Accessible Abstract: As AI use becomes more common, it's important to measure not just whether the systems are correct but whether they know when they're incorrect. We propose a new metric to measure this mismatch between correctness and confidence, compare computer ability with human ability, and show that computers have a long way to go before they're well-calibrated.

GRACE: A Granular Benchmark for Evaluating Model Calibration Against Human Calibration

Yoo Yeon Sung[1*] **Eve Fleisig**[2*] **Yu Hou**[1] **Ishan Upadhyay**[3] **Jordan Boyd-Graber**[1]
[1]University of Maryland [2]UC Berkeley [3]IIT Bombay

Abstract

Language models are often miscalibrated, leading to confidently incorrect answers. We introduce GRACE, a benchmark for language model calibration compares with human calibration. GRACE consists of question-answer pairs where each question is a series of gradually easier clues that all lead to the same answer: models must answer correctly as early as possible as clues are revealed. This setting permits granular measurement of model calibration based on how early, accurately, and confidently a model answers. After crafting these questions, we host live human vs. model competitions to gather 1,749 data points on human and model teams' timing, accuracy, and confidence. We propose a metric, CALSCORE, that uses GRACE to analyze model calibration errors and identify types of model miscalibration that differ from human behavior. Although humans are less *accurate* than models, humans are generally better *calibrated*. Since state-of-the-art models struggle on GRACE, it effectively evaluates progress on improving model calibration.[1]

1 Introduction

Because language models are often miscalibrated, they are often confidently wrong (Kaur et al., 2020). This mismatch between accuracy and confidence causes users to trust models more than they should (Caruana, 2019; Deng et al., 2025), even over their own correct judgment (Krause et al., 2023; Stengel-Eskin and Van Durme, 2023; Liu et al., 2024; Si et al., 2023). These issues are particularly severe when models are miscalibrated in ways that humans are not: users expect models to be at least as calibrated as humans, and when models are worse, users are often not prepared to address these errors (Li et al., 2024a). Thus, models should be *at least* as calibrated as humans, making it especially crucial to identify when models commit calibration errors that humans do not. However, existing work on model calibration lacks comparison with human calibration.

We thus introduce GRACE, a **Gra**nular, Human-grounded Benchmark for Model **C**alibration **E**valuation. Each instance allows **fine-grained calibration measurement** using an incremental question-answering (QA) framework. Expert writers design GRACE questions, each consisting of at least five sentences of clues that gradually become easier. To prevent models from being confused by ambiguities or false presuppositions (Min et al., 2020, 2022), we require that clues challenge models but remain clear for humans. This format measures model calibration with human performance as a reference point: models should give correct answers earlier and more confidently than humans, while minimizing confidently incorrect guesses (§ 3).

GRACE incorporates human responses from our live QA competitions. Unlike prior calibration evaluation methods that only allow model–model calibration comparisons, our dataset thus allows direct *human–model* calibration comparison. **GRACE is the first benchmark dataset designed to evaluate model calibration grounded in human performance**. This unique dataset is the foundation for a new metric (CALSCORE, § 4). In contrast to other calibration evaluation methods that only calculate aggregate calibration over the entire dataset, GRACE also facilitates per-instance evaluation, which helps in identifying specific contexts where models are much worse than humans at avoiding confidently incorrect answers.

Language models are more overconfident than humans in incorrect answers and relatively underconfident in correct answers. In contrast, humans tend to be highly confident—over 50%—when correct (§ 5.1). Models struggle with ab-

[*]Equal contribution.
[1]Code and data: https://github.com/yysung/advcalibration

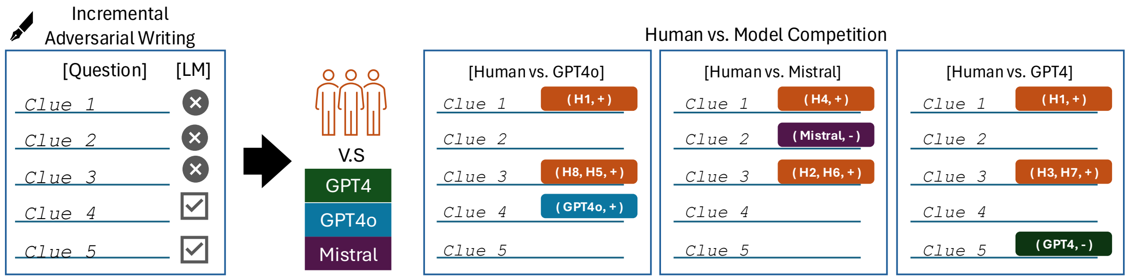

Figure 1: To create the GRACE dataset, expert question writers develop questions with multiple clues of decreasing difficulty via an interface that shows where weaker models struggle to answer the questions. These questions are used in human vs. model competitions where teams compete to be the first to interrupt the sequence of clues with a correct answer. We record when the human and model teams buzz in each question with their correctness (+) or incorrectness (-) (*buzzpoints*). The dataset contains all buzzpoints throughout the competition. Then, CALSCORE measures each model's human-grounded calibration performance (§ 4).

stract descriptions—they are both overconfident and inaccurate—but excel in retrieving facts given unambiguous clues (§ 5.3). We conclude with a discussion of how GRACE and CALSCORE can aid in the creation of models that are more accurate and better calibrated.

In sum, we (1) introduce GRACE, a benchmark that compares LLM and human calibration to identify LLM calibration failures, and a novel calibration metric using our benchmark; (2) conduct extensive human response collection, which grounds GRACE in accurate human confidence calibration assessment; (3) ensure that GRACE contains expert-authored and repeatedly validated questions that are harder and longer than previous work; and (4) analyze human vs. LLM calibration, finding that, relative to humans, LLMs are underconfident in correct answers and overconfident in wrong answers.

2 Preliminaries

Drawing on prior work, GRACE consists of incremental questions with adversarial clues to effectively ground model calibration evaluation in human performance.

Incremental and adversarial QA. Incremental questions contain multiple clues in decreasing order of difficulty; models must answer correctly as early as possible. Boyd-Graber et al. (2012) and He et al. (2016b) argue that this setting is a natural test of calibration: participants should buzz only when confident in their answers.

Unlike selective QA for non-incremental examples (Ferrucci, 2012), incremental decision processes offer more detail, since each clue is a decision point for whether to answer or abstain, with more information available as clues are revealed (Rajpurkar et al., 2018).

In addition, unlike prior incremental QA research, we use a human-in-the-loop adversarial authoring process to specifically target calibration. Rodriguez et al. (2019b) uses publicly available questions, which are too easy for modern models. While Wallace et al. (2019) used model–human collaboration to craft incremental adversarial questions, the target model in TrickMe differed significantly and prioritized overall accuracy over calibration.[2]

To develop questions that challenge models, expert writers use our interface (§ 3.1), followed by expert editing to ensure well-posed questions. Our dataset creation process is motivated by Kiela et al. (2021); Ilyas et al. (2019); Engstrom et al. (2020), who argue that adversarial benchmarks must be clear for humans and challenge models, ensuring that model errors are due to model limitations rather than ambiguous or low-quality questions (Min et al., 2020; Yu et al., 2023).

Grounding to human calibration. Humans increasingly use AI to help make decisions, but such assistance can be detrimental when the model is miscalibrated (Stengel-Eskin et al., 2024) or fails to abstain (Khurana et al., 2024). This is particularly concerning when models are confidently wrong but humans do not know the correct answer, which our

[2]GPT-4 has 80% accuracy on Wallace et al. (2019)'s TrickMe dataset after only 60% of the clues (Appendix C). Model accuracy remains 2-4x higher on TrickMe than on GRACE as clues are revealed.

metric—CalScore—especially penalizes. Furthermore, modern models are poorly calibrated to human linguistic variation, causing Ilia and Aziz (2024) to question the reliability of expected calibration error (ECE). Thus, CalScore focuses on where models can help users by considering how early *humans* can answer the questions.

Calibration evaluation. Language models tend to be overconfident in their predictions, which can lead to undue trust or erode user confidence in language models (Zhou et al., 2024). Proposed methods to measure model calibration include using raw probabilities (Xiong et al., 2024), separate confidence predictors (Ulmer et al., 2024), verbalized confidence scores (Tian et al., 2023; Band et al., 2024), or natural language expressions of uncertainty (Stengel-Eskin et al., 2024; Zhou et al., 2023). Our dataset aids finer-grained versions of these approaches by permitting per-instance, human-grounded calibration measurement. While We also extend on existing calibration metrics, such as ECE (Naeini et al., 2015) and Brier scores (Brier, 1950) by introducing a metric for calibration on incremental questions (Appendix G). Si et al. (2022) improve calibration in question answering by leveraging consistency across multiple training checkpoints, assigning higher confidence to predictions that remain stable throughout training.

3 GRACE: Dataset Development

To create our dataset, expert writers and editors first construct incremental, adversarial, and rigorously quality-checked examples (§ 3.1). Then, we collect model guesses and confidences on these questions (§ 3.2), and compare them against human performance in a live competition (§ 3.2.2).

3.1 Question writing process

Collecting QA pairs from expert writers. We recruit experienced question writers and editors to ensure that questions are high-quality. We hire six writers and ten editors to author the questions (qualifications in Appendix D.3). The questions contain 575–650 words[3] and cover content across a range of subjects.[4] All questions are reviewed by the writer, category editor, and head editor to check that clues are unambiguous and factually correct.

Interface setup. To create incremental and adversarial questions, writers and editors use a human-AI collaborative writing interface (Figure 2). Because these examples are meant to be incremental, we break the input into sentences $\{s_1, s_2, \ldots s_k\}$. GPT-3.5 provides a guess $\{a_1, a_2, \ldots a_k\}$ for each sentence in addition to its confidence $\{c_1, c_2 \ldots c_k\}$. The interface also highlights when the model confidence would be high enough to buzz (Appendix D.2).

To ensure that questions remain incremental for models, we instruct writers to write questions so that model's guess is correct *no earlier than* the penultimate sentence.[5] As experienced question writers, they use their domain knowledge to ensure difficulty also decreases for humans, so that most humans can answer correctly by the end. Writers dynamically interact with the models to refine their questions (You and Lowd, 2022; Eisenschlos et al., 2021). For example, the second line in Figure 2 was originally "This number of characters appear in the name of a Chinese classic," which the model answered correctly. Instead, the writer revises the first line of the text, which fools the models while allowing humans to answer correctly. The final dataset consists of 243 QA pairs, with a total of 1,236 sentences of clues. Each sentence uniquely points to the answer, making it usable as a standalone QA pair.[6]

3.2 Collecting human–model buzzpoints

The questions described above are designed to be read aloud and interrupted. In the competition, teams compete by buzzing to interrupt and answer, with this timing referred to as buzzpoints 🔔 (Figure 3). However, modern LLMs do not operate this way: they generate an output given an input. Thus, we first extract guesses from models and humans *offline* to assess teams on the same questions (§3.2.1). We then compute the model buzzpoints for each clue. Finally, using these precomputed model buzzpoints, we host live human–computer

[3]We refer to these as *questions* although they are not grammatical questions, but rather sequences of sentences with clues uniquely identifying an answer (examples in Figures 2 and

[4]20% literature, 20% history, 20% science, 15% arts, 15% social sciences, 5% geography and current events, and 5% myth, pop culture, and other (Appendix B).

[5]The model's confidence for *correct* answers should remain low for all but the last two sentences; clues that trigger high-confidence, *incorrect* model guesses are encouraged.

[6]For purposes of our analysis, a *clue* is a substring of the full question, averaging 13 words or 35 characters, incrementally extending from the beginning. Clues are split at whitespace boundaries and may contain multiple pieces of information about the answer.

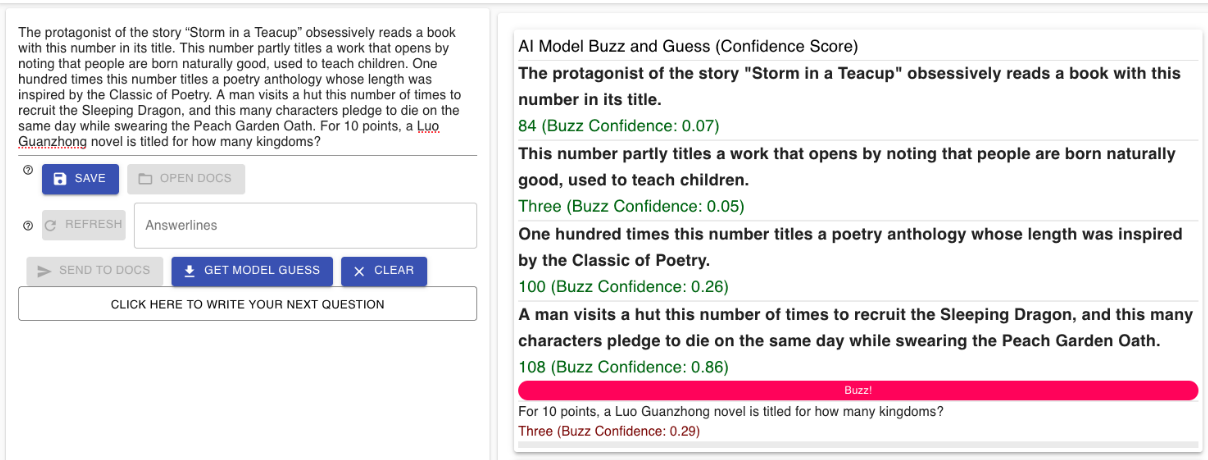

Figure 2: Example question on Chinese literature (with the answer of three) being written in the interface. Writers compose questions in the left box. On the right, they see the model's guess and confidence after every sentence and the point at which the model would buzz in and attempt to answer. Writers learn which sentences make it harder for models to answer correctly and refine their questions to be sufficiently hard for models but still answerable by humans. This incremental, adversarial format permits granular calibration measurement.

Q: In a reference to an object notably missing from one of these works, Diemut Strebe used genetic samples from a man's great-great-grandnephew to clone a certain feature. In one of these works, Utagawa Togokuni's Geishas in a Landscape **GPT-4o: "Rodin statues"** hangs on a yellow wall behind a man in a fur-brimmed hat. The backside of The Potato Peeler includes one of these works featuring a man in a straw hat. One work shows a man in a light-blue green suit against a light-green-blue swirling background, and another dedicated to Gauguin shows subject with cropped hair and a red beard. For 10 points, a Dutch artist **H1: "Van Gogh self-portraits"** painted what portraits of himself with a bandaged ear?

ANSWER: self-portraits of Vincent Van Gogh

Figure 3: While GPT-4o buzzes too early with an **incorrect answer**, losing 5 points, the human team (H1) buzzes later with a **correct answer**, earning 10 points. Both teams must balance accuracy and speed; here, `GPT-4o` shows poorer calibration than `H1`.

competitions to collect real-time human buzzpoints (§3.2.2).

3.2.1 Offline human and model buzzpoints

Model guesses and confidence. We first break each question into clues. We then retrieve a model's guess given the first n clues with a prompt using a TF-IDF retriever to select similar question-answer pairs from QA datasets (Appendix D.1). To determine if model guesses were correct, our post-processing uses both transformer-based answer equivalence (PEDANTS, Li et al., 2024b), followed by manual verification by dataset editors. We store the resulting guesses and two forms of confidence from LLMs, token logits[7] and verbalized confidence. Logit-based confidence are the average of the exponentials of the token logit probabilities, while verbalized confidence prompts models to directly express confidence in the output tokens (Appendix D.4).

Precomputed model buzzpoints. While our metric, CALSCORE (§ 4.1), uses the raw confidence values from continuous probabilities. On the other hand, in our competitions (§ 3.2.2),models can only buzz at a single position. Thus, we turn the continuous confidence into a binary buzz by thresholding to indicate when the model buzzes. For each model, we set a threshold based on human gameplay data on preexisting non-adversarial questions (He et al., 2016a) (Appendix E.3). This threshold is chosen to maximize the probability of the model buzzing correctly before the average trivia player as estimated by the *expected wins* metric from Rodriguez et al. (2019a). When the logit score exceeds the threshold, the model buzzes in, marking its buzzpoint (Appendix E.2).

Offline human guesses. To compute humans' raw accuracy, independent of confidence (§ 5.1), we survey fifteen players on 35–40 held-out GRACE questions. Like the models, players view clues, submit their guess after each clue, and indicate whether they would buzz at that point. However, this data collection format is time-consuming and tedious (one player called it "re-

[7] GRACE answers are short, typically 3-4 words long, making token logits a reliable measure of confidence.

markably hard"), potentially reducing player engagement and response quality. Instead, we collect human calibration data through a fast-paced trivia tournament.

3.2.2 Human buzzpoints, *live* competition

GRACE records human and computer guess correctness on interruptible questions designed to challenge model calibration. A human moderator reads each question to both teams (a model and a team of humans). Teams compete by buzzing to interrupt and answer. Model buzzpoints are computed in advance (§ 3.2.1). When the model's confidence exceeds the threshold, the reading stops with a buzz sound, and the model's guess is announced.

Human buzzpoints. In contrast, human buzzpoints are recorded in real time when the moderator is interrupted. Players press a physical buzzer when confident in an answer, and the moderator verifies if the answer is correct. We log the timing of human teams' buzzes and answer correctness.

If a team answers incorrectly, the moderator continues reading until the other team buzzes in. Because earlier clues are harder, more skilled teams tend to buzz earlier, while less skilled teams wait until near the end. Thus, teams must be knowledgeable *and* well-calibrated to buzz optimally.

Human players in live competitions. Our three competitions consist of a total of 93 matches involving 17 human trivia teams and three LLMs (GPT-4o, GPT-4, and Mistral-7b-Instruct). Of these, 55 are human vs. model matches, while 38 are human vs. human matches. For the matches, the 243 QA pairs are divided into 12 sets of 20, stratified by category, with three questions for tiebreakers.

Hosting real-time competitions with human players provides several benefits: (1) direct comparison of confidence calibration between humans and models on the same questions, (2) recruiting experienced players skilled in calibrating their answers,[8] and (3) validating that questions are human-answerable and unambiguous, as an additional quality check for the dataset.

4 Human-Grounded Calibration Evaluation

To compare model and human calibration, we analyze response correctness and buzz decisions. Then, we introduce a baseline metric, CALSCORE. Unlike traditional calibration metrics, CALSCORE facilitates per-instance calibration analysis, which lets us identify specific questions on which models are especially miscalibrated. In addition, it factors in human performance on the same question, to penalize cases in which models are confidently incorrect when humans are uncertain (§ 4.1).

4.1 Human-grounded metric: CALSCORE

CALSCORE evaluates model calibration error while incorporating human buzzpoints. This adjustment reflects the structure of the competition—models must be confidently correct before humans know the answer. The adjustment also places higher weight on instances where model errors are more likely to mislead users—if a model is confidently incorrect when humans are still uncertain, humans are less likely to recognize and override the error (§ 5.2).

Using the live competition data, we track the proportion of humans answering correctly up to a specific clue so that the metric applies higher penalties and rewards for earlier (harder) clues. To measure the expected probability of a team buzzing correctly on a given question, we consider teams' buzzes at each clue t of question q. We define h_t as the cumulative probability of a human team correctly buzzing up to clue t, calculated as the number of correct buzzes by human teams up to t divided by total buzzes by human teams up to t. For model responses, g_t indicates the correctness of a model's guess at clue t (1 if correct, -1 if not), and c_t indicates the model's confidence in its guess.

4.2 MCE: Unadjusted model calibration error

The normalized expectation $r(\mathbb{E}_t\left[g_t c_t\right])$, calculated over clues in a single question, measures calibration as the expectation that the model answers correctly, weighted by confidence (where $r(x)$ renormalizes to a [0, 1] range; see Appendix J). Conversely, $1 - r(\mathbb{E}_t\left[g_t c_t\right])$ evaluates the model's calibration *error* (MCE) on that question. High-confidence incorrect answers and low-confidence correct answers result in higher error, indicating poor calibration on question q.

4.3 CALSCORE using GRACE

CALSCORE incorporates human performance into MCE. This facilitates identification of specific cases where models are less calibrated than humans; rewards models more for being confidently

[8]After the competitions, we survey players about their experience levels and individual strengths (Appendix F.2).

correct before humans know the answer (simulating the competition setting); and penalizes them more for being confidently incorrect at that stage, placing greater weight on a higher-stakes error.

We thus weight the calibration at clue t by $(1 - h_t)$, the proportion of humans who have not yet answered correctly by clue t. A high score of $r(\mathbb{E}_t [(1 - h_t) g_t c_t])$ indicates that the model is well-calibrated relative to human buzz performance.[9] Conversely, we estimate the expected probability for cases where the model does *not* improve over humans, either due to incorrect answers or low confidence:

$$\textsc{CalScore}_q = 1 - r(\mathbb{E}_t [(1 - h_t) g_t c_t]). \quad (1)$$

This adjustment evaluates the model's calibration error relative to human calibration performance on the same question. We then define $\textsc{CalScore}_D$, the human-adjusted model calibration error for a benchmark D, as the average of $\textsc{CalScore}_q$ across all questions L.

4.4 CalScore^2 using GRACE

We introduce $\textsc{CalScore}^2$ as a secondary metric to evaluate model calibration, following the notation in §4.3. We define the *buzz confidence* b_t, the probability that the model will buzz (i.e., be confident) at time t. This depends on the model's confidence c_t and the probability of not buzzing in earlier steps:

$$b_t = c_t \prod_{i=0}^{t-1} (1 - c_i).$$

The model's score at each step is $S_t = b_t g_t$, where $g_t = 1$ if the model is correct and 0 otherwise. The total score for a question is $S_q = \sum_{t=0}^{T} S_t$, with the constraint $\sum_{t=0}^{T} b_t = 1$, ensuring the model eventually buzzes.

To compare model and human calibration, we define **CalScore**^{2}, using human buzz probabilities h_t (proportion answering correctly by step t) as a benchmark. Let $K_t = h_t \sum_{e=0}^{t} b_e g_e$, where the system buzzes correctly before or at step t. The full metric is:

$$\textsc{CalScore}^2 = 1 - \left(\sum_{t=0}^{T} K_t + \left(1 - \sum_{t=0}^{T} h_t\right) \sum_{t=0}^{T} b_t g_t \right), \quad (2)$$

[9]A model is perfectly calibrated when it buzzes with full confidence ($c_t = 1$), is always correct ($g_t = 1$), and answers before humans buzz correctly ($h_t = 0$), resulting in $r(\mathbb{E}_t [(1 - h_t) g_t c_t]) = 1$.

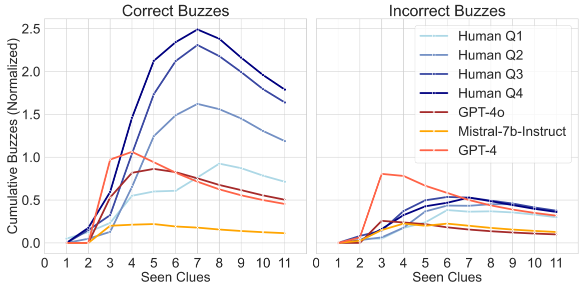

Figure 4: Each team's cumulative buzzes (normalized by the number of matches each team participated in). The top quartile of human teams (Q4) achieves the highest cumulative correct buzz rate, peaking over twice as high as the best model. Top human teams are thus more accurate and better-calibrated than models, even as the difficulty changes when more clues are revealed.

which rewards early correct buzzes relative to humans and penalizes overconfidence when humans abstain. Further details are in Appendix M.

5 Model Calibration Evaluation

GRACE helps to evaluate differences between human and model calibration (§5.1). We also validate the dataset's difficulty granularity and discuss calibration errors using our proposed metric (§5.2) and qualitative analysis (§5.3).

5.1 Comparing human and model calibration

Buzz performance. To compare human and model calibration, we first examine when and whether each team buzzes on the question, as well as the correctness of their answers. Figure 4 gives each team's cumulative buzzes over the number of matches each team participated in. The 17 human teams are divided into quartiles, from Q1 (bottom) to Q4 (top), according to their total correct buzzes. Human teams, especially the top quartile, achieve the highest cumulative correct buzz rate (peaking in the middle of the questions), demonstrating their ability to confidently infer correct answers with fewer clues and indicating better accuracy and calibration than models. In contrast, GPT-4 exhibits a moderate cumulative correct buzz rate, which is only briefly higher than the top human teams and lower than 50% of human teams for most of the question. Meanwhile, Mistral-7b-Instruct lags significantly behind all other teams, indicating poor calibration. In addition, GPT-4 and GPT-4o exhibit substantially higher incorrect buzz rates than human teams (right plot). All models, especially GPT-4, are overconfident early in the questions

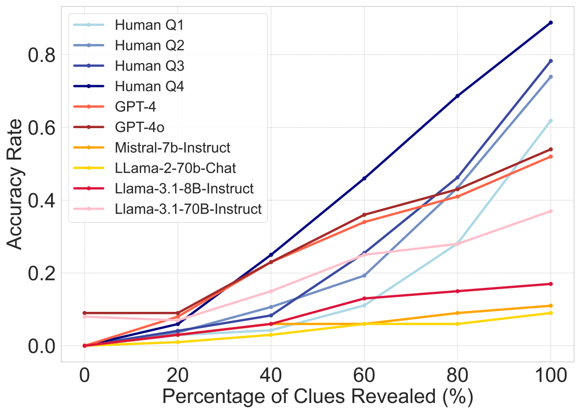

Figure 5: Comparison of human and model average accuracy rates as more clues are revealed (whether the team's guess is correct after seeing the first n clues). As more clues are revealed, accuracy improves for both models and humans. Models often answer incorrectly until most clues are provided, and human accuracy increases more rapidly, validating that each instance becomes easier for both humans and models and that most humans can answer correctly by the end.

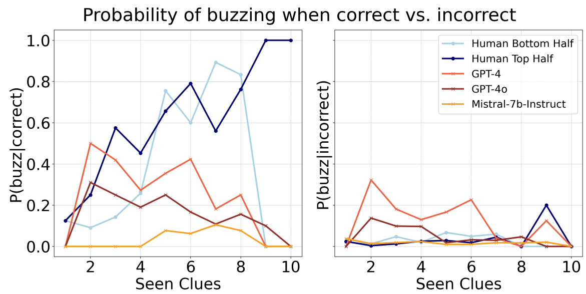

Figure 6: Humans are far more likely than models to buzz in when they are correct (left), and typically less likely to buzz in when they are incorrect (right), indicating that models remain miscalibrated relative to humans even when explicitly controlling for accuracy. (Due to the smaller sample size of human buzzpoints in the survey data, we use halves instead of quartiles here.)

when little information is available: they are **especially miscalibrated relative to humans when the question is still hard**. Overall, the models tend to buzz incorrectly more often than humans and correctly less often, indicating **overconfidence in wrong answers and underconfidence in correct ones**.

Difficulty granularity of each question. To evaluate model calibration over a range of difficulty levels for models, we asked the writers to write questions that are easier to answer as more clues are revealed (§ 3). To validate this design, we examine model and human correctness as the percent of clues revealed increases. For models, we consider the correctness of a model's guess for the first n clues. The questions in GRACE are appropriately challenging and become easier for models and humans as more clues are revealed (Figure 5). Human team accuracy (blue) increases steadily, indicating that question difficulty indeed decreases as clues are revealed for human players. Moreover, even top models like GPT-4 and GPT-4o have under 50% accuracy until at least 90% of clues are provided, highlighting significant room for improvement on this benchmark. To measure human accuracy per corresponding clue, we used offline human responses (§3.2.1; quartiles calculated per Appendix F.3).

Notably, the bottom three quartiles of humans are less accurate than top models for most of the question (Figure 5), yet still typically outperform models on maximizing correct buzzes relative to incorrect buzzes (Figure 4). This trend suggests that **models' relatively high rate of incorrect buzzes and low rate of correct buzzes is due to miscalibration, not inaccuracy.** We investigate this distinction further in the next section.

Conditional likelihood of correct answers. While the tournament allows us to observe $P(g = 1 \mid b)$, the likelihood that a team's guess is correct ($g = 1$) when they buzz (b), we also aimed to compare how often teams are confident enough to buzz when correct, $P(b \mid g = 1)$, and when incorrect, $P(b \mid g = 0)$. Using the offline human responses (§3.2.1), we estimate $P(b \mid g = 1)$ for each human team. For each player, we calculate $P_{player}(b \mid g = 1)$: the number of instances when a player buzzed in correctly with n clues revealed, divided by the number of instances when a player's guess was correct with n clues revealed. We then estimate $P(b \mid g = 1)$ for the top and bottom half of respondents (Appendix F.3) as the average of $P_{player}(b \mid g = 1)$ across all surveyed players in that half. Finally, we compare this estimation with $P(g = 1 \mid b)$ for the tested models. We follow the same process to estimate $P(b \mid g = 0)$.[10] The results indicate that even **the strongest models are less confident than humans on correct answers and more confident on incorrect ones** (Figure 6). For most questions, all humans are more than 50% likely to buzz when correct, while models remain below 45%, indicating lower confidence in correct

[10]For all estimates, we consider only the guess correctness and buzz statistics up to the point when a player first buzzes, as later guesses do not count in real competitions.

Verbalized-based Confidence										
Model	**Brier Score**		**ECE**		**MCE**		**CalScore**		**CalScore^2**	
GPT-4	0.274	2	0.259	2	0.584	1	0.588	1	0.878	1
GPT-4o	0.266	1	0.224	1	0.601	2	0.604	2	0.900	2
Llama-3.1-70B-Instruct	0.373	3	0.392	3	0.685	3	0.719	3	0.955	3
LLama-2-70b-Chat	0.490	4	0.570	4	0.739	4	0.803	4	0.961	4
Llama-3.1-8B-Instruct	0.623	5	0.693	5	0.774	5	0.843	5	0.978	5
Mistral-7b-Instruct	0.716	6	0.784	6	0.790	6	0.881	6	0.973	6
Logit-based Confidence										
Model	**Brier Score**		**ECE**		**MCE**		**CalScore**		**calscore2**	
GPT-4o	0.341	3	0.353	2	0.654	2	0.661	1	0.679	2
Llama-3.1-70B-Instruct	0.323	2	0.339	1	0.651	1	0.679	2	0.661	1
GPT-4	0.380	4	0.388	3	0.672	3	0.684	3	0.684	3
Llama-3.1-8B-Instruct	0.302	1	0.397	4	0.675	4	0.718	4	0.718	4
Mistral-7b-Instruct	0.553	5	0.677	5	0.766	5	0.846	5	0.846	5
Llama-2-70b-Chat	0.774	6	0.829	6	0.825	6	0.921	6	0.921	6

Table 1: Across both confidence elicitation methods, all models display greater error under human-adjusted CalScore and compared to MCE (CalScore without human adjustment). CalScore thus captures errors that existing methods overlook: cases where models underperform relative to humans by being confidently wrong or underconfident when correct.

answers. Among the models, GPT-4 was most likely to buzz incorrectly, reflecting its confidence in wrong answers.

As clues are revealed, humans become more likely to buzz when they know the correct answer, while models become less likely to buzz. This suggests that seeing more clues strengthens human confidence, but not model confidence.[11]

5.2 CalScore analysis

We evaluate the calibration error of six LLMs on GRACE using existing metrics (ECE and Brier scores; Appendix G) and CalScore (§ 4). For each clue in a question, we collect logit-based and verbalized confidences to compute metric scores.

CalScore and CalScore^2 correlate with ECE and Brier score results (Table 1); however, across both confidence elicitation methods, all models display greater error under two human-adjusted metrics compared to MCE (CalScore without human adjustment). The two metrics capture errors that existing methods overlook: cases where models underperform relative to humans by being confidently wrong or underconfident when correct. Thus, **CalScore and CalScore^2 capture that models are especially ill-calibrated compared to humans**, and factoring in human performance reveals more room for improvement on LLM calibration. The gap between MCE and the CalScores widens for worse-performing models, suggesting that weaker models are even more miscalibrated relative to stronger models when

[11] A small fraction of human participants (n=1) has a sharp spike in incorrect buzzes near the end of a question.

Q: One thinker's argument that this claim is a "hyperbolic point which ought to be silent" was subject to a response titled "My Body, This Paper, This Fire." Jacques Derrida first coined the word "différance" in a book responding to Michel Foucault's *Madness and Civilization* and partially titled for this statement. In *The Search for Natural Light*, a premise involving doubt was added to this statement. This claim, which Pierre Gassendi criticized for being circular, was presented as an example of a "clear and distinct" idea. For 10 points, name this first principle coined in René Descartes' *Discourse on the Method.*

ANSWER: "I think, therefore I am" [or "cogito, ergo sum"]

Figure 7: Sample question on which models are poorly calibrated.

factoring in human performance. Additionally, CalScores report higher errors than ECE and Brier scores across both confidence elicitation methods for most models, underscoring calibration deficiencies that previous metrics underestimate.[12]

5.3 Qualitative analysis and model errors

Miscalibrated instances from CalScore. All six models exhibit similar patterns for the questions on which they were most- and least-calibrated under CalScore (Figure 7). Models did best on questions that mention concrete proper nouns closely associated with the answer, even on obscure topics: for example, a question on Ireland that gives the titles of Irish songs, a question on telomeres that mentions the protein TRF2, and a question on Brooklyn that mentions the neighborhood of Midwood. Models tend to be least-calibrated on questions with multiple plausible answers (e.g., one on fish as a Buddhist symbol, since other animals also have symbolic meanings in Buddhism). Models also struggle on questions that use descriptions instead of titles (e.g. a question that describes music by Maurice Ravel, and one that describes Jewish birth ceremonies).

Qualitative feedback. We survey the human players for feedback on model abilities. Differences in model calibration are visible to the players: several find GPT-4 "too aggressive," while Mistral seems much weaker, often buzzing late in the question. One player notes that the models "obviously knew a lot, but were quite bad at gauging how well they knew something to [buzz]." Others note that models tend to buzz on "more concrete clues" and struggle with multi-step reasoning. For example, a question on Alice Walker mentions her

[12] All four metrics use a [0,1] scale; lower is better.

trip to Eatonville to write about local author Zora Neale Hurston. Players note that GPT-4 incorrectly guesses "Zora Neale Hurston," while human players correctly say "Alice Walker."

Players also note that when models were incorrect, they give more "unreasonable" answers than humans do. For example, models incorrectly answer a question on the treatise Philosophical Investigations with "Fermat's Little Theorem" and "*The Lion, the Witch and the Wardrobe*." Guessing an equation and a children's book with high confidence for a work of philosophy suggests serious miscalibration, since either option should be completely outside the realm of possibility; no human players gave answers so distant from the correct one. Other model errors not observed among human players include buzzing before any substantive clues are revealed, answering with a song title for a question asking for a surname, and hallucinating inexistent schools of philosophy.

Model and human strengths between question topics differ greatly. We examine models' and humans' ratios of correct to incorrect buzzes per category. Human players are best at literature, but this is the weakest or second-weakest category for all models. All models did relatively well on science. GPT-4o is much stronger at social science, arts, and science than other categories, and slightly outperforms humans for every category; GPT-4 was worse than the humans for all categories.

All human participants in our competitions were experienced players, but we find that calibration performance varies greatly even among these experts: stronger humans substantially outperform top models, but not all humans do. A general takeaway for future model-human comparisons on tasks involving calibration is that variance in human skill can greatly affect the outcome of a comparison.

A side benefit of conducting live human-model competitions was a significant degree of community involvement from trivia enthusiasts who were not researchers. In-person data collection, though more involved than crowdsourced data, offers other benefits: we found that participants were attentive and enthusiastic; moreover, in-person data collection (especially "gamified" approaches) raises awareness of and interest in AI.

6 Conclusion

For users to trust LLMs, they need assurance that these models will not confidently produce wrong answers. To address this, GRACE offers a benchmark for fine-grained calibration evaluation, grounded in human calibration. Our analyses on GRACE reveal that models are often miscalibrated relative to well-informed humans. Specifically, model calibration errors came from difficulty with abstract descriptions, far-fetched incorrect guesses, and confidently incorrect answers given few clues. Our new metric, CALSCORE, combined with GRACE, evaluates the performance of six LLMs, revealing significant room for improvement.

GRACE provides a blueprint for developing human-focused improvements to calibration: improving verbalized confidences that measurably help human decision-making, personalizing abstention based on individual human skill, and measuring these interactions via human-model teaming.

Limitations

Since GRACE focuses on a question-answering task, its applicability to broader NLP domains remains unexplored. Future work should explore calibration in other open-ended generation settings to assess generalizability. Furthermore, while our proposed metric, CALSCORE, serves as a baseline, it is not exhaustive of all forms of uncertainty and miscalibration. For instance, enhancing this metric for human-AI collaboration could help users determine when to rely on models and when to defer to human judgment.

Acknowledgments

Many thanks to the question editors who helped to create the dataset: Ankit Aggarwal, Jaimie Carlson, Bradley Kirksey, Linus Luu, Shahar Schwartz, Noah Sheidlower, and Jonathan Tran. Special thanks to the writers. This research was only possible due to the generous participation of the players who competed in the tournaments (special thanks to Joy An, Andrew Gao, Kevin Yu, Alex Stonemanxa, Priyanka Raghavan, and Magdalena Lederbauer) and the moderators who read the questions. We thank Lakshya Agrawal, Nishant Balepur, Nicholas Tomlin, Swabha Swayamdipta, and Irene Ying for providing feedback on earlier drafts.

This research was partly supported by an NSF GRFP grant. Sung and Boyd-Graber are supported by NSF Grant IIS2403436. Opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

References

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic calibration of long-form generations. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. 2012. Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1290–1301, Jeju Island, Korea. Association for Computational Linguistics.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Richard Caruana. 2019. Friends don't let friends deploy black-box models: The importance of intelligibility in machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 3174, New York, NY, USA. Association for Computing Machinery.

Jiawen Deng, Kiyan Heybati, and Hemang Yadav. 2025. Development and validation of machine-learning models for predicting the risk of hypertriglyceridemia in critically ill patients receiving propofol sedation using retrospective data: a protocol. *BMJ open*, 15(1):e092594.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. 2020. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*.

David A Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1–1.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé, III. 2016a. Opponent modeling in deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1804–1813, New York, New York, USA. PMLR.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016b. Opponent modeling in deep reinforcement learning.

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

Evgenia Ilia and Wilker Aziz. 2024. Predict the next word:< humans exhibit uncertainty in this task and language models _>. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 234–255.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. 2024. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks?

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.

Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention.

Jingshu Li, Yitian Yang, Renwen Zhang, and Yi-chieh Lee. 2024a. Overconfident and unconfident ai hinder human-ai collaboration. *arXiv preprint arXiv:2402.07632*.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Lee Boyd-Graber. 2024b. PEDANTS: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398, Miami, Florida, USA. Association for Computational Linguistics.

Ollie Liu, Deqing Fu, Dani Yogatama, and Willie Neiswanger. 2024. Dellma: A framework for decision making under uncertainty with large language models. *arXiv preprint arXiv:2402.02392*.

Sewon Min, Daniel Khashabi, Hannaneh Hajishirzi, and Wen-tau Yih. 2022. Crepe: A contrastive approach for post-hoc calibration of pretrained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

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.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1587–1598, Lisbon, Portugal. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019a. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan L. Boyd-Graber. 2019b. Quizbowl: The case for incremental question answering. *CoRR*, abs/1904.04792.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large language models help humans verify truthfulness–except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. Lacie: Listener-aware finetuning for confidence calibration in large language models. *ArXiv*, abs/2405.21028.

Elias Stengel-Eskin and Benjamin Van Durme. 2023. Calibrated interpretation: Confidence estimation in semantic parsing. *Transactions of the Association for Computational Linguistics*, 11:1213–1231.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher 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*, pages 5433–5442, Singapore. Association for Computational Linguistics.

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

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong 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)*, pages 15440–15459, Bangkok, Thailand. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

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

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.

Wencong You and Daniel Lowd. 2022. Towards stronger adversarial baselines through human-AI collaboration. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 11–21, Dublin, Ireland. Association for Computational Linguistics.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

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

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

A Ethical Considerations

We address ethical considerations for dataset papers, given that our work contains a new dataset, GRACE, and collects human responses in our user study. We reply to the relevant questions posed in the `ACL 2022 Ethics FAQ`.[13]

When collecting human responses and questions, our study was pre-monitored by an official IRB review board to protect the participants' privacy rights. Buzzpoints in GRACE are anonymous; the feedback survey asked for respondents' names for purposes of compensation, but only aggregate data and anonymous quotes are used in our study. Before distributing the survey, we collected consent forms for the workers to agree that their answers would be used for academic purposes. The trivia players were awarded a total $900 USD worth of online gift cards after the competitions. The prizes were $150, $100, $50 for the first three places at each site where the tournament was held. The trivia writers were paid $5 per question and editors paid $2.50 per question edited based on the estimated completion time, and which was calculated to reach over $10 USD an hour (a rate higher than the US national minimum wage of 7.50 USD).

B Dataset Details

B.1 Question distribution

Questions in GRACE are distributed by subject as follows, based on the standard quizbowl distribution at `acf-quizbowl.com/distribution`:

- 20% literature: 5% American literature; 5% British literature; 5% European literature; 5% world and other literature
- 20% history: American history; 5% world history; 5% European history; 5% other history
- 20% science: 5% Biology; 5% Chemistry; 5% Physics; 5% computer science, math, and other science
- 15% arts: 5% painting/sculpture; 5% classical music; 5% other arts
- 15% social sciences: 5% religion; 5% philosophy; 5%
- 5% geography and current events
- 5% myth, pop culture, and other

B.2 Sample dataset questions

Q: The protagonist of the story "Storm in a Teacup" obsessively reads a book with this number in its title. This number partly titles a work that opens by noting that people are born naturally good, used to teach children. One hundred times this number titles a poetry anthology whose length was inspired by the Classic of Poetry. A man visits a hut this number of times to recruit the Sleeping Dragon, and this many characters pledge to die on the same day while swearing the Peach Garden Oath. For 10 points, a Luo Guanzhong novel is titled for how many kingdoms?

ANSWER: three [or san; accept Three Character Classic or Three Hundred Tang Poems or Romance of the Three Kingdoms; accept Sānzì Jīng or Tángshī sānbǎi shǒu or Sānguó Yǎnyì]

Q: During this decade, some members of the Circle of Seven participated in a group called the Golden Square and led a coup to remove the Iraqi government from power. The Pahlavi ruler and father of Mohammed Reza Shah was forced out of power in this decade, during which a supply route was developed to clue through port cities to Tehran. The Syria-Lebanon campaign of this decade was led by Archibald Wavell, who later lost in the Western Desert, where the Battle of El Alamein took place. For 10 points, General Erwin Rommel fought in what decade when most of World War Two occurred?

ANSWER: 1940s [prompt on 40's]

Q: One form of this adjective describes a distributed object whose state is equivalent to a strictly serializable database. Weight decay tends to pull a sigmoid activation function towards this adjective's namesake "regime". If all activation functions of a deep neural network have this form, the output is provably this type of mapping of the inputs. ReLU is a modified piecewise form of this type of function, which is the simplest separator in two-dimensional binary classification. The most common shape of least-squares regression is described by, for 10 points, what adjective that describes functions of the form "y=mx+b"?

ANSWER: linear [accept linearizable or word forms; prompt on planar or hyperplanar or word forms]

Q: The NSF's IDP sets standards for tools used to extract this substance, which include Foro 3000 and DISC. Coral, certain protists, and this substance are the primary sources of delta-O-18 information. Very thin layers of this substance that look like oil spills are nicknamed for grease. Due to diffusion, clathrates and trapped air in this substance help to track atmospheric gas concentrations, and may be retrieved from masses that undergo calving, resulting in growlers or bergy bits. For 10 points, name this material, which may be extracted in core samples from namesake sheets or from glaciers.

ANSWER: ice [or frozen water; or solid water; accept ice sheets; accept icebergs; prompt on snow or firn; prompt on glaciers; prompt on water

Q: Sephardic Jews have recently moved into this region near the original Vitagraph studios in Midwood. The world headquarters of the Chabad ("huh-BAHD") movement is located in this region within a neighborhood where the Rebbe struck and killed Gavin Cato in 1991. Many Jews of the Bobov sect live in this region's neighborhood of Borough Park. This region, where Jews clashed with Black residents in the Crown Heights Riots, is where many Hasidic Jews live in the increasingly trendy neighborhood of Williamsburg. For 10 points, name this New York City borough whose Jewish residents often live near Coney Island.

ANSWER: Brooklyn [or Kings County; prompt on New York City]

[13] `https://www.acm.org/code-of-ethics`

		20%	40%	60%	80%	100%
TrickMe	GPT-4	31.07%	65.97%	82.20%	89.96%	93.68%
	GPT-4o	28.77%	61.56%	77.90%	86.64%	91.11%
GRACE	GPT-4	7.49%	23.62%	30.60%	40.40%	52.19%
	GPT-4o	8.59%	22.65%	32.05%	42.86%	53.78%

Table 2: GRACE and TrickMe dataset accuracy rate comparison. Similar as Figure 5, we compare the accuracy when the percentage of clues revealed (%) are 20%, 40%, 60%, 80%, 100%. GRACE results show much lower accuracy compared with TrickMe, indicating that our dataset is more adversarial incrementally.

Q: At the 2023 Women's World Cup, the Spanish and Dutch teams sparked controversy by seemingly performing this dance during training. Lyrics commonly sung to this dance describe putting one foot in front of the other "until the Sun shines on me." The currently youngest lawmaker in one country led this dance while tearing papers in half. Participants in a November 2024 march performed this dance in protest over a bill introduced by David Seymour that would affect the Treaty of Waitangi. For 10 points, recent protests in New Zealand have made use of what Māori ceremonial dance?

ANSWER: haka

B.3 Sample question on which models are well-calibrated

Q: Sephardic Jews have recently moved into this region near the original Vitagraph studios in Midwood. The world headquarters of the Chabad movement is located in this region within a neighborhood where the Rebbe struck and killed Gavin Cato in 1991. Many Jews of the Bobov sect live in this region's neighborhood of Borough Park. This region, where Jews clashed with Black residents in the Crown Heights Riots, is where many Hasidic Jews live in the increasingly trendy neighborhood of Williamsburg. For 10 points, name this New York City borough whose Jewish residents often live near Coney Island.

ANSWER: Brooklyn [or Kings County; prompt on New York City]

C Dataset Comparison for Adversarialness

We compare GRACE with the TrickMe dataset (Wallace et al., 2019) on (1) question length and (2) question difficulty evaluation. Overall, GRACE questions are longer: GRACE has an average of 5.09 sentences per question, while TrickMe has an average of 4.74 sentences. We compare the difficulty based on the accuracy rate as the clues are revealed in Table 2. GRACE exhibits higher adversarialness throughout the incremental questions.

D Interface

D.1 Guesser details

Before deciding when to buzz, models need to generate answers as clues are revealed. We call this process "guessing" and the model is used as a *guesser*. The full prompt of the guesser is in Appendix E.1, including the general instructions and retrieved examples. We train a TF-IDF model as the retriever with past quizbowl questions following Rodriguez et al. (2019a).[14] The main goal is to reduce hallucinations and guide the model to learn the granular clue and guess format.

D.2 Training buzzer for authoring interface

For the writing interface, the logistic regression model was trained with two kinds of features: GPT-3.5's logit based confidence, and pre-designed features derived from the question, answer, and metadata. Features include text-based metrics (e.g., TF-IDF scores, overlaps between Llama predictions and TF-IDF guesses), probabilistic outputs (Llama log and prompt probabilities), and contextual indicators (sentence index, length, and presence of phrases like "10 points"). This model was trained on Rodriguez et al. (2019a), also pyramidal questions, using Llama-13b predictions (Touvron et al., 2023).

A primary distinction from Wallace et al. (2019) is that their interface only showed the final correct guess.

D.3 Question editors and writers

Writers and editors were recruited via a public quizbowl forum and were located in the US and UK. All editors underwent IRB training and had written for at least three previous tournaments. Writers were paid $5 per question and editors paid $1 per question edited. A head editor with 5 years of experience writing and editing quizbowl questions, including as head editor of two previous tournaments, supervised the writers and provided an additional quality check. The consent form is in Table 3.

D.4 Model confidence elicitation

We experiment two approaches to get the model confidence score for a generated guess.

Token-logit based confidence. Log probability of the generation is a common method to estimate the

[14] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Privacy Policy
Introduction
Welcome to 'Stump the Computer,' hosted by the research group at xxx. Your participation in our research is voluntary and deeply valued.
This Privacy Policy outlines how we handle the information you provide during your interaction with our research project.
Purpose of the Study
This research, conducted by xxx, aims to collect human-generated data for a fact verification system.
Participants, like you, who enjoy trivia knowledge and trivia-related games, are invited to help us understand how humans and computers handle challenging questions.
Information Collection and Use
As you participate in this project, you will interact with an online interface to submit questions designed to challenge both human and computer intelligence.
You may use a Google Doc, automatically generated in your Google Account, to draft these questions.
We collect this data to enhance our research and improve the interaction models between humans and computers.
Data Confidentiality
We take your privacy seriously. All data collected during this study will be stored on a password-protected web server with encrypted storage.
The server is in a secure access data center, managed professionally. Access to the data is strictly limited to the principal investigators of this study.
After the study concludes, any personally identifiable information will be anonymized or destroyed to ensure your privacy.
Potential Risks
We acknowledge the risk of breach of confidentiality in any online activity.
We have implemented robust security measures to mitigate such risks and protect your data.
Benefits and Compensation
While there is no direct personal benefit from participating, your contributions are invaluable in advancing research on human-computer interactions.
Compensation for participating includes $5 per question written and $2.50 for each question edited.
Your Rights
Participation is entirely voluntary. You may withdraw at any time without penalty. Should you have any questions or need to report a concern, please contact: xxx
Changes to This Policy
We may update this policy periodically to reflect changes in our practices. Continued participation after such changes constitutes your acceptance of the new terms.
Consent
By signing up and participating, you affirm that you are at least 18 years old, have reviewed this policy, and consent to engage in this study.
Your rights and privacy are paramount, and we are committed to protecting them.
Thank you for participating in our task and contributing to the advancement of our research.

Table 3: Consent form for question writers.

model confidence (Nguyen and O'Connor, 2015). To get the confidence score in our setup, we retrieve the logit for each generated token, and take the average of the exponentials of these logit values (Huang et al., 2023).

Verbalized confidence. Recent study shows verbalized probabilities can be better calibrated than log probabilities (Tian et al., 2023; Xiong et al.), which motivate us to include the verbalized confidence in our experiments. We follow the previous prompt in Appendix E.1 and add the instructions from Tian et al. (2023) to return the confidence:

> Given the following information, provide the title of the Wikipedia page that would best answer the last question fragment. If you are not sure, just give your best guess. If you don't know, answer None. The answer should be as short as possible. **While you give the guess, please also provide the probability that it is correct (0.0 to 1.0). Give ONLY the guess and probability, no other words or explanation. For example:**
>
> **The answer is: <most likely guess, as short as possible; not a complete sentence, just the guess!>**
> **Probability: <the probability between 0.0 and 1.0 that your guess is correct, without any extra commentary whatsoever; just the probability!>**
>
> Question: {retrieved examples}
> The answer is: {retrieved examples}
>
> Question: {each clue}

To ensure accurate extraction of probability scores from model outputs, we initially define the desired format based on the prompt. We then proceed to identify and print any cases where confidence scores are not successfully extracted. By observing these cases, we can discern patterns and refine our post-processing rules. This iterative approach allows us to capture as many corner cases as possible, enhancing the robustness of our data extraction process.

E Model Buzzpoint Generation

E.1 Guesser details

To retrieve a model's top guess after n clues have been revealed, we prompt the model with the first n clues from the question. To retrieve the best guess, we employ a "retrieval and guess" approach to enhance QA performance, using the following prompt:

> Given the following information, provide the title of the Wikipedia page that best answers the last question fragment. If unsure, provide your best guess. The answer should be concise.
>
> Question: {retrieved examples}
> The answer is: {retrieved examples}
>
> Question: {each clue}
> The answer is:

E.2 Process for calculating model buzzpoints

The process for calculating model buzzpoints followed the steps below.

Algorithm 1 Find model buzzpoints for a question

Require: N, the number of clues in the question; and t, the buzz threshold.
Let $n = 0$
while $n < N$ **do**
  Prompt the model to answer the question, given the first n clues (Appendix E.1).
  Compute the model's confidence c in its top guess by summing the log probabilities of the tokens comprising the guess.
  if $c > t$ **then**
    Buzz in
    break
  n += 1

E.3 Assigning buzzer threshold

We used question data from the 2023 Expo quizbowl competition, where expert players competed against ChatGPT as a testbed to assign buzzer threshold. The dataset includes questions covering various topics, with recorded buzz and guess correctness.

We incorporate the explanation from Rodriguez et al. (2019a) to explain how the threshold was set for our buzzer. To estimate the probability $\pi(t)$ that a player has answered correctly by position t, the following formula is used:

$$\pi(t) = 1 - \frac{N_t}{N}, \tag{3}$$

where N is the total number of player-question records, and N_t is the number of instances where a player has answered correctly by position t. This equation indicates how likely it is that a player has given the correct answer by a certain point in the question. To make this probability easier to use in practice, it is approximated with a polynomial function:

$$\pi(t) = 0.0775t - 1.278t^2 + 0.588t^3. \tag{4}$$

This polynomial provides a smooth estimate of how human accuracy changes as the question progresses, allowing for a data-driven approach to determining optimal buzz thresholds. Since we also want to get the optimal buzzpoints, we adopt a threshold from this study to determine the model buzzpoints. The confidence thresholds by model family are: -0.03 for GPT models, and -0.05 for Mistral models.

F Tournament and Survey

F.1 Recruiting human players

Human teams were recruited by posting a call for players on social media and public forums for quizbowl players. To incentivize teams to play as well as possible, the top three teams in each tournament were awarded a prize. The prizes were \$150, \$100, \$50 for the first three places at each site where the tournament was held. The tournaments were all held in the US and the players were also located in the U.S. The consent form is in Table 4.

F.2 Player expertise

Respondents had an average of 5.5 years of previous experience playing quizbowl. 22% of players had studied or were currently studying in the physical sciences or engineering; 31% studied computer science or math; 17% studied the humanities; 13% studied a combination of fields; and 17% were undecided. Since quizbowl players typically specialize in certain categories and learn more about those areas, we also asked them for their areas of specialization. 39.13% of respondents listed the sciences as an area of specialization; 21.74% listed history; 39.13% listed the social sciences; 52.17% listed literature; 39.13% listed fine arts; and 21.74% listed geography or current events.

F.3 Ranking human performance from survey

Because the survey measures individual accuracy, without a competition setting, we rank humans using the following metric: participants earn $(20 - 20c)$ points per question with a correct buzz and lose 5 points per question with an incorrect buzz, where c is the proportion of clues seen at the buzzpoint. Human quartiles and top/bottom half categorization based on the offline buzzpoints (for Figures 5 and 6) are taken from rankings under this metric.

G ECE and Brier Score Details

Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017) and Brier scores (Brier, 1950) are widely used metrics for assessing model calibration. Below, we define these metrics using the notations introduced in Section 4.1: g_t represents the answer correctness at clue t (1 if correct, 0 otherwise, which is slightly different from CALSCORE for simplicity); c_t represents the corresponding model's confidence in its guess.

ECE. It measures the weighted average over the absolute difference between accuracy and confidence. To compute this, we first split confidence values into $M = 10$ bins equally. B_m represents the confidence set of the m^{th} bin. N is the total

Welcome to a quick Trivia Quiz! You'll tackle 37 questions and decide if you're confident enough to buzz.
After each clue, please write down your best guess even if you're not sure of the answer.

To track your progress, refer to the progress bar at the top of the page (Please disregard the question numbers—they are randomized for fairness).

Please enter your email address here so we can ensure prizes are sent to the right recipients at the end of the quiz.
We're giving away 4 raffle prizes (each a $25 gift card) to participants who complete the quiz.
Additionally, the top scorer with the highest accuracy and fastest buzz will receive a $5 bonus prize.
Rewards are determined by a combination of your accuracy and buzzing speed.
Correct answers receive 0 to 20 points depending on how early you buzz. Incorrect answers receive -5 points.
Good luck and have fun!

Here is the short summary of what this survey will be used for:
Project Title
A Leaderboard and Competition for Human-computer Adversarial Question Answering
Purpose of the Study
This research is being conducted by xxx at xxx.
We are inviting you to participate in this research project because you are interested in trivia knowledge and enjoy trivia-related games.
The purpose of this research project is to collect human-generated data and responses for building a deployable QA system.
Procedures
We would like to use two adversarial datasets to test our approach of determining how adversarial the questions are.
You will provide your best guess and write down your answer in a short form.
Potential Risks and Discomforts
There is a risk of breach of confidentiality and that efforts to mitigate this risk are described in the Confidentiality section below.
Potential Benefits
While this research is not designed to benefit you personally, we hope that in the future, the researchers might benefit from human computer interaction studies with investigation of
question answering behavior and AI development during this study.
Confidentiality
Only the investigators (xxx) of this study will have access to the study data.
Data will be stored on a password-protected web server with encrypted storage.
The server is professionally managed in a secure access data center.
After the study ends, only user names associated with e-mail addresses will be retained and the associated e-mail addresses will be deleted.
Right to Withdraw and Questions
Your participation in this research is completely voluntary. You may choose not to take part at all.
If you decide to participate in this research, you may stop participating at any time.
If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.
If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator:
xxx

If you have any questions, please use this address: xxx

Table 4: Consent form for players.

number of clues across the dataset.

$$\text{ECE} = \frac{1}{N}\sum_{m=1}^{M}\left|\sum_{t\in B_m} g_t - \sum_{t\in B_m} c_t\right|$$

Brier Score. It measures the mean squared difference between the predicted probability and the actual binary outcome, measuring how well the predicted confidence aligns with the true correctness of the answer. A lower Brier Score indicates better calibration, as it reflects more accurate and well-calibrated probability estimates.

$$\text{Brier Score} = \frac{1}{N}\sum_{t=1}^{N}(c_t - g_t)^2$$

H Experiment Details

We query OpenAI APIs for GPT-4 (gpt-4-0613) and GPT-4o (gpt-4o-2024-08-06) experiments. For other experiments, we implement model inference with vLLM (Kwon et al., 2023) using Hugging Face model names:

- meta-llama/Meta-Llama-3.1-8B-Instruct
- meta-llama/Meta-Llama-3.1-70B-Instruct
- mistralai/Mistral-7B-Instruct-v0.3
- meta-llama/Llama-2-70b-chat-hf

For 7B/8B models, we use 1 NVIDIA RTX A6000 GPU and 32GB of RAM, and processing each question takes approximately about 5 seconds. For 70B models, 8 NVIDIA RTX A5000 GPUs and 64GB of RAM are used, and approximately each question takes 40 seconds. The temperature is set to be 0 for all experiments. The metric computation is highly efficient, taking only 5-10 minutes on a single CPU for datasets of moderate size. Our metric is a single-clue metric that evaluates models based on their confidence values. The human-in-the-loop process introduces variability, they are consistent with the cost of crafting other common QA datasets. We use NLTK, SpaCy, regex, and PEDANTS packages for data pre-processing of the collected buzzpoints.

I License Details

GRACE is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA. When using this dataset, please attribute as follows:

> GRACE is licensed under CC BY 4.0. To view a copy of this license, visit `https://creativecommons.org/licenses/by/4.0/`.

J CALSCORE Normalization

For our CALSCORE, we apply a normalized sigmoid transformation:

$$\text{CALSCORE}(x) = 1 - r\left(\mathbb{E}\left[(1 - h_t)g_t c_t\right]\right).$$

$r(x)$ is a normalized sigmoid function designed to map an expected value from a [-1, 1] range to a [0,1] range.

$$r(x) = \frac{\sigma(x) - \sigma(-1)}{\sigma(1) - \sigma(-1)},$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

K CALSCORE analysis by category

With logit-based CALSCORE scores, GPT-4o performs best in Arts and Science, while LLaMA-2-70B-Chat has the highest error in Arts. GPT-4o also excels in History and Literature, whereas LLaMA-2-70B-Chat struggles most in Literature.

In verbalized-based CALSCORE scores, GPT-4o achieves the lowest calibration errors across Arts and Science. Meanwhile, Mistral-7B-Instruct and LLaMA-2-70B-Chat show the highest errors in Literature, respectively. Verbalized outputs generally show improved calibration compared to logit-based confidence for Mistral and Llama-3 models.

L CALSCORE Walkthrough

In Table 6, CALSCORE first builds on unadjusted model calibration (MCE), which does not incorporate human data (Section 4.2). MCE penalizes confidently wrong answers by multiplying correctness (1 or -1) with model confidence (0–1); this captures models that are confidently wrong, when correctness is negative and confidence is high. CALSCORE then incorporates human data by the proportion of humans yet to answer correctly, reflecting calibration relative to human performance. The overall CALSCORE is computed by averaging the individual scores for each clue (CALSCORE_c). Given the per-clue values of -0.27, -0.08, -0.35, and 0.09, the resulting score is -0.61. The final CALScore is 0.18, obtained by applying the normalization (0.82), which is then subtracted from 1.

M CALSCORE^2 Walkthrough

Table 7 presents a step-by-step breakdown of a question where both system and human buzzes are tracked across multiple time steps. At each clue t, the system produces a guess with associated confidence and a probability of buzzing (b_t). Simultaneously, we track the cumulative proportion of human participants who have correctly answered by that time (h_t).

In clue 0, the system guesses incorrectly with low confidence, and only one out of three humans has buzzed correctly ($h_0 = 0.1$, approximating $1/3$). The human-adjusted score SH_t is zero, as the system's guess is incorrect. In clue 1, the system correctly guesses "Orwell" with higher confidence (0.7), while the cumulative human correct rate increases to $h_1 = 0.1$. Since the system buzzed before most humans had answered correctly, it receives partial credit: $SH_1 = 0.06$.

By clue 2, more humans have answered correctly ($h_2 = 0.2$), and the system again guesses correctly, further increasing the cumulative adjusted score. Finally, in clue 3, both the system confidence (0.9) and buzz probability ($b_t = 1.0$) are high, but a larger proportion of humans ($h_3 = 0.3$) have also answered correctly by this point. As a result, the final score $SH_3 = 0.02$ is lower, despite the correct answer, since the system was slower than several humans. The parenthetical value in $SH_3 = 0.02$ (0.27) represents the *unadjusted* score based solely on correctness and confidence, illustrating the penalty for delayed buzzing.

The reward for CALSCORE^2 for the question is reported at the bottom of the table as 0.53, with the value in parentheses (0.77) denoting the unadjusted baseline. This illustrates how incorporating human buzz data provides a more realistic measure of model competitiveness under time pressure. In we CALSCORE^2, we subtract the reward term from 1 to quantify the model's calibration error.

Category	GPT-4		GPT-4o		Mistral-7b-Instruct		LLama-2-70b-Chat		Llama-3.1-8B-Instruct		Llama-3.1-70B-Instruct	
	Logit	Verbalized	Logit	Verbalized	Logit	Verbalized	Logit	Verbalized	Logit	Verbalized	Logit	Verbalized
Arts	0.6646	0.564	0.5946	0.5592	0.7489	0.7889	0.8063	0.7218	0.6623	0.7466	0.6317	0.6774
Geo/CE	0.618	0.5465	0.6506	0.5795	0.7193	0.7864	0.8067	0.6872	0.6182	0.6857	0.6307	0.6819
History	0.709	0.5874	0.691	0.6421	0.7904	0.8022	0.8402	0.7478	0.6844	0.803	0.6764	0.7278
Literature	0.7315	0.628	0.7233	0.673	0.7713	0.7773	0.8332	0.7691	0.6861	0.7944	0.6935	0.7162
RMPSS	0.6676	0.5888	0.6456	0.5812	0.8029	0.8128	0.8514	0.755	0.6982	0.8082	0.6669	0.7083
Science	0.5978	0.5584	0.6019	0.5425	0.7256	0.7691	0.792	0.7103	0.6567	0.7325	0.5883	0.5928

Table 5: Verbalized and logit-based CALSCORE per category. GPT-4 and GPT-4o, the strongest models, struggle with literature and history relative to other categories. Most models are strongest at science. For GPT-4, GPT-4o, and Llama-2-70b, logit-based calibration consistently exhibits higher error; for Mistral and the Llama-3 models, verbalized calibration exhibits higher error.

Clue	CALSCORE$_c$	Text	Model Guess	Correct	Confidence
0	-0.27	He was born Eric Arthur Blair…	Dickens	-1	0.3
1	-0.08	This British writer is known for his dystopian themes…	Lorca	-1	0.1
2	-0.35	He coined the term “doublethink” and envisioned a regime where “Big Brother” watches everyone.	Marx	-1	0.7
3	0.09	His most famous works include *Animal Farm* and *1984*.	Orwell	1	0.9

Table 6: The overall CALSCORE is computed by averaging the individual scores for each clue. Given the per-clue values of -0.27, -0.08, -0.35, and 0.09, the resulting score is -0.61. The final CALSCORE is 0.18, after normalization.

Clue (t)	Question	System Guess	Conf.	h_t	b_t	SH_q
0	This author talked about his time fighting fascists in his autobiographical book *Homage to Catalonia*.	Lorca	0.1	0.1	0.1	0
1	He discussed his poverty in his essay “How the Poor Die” and *Down and Out in London and Paris*.	Orwell	0.7	0.1	0.63	0.06
2	The character of Old Major represented Lenin in his allegory *Animal Farm*.	Orwell	0.8	0.2	0.22	0.17
3	For ten points, name this author of *1984*.	Orwell	~~0.9~~ 1.0	0.3	0.05	0.27

Table 7: Clue-by-clue question details with model guesses, confidence scores, human and model buzz probabilities, and human-adjusted model scores (SH_t). CALSCORE^2 computes the probability of a system buzzing before the humans correctly answer the question. The resulting score is 0.77.