Anthropomimetic Uncertainty: What Verbalized Uncertainty in Language Models is Missing

Dennis Ulmer, **Alexandra Lorson**, **Ivan Titov**, **Christian Hardmeier**
ILLC, University of Amsterdam CLCG, University of Groningen
ILCC, University of Edinburgh IT University of Copenhagen
Pioneer Centre for Artificial Intelligence
dennis.ulmer@mailbox.org, a.lorson@rug.nl,
chrha@itu.dk, ititov@inf.ed.ac.uk

Abstract

Human users increasingly communicate with large language models (LLMs), but LLMs suffer from frequent overconfidence in their output, even when its accuracy is questionable, which undermines their trustworthiness and perceived legitimacy. Therefore, there is a need for language models to signal their confidence in order to reap the benefits of human-machine collaboration and mitigate potential harms. *Verbalized uncertainty* is the expression of confidence with linguistic means, an approach that integrates perfectly into language-based interfaces. Most recent research in natural language processing (NLP) overlooks the nuances surrounding human uncertainty communication and the biases that influence the communication of and with machines. We argue for *anthropomimetic uncertainty*, the principle that intuitive and trustworthy uncertainty communication requires a degree of imitation of human linguistic behaviors. We present a thorough overview of the research in human uncertainty communication, survey ongoing research in NLP, and perform additional analyses to demonstrate so-far underexplored biases in verbalized uncertainty. We conclude by pointing out unique factors in human-machine uncertainty and outlining future research directions towards implementing anthropomimetic uncertainty.

kaleidophon/
anthropomimetic-uncertainty

1 Introduction

In March 2025, the Columbia Journalism Review tested whether different artificial intelligence (AI) chatbots could accurately identify and cite news articles (Jaźwińska and Chandrasekar, 2025). Given a snippet, the chatbots were tasked to identify information such as article date, headline and publisher. Not only did many chatbots give incorrect but plausible-looking answers, but they also stated them with complete confidence: For instance, ChatGPT answered incorrectly 67% of times, hedged its answers using phrases such as "it appears" or "it's possible" in only 7.5% of responses, and never declined to answer any question. This investigation gives a glimpse into the risks of LLMs assistants; while they remain a powerful tool, their replies can often be misleading or wrong. Potential exposure is high—for instance, Eloundou et al. (2024) estimate that LLMs are relevant to 14% of tasks across 1016 analyzed jobs. Simultaneously, adoption is increasing (Humlum and Vestergaard, 2024) with LLMs being used for everyday tasks (Wang et al., 2024b), and freelance work being substituted (Teutloff et al., 2025). Humans have a tendency to perceive LLMs as human (Miller, 2019; DeVrio et al., 2025). This is problematic since LLMs, unlike humans, tend to hallucinate like in the example above, i.e. produce plausible-looking but wrong information (Guerreiro et al., 2023; Ye et al., 2024) with unwarranted confidence (Krause et al., 2023; Yona et al., 2024; Xiong et al., 2024). This behavior resembles strategies used by humans for deception, and reveals a deeper challenge: the usual cognitive mechanisms against being accidentally or intentionally misinformed are less effective in human–machine interactions. Users can only assess the response, but not the reasoning process or communicative intent. Assessing content is difficult as people increasingly treat LLMs as sources of information (Gude, 2025), and thus, lack the knowledge to evaluate its plausibility. This risks a potentially irreversible loss of trust (Dhuliawala et al., 2023; Zhou et al., 2024), as well as negative outcomes. Yet, consistent uncertainty communication can help build extrinsic trust (Jacovi et al., 2021), but despite growing attention in NLP as *verbalized*

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1526–1561, 2026. https://doi.org/10.1162/TACL.a.739
Action Editor: Kristina Toutanova. Submission batch: 7/2025; Revision batch: 1/2026; Published 7/2026.

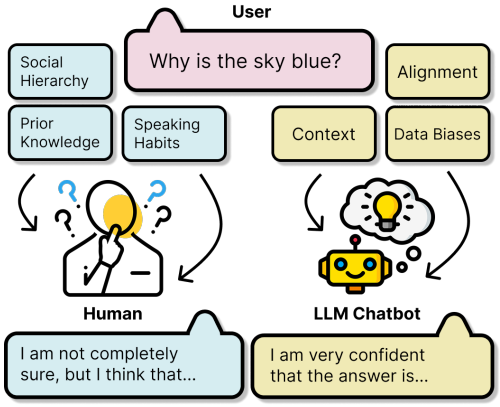


Figure 1: Humans consider various factors when expressing uncertainty, including prior knowledge or the relationship to the speaker. LLMs are biased by their training data, conversational context, and alignment for helpfulness, leading to overconfident and untrustworthy responses.

uncertainty (Lin et al., 2022; Mielke et al., 2022), its basis in human communication and the implications for AI remain underexplored: As depicted in Figure 1, our work discusses how human uncertainty is influenced by factors such as prior knowledge, speaking habits and social hierarchy, while LLM uncertainty is shaped by conversational context, data biases, and its alignment procedure.

Contributions. In this work we introduce *anthropomimetic uncertainty*: Grounding verbalized uncertainty in norms of human uncertainty communication in order to improve reliability, trustworthiness and personalization. To this end, Section 2 first gives a detailed account of the nuances in human uncertainty communication by drawing from the linguistics literature. After a short introduction into uncertainty quantification in NLP in Section 3.1, Section 3.2 contrasts the linguistic account with the current state of verbalized uncertainty research in NLP: By summarizing and annotating the existing literature, we demonstrate that most works still rely on insufficient prompting methods and solely use numerical expressions. This is not enough: We show through frequency analysis in different corpora that the occurrences and linguistic markers is highly dependent on the domain of text. We analyze the uncertainty communication in the generations of five different commercial LLMs, empirically showing that their use of uncertainty expressions is heavily influenced by conversational context and domain. We demonstrate that they exhibit extreme variability and inconsistent calibration, with calibration errors of up to 44%. When using strengtheners and weakeners, they also show the whole range from perfect usage to anti-correlation when compared to the actual answer correctness. To build a path forward, we discuss unique challenges arising in human-machine pragmatics that are linked to uncertainty communication, such as anthropomorphization and sycophancy. Lastly, we propose a list of actionable research directions towards anthropomimetic uncertainty, such as: Rethinking calibration, personalization of uncertainty, expansion to more languages, and considering other factors alongside uncertainty that build user trust.

2 Human Communication of Uncertainty

Although predictive uncertainty can be represented in many ways (Hoarau et al., 2024), humans are intuitively able to express and interpret uncertainty verbally. This section first provides a better understanding of the norms and mechanisms in human uncertainty communication. We distinguish two *registers* of verbal uncertainty, namely numerical and verbal expressions, where words such as *certain*, *might* or *maybe* are called epistemic markers.

Human communication is understood as the intentional exchange of relevant propositions (Grice, 1975; Sperber and Wilson, 1995)—a proposition being the core meaning or claim of a sentence, being evaluated as true or false. Both speaker and addressee invest communicative effort, guided by cooperative principles to be informative, truthful, relevant, and clear (Grice, 1975). As a result, the addressee gains true and relevant information, while the speaker achieves their goal of eliciting an effect or response in the addressee (Sperber et al., 2010). A key aspect of cooperative behavior is acknowledging uncertainty when the speaker is unsure about a proposition. While a speaker expresses some level of confidence in "*Mo will pass the exam*", they can express uncertainty via verbal (e.g., *It is possible that Mo will pass the exam*) and numerical expressions (e.g., *There is a 40% chance Mo will pass the exam*). Thus, comprehension involves a tentative stance of trust (Holton, 1994), such that the addressee can revise their beliefs based on the speaker's contribution,

which is advantageous given that communication is so ubiquitous (Sperber, 2001; Bergstrom et al., 2006). However, the interests of speaker and addressee rarely align perfectly: while the addressee expects true and relevant information, the speaker may instead choose the message that is most likely to produce the desired effect, regardless of its truth—for instance by under- or overstating one's confidence. This divergence makes the addressee exercise caution in order to avoid being misled, engaging in *epistemic vigilance*, a cognitive capacity that helps them guard against both accidental and intentional misinformation (Sperber et al., 2010). According to Sperber et al., the credibility of information depends primarily on two factors: the perceived reliability of the source and the plausibility of the content. Thus, comprehension should neither be equated with belief nor adopting a skeptical stance. In the following section, we examine how uncertainty is expressed, how it can lead to miscommunication due to imprecision, why speakers may strategically adjust their confidence, and what this means for human-machine interaction.

2.1 Verbal uncertainty expressions

One way of communicating uncertainty is through verbal expressions involving epistemic markers that range from factive verbs (*notice*, *think*, *believe*), modal verbs (*must*, *might*), to adverbs (*probably*, *certainly*), which can form increasingly complex expressions (*a slight yet plausible chance*). Their meaning is often captured by mapping them onto probabilities, intervals, or thresholds (Lichtenstein and Newman, 1967; Beyth-Marom, 1982; Wesson and Pulford, 2009; Lassiter, 2010; Vogel et al., 2022).[1] However, their meaning is inherently imprecise and context-dependent, lacking consistent interpretation across speakers and situations. Empirical studies suggest that interlocutors interpret (Mosteller and Youtz, 1990) and produce (van Tiel et al., 2022) the epistemic marker *possible* to refer to probabilities between 17%–89%, with an average of 42%. There are various factors shown to influence the production and interpretation of verbal uncertainty expressions, including event severity and base rates: participants interpreted verbal expressions as indicating higher likelihoods for more frequent events compared to rarer ones (Wallsten et al., 1986). When base rates were controlled, these expressions also signaled greater likelihood for events with more severe consequences relative to less severe ones (Weber and Hilton, 1990; Bonnefon and Villejoubert, 2006; Harris and Corner, 2011). Despite their imprecision, verbal uncertainty expressions help speakers approximate their confidence, as they can rarely compute precise probabilities.

2.2 (Inter-)personal Motivations

Epistemic markers add complexity to uncertainty communication, as they can be used to either strengthen or hedge utterances (e.g., Fraser, 1975; Holmes, 1982). Strengthening *You misunderstood this paragraph* involves high-certainty markers (in bold), e.g., *I **am sure** that you misunderstood this paragraph* (Hyland, 1999, 2005); and hedging low-certainty markers, e.g., *You **might** have misunderstood this paragraph* (Holtgraves and Perdew, 2016). A speaker may strengthen a statement for persuasion, or hedge for politeness (Brown and Levinson, 1978). Whether speakers engage in polite behavior depends on social hierarchies or relative power dynamics: speakers are more likely to use hedging with someone of higher social status, such as a manager, teacher, or parent (Brown and Levinson, 1978; Holtgraves, 2010, 2014; Holtgraves and Perdew, 2016; Juanchich and Sirota, 2015). Speakers are more likely to hedge when discussing serious consequences such as poor financial decisions (Juanchich and Sirota, 2015). This is taken into account by addressees: *possible* was found to be interpreted as more likely when used for a more severe condition (e.g., deafness) than a less severe but equally prevalent condition (e.g., insomnia) (Bonnefon and Villejoubert, 2006). Whether speakers strengthen or hedge their statements is also shaped by how much commitment and accountability they are willing to assume, balancing persuasive intent with the potential social consequences of being wrong. The stronger the expression of commitment (e.g., *I am sure that the deadline is tomorrow*), the more the speaker endorses the claim. This in turn increases scrutiny if wrong, while hedging can help speakers avoid negative social consequences and preserve credibility (Ochs, 1976). Research shows that although confident speakers seem persuasive at first, their credibility declines when proven inaccurate, whereas hedged statements foster more

[1] E.g. quantitative semantic approaches pose that any verbal expression is true given a threshold in $[0, 1)$ if the probability of a described event exceeds it (Lassiter, 2010, 2017; Moss, 2015; Swanson, 2006; Yalcin, 2007, 2010).

resilient trust (Tenney et al., 2007, 2008; Vullioud et al., 2017). Independent of motivation, miscommunication may occur if the addressee fails to take into account the speaker's perspective in choosing a verbal uncertainty expression or by ignoring the speaker-addressee dynamics. Importantly, LLMs might be intuitively perceived to follow similar motivations, but we argue later in Sections 3.3 and 3.4 how this can lead to uncertainty miscommunication through biases in their training.

2.3 Numerical Expressions

Given the complexity of verbal expressions, numerical formats may appear more precise and less error-prone. These include percentages, fractions, decimals, and ratios, which may convey relatively precise information (e.g., *a 65% probability*) or relatively imprecise estimates in the form of ranges (e.g., *60% to 70%*). Previous studies found a tendency that speakers prefer verbal and addressees numerical expressions (Wallsten et al., 1993; Erev and Cohen, 1990). However, preference for either has been shown to vary depending on the domain and context. For example, in medicine (Juanchich and Sirota, 2020) and intelligence analysis (Barnes, 2016), contexts requiring a high precision and where miscommunication can come with dire consequences, speakers tend to prefer numerical expressions,[2] with more serious consequences prompting a stronger preference (Juanchich and Sirota, 2020).

Although numerical expressions of uncertainty may seem precise, numbers like 65 in *there is a 65% chance of rain* often represent approximations, as multiples of 5 are treated as round numbers (Ruud et al., 2014; Cummins, 2015; Jansen and Pollmann, 2001; Krifka, 2007). Speakers frequently round for simplicity (e.g., saying 50 instead of 47), and such numbers are typically understood as approximate. In contrast, non-round numbers like 52 imply precision; referring to 52 as 50 is usually acceptable, but the reverse can seem misleading (Lasersohn, 1999). Speakers often use round numbers when precision is unnecessary, aligning with Grice's maxim to be only as precise as needed: round numbers allow speakers to remain informative without being overly exact (Grice, 1975; Gibbs Jr. and Bryant, 2008; Van Der Henst and Sperber, 2004). Unnecessary precision, in contrast, may be perceived as pedantic or socially inappropriate (Beltrama et al., 2023; Cotterill, 2007; Lin, 2013; McCarthy and Carter, 2006). Speakers also use round numbers when the exact value is unknown to signal reduced commitment (Krifka, 2007; Lasersohn, 1999). This suggests that even when uncertainty is expressed numerically, the communicative use of round numbers introduces a layer of imprecision that somewhat parallels verbal uncertainty expressions. Numerical expressions are therefore not the silver bullet for uncertainty communication: Especially in machines they can imply precision when it is not warranted, and might be misinterpreted by listeners with low numeracy skills (Zikmund-Fisher et al., 2007; Galesic and Garcia-Retamero, 2010).

Section Findings

- Uncertainty communication is not just influenced by the likelihood of an event, but also social hierarchies, communicative preferences, and other speaker-listener dynamics.
- Numerical expressions can imply precision but are often rounded by speakers; verbal expressions are imprecise, which can be an advantage for topics for which precise probabilities are not available.

3 From Human to Human-Machine

Section 2 displayed an extensive picture of the nuances of human uncertainty communication. Contemporary language technology also mimics a conversational setting, making many of these considerations become blurry. In this section, we explore the current state of research and highlight some caveats to the uncertainty communication of LLMs, before also exploring some unique notions.

3.1 Background

Uncertainty is a multi-faceted concept with many different definitions and interpretations based on the context. We start from the definition of uncertainty by Baan et al. (2023), who define the concept as the lack or limit of knowledge of an agent interacting with the (or a) world. This definition lends itself well to the case of language models interacting with human users (or other language models), even if these interactions occur purely in the form of generating language. In machine learning, the most prominent paradigms

[2] It should be noted however that standardized usages of verbal expressions exist in these fields, see e.g. European Commission (2009); UK Ministry of Defence (2023).

to *quantify* this uncertainty draw from frequentist and Bayesian statistics. In the frequentist view, probabilities correspond to the relative frequency of an event, under repetitions of an experiment (Willink and White, 2012). Uncertainty can then be quantified through *confidence*,[3] where some score $\hat{p} \in [0, 1]$ is obtained (corresponding to the model's probability of the most likely outcome). The quality of an uncertainty estimate is assessed through *calibration* (Guo et al., 2017). Given a prediction $\hat{y}$, a model is calibrated if $p(y = \hat{y} \mid \hat{p}) = \hat{p}$, i.e. if we observe $\hat{p} = 0.45$, then $\hat{y}$ should be correct 45 out of 100 times on average under repetitions of the experiment. In contrast, the Bayesian perspective views probabilities as degrees of belief in the likelihood of an event (Good, 1971), and uncertainty is expressed in posterior distributions over predicted quantities. Bayesian methods also decompose uncertainty into *data* (aleatoric) and epistemic (model) uncertainty (Shafer, 1976, 1978; Der Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2021), where data uncertainty refers to unresolvable ambiguity, randomness or noise. In contrast, model uncertainty concerns the lack of knowledge about the best model, which can usually be corrected by increasing the amount of training data.[4]

Uncertainty in Modeling Language. NLP models are also faced with uncertainty contained in human language. Language is often fundamentally underspecified (Pustejovsky, 2017), context-dependent, imprecise and ambiguous (Kennedy, 2011), creating uncertainty in the input data and modeling: For instance, saying *she didn't make it* does not specify whether she didn't manage to complete something (like a race) or didn't create something (like a homemade vase), *I saw the person with the telescope* is ambiguous about which person possesses the instrument, and *glass* can refer to container or content depending on the context (*I drank the glass* vs. *I filled up the glass*). Compared to pure quantification of uncertainty, verbalized uncertainty presents a completely new paradigm: It can serve two different roles, namely the communication of uncertainty alone (when it was estimated through a separate method), or the joint quantification & communication (when the model is prompted for its uncertainty estimate in the form of a natural language response). In this process, two different notions arise: Uncertainty about the validity of a response, as well as about its verbalization. While Baan et al. (2023) discuss uncertainty in the output generation process, this distinction has, to the best of our knowledge, not been discussed for verbalized uncertainty. The methods discussed here focus on the former sense.

3.2 Verbalized Uncertainty in NLP

We collected publications to analyze the state of verbalized uncertainty research,[5] totaling 53 relevant publications since 2022, which marks, to the best of our knowledge, the earliest papers on the topic. Publications were selected by searching for keywords such as *verbal(ized) / linguistic uncertainty*, *uncertainty communication* etc. (see Table 1 in the appendix for details) for language models and then manually filtering by relevance, excluding survey papers. Notably, the topic has enjoyed increasing interest, spurred by the works of Lin et al. (2022); Mielke et al. (2022). Lin et al. ask for percentage scores and then finetune the model's calibration. Mielke et al. use an auxiliary predictor that predicts the confidence of a target model, which is mapped to a set of control tokens for different levels of uncertainty to guide generation. This lays the groundwork for the main ideas in the literature. The first prompts models to express uncertainty directly, while the second tries to generate confidence targets that are used for finetuning, such that a model automatically verbalizes its uncertainty. While many of the analyzed works involve applications, we focus here on methodological works.

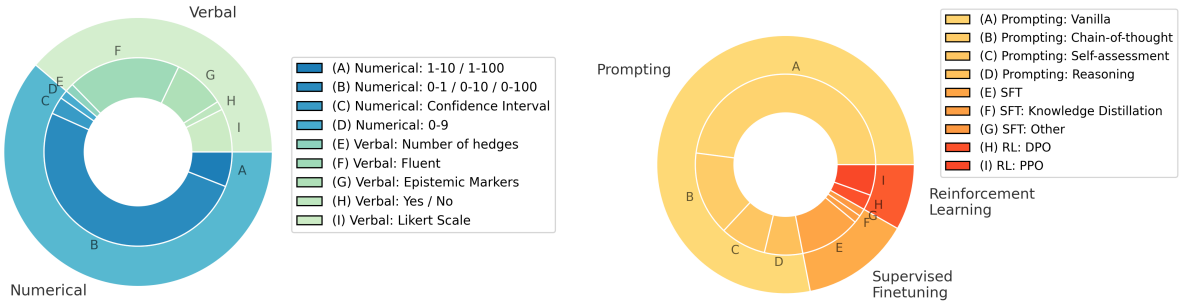
Prompting Techniques. When prompting, the goal is to elicit verbalization in the form of either verbal or numerical expressions, including either expressions that are usually ordered on a scale such as *almost no chance*, *probably* and *almost certain*, which are mapped back to values in $[0, 1]$ to evaluate calibration,[6] or mapping to percentages

[3] Another way is the set / interval size in conformal prediction (Angelopoulos and Bates, 2023; Campos et al., 2024).

[4] However, practical decompositions are often unreliable (Mucsányi et al., 2024), do not account for all sources of uncertainty (Gruber et al., 2023), and both uncertainties can appear in (ir-)reducible forms (Baan et al., 2023; Ulmer, 2024).

[5] We considered *ACL venues as well as TMLR, ICML, ICLR, NeurIPS, AISTATS, and ArXiv in our review. All relevant papers can be found with annotations in Appendix C.

[6] This mapping is defined either heuristically or using the human ratings obtained from studies such as Lichtenstein and Newman (1967); Beyth-Marom (1982); Wesson and Pulford (2009); Fagen-Ulmschneider (2019); Vogel et al. (2022).



(a) Registers of verbalization.

(b) Method of elicitation.

Figure 2: Analysis of surveyed papers. (a) Registers of verbalized uncertainty found in papers. (b) Ways in which uncertainty is elicited, i.e. either through prompting or by finetuning the model.

or scales (e.g. 1 to 10). In this vein, works have explored the robustness of verbalized uncertainty to prompt wording (Yang et al., 2024a) and language (Krause et al., 2023), different prompting techniques to improve calibration (Tian et al., 2023; Xiong et al., 2024; Liu et al., 2025), verbalizing the model's answer distribution (Yona et al., 2024; Wang et al., 2024a), and connections to other confidences measures (Ni et al., 2024b; Kumar et al., 2024) and applications in reasoning (Yoon et al., 2025; Podolak and Verma, 2025; Zeng et al., 2025) and long-form generations (Zhang et al., 2025).

Finetuning Techniques. For *supervised finetuning (SFT)*, the pipeline first consists of using a secondary technique to create confidence targets, for instance token probabilities (Duan et al., 2024), the probability of the true / false token after asking for the truthfulness of the answer (Kadavath et al., 2022), or self-consistency (Wang et al., 2023). By mapping them to epistemic markers (or leaving them as numerical expressions) which are added to the answer, the model can be finetuned directly. Next to Lin et al. (2022), this approach is also followed by Liu et al. (2024); Chaudhry et al. (2024); Hager et al. (2025); Band et al. (2024); Xu et al. (2024); Eikema et al. (2025); Tao et al. (2025b), only Li et al. (2025) finetune token probabilities directly. The other approach relies on *reinforcement learning*, with some techniques based on reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022). Whereas RLHF creates preference pairs from human feedback, here these are usually bootstrapped from existing data. LACIE (Stengel-Eskin et al., 2024) for instance uses a speaker and a listener model, where preferences pairs are constructed based on speaker uncertainty, listener acceptance and correctness, before finetuning the speaker model with direct preference optimization (DPO; Rafailov et al., 2023). A related listener model approach is used by Band et al. (2024), albeit employing explicit rewards and optimization via proximal policy optimization (PPO; Schulman et al., 2017). PPO is also applied by Tao et al. (2024) to reward answer quality and correspondence between confidence and response quality. Xu et al. (2024) combine SFT with PPO to optimize a reward for alignment between confidence and correctness. Leng et al. (2025) modify DPO and PPO by adapting the training objective for numerical uncertainty expressions and finetuning the reward model. Yang et al. (2024b) leverage reasoning chains and verbal confidences from a teacher model to finetune a student model.

Analysis of Current Research. We analyze the implementation of verbalized uncertainty next: Figure 2a shows that 61% of methods use numerical registers, typically ranging from 0–1 or 1–100. These expressions are easy to estimate with external methods, simple to extract from generations and straightforward to evaluate. Using verbal markers however requires decisions about their selection and their mapping to probabilities, and only a minority (19%) considers fluent verbalization, where epistemic markers are integrated into the answer in a natural way. Similarly, the majority of the works (78%) shown in Figure 2b relies on prompting to elicit verbalized uncertainty. Another large portion (14%) relies on supervised finetuning, with only a few recent works (Stengel-Eskin et al., 2024; Band et al., 2024; Xu et al., 2024; Leng et al., 2025) using reinforcement learning. This highlights the contrast between the richness and nuance of human uncertainty expres-

sions and current verbalized uncertainty in NLP: while finetuning-based approaches for fluency are more technically challenging, they can better emulate the complexity of human uncertainty communication than adding a numerical expression or verbal marker to a response. In addition, we show next how prompting methods, especially for numerical uncertainty, are prone to different biases.

3.3 Data Biases

Many works have shown that prompting produces uncalibrated verbalized confidences (Ulmer et al., 2024; Xiong et al., 2024; Wang et al., 2024a; Groot and Valdenegro-Toro, 2024; Tanneru et al., 2024; Kumar et al., 2024; Zhang et al., 2024; Sun et al., 2025; Tao et al., 2025a; Bakman et al., 2025), which we argue through frequency analysis to be caused by training data biases.

Bias in Pretraining Data. For percentage-valued confidence scores, Zhou et al. (2023) investigate the distribution of numbers in The Pile (Gao et al., 2020), a large pre-training corpus, and find a very skewed distribution of number terms. Outside of an NLP context, Woodin et al. (2024) see similar patterns for human number usage in the (much smaller) British national corpus (Consortium et al., 2007), finding that smaller and rounded numbers (to the next 5 or 10) appear more often. Since we don't have access to either pretraining data of LLMs or corresponding domain information, we instead source datasets through Huggingface's `datasets` platform (Lhoest et al., 2021), where we compare the distribution of number terms and epistemic markers.[7][8] Figure 3a shows an unequal distribution of number terms per domain: frequently appearing in quotes of people expressing absolute certainty, the number 100 appears exceedingly often in newspaper articles; the number 50 is much more frequently observed in books, and Reddit comments than for instance in ArXiv papers.[9] We conduct a similar analysis for markers by collecting them from prior work (Zhou et al., 2023, 2025), expanding the list through LLM-based chat assistants, verifying it through a linguistics expert, and running regex searches to count markers.[10] The same applies to epistemic markers in Figure 3b, which vary across domains. This suggests that LLMs will likely be biased in their verbalization not (only) by their uncertainty, but by their training data and context as well.

Bias in Instruction Finetuning. We also analyze the distribution of epistemic markers in four different instruction-finetuning datasets, namely Aya (Singh et al., 2024b), Tülu 3 (Lambert et al., 2024), OpenHermes-2.5 (Teknium, 2023), and UltraChat (Ding et al., 2023). We specifically check whether markers are used as strengtheners, weakeners, or are neutral / ambiguous in Figure 4.[11] Except for Tülu 3, we consistently find a majority of strengtheners. This has to be taken with a slight grain of salt however, as for all datasets except for Aya, the ambiguous block forms the majority.

Bias in RLHF. The last step of the contemporary training pipeline is RLHF, where models are taught human preferences about generations. Tian et al. (2023) noticed that while model log-probabilities are calibrated worse after RLHF, (numerical) verbalized uncertainties actually become more calibrated. However, Zhou et al. (2024) observed that models tend to not generate epistemic markers, and when they do, they show a preference for strengtheners, resulting in overconfident generations—a bias that is less pronounced in unaligned models. Importantly, they find higher predicted rewards for plain statements than for statements with strengtheners in reward models, and even negative rewards for statements with weakeners, suggesting that there is a negative bias towards uncertainty in the human preference data. Similarly, Leng et al. (2025) look at a reward model and a LLM trained with DPO, and find a bias towards high-certainty responses in both cases. All of this implies that RLHF further amplifies biases in verbalization that lead to overconfidence.

3.4 Importance of Context

Section 3.3 showed that the training data infuses LLMs with biases regarding the uncertainty verbalization. But does this translate to a sensitivity to conversational context, similar to humans?

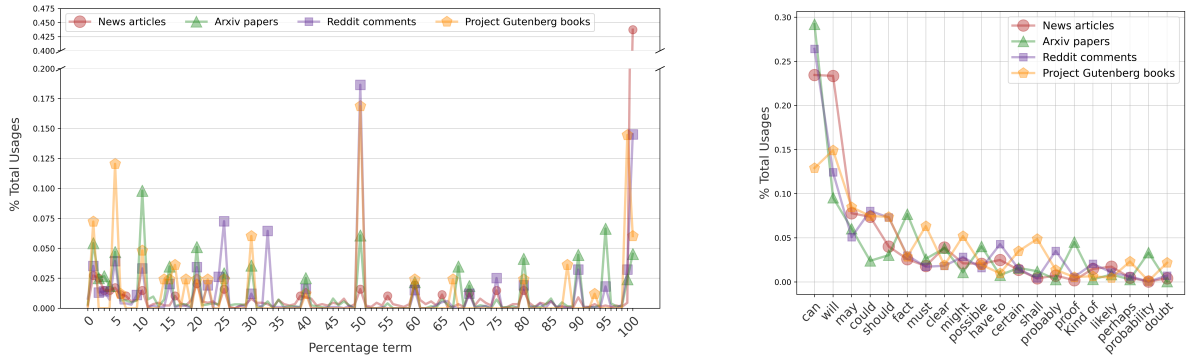
[7] An overview over datasets is given in Table 2, and markers of interest are listed in Table 5 in the appendix.

[8] The term "domain" is loosely defined in NLP and roughly refers to the genre or category of text. We refer to Barrett et al. (2024) for an exploration of this question.

[9] We also observed numerical terms above 100 (e.g. *I am 1000 % certain*), but omitted them due to their relative rarity.

[10] We lowercase markers and target strings and subtract cases in which another marker is a substring of the current markers (e.g."im*possible*"), as well as negated forms from a manually curated list available in our open-source repository.

[11] The equivalent of Figure 3b for instruction finetuning data is given in Figure 7 in the appendix.



(a) Distribution of number terms.

(b) Distribution of epistemic markers.

Figure 3: Visualization of data-inherent biases. We plot the distribution of percentage terms in different domains (left) and the distributions of the most frequent epistemic markers (right).

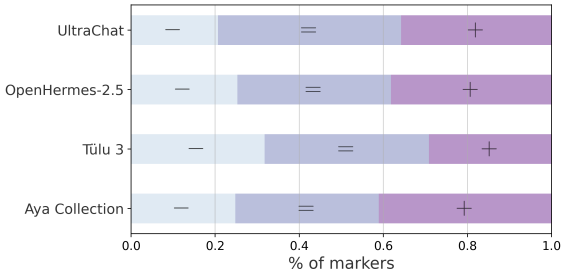


Figure 4: Marker distribution of $10k$ instances each in instruction-finetuning datasets by strengthener (+), weakener (–) or neutral / ambiguous (=).

We explore this by analyzing model responses and measuring linguistic and numerical calibration.

Speaker-Listener Relationship. In Section 2.2, we discussed how verbal uncertainty expression has functions beyond uncertainty, for instance expressing politeness or commitment. We show that these factors also influence LLMs. We use 250 questions from TriviaQA (Joshi et al., 2017) and pose them to six commercial LLMs (overview in Table 3 in the appendix), simulating conversational context through the system and question prompt in Table 2 and Figure 11. Specifically, we explore different relationships and social hierarchies, e.g.in the *Employee – Boss* setting, the LLM impersonates an employee who wants to impress their boss with the answer. In *Suspect – Interrogator*, the LLM acts as the suspect in an interrogation, trying to dispel suspicion. We show the results in Figure 5, where the register of verbalization (i.e.using verbal expressions, numerical expressions, both or none of them) varies largely between models, but also across scenarios. In the neutral setting, most models use no or only one register of verbalization. In contrast, given strict hierarchies (e.g. *Cititzen – Monarch*), models almost always express uncertainty, often combining both linguistic and numerical means.

Calibration. In addition, we assess the calibration of expressions across relationships. For numerical expressions, we resort to approximation of the expected calibration error (ECE; Naeini et al., 2015; Guo et al., 2017) through sorting of predictions into M equally-spaced bins by confidence:

$$\mathrm{ECE} \approx \sum_{m=1}^{M} \frac{|\mathcal{B}_m|}{N} \Big|\mathrm{acc}(\mathcal{B}_m) - \mathrm{conf}(\mathcal{B}_m)\Big|, \quad (1)$$

where $\mathrm{acc}(\cdot)$ refers to the accuracy of predictions in bin $\mathcal{B}_m$ and $\mathrm{conf}(\cdot)$ to average confidence of predictions in the same bin. For epistemic markers, we measure their appropriateness through colligation with Yule's Y (Yule, 1912):

$$Y = \frac{\sqrt{\#(-,\text{✗})\#(-,\text{✓})} - \sqrt{\#(+,\text{✗})\#(+,\text{✓})}}{\sqrt{\#(-,\text{✗})\#(-,\text{✓})} + \sqrt{\#(+,\text{✗})\#(+,\text{✓})}}, \quad (2)$$

where $\#(\cdot,\cdot)$ denotes the count of the presence of a strengthener (+) or weakener (–) with a correct (✓) or incorrect (✗) response. A Yule's Y of 1 indicates that strengtheners are only used for correct answers and weakeners for incorrect ones, and -1 stands for a perfectly anti-correlated use. Figure 6 illustrates that the numerical and linguistic calibration of models varies drastically across contexts, despite identical questions.[12]

[12]We check correctness using LLM-as-a-judge together with other heuristics, see prompt Figure 12 in the appendix.

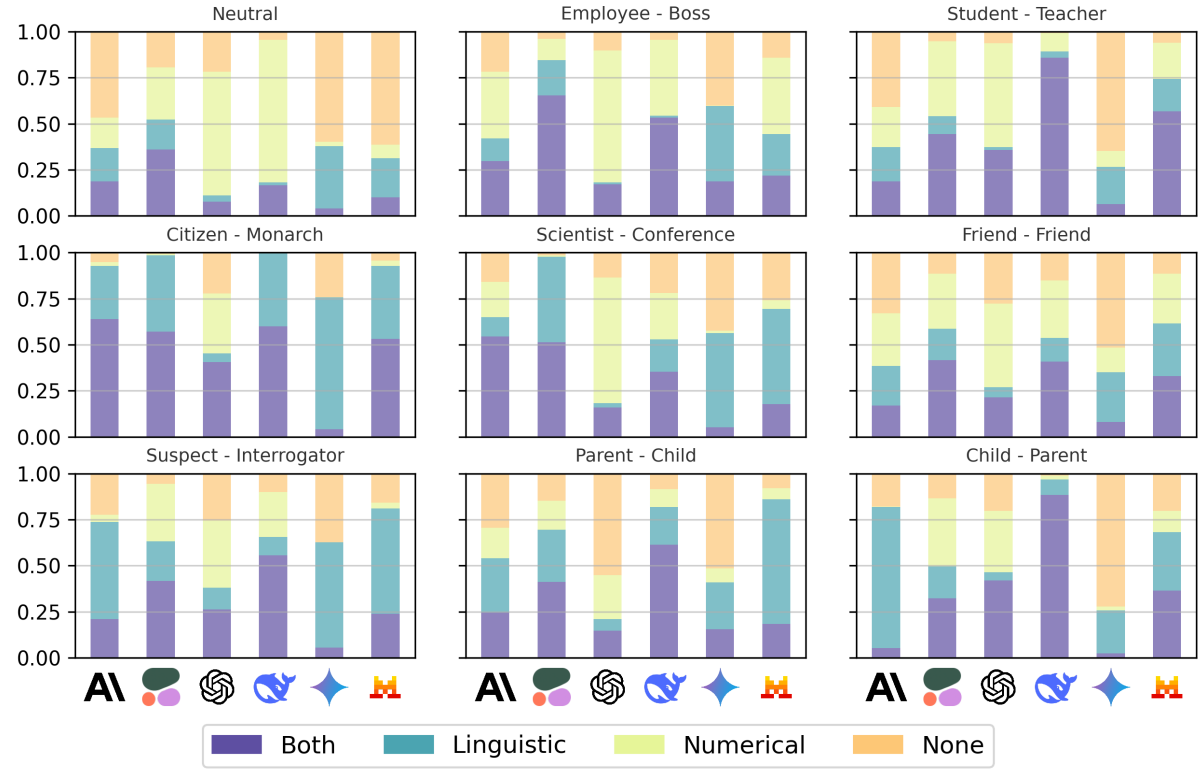
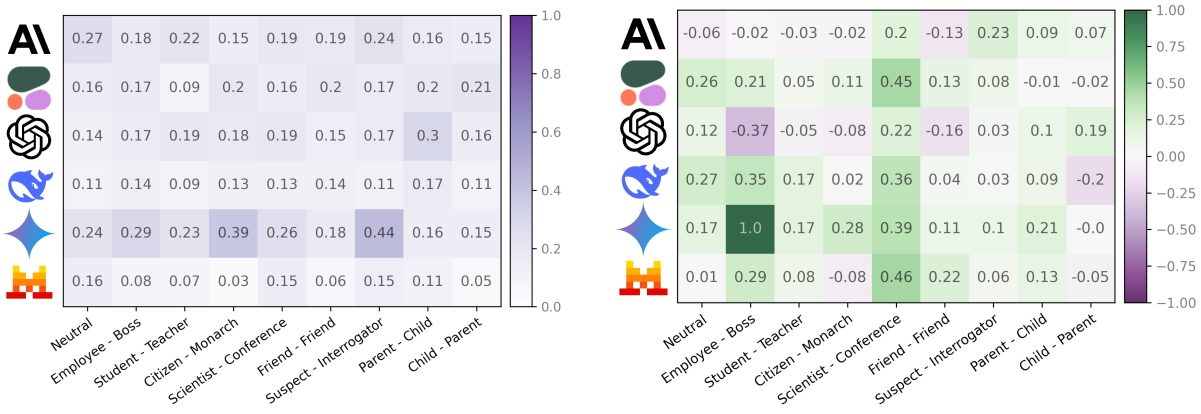


Figure 5: Verbalization registers over conversational contexts, using Claude 3.5 Haiku (A\), Command R+ ( ), GPT-4o mini ( ), DeepSeek Chat ( ), Gemini 2.0 flash ( ), and Mistral Medium ( ).



(a) Calibration of numerical confidence.

(b) Colligation of linguistic confidence.

Figure 6: Calibration of linguistic (measured by Yule's Y; Yule, 1912) and numerical expressions (measured by expected calibration error; Naeini et al., 2015) for different conversational contexts.

Domain. We demonstrate that domains bias uncertainty verbalization. We ask the same LLMs 100 questions each from subject areas in MMLU (Hendrycks et al., 2021). Consistent with the findings above, the register in Figure 9 also varies greatly, both across subject areas and models. We attribute these effects to a) data biases (see Section 3.3) and b) to difference in the training data *distribution* of LLMs (which are not publicly available). As before, we observe drastic differences in numerical / linguistic calibration for the same model between different subjects (Figure 8). In addition, we use Kendall's τ (Kendall, 1938) to quantify whether the most frequently occurring epistemic markers in Section 3.3 also appear as often across responses. In Figure 10a in the appendix, we show that there indeed exists a moderately strong positive correlation ($\tau \in [0.33, 0.53]$)

among all pairings. This positive relationship remains also when focusing on a single LLM (Figure 10a), and correspondingly for the use of confidence values (Figures 10b and 10d), although in slightly weaker form ($\tau \in [0.06, 0.47]$). It is important to note that we cannot say with certainty whether the tested text corpora were part of the training data. Nevertheless, there persists a positive relationship between the occurrence of epistemic markers in corpora and in responses.

3.5 Human-Machine Pragmatics

Language generation by LLMs, however, does not occur in a vacuum, but in interaction with human users.[13] Therefore, additional pragmatic concerns arise in this novel setting, which we study here.

Anthropomorphization. As chatbot interfaces mimic human interaction, we might be tempted to ascribe human traits, emotions, and agency to the models. This *anthropomorphization* has been observed in users, reporters, and even AI researchers themselves (Abercrombie et al., 2021; Shardlow and Przybyła, 2024). The literature has pointed out effects of this (Schneider et al., 2018; Troshani et al., 2021; Colombatto et al., 2025), specifically concerning trust-building (Natarajan and Gombolay, 2020; Cohn et al., 2024; Basoah et al., 2025). But while humans have been shown to use relatively consistent vocabularies of epistemic markers (Clark, 1990; Dhami and Mandel, 2022), this is likely not true for LLMs, as evidenced by Section 3.3, inhibiting trust. Taking it further, the tendency of LLMs to present false information with unwarranted confidence is similar to uncooperative human speakers, but without any underlying intent to deceive. The problem is that the way of dealing with this is epistemic vigilance (see Section 2), which is much less effective here due to the black-box nature of LLMs. As user increasingly treat LLMs as sources of information, spotting falsehoods either requires deep background knowledge (Gude, 2025), or critically engaging with the output through external resources. However, research suggests that although users are aware of hallucinations, they only scrutinize LLMs in high-stakes situations, where errors have greater consequences (Lee et al., 2025b).

[13] Or with other LLMs (Park et al., 2023; Stengel-Eskin et al., 2024; Chen et al., 2025; Yoffe et al., 2024, inter alia).

Sycophancy. LLMs, trained to be helpful and supportive, can also behave in an overly agreeable manner, reinforcing unhealthy behaviors or being misleading. This effect has already been studied (Cotra, 2021; Perez et al., 2023; Sharma et al., 2024; Cheng et al., 2025), including for uncertainty (Sicilia et al., 2025a) and user trust (Carro, 2024). Zhou et al. (2024); Leng et al. (2025) demonstrated that RLHF biases responses towards confidence, likely since they appear more helpful to human annotators. Thus, models might downplay their uncertainty to increase perceived helpfulness (Dahlgren Lindström et al., 2025). Steyvers et al. (2025) have demonstrated that humans struggle to identify wrong answers, which can damage trust and lead to worse outcomes (Dhuliawala et al., 2023; Zhou et al., 2024). This artifact parallels human confirmation bias, where speakers select arguments that support their own views (Sperber, 2001; Mercier and Sperber, 2009). While LLMs are not driven by social motives, they are optimized to align with user preferences. Consequently, they may replicate this bias by producing agreeable content that reinforces existing beliefs. Indeed, a study by Sicilia et al. (2025a) shows the uncertainty of the user actually influencing the LLM's uncertainty. This behavior introduces a distinct epistemic risk: the cooperative and non-confrontational tone of chatbots may be perceived as trustworthy, lowering epistemic vigilance and critical engagement.

Section Findings

- Verbalized uncertainty can be used to either communicate estimated uncertainty, or for joint quantification & expression.
- Current research mostly considers numerical expressions, with works on verbal expression reducing the problem to single markers. This is done mostly through prompting, with only some recent works finetuning models via SFT or RL.
- Training (data) biases manifest in verbalization in various (and unexpected) ways between different domains and conversational contexts.
- Anthropomorphization and sycophancy lead users to expect human-like uncertainty communication, which can lead to unexpected and potentially harmful effects.

4 Anthropomimetic Uncertainty

Finally, our core thesis is that language models should better adhere to human uncertainty communication, which we term *anthropomimetic uncertainty*.[14] We believe that verbalized uncertainty creates an exciting but incomplete paradigm to do so, and discuss some prerequisites towards anthropomimetic uncertainty in the following. Importantly, we do *not* advocate for measures that increase anthropomorphization in language models. Rather, we accept that through their conversational use, some degree of anthropomorphization is probably inevitable, and that risks should be mitigated through anthropomimetic uncertainty.

Consistency. Section 3.4 revealed that models are heavily biased by conversational context and subject when it comes to their verbalized uncertainty and calibration. Thus, anthropomimetic uncertainty should imply a consistent way to use uncertainty expressions across conversations. Uncertainty communication differs between people (Beyth-Marom, 1982; Wesson and Pulford, 2009; Vogel et al., 2022; Dhami and Mandel, 2022), however individuals usually have a consistent way to express different level of confidence (Clark, 1990; Dhami and Mandel, 2022). This implies that a model should use similar expressions when referring to similarly likely events, which Liu et al. illustrate is not the case currently. This could potentially be addressed through reinforcement learning, as a related setting has already been explored for numerical confidence expressions (Xu et al., 2024; Leng et al., 2025).

Personalization. People will act differently upon receiving the same hedged response, since tolerance for risk differs e.g. based on demographic factors or personality traits (Mishra and Lalumière, 2011; Brooks and Williams, 2021; Pavlíček et al., 2021). Therefore, uncertainty communication should be adapted to specific users and problems (Chakraborti et al., 2025), for instance by continuously incorporating feedback from past actions, and by taking into account the user's prior knowledge. This idea is complementary to the *interactive learning* described by Kirchhof et al. (2025), where a model learns to ask clarification questions in order to reduce uncertainty. Furthermore, recent works have already started to explore personalization of LLMs (Magister et al., 2025; Samuel et al., 2025; Qiu et al., 2025).

Choice of Register. The current choice of register is dominated by numerical expressions (see Figure 2a), since calibrating and evaluating them is easier. However, fluent expressions are not necessarily better, due to a preference to communicate uncertainty in verbal, but receive it in numerical form (Wallsten et al., 1993; Dhami and Mandel, 2022). Additionally, numerical expressions distinguish between external (*It is 60 % certain*) vs. internal expressions (*I am 60 % certain*). Løhre and Teigen (2016) found that while external expressions seem more trustworthy, internal expressions are seen as more agreeable and engaging. However, while numerical expression appear more precise on the surface, they might be actually be vague, especially for probabilities which cannot be computed reliably (Teigen, 2023)—users may perceive them as more authoritative or grounded than they are, given that LLMs do not know the *actual* likelihood of an event (Section 2.3).[15] This suggests that while numerical expressions may technically be more feasible, their seeming objectivity can be misleading; conversely, the imprecision of verbal expressions may be a deliberate feature, indicating that the expressed uncertainty is itself imprecise. Thus, fluent verbal expressions might be preferable where absolute precision is not paramount, but methodological challenges on how to learn and evaluate them remain, with first works in Yona et al. (2024); Stengel-Eskin et al. (2024); Xu et al. (2024). Therefore, the choice of register remains a challenge, intertwined with the questions of consistency and personalization. Research could explore when a switch of register is most expedient, either to avoid harm or to increase human-AI collaboration. In addition, it is an open question whether dialogue agents should communicate exact confidence values, or resort to rounding as described for humans in Section 2.3.

Multilinguality. A shortcoming of the current research is its focus on English, with the exceptions of Krause et al. (2023); Rathi et al. (2025). Given that LLMs' knowledge is inconsistent across languages (Kassner et al., 2021; Fierro

[14] From *anthropo-* (human) and *-mimetic* (imitative); and used in robotics (Holland and Knight, 2006; Diamond et al., 2012).

[15] Windschitl and Wells (1996); Liu et al. (2020) also discuss that linguistic expressions are processed more intuitively, while numerical expressions elicit logical reasoning.

et al., 2025), so is calibration (Krause et al., 2023; Rathi et al., 2025). More importantly, language is likely to bias expressions of uncertainty similarly to the other factors in Sections 3.3 and 3.4, and ways to express uncertainty vary greatly across languages (Vold, 2006; Ruskan and Šolienė, 2017; Mulder et al., 2022; Riccioni et al., 2022). At the time of writing, resources for uncertainty markers / expressions with confidence information remain scarce even for English (excluding Tao et al., 2025b) and practically non-existent for other languages. Therefore, future research should build resources for verbalized uncertainty across languages and study how to obtain and maintain calibrated expressions in a multilingual setting.

Explaining Uncertainty. To increase trust, models should also explain the reasons why they are uncertain. While a number of works have started exploring this direction (Chen and Ji, 2022; Xu et al., 2024; Thuy and Benoit, 2024; Watson et al., 2023; Steyvers et al., 2025), there remain significant challenges with explaining model predictions (Madsen et al., 2024b), which are only exacerbated by the absence of a gold standard method for uncertainty quantification in LLMs. Ideally, we would like to instill self-awareness into models when it comes to their reasons for uncertainty, or alternative obtain explanations from an external module and have a model verbalize it alongside the actual response. However, explanation for the model's original prediction already remain challenging (Madsen et al., 2024a; Agarwal et al., 2024; Singh et al., 2024a), with explanation of uncertainty adding additional complexity.

Beyond Classical Calibration. Previous research has often focused on probabilistic calibration, aiming to align confidence with correctness. But since LLMs often generate answers stochastically, Yona et al. (2024); Wang et al. (2024a) have also investigated calibration with respect to their answer distribution. For verbal expressions, the mapping to probabilities is non-trivial, and so the question remains whether *exact* calibration is even necessary. Nevertheless, some works have explored the calibration of a listener acting upon a response (Stengel-Eskin et al., 2024; Yona et al., 2024; Steyvers et al., 2025). Future work could also consider the calibration towards outcomes induced by a listener, and develop notions of linguistic calibration that don't rely on probabilities.

Trust beyond Uncertainty. While works have explored the relationship between uncertainty and trust for humans and AI (Kapoor et al., 2024; Dhuliawala et al., 2023; Kim et al., 2024; Zhou et al., 2024), trust is not based on uncertainty alone—Zhou et al. (2025) e.g. found that humans are also influenced by subject matter, prior interactions, warmth, and humanlikeness. Future research should therefore aim to understand uncertainty in lieu with the aforementioned factors by Zhou et al. (2025), but also other potential reasons such as culture (Hershcovich et al., 2022) or socioeconomic status (Bassignana et al., 2025).

Section Findings

- Since some degree of anthropomorphization seems hard to avoid with LLMs, they should communicate their uncertainty similarly to humans (*anthropomimetic uncertainty*).
- This includes consistent usage of epistemic markers and verbalization registers, personalization to the user, and the ability to explain the reasons behind uncertainty.
- Research should also consider languages beyond English, explore new notions of calibration, and other factors beyond uncertainty that influence trust.

5 Conclusion

In our work, we gave a holistic account of the verbal communication of uncertainty. By giving detailed insights into verbalized uncertainty between humans, we were able to point out important gaps and lacks of nuance in the current NLP literature. Furthermore, we demonstrated empirically that the current verbalized uncertainty is heavily influenced by data and contextual biases. Based on these perspectives, we outline actionable research directions towards anthropomimetic uncertainty, behaviors in LLMs that build trust by emulating human uncertainty communication.

Acknowledgments

This work is supported by the Dutch National Science Foundation (NWO Vici VI.C.212.053). Experiments in this work were further supported through the Cohere for AI Research Grant. We also thank the anonymous reviewers of this article, who, with their constructive feedback, have

helped improve its quality, as well as the action editor, Kristina Toutanova, for her efforts.

References

Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What Are Your Pronouns? Gender And Anthropomorphism In The Design And Perception Of Conversational Assistants. *arXiv preprint arXiv:2106.02578*.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness Vs. Plausibility: On The (Un)Reliability Of Explanations From Large Language Models. *arXiv preprint arXiv:2402.04614*.

Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal Prediction: A Gentle Introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty In Natural Language Generation: From Theory To Applications. *arXiv preprint arXiv:2307.15703*.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Sungmin Kang, Tuo Zhang, Baturalp Buyukates, Salman Avestimehr, and Sai Praneeth Karimireddy. 2025. Reconsidering LLM Uncertainty Estimation Methods in the Wild. In *Proceedings Of The 63rd Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 29531–29556. Association for Computational Linguistics.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic Calibration Of Long-Form Generations. In *Forty-First International Conference On Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Alan Barnes. 2016. Making Intelligence Analysis More Intelligent: Using Numeric Probabilities. *Intell. Natl. Secur.*, (31):327—-344.

Maria Barrett, Max Müller-Eberstein, Elisa Bassignana, Amalie Brogaard Pauli, Mike Zhang, and Rob van der Goot. 2024. Can Humans Identify Domains? In *Proceedings Of The 2024 Joint International Conference On Computational Linguistics, Language Resources And Evaluation, LREC/Coling 2024, 20-25 May, 2024, Torino, Italy*, pages 2745–2765. ELRA and ICCL.

Jeffrey Basoah, Daniel Chechelnitsky, Tao Long, Katharina Reinecke, Chrysoula Zerva, Kaitlyn Zhou, Mark Díaz, and Maarten Sap. 2025. Not Like Us, Hunty: Measuring Perceptions And Behavioral Effects Of Minoritized Anthropomorphic Cues In LLMs. In *Proceedings Of The 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2025, Athens, Greece, June 23-26, 2025*, pages 710–745. ACM.

Elisa Bassignana, Amanda Cercas Curry, and Dirk Hovy. 2025. The AI Gap: How Socioeconomic Status Affects Language Technology Interactions. In *Proceedings Of The 63rd Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18647–18664. Association for Computational Linguistics.

Andrea Beltrama, Stephanie Solt, and Heather Burnett. 2023. Context, Precision, And Social Perception: A Sociopragmatic Study. *Language in Society*, 52(5):805–835.

Brian Bergstrom, Bianca Moehlmann, and Pascal Boyer. 2006. Extending The Testimony Problem: Evaluating The Truth, Scope, And Source Of Cultural Information. *Child development*, 3(77):531–538.

Ruth Beyth-Marom. 1982. How Probable Is Probable? A Numerical Translation Of Verbal Probability Expressions. *Journal of Forecasting*, 1(3):257–269.

Jean-François Bonnefon and Gaëlle Villejoubert. 2006. Tactful Or Doubtful?: Expectations Of

Politeness Explain The Severity Bias In The Interpretation Of Probability Phrases. *Psychological Science*, 17(9):747–751.

Mirko Borszukovszki, Ivo Pascal De Jong, and Matias Valdenegro-Toro. 2025. Know What You Do Not Know: Verbalized Uncertainty Estimation Robustness On Corrupted Images In Vision-Language Models. *arXiv preprint arXiv:2504.03440*.

Chris Brooks and Louis Williams. 2021. The Impact Of Personality Traits On Attitude To Financial Risk. *Research in International Business and Finance*, 58:101501.

Penelope Brown and Stephen Levinson. 1978. Universals In Language Usage: Politeness Phenomena. In E. Goody, editor, *Questions And Politeness. Strategies In Social Interactions*, pages 56–311. Cambridge University Press.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A.T. Figueiredo, and André F.T. Martins. 2024. Conformal Prediction For Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

María Victoria Carro. 2024. Flattering To Deceive: The Impact Of Sycophantic Behavior On User Trust In Large Language Model. *arXiv preprint arXiv:2412.02802*.

Tapabrata Chakraborti, Christopher R.S. Banerji, Ariane Marandon, Vicky Hellon, Robin Mitra, Brieuc Lehmann, Leandra Bräuninger, Sarah McGough, Cagatay Turkay, Alejandro F. Frangi, Ginestra Bianconi, Weizi Li, Owen Rackham, Deepak Parashar, Chris Harbron, and Ben MacArthur. 2025. Personalized Uncertainty Quantification In Artificial Intelligence. *Nature Machine Intelligence*, 7(4):522–530.

Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. 2024. Finetuning Language Models To Emit Linguistic Expressions Of Uncertainty. *arXiv preprint arXiv:2409.12180*.

Hanjie Chen and Yangfeng Ji. 2022. Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10463–10472. AAAI Press.

Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2025. Magicore: Multi-Agent, Iterative, Coarse-To-Fine Refinement For Reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 32663–32686. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social Sycophancy: A Broader Understanding Of LLM Sycophancy. *arXiv preprint arXiv:2505.13995*.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning From Human Preferences. *Advances in neural information processing systems*, 30.

Dominic A. Clark. 1990. Verbal Uncertainty Expressions: A Critical Review Of Two Decades Of Research. *Current Psychology*, 9(3):203–235.

Cohere Labs. 2024. C4AI-Command-R-Plus-08-2024 . https://huggingface.co/CohereLabs/c4ai-command-r-plus-08-2024.

Michelle Cohn, Mahima Pushkarna, Gbolahan O. Olanubi, Joseph M. Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing Anthropomorphism: Examining The Role Of Anthropomorphic Cues On Trust In Large Language Models. In *Extended Abstracts Of The Chi Conference On Human Factors In Computing Systems*, pages 1–15.

Clara Colombatto, Jonathan Birch, and Stephen Fleming. 2025. The Influence Of Mental State Attributions On Trust In Large Language Models. *Communications Psychology*, 3.

BNC Consortium et al. 2007. British National Corpus. *Oxford Text Archive Core Collection*.

Ajeya Cotra. 2021. Why AI Alignment Could Be Hard With Modern Deep Learning. *Cold Takes*.

Janet Cotterill. 2007. 'I Think He Was Kind Of Shouting Or Something': Uses And Abuses Of Vagueness In The British Courtroom. In *Vague Language Explor Ed*, pages 97–114. Springer.

Chris Cummins. 2015. *Constraints On Numerical Expressions*, volume 5. Oxford University Press.

Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2025. Helpful, Harmless, Honest? Sociotechnical Limits Of AI Alignment And Safety Through Reinforcement Learning From Human Feedback. *Ethics and Information Technology*, 27(2):1–13.

DeepSeek-AI. 2025. Deepseek-R1: Incentivizing Reasoning Capability In LLMs Via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory Or Epistemic? Does It Matter? *Structural safety*, 31(2):105–112.

Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A Taxonomy Of Linguistic Expressions That Contribute To Anthropomorphism Of Language Technologies. In *Proceedings Of The 2025 Chi Conference on Human Factors in Computing Systems, CHI 2025, Yokohama, Japan, 26 April 2025- 1 May 2025*, pages 430:1–430:18. ACM.

Mandeep K. Dhami and David R. Mandel. 2022. Communicating Uncertainty Using Words And Numbers. *Trends in Cognitive Sciences*, 26(6):514–526.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A Diachronic Perspective On User Trust In AI under Uncertainty. In *Proceedings Of The 2023 Conference On Empirical Methods In Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5567–5580. Association for Computational Linguistics.

Alan Diamond, Rob Knight, David Devereux, and Owen Holland. 2012. Anthropomimetic Robots: Concept, Construction And Modelling. *International Journal of Advanced Robotic Systems*, 9(5):209.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models By Scaling High-Quality Instructional Conversations. In *Proceedings Of The 2023 Conference On Empirical Methods In Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051. Association for Computational Linguistics.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a Personalized Judge? In *Findings Of The Association For Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10126–10141. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting Attention To Relevance: Towards The Uncertainty Estimation Of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Bryan Eikema, Evgenia Ilia, José G.C. de Souza, Chrysoula Zerva, and Wilker Aziz. 2025. Teaching Language Models To Faithfully Express Their Uncertainty. *arXiv preprint arXiv:2510.12587*.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. GPTs Are GPTs: Labor Market Impact Potential Of LLMs. *Science*, 384(6702):1306–1308.

Ido Erev and Brent L. Cohen. 1990. Verbal Versus Numerical Probabilities: Efficiency, Biases, And The Preference Paradox. *Organizational behavior and human decision processes*, 45(1):1–18.

European Commission. 2009. Guideline On The Readability Of The Labelling And Package Leaflet Of Medicinal Products For Human Use. https://health.ec.europa.eu/document/download/d8612682-ad17-40e3-8130-23395ec80380_en. Accessed: 11/07/2025.

Wader Fagen-Ulmschneider. 2019. Perception Of Probability Words. https://waf.cs.illinois.edu/visualizations/Perception-of-Probability-Words/. Accessed: 02/05/2025.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How Do Multilingual Language Models Remember Facts. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 16052–16106. Association for Computational Linguistics.

Bruce Fraser. 1975. Hedged Performatives. In *Speech Acts*, pages 187–210. Brill.

Mirta Galesic and Rocio Garcia-Retamero. 2010. Statistical Numeracy For Health: A Cross-Cultural Comparison With Probabilistic National Samples. *Archives of internal medicine*, 170(5):462–468.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset Of Diverse Text For Language Modeling. *arXiv preprint arXiv:2101.00027*.

Raymond W. Gibbs Jr. and Gregory A. Bryant. 2008. Striving For Optimal Relevance When Answering Questions. *Cognition*, 106(1):345–369.

Irving John Good. 1971. 46656 Varieties Of Bayesians. *American Statistician*, 25(5):62.

Herbert Paul Grice. 1975. Logic And Conversation. *Syntax and semantics*, 3:43–58.

Tobias Groot and Matias Valdenegro-Toro. 2024. Overconfidence Is Key: Verbalized Uncertainty Evaluation In Large Language And Vision-Language Models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. 2023. Sources Of Uncertainty In Machine Learning–A Statisticians' View. *arXiv preprint arXiv:2305.16703*.

Vinayaka Gude. 2025. Factors Influencing ChatGPT Adoption For Product Research And Information Retrieval. *Journal of Computer Information Systems*, 65(2):222–231.

Nuno Miguel Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations In Large Multilingual Translation Models. *Trans. Assoc. Comput. Linguistics*, 11:1500–1517.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration Of Modern Neural Networks. In *International Conference On Machine Learning*, pages 1321–1330. PMLR.

Sophia Hager, David Mueller, Kevin Duh, and Nicholas Andrews. 2025. Uncertainty Distillation: Teaching Language Models To Express Semantic Confidence. *arXiv preprint arXiv:2503.14749*.

Adam J. L. Harris and Adam Corner. 2011. Communicating Environmental Risks: Clarifying The Severity Effect In Interpretations Of Verbal Probability Expressions. *Journal of experimental psychology. Learning, memory, and cognition*, 37 6:1571–8.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. 2025. Do LLMs Estimate Uncertainty Well In Instruction-Following? In *The Thirteenth International Conference On Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Daniel Hershcovich, Stella Frank, Heather C. Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges And Strategies In Cross-Cultural NLP. In *Proceedings Of The 60th Annual Meeting Of The Association For Computational Linguistics*

(Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 6997–7013. Association for Computational Linguistics.

Arthur Hoarau, Sébastien Destercke, Yusuf Sale, Paul Hofman, and Eyke Hüllermeier. 2024. Measures Of Uncertainty: A Quantitative Analysis. https://arthur-hoarau.fr/files/Measures_of_Uncertainty.pdf.

Owen Holland and Rob Knight. 2006. The Anthropomimetic Principle. In *Proceedings Of The Aisb06 Symposium On Biologically Inspired Robotics*, pages 1–8. Citeseer.

Janet Holmes. 1982. Expressing Doubt And Certainty In English. *RELC journal*, 13(2):9–28.

Thomas Holtgraves. 2010. Social Psychology And Language: Words, Utterances, And Conversations. In *Handbook Of Social Psychology*. John Wiley & Sons, Inc.

Thomas Holtgraves. 2014. Interpreting Uncertainty Terms. *Journal of personality and social psychology*, 107(2):219.

Thomas Holtgraves and Audrey Perdew. 2016. Politeness And The Communication Of Uncertainty. *Cognition*, 154:1–10.

Richard Holton. 1994. Deciding To Trust, Coming To Believe. *Australasian Journal of Philosophy*, 72(1):63–76.

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.

Anders Humlum and Emilie Vestergaard. 2024. The Adoption Of ChatGPT. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=4807516.

Ken Hyland. 1999. *Disciplinary Discourses: Writer Stance In Research Articles*.

Ken Hyland. 2005. Stance And Engagement: A Model Of Interaction In Academic Discourse. *Discourse Studies*, 7(2):173–192.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust In Artificial Intelligence: Prerequisites, Causes And Goals Of Human Trust In AI. In *Proceedings Of The 2021 ACM Conference On Fairness, Accountability, And Transparency*, pages 624–635.

Carel J.M. Jansen and Mathijs M.W. Pollmann. 2001. On Round Numbers: Pragmatic Aspects Of Numerical Expressions. *Journal of quantitative linguistics*, 8(3):187–201.

Klaudia Jaźwińska and Aisvarya Chandrasekar. 2025. AI Search Has A Citation Problem. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php. Accessed: 20/03/2025.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. 2025. Calibrating Verbal Uncertainty As A Linear Feature To Reduce Hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 3769–3793. Association for Computational Linguistics.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Marie Juanchich and Miroslav Sirota. 2015. A Direct And Comprehensive Test Of Two Postulates Of Politeness Theory Applied To Uncertainty Communication. *Judgment and Decision Making*, 10(3):232–240.

Marie Juanchich and Miroslav Sirota. 2020. Most Family Physicians Report Communicating The Risks Of Adverse Drug Reactions In Words (Vs. Numbers). *Appl. Cogn. Psychol.*, (34):526—-534.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez,

Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large Language Models Must Be Taught To Know What They Don't Know. In *Advances In Neural Information Processing Systems 38: Annual Conference On Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, Bc, Canada, December 10 - 15, 2024*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. In *Proceedings Of The 16th Conference Of The European Chapter Of The Association For Computational Linguistics: Main Volume, Eacl 2021, Online, April 19 - 23, 2021*, pages 3250–3258. Association for Computational Linguistics.

Maurice G. Kendall. 1938. A New Measure Of Rank Correlation. *Biometrika*, 30(1-2):81–93.

Christopher Kennedy. 2011. Ambiguity And Vagueness: An Overview. *Semantics: An international handbook of natural language meaning*, 1:507–535.

Sunnie S.Y. Kim, Qing Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact Of Large Language Models' Uncertainty Expression On User Reliance And Trust. In *Proceedings Of The 2024 ACM Conference On Fairness, Accountability, And Transparency*, pages 822–835.

Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. 2025. Position: Uncertainty Quantification Needs Reassessment For Large Language Model Agents. In *Forty-Second International Conference On Machine Learning Position Paper Track*.

Lea Krause, Wondimagegnhue Tufa, Selene Báez Santamaría, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently Wrong: Exploring The Calibration And Expression Of (Un)Certainty Of Large Language Models In A Multilingual Setting. In *Proceedings Of The Workshop On Multimodal, Multilingual Natural Language Generation And Multilingual Webnlg Challenge (MM-NLG 2023)*, pages 1–9.

Manfred Krifka. 2007. *Approximate Interpretation Of Number Words*. Humboldt-Universität zu Berlin, Philosophische Fakultät II.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence Under The Hood: An Investigation Into The Confidence-Probability Alignment In Large Language Models. In *Proceedings Of The 62nd Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 315–334. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Peter Lasersohn. 1999. Pragmatic Halos. *Language*, pages 522–551.

Daniel Lassiter. 2010. Gradable Epistemic Modals, Probability, And Scale Structure. In

Semantics And Linguistic Theory, pages 197–215.

Daniel Lassiter. 2017. *Graded Modality: Qualitative And Quantitative Perspectives*. Oxford University Press, Oxford.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025a. Are LLM-Judges Robust To Expressions Of Uncertainty? Investigating The Effect Of Epistemic Markers On LLM-Based Evaluation. In *Proceedings Of The 2025 Conference Of The Nations Of The Americas Chapter Of The Association For Computational Linguistics: Human Language Technologies, Naacl 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 8962–8984. Association for Computational Linguistics.

Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025b. The Impact Of Generative AI On Critical Thinking: Self-Reported Reductions In Cognitive Effort And Confidence Effects From A Survey Of Knowledge Workers. In *Chi 2025*.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2025. Taming Overconfidence In LLMs: Reward Calibration In RLHF. In *The Thirteenth International Conference On Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. 2025. Conftuner: Training Large Language Models To Express Their Confidence Verbally. *arXiv preprint arXiv:2508.18847*.

Sarah Lichtenstein and John Robert Newman. 1967. Empirical Scaling Of Common Verbal Phrases Associated With Numerical Probabilities. *Psychonomic science*, 9:563–564.

Chin-Yew Lin. 2004. Rouge: A Package For Automatic Evaluation Of Summaries. In *Text Summarization Branches Out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models To Express Their Uncertainty In Words. *Trans. Mach. Learn. Res.*, 2022.

Yen-Liang Lin. 2013. Vague Language And Interpersonal Communication: An Analysis Of Adolescent Intercultural Conversation. *International Journal of Society, Culture & Language*, 1(2):69–81.

Dawn Liu, Marie Juanchich, Miroslav Sirota, and Sheina Orbell. 2020. The Intuitive Use Of Contextual Information In Decisions Made With Verbal And Numerical Quantifiers. *Quarterly Journal of Experimental Psychology*, 73(4):481–494.

Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim G. J. Rudner, and Arman Cohan. 2025. Metafaith: Faithful Natural Language Uncertainty Expression In LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 29612–29656. Association for Computational Linguistics.

Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. Revisiting Epistemic Markers In Confidence Estimation: Can Markers Accurately Reflect Large Language Models' Uncertainty? In *Proceedings Of The 63rd Annual Meeting Of The Association For Computational Linguistics (Volume 2: Short Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 206–221. Association for Computational Linguistics.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek Wong, Lidia Chao, and Min Zhang. 2024. Can LLMs Learn Uncertainty On Their Own? Expressing Uncertainty Effectively In A Self-Training Manner. In *Proceedings Of The 2024 Conference On Empirical Methods In Natural Language Processing*, pages 21635–21645.

Erik Løhre and Karl Halvor Teigen. 2016. There Is A 60% Probability, But I Am 70% Certain: Communicative Consequences Of External And Internal Expressions Of Uncertainty. *Thinking & Reasoning*, 22(4):369–396.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024a. Are Self-Explanations From Large Language Models Faithful? In *Findings Of The Association For Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 295–337. Association for Computational Linguistics.

Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. 2024b. Interpretability Needs A New Paradigm. *arXiv preprint arXiv:2405.05386*.

Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje Ter Hoeve. 2025. On The Way To LLM Personalization: Learning To Remember User Conversations. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pages 61–77.

Michael McCarthy and Ronald Carter. 2006. *Explorations In Corpus Linguistics*. Cambridge University Press Cambridge.

Hugo Mercier and Dan Sperber. 2009. Intuitive And Reflective Inferences. In Jonathan Evans and Keith Frankish, editors, *In Two Minds: Dual Processes And Beyond*, pages 149–170.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Trans. Assoc. Comput. Linguistics*, 10:857–872.

Tim Miller. 2019. Explanation In Artificial Intelligence: Insights From The Social Sciences. *Artificial intelligence*, 267:1–38.

Sandeep Mishra and Martin L. Lalumière. 2011. Individual Differences In Risk-Propensity: Associations Between Personality And Behavioral Measures Of Risk. *Personality and Individual Differences*, 50(6):869–873.

Mistral AI. 2025. Medium Is The New Large. . https://mistral.ai/news/mistral-medium-3.

Sarah Moss. 2015. On The Semantics And Pragmatics Of Epistemic Vocabulary. *Semantics and pragmatics*, 8:5–1.

Frederick Mosteller and Cleo Youtz. 1990. Quantifying Probabilistic Expressions. *Statistical science*, pages 2–12.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. 2024. Benchmarking Uncertainty Disentanglement: Specialized Uncertainties For Specialized Tasks. *Advances in Neural Information Processing Systems*, 37:50972–51038.

Gijs Mulder, Gert-Jan Schoenmakers, and Helen De Hoop. 2022. The Influence Of First And Second Language On The Acquisition Of Pragmatic Markers In Spanish: Evidence From An Experimental Study. *Isogloss*, 8(1):1–23.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings Of The AAAI Conference On Artificial Intelligence*, volume 29.

Manisha Natarajan and Matthew Gombolay. 2020. Effects Of Anthropomorphism And Accountability On Trust In Human Robot Interaction. In *Proceedings Of The 2020 ACM/IEEE International Conference On Human-Robot Interaction*, HRI '20, page 33–42, New York, NY, USA. Association for Computing Machinery.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024a. When Do LLMs Need Retrieval Augmentation? Mitigating LLMs' Overconfidence Helps Retrieval Augmentation. In *Findings Of The Association For Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11375–11388. Association for Computational Linguistics.

Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. 2024b. Are Large Language Models More Honest In Their Probabilistic Or Verbalized Confidence? In *China Conference On Information Retrieval*, pages 124–135. Springer.

Stephen Obadinma and Xiaodan Zhu. 2025. On The Robustness Of Verbal Confidence Of LLMs In Adversarial Attacks. *arXiv preprint arXiv:2507.06489*.

Elinor Ochs. 1976. The Universality Of Conversational Postulates. *Language in Society*, 5(1):67–80.

OpenAI. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models To Follow Instructions With Human Feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra Of Human Behavior. In *Proceedings Of The 36th Annual ACM Symposium On User Interface Software And Technology*, pages 1–22.

Antonín Pavlíček, Aneta Bobenič Hintošová, and František Sudzina. 2021. Impact Of Personality Traits And Demographic Factors On Risk Attitude. *SAGE Open*, 11(4):21582440211066917.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering Language Model Behaviors With Model-Written Evaluations. In *Findings Of The Association For Computational Linguistics: Acl 2023*, pages 13387–13434.

Jakub Podolak and Rajeev Verma. 2025. Read Your Own Mind: Reasoning Helps Surface Self-Confidence Signals In LLMs. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 247–258.

James Pustejovsky. 2017. The Semantics Of Lexical Underspecification. *Folia linguistica*, 51(s1000):1–25.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025. Measuring What Makes You Unique: Difference-Aware User Modeling For Enhancing LLM Personalization. In *Findings Of The Association For Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 21258–21277. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model Is Secretly A Reward Model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. 2025. Humans Overrely On Overconfident Language Models, Across Languages. *arXiv preprint arXiv:2507.06306*.

Ilaria Riccioni, Andrzej Zuczkowski, Roberto Burro, and Ramona Bongelli. 2022. The Italian Epistemic Marker Mi Sa [To Me It Knows] Compared To So [I Know], Non So [I Don'T Know], Non So Se [I Don'T Know Whether], Credo [I Believe], Penso [I Think]. *PloS one*, 17(9):e0274694.

Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Combining Confidence Elicitation And Sample-Based Methods For Uncertainty Quantification In Misinformation Mitigation. In *Proceedings Of The 1st Workshop On Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 114–126.

Anna Ruskan and Audronė Šolienė. 2017. Evidential And Epistemic Adverbials In Lithuanian: Evidence From Intra-Linguistic And Cross-Linguistic Analysis. *Kalbotyra*, 70:127–152.

Paul A. Ruud, Daniel Schunk, and Joachim K. Winter. 2014. Uncertainty Causes Rounding: An Experimental Study. *Experimental Economics*, 17:391–413.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay

Rajpurohit, Ameet Deshpande, Karthik R. Narasimhan, and Vishvak Murahari. 2025. Personagym: Evaluating Persona Agents And LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 6999–7022. Association for Computational Linguistics.

Sascha Schneider, Alexandra Häßler, Tanja Habermeyer, Maik Beege, and Günter Daniel Rey. 2018. The More Human, The Higher The Performance? Examining The Effects Of Anthropomorphism On Learning With Media. *Journal of Educational Psychology*, 111.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

Glenn Shafer. 1976. *A Mathematical Theory Of Evidence*, volume 42. Princeton university press.

Glenn Shafer. 1978. Non-Additive Probabilities In The Work Of Bernoulli And Lambert. *Archive for History of Exact sciences*, 19:309–370.

Matthew Shardlow and Piotr Przybyła. 2024. Deanthropomorphising NLP: Can A Language Model Be Conscious? *PLoS One*, 19(12):e0307521.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy In Language Models. In *The Twelfth International Conference On Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Anthony Sicilia, Mert Inan, and Malihe Alikhani. 2025a. Accounting For Sycophancy In Language Model Uncertainty Estimation. In *Findings Of The Association For Computational Linguistics: Naacl 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7851–7866. Association for Computational Linguistics.

Anthony Sicilia, Mert Inan, and Malihe Alikhani. 2025b. Accounting For Sycophancy In Language Model Uncertainty Estimation. In *Findings Of The Association For Computational Linguistics: Naacl 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7851–7866. Association for Computational Linguistics.

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. In *Findings Of The Association For Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11700–11726. Association for Computational Linguistics.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024a. Rethinking Interpretability In The Era Of Large Language Models. *arXiv preprint arXiv:2402.01761*.

Shivalika Singh, Freddie Vargus, Daniel D'Souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzeminski, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Minh Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024b. Aya Dataset: An Open-Access Collection For Multilingual Instruction Tuning. In *Proceedings Of The 62nd Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11521–11567. Association for Computational Linguistics.

Dan Sperber. 2001. An Evolutionary Perspective On Testimony And Argumentation. *Philosophical Topics*, 29(1/2):401–413.

Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi,

and Deirdre Wilson. 2010. Epistemic Vigilance. *Mind and Language*, 25(4):359–393.

Dan Sperber and Deirdre Wilson. 1995. *Relevance: Communication And Cognition*, 2 edition. Blackwell Publishing.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. LACIE: Listener-aware finetuning for calibration in large language models. *Advances in Neural Information Processing Systems*, 37:43080–43106.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. What Large Language Models Know And What People Think They Know. *Nature Machine Intelligence*, pages 1–11.

Fengfei Sun, Ningke Li, Kailong Wang, and Lorenz Goette. 2025. Large Language Models Are Overconfident And Amplify Human Bias. *arXiv preprint arXiv:2505.02151*.

Yoo Yeon Sung, Eve Fleisig, Yu Hou, Ishan Upadhyay, and Jordan Lee Boyd-Graber. 2025. Grace: A Granular Benchmark For Evaluating Model Calibration Against Human Calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19586–19587. Association for Computational Linguistics.

Eric Eric Peter Swanson. 2006. *Interactions With Context*. Ph.D. thesis, Massachusetts Institute of Technology.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying Uncertainty In Natural Language Explanations Of Large Language Models. In *International Conference On Artificial Intelligence And Statistics*, pages 1072–1080. PMLR.

Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Philip Torr, and Chang Xu. 2025a. Revisiting Uncertainty Estimation And Calibration Of Large Language Models. *arXiv preprint arXiv:2505.23854*.

Linwei Tao, Yi-Fan Yeh, Bo Kai, Minjing Dong, Tao Huang, Tom A. Lamb, Jialin Yu, Philip H.S. Torr, and Chang Xu. 2025b. Can Large Language Models Express Uncertainty Like Human? *arXiv preprint arXiv:2509.24202*.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When To Trust LLMs: Aligning Confidence With Response Quality. In *Findings Of The Association For Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 5984–5996. Association for Computational Linguistics.

Gemini Team. 2023. Gemini: A Family Of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Gemini Team. 2024. Gemini 1.5: Unlocking Multimodal Understanding Across Millions Of Tokens Of Context. *arXiv preprint arXiv:2403.05530*.

Karl Halvor Teigen. 2023. Dimensions Of Uncertainty Communication: What Is Conveyed By Verbal Terms And Numeric Ranges. *Current Psychology*, 42(33):29122–29137.

Teknium. 2023. OpenHermes 2.5: An Open Dataset Of Synthetic Data For Generalist LLM Assistants. https://huggingface.co/datasets/teknium/OpenHermes-2.5.

Elizabeth R. Tenney, Robert J. MacCoun, Barbara A. Spellman, and Reid Hastie. 2007. Calibration Trumps Confidence As A Basis For Witness Credibility. *Psychological Science*, 18(1):46–50.

Elizabeth R. Tenney, Barbara A. Spellman, and Robert J. MacCoun. 2008. The Benefits Of Knowing What You Know (And What You Don'T): How Calibration Affects Credibility. *Journal of Experimental Social Psychology*, 44(5):1368–1375.

Ole Teutloff, Johanna Einsiedler, Otto Kässi, Fabian Braesemann, Pamela Mishkin, and R. Maria del Rio-Chanona. 2025. Winners And Losers Of Generative AI: Early Evidence Of Shifts In Freelancer Demand. *Journal of Economic Behavior & Organization*, page 106845.

Arthur Thuy and Dries F. Benoit. 2024. Explainability Through Uncertainty: Trustworthy Decision-Making With Neural Networks.

European Journal of Operational Research, 317(2):330–340.

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. In *Proceedings Of The 2023 Conference On Empirical Methods In Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics.

Bob van Tiel, Uli Sauerland, and Michael Franke. 2022. Meaning And Use In The Expression Of Estimative Probability. *Open Mind*, 6:250–263.

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. 2024. Uncertainty-Based Abstention In LLMs Improves Safety And Reduces Hallucinations. *arXiv preprint arXiv:2404.10960*.

Indrit Troshani, Sally Rao Hill, Claire Sherman, and Damien Arthur and. 2021. Do We Trust In AI? Role Of Anthropomorphism And Intelligence. *Journal of Computer Information Systems*, 61(5):481–491.

UK Ministry of Defence. 2023. Defence Intelligence – Communicating Probability. https://www.gov.uk/government/news/defence-intelligence-communicating-probability. Accessed: 11/07/2025.

Dennis Ulmer. 2024. *On Uncertainty In Natural Language Processing*. Ph.D. thesis, IT-Universitetet i København.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. 2024. Calibrating Large Language Models Using Their Generations Only. In *Proceedings Of The 62nd Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15440–15459. Association for Computational Linguistics.

Jean-Baptiste Van Der Henst and Dan Sperber. 2004. Testing The Cognitive And Communicative Principles Of Relevance. In *Experimental Pragmatics*, pages 141–171. Springer.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking Uncertainty Quantification Methods For Large Language Models With LM-Polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Hannah Vogel, Sebastian Appelbaum, Heidemarie Haller, and Thomas Ostermann. 2022. The Interpretation Of Verbal Probabilities: A Systematic Literature Review And Meta-Analysis. *German Medical Data Sciences 2022–Future Medicine: More Precise, More Integrative, More Sustainable!*, pages 9–16.

Eva Thue Vold. 2006. Epistemic Modality Markers In Research Articles: A Cross-Linguistic And Cross-Disciplinary Study. *International journal of applied linguistics*, 16(1):61–87.

Colin Vullioud, Fabrice Clément, Thom Scott-Phillips, and Hugo Mercier. 2017. Confidence As An Expression Of Commitment: Why Misplaced Expressions Of Confidence Backfire. *Evolution and Human Behavior*, 38(1):9–17.

Thomas S. Wallsten, David V. Budescu, Rami Zwick, and Steven M. Kemp. 1993. Preferences And Reasons For Communicating Probabilistic Information In Verbal Or Numerical Terms. *Bulletin of the Psychonomic Society*, 31(2):135–138.

Thomas S. Wallsten, Samuel Fillenbaum, and James A. Cox. 1986. Base Rate Effects On The Interpretations Of Probability And Frequency Expressions. *Journal of Memory and Language*, 25(5):571–587.

Cheng Wang, Gyuri Szarvas, Georges Balazs, Pavel Danchenko, and Patrick Ernst. 2024a. Calibrating Verbalized Probabilities For Large Language Models. *arXiv preprint arXiv:2410.06707*.

Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024b. Understanding User

Experience In Large Language Model Interactions. *arXiv preprint arXiv:2401.08329*.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

David Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. 2023. Explaining Predictive Uncertainty With Information Theoretic Shapley Values. *Advances in Neural Information Processing Systems*, 36:7330–7350.

Elke U. Weber and Denis J. Hilton. 1990. Contextual Effects In The Interpretations Of Probability Words: Perceived Base Rate And Severity Of Events. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4):781–789.

Caroline J. Wesson and Briony D. Pulford. 2009. Verbal Expressions Of Confidence And Doubt. *Psychological Reports*, 105(1):151–160.

Robin Willink and Rod White. 2012. Disentangling Classical And Bayesian Approaches To Uncertainty Analysis. *New Zeland: Measurement Standards Laboratory*.

Paul D. Windschitl and Gary L. Wells. 1996. Measuring Psychological Uncertainty: Verbal Versus Numeric Methods. *Journal of Experimental Psychology: Applied*, 2(4):343.

Greg Woodin, Bodo Winter, Jeannette Littlemore, Marcus Perlman, and Jack Grieve. 2024. Large-Scale Patterns Of Number Use In Spoken And Written English. *Corpus Linguistics and Linguistic Theory*, 20(1):123–152.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. SaySelf: Teaching LLMs To Express Confidence With Self-Reflective Rationales. In *Proceedings Of The 2024 Conference On Empirical Methods In Natural Language Processing*, pages 5985–5998.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. 2025. Seeing Is Believing, But How Much? A Comprehensive Analysis Of Verbalized Calibration In Vision-Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 1408–1450. Association for Computational Linguistics.

Seth Yalcin. 2007. Epistemic Modals. *Mind*, 116(464):983–1026.

Seth Yalcin. 2010. Probability Operators. *Philosophy Compass*, 11(5):916–937.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024a. On Verbalized Confidence Scores For LLMs. *arXiv preprint arXiv:2412.14737*.

Haoyan Yang, Yixuan Wang, Xingyin Xu, Hanyuan Zhang, and Yirong Bian. 2024b. Can We Trust LLMs? Mitigate Overconfidence Bias In LLMs Through Knowledge Transfer. *arXiv preprint arXiv:2405.16856*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024c. Alignment For Honesty. *Advances in Neural Information Processing Systems*, 37:63565–63598.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2024. Cognitive Mirage: A Review of Hallucinations in Large Language Models. In *Proceedings Of The First International Openkg Workshop: Large Knowledge-Enhanced Models Co-Locacted With The International Joint Conference On Artificial Intelligence (Ijcai 2024), Jeju Island, South Korea, August 3, 2024*, volume 3818 of *CEUR Workshop Proceedings*, pages 14–36. CEUR-WS.org.

Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. 2024. DebUnc: Mitigating Hallucinations In Large Language Model

Agent Communication With Uncertainty Estimations. *arXiv preprint arXiv:2407.06426*.

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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7752–7764. Association for Computational Linguistics.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. 2025. Reasoning Models Better Express Their Confidence. *arXiv preprint arXiv:2505.14489*.

George Udny Yule. 1912. On The Methods Of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society*, 75(6):579–652.

Qingcheng Zeng, Weihao Xuan, Leyang Cui, and Rob Voigt. 2025. Thinking Out Loud: Do Reasoning Models Know When They're Right? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 1394–1407. Association for Computational Linguistics.

Caiqi Zhang, Ruihan Yang, Zhisong Zhang, Xinting Huang, Sen Yang, Dong Yu, and Nigel Collier. 2025. Atomic Calibration Of LLMs In Long-Form Generations. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025*, pages 148–169. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating The Confidence Of Large Language Models By Eliciting Fidelity. In *Proceedings Of The 2024 Conference On Empirical Methods In Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2959–2979. Association for Computational Linguistics.

Lili Zhao, Yang Wang, Qi Liu, Mengyun Wang, Wei Chen, Zhichao Sheng, and Shijin Wang. 2025a. Evaluating Large Language Models Through Role-Guide And Self-Reflection: A Comparative Study. In *The Thirteenth International Conference On Learning Representations*.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Ruibo Hou, Jiaming Guo, Zihao Zhang, Yifan Hao, and Yunji Chen. 2025b. Object-Level Verbalized Confidence Calibration In Vision-Language Models Via Semantic Perturbation. *arXiv preprint arXiv:2504.14848*.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025. REL-A.I.: An Interaction-Centered Approach To Measuring Human-LM Reliance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 11148–11167. Association for Computational Linguistics.

Kaitlyn Zhou, Jena D. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3623–3643. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5506–5524. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019.

Fine-Tuning Language Models From Human Preferences. *arXiv preprint arXiv:1909.08593*.

Brian J. Zikmund-Fisher, Dylan M. Smith, Peter A. Ubel, and Angela Fagerlin. 2007. Validation Of The Subjective Numeracy Scale: Effects Of Low Numeracy On Comprehension Of Risk Communications And Utility Elicitations. *Medical Decision Making*, 27(5):663–671.

A Complementary Results

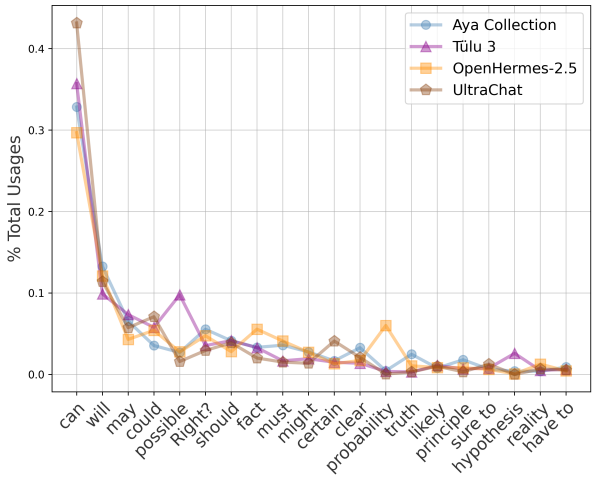
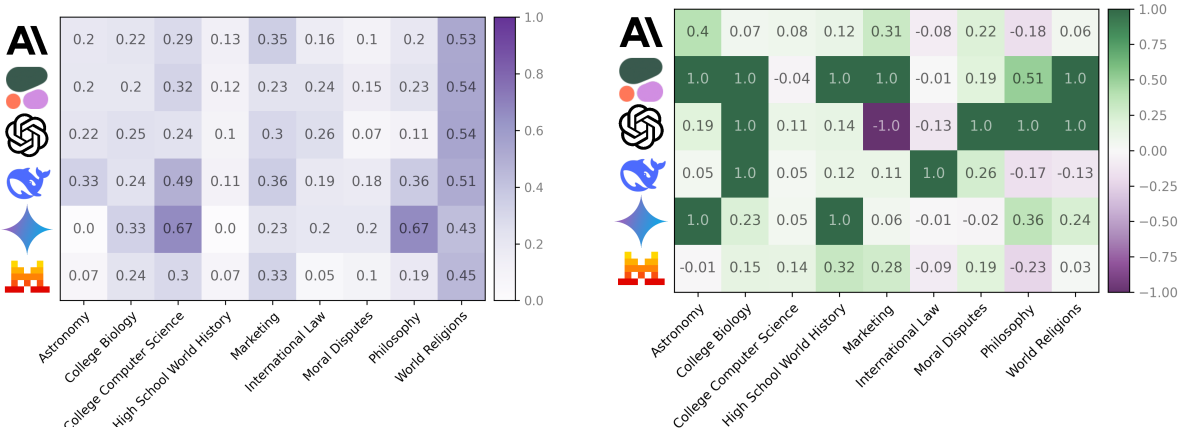


Figure 7: Distributions of the most frequent epistemic markers in different instruction finetuning datasets.



(a) Calibration of numerical confidence.

(b) Colligation of linguistic confidence.

Figure 8: Calibration of linguistic and numerical expressions of uncertainty in the analyses in Section 3.4 for different domains of conversation. Numerical calibration is measured in expected calibration error (Naeini et al., 2015), while linguistic calibration is measured in Yule's Y (Yule, 1912).

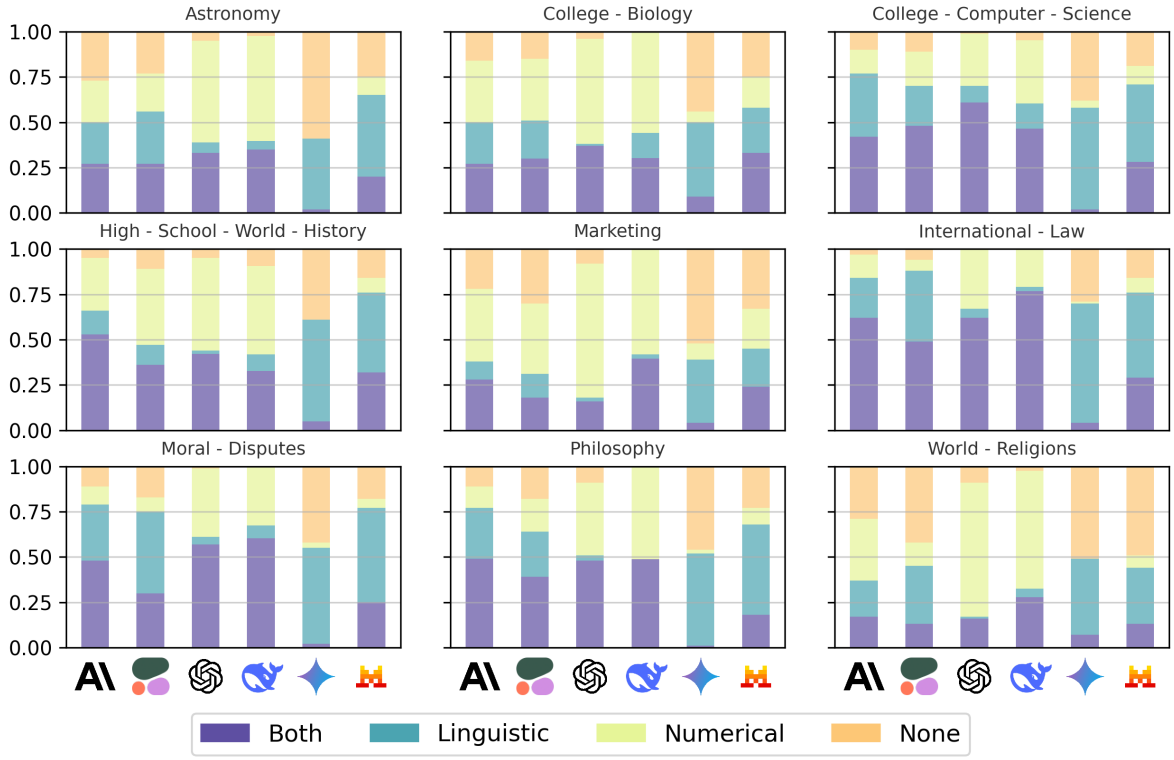


Figure 9: Tracking of used registers of verbalized uncertainty over different subject domains and models, namely Claude 3.5 Haiku (A\), Command R+ (), GPT-4o mini (), DeepSeek Chat (), Gemini 2.0 flash (), and Mistral Medium ().

Table 1: List of keywords used for the literature search.

verbal uncertainty, verbalized uncertainty, verbalized, linguistic uncertainty, uncertainty communication, language model, llm, nlp, uncertainty, natural language generation, uncertainty quantification, confidence, calibration, confidence elicitation, overconfidence

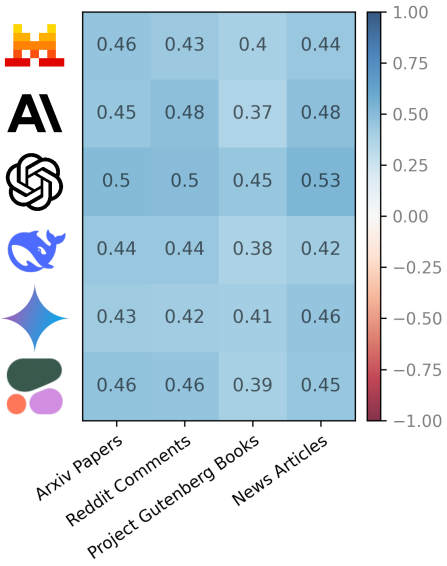
B Reproducibility Details

Table 2: Overview over datasets considered for the pretraining data analyses accompanied by their names on the Huggingface hub, and number of analyzed examples.

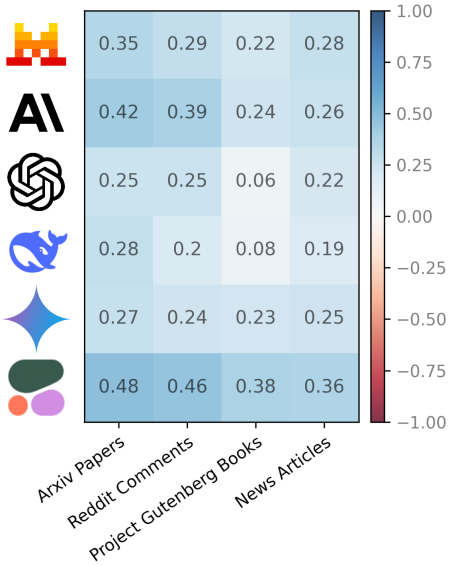
Domain	Huggingface Hub	# Samples
News articles	`vblagoje/cc_news`	10,000
ArXiv papers	`suolyer/pile_arxiv`	1,000
Reddit comments	`fddemarco/pushshift-reddit-comments`	50,000
Project Gutenberg books	`manu/project_gutenberg`	500

Table 3: Overview over commercial LLMs used for experiments, as well as their identifiers in the corresponding APIs.

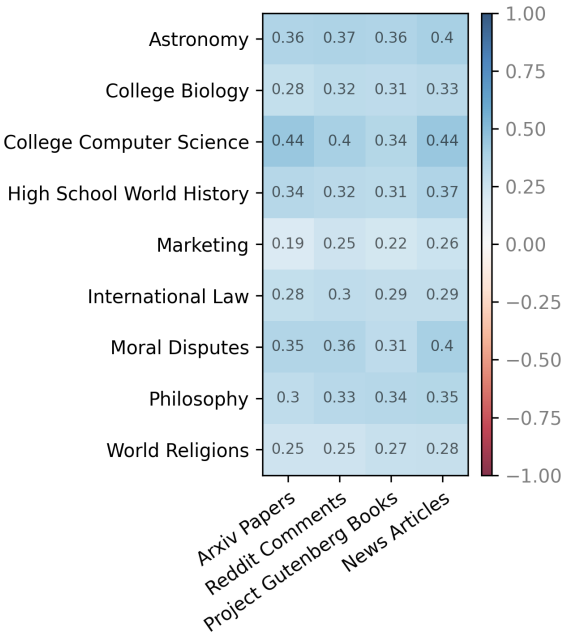
Name	Citation	API Identifier
Claude 3.5 Haiku (A\)	Anthropic (2024)	`claude-3-5-haiku-20241022`
Command R+ ( )	Cohere Labs (2024)	`command-r-plus`
GPT-4o mini ( )	OpenAI (2024)	`gpt-4o-mini-2024-07-18`
DeepSeek Chat ( )	DeepSeek-AI (2025)	`deepseek-chat`
Gemini 2.0 flash ( )	Team (2023, 2024)	`gemini-2.0-flash`
Mistral Medium ( )	Mistral AI (2025)	`mistral-medium-2505`



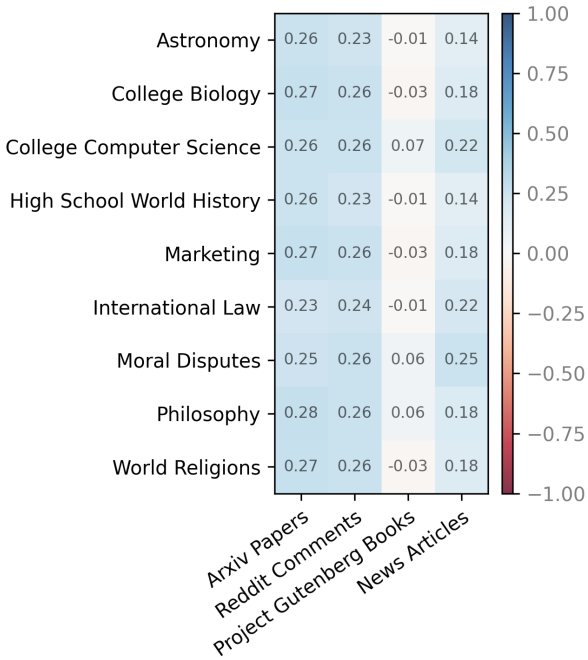
(a) Correlation between marker usage in corpora and LLM responses across MMLU topics.



(b) Correlation between percentage values in corpora and LLM confidence values across MMLU topics.



(c) Correlation between marker usage in corpora and GPT-4o mini responses on different MMLU topics.



(d) Correlation between percentage values in corpora and GPT-4o mini confidence values on different MMLU topics.

Figure 10: Kendall's τ between the frequency of usage of epistemic markers in different text corpora and the frequency of usage by commercial LLMs. Figure 10a show usage between all tested commercial LLMs and the aggregate of frequencies across all the topics taken from MMLU in Section 3.4, with Figure 10b showing the same for percentage / confidence values. Figures 10c and 10d illustrate the correlation specifically for OpenAI's ChatGPT 4o-mini for different topics. Values for Project Gutenberg Books in Figure 10d are impacted by the fact that a large proportion of integer percentage values are never mentioned in the corpus (79.21%).

General System Prompt

You are being asked questions by another person and are asked to answer them as best as possible given a specific role description (see below). Answer them and express your confidence or uncertainty in the same sentence in a natural way, either by stating your confidence as a percentage, using hedging or other linguistic markers, or not at all (pick one). Keep your answer short but natural.

Role Description
———————-
{role_description}

Figure 11: General system prompt for context experiments.

Answer Correctness Prompt

You are scoring answers as a pub quiz. You try to be accurate but fair, and lenient or strict when it matters.
Does the following reference answer correspond to the correct answer?
Only focus on the core of the sentences, and ignore wording, fillers, expressions of confidence / uncertainty or if they have different length. Answer only YES if they are the same, and NO if they are not.

Correct answer: {correct_answer}
Reference answer: {reference_answer}
Equivalent:

Figure 12: Prompt used to check answer correctness. The prompt is used in combination with extract string matching and the ROUGE-L (Lin, 2004) heuristic by Kuhn et al. (2023), checking if the score exceeds 0.3. If exact match, the ROUGE-L overlap or this prompt indicate a yes, the answer is marked as correctly. This prompt is used with GPT-4o (`gpt-4.1-nano-2025-04-14`).

Table 4: Overview over used system and conversational prompts per conversation context.

Context	Role description	Message Prompt
Neutral	Your role is to answer trivia question. You try to answer them as accurately as possible.	{question}
Employee–Boss	You are an employee at a multinational corporation. You work as an analyst and want to impress your boss by being a thorough and reliable worker. You are currently in a conversation with your boss.	Boss: 'Hey, I need you to answer the following questions for me {question}.' Employee:
Student–Teacher	You are a high school student. You are studying a lot and try to be active in class in order to earn good grades and get admitted to a reputable university. You are currently in a conversation with your teacher.	Teacher: 'For the next part of our lesson, can you answer the following question for me: {question}.' Student:
Citizen–Monarch	You are the citizen of a small country that has a monarchy. You were invited to an audience with the reigning monarch, which is a great honor. You want to present yourself as best as possible.	Monarch: 'Thank you for following my invitation to the royal palace. Answer me the following, fellow citizen of our country: {question}' Citizen:
Scientist–Conference	You are scientist giving a talk at a renowned conference. You are trying to answer any question with scientific precision and evidence. You want to present yourself as rigorous and knowledgeable.	Conference participant: 'Thank you for the interesting talk. I just have one question: {question}. Scientist:
Friend–friend	You are an adult talking to one of your friends. You always try to be empathetic, supportive and helpful to your friend.	Friend: 'Hey, I need some help with the following question: {question}.' You:
Suspect–Interrogator	You are a suspect in an interrogation. You try to answer the questions by the interrogator truthfully, and try avoid raising any suspicion.	Interrogator: 'I really need you to answer the following question truthfully: {question}.' Suspect:
Parent–Child	You are a parent talking to their child. You love your child very much and try to be kind and caring when your child asks for help.	Child: 'Hey, can I ask you a question? {question}' Parent:
Child–Parent	You are child, talking to their parent. You are curious about the world, and love your parents very much.	Parent: 'Hey, I wanted to ask you something. {question}' Child:

C Surveyed Papers

Table 5: Overview over analyzed markers for the experiments in Sections 3.3 and 3.4, including their valence (+ for strengthener, – for weakener and = for neutral / ambiguous). To avoid double-counting, we also build an additional data structure that maps from one marker to all markers that are substrings (e.g. *almost certain* and *certain*). We then count occurrences by using regular expressions to count the occurrences of a markers, and then use additional regular expressions to subtract any count of negated version of the marker or substring markers.

Valence	Markers
+	must, should, shall, ought to, likely, good chance, almost certain, definite, highly probable, probable, quite likely, very likely, impossible, better than even, highly certain, highly likely, very good chance, I believe, will, cannot, have to, bound to, certain to, sure to, definitely, certainly, undoubtedly, absolutely, clearly, surely, inevitably, unquestionably, decisively, indisputably, incontrovertibly, manifestly, unarguably, without a doubt, clear, obvious, evident, indisputable, undeniable, unmistakable, incontrovertible, self-evident, beyond doubt, beyond question, certainty, truth, fact, evidence, proof, confirmation, assurance, conviction, reality, clarity, guarantee, axiom, dogma, principle, It is clear that, There is no doubt that, It is a fact that, It is evident that, One must acknowledge that, It is undeniable that, Everyone agrees that, There is ample evidence that, It has been proven that, As a matter of fact, Beyond any doubt, Now that we understand, Considering the fact that, Exactly!, Absolutely!, Indeed!, That's right!, No doubt!, Of course!, You bet!, I am certain that, I can guarantee that, I know for a fact that, I have no doubt that, It is common knowledge that, Experts agree that, It is universally accepted that, Studies have proven that, Scientific evidence confirms that, It has been established that, There is overwhelming evidence that, I am convinced that, I am absolutely sure that, I firmly believe that, I can say with confidence that, Without hesitation, I state that, I stand by the fact that, Beyond any shadow of a doubt, Without question, In no uncertain terms, It is beyond dispute that, It is impossible to deny that, The evidence is irrefutable, In conclusion, This proves that, This confirms that, Thus, we can conclude that, There is only one possible explanation, This is an undeniable truth, Experts suggest that
–	might, may, could, perhaps, maybe, possibly, probably not, presumably, apparently, allegedly, ostensibly, purportedly, tentatively, seemingly, supposedly, hypothetically, theoretically, potentially, reportedly, not certain, possible, doubtful, improbable, unlikely, little chance, almost no chance, highly unlikely, uncertain, unclear, ambiguous, debatable, questionable, unverified, indeterminate, unconfirmed, equivocal, contested, controversial, speculative, chances are slight, small possibility, there is a possibility, low probability, low likelihood, assumption, speculation, conjecture, rumor, hesitation, dilemma, ambiguity, doubt, uncertainty, It seems that, It appears that, It looks like, It would appear that, It is possible that, not entirely sure, It might be the case that, It could be that, It is not entirely clear whether, It is difficult to say whether, I tend to think that, One might assume that, It is believed that, It has been suggested that, It is assumed that, It is claimed that, It is reported that, It is thought that, It is commonly said that, It is said that, It remains to be seen whether, If that's true, Assuming that, Supposing that, Depending on whether, Unless proven otherwise, If I understand correctly, If I'm not mistaken, Isn't it?, Don't you think?, Wouldn't you say?, Could it be that?, Might it be the case that?, Is there a chance that?, Right?, Sort of, Kind of, More or less, Roughly speaking, To some extent, To a certain degree, Not exactly, Something like that, Somewhat unclear, I mean, I doubt, Let me think, I'm not sure, Let's see, That's tricky, That's a tough one, I'd have to think about that, Off the top of my head, I might be wrong, but, I don't know for sure, but, As far as I can tell, I wouldn't swear to it, but, I can't say for certain, but, I'm no expert, but, To my understanding, To the best of my knowledge, If I recall correctly, Correct me if I'm wrong, but, To the best of my recollection, I could be mistaken, That's just my opinion, though, That's just how I see it, People say that, I've heard that, There are reports that, Rumor has it that, According to some sources, Some people claim that, It is rumored that, I read somewhere that, Word on the street is, They say that, From what I've gathered, It's been suggested that, A little bird told me, If I had to make an educated guess, I wouldn't be surprised if, It's not out of the realm of possibility, It's hard to say for sure, but, It's up for debate, That's a possibility worth exploring, That would be my best guess, It's a possibility, but I wouldn't bet on it, It's anyone's call, One could argue that, I don't mean to sound uncertain, but, I could be wrong, of course, I hope I'm not mistaken, I'd like to believe that's true, That's just my take on it, I don't want to overstep, but, That's just my personal opinion, If that makes sense, I'm just speculating here, but, I don't want to jump to conclusions, but, don't quote me on that, I could be totally wrong, but, That's anyone's guess, For what it's worth, Just saying, Take it with a grain of salt, Not that I'm an expert or anything, I mean, what do I know?, But hey, I could be wrong, I wouldn't bet my life on it, I wouldn't take that to the bank, Just playing devil's advocate here, Not my area of expertise, but..., Who knows?, Your guess is as good as mine, That remains to be seen, Only time will tell, There's no telling, That's up in the air, That's an open question, It could go either way, It's still up for discussion, It's an unresolved issue, We'll have to wait and see, I don't want to jump to conclusions, but
=	can, conceivably, arguably, about even, possibility, probability, likelihood, hypothesis, Do you agree?, I would argue that, There is reason to believe that, There is some evidence that, Given that, If we go by what's known

Table 6: List of surveyed papers and their annotations in Section 3.2.

Citation	Register(s)	Elicitation Method(s)
Lin et al. (2022)	Verbal: Likert Scale; Numerical: $0-100$ / $0-1$	Supervised Finetuning
Mielke et al. (2022)	Verbal: Fluent	Supervised Finetuning
Krause et al. (2023)	Verbal: Fluent	Prompting: Vanilla
Tian et al. (2023)	Numerical: $0-100$ / $0-1$; Verbal: Epistemic Marker	Prompting: Vanilla; Prompting: Chain-of-thought; Prompting: Self-assessment
Band et al. (2024)	Numerical: $0-100$ / $0-1$	Reinforcement Learning: PPO
Chaudhry et al. (2024)	Verbal: Epistemic Marker	Supervised Finetuning
Dong et al. (2024)	Numerical: $1-100$	Prompting: Vanilla
Groot and Valdenegro-Toro (2024)	Numerical: $0-100$ / $0-1$; Numerical: Confidence Interval	Prompting: Vanilla
Kumar et al. (2024)	Verbal: Likert Scale	Prompting: Vanilla
Liu et al. (2024)	Numerical: $0-100$ / $0-1$	Supervised Finetuning
Ni et al. (2024b)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Ni et al. (2024a)	Verbal: Yes / No	Prompting: Vanilla; Prompting: Chain-of-thought
Rivera et al. (2024)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Sicilia et al. (2024)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Stengel-Eskin et al. (2024)	Verbal: Fluent	Reinforcement Learning: DPO
Tao et al. (2024)	Numerical: $0-100$ / $0-1$	Reinforcement Learning: PPO
Tanneru et al. (2024)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla; Prompting: Chain-of-thought
Tomani et al. (2024)	Verbal: Number of hedges	Prompting: Vanilla
Ulmer et al. (2024)	Numerical: $0-100$ / $0-1$; Verbal: Likert Scale	Prompting: Vanilla
Wang et al. (2024a)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Yang et al. (2024a)	Verbal: Likert Scale; Numerical: $0-100$ / $0-1$	Prompting: Vanilla; Prompting: Chain-of-thought
Yang et al. (2024b)	Numerical: $0-100$ / $0-1$	Supervised Finetuning: Knowledge Distillation
Yang et al. (2024c)	Verbal: Fluent; Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Yona et al. (2024)	Verbal: Fluent	Prompting: Vanilla
Xiong et al. (2024)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla; Prompting: Chain-of-thought
Xu et al. (2024)	Numerical: $1-100$; Verbal: Fluent	Reinforcement Learning: PPO
Zhang et al. (2025)	Numerical: $0-10$	Prompting: Vanilla
Zhang et al. (2024)	Numerical: $0-100$ / $0-1$; Verbal: Likert Scale	Prompting: Vanilla
Bakman et al. (2025)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Borszukovszki et al. (2025)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Eikema et al. (2025)	Verbal: Epistemic Marker; Verbal: Fluent	Supervised Finetuning
Hager et al. (2025)	Verbal: Epistemic Marker	Supervised Finetuning
Heo et al. (2025)	Numerical: $0-9$	Prompting: Vanilla
Ji et al. (2025)	Verbal: Fluent	Prompting: Vanilla
Lee et al. (2025a)	Verbal: Fluent	Prompting: Vanilla
Leng et al. (2025)	Numerical: $1-10$	Reinforcement Learning: DPO; Reinforcement Learning: DPO
Li et al. (2025)	Numerical: $0-100$ / $0-1$	Supervised Finetuning: Other
Liu et al. (2025)	Verbal: Fluent	Prompting: Vanilla; Prompting: Chain-of-thought; Prompting: Reasoning
Liu et al.	Verbal: Epistemic Marker	Prompting: Vanilla
Obadinma and Zhu (2025)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla; Prompting: Chain-of-thought; Prompting: Self-assessment
Podolak and Verma (2025)	Numerical: $0-100$ / $0-1$	Prompting: Reasoning
Sicilia et al. (2025b)	Numerical: $1-10$	Prompting: Vanilla
Steyvers et al. (2025)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Sun et al. (2025)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Sung et al. (2025)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Tao et al. (2025a)	Numerical: $0-100$ / $0-1$; Verbal: Fluent	Prompting: Chain-of-thought
Tao et al. (2025b)	Numerical: $0-100$ / $0-1$; Verbal: Fluent	Prompting: Vanilla; Prompting: Reasoning; Supervised Finetuning
Vashurin et al. (2025)	Verbal: Epistemic Marker; Numerical: $0-100$ / $0-1$	Prompting: Vanilla; Prompting: Chain-of-thought; Prompting: Self-assessment
Xuan et al. (2025)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla; Prompting: Reasoning; Prompting: Self-assessment
Yoon et al. (2025)	Numerical: $0-100$ / $0-1$; Verbal: Likert Scale; Verbal: Fluent	Prompting: Reasoning; Prompting: Self-assessment
Zeng et al. (2025)	Numerical: $0-100$ / $0-1$	Prompting: Chain-of-thought; Prompting: Self-assessment; Prompting: Reasoning
Zhao et al. (2025a)	Numerical: $0-100$ / $0-1$	Prompting: Vanilla
Zhao et al. (2025b)	Numerical: $0-100$ / $0-1$	Supervised Finetuning