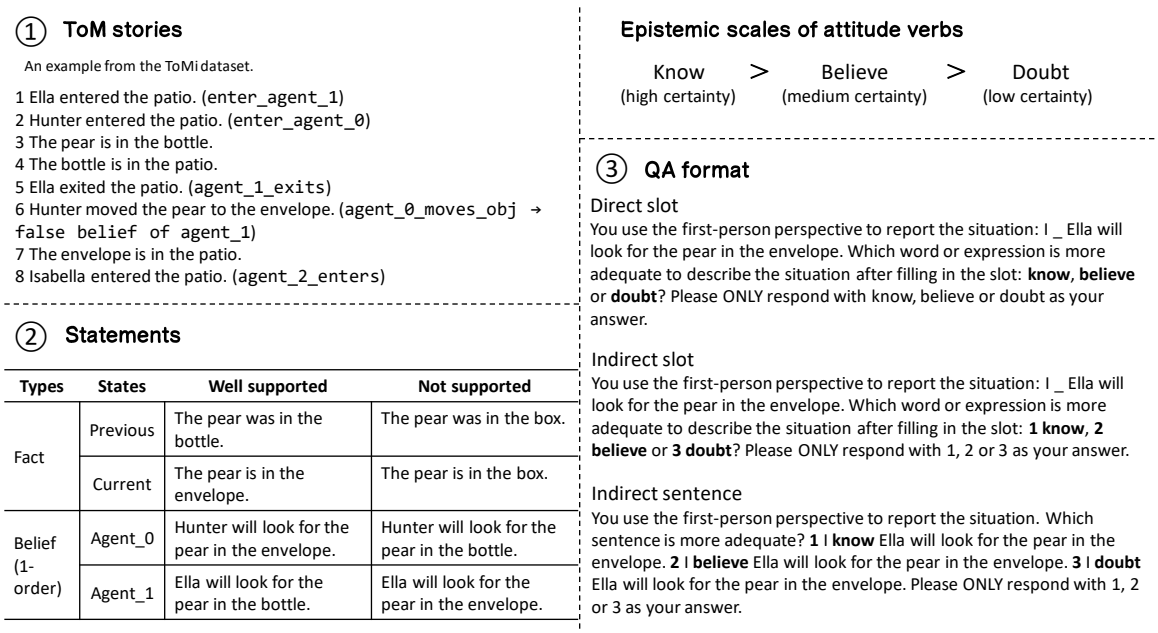

Figure 4: Stimuli in experiment 2 to assess attitude verbs.

we change the questions and test the knowledge of LLMs in selecting attitude verbs (*know*, *believe*, and *doubt*) to report facts or beliefs in different statements. To answer the question correctly, the model needs to choose the right verb based on the type of statement (fact/belief) and the strength of evidence. [3]

Thirty stories were selected randomly from the ToMi dataset (Le et al., 2019; Sclar et al., 2023). For each story, we constructed eight statements: two statements on previous facts; two statements on current facts, two statements on 1-order beliefs of `agent 0`, and two statements on 1-order beliefs of `agent 1`. All paired statements in these four subtypes are contrastive in low/not low certainty (see examples in Figure 4). In addition, there are also three question and answer formats: direct slot, indirect slot, and indirect sentence, which is similar to Experiment 1.

We measure the performance with *accuracy*, *paired accuracy*, and *joint accuracy*. For paired accuracy, we count the answer as correct if questions in a pair of low and not low certainty conditions are both answered correctly, since these prompts are constructed in a controlled way. For joint accuracy, we count the answer as correct if questions about eight statements of the same ToM story are all answered correctly.

5.2 Statistical Analysis and Results

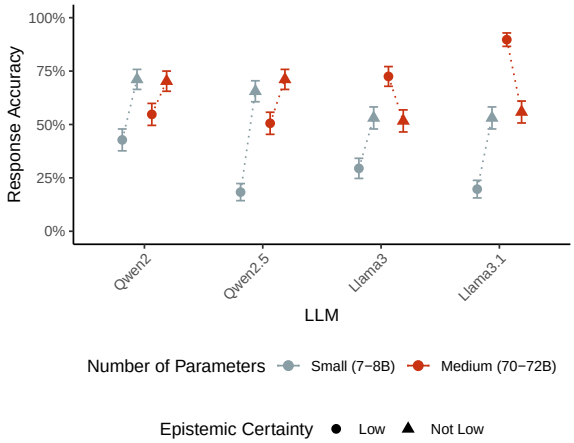

Figure 5: Response accuracy in Experiment 2 by LLM, number of parameters, and modal condition. Error bars represent 95% confidence intervals.

Table 2 reports mean accuracies and paired accuracies for the assessed LLMs in Experiment 2. Again, LLMs with a medium number of parameters (70–72B) perform noticeably better than their counterparts with a small number of parameters (7–8B), by about 20 percentage points in accuracy (60.8–72.8% vs. 36.4–56.9%).

We use logistic regression to statistically model the effects of different factors on response accuracy.

[3] Negation will change the values of an epistemic scale (Horn, 1989) and introduce more complex inference. For example, I am *certain that not P* vs. I am *not certain that P*. To keep the setting simple and focus on attitude verbs, we exclude negation here.

LLM	Acc	Paired Acc	Joint Acc
Qwen2-7B	56.9	38.6	0
Qwen2.5-7B	41.9	16.1	0
Llama3-8B	41.3	28.9	0
Llama3.1-8B	36.4	12.2	0
Qwen2-72B	62.5	42.2	4.4
Qwen2.5-72B	60.8	46.4	5.6
Llama3-70B	62.1	40.3	0
Llama3.1-70B	72.8	54.2	0

Table 2: Mean accuracy, paired accuracy, and joint accuracy in percent, for different LLMs in Experiment 2.

Specifically, we are interested in whether an LLM's number of parameters, the epistemic certainty of the attitude verb (low for *doubt*, not low for *believe* and *know*), the type of statement (belief- or fact-based), and finally the QA format (direct/indirect slot, indirect sentence) each modulate the response accuracies on the ToM task in some meaningful way. Appendix C.2 provides more details on how these statistical analyses were performed.

For Qwen2, the number of parameters does not show an effect on response accuracy in the ToM task ($b = 0.08$, $SE = 0.15$, $p = .604$). Yet, for each of the remaining three LLM classes, an increase in the number of parameters (from small to medium) does in fact lead to a significant increase in response accuracy (Qwen2.5: $b = 1.65$, $SE = 0.18$, $p < .001$; Llama3: $b = 1.88$, $SE = 0.19$, $p < .001$; Llama3.1: $b = 2.70$, $SE = 0.19$, $p < .001$).

Figure 5 shows that, for all LLMs with a small number of parameters, target verbs with a relatively high epistemic certainty (*believe* or *know*) are associated with higher response accuracies, compared to a target verb with low epistemic certainty (*doubt*). This trend also holds for the Qwen2/Qwen2.5 models with a medium number of parameters, but is interestingly reversed for the Llama3/Llama3.1 models with a medium number of parameters. These effect patterns are also reflected in the regression estimates for the main effect of epistemic certainty (Qwen2: $b = 1.64$, $SE = 0.16$, $p < .001$; Qwen2.5: $b = 3.08$, $SE = 0.22$, $p < .001$; Llama3: $b = 1.36$, $SE = 0.32$, $p < .001$; Llama3.1: $b = -0.55$, $SE = 0.17$, $p = .001$) and the interaction between number of parameters and epistemic certainty (Qwen2: $b = -1.61$, $SE = 0.33$, $p < .001$; Qwen2.5: $b = -2.99$, $SE = 0.44$, $p < .001$; Llama3: $b = -5.30$, $SE = 0.64$, $p < .001$; Llama3.1: $b = -4.78$, $SE = 0.34$, $p < .001$).

The main effect of statement type (belief- vs. fact-based) is large and highly significant across LLM classes (Qwen2: $b = 3.02$, $SE = 0.17$, $p < .001$; Qwen2.5: $b = 3.79$, $SE = 0.22$, $p < .001$; Llama3: $b = 4.79$, $SE = 0.34$, $p < .001$; Llama3.1: $b = 2.74$, $SE = 0.19$, $p < .001$), indicating substantially higher response accuracies for fact-based statements. Figure 8 in Appendix D.2 confirms this visually.

Other included predictors fail to show consistent and salient effects across the assessed LLMs.

To summarize, the logistic regression analyses allow for three key observations about the assessed LLMs' behavior on the ToM task: (1) LLMs with a medium (rather than small) number of parameters tend to respond more accurately. (2) Relatively high epistemic certainty (*believe* or *know*) leads to higher accuracy, but note that the Llama3/Llama3.1 models with a medium number of parameters display the opposite effect. (3) LLMs systematically show higher accuracies on fact-based statements than on belief-based statements.

6 Discussion

6.1 Comparing the Behavior of Epistemic Modality in Human and Machines

Epistemic modal verbs Ozturk and Papafragou (2015) tested children with the *hidden object task* and evaluated their knowledge of modal expressions with *may* and *have to*. It is reported that children between the ages of 4 and 5 years have a basic understanding of epistemic semantic modals, but their knowledge of epistemic modals is sensitive to the syntactic–semantic context (statement vs. question prompts). Specifically, children were better at evaluating statements than at answering questions. Similarly, the performance of LLMs is also affected by prompt formats (see Figure 7). As a control group, adults achieve 97-100% accuracy in different stories. By contrast, there is still a certain gap between the performance of LLMs and adults.

Attitude verbs Different from words referring to concrete objects or visible actions, attitude verbs describe internal states of mind and leave few cues in the physical world. These characteristics pose challenges for native language learners. It is not until well into preschool that children begin to show adult-like performance on tasks involving attitude verbs (Hacquard and Lidz, 2022). There are different hypotheses and theories about the learning of attitude verbs (Landau and Gleitman, 1985; Gleitman, 1990; Diessel and Tomasello, 2001; Montgomery, 2002; Papafragou et al., 2007; Israel, 2008;

Becker and Estigarribia, 2013; Hacquard and Lidz, 2019). Such accounts, in turn, can also raise important questions for scientists in different research communities:

- For researchers of machine learning and computational linguistics, who are more interested in how LLMs learn and use attitude or mental verbs, the potential questions include: (1) Can a language model learn the meaning of attitude verbs without consciousness? (2) Are existing LLM training pipelines, such as self-supervised pretraining + supervised finetuning (SFT) + reinforcement learning from human feedback (RLHF), sufficient to learn the meaning of attitude verbs? (3) How do LLMs benefit from theories of children's semantic development (e.g., learning from interactions in pragmatically informative environments)?

- For cognitive scientists who specialize in developmental psycholinguistics and language acquisition, LLM's training mechanism may also provide a computational modeling perspective to validate or challenge existing learning theories of attitude verbs. For example, LLM pretraining employs self-supervised learning and does not explicitly use a priori syntactic categories. Does this imply that the cues of statistical distribution contribute to learning and that a priori syntactic categories are not necessary?

6.2 Future Work

Forms of Modality in Low-Resource Languages There are different means to express modal semantics in non-English languages: modal affixes, modal case, etc. (de Haan, 2006). These morphological variations in low-resource languages impose challenges for building truthful multilingual LLMs, and remain to be tested in the future.

Enriching Benchmarks with Multimodal Evidence and Complex Reasoning (1) We used text-based stories in both experiments, but future work could test epistemic reasoning in multimodal environments, especially as embodied intelligence becomes increasingly prominent. (2) In order to avoid interference from unnecessary world knowledge, we controlled the complexity of reasoning. However, LLMs still need to improve how they handle conflicting evidence (Kazemi et al., 2023; Wan et al., 2024).

7 Conclusion

In this paper, we evaluate the semantic knowledge of epistemic modality in open-weights LLMs through controlled stories, and show their limited performance in generating appropriate epistemic expressions. This implies that responses containing epistemic uncertainty from LLMs may be unreliable. Insufficient semantic knowledge of epistemic modality is a potential reason why LLMs are not good at truthfully expressing uncertainty. Therefore, to build rational LLMs, we should not only improve the methods of uncertainty estimation and calibration, but also enrich the semantic representation of epistemic modality.

Acknowledgments

We are grateful to Jiangtian Li and the anonymous reviewers for helpful feedback.

Limitations

We follow the behavioral approach to evaluate the semantic knowledge of epistemic modality in LLMs, and there are several limitations in this work. (1) Children's acquisition profiles and data from adult groups in existing literature on semantic development (Noveck et al., 1996; Ozturk and Papafragou, 2015; Hacquard and Lidz, 2022) provide indirect evidence that adults can achieve high performance on the tasks in this paper. However, we did not test human participants directly. (2) We did not leverage logit probabilities to design new metrics of intrinsic uncertainty, nor did we build mapping between model logits and self-reported human responses. (3) Our work focuses on the English language, and we did not study whether subword tokenization in LLMs can handle the morphological encoding of modality in low-resource languages.

References

Alexandra Y Aikhenvald. 2004. *Evidentiality*. Oxford University Press.

AI@Meta. 2024. Llama 3 model card.

Htrotugu Akaike. 1973. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Misha Becker and Bruno Estigarribia. 2013. Harder words: Learning abstract verbs with opaque syntax. *Language Learning and Development*, 9(3):211–244.

Catarina G Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8467–8502, Miami, Florida, USA. Association for Computational Linguistics.

Paul Bloom. 2002. *How children learn the meanings of words*. MIT press.

Kasper Boye. 2012. *Epistemic meaning: A crosslinguistic and functional-cognitive study*, volume 43. Walter de Gruyter.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Strang Burton and Lisa Matthewson. 2015. Targeted construction storyboards in semantic fieldwork. In M. Ryan Bochnak and Lisa Matthewson, editors, *Methodologies in Semantic Fieldwork*, chapter 5, pages 135–156. Oxford University Press.

Joan Bybee, Revere Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press.

Joan L Bybee. 1985. *Morphology: A study of the relation between meaning and form*. John Benjamins Publishing Company.

Herbert H Clark and Catherine R Marshall. 1981. Definite knowledge and mutual knowledge. In Aravind K Joshi, Bonnie L Webber, and Ivan A Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press.

Jennifer Coates. 1983. *The semantics of the modal auxiliaries*. Croom Helm.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ferdinand de Haan. 2006. Typological approaches to modality. In Wolfgang Klein and Stephen Levinson, editors, *The Expression of Modality*, pages 27–69. Mouton de Gruyter.

Andreas Demetriou, Antigoni Mouyi, and George Spanoudis. 2010. The development of mental processing. In Richard M. Lerner., editor, *The Handbook of Life-Span Development: Cognition, Biology, and Methods*, page 36–55. Wiley.

Holger Diessel and Michael Tomasello. 2001. The acquisition of finite complement clauses in english: A corpus-based analysis. *Cognitive Linguistics*, 12(2).

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Wade Fagen-Ulmschneider. 2019. Perception of probability words. Accessed: Feb-01-2025.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Anastasia Giannakidou and Alda Mari. 2021. *Truth and veridicality in grammar and thought: Mood, modality, and propositional attitudes*. University of Chicago Press.

Lila Gleitman. 1990. The structural sources of verb meanings. *Language acquisition*, 1(1):3–55.

Alison Gopnik and Janet W Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pages 26–37.

Erin Grant, Aida Nematzadeh, and Thomas L Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *39th Annual Meeting of the Cognitive Science Society: Computational Foundations of Cognition, CogSci 2017*, pages 427–432. The Cognitive Science Society.

Valentine Hacquard and Jeffrey Lidz. 2019. Children's attitude problems: Bootstrapping verb meaning from syntax and pragmatics. *Mind & Language*, 34(1):73–96.

Valentine Hacquard and Jeffrey Lidz. 2022. On the acquisition of attitude verbs. *Annual Review of Linguistics*, 8(1):193–212.

William Hirst and Joyce Weil. 1982. Acquisition of epistemic and deontic meaning of modals. *Journal of child language*, 9(3):659–666.

Wesley Holliday, Matthew Mandelkern, and Cedegao Zhang. 2024. Conditional and modal reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3800–3821.

Laurence R. Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Michael Israel. 2008. Mental spaces and mental verbs in early child english. In Andrea Tyler, Yiyoung Kim, and Mari Takada, editors, *Language in the context of use : discourse and cognitive approaches to language*, pages 199–232. Mouton de Gruyter.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Emily Jin, Zhuoyi Huang, Jan-Philipp Fränken, Weiyu Liu, Hannah Cha, Erik Brockbank, Sarah A Wu, Ruohan Zhang, Jiajun Wu, and Tobias Gerstenberg. 2024. MARPLE: A benchmark for long-horizon inference. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems*, 36:39052–39074.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Barbara Landau and Lila R Gleitman. 1985. *Language and experience: Evidence from the blind child*. Harvard University Press.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. 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 5872–5877.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Derek E Montgomery. 2002. Mental verbs and semantic development. *Journal of Cognition and Development*, 3(4):357–384.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400.

Ira A Noveck, Simon Ho, and Maria Sera. 1996. Children's understanding of epistemic modals. *Journal of child language*, 23(3):621–643.

Ozge Ozturk and Anna Papafragou. 2015. The acquisition of epistemic modality: From semantic meaning to pragmatic interpretation. *Language learning and development*, 11(3):191–214.

Frank Robert Palmer. 1986. *Mood and modality*. Cambridge University.

Anna Papafragou. 2002. Modality and theory of mind: Perspectives from language development and autism. In *Modality and its Interaction with the Verbal System*, pages 185–204. John Benjamins Publishing Company.

Anna Papafragou, Kimberly Cassidy, and Lila Gleitman. 2007. When we think about thinking: The acquisition of belief verbs. *Cognition*, 105(1):125–165.

Paul Portner. 2009. *Modality*. Oxford Surveys in Semantics and Pragmatics. Oxford University Press.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. Llamas know what gpts don't show: Surrogate models for confidence estimation. *arXiv preprint arXiv:2311.08877*.

Damien Sileo and Marie-francine Moens. 2023. Probing neural language models for understanding of words of estimative probability. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 469–476, Toronto, Canada. Association for Computational Linguistics.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Johan Van der Auwera and Vladimir A Plungian. 1998. Modality's semantic map. *Linguistic Typology*.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454.

Cesko C. Voeten. 2023. *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*. R package version 2.11.

Thomas S Wallsten, David V Budescu, Amnon Rapoport, Rami Zwick, and Barbara Forsyth. 1986. Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, 115(4):348.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

Sanne JW Willems, Casper J Albers, and Ionica Smeets. 2019. Variability in the interpretation of dutch probability phrases-a risk for miscommunication. *arXiv preprint arXiv:1901.09686*.

Sarah A Wu, Erik Brockbank, Hannah Cha, Jan-Philipp Fränken, Emily Jin, Zhuoyi Huang, Weiyu Liu, Ruohan Zhang, Jiajun Wu, and Tobias Gerstenberg. 2024. Whodunnit? inferring what happened from multimodal evidence. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori B Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524.

A More Background on Epistemic Meaning

Evidence is also often referred to by the terms *evidentiality* (Aikhenvald, 2004) or *evidentials* (Bybee et al., 1994). Direct evidence covers "firsthand/visual/auditory/participatory evidence". Within indirect evidence, *reportive* and *inferential evidence* can be distinguished from one another: reportive evidence is about "secondhand/hearsay/quotative evidence"; inferential evidence covers "inferential/assumptive evidence".

Full support has the strongest strength on the epistemic modal scale, and is equivalent to "knowledge/certainty (that not)/ epistemic impossibility". Neutral support covers "epistemic possibility/(complete) uncertainty". Partial support lies between full support and neutral support, and covers "probability/(relative weak) uncertainty/(un)likely/epistemic necessity".

A *semantic map* is a visual representation of cross-linguistic patterns or regularities in semantic structures, which maps how various languages categorize meaning in a specific domain. The categories of evidence and commitment constitutes a continuous region in the semantic map of epistemic expressions (see Figure 1). We use nodes in the map to show the relation of two experiments in this paper.

B Contrast Coding for Experiments

B.1 Experiment 1

B.2 Experiment 2

C Additional Information on Logistic Regression Analyses

C.1 Experiment 1

We consider all main effects of the factors number of parameters, story type, modal condition, and QA format, in addition to each two-way interaction between number of parameters and any of the

Number of Parameters		Medium
Small		−0.5
Medium		0.5
Modal Condition		Necessity
Possibility		−0.5
Necessity		0.5
Story Type	1-Shot	Counterfactual
Base	−0.5	−0.5
1-Shot	0.5	0
Counterfactual	0	0.5
QA Format	Ind. Slot	Ind. Sent
Direct Slot	−0.5	−0.5
Indirect Slot	0.5	0
Indirect Sentence	0	0.5

Table 3: Contrast coding for statistical analyses of Experiment 1. The factors are number of parameters, story type, modal condition, and QA format.

Number of Parameters			Medium
Small			−0.5
Medium			0.5
Epistemic Certainty			Not Low
Low			−0.5
Not Low			0.5
Statement Type	Fact	Agent 1	Current
Belief (Agent 0)	−0.5	−0.5	0
Belief (Agent 1)	−0.5	0.5	0
Fact (Previous)	0.5	0	−0.5
Fact (Current)	0.5	0	0.5
QA Format		Ind. Slot	Ind. Sent
Direct Slot		−0.5	−0.5
Indirect Slot		0.5	0
Indirect Sentence		0	0.5

Table 4: Contrast coding for statistical analyses of Experiment 2.

remaining factors. We focus specifically on the interactions between number of parameters and any other factor because it is plausible to assume that at scale (i.e., with more parameters) LLMs display clear qualitative differences in their response patterns. This, in turn, may modulate the effect patterns associated with any of the remaining factors.

In order to probe the relevance of potential random effects, we applied forward model selection based on the Akaike Information Criterion (AIC; Akaike, 1973), as implemented in the *buildmer* package (Voeten, 2023) for the R programming language (R Core Team, 2023). We assessed random intercepts and random slopes for number of parameters, varying by LLM, by template, and by item nested within template. However, the model selection indicated that none of the assessed random effect terms substantially improved goodness of model fit as measured by the AIC. That is why eventually we fitted simple logistic regression models, without any random effects, whose results we report below.

In some cases, AIC-based model selection led to dropping certain fixed-effect interaction terms, as their inclusion did not improve the overall goodness of fit enough to offset the AIC penalty for additional parameters.

All factors in the logistic regression models were contrast-coded using sum-to-zero effect coding. The precise coding scheme applied for each factor is reported in Table 3, in B.1. The regression results are summarized in Table 7 (Qwen2), Table 8 (Qwen2.5), Table 9 (Llama3), and Table 10 (Llama3.1), respectively. Receiver operating characteristic (ROC) curves for the optimally fitting logistic regression models for each LLM class are shown in Figure 10, Appendix G.2

C.2 Experiment 2

In the logistic regression analyses for Experiment 2, the factors number of parameters, epistemic certainty, statement type, and QA format are treated as main effects. They are coded using sum-to-zero effect coding (see details on contrast coding in Table 4, Appendix B.2). The two-way interactions between number of parameters and any of the remaining factors are also assessed.

A separate logistic regression model is fitted for each examined class of LLMs (Qwen2, Qwen2.5, Llama3, Llama3.1). Using the *buildmer* package (Voeten, 2023) for R, we conduct an AIC-based model selection process. Its main purpose is to check whether it is necessary to include any random effects which would be theoretically justified by the design. We consider by-item random intercepts as well as by-item random slopes for the factor number of parameters. As the model selection analyses reveal, however, in all cases the retained optimal model does not include any random effects, i.e., is just a simple logistic regression model.

In Appendix F.2, all results of the logistic regression analyses for Experiment 2 are reported; see Tables 11 (Qwen2), 12 (Qwen2.5), 13 (Llama3), and 14 (Llama3.1), respectively. ROC curves for the optimally fitting logistic regression models for each LLM class are shown in Figure 11, Appendix G.2.

D Further Plots of Response Accuracy by Condition

D.1 Experiment 1

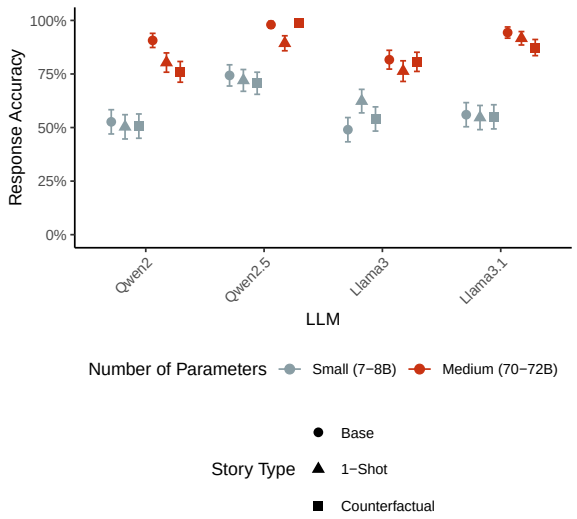

Figure 6: Response accuracy in Experiment 1 by LLM, number of parameters, and story type. Error bars represent 95% confidence intervals.

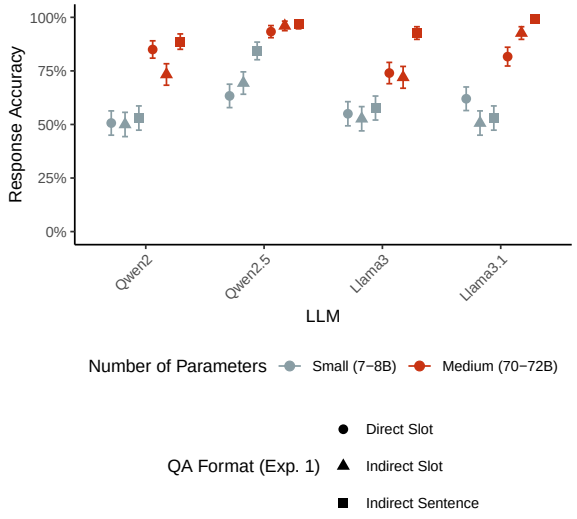

Figure 7: Response accuracy in Experiment 1 by LLM, number of parameters, and QA format. Error bars represent 95% confidence intervals.

D.2 Experiment 2

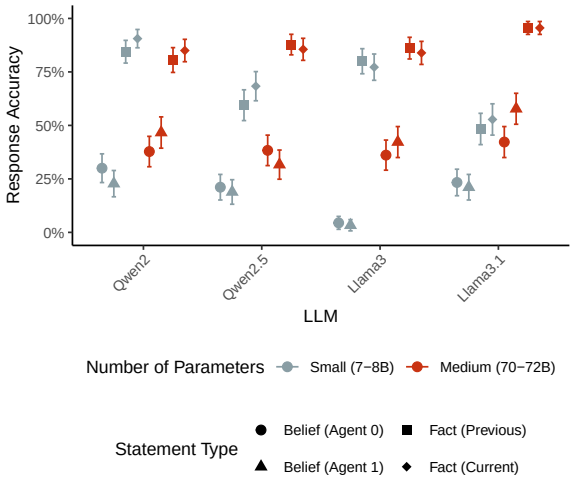

Figure 8: Response accuracy in Experiment 2 by LLM, number of parameters, and statement type. Error bars represent 95% confidence intervals.

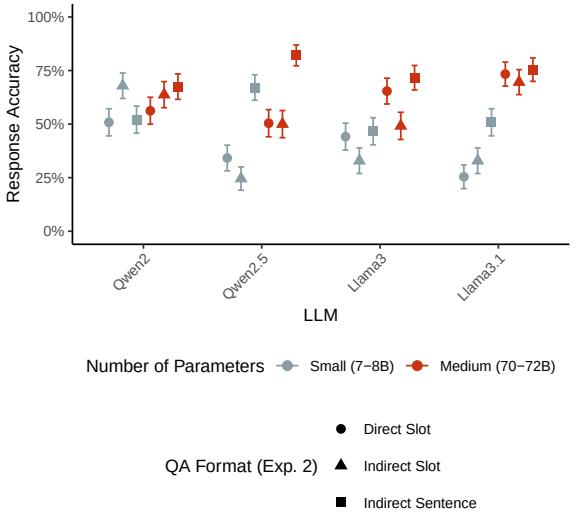

Figure 9: Response accuracy in Experiment 2 by LLM, number of parameters, and modal condition. Error bars represent 95% confidence intervals.

E Story Templates and Examples

E.1 Experiment 1

There are five templates to generate stories in Experiment 1: *the hidden object*, *Whodunnit*, *city travel*, *the grocery store's promotion*, and *fictional characters*. *The hidden object* is a classic task from Hirst and Weil (1982), Noveck et al. (1996), Ozturk and Papafragou (2015). The textual template of *Whodunnit* is inspired by Wu et al. (2024) and Jin et al. (2024). The others are designed by ourselves.

Types	N/P condition	QA format	Prompt	Answer
base	N	direct slot	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box and not in the purple box. The peanut _ be in the blue box. Which word or expression is more adequate to describe the situation after filling in the slot: may or has to? Please ONLY respond with may or has to as your answer.	has to
base	P	direct slot	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box. The peanut _ be in the blue box. Which word or expression is more adequate to describe the situation after filling in the slot: may or has to? Please ONLY respond with may or has to as your answer.	may
base	N	indirect slot	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box and not in the purple box. The peanut _ be in the blue box. Which word or expression is more adequate to describe the situation after filling in the slot: 1 may or 2 has to? Please ONLY respond with 1 or 2 as your answer.	2
base	P	indirect slot	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box. The peanut _ be in the blue box. Which word or expression is more adequate to describe the situation after filling in the slot: 1 may or 2 has to? Please ONLY respond with 1 or 2 as your answer.	1
base	N	indirect sentence	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box and not in the purple box. Which sentence is more adequate to describe the situation? 1 The peanut may be in the blue box. 2 The peanut has to be in the blue box. Please ONLY respond with 1 or 2 as your answer.	2
base	P	indirect sentence	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box. Which sentence is more adequate to describe the situation? 1 The peanut may be in the blue box. 2 The peanut has to be in the blue box. Please ONLY respond with 1 or 2 as your answer.	1
1-shot	N	direct slot	There are three boxes in a room: a yellow box, a red box and a gray box. A peanut is hidden in one of these boxes. If the peanut is not in the gray box, the peanut may be in the yellow box or red box. If the peanut is not in the gray and the red box, the peanut has to be in the yellow box. Now there is another case. There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box and not in the purple box. The peanut _ be in the blue box. Which word or expression is more adequate to describe the situation after filling in the slot: may or has to? Please ONLY respond with may or has to as your answer.	has to
1-shot	P	direct slot	There are three boxes in a room: a yellow box, a red box and a gray box. A peanut is hidden in one of these boxes. If the peanut is not in the gray box, the peanut may be in the yellow box or red box. If the peanut is not in the gray and the red box, the peanut has to be in the yellow box. Now there is another case. There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box. The peanut _ be in the blue box. Which word or expression is more adequate to describe the situation after filling in the slot: may or has to? Please ONLY respond with may or has to as your answer.	may

1-shot	N	indirect slot	There are three boxes in a room: a yellow box, a red box and a gray box. A peanut is hidden in one of these boxes. If the peanut is not in the gray box, the peanut may be in the yellow box or red box. If the peanut is not in the gray and the red box, the peanut has to be in the yellow box. Now there is another case. There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box and not in the purple box. The peanut _ be in the blue box. Which word or expression is more adequate to describe the situation after filling in the slot: 1 may or 2 has to? Please ONLY respond with 1 or 2 as your answer.	2
1-shot	P	indirect slot	There are three boxes in a room: a yellow box, a red box and a gray box. A peanut is hidden in one of these boxes. If the peanut is not in the gray box, the peanut may be in the yellow box or red box. If the peanut is not in the gray and the red box, the peanut has to be in the yellow box. Now there is another case. There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box. The peanut _ be in the blue box. Which word or expression is more adequate to describe the situation after filling in the slot: 1 may or 2 has to? Please ONLY respond with 1 or 2 as your answer.	1
1-shot	N	indirect sentence	There are three boxes in a room: a yellow box, a red box and a gray box. A peanut is hidden in one of these boxes. If the peanut is not in the gray box, the peanut may be in the yellow box or red box. If the peanut is not in the gray and the red box, the peanut has to be in the yellow box. Now there is another case. There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box and not in the purple box. Which sentence is more adequate to describe the situation? 1 The peanut may be in the blue box. 2 The peanut has to be in the blue box. Please ONLY respond with 1 or 2 as your answer.	2
1-shot	P	indirect sentence	There are three boxes in a room: a yellow box, a red box and a gray box. A peanut is hidden in one of these boxes. If the peanut is not in the gray box, the peanut may be in the yellow box or red box. If the peanut is not in the gray and the red box, the peanut has to be in the yellow box. Now there is another case. There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hidden in one of these boxes. It is known that the peanut is not in the black box. Which sentence is more adequate to describe the situation? 1 The peanut may be in the blue box. 2 The peanut has to be in the blue box. Please ONLY respond with 1 or 2 as your answer.	1
counter-factual	N	direct slot	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hiding in the black box, because yesterday Tom moved the peanut from the purple box to the black box. If he hadn't moved the peanut yesterday, imagine which box the peanut would be in today. The peanut _ be in the purple box. Which word or expression is more adequate to describe the situation after filling in the slot: may or has to? Please ONLY respond with may or has to as your answer.	has to
counter-factual	P	direct slot	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hiding in the black box, because yesterday Tom moved the peanut from one of the other two boxes to the black box. If he hadn't moved the peanut yesterday, imagine which box the peanut would be in today. The peanut _ be in the purple box. Which word or expression is more adequate to describe the situation after filling in the slot: may or has to? Please ONLY respond with may or has to as your answer.	may
counter-factual	N	indirect slot	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hiding in the black box, because yesterday Tom moved the peanut from the purple box to the black box. If he hadn't moved the peanut yesterday, imagine which box the peanut would be in today. The peanut _ be in the purple box. Which word or expression is more adequate to describe the situation after filling in the slot: 1 may or 2 has to? Please ONLY respond with 1 or 2 as your answer.	2
counter-factual	P	indirect slot	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hiding in the black box, because yesterday Tom moved the peanut from one of the other two boxes to the black box. If he hadn't moved the peanut yesterday, imagine which box the peanut would be in today. The peanut _ be in the purple box. Which word or expression is more adequate to describe the situation after filling in the slot: 1 may or 2 has to? Please ONLY respond with 1 or 2 as your answer.	1

counter-factual	N	indirect sentence	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hiding in the black box, because yesterday Tom moved the peanut from the purple box to the black box. If he hadn't moved the peanut yesterday, imagine which box the peanut would be in today. Which sentence is more adequate to describe the imagined situation? 1 The peanut may be in the purple box. 2 The peanut has to be in the purple box. Please ONLY respond with 1 or 2 as your answer.	2
counter-factual	P	indirect sentence	There are three boxes in a room: a black box, a purple box and a blue box. A peanut is hiding in the black box, because yesterday Tom moved the peanut from one of the other two boxes to the black box. If he hadn't moved the peanut yesterday, imagine which box the peanut would be in today. Which sentence is more adequate to describe the imagined situation? 1 The peanut may be in the blue box. 2 The peanut has to be in the blue box. Please ONLY respond with 1 or 2 as your answer.	1

E.2 Experiment 2

A ToM story with eight statements in direct slot format is demonstrated below.

Statements	**Certainty**	**Prompt**	**Answer**
Fact (previous)	Not low	You witness a sequence of events: 1 Hannah entered the hall. 2 Amelia entered the hall. 3 The corn is in the drawer. 4 The drawer is in the hall. 5 Amelia hates the grapes. 6 Amelia hates the stockings. 7 Hannah exited the hall. 8 Amelia moved the corn to the bathtub. 9 The bathtub is in the hall. 10 Hannah entered the hall. 11 Charlotte entered the hall. You use the first-person perspective to report the situation: I _ the corn was in the drawer. Which word or expression is more adequate to describe the situation after filling in the slot: know, believe or doubt? Please ONLY respond with know, believe or doubt as your answer.	know
Fact (previous)	Low	You witness a sequence of events: 1 Hannah entered the hall. 2 Amelia entered the hall. 3 The corn is in the drawer. 4 The drawer is in the hall. 5 Amelia hates the grapes. 6 Amelia hates the stockings. 7 Hannah exited the hall. 8 Amelia moved the corn to the bathtub. 9 The bathtub is in the hall. 10 Hannah entered the hall. 11 Charlotte entered the hall. You use the first-person perspective to report the situation: I _ the corn was in the box. Which word or expression is more adequate to describe the situation after filling in the slot: know, believe or doubt? Please ONLY respond with know, believe or doubt as your answer.	doubt
Fact (current)	Not low	You witness a sequence of events: 1 Hannah entered the hall. 2 Amelia entered the hall. 3 The corn is in the drawer. 4 The drawer is in the hall. 5 Amelia hates the grapes. 6 Amelia hates the stockings. 7 Hannah exited the hall. 8 Amelia moved the corn to the bathtub. 9 The bathtub is in the hall. 10 Hannah entered the hall. 11 Charlotte entered the hall. You use the first-person perspective to report the situation: I _ the corn is in the bathtub. Which word or expression is more adequate to describe the situation after filling in the slot: know, believe or doubt? Please ONLY respond with know, believe or doubt as your answer.	know
Fact (current)	Low	You witness a sequence of events: 1 Hannah entered the hall. 2 Amelia entered the hall. 3 The corn is in the drawer. 4 The drawer is in the hall. 5 Amelia hates the grapes. 6 Amelia hates the stockings. 7 Hannah exited the hall. 8 Amelia moved the corn to the bathtub. 9 The bathtub is in the hall. 10 Hannah entered the hall. 11 Charlotte entered the hall. You use the first-person perspective to report the situation: I _ the corn is in the box. Which word or expression is more adequate to describe the situation after filling in the slot: know, believe or doubt? Please ONLY respond with know, believe or doubt as your answer.	doubt

Belief (agent 0)	Not low	You witness a sequence of events: 1 Hannah entered the hall. 2 Amelia entered the hall. 3 The corn is in the drawer. 4 The drawer is in the hall. 5 Amelia hates the grapes. 6 Amelia hates the stockings. 7 Hannah exited the hall. 8 Amelia moved the corn to the bathtub. 9 The bathtub is in the hall. 10 Hannah entered the hall. 11 Charlotte entered the hall. You use the first-person perspective to report the situation: I _ Amelia will look for the corn in the bathtub. Which word or expression is more adequate to describe the situation after filling in the slot: know, believe or doubt? Please ONLY respond with know, believe or doubt as your answer.	believe
Belief (agent 0)	Low	You witness a sequence of events: 1 Hannah entered the hall. 2 Amelia entered the hall. 3 The corn is in the drawer. 4 The drawer is in the hall. 5 Amelia hates the grapes. 6 Amelia hates the stockings. 7 Hannah exited the hall. 8 Amelia moved the corn to the bathtub. 9 The bathtub is in the hall. 10 Hannah entered the hall. 11 Charlotte entered the hall. You use the first-person perspective to report the situation: I _ Amelia will look for the corn in the drawer. Which word or expression is more adequate to describe the situation after filling in the slot: know, believe or doubt? Please ONLY respond with know, believe or doubt as your answer.	doubt
Belief (agent 1)	Not low	You witness a sequence of events: 1 Hannah entered the hall. 2 Amelia entered the hall. 3 The corn is in the drawer. 4 The drawer is in the hall. 5 Amelia hates the grapes. 6 Amelia hates the stockings. 7 Hannah exited the hall. 8 Amelia moved the corn to the bathtub. 9 The bathtub is in the hall. 10 Hannah entered the hall. 11 Charlotte entered the hall. You use the first-person perspective to report the situation: I _ Hannah will look for the corn in the drawer. Which word or expression is more adequate to describe the situation after filling in the slot: know, believe or doubt? Please ONLY respond with know, believe or doubt as your answer.	believe
Belief (agent 1)	Low	You witness a sequence of events: 1 Hannah entered the hall. 2 Amelia entered the hall. 3 The corn is in the drawer. 4 The drawer is in the hall. 5 Amelia hates the grapes. 6 Amelia hates the stockings. 7 Hannah exited the hall. 8 Amelia moved the corn to the bathtub. 9 The bathtub is in the hall. 10 Hannah entered the hall. 11 Charlotte entered the hall. You use the first-person perspective to report the situation: I _ Hannah will look for the corn in the bathtub. Which word or expression is more adequate to describe the situation after filling in the slot: know, believe or doubt? Please ONLY respond with know, believe or doubt as your answer.	doubt

F Statistical Coefficients From Logistic Regression Analyses

F.1 Experiment 1

	Correct Response			
Predictor	b	SE	z	p
(Intercept)	0.87	0.06	14.37	< .001
Number of Parameters [Medium > Small]	1.63	0.12	13.46	< .001
Modal Condition [Necessity > Possibility]	0.87	0.12	7.59	< .001
Story Type 1-Shot [> Base]	−0.21	0.16	−1.31	< .001
Story Type Counterfactual [> Base]	−0.50	0.16	−3.15	.002
QA Format Indirect Slot [> Direct Slot]	−0.66	0.16	−4.23	< .001
QA Format Indirect Sentence [> Direct Slot]	0.55	0.17	3.18	.002
Number of Parameters × Modal Condition	−0.93	0.23	−4.05	< .001
Number of Parameters × Story Type Counterfactual	−1.03	0.29	−3.52	< .001
Number of Parameters × QA Format Indirect Slot	−1.11	0.31	−3.53	< .001
Number of Parameters × QA Format Indirect Sentence	0.78	0.35	2.26	.024
Observations	1800			
R^2_{Tjur}	.186			
AIC	1956.6			

Table 7: Logistic regression results on **Qwen2-7B/72B** data from Experiment 1, derived from the optimal regression model selected by AIC.

	Correct Response			
Predictor	b	SE	z	p
(Intercept)	2.67	0.18	14.86	< .001
Number of Parameters [Medium > Small]	3.15	0.36	8.84	< .001
Modal Condition [Necessity > Possibility]	1.99	0.32	6.31	< .001
Story Type 1-Shot [> Base]	−1.41	0.28	−5.05	< .001
Story Type Counterfactual [> Base]	0.81	0.39	2.09	.036
QA Format Indirect Slot [> Direct Slot]	−0.31	0.20	−1.52	.128
QA Format Indirect Sentence [> Direct Slot]	1.30	0.23	5.78	< .001
Number of Parameters × Modal Condition	1.54	0.63	2.44	.015
Number of Parameters × Story Type 1-Shot	−2.74	0.56	−4.90	< .001
Number of Parameters × Story Type Counterfactual	1.99	0.78	2.57	.010
Observations	1800			
R^2_{Tjur}	.206			
AIC	1246.0			

Table 8: Logistic regression results on **Qwen2.5-7B/72B** data from Experiment 1, derived from the optimal regression model selected by AIC.

F.2 Experiment 2

Predictor	Correct Response b	SE	z	p
(Intercept)	1.30	0.10	13.20	< .001
Number of Parameters [Medium > Small]	2.17	0.20	11.08	< .001
Modal Condition [Necessity > Possibility]	1.79	0.19	9.44	< .001
Story Type 1-Shot [> Base]	0.04	0.17	0.23	.822
Story Type Counterfactual [> Base]	−0.01	0.16	−0.04	.971
QA Format Indirect Slot [> Direct Slot]	−0.79	0.17	−4.69	< .001
QA Format Indirect Sentence [> Direct Slot]	1.34	0.20	6.86	< .001
Number of Parameters × Modal Condition	3.66	0.38	9.63	< .001
Number of Parameters × Story Type 1-Shot	−1.02	0.30	−3.44	< .001
Number of Parameters × QA Format Indirect Slot	−1.19	0.34	−3.51	< .001
Number of Parameters × QA Format Indirect Sentence	2.27	0.39	5.79	< .001
Observations	1800			
R^2_{Tjur}	.189			
AIC	1866.7			

Table 9: Logistic regression results on **Llama3-8B/70B** data from Experiment 1, derived from the optimal regression model selected by AIC.

Predictor	Correct Response b	SE	z	p
(Intercept)	1.70	0.11	15.40	< .001
Number of Parameters [Medium > Small]	2.96	0.22	13.45	< .001
Modal Condition [Necessity > Possibility]	1.14	0.17	6.82	< .001
Story Type 1-Shot [> Base]	−0.02	0.17	−0.14	.887
Story Type Counterfactual [> Base]	−0.52	0.19	−2.67	.008
QA Format Indirect Slot [> Direct Slot]	−0.41	0.18	−2.33	.020
QA Format Indirect Sentence [> Direct Slot]	1.54	0.28	5.61	< .001
Number of Parameters × Modal Condition	1.16	0.33	3.48	< .001
Number of Parameters × Story Type Counterfactual	−0.98	0.36	−2.70	.007
Number of Parameters × QA Format Indirect Sentence	3.43	0.48	7.19	< .001
Observations	1800			
R^2_{Tjur}	.210			
AIC	1659.4			

Table 10: Logistic regression results on **Llama3.1-8B/70B** data from Experiment 1, derived from the optimal regression model selected by AIC.

Predictor	Correct Response b	SE	z	p
(Intercept)	0.62	0.07	8.56	< .001
Number of Parameters [Medium > Small]	0.08	0.15	0.52	.604
Epistemic Certainty [Not Low > Low]	1.64	0.16	9.98	< .001
Statement Type Fact [> Belief]	3.02	0.17	17.34	< .001
Statement Type Agent 1 [> Agent 0]	−0.05	0.18	−0.31	.757
Statement Type Current [> Previous]	0.49	0.23	2.16	.031
QA Format Indirect Slot [> Direct Slot]	0.99	0.21	4.82	< .001
QA Format Indirect Sentence [> Direct Slot]	−0.14	0.20	−0.70	.481
Number of Parameters × Epistemic Certainty	−1.61	0.33	−4.89	< .001
Number of Parameters × Statement Type Fact	−2.06	0.35	−5.92	< .001
Number of Parameters × Statement Type Agent 1	0.88	0.35	2.49	.013
Number of Parameters × Statement Type Current	−0.33	0.45	−0.73	.466
Number of Parameters × QA Format Indirect Slot	−1.72	0.41	−4.19	< .001
Number of Parameters × QA Format Indirect Sentence	1.37	0.39	3.47	< .001
Observations	1440			
R^2_{Tjur}	.384			
AIC	1354.3			

Table 11: Logistic regression results on **Qwen2-7B/72B** data from Experiment 2, derived from the optimal regression model selected by AIC.

Predictor	Correct Response b	SE	z	p
(Intercept)	−0.01	0.09	−0.10	.921
Number of Parameters [Medium > Small]	1.65	0.18	9.15	< .001
Epistemic Certainty [Not Low > Low]	3.08	0.22	14.08	< .001
Statement Type Fact [> Belief]	3.79	0.22	17.28	< .001
Statement Type Agent 1 [> Agent 0]	−0.38	0.22	−1.75	.081
Statement Type Current [> Previous]	0.39	0.25	1.59	.113
QA Format Indirect Slot [> Direct Slot]	−2.65	0.27	−9.73	< .001
QA Format Indirect Sentence [> Direct Slot]	3.99	0.27	14.58	< .001
Number of Parameters × Epistemic Certainty	−2.99	0.44	−6.85	< .001
Number of Parameters × Statement Type Fact	−1.01	0.44	−2.31	.021
Number of Parameters × Statement Type Current	−1.22	0.49	−2.48	.013
Number of Parameters × QA Format Indirect Slot	2.01	0.54	3.70	< .001
Number of Parameters × QA Format Indirect Sentence	−1.53	0.55	−2.81	.005
Observations	1440			
R^2_{Tjur}	.597			
AIC	1009.2			

Table 12: Logistic regression results on **Qwen2.5-7B/72B** data from Experiment 2, derived from the optimal regression model selected by AIC.

Predictor	Correct Response b	SE	z	p
(Intercept)	−0.20	0.09	−2.11	.035
Number of Parameters [Medium > Small]	1.88	0.19	9.87	< .001
Epistemic Certainty [Not Low > Low]	1.36	0.32	4.31	< .001
Statement Type Fact [> Belief]	4.79	0.34	14.07	< .001
Statement Type Agent 1 [> Agent 0]	0.21	0.22	0.97	.331
Statement Type Current [> Previous]	−0.22	0.22	−0.99	.325
QA Format Indirect Slot [> Direct Slot]	−1.76	0.23	−7.81	< .001
QA Format Indirect Sentence [> Direct Slot]	1.26	0.22	5.60	< .001
Number of Parameters × Statement Type Fact	−4.50	0.67	−6.73	< .001
Number of Parameters × Epistemic Certainty	−5.30	0.64	−8.34	< .001
Observations	1440			
R^2_{Tjur}	.546			
AIC	1064.9			

Table 13: Logistic regression results on **Llama3-8B/70B** data from Experiment 2, derived from the optimal regression model selected by AIC.

Predictor	Correct Response b	SE	z	p
(Intercept)	0.59	0.09	6.32	< .001
Number of Parameters [Medium > Small]	2.70	0.19	14.57	< .001
Epistemic Certainty [Not Low > Low]	−0.55	0.17	−3.26	.001
Statement Type Fact [> Belief]	2.74	0.19	14.73	< .001
Statement Type Agent 1 [> Agent 0]	0.43	0.20	2.18	.029
Statement Type Current [> Previous]	0.19	0.22	0.87	.386
QA Format Indirect Slot [> Direct Slot]	−0.46	0.20	−2.23	.026
QA Format Indirect Sentence [> Direct Slot]	1.06	0.21	5.00	< .001
Number of Parameters × Epistemic Certainty	−4.78	0.34	−14.24	< .001
Number of Parameters × Statement Type Fact	2.28	0.37	6.14	< .001
Number of Parameters × Statement Type Agent 1	1.17	0.40	2.97	.003
Number of Parameters × QA Format Indirect Sentence	−1.22	0.37	−3.27	.001
Observations	1440			
R^2_{Tjur}	.465			
AIC	1217.9			

Table 14: Logistic regression results on **Llama3.1-8B/70B** data from Experiment 2, derived from the optimal regression model selected by AIC.

G ROC Curves for Logistic Regression Models

G.1 Experiment 1

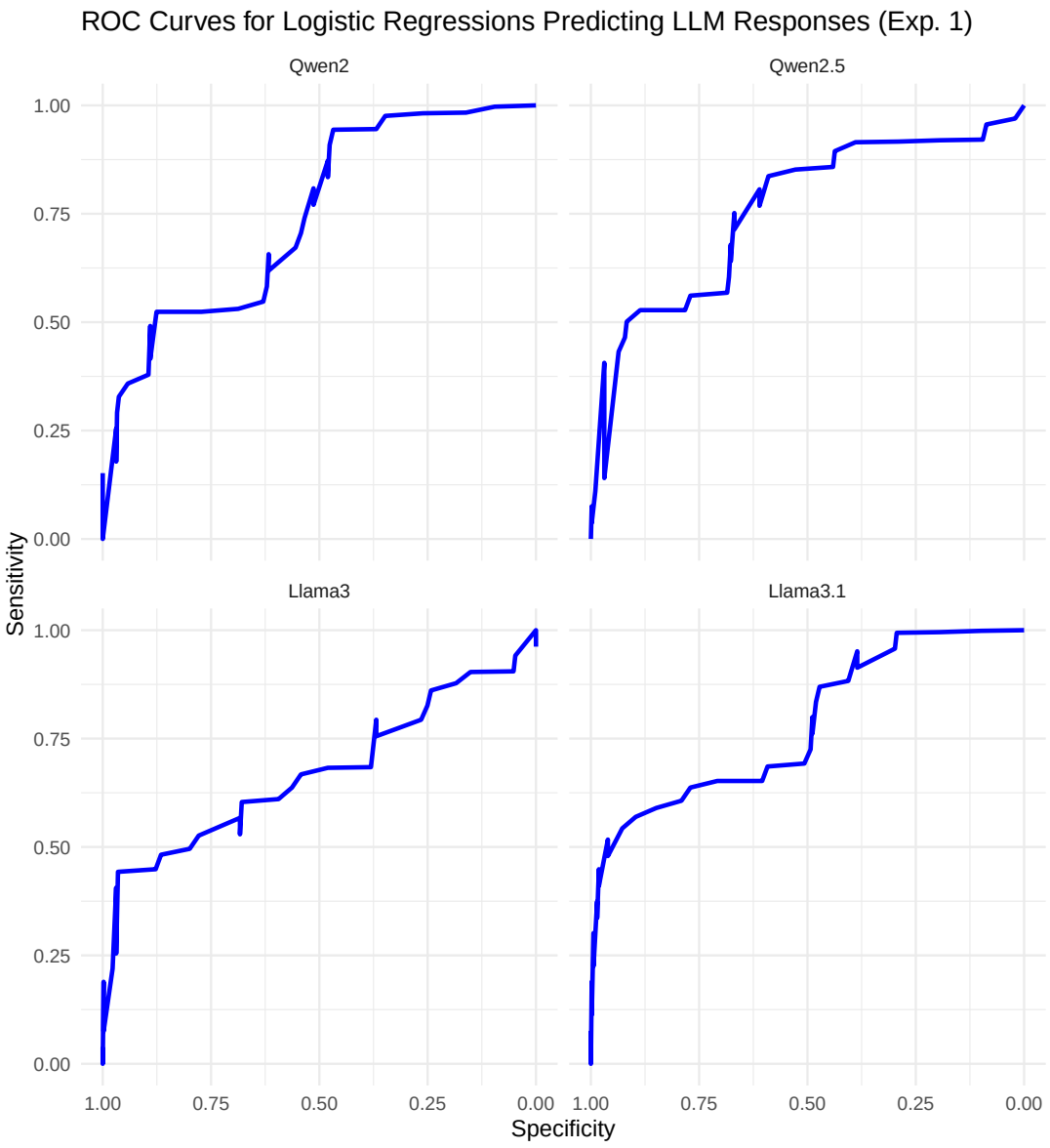

Figure 10: Receiver operating characteristic (ROC) curves for logistic regression models predicting LLM response data from Experiment 1.

G.2 Experiment 2

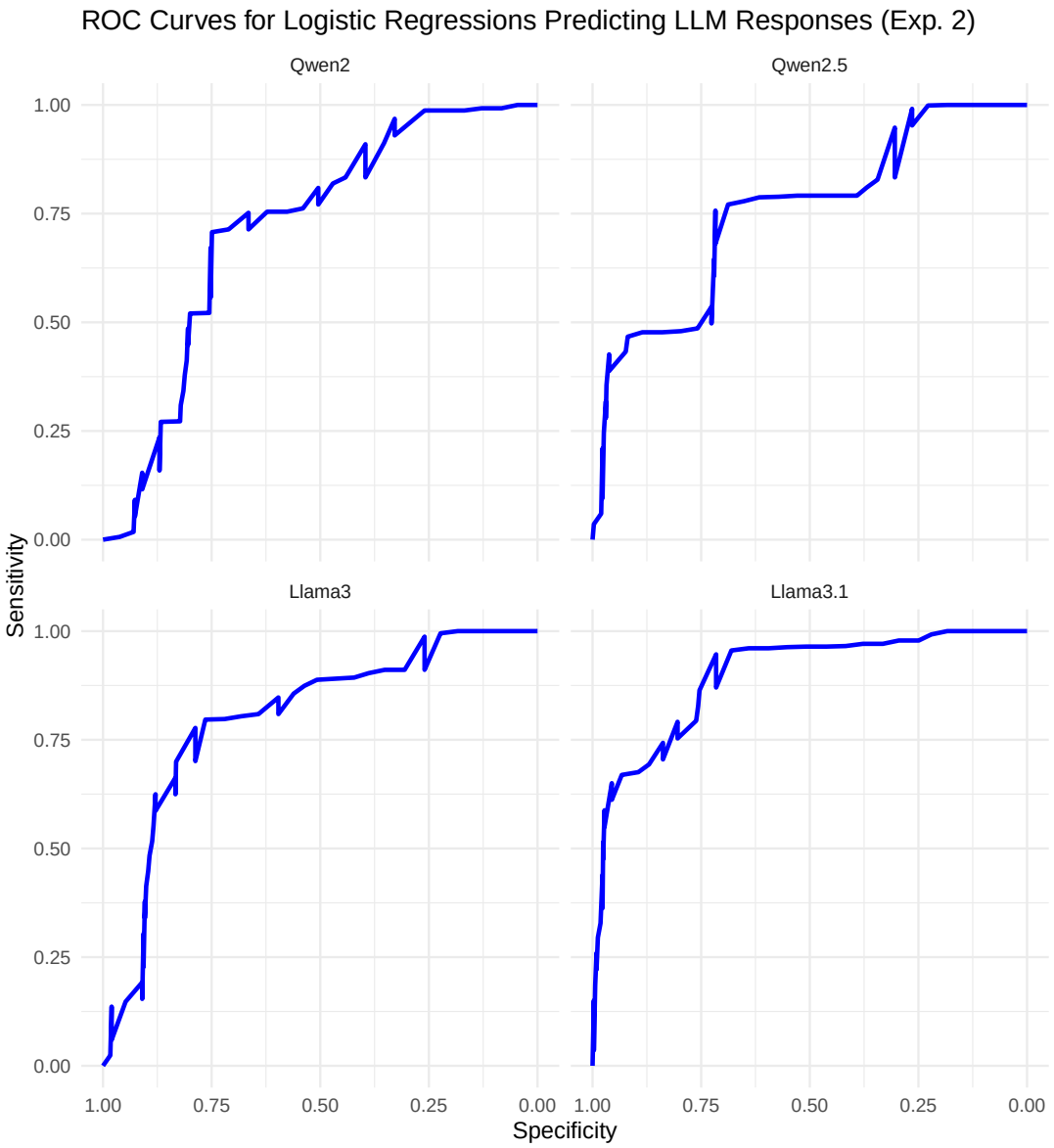

Figure 11: Receiver operating characteristic (ROC) curves for logistic regression models predicting LLM response data from Experiment 2.