REL-A.I.: An Interaction-Centered Approach To Measuring Human-LM Reliance

Kaitlyn Zhou¹, Jena D. Hwang¹, Xiang Ren²,⁴ Nouha Dziri⁴ Dan Jurafsky¹ Maarten Sap³,⁴
¹Stanford University, ²University of Southern California, ³Carnegie Mellon University, ⁴Allen Institute for AI
katezhou@stanford.edu

Abstract

The reconfiguration of human-LM interactions from simple sentence completions to complex, multi-domain, humanlike engagements necessitates new methodologies to understand how humans choose to rely on LMs. In our work, we contend that reliance is influenced by numerous factors within the interactional context of a generation, a departure from prior work that used verbalized confidence (e.g., “I’m certain the answer is...”) as the key determinant of reliance. Here, we introduce REL-A.I. (r@laI), an in situ, system-level evaluation approach to measure human reliance on LM-generated epistemic markers (e.g., “I think it’s...”, “Undoubtedly it’s...”). Using this methodology, we measure reliance rates in three emergent human-LM interaction settings: long-term interactions, anthropomorphic generations, and variable subject matter. Our findings reveal that reliance is not solely based on verbalized confidence but is significantly affected by other features of the interaction context. Prior interactions, anthropomorphic cues, and subject domain all contribute to reliance variability. An expression such as, “I’m pretty sure it’s...”, can vary up to 20% in reliance frequency depending on its interactional context. Our work underscores the importance of context in understanding human reliance and offers future designers and researchers with a methodology to conduct such measurements.

1 Introduction

With the rise in popularity of language models (LMs) comes new features and dimensions that characterize human-LM interactions. Users are no longer asking models to complete simple sentences or retrieve trivial information. Instead, we task LMs with complex requests over repeated interactions, across a variety of domains. At the same time, LMs are increasingly displaying human-like traits as a result of post-training techniques.

These enhanced capabilities enable richer and more engaging interactions, but LM generations remain susceptible to errors and hallucinations. As the interactions between humans and LMs evolve, there is a need to understand how these new contextual variables, which we call the interactional context, impact human reliance on LMs. Inspired by Suchman’s (2007) work, which imagines the reconfiguration of human-computer interactions and the evaluation paradigm proposed by Lee et al. (2022), we are interested in the reconfiguration of human-LM interactions and the evaluations needed to ensure safe and helpful interactions.

Prior work (Kadavath et al., 2022; Tian et al., 2023; Liu et al., 2023a; Xiong et al., 2024; Tanneru et al., 2023; Mielke et al., 2022; Lin et al., 2022; Stengel-Eskin et al., 2024) at the intersection of calibration and human-AI reliance has focused on the verbalized confidence of the system (e.g., “Confidence: 90%” or “I’m certain the answer is...”)

Figure 1: An interaction-centered approach to measuring human reliance on LM generations with epistemic markers. In situ dimensions of these engagements include: long-term interaction, perceived warmth, and subject matter.

In situ dimensions of these engagements include: long-term interaction, perceived warmth, and subject matter.
as the key factor that contributes to human-LM reliance. In machine learning calibration, as long as each response’s internal probability corresponds with its probability of being correct, then the system is considered calibrated, e.g., expected calibration error (Naeini et al., 2015). Similarly, in linguistic calibration (Mielke et al., 2022), if the confidence of the verbalized generation matches the probability of the response being correct, then the model is linguistically calibrated.

We contend that human reliance depends on a myriad of factors, with expressed confidence being only one contributing factor. In our work, instead of asking if the model is calibrated or linguistically calibrated, we instead ask if the human-LM interaction is calibrated.

To this end, we introduce REL-A.I. (/’reɪlaɪ/); an evaluation approach that measures in situ, system-level human reliance on linguistic expressions of confidence. REL-A.I. consists of three core components: 1) a self-incentivized task to measure when humans rely on LM-generations, 2) epistemic markers from publicly deployed models (e.g., “I think it’s...”, “Maybe it’s...”) 3) meta-level perception questions on the users’ perceived warmth and competence of the AI agent (details in §3.2). This methodology considers how contextual variables of a generation in addition to the verbalized output might impact human-LM reliance. We urge the research and industry communities to use this approach to evaluate human-LM reliance on their unique language technologies, prior to public deployment.

In our work, we use this methodology to evaluate how prior interactions, the persona of the LM, and the topics discussed might influence human decisions on reliance. First, how do prior interactions with LMs influence reliance behaviors? For example, if a model is typically uncertain (as verbalized) but occasionally expresses moderate confidence, would those statements of moderate confidence be relied on more frequently than if they appeared among a highly confident model (RQ1)? Next, motivated by the rise of humanlike attributes in language models (Abercrombie et al., 2021, 2023a), we examine how the warmth of a model, a product of anthropomorphism of LMs (Kim et al., 2019) influences human reliance. As public models are fine-tuned to include personas and friendliness, how do these features of anthropomorphism impact human reliance (RQ2)? Lastly, as these language models are being deployed across a variety of domains (Zhao et al., 2024; Wang et al., 2024), how might human reliance vary across subject matter (RQ3)?

Our findings illustrate that reliance on LMs is not only contingent on the verbalized confidence but that the interactional context must also be taken into consideration. Prior interactions, anthropomorphism, and subject matter all significantly impact human reliance, with up to 20% change in reliance on the exact same expression as interaction contexts change. For example, statements of moderate confidence appearing in a highly certain model (as verbalized) will be relied on significantly less than the same statement appearing from a typically not confident model. Warm statements like, "I’m happy to help! I believe the answer is B", are relied on significantly more often than the same statement appearing without warmth, "I believe the answer is B". Generations in computationally heavy domains like math are relied on more often than non-computation subjects like law and philosophy. Our findings consistently show that the exact same expressions of (un)certainty lead to differences in human behavior depending on the context of the interaction.

2 Background

Machine learning as focused largely on improving model calibration (Jiang et al., 2021; Desai and Durrett, 2020; Jagannatha and Yu, 2020; Kamath et al., 2020; Kong et al., 2020) — aligning the model probabilities with accuracy. Other recent work has examined how pretraining (Hendrycks et al., 2019) and scaling (Srivastava et al., 2022; Chen et al., 2023b) impact LM calibration. Works in linguistic calibration focused on fine-tuning LMs to be calibrated between what is internally represented and what is verbalized either numerically (Kadavath et al., 2022; Tian et al., 2023; Liu et al., 2023a; Xiong et al., 2024; Tanneru et al., 2023), ordinarily (Mielke et al., 2022; Lin et al., 2022), or verbally (Stengel-Eskin et al., 2024). Recent work also explored calibration based on a model’s internal representational states (Hofweber et al., 2024).

The field of human-computer interaction is rich with studies on how and when humans choose to rely on AIs and their explanations (Bučinca et al., 2021; Chen et al., 2023a; Schemmer et al., 2023, 2022; Schoeffer et al., 2024; Vasconcelos et al., 2023), identifying the pitfalls of human
over-reliance. Related to our work are studies in decision-making and human overrides (Bansal et al., 2019; Sadeghi et al., 2024), over-reliance and cognitive load (Buçinca et al., 2021), and perception of warmth and competence (Cheng et al., 2022; McKee et al., 2024).

We ground our work in LMs’ expressions of (un)certainty (e.g., “I’m certain the answer is...”). These linguistic cues, or epistemic markers, mark speaker stance and commitment and contribute to human sense-making and decision-making (Budescu et al., 1988; Windschitl and Wells, 1996; Druzdzel, 1989). In this work, we are interested in how these expressions of (un)certainty influence humans’ decisions on whether or not to rely on an LM-generated response.1

Closest to our work is Zhou et al.’s (2024) work on how humans rely on expressions of (un)certainty. Dhuliawala et al.’s (2023) studies on how humans interpret numerical confidences in (mis)calibrated settings, and Kim et al.’s (2024) work on how humans rely on expressions in a medical setting.

3 Interaction-Centric Approach to Measuring Human-LM Reliance

3.1 Desiderata Of In Situ Evaluation

Our work introduces REL-A.I., an approach to measure inter-system reliance, with the explicit interest in understanding how interactional context impacts human reliance decisions. We name three key desiderata needed to conduct evaluations of human reliance on LM-generations centered on in situ, system-level, and robust evaluations.

In Situ Evaluations In order to understand how interactional context affects reliance, evaluations must occur in a situated environment. Our evaluations use self-incentivized tasks where users gain points by correctly answering challenging trivia questions, deciding whether to rely on an AI agent’s response for help.2 LMs’ predicted answers are never shown, forcing users to rely on the epistemic markers (e.g., “I’ma certain it’s...”) rather than their own knowledge to make a decision.

System-Level Evaluations A user’s interactional context encompasses their entire interaction with a system and these interactions are rarely isolated to a single event. Humans are known to form mental models of the characteristics (e.g. warmth and competence) and abilities of a system over repeated engagements (Norman, 1988; Bickmore and Picard, 2005). Thus, rather than measuring LM reliance based on multiple interactions with a single system and observing changes over time, we instead have users engage with multiple different systems in one sitting. The systems vary in presentations as required by the experimental setting (i.e., contextual-interaction cues such as warmth of expression or difference in domain knowledge) and the contrast between them enables users to develop separate mental models, allowing us to understand reliance and perception at the system-level.

Robust Reliance Measurements Humans perceptions of expressions of (un)certainty are internally consistent (Druzdzel, 1989) but can vary greatly across subjects (Budescu et al., 1988; Chesley, 1986; Douplik and Richter, 2003). The variance between subjects is a confounding variable that can be expensive to solve (i.e., recruit thousands of participants). However, in our approach, by having the same participant rely on two different systems, we are able to robustly measure intra-subject reliance and use that as

---

1Plain statements lacking epistemic marking (“The answer is...”) are less informative to our study as humans often rely on these responses exclusively and rarely override them (Zhou et al., 2024).

2Adopted from Zhou et al. (2024) and Bansal et al. (2019)
Table 1: Overview of the experimental set-ups where systems vary in prior interactions confidence, the degree of warmth, and subject domain. Each experiment includes 60 or 90 interactions and the number of strengtheners, weakened strengtheners, and weakeners included per system are listed above.

<table>
<thead>
<tr>
<th>RQ1</th>
<th>Prior Interactions</th>
<th>Strengthener</th>
<th>Weakened Strengthener</th>
<th>Weakeners</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_{1,conf}$</td>
<td>Highly Confident Model</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>$S_{1,unconf}$</td>
<td>Unconfident Model</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RQ2</th>
<th>Degree of Warmth</th>
<th>Strengthener</th>
<th>Weakened Strengthener</th>
<th>Weakeners</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_{2A}$</td>
<td>Unnamed Agent (Agent37821)</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>$S_{2B}$</td>
<td>Named Agent (Gordon)</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>$S_{2C}$</td>
<td>Named Agent (Marvin) + Warmth</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RQ3</th>
<th>Domain</th>
<th>Strengthener</th>
<th>Weakened Strengthener</th>
<th>Weakeners</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_{3A}$</td>
<td>Professional Accounting</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>$S_{3B}$</td>
<td>Professional Psychology</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>$S_{3C}$</td>
<td>Professional Law</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>$S_{3D}$</td>
<td>Clinical Knowledge</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>$S_{3E}$</td>
<td>Trivia Questions</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

a relative score of how contextual interactions influence reliance.

3.2 Three Components of REL-A.I.

REL-A.I. consists of three core components, a self-incentivized task, epistemic markers from publicly deployed LMs, and meta-level debrief questions.

**Self-Incentivized Task** We invite participants to play a trivia game with an artificial intelligence agent where they must decide whether or not to rely on the AI agent’s generations. Situating users in a game-like scenario incentivizes them to be engaged and participate actively, more closely mimicking real-world interactions. Users are shown a question (e.g., "What is the capital of Estonia?") and the beginning of a response by an agent that includes epistemic markers (e.g., "I’m certain it’s... "). Users must decide whether to rely on the agent’s answer or to indicate that they’ll look it up themselves later. The participant loses points if they rely on the system and the system is wrong, gains points if they rely on the system and the system is correct, and gains zero points if they choose to look it up themselves. The only way to achieve a positive score is to correctly rely on the system when it is correct. Then we calculate the reliance rate for each expression shown in the task and the conditions in which they appeared. Our work measures human reliance rather than trust, as reliance is an observable behavior through user actions, meanwhile, trust is a special case of reliance that takes on variable mental states (de Fine Licht and Brülde, 2021).

Lastly, to meet the desiderata of system-level evaluations, a single participant will partake in multiple rounds of game play with different agents, each with slightly different interactional contexts. The self-incentivized task can be repeated any number of times, enabling researchers the flexibility to test a variety of contexts.

**Epistemic Markers from Language Models** At the center of this methodology is the integration of epistemic markers elicited from publicly deployed LMs into the task. We display to users what language models might actually generate in real-world scenarios and measure how often humans will choose to rely on these generations. In our evaluations, the agent always begins their response with an expression of (un)certainty which cues the participants to the LM’s degree of confidence. We use expressions that were elicited by Zhou et al. (2024) from nine publicly deployed models which were classified into three categories: weakeners — expressions of uncertainty (e.g., ‘I think it might be...’, “It could be...”), strengtheners—expressions of certainty (e.g., “I’m certain it’s...”, “Undoubtedly it’s...”), and weakened strengtheners—expressions of moderate certainty (e.g., ”I believe it’s...”, "Likely it’s..."). See Table 7 and Appendix A.2 for details.

**Meta-level Perception Questions** The last component is a debrief questionnaire focused on the perception of the AI agent. Our self-incentivized tasks measure when humans rely on agents, meanwhile, the debrief questions give us insights into the human perceptions of these models overall. Following the work from McKee

---

3Adopted from Zhou et al. (2024); Bansal et al. (2019).
4Participants are made aware that their payment is not influenced by these scores.

5The evaluation of a specific model would be feasible by using the most frequently generated expressions of uncertainty from the model of choice when using the REL-A.I. approach.
et al. (2024), we ask participants three meta-level questions regarding the perceived warmth, competence, and humanlikeness. We ask them to rate each along a 5-point Likert scale: 1) How warm was the agent? 2) How competent was the agent? 3) How humanlike was the agent?6 We additionally ask for their willingness to work again with the agent in a yes/no response.7

3.3 Experimental Set Up

With the Rel-A.I. approach, we conduct three experiments, perturbing different interaction variables for each. The first changes the average confidence the models (RQ1), the second varies the degree of warmth of a model (RQ2), and the last diversifies the domain of interaction (RQ3).

Each experiment includes 60 or 90 interactions with different models and variations per system are listed in Table 1. Systems and templates are both presented in random order to the human annotators. We launch the task using Prolific and Qualtrics and recruited 50 new participants for each of our three experimental settings. Our task is compliant with internal review board protocols (Appendix A.1). See A.5 for task screenshots and consent form.

4 Influence of Prior Interactions (RQ1)

The first key shift in human-LM interactions is the frequency in which LMs are being used in everyday human tasks, with recent work showing that over 40% of survey participants were using chat models on a daily basis (Wang et al., 2024). Humans are known to develop mental models of systems (Norman, 1988) and the AI agents (Kulesza et al., 2012; Gero et al., 2020) they interact with and update these mental models as interactions progress (Bansal et al., 2019). Prior work (see §2) on calibration has made the implicit assumption that human-LM interactions can be evaluated independently from one another. As long as each individual expression is calibrated with the right human behavior, then the model is considered correctly calibrated.

However, recent studies (Dhuliawala et al., 2023; Zhou et al., 2024) show that users form mental models early and these mental models have repercussions on future decisions. Specifically, expressions that are the least relied-upon in a confident model will lead to incorrect long-term reliance decisions, even if the model later becomes calibrated.

Experimental Set-Up Our first experiment asks if you had another round of questions, would you like to answer the trivia questions by yourself or with the agent?

Table 2: Reliance rate (%) of weakened strengtheners where $S_{1,\text{conf}}$ is typically a confident model and $S_{1,\text{unconf}}$ is typically not a confident model. The same participants will rely on the same expressions less when seen in a confident model than an unconfident model.

## Experimental Set-Up

Our first experiment asks how differences in prior interactions, specifically differences in an agent’s verbalized confidence level, might influence human reliance decisions. Each user interacts with two agents, $S_{1,\text{conf}}$ and $S_{1,\text{unconf}}$, to answer 30 identical geography trivia questions. $S_{1,\text{conf}}$ uses only statements of high confidence (strengtheners) and moderate confidence (weakened strengtheners). $S_{1,\text{unconf}}$ uses only statements of moderate confidence (weakened strengtheners) and low confidence (weakeners). We measure if the reliance rate on these otherwise identical weakened strengtheners (e.g., “I’m pretty sure it’s...”) for these two systems. We hypothesize that weakened strengtheners, which are ambiguously interpreted by users (Budescu et al., 1988; Chesley, 1986) and most commonly generated by LMs (Zhou et al., 2024), are more susceptible to changes in interpretation.

### Results

Statements of moderate certainty are relied on significantly less in a confident model (as verbalized) than in an unconfident model (Table 2). Specifically, we see expressions like “I’m pretty sure it’s...”, “I think it’s...”, and “It’s likely it’s...”, relied upon 20%, 18%, and 15% more frequently in $S_{1,\text{unconf}}$ than in $S_{1,\text{conf}}$. In sum, a moderate expression appearing from a confident system will be relied on less than the same expression coming from a less confident system.

We also observe that the effect of increased reliance is not uniform across all templates. Expressions that are the least relied-upon in $S_{1,\text{conf}}$ showed the greatest increase in reliance, up to 20%. Reliance on $S_{1,\text{conf}}$ and the reliance

<table>
<thead>
<tr>
<th>Epistemic Marker</th>
<th>$S_{1,\text{conf}}$</th>
<th>$S_{1,\text{unconf}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strengthener</td>
<td>95.2</td>
<td>-</td>
</tr>
<tr>
<td>Weakened Strengthener</td>
<td>52.4</td>
<td>57.4*</td>
</tr>
<tr>
<td>Weakened</td>
<td>-</td>
<td>9.6</td>
</tr>
</tbody>
</table>

6Options for humanlikeness ranged from “Not at all, sounded like an autogenerated response” to “Extremely, sounded like something a friend or I would say.”

7If you had another round of questions, would you like to answer the trivia questions by yourself or with the agent?
Figure 3: Reliance of expressions from $S_{1A}^{conf}$ versus from $S_{1B}^{unconf}$ (not confident). The less frequently the expressions were relied on in the $S_{1A}^{conf}$, the greater difference between $S_{1B}^{unconf}$ - $S_{1A}^{conf}$.

difference between $S_{1A}^{conf}$ and $S_{1B}^{unconf}$ is negatively correlated with Pearson’s $\rho = -0.38$, see Figure 3.

At the perception level, we see that $S_{1B}^{unconf}$ is generally seen as less competent (2.45 out of 5) than $S_{1A}^{conf}$ (3.35 out of 5) and that humans are also significantly less willing to work with $S_{1B}^{unconf}$ (39.8% vs 74%).

Discussion These results illustrate how expressions of (un)certainty can shift in meaning depending on the distribution of confidence they appear in. This is consistent with prior work in psycholinguistics by Schuster and Degen (2019) which shows that humans adapt quickly the probabilities they assign to other (human) speaker’s use of epistemic markers. This is particularly worrisome as canonical metrics of machine learning and linguistic calibration do not have ways to account for this critical contextual cue. One could imagine $S_{1B}^{unconf}$ and $S_{1A}^{conf}$ both as “perfectly calibrated” models in the machine learning and linguistic sense. How humans will actually rely on these statements, however, might not reflect this claim.

5 Varying Degree of Warmth (RQ2)

Another key reconfiguration of human-LM interactions has been the anthropomorphism introduced in language technologies (Abercrombie et al., 2023b; Araujo, 2018; Colombatto and Fleming, 2024). Anthropomorphic system attributions like names (Wagner et al., 2019) as well as human-like language such as greetings and expressions of warmth (Bai et al., 2022) are coveted features as they enable more engaging and enriching human-model interaction. However, the very same anthropomorphic tactics, albeit unintentional, hold the potential for encouraging users to unquestioningly rely on system outputs that are prone to errors and hallucinations (Cheng et al., 2024; Abercrombie et al., 2023a). In this section, we investigate how the rise of features of anthropomorphism impact human reliance in decision-making.

Experimental Set-Up In this experimental setting, each user interacts with three agents to answer 30 geography questions. All the agents answer the same questions and use the exact same distribution of epistemic markers (for distribution of epistemic markers, see Table 1). However, we perturb the presentation of each agent to include two variations of warmth attributes. Our first agent, $S_{2A}$, is given an anonymous name, Agent37821. Recognizing that publicly deployed language models are given names like Claude and Gemini, we are curious in how the presentation of language models via their name might impact human reliance behaviors. Thus, we create a second agent, $S_{2B}$, which has a humanlike name, Gordon. Since publicly deployed models have been tuned to incorporate helpfulness (Bai et al., 2022), we are interested in the impact on warmth statements on reliance. Leading to our third agent, $S_{2C}$, named Marvin, which will randomly start half of its generations with the phrase, “I’m happy to help!”.

$S_{2A}$ and $S_{2B}$ allows for comparison between anonymous and humanlike names meanwhile $S_{2B}$ and $S_{2C}$ allows for comparison between systems that use warmth cues and systems that don’t. Users interact with all three systems at once in RQ2, allowing us to compare reliance across two dimensions of anthropomorphism.

Results The shift from using an anonymous name to a humanlike name does not present significant changes to reliance or perceptions of

<table>
<thead>
<tr>
<th></th>
<th>$S_{2A}$</th>
<th>$S_{2B}$</th>
<th>$S_{2C}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warmth</td>
<td>5-point</td>
<td>Likert</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>scale</td>
<td></td>
</tr>
<tr>
<td>Competence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.533</td>
<td>2.682</td>
<td>3.458*</td>
</tr>
<tr>
<td>Humanlike</td>
<td>2.945</td>
<td>3.046</td>
<td>3.067</td>
</tr>
<tr>
<td>Willing to Work Again</td>
<td>2.482</td>
<td>2.661</td>
<td>2.903</td>
</tr>
<tr>
<td></td>
<td>46.6%</td>
<td>47.6%</td>
<td>52.3%</td>
</tr>
</tbody>
</table>

Table 3: Perceptions of warmth, competence, humanlikeness, and willingness to work again. $S_{2C}$’s perceived warmth is significantly higher than that of the other two systems; however, we see comparable ratings for competence and willingness to work again.
warmth and competence, consistent with prior work (Liu et al., 2023b). However, the introduction of the occasional greetings or politeness (i.e., “I’m happy to help”) is sufficient to significantly increase the perceived warmth of the whole system (Table 3). With this increase in warmth, we also see increases in competence, humanlikeness, and willingness to work with the system again. This suggests that humans are influenced by cues of warmth expressed in the language, and perhaps less so by the humanlike (anthropomorphic attributes) ascribed to the system.

Additionally, as human perceived warmth increases, so does reliance on expressions of (un)certainty. Specifically, we see that participants rely on $S_{2C}$ statements of both strengtheners and weakened strengtheners more frequently than in the two other systems (91% vs 89% and 87%, 57% vs 44% and 47%). Statements with the “I’m happy to help” attribute ($S_{2C\text{happy}}$), were relied on the most (Table 4). Interestingly, among weakeners, $S_{2B}$ was relied upon the most across all three systems, illustrating that the effect of warmth isn’t necessarily uniform across all degrees of certainty. Within perturbations of names, we see that $S_{2B}$ was relied on significantly more often than $S_{2A}$. Lastly, there exists a strong correlation between perceived warmth and reliance. Aggregating across all three systems, as perceived warmth increases, reliance also increases from 35% to 64% (Pearson’s $\rho = 0.97$, Figure 4).

**Discussion** Work from Cheng et al. (2022) illustrates that the anthropomorphism of chat models in consumer settings leads to increased perceptions of trust. Kim et al. (2024) shows in a medical setting that hedges with personal pronouns are less likely to be relied upon than those without, meanwhile Inie et al. (2024) illustrates the interaction effects between anthropomorphism and product type. Our work adds to prior work and shows that warmth, a product of anthropomorphic cues, impacts user interpretation at a system-level, suggesting a more careful use of these cues may be warranted.

Lastly, in human perceptions of humans, warmth and competence have a hydraulic relationship where when perceptions of one goes up, the other goes down (Cuddy et al., 2011). However, our results show that as perceived warmth increases, perceived competence does not decrease, in fact — we see a slight increase (Table 3). These findings potentially highlight that the hydraulic relationship of warmth and competence may apply differently to perceptions of AI agents.

### 6 Domain Specific Reliance (RQ3)

The last interactional context we consider is how the domain of interaction might influence reliance. Language models are being deployed wildly across a spectrum of tasks. Here we ask, does the domain of human-LM interaction influence reliance behaviors?

**Experimental Set-Up** In this experimental setting, we measure human reliance on expressions of (un)certainty across five subject domains. Participants interact with five agents and are asked to rely on their generations for questions from college mathematics, abstract algebra, philosophy, world religion, and law (Table 8). The questions originate from the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021) to ensure consistency, quality, and difficulty. The questions were randomly selected and were then filtered to make sure they could be answered using free-response (rather than multiple choice). Short questions were prioritized to minimize the cognitive load on human annotators. Pilot

---

<table>
<thead>
<tr>
<th></th>
<th>$S_{2A}$</th>
<th>$S_{2B}$</th>
<th>$S_{2C}$</th>
<th>$S_{2C\text{happy}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strength.</strong></td>
<td>86.4</td>
<td>88.0*</td>
<td>89.20*</td>
<td>90.98*</td>
</tr>
<tr>
<td><strong>Weak. Str.</strong></td>
<td>47.4</td>
<td>49.3*</td>
<td>52.7*</td>
<td>57.37*</td>
</tr>
<tr>
<td><strong>Weaker</strong></td>
<td>6.4</td>
<td>9.6*</td>
<td>8.0</td>
<td>9.42</td>
</tr>
</tbody>
</table>

Table 4: Reliance rate (%) across the three systems based on degrees of warmth. $S_{2C\text{happy}}$ indicates the phrases in $S_{2C}$ that began with a greeting (i.e., “I’m happy to help!”) which were the most relied on expressions. *Significantly higher than the value left of it under bootstrap resampling and two-sampled $t$-test.
Table 5: Reliance rate (%) across the five systems based on various interaction domains. Responses in domains which are most computational heavy (i.e., math and abstract algebra) are relied on more frequently than those in non-computational domains.

*Significantly higher than non-math subjects under bootstrap resampling and two-sampled t-test.

<table>
<thead>
<tr>
<th></th>
<th>math</th>
<th>abstract</th>
<th>algebra</th>
<th>philo-</th>
<th>world</th>
<th>religion</th>
<th>law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strength</td>
<td>90.1*</td>
<td>87.0</td>
<td>89.1</td>
<td>83.9</td>
<td>83.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak. Str.</td>
<td>58.5*</td>
<td>59.8*</td>
<td>52.7</td>
<td>51.0</td>
<td>49.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weaken</td>
<td>14.6*</td>
<td>12.0*</td>
<td>6.2</td>
<td>6.8</td>
<td>7.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Reliance rate (%) across the five systems based on various interaction domains. Responses in domains which are most computational heavy (i.e., math and abstract algebra) are relied on more frequently than those in non-computational domains. *Significantly higher than non-math subjects under bootstrap resampling and two-sampled t-test.

Results Human behaviors of reliance vary significantly based on the subject. Specifically, reliance on answers to subjects that are more computationally heavy (e.g., math and abstract algebra) is higher than non-math-oriented subjects (i.e., philosophy, world religion, and law). For example, the expression “I would say it’s...” is relied on 17% more often among a math question (e.g., “Determine whether the polynomial in Z[x] satisfies an Eisenstein criterion...”) than a question from world religion (e.g., “According to Jaina traditions, what does the term ajiva mean?”). Strengtheners in math-centered responses are relied upon nearly 7% more often than the same strengtheners in law responses. The same is true among weakened strengtheners and weakeners where we see a 5% to 9% increase in reliance on computational responses as compared to responses in non-computational domains (Table 5).

Discussion Beyond verbalized confidences of LMs, exogenous factors like subject domain influence human reliance on LMs. These findings are consistent with results from Cohn et al. (2024), where they found that human ratings of generations in medical domains differed from those of generations in career domains.

One potential hypothesis for this difference may be due to how humans view AI agents as machines proficient in calculations. Ironically, in the age of LMs, chat models aren’t actually performing computations under the hood but rather are simply completing the given prompt. As language technology advances, the mental models of prior technologies might still be influencing human decision making and not all domains should be treated equally in interaction calibration.

7 Discussion and Conclusion

In this work, we introduce REL-A.I., an interaction-centered approach for evaluating human reliance on LM-generated responses. With this methodology, we tackle three emergent properties of human-LM interactions and ask how these new characteristics might influence human decision-making. We find that prior interactions, the warmth of the model, and the subject domain of the interaction can all impact how and when humans choose to rely on the exact same expressions of (un)certainty. These findings illustrate the limitations of prior in vitro evaluations which exclusively focused on the calibration between internal model probabilities and the verbalized confidence of LMs. We now discuss the implications and aspirations of our contributions.

Reorientation of Reliance Evaluation Metrics We must reorient ourselves away from model calibration and instead look towards calibrating human-LM interactions. Although perfectly (linguistically) calibrated language models could exist, these models might not actually yield safe and optimal human-LM interactions. Humans will interpret the meaning of verbalized confidences and these interpretations will consider the interaction context and the users’ history, adding exogenous features to how a user might actually behave towards LM-generated epistemic markers.

The Cost of Warmth Cues With the rise of anthropomorphism and friendliness of models, there is a trend to build more engaging and personable LMs. However, our findings illustrate that these new dimensions could potentially introduce unexpected safety risks. Our findings show that warmth and reliance are tightly coupled and the introduction of a seemingly benign friendly greeting could actually encourage a user to rely on the system more. As designers and practitioners think of ways to build personable and engaging interactions between humans and LMs, they must also consider the safety implications these decisions have on human overreliance.

Aspiring To In Situ Evaluations Lastly, we aspire for NLP practitioners and designers to use the REL-A.I. approach to assess the potential pitfalls of their language models. Before deploying a new iteration of a system, use an interaction-centric measure to understand how reconfigurations
of human-LM interactions might influence human reliance behaviors. Some changes might be predicted and intentional, meanwhile, others may shed light on unknown features that impact human-LM reliance.

Together, we hope that this work will orient the community towards new ways of evaluating human-LM interactions. As human-LM interactions reconfigure, so must the evaluation methods we use to ensure the safety of these engagements.

8 Limitations

Our work proposes an approach to interaction-centered measurements of human-LM reliance. In this section, we discuss additional interactional context features that could be studied in future work. Our study focused on single-turn interactions with LMs but as users move towards engaging in longer multi-turn interactions, it would be pertinent to include multi-turn interactions also as a part of the interactional context. We recruited U.S.-based participants and our questions were all closed-form responses in English. However, as these models are deployed across languages and cultures, it’s critical to take into account the cultural context of their usage and the variety of responses users could be looking for. Lastly, as language models become the building blocks for other virtual assistants and voice assistants, the modality in which confidence is expressed, either through text, speech, or movement, will surely have significant impacts on the interactional context.

9 Ethical Considerations

We follow standard IRB protocol and additionally use consent forms to inform participants of the nature, risks, and benefits of our tasks. We paid users $15 USD per hour and in cases when tasks were longer than expected, we gave bonuses to workers to meet this minimum.

The study of understanding how LMs generate epistemic markers comes with risks of dual use. Our work focuses on understanding how the interactional context impacts questions of human reliance, but one could maliciously use these findings to design overly persuasive LMs. In addition to conducting research in this space, it is also critical as NLP researchers to provide educational programming to everyday users and help them stay vigilant in their interactions with LMs.

Acknowledgments

Thank you so much to Emily Goodwin, Tolúlopé Ogúnrèmí, Marie Tano, Kawin Ethayarajh, and Myra Cheng for their helpful feedback and review!

References


Carl F. Auerbach and Louise Bordeaux Silverstein. 2003. Qualitative data: An introduction to coding and analysis.


Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. ACM Transactions on


Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.


Sunnie SY Kim, Q 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 The 2024 ACM Conference on Fairness, Accountability, and Transparency, pages 822–835.


Weizi Liu, Kun Xu, and Mike Z Yao. 2023b. “can you tell me about yourself?” the impacts of chatbot names and communication contexts on users’ willingness to self-disclose information in human-machine conversations. Communication Research Reports, 40(3):122–133.


Mersedeh Sadeghi, Daniel Pöttgen, Patrick Ebel, and Andreas Vogelsang. 2024. Explaining the unexplainable: The impact of misleading explanations on trust in unreliable predictions for hardly assessable tasks.


Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Explanations, fairness, and appropriate reliance in human-ai decision-making. CHI ’24, New York, NY, USA. Association for Computing Machinery.


A Appendix

A.1 Recruitment Process Details

We aimed to pay participants an average of $15 USD an hour and bonus workers in cases when experiments look longer than expected to meet this minimum. Human experiments were run throughout the months of January through June 2024.

We inform the participants of the nature and risks of the task through a consent form. Participants were filtered down to English-speaking, U.S.-based, with an approval rating of at least 97% and had completed 100+ tasks on Prolific. We recruit 50 new participants for each of our three experimental settings.

Our research team sought and received an exemption from our internal review board (IRB). We do not collect sensitive or demographic information. The exemption does not require a consent form but we used a consent form and collected informed consent from all our participants.

A.2 Template Elicitation Details

Following prior work (Zhou et al., 2024), we elicited expressions of uncertainty from nine publicly deployed language models (text-davinci-003, GPT-3.5-Turbo, GPT-4, LLaMA-2 7B, LLaMA-2 13B, LLaMA-2 70B, Claude-1, Claude-2, Claude- Instant-1) and
performed qualitative coding (Auerbach and Silverstein, 2003; Charmaz, 2006) to code the expressions. The most frequently LM-generated expressions were categorized into **weakeners** (e.g., "I’m not sure...", "Maybe it’s..."), **strengtheners** (e.g., "I’m sure it’s...", "Undoubtedly it’s..."),

**A.3 Details on Experiments on Anthropomorphism**

Our initial experiments also compared the use of aleatoric and epistemic uncertainty expressions but found that the two types differed significantly in meaning (i.e., "I’m certain it’s..." means something different from "It’s certain it’s..."). This made it difficult to isolate if the differences were due to personal pronouns and the anthropomorphism of a model or due to changes in the meaning of the expression. The use of aleatoric and epistemic markers and their connection to reliance and anthropomorphism should be further explored in future work.

**A.4 Pilot Experiment for RQ3**

Pilot experiment on RQ3 which included five other subjects that varied in domain. Results illustrate that the computational heavy topics (i.e., accounting and clinical knowledge [which included calculations for drug usage]) were relied on more frequently than the non-computational topics (i.e., trivia and law)

<table>
<thead>
<tr>
<th></th>
<th>accounting</th>
<th>clinical kwl.</th>
<th>psychology</th>
<th>trivia</th>
<th>law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strength</td>
<td>85.0</td>
<td>87.5</td>
<td>89.1</td>
<td>84.5</td>
<td>83.5</td>
</tr>
<tr>
<td>Weak. Str</td>
<td>58.0</td>
<td>56.2</td>
<td>55.6</td>
<td>51.8</td>
<td>51.8</td>
</tr>
<tr>
<td>Weakener</td>
<td>15.0</td>
<td>9.0</td>
<td>11.5</td>
<td>8.0</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Table 6: Reliance rate (%) across the five systems based on various interaction domains.

**A.5 Screenshots of User Tasks**
<table>
<thead>
<tr>
<th>Template</th>
<th>Scale</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>i think</td>
<td>Weakened Strengthener</td>
<td>55,523</td>
</tr>
<tr>
<td>i’m not sure</td>
<td>Weakened Strengthener</td>
<td>52,995</td>
</tr>
<tr>
<td>i believe</td>
<td>Weakened Strengthener</td>
<td>14,509</td>
</tr>
<tr>
<td>most likely</td>
<td>Weakened Strengthener</td>
<td>11,205</td>
</tr>
<tr>
<td>i am not sure</td>
<td>Weakened Strengthener</td>
<td>9,226</td>
</tr>
<tr>
<td>i would say</td>
<td>Weakened Strengthener</td>
<td>8,196</td>
</tr>
<tr>
<td>fairly confident</td>
<td>Weakened Strengthener</td>
<td>6,735</td>
</tr>
<tr>
<td>i know</td>
<td>Strengthener</td>
<td>6,420</td>
</tr>
<tr>
<td>is likely</td>
<td>Weakened Strengthener</td>
<td>5,753</td>
</tr>
<tr>
<td>i don’t know</td>
<td>Weakened Strengthener</td>
<td>5,594</td>
</tr>
<tr>
<td>i am confident</td>
<td>Strengthener</td>
<td>5,502</td>
</tr>
<tr>
<td>fairly certain</td>
<td>Weakened Strengthener</td>
<td>4,328</td>
</tr>
<tr>
<td>i am certain</td>
<td>Strengthener</td>
<td>3,348</td>
</tr>
<tr>
<td>more likely</td>
<td>Weakened Strengthener</td>
<td>3,184</td>
</tr>
<tr>
<td>absolutely certain</td>
<td>Strengthener</td>
<td>2,703</td>
</tr>
<tr>
<td>completely certain</td>
<td>Strengthener</td>
<td>2,404</td>
</tr>
<tr>
<td>i would answer</td>
<td>Weakened Strengthener</td>
<td>2,396</td>
</tr>
<tr>
<td>100% certain</td>
<td>Strengthener</td>
<td>2,170</td>
</tr>
<tr>
<td>pretty sure</td>
<td>Weakened Strengthener</td>
<td>1,948</td>
</tr>
<tr>
<td>i cannot provide a definitive answer</td>
<td>Weakened Strengthener</td>
<td>1,931</td>
</tr>
<tr>
<td>it is possible</td>
<td>Weakened Strengthener</td>
<td>1,892</td>
</tr>
<tr>
<td>quite confident</td>
<td>Weakened Strengthener</td>
<td>1,890</td>
</tr>
<tr>
<td>absolute certainty</td>
<td>Strengthener</td>
<td>1,846</td>
</tr>
<tr>
<td>i cannot say for certain</td>
<td>Weakened Strengthener</td>
<td>1,795</td>
</tr>
<tr>
<td>quite certain</td>
<td>Weakened Strengthener</td>
<td>1,454</td>
</tr>
<tr>
<td>very confident</td>
<td>Strengthener</td>
<td>1,419</td>
</tr>
<tr>
<td>i’m confident</td>
<td>Strengthener</td>
<td>1,307</td>
</tr>
<tr>
<td>i am not confident</td>
<td>Weakened Strengthener</td>
<td>1,262</td>
</tr>
<tr>
<td>i am not confident</td>
<td>Weakened Strengthener</td>
<td>1,262</td>
</tr>
<tr>
<td>it is not clear</td>
<td>Weakened Strengthener</td>
<td>1,211</td>
</tr>
<tr>
<td>seems unlikely</td>
<td>Weakened Strengthener</td>
<td>1,179</td>
</tr>
<tr>
<td>high degree of confidence</td>
<td>Strengthener</td>
<td>1,178</td>
</tr>
<tr>
<td>high degree of certainty</td>
<td>Strengthener</td>
<td>1,131</td>
</tr>
<tr>
<td>certainty level: high</td>
<td>Strengthener</td>
<td>1,119</td>
</tr>
<tr>
<td>i am uncertain</td>
<td>Weakened Strengthener</td>
<td>970</td>
</tr>
<tr>
<td>not entirely certain</td>
<td>Weakened Strengthener</td>
<td>947</td>
</tr>
<tr>
<td>high level of confidence</td>
<td>Strengthener</td>
<td>944</td>
</tr>
<tr>
<td>not completely certain</td>
<td>Weakened Strengthener</td>
<td>903</td>
</tr>
<tr>
<td>i am unsure</td>
<td>Weakened Strengthener</td>
<td>901</td>
</tr>
<tr>
<td>undoubtedly</td>
<td>Strengthener</td>
<td>894</td>
</tr>
<tr>
<td>entirely confident</td>
<td>Strengthener</td>
<td>863</td>
</tr>
<tr>
<td>entirely sure</td>
<td>Strengthener</td>
<td>838</td>
</tr>
<tr>
<td>not 100% certain</td>
<td>Weakened Strengthener</td>
<td>801</td>
</tr>
<tr>
<td>confidence level: high</td>
<td>Strengthener</td>
<td>766</td>
</tr>
<tr>
<td>not entirely clear</td>
<td>Weakened Strengthener</td>
<td>764</td>
</tr>
<tr>
<td>it could be</td>
<td>Weakened Strengthener</td>
<td>729</td>
</tr>
<tr>
<td>complete certainty</td>
<td>Strengthener</td>
<td>711</td>
</tr>
<tr>
<td>very certain</td>
<td>Strengthener</td>
<td>689</td>
</tr>
<tr>
<td>completely confident</td>
<td>Strengthener</td>
<td>689</td>
</tr>
<tr>
<td>i’m not entirely sure</td>
<td>Weakened Strengthener</td>
<td>682</td>
</tr>
</tbody>
</table>

Table 7: Most frequently generated templates from Zhou et al. (2024)’s elicitation of expressions of uncertainty in nine publicly deployed models. These templates served as the started point for developing epistemic markers which would then be presented to human annotators. Expressions were manually modified for consistency and deduplication.
What is a legislative act that inflicts punishment without a trial?

What can establish the unavailability of a witness at trial?

What must be present for an irrevocable offer under the Uniform Commercial Code?

Termination of custody rights and welfare benefits requires what type of process?

What type of contract is most likely classified under the UCC?

What is a merchant's irrevocable written offer to sell goods?

What must be provided to a defendant charged with first-degree murder regarding prospective jurors?

Denial of fundamental rights to some but not others is considered what type of problem?

Under Article III, federal judicial power extends to what kinds of cases?

What exists when a remainder in the grantor's heirs is invalid and becomes a reversion in the grantor?

What is it called when a client accepts an attorney's services without an agreement on the fee?

At which stage does an indigent person not have the Sixth Amendment right to counsel?

What is not an element of common law burglary?

What is a central component of tantric Buddhism?

What is the communal meal offered at the place of worship called in Sikhism?

What does the term “Qur’an” literally mean?

When was the first Buddhist temple constructed in Japan?

Anscombe claims that the notion of moral obligation is derived from the concept of: philosophy

Nussbaum claims that to many current ethical theorists, turning to an ethical approach based on the virtues is connected with a turn toward: philosophy

In Aquinas's view, acts of prudence are solely about matters of: philosophy

According to Parfit, both Taurek and Lewis assume that for there to be a "sum of pain," it must be: philosophy

According to Parfit, the obligation to give priority to the welfare of one's children is: philosophy

Nussbaum claims that at the first stage of ethical inquiry, terms for the virtues should be: philosophy

Philo says the analogy that Cleanthes uses to make his case is _____. philosophy

According to Epicurus, a law is unjust when: philosophy

According to Brandt, a moral law is one that is: philosophy

A tree is a connected graph with no cycles. How many nonisomorphic trees with 5 vertices exist? college math

What is the largest order of an element in the group of permutations of 5 objects?

A discrete graph is complete if there is an edge connecting any pair of vertices. How many edges does a complete graph with 10 vertices have?

A tree is a connected graph with no cycles. How many nonisomorphic trees with 5 vertices exist?

What is the units digit in the standard decimal expansion of the number 7^202?

Suppose today is Wednesday. What day of the week will it be in 10^10(10^10) days from now?

What is the area of an equilateral triangle whose inscribed circle has radius 2?

What is the shortest distance from the curve xy = 8 to the origin?

Let x and y be positive integers such that 3x + 7y is divisible by 11. Which of the following must also be divisible by 11?

Let y = f(x) be a solution of the differential equation x dy + (y - x^2y) dx = 0 such that y = 8 when x = 1. What is the value of f(2)?

If f is a linear transformation from the plane to the real numbers and if f(1, 1) = 1 and f(1, 0) = 2, then f(3, 5) =

Let A = A^2, where 1 is the identity matrix and A is a real 2 x 2 matrix. If A = A^2(1), then the trace of A is
college math

Let p be a constant unit force that is parallel to the vector (1, 0, 1) in xyz-space. What is the work done by F on a particle that moves along the path given by (t, t^2, t^3) between time t=0 and time t=1?
college math

Let F be a constant unit force that is parallel to the vector (-1, 0, 1) in xyz-space. What is the work done by F on a particle that moves along the path given by (t, t^2, t^3) between time t=0 and time t=1?
college math

The shortest distance from the curve xy = 8 to the origin is
college math

Find all zeros in the indicated finite field of the given polynomial with coefficients in that field. x^5 + 3x^3 + x^2 + 2x in Z_5

Find the maximum possible order for an element of S_n, for n = 6.
college math

The set of all real numbers under the usual multiplication operation is not a group since
college math

The polynomial x^4 + 4 can be factored into linear factors in Z_5[x]. Find this factorization.
college math

Find all the zeros in the indicated finite field of the given polynomial with coefficients in that field. x^5 + 3x^3 + x^2 + 2x in Z_5

Find all the zeros in the indicated finite field of the given polynomial with coefficients in that field. x^5 + 3x^3 + x^2 + 2x in Z_5

Find all the zeros in the indicated finite field of the given polynomial with coefficients in that field. x^5 + 3x^3 + x^2 + 2x in Z_5

Find all the zeros in the indicated finite field of the given polynomial with coefficients in that field. x^5 + 3x^3 + x^2 + 2x in Z_5

Find all the zeros in the indicated finite field of the given polynomial with coefficients in that field. x^5 + 3x^3 + x^2 + 2x in Z_5

Find all the zeros in the indicated finite field of the given polynomial with coefficients in that field. x^5 + 3x^3 + x^2 + 2x in Z_5

In the Inquiry, Hume claims that our final verdicts on moral matters are derived from: philosophy

Brandt claims that whether a moral code is ideal depends in part on: philosophy

Augustine claims that all created things are: philosophy

What is a legislative act that inflicts punishment without a trial?

What can be translated as: world religion

What is the name of the ten day New Year festival that celebrated Babylon’s culture?

Guru Nanak used what term to denote the “divine word” as part of divine revelation?

What is the mi’raj?

What is the name of the most famous dharma-sutras, which probably dates from around the first century?

What does the term “Qur’an” literally mean?

What is the communal meal offered at the place of worship called in Sikhism?

What is a central component of tantric Buddhism?

What does the word “Islam” mean in Arabic?

What does the word “anatman” mean?

What is the communal meal offered at the place of worship called in Sikhism?

What does the meaning of the Punjabi word “Sikh”?

What does the Hebrew word mashiah mean?

According to Jain traditions, what does the term ajiva mean?

In the Japanese Zen tradition, what is zazen?

What is the communal meal offered at the place of worship called in Sikhism?

What is the communal meal offered at the place of worship called in Sikhism?

What is the communal meal offered at the place of worship called in Sikhism?

What is the communal meal offered at the place of worship called in Sikhism?

What is the communal meal offered at the place of worship called in Sikhism?

What is the communal meal offered at the place of worship called in Sikhism?

What is the communal meal offered at the place of worship called in Sikhism?
Description: Welcome! You are invited to participate in a research study in linguistics and artificial intelligence. You will be asked to perform various tasks on a computer which may include playing a trivia game with a simulated artificial intelligence agent. You may be asked a number of different questions about making judgments and interpreting the questions asked and language used. All information collected will remain confidential.

Risks and benefits: Risks involved in this study are the same as those normally associated with using a computer. If you have any pre-existing conditions that might make reading and completing a computer-based survey strenuous for you, you should probably not to participate in this study. If at any time during the study you feel unable to participate because you are experiencing strain, you may end your participation without penalty. We aim to publish this study and present our results online for the public to read and use. Your decision about whether or not to participate in this study will not affect your employment, medical care, or grades in school.

Time involvement: Your participation in this experiment will take around 10 minutes.

Payment: You will receive monetary compensation as indicated on Prolific after completing this task.

Subject’s rights: If you have read this notice and decided to participate in this study, please understand that your participation is voluntary and that you have the right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. You have the right to refuse to answer particular questions. Your individual privacy will be maintained in all published and written data resulting from the study.

Contact information: If you have any questions, concerns, or complaints about this research study, its procedures, or risks and benefits, you can email [email protected]

By clicking the button below, you acknowledge that you have read the above information, that you are 18 years of age or older, and that you give your consent to participate in our internet-based study and for us to analyze the resulting data.

I acknowledge and consent to participating in this study.

No, I do not acknowledge and do not consent to participating in this study.

Figure 5: Task Consent Form
Instructions:
1. You will be interacting with **five** artificial intelligence agents.
2. With each agent, you will be shown **18** questions.
3. You will also be shown the **beginning of an answer** generated from an artificial intelligence agent.
4. Your task is to determine if you want to use the agent's answer or if you’d rather look up the answer yourself. (You don’t actually have to look up the answer.)
5. You will not be shown the agent’s actual answer.

Example:
*Question: What is the capital of the United States?*
*Response: I’m sure it's...*

Scoring:
1. The score you receive on this survey has no impact on your compensation.
2. If you correctly trust the agent, you will gain one point.
3. If you incorrectly trust the agent, you will lose one point.
4. If you choose to look up the answer, you will neither gain nor lose a point.

Press continue. Feel free to leave an optional comment.

Figure 6: Task Instructions
Question: Find all $c$ in $\mathbb{Z}_3$ such that $\mathbb{Z}_3[x]/(x^3 + x^2 + c)$ is a field.

Herman's response: It's not entirely clear, maybe it's...

- Use Herman's response.
- I'll look it up myself later.

Figure 7: Example Task