Evaluating Theory of (an uncertain) Mind:
Predicting the Uncertain Beliefs of Others from Conversational Cues

Anthony Sicilia and **Malihe Alikhani**
Khoury College of Computer Sciences, Northeastern University
{sicilia.a, m.alikhani}@northeastern.edu

Abstract

Typically, when evaluating Theory of Mind, we consider the beliefs of others to be binary: held or not held. But what if someone is unsure about their own beliefs? How can we quantify this uncertainty? We propose a new suite of tasks, challenging language models (LMs) to model the uncertainty of participants in a dialogue. We design these tasks around conversation forecasting, where the goal is to predict the probability of an unobserved conversation outcome. Uniquely, we view conversation agents themselves as forecasters, asking an LM to predict the uncertainty of an individual from their language use. We experiment with scaling methods, bagging, and demographic context for this regression task, conducting experiments on three dialogue corpora (social, negotiation, task-oriented) with eight LMs. While LMs can explain up to 7% variance in the uncertainty of others, we highlight the difficulty of the tasks and room for future work, especially in tasks that require explicit shifts in perspective.

1 Introduction

Theory-of-mind (ToM) and, specifically, false belief prediction are vital for planning and decision-making in conversation (Ho et al., 2022). While beliefs are often treated as existing in a binary state (held or not held), there are situations where an individual's belief is better represented more flexibly (e.g., held, not held, or *unsure*), capturing uncertainty or belief intensity. For instance, intelligent tutoring systems need to model student uncertainty about course materials to provide effective feedback (Forbes-Riley and Litman, 2009; Jraidi and Frasson, 2013), and in task-oriented settings, people may even be unsure of their goals (Sicilia et al., 2023), impeding success when this uncertainty is not considered (see Figure 1). Meanwhile, recent work suggests that the perceived uncertainty in AI systems does not always align with human uncertainty in dialogue (Testoni and Fernández, 2024).

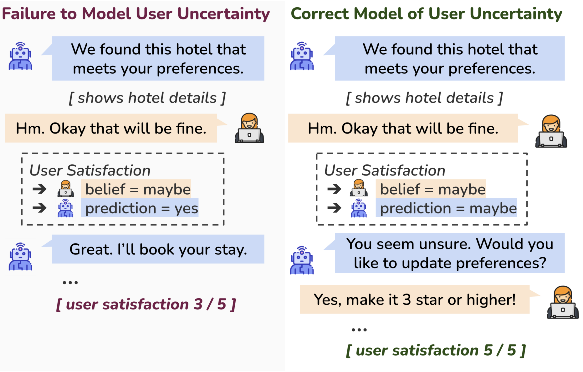
Failure to Model User Uncertainty
We found this hotel that meets your preferences.
[shows hotel details]
Hm. Okay that will be fine.
User Satisfaction
➔ belief = maybe
➔ prediction = yes
Great. I'll book your stay.
...
[user satisfaction 3 / 5]
Correct Model of User Uncertainty
We found this hotel that meets your preferences.
[shows hotel details]
Hm. Okay that will be fine.
User Satisfaction
➔ belief = maybe
➔ prediction = maybe
You seem unsure. Would you like to update preferences?
Yes, make it 3 star or higher!
...
[user satisfaction 5 / 5]

Figure 1: Recognizing uncertainty in others can influence AI dialogue strategies, ultimately improving task-success. Here, an AI assistant recognizes user uncertainty and probes to resolve it, eventually increasing user satisfaction. We formalize tasks to assess model ability to recognize uncertainty from language cues.

This paper studies whether language models can recognize an individual's uncertainty from the language they use in conversation. We study this specific ToM capacity in language models, using conversation forecasting as a tool. Whereas existing forecasting tasks (Sokolova et al., 2008; Zhang et al., 2018; Sicilia et al., 2024) focus on predicting aleatoric factors of uncertainty, which are inherent to the data and independent of perspective, we focus on predicting subjective and individual factors of uncertainty held by each conversation participant. In particular, we ask models to forecast *the uncertainty of others' beliefs* as well as *the uncertainty of others' beliefs about others*.

As noted, the ToM tasks we study emphasize an important capability for conversational agents, like language models: their ability to reason about others' mental states, and particularly, uncertainty. Conversational agents use these skills to collaborate and achieve goals, making these intimately related to communicative grounding – the collaborative process whereby language is used to establish mutual understanding (Clark and Brennan, 1991).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8007–8021
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

While the latter focuses on how shared knowledge can be established through direct evidence, such as acknowledgment (e.g., *okay, got it*), ToM enables reasoning about beliefs without explicit validation, instead using language cues to make assumptions. Both processes can work together during a conversation, as when recognizing uncertainty triggers behaviors like acknowledgment (Nilsenová, 2001).

To formalize belief uncertainty, we build on a traditional statistical view, where uncertainty is measured via probability (Bröcker, 2009), granting us an established framework to design our tasks. Belief uncertainty can be measured on a ternary scale (yes, no, maybe) or more general spectrum (e.g., a Likert scale) and we discuss some strategies to calibrate these human annotations to real-world probabilities. Meanwhile, to capture differences in individual perspective, we disentangle this probabilistic notion into two components – the epistemic (subjective) and the aleatoric (ground-truth) – common to modern studies of uncertainty in machine learning (Hüllermeier and Waegeman, 2021). Interestingly, our formal setup defines a series of regression tasks, allowing us to explore the relatively unexplored area of continuous inference with language models (Vacareanu et al., 2024). In this context, we study the relation between traditional methodologies, like bagging, and recent language modeling methods, like self-consistent chain-of-thought (Wang et al., 2023) for the first time.

In initiating this evaluation of ToM about uncertainty, we offer a few contributions:

1. we formalize "false uncertainty" – a concept akin to false belief – and connect it to Theory of Mind and forecasting (§ 2.2), using this to motivate a new task suite (§ 3.2);
2. we propose new methods to forecast others' uncertainty with language models (§ 4), studying ways to use language models for regression and calibrate probability estimates with continuous labels (rather than discrete);
3. we study impacts of individual demographics, goals, and other context on model ToM (§ 5).

From experiments (§ 5) across three corpora (social, negotiation, and task-oriented) and eight models, our findings suggest that language models are able to explain some of the variance in others' uncertainty (up to 7%). Yet, we also observe the difficulty of this task, even for humans, making code open-source to promote progress.[1]

2 Conversation Forecasting and ToM

We focus on the setup of Sicilia et al. (2024) where an agent observes a conversation and is asked to express their uncertainty about a potential outcome for this conversation; e.g., "*How much does Speaker A like Speaker B?*" or "*Will the negotiation result in a deal?*" As implied, the conversation is just a partial window into the true (or, eventual) ground-truth. Hidden information, such as future events or mental states, creates an inherent randomness about reality, which may not be fully determined by the available evidence. In this context, we assume a (human) agent forms a mental model capturing their uncertainty about the outcome – a "forecast" about whether the outcome will occur.

2.1 Comparing Forecasts with Ground-Truth

Given a (potentially partial) conversation and any accompanying evidence about the situation (e.g., interlocutor context), the forecasting agent expresses their uncertainty P about the outcome of interest. For now, we assume P is a probability estimate, but later allow other expressions of uncertainty (§ 3.1). The forecast P is evaluated by Brier score:

$$\mathrm{BS} = (P - O)^2 \tag{1}$$

where O is a binary indicator of the outcome (e.g., 1 if a deal occurs, 0 else). Forecasters with accurate uncertainty estimates (agreeing exactly with the distribution of O) will have a lower Brier score than other, less accurate forecasters. Brier score also ranks sub-optimal forecasts with consideration of both calibration and variance (Bröcker, 2009).

2.2 The Missing Building Blocks for ToM

We observe that Brier score, alone, does not capture the full story about an agent's uncertainty P. Indeed, the Brier score measures two individual aspects of uncertainty:

$$\mathbf{E}[\mathrm{BS}] = \underbrace{\mathbf{Var}[O]}_{\text{aleatoric uncertainty}} + \underbrace{\mathbf{E}[(P-p)^2]}_{\text{epistemic uncertainty}} \tag{2}$$

where p is the probability $O = 1$. While the outcome variance captures the inherent randomness of the forecasting task, the latter quantifies the forecaster's excess errors that should not be attributed to this randomness. These model-specific aspects of error are the **epistemic uncertainty** (Lahlou et al., 2022; Hüllermeier and Waegeman, 2021).

[1] https://github.com/anthonysicilia/forecasting-tom

CaSiNo	CANDOR	MultiWOZ
S1: Nice to interact with you! **S2**: Same here. I hope to get some items for my family... **S1**: Yes, let's make a deal that benefits both families. ...*negotiation continues* **S2**: I get it, but I guess we will have to compromise... **S1**: I'd like the option of 2 food packs. **S2**: In that case I'll take additional firewood and water **S1**: Sounds fair...	**S1**: our office is neat. We have people with all sorts of different backgrounds... **S2**: mm **S1**: theater, **S2**: hmm, **S1**: um like business **S2**: mhm **S1**: the managing partner of our office. **S2**: Yes. **S1**: Her degree is in art **S2**: wow. ...*conversation continues*	**S1**: what type of attraction? **S2**: I don't know, let's say a museum. I need the address too. **S1**: castle galleries is quite nice, they are at unit su43, grande arcade, saint andrews st. Free admission too **S2**: Free is the right price tag for me. I appreciate all your help, that's it for today. Have a great day! **S1**: Glad I could help, have a great day. **S2**: You too and thank you for your help. I'm looking forward to a nice day.
*How certain is **S1** they are more satisfied than would occur by chance?* → Ground-truth $P = 11\%$ → Predicted $\hat{P} = 40\%$	*How certain is **S2** they like **S1** more than would occur by chance?* → Ground-truth $P = 27\%$ → Predicted $\hat{P} = 90\%$	*How certain is **S2** they are more satisfied than would occur by chance?* → Ground-truth $P = 81\%$ → Predicted $\hat{P} = 90\%$

Table 1: Dialogues adapted from examples in each corpus. Below these, model prompts are shown for the 1st-Order ToM task, demonstrating "more than chance" calibration strategy, ground-truth uncertainty, and estimates by Meta's Llama 3.1 70B.

Integrating ToM in Forecasting Uniquely, we consider the epistemic uncertainty of human interlocutors (treated as forecasters) in a conversation. This dual interpretation captures the individual aspects of an interlocutor's uncertainty by comparing their forecast to ground-truth. Precisely, it measures the fluctuations in uncertainty caused by the interlocutor themselves – their knowledge, perceptions, and biases – rather than those (fluctuations) which may be attributed to changes in ground-truth. Based on the epistemic uncertainty, we define an interlocutor's **false uncertainty** as:

$$\mathrm{FUn} = P - p. \tag{3}$$

False uncertainty similarly captures subjective fluctuations, but preserves the direction of this subjectivity, distinguishing between positive (over-confident) or negative (under-confident) forms of uncertainty. Quantifying false uncertainty will be the primary motivation for our task design. While works have focused on improving the quality of a forecast (the Brier score), ours is first to propose quantification of other interlocutors' uncertainty.

2.3 Related Studies of Theory-of-Mind

Theory-of-Mind is often evaluated (in language models) using question-answering (Nematzadeh et al., 2018; Le et al., 2019; Sap et al., 2022); which, for instance, mimics common ToM evaluations from psychology, like the Sally-Anne test. Other proposals study machine ToM in situated and collaborative environments (Bara et al., 2021; Ma et al., 2023b; Li et al., 2023), focus on higher-order ToM (Wu et al., 2023), or consider ToM beyond (more common) belief/false belief anticipation (van Duijn et al., 2023). Inference-time methods to improve ToM in language models have also been studied (Takmaz et al., 2023; Sclar et al., 2023). Yet, whenever beliefs are involved, they are typically assumed to held or not held. Our forecasting setting resolves this by emphasizing the potential uncertainties attached to a belief.

While our work is the first to operationalize the forecasting setting for purpose of a dedicated theory-of-mind evaluation suite, it is important to acknowledge myriad other pragmatic reasoning tasks that can also involve modeling of belief uncertainty (Fried et al., 2023). For example, these include reference games (Monroe et al., 2017) and certain goal-oriented dialogue tasks (Haber et al., 2019). Other formal models of (uncertain) communication, such as the Rational Speech Acts framework, are also well-equipped to handle notions of uncertainty in a ToM context (Goodman and Stuhlmüller, 2013). These tasks and frameworks are related to our current framing, but our use of the forecasting setting enables a more precise and focused task definition.

Lastly, it is important to discuss the formal definition of Theory-of-Mind. Quesque and Rossetti (2020) suggest ToM evaluations should require representation of a mental state that differs from one's own (*non-merging criterion*) and ensure task-success cannot be based on lower-level processes (*mentalizing criterion*). For example, inferring someone's uncertainty based purely on one's own opinion or observed environmental factors would not constitute ToM. Since language models do not have mental states, non-merging can be achieved by information-asymmetry (Kim et al., 2023), which forces the model to predict distinct (non-merged) perspectives. In forecasting, this distinction lies in the interlocutor uncertainty P and ground-truth p, whenever $P \neq p$. As for *mentalizing*, this criterion is undermined by spurious data correlation, to which language models are susceptible (Kim et al., 2023; Shapira et al., 2024). Our current, natural conversation corpora (§ 3.3) do not necessarily exclude the possibility that models are actually using spurious features to make their predictions. Albeit, it has been suggested socially situated tasks, like ours, can mitigate potential confounding (Ma et al., 2023a). Using controlled data can completely remove confounding, which we leave as future work.

3 New Uncertainty Quantification Tasks

3.1 Human Expressions of Uncertainty

As we are (uniquely) interested in humans as conversation forecasters, probability annotations are not necessarily the most effective way to elicit uncertainty or intensity of belief. Indeed, most of the corpora we study (§ 3.3) annotates belief intensity on a Likert scale; e.g., "on a scale from 1 to 10, how much do you think Speaker B likes you." We focus on probability estimates because these can be compared to ground-truth world states; i.e., whether B actually "likes," to enforce non-merging (§ 2.2). Without "the world" or "reality" as reference, we have no way to define subjective, or false, uncertainty. Thus, we map human expressions to probability estimates to enable comparison.

Calibration Strategy: "More Than Chance" Mapping verbal or quasi-continuous expressions of belief uncertainty to probabilities is a calibration problem; e.g., it has been approached for language models using scaling (Tian et al., 2023). In this work, we enable calibration by making a slight semantic change to the outcome of interest. Instead of studying "whether Speaker B likes Speaker A" we study "whether Speaker B likes Speaker A more than would occur by chance." This alteration ties belief intensity annotations to a ground-truth outcome that is observable in data. Precisely, following the colloquial meaning of "more than chance" in the statistics literature, the ground-truth probability is defined by a p-value for the magnitude of the belief, computed from the data. In turn, appending "more than chance" defines both ground-truth outcome probabilities and an appropriate calibration function for human expressions of intensity. We provide details in § A.1.

3.2 Uncertainty Quantification (UQ) Tasks

1st-Order ToM Uncertainty (1TUQ) To quantify false uncertainty, one first needs to quantify an interlocutor's base uncertainty about their belief (i.e., the forecast P). Aptly, our first task evaluates a language model's ability to quantify the base uncertainty of others. For instance, suppose an interlocutor A expresses their uncertainty about "whether A is happy" and this is calibrated to a probability forecast P.[2] The language model's task is to make a prediction $\hat{P}$ about the uncertainty of A's belief. We evaluate this prediction using regression metrics; e.g., the correlation between P and $\hat{P}$, the absolute error, and the explained variance.

2nd-Order ToM Uncertainty (2TUQ) Besides their own beliefs, interlocutors also hold uncertainty about the beliefs of others. For instance, an interlocutor A can express their uncertainty about "whether interlocutor C likes A". Then, TUQ tasks the language model with quantifying the uncertainty of A about C's belief. Similar to the first-order task, we evaluate a language model's prediction by comparing it to A's true uncertainty.

False Uncertainty (FUnQ) Finally, we ask language models to directly quantify an interlocutor's false uncertainty. In essence, this requires them to quantify both the interlocutor's uncertainty about a belief as well as the ground-truth probability that the belief is true (the outcome probability p). For instance, P may be a forecast about "whether interlocutor C likes A" and p may be the ground-truth probability that "C likes A." The language model is tasked with quantifying $\mathrm{FUn} = P - p$, and we evaluate this estimation using regression metrics.

3.3 Corpora and Basic Prompts

CaSiNo is a corpus of negotiations about camp-resource allocation (Chawla et al., 2021). Interlocutors barter over available resources, such as fire-wood and water, based on (assigned) resource preferences. Performance-based monetary incentives stimulate competitive behaviors. Interlocutors indicate their satisfaction with the final deal on a 5-point scale. For an interlocutor A, we ask language

[2] Recall, belief intensity needs calibration to a world outcome to make sense as a probability; e.g., "A tells friends about happiness." We use an outcome observable in the data, i.e. "whether A is happier than would occur by chance."

models to predict “how certain A is that they are more satisfied than would occur by chance.” Precise details are in § A.3. This formulation allows us to evaluate language models for 1st Order ToM uncertainty quantification (1TUQ). The average number of tokens in a conversation is 320.

CANDOR is a corpus of spoken conversations between strangers, conducted over video communication platform (Reece et al., 2023). Conversations are social in nature with minimum time constraints and an assigned goal of “getting to know each other.” Exit interviews (conducted privately) ask interlocutors to quantify how much they like each other on a 7-point scale, as well as how much they *think* their conversation partner likes them. For two interlocutors A and B, we ask language models to predict “how certain is B that they like A more than would occur by chance.” As with CaSiNo, this lets us evaluate language models at the first-order task (1TUQ). Because of the available data, we also ask language models to predict “how certain is A that B likes A more than would occur by chance.” As we discuss in § 4, this lets us evaluate models at the second-order task (2TUQ) and False Uncertainty Quantification (FUnQ). The average token-count is 11K, but we only show models the first 5K.

MultiWOZ is a task-oriented Wizard-of-Oz corpus wherein one human plays the role of a conversational booking system (for hotels, restaurants, etc.) and the other plays as user (Eric et al., 2020). Additional annotations (Sun et al., 2021) designate perceived satisfaction of the user by crowd-workers on a 5-point scale. This annotation is less organic than previous datasets, but it can be considered a representative proxy, capturing how annotators might feel if they were in the user’s position. As such, we ask language models to predict “how certain the user is that they are more satisfied than would occur by chance” using the average crowd-worker annotation as ground-truth. This allow us to evaluate 1TUQ. The average token-count is 460.

4 Methods

4.1 Forecasting the Uncertainty of Beliefs

Direct Forecasting (`DF`, Sicilia et al., 2024) is a good “out-of-the-box” method for uncertainty-aware conversation forecasting with language models. Adapted to our belief anticipation problem, we prompt the language model to express its predicted uncertainty for the interlocutor on a 10-point scale. We parse the prediction directly from the model’s sampled completion and divide by 10 to get an estimate $\hat{P}$ for the interlocutor’s true forecast P. In general, we use the Chain of Thought (CoT) strategy proposed by Kojima et al. (2022), asking the model to approach the prediction “step-by-step.”

4.1.1 Post-Hoc Scaling

Post-hoc scaling (calibration) tends to improve direct forecasts (Tian et al., 2023; Sicilia et al., 2024), requiring only a small amount of data. Notably, our ToM tasks work with continuous uncertainty annotations in place of traditional, discrete outcome annotations. We propose new scaling methods to accommodate our data.

Platt Scaling: `DF (PS)` One option is to assume the relationship between the true uncertainty P and the predicted uncertainty $\hat{P}$ is linear in the logits; e.g., this is common in soft classification (Platt et al., 1999). In our new setting,

$$\mathrm{logit}(P) \approx \alpha \cdot \mathrm{logit}(\hat{P}) + \beta. \tag{4}$$

The new (re-scaled) forecast is:

$$\hat{P}_{\mathrm{PS}} = \mathrm{expit}\big(\alpha \cdot \mathrm{logit}(\hat{P}) + \beta\big) \tag{5}$$

where α, β are the MLE estimates of Eq. (4).

Linear Scaling: `DF (LS)` We also suggest linear scaling, which instead learns a direct linear map:

$$\hat{P}_{\mathrm{LS}} = \mathrm{clip}(\alpha \cdot \hat{P} + \beta, 0, 1). \tag{6}$$

4.1.2 Fine-Tuning a Regression Head (`FT`)

In place of direct forecasts, fine-tuning a classification head on a language model’s latent features can help boost performance in soft classification (Kadavath et al., 2022). Again, since our annotations are continuous, we slightly modify this, replacing the classification head with a regression head. Specifically, using the same prompt as `DF`, the language model encodes latent features $\mathbf{x}$ and inference is conducted as:

$$\hat{P}_{\mathrm{FT}} = f_\theta(\mathbf{x}) \tag{7}$$

where f_θ is the regression head. We test a linear regression head denoted `FT (L)`, a 2-layer ReLU-network head denoted `FT (NN)`, and a random forest head denoted `FT (RF)`. Details are in § A.4.

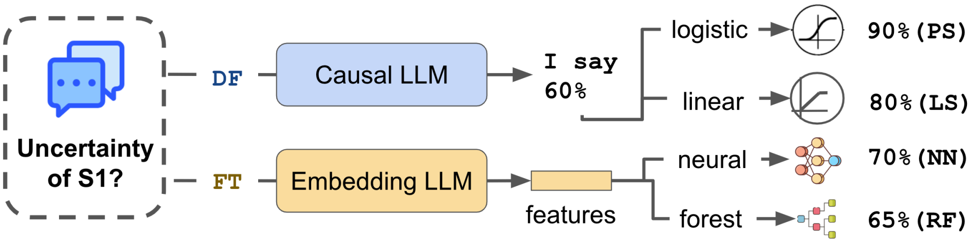

Figure 2: Comparison of workflow for direct forecasting with calibration (DF) and fine-tuning with regression-heads (FT). The Bag-of-Thoughts (BoT) method is applied during the causal LLM step, aggregating predictions before any logistic or linear scaling. If included, demographic information (DEM) is used to modify prompts at the start of the workflow.

4.1.3 Bias and Variance

Classically, for a fixed probability P, the MSE of a corresponding estimate $\hat{P}$ can be decomposed:

$$\mathbf{E}[(P - \hat{P})^2] = \mathbf{Var}[\hat{P}] + \mathbf{Bias}^2(P, \hat{P}) \quad (8)$$

where $\mathbf{Bias}(P, \hat{P}) = \mathbf{E}[\hat{P}] - P$. This points to two possible ways we can reduce error, discussed next.

Bagging (BoT) Bagging or Bootstrap Aggregating trains many models on random samples from the same data and averages the predictions of all the models. For instance, this is how random forests (RF) are trained (Breiman, 2001). It is a known variance reduction strategy, and is why we explore RF as a regression head. Another strategy we can take, with language models, is to "bag" the Chain-of-Thought (CoT) inferences generated when we prompt models to "think step-by-step." We suggest re-sampling the direct forecasts produced by this CoT prompt many times ($n = 10$) and averaging these to make an inference. The changes in sampling distribution triggered by each new explanation make this distinct from greedy decoding (see Table 7). It is akin to traditional bagging, except we re-sample model explanations, instead of training data. We call this approach a "Bag of Thoughts" (**BoT**). Notably, our strategy is similar to, yet distinct from, self-consistent decoding in classification (Wang et al., 2023), where a majority vote is used to aggregate multiple CoT inferences. With majority votes, variance reduction is not necessarily a plausible motivation, since the bias-variance trade-off is not well defined (Brown and Ali, 2024) and reducing variance can actually increase error (James, 2003). In contrast, our method uniquely connects CoT aggregation to variance reduction.

Demographic Data (DEM) provides important background information about the interlocutor in question. Particularly, we consider the age, sex, race, and education of the interlocutor. We hypothesize these characteristics reduce prediction bias because they add situational context, an important aspect of Theory of Mind (Ma et al., 2023b). Moreover, language models are known to inherent and propagate certain social biases (Gallegos et al., 2023) and improved context representations – such as those achieved by making demographics clear – can be an effective means to mitigate the biases in generative inference (Sicilia and Alikhani, 2023).

4.2 Forecasting False Uncertainty

A straightforward way to predict false uncertainty is to have the model shift perspectives across two inference steps. For instance, in CANDOR, we predict A's uncertainty about "whether C likes A..." and then predict the ground-truth probability that "C likes A...", shifting perspectives from A to the outside world (C's belief, in this case). Denoting these $\hat{P}$ and $\hat{p}$, respectively, we combine estimates:

$$\hat{\mathrm{FUn}} = \hat{P} - \hat{p}. \quad (9)$$

The same strategies discussed in § 4.1 can be used to make the individual components of this inference, learning $\hat{p}$ and $\hat{P}$, separately. Alternatively, one can estimate $\hat{\mathrm{FUn}}$ directly. This makes the most sense in the fine-tuning setting, where both latent representations (from the $\hat{p}$ and the $\hat{P}$ prompt) can interact in a (non-linear) regression head to improve inference. -J denotes this joint strategy.

5 Experiments

We use the 3 datasets/prompting schemes discussed in § 3.3. More details on prompts are in § A.3. We use 5 different random seeds to create 5 distinct train/test splits. For training, $n = 100$ unless otherwise noted. Models are listed in tables with version numbers and inference strategies detailed in § A.2

Metrics We report standard regression metrics including the Pearson (linear) correlation R, the Spearman (rank) correlation ρ, the mean absolute

`method`	`xl`	MAE	R^2	min	max
`DF`	✗	44.9	-370	-620	-110
`DF (LS)`	✗	**22.2**	**1.5**	-4.1	7.5
`DF (PS)`	✗	31.9	-180	-740	-1.8
`FT (L)`	✗	22.5	-1.0	-2.3	1.1
`DF (LS)`	✓	22.1	**2.1**	-5.3	12.5
`FT (L)`	✓	22.2	0.1	-5.1	3.7
`FT (NN)`	✓	22.3	-9.2	-30.5	5.4
`FT (RF)`	✓	**21.9**	1.3	-5.4	7.6

Table 2: Regression performance on first-order Theory of Mind uncertainty quantification (1TUQ). No BoT or demographic data is used. Direct forecasts (`DF`) are linearly correlated ($R = 0.14$) before scaling, but only linear post-hoc scaling (`LS`) calibrates these forecasts to be good predictors (R^2 up to 12.5%). A slight non-linear, monotone relationship also exists ($\rho = 0.16$), but this is not well-modeled by Platt scaling (`PS`). Tuning a regression head (`FT`) in place of direct forecasting (`DF`) does not explain more variance, even with 8x more data (`xl`). With the same data, `DF (LS)` performs best.

error MAE, and the % of variance in the test data explained by the predictions R^2. Explaining more variance is better, but it's not typical to explain all of it ($R^2 = 100\%$). For reference, explained variance in (traditional) forecasting tasks with language models is low (up to 10% Brier Skill Score – a type of explained variance for soft classifiers, Sicilia et al., 2024). Generally, we use train data to compute the mean when estimating variance on the test set (called "out-of-sample" R^2), which provides a fair evaluation on held-out sets. In this context, R^2 can also be interpreted as percent improvement compared to a constant mean prediction. For MAE, we report the *additive* error in % probability (e.g., $|0.2 - 0.4| \times 100\% = 20\%$). Finally, metrics are micro-averaged over all data splits.

5.1 Results & Analysis

We structure our results using a research question (RQ) / answer (A) format with trailing discussions.

> *RQ1: Can models predict the uncertainty of others from conversation cues?*
> *A: No. Inference "out-of-the-box" is poor. Some simple methods do improve.*

Comparison of Scaling Methods Table 2 reports regression metrics for first-order ToM UQ (1TUQ) split according to different methods of inference. While direct forecasts are ineffective "out-of-the-box," linear scaling (`DF LS`) with 100 data points can improve scores to a positive explained variance, on average. These results suggest a consistent (if slight) linear relationship between the language model's inferences and the interlocutors' true uncertainty. Explained variance sometimes exceeds 7%, or with more data, 12%. Contrary to conventional wisdom (using soft classifiers to forecast outcomes), a logit-linear relationship between the model's inferences and it's target seems unlikely, due to the poor performance of `DF PS`.

> *RQ2: Does variance reduction via bagging improve inference capability?*
> *A: Yes. Random forests trained on language model embeddings show promise. The proposed Bag of Thoughts (BoT) strategy also improves inference.*

Variance Reduction Strategies Use of bagging in fine-tuning (i.e., via random forests) did improve performance as anticipated, compared to other tuning strategies. We recall, bagging is a known variance reduction strategy, which can ultimately reduce errors by this mechanism. Another variance reduction strategy we propose is Bag of Thoughts (BoT). Table 3 reports ablation study of BoT for first-order TUQ, limited to CANDOR and CaSiNo. Ablation is also reported for second-order TUQ, limited to CANDOR, in Table 5. Findings show BoT has positive impact on small models on average, with particular models/setups seeing substantial gain (2% bump for `Gemma 7B` on 1TUQ, more for `GPT 3.5` on 2TUQ). Performance is amplified more so in Table 4 (includes MultiWOZ). Averaged across all 1TUQ data, BoT allows small models to surpass some large models (particularly, `Llama3 70B`). We did not try BoT for large models, as their lower throughput (tokens/second) made repeated sampling time consuming. Comparison between BoT and greedy decoding is in Table 7.

Why BoT Works Because we use BoT on direct forecasts, and *then* scale them, BoT actually reduces variance in the feature space of the linear scaling function (not necessarily the predictions). Indeed, comparing before/after BoT shows an increase in the standard deviation of the prediction (+0.5% proba.). Meanwhile, in feature space (the pre-scaled forecast), the STD decreases by 1.5% probability. Our hypothesis is that variance reduction in feature space (by BoT) actually increases the signal-to-noise-ratio, mitigating the effects of outlier inferences from the model. Increase correlation between pre-scaled forecasts and ground-truth after applying BoT (+0.03) may confirm this.

> *RQ3: Can interlocutor demographic information be used to improve ToM UQ?*
> *A: Yes, depending on model size.*

BoT	**DEM**	`Llama3 8B`	`Mix 8x7B`	`Gemma 7B`	`GPT 3.5`	`Avg`	`Llama3 70B`	`Mix 8x22B`	`GPT 4o`
✗	✗	0.7	1.3	0.0	-0.3	0.4	0.1	2.5	2.2
✗	✓	-1.5	0.7	0.7	-1.0	-0.3	**0.7**	**3.2**	**3.3**
✓	✗	**0.8**	**1.4**	**1.9**	0.0	1.0	✗	✗	✗
✓	✓	0.8	0.6	0.7	**0.6**	0.7	✗	✗	✗

Table 3: R^2, i.e., % explained variance, of direct forecasts (`DF LS`) micro-averaged across CANDOR and CaSiNo for 1TUQ task. These results ablate use of demographic data in prompt and BoT method (§ 4.1.3). Demographic context (**DEM**) helps larger models, while smaller models fail to effectively use it. **BoT** tends to help, or have no effect, on small models, regardless of demographic context. We test BoT for small models only due to inference cost constraints.

Dataset	`Llama3 8B`	`Mix 8x7B`	`Gemma 7B`	`GPT 3.5`	`Llama3 70B`	`Mix 8x22B`	`GPT 4o`	`Avg`	`Hum`
CANDOR	-0.2	**2.3**	0.2	0.1	-0.4	**1.0**	0.6	0.5	2.7
CaSiNo	1.7	0.5	**3.6**	0.0	0.6	**3.9**	3.7	2.0	✗
MultiWOZ	**4.8**	2.6	4.4	3.3	4.5	2.8	**6.8**	4.2	✗
Avg	2.1 (0.6)	1.8 (1.6)	2.7 (0.7)	1.1 (0.0)	1.5	2.5	**3.7**		

Table 4: Explained variance R^2 for direct forecasts (`DF LS`) on 1TUQ task separated by data and model. BoT is used for small models only, with ablation in parentheses. Models show varied success across different corpora, meanwhile BoT improves small models to outperform others 10x larger. `Hum` denotes human R^2, after linear scaling `LS` (MAE=17.7).

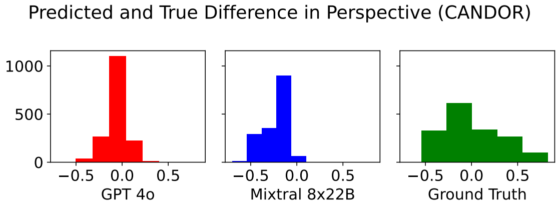

Figure 3: Differences in perspective captured by false uncertainty (FUn) on CANDOR. Model predictions (without scaling) and ground-truth values are shown. Humans exhibit highly variable differences in perspective, whereas models tend to show negative bias, underestimating FUn, with less overall variance. Predictions concentrated near zero show that models fail to distinguish between interlocutor mental states.

Use of Demographic Context In Table 3 we ablate the role of including demographic data in the prompt (**DEM**), limited to first-order TUQ on CANDOR and CaSiNo. With or without BoT, adding demographics tends to hurt performance of small models (0.7% and 0.4% drop in average R^2, respectively). Meanwhile, the scaled inferences of larger models are all improved by including demographics. In similar ablation for second-order TUQ (Table 5), we did find less conclusive evidence of a distinction between smaller and larger models use of demographic context. For instance, without BoT, demographics seem to help both model groups.

Demographics and Bias Our initial hypothesis was the bias reduction was the principle mechanism by which demographic context could reduce error. This is consistent (for large models, 1TUQ) with observed reduction in bias after including demographics (-0.1%). The limited effect size does suggest potential for other factors. For instance, similar to variance reduction, interplay between demographic data and scaling may play a role. On the other hand, models (in general) may be ineffective in using demographic context.

> *RQ4: What factors of conversation context impact recognition of uncertainty?*
> *A: Conversation length and speaker motives may play a role. Models generally have trouble with perspective shift.*

Data Comparison Table 4 reports explained variance for 1TUQ for `DF LS`, split by model and dataset. We observe CANDOR to be the most difficult dataset for 1TUQ, followed by CaSiNo, then MultiWOZ. One hypothesis for the difficulty of CANDOR is the length of it's conversations, which average more than 11K tokens (GPT-2 tokenizer). This may also be compounded because dialogue is between strangers. Small, but important, nuances can become dominated other – perhaps, superficially polite – interactions. The reality that humans can hide their true mental states may also explain increased difficulty in CaSiNo, a negotiation corpus. Rather than “acting” to be polite, interlocutors in the CaSiNo corpus hide motives and intentions, as a strategy, to receive a better deal. In contrast, in the collaborative and task-oriented MultiWOZ corpus, interlocutors have incentive to reveal many aspects of their mental state; e.g., to indicate satisfactory constraints for their booking task.

Why FUnQ is Hard Most methods exhibit poor performance on the False Uncertainty Quantification (FUnQ) task. One possibility is that this difficulty may, in part, come because model errors

BoT	**DEM**	Llama 3 8B	Mix 8x7B	Gemma 7B	GPT 3.5	Avg	Llama 3 70B	Mix 8x22B	GPT 4o
✗	✗	-1.4	-0.9	-0.7	-0.4	-0.9	0.4	1.0	0.1
✗	✓	0.0	1.4	-0.7	-0.2	0.1	**0.7**	**1.8**	**0.7**
✓	✗	**0.8**	0.6	-1.8	**1.8**	0.4	✗	✗	✗
✓	✓	0.6	**1.5**	**-0.6**	-0.3	0.3	✗	✗	✗

Table 5: R^2, i.e., % explained variance, of direct forecasts (DF LS) for 2TUQ task on CANDOR. As before, **BoT** helps smaller models, often pushing them to perform at the level of larger counterparts. In second-order ToM UQ, demographic data (**DEM**) appears to help most models. Albeit, concurrent use of BoT (small models) complicates this result with varied performance.

method	xl	MAE	R^2	$\min R^2$	$\max R^2$
DF (LS)	✗	24.2	-0.9	-2.5	-0.1
DF (LS)	✓	24.3	-1.9	-7.0	-0.2
FT (RF)	✓	24.8	-5.2	-8.9	-2.2
FT (RF-J)	✓	23.9	0.3	-1.7	1.4

Table 6: Regression metrics for False Uncertainty prediction (FUnQ) on CANDOR. Even with more data, neither direct forecasting (LS) nor fine-tuning is able to explain variance in the ground-truth false uncertainty. False uncertainty prediction is a more difficult task, requiring models to perspective shift, from the interlocutor's uncertainty to that of the outside world.

compound across multiple inference steps; e.g., the inference for interlocutor's uncertainty $\hat{P}$ and the inference for the ground-truth probability $\hat{p}$. Indeed, one data point in favor of this hypothesis is the positive explained variance of the joint fine-tuning procedure (FT RF-J), which conducts non-linear inference over the embedding of both prompts (i.e., to infer $\hat{P}$ and $\hat{p}$) and then produces a single estimate for the difference $P - p$. On the other hand, another problem may come from a models' inability to flexibly shift perspectives (e.g., from the mental state of Speaker A to the mental state of Speaker B). Figure 3 shows a qualitative analysis, comparing distributions of False Uncertainty (both ground-truth and base predictions by models). Results lend evidence to the hypothesis that models simply fail to distinguish between interlocutor perspectives due to predicted FUn concentration near 0, among other pitfalls.

RQ5: How do humans compare?
A: Slightly better than language models.

Human Performance Because of available annotations (§ 3.3), we infer human performance at first-order ToM UQ on CANDOR. Interestingly, linear scaling also improves the performance for human forecasts, which may be suggestive of individual baselines for how people express their uncertainty (or, intensity) about beliefs. Human performance is not drastically higher than models ($R^2 = 2.7\%$, MAE=17.7), which is again suggestive of the difficulty of this corpus (recall, our data comparison).

RQ6: Can uncertainty estimates improve model inference at routine ToM?
A: Yes. § A.6 shows uncertainty estimates can improve F1 at belief classification.

6 Conclusions

This paper details tasks and methods to explore if language models can infer subjective uncertainties from language. We connect this capacity to communicative grounding and Theory-of-Mind, suggesting the ability to infer an interlocutor's uncertainty from linguistic cues is fundamental to both. Methodologically, we capture this by proposing a continuous analog of false belief recognition (i.e., false uncertainty quantification) and discuss various techniques that map from model representations of uncertainty to estimates of interlocutor belief. Our high-level findings suggest that precisely quantifying the uncertainty of others can be difficult for both models and humans. While top-performing models do reasonably well in structured goal-oriented settings like MultiWOZ, explaining about 7% variance in belief uncertainty, top-performance degrades to only 1% in CANDOR, a dataset characterized by long, informal, and socially-motivated conversations. Here, both humans and models struggle, likely due to unclear conversation goals and ambiguous belief expression – humans still beat models by a small margin. We also test ways to improve model performance, showing mixed effects for different models.

Potential for Progress Although available comparisons to human performance suggest constrained headroom, qualitative analyses show some consistent model errors that future methods can aim to address. These include: failure to distinguish between individual perspectives, underrepresention of population-level variances in belief uncertainty, and persistent biases. Our results also suggest dependencies between individual expressions of uncertainty and demographics, which may be of interest to broad research communities.

Limitations

As noted in our conclusions, there are many aspects of this research which call for continued development. For instance, we study relatively simple fine-tuning strategies and do not explore downstream applications directly in our experiments. The purpose of this work is to propose an evaluation methodology and some initial algorithms. We acknowledge the limited scope of our initial experiments, with respect to these topics of future work.

Moreover, while we do study three diverse corpora, generalization of our findings to new data is not guaranteed. Different outcomes or corpora may show worse results from the studied language models. Scientific conclusions drawn about this data may not generalize to new corpora either. Reproducibility studies and greater data collection is needed to allow the tasks we propose to be studied at a larger scale, to mitigate concerns about generalization of methods/findings.

While we do motivate this work from the point-of-view of ToM, we do recall some potential caveats we brought up when outlining ToM evaluation criteria (**ToM Criteria**). Namely, our setup does not explicitly control for potential spurious correlations between the dialogues we use as features and the target predictions we evaluate as a proxy for ToM capability. For this reason, one should be careful to interpret our results as suggesting language models have (any level) of ToM capability. With that said, future research can incorporate more data controls (without changing our task designs) to help mitigate any present limitations from the corpora we study. It should also be noted that it is not clear whether ToM is a capability that can ever truly exist in language models, at least in a human sense. Exactly what counts as "reasoning" about human's mental states and what is purely "statistical parrot"-like behavior is a topic of debate, which this paper does not aim to address. Simply, we aim to propose tasks that can evaluate (or, approach evaluation) of abilities traditionally associated with ToM in humans.

Finally, there are many other aspects of ToM, besides anticipation of beliefs and false beliefs, which we do not cover in this paper. Some mentioned related works include study of other sub-topics, under the broad umbrella of ToM. We focus on false beliefs because of the practical value that considering uncertainty of beliefs can have.

Ethics

The data we use in this study is either publicly available or available at the request of the respective dataset authors. Our use of this data is consistent with any license or terms of use attached to the data. Some of the data used in this study may contain personally identifying information. Appropriate care should be taken, whenever re-creating our evaluation setup, about the consequences of this fact. Details on these and other ethical considerations are discussed by the datasets original authors.

The models we study, and therefore methods we propose (built on these models), may have social biases inherited from their training data. Although bias mitigation and other safety protocols can be put in place to mitigate concerns, using these models can also introduce (unexpected) biases however they are deployed. For instance, if used in the applications we suggest, they might change a dialogue systems policy, and this change could disproportionately impact a subset of users. The scale of impact is made greater by the ability to use our methods without much human supervision. Thus, deployment of these models should consider the potential impact on users and other far-reaching consequences that improperly moderated use of our models can create.

Acknowledgments

This research was supported in part by Other Transaction award HR0011249XXX from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program. Thanks to Mert Inan and Dilek Hakkani-Tur for helpful feedback.

References

AI@Meta. 2024. Llama 3 model card.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Jochen Bröcker. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519.

Gavin Brown and Riccardo Ali. 2024. Bias/variance is not the same as approximation/estimation. *Transactions on Machine Learning Research*.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Kate Forbes-Riley and Diane Litman. 2009. Adapting to student uncertainty improves tutoring dialogues. In *Artificial intelligence in education*, pages 33–40. IOS Press.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640.

Dan Fu, Simran Arora, Jessica Grogan, Isys Johnson, Evan Sabri Eyuboglu, Armin Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. 2024. Monarch mixer: A simple sub-quadratic gemm-based architecture. *Advances in Neural Information Processing Systems*, 36.

Dan Fu, Simran Arora, and Christopher Ré. 2023. Monarch mixer: Revisiting bert, without attention or mlps.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The photobook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

Mark K Ho, Rebecca Saxe, and Fiery Cushman. 2022. Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11):959–971.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Gareth M James. 2003. Variance and bias for general loss functions. *Machine learning*, 51:115–135.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Imène Jraidi and Claude Frasson. 2013. Student's uncertainty modeling through a multimodal sensor-based approach. *Journal of Educational Technology & Society*, 16(1):219–230.

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*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

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.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor I Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2022. Deup: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*.

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, Hong Kong, China. Association for Computational Linguistics.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023a. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 15–26.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023b. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

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, Brussels, Belgium. Association for Computational Linguistics.

Marie Nilsenová. 2001. Uncertainty in the common ground.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

François Quesque and Yves Rossetti. 2020. What do theory-of-mind tasks actually measure? theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science advances*, 9(13):eadf3197.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

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, St. Julian's, Malta. Association for Computational Linguistics.

Anthony Sicilia and Malihe Alikhani. 2023. Learning to generate equitable text in dialogue from biased training data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2898–2917.

Anthony Sicilia, Yuya Asano, Katherine Atwell, Qi Cheng, Dipunj Gupta, Sabit Hassan, Mert Inan, Jennifer Nwogu, Paras Sharma, and Malihe Alikhani. 2023. Isabel: An inclusive and collaborative task-oriented dialogue system. *Alexa Prize TaskBot Challenge 2 Proceedings*.

Anthony Sicilia, Hyunwoo Kim, Khyathi Raghavi Chandu, Malihe Alikhani, and Jack Hessel. 2024. Deal, or no deal (or who knows)? forecasting uncertainty in conversations using large language models. *arXiv preprint arXiv:2402.03284*.

Marina Sokolova, Vivi Nastase, and Stan Szpakowicz. 2008. The telling tail: Signals of success in electronic negotiation texts. In *Proceedings of the Third*

International Joint Conference on Natural Language Processing: Volume-I.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2499–2506.

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernandez. 2023. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4198–4217, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Alberto Testoni and Raquel Fernández. 2024. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–275, St. Julian's, Malta. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Robert Vacareanu, Vlad-Andrei Negru, Vasile Suciu, and Mihai Surdeanu. 2024. From words to numbers: Your large language model is secretly a capable regressor when given in-context examples. *arXiv preprint arXiv:2404.07544*.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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, Singapore. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

A Appendix

A.1 Extreme Value Uncertainty

Given an annotation m about an interlocutor A's magnitude of belief, we consider the forecasting problem with outcome $\gamma =$ "whether A's magnitude of belief is more extreme than would be observed by chance." Then, in the context of the full dataset, the magnitude annotation m implicitly defines the ground-truth probability of our outcome:

$$p = \mathbf{P}\{m > M\} \tag{10}$$

where M is sampled from all dialogue annotations, and thus, p can be computed from data. A separate agent (e.g., another interlocutor B) can annotate their own perception m' of A's belief, which is an expression of uncertainty/intensity of their belief about A's belief. This can then be calibrated to a probability estimate about the outcome γ, using the same formula:

$$P = \mathbf{P}\{m' > M\}. \tag{11}$$

The important qualities of this outcome formulation are that: (a) it implicitly defines both uncertainty annotations and (ground-truth) calibration functions, which are not available in typical forecasting problems; and (b) it is general, since it implicitly models certainty about any belief for which we have magnitude annotations.

The semantics of the outcome are, in fact, not much different than a more typical decision "whether A believes _____" instead asking a question about relativity of belief, to induce the (implicit) certainty annotations from those (magnitude annotations) that already exist. As a caveat, this outcome format does not work for calibrating uncertainty/intensity of beliefs about many types of "future events"; e.g., whether a deal will occur. In these contexts, the human expression may need to be calibrated with data in order to map human expressions of belief intensity to the same scale as ground-truth outcome probabilities for comparison.

A.2 Models

Direct forecasting (`DF`) is conducted with `Llama3 8B` and `70B` (AI@Meta, 2024), `Mixtral 8x7B` and `x22B` (v0.1 Jiang et al., 2024), `Gemma 7B` (Team et al., 2024), `GPT 3.5` (turbo-0125, OpenAI) and `GPT-4o` (2024-05-13, OpenAI). All models are instruction-tuned (chat) versions. We use default sampling parameters, given on the API or model repository. For fine-tuning (`FT`), we use latent representations from a pre-trained masked language model, specifically fine-tuned for long-context embedding (M2-BERT, Fu et al., 2024), which regularly beats much larger models at embedding tasks (Fu et al., 2023). We use Together AI and Open AI APIs for inference.

A.3 Prompts

We use a common system prompt for all models:

You are TheoryOfMindGPT, an expert language model at using your theory-of-mind capabilities to predict the beliefs and actions of others in human conversations. You will be given a potentially unfinished conversation between two speakers. Put yourself in the mindset of the speakers and try to reason about the requested conversation outcome. Use the keyword "CERTAINTY" to report your prediction for the outcome of interest. Report your answer on a scale from 1 to 10 with 1 indicating "not likely at all" and 10 indicating "almost certainly". For example, "CERTAINTY = 7".

The user-role prompt is also common, with slight variations by corpora, or inclusion of demographics. Here is an example for MultiWOZ.

In the following conversation segment, a human user is interacting with an AI task assistant. `**insert conversation, set off by white space**`. *Now, fast-forward to the end of the conversation. How certain is the user that they (the user) are more satisfied than would occur by chance? Let's think step by step, but keep your answer concise (less than 100 words).*

We found that "Let's think step by step" increased the rate of an explanation associated with the answer. This agrees with the findings of Kojima et al. (2022). We also found that "keep your answer concise (less than 100 words)" was important to prevent models from going on too long without providing an answer (more than 256 tokens).

A.4 Optimization

We used `scikit-learn` to implement all scaling and fine-tuning algorithms (Pedregosa et al., 2011). Ordinary least squares is used to optimize both scaling methods, while SGD is used for the linear fine-tuning method. The neural regression head has a single hidden layer of dimension 100. The random forest has 100 trees of maximum depth 5. All other optimization parameters (e.g., for regularization) are the defaults of the library.

	Llama3	Mixtral	Gemma	GPT	Avg
greedy	0.5	2.2	2.2	-0.7	1.0
BoT	**4.8**	**2.6**	**4.4**	**3.3**	**3.8**

Table 7: Comparison of greedy sampling (temperature = 0) and BoT, illustrating distinction between them. Results show explained variance on MultiWOZ (1TUQ). BoT beats greedy sampling for all models. As final model prediction is conditioned on preceding explanation, setting temperature to 0 does not provide an adequate summary of the true (intractable) sampling distribution. BoT, on the other hand, approximates the mean of the true sampling distribution and provably lowers the variance of the model inference.

A.5 Model Comparison

Table 4 reports explained variance for 1TUQ for DF LS, split by model and dataset. GPT-4o and Mix 8x22B offer the best performance with GPT beating out Mixtral for first place, primarily on MultiWOZ. Interestingly, Gemma 7B (with BoT) outperforms two of the larger models, on average. The performance of Llama3 70B was also surprising, as it is often improved by much smaller models (if they use BoT). Table 5 also reports explained variance, split by model, for 2TUQ. Here, Gemma does not perform as well and neither does GPT-4o (i.e., the fair comparison across tables is *without* demographic data). The most successful models are the Mixtral models, suggesting a unique advantage from their training data (closed-source) or their MoE architectures. GPT 3.5 also shows promise in 2TUQ, under one setting (BoT, no demographics), but is less robust to perturbations among settings.

A.6 Case Study: Does Considering Uncertainty Improve ToM Predictions Outright

Throughout the paper, we have argued for the importance of estimating the uncertainty in others' beliefs, pointing to substantial existing literature as well as a few motivating examples. Here, we show how reasoning about others' uncertainty can even help language models to improve their accuracy at a traditional ToM task (i.e., an existing belief prediction task). Specifically, we use a belief prediction task built on one of the experimental corpora from § 5: CaSiNo. In this campsite negotiation corpora, annotations for satisfaction are provided on a Likert scale, but have clear semantic descriptions (e.g., "very satisfied"), making it easy to create a binary labeling scheme for the outcome "speakers are satisfied" or not. We use the scheme outlined by Sicilia et al. (2024) in their conversation forecasting work, where a conversation is labeled 1 if both speakers are satisfied and 0 otherwise. We ask the model to make this prediction in two ways: (1) with an estimate of the speakers' uncertainty about this outcome and (2) without this estimate of uncertainty, making a simple binary prediction. For the uncertainty estimate, all answers greater than 5 are mapped to a prediction of 1 (i.e., a prediction that both users are satisfied). Overall, we use a largely similar prompt as shown in § A.3 and used in our previous experiments. Results in Table 8 are promising, showing that a language model's inferences can be more accurate when they reason about the uncertainty of others' beliefs to make predictions (rather than making a binary choice). This is especially true for the smaller Llama 3 model. These results are in line with our central argument that considering uncertainty in ToM is an important skill to evaluate.

	ACC		F1	
uncertainty	✗	✓	✗	✓
Llama 3 8B	60.5	65.5	71.7	77.8
Llama 3 70B	68.5	70.5	80.0	82.2

Table 8: Accuracy and F1 of Llama 3 series models on **CaSiNo** corpora. We consider a binary conversation outcome – whether both speakers are satisfied with the negotiation – as suggested by Sicilia et al. (2024), to test our uncertainty estimates on a simple ToM belief prediction task. Results are shown with and without use of uncertainty estimation (§ 4.1) to make the prediction. These results show how reasoning about others' uncertainty can help language models, even in more traditional ToM tasks.