Confidence Calibration in Large Language Models

Anonymous ACL submission

Abstract

We investigate the calibration of large language models' (LLMs') confidence across diverse tasks. The results of our preregistered study show that the current crop of LLMs are, like people, too sure they are right: confidence exceeds accuracy, on average. However, this tendency toward overconfidence is moderated by a powerful hard-easy effect, wherein overconfidence is greatest on difficult tests; by contrast, easy tests actually show substantial underconfidence. We develop *LifeEval*, a test for evaluating model calibration across levels of difficulty.

1 Introduction

Large Language Models have seen widespread adoption due to their ability to provide useful information through natural language (Bick et al., 2024). However, LLMs' usefulness as guides, teachers, and advisers depends on their provision of truthful and accurate information (Afroogh et al., 2024). Hallucination, in which an LLM confidently reports falsehoods, fundamentally undermines their value (Kalai et al., 2025). That is why a proviso warns ChatGPT users, "ChatGPT can make mistakes. Check important info" (OpenAI, 2025a). Other LLMs come with similar warnings.

Ideally, an LLM ought to provide only truthful information. This is, of course, unrealistic for at least two reasons. First, it ignores the complexity of irreducible uncertainties. Few things can be known with certainty and perfect Bayesian rationality only provides probabilistic credences. Second, it neglects the limits in the LLM's information (Tripathi et al., 2025). LLMs generally lack access to verifiable ground truth, and must rely on the imperfect information available to them.

Accepting these constraints, a more realistic possibility is well-calibrated confidence. That is, the LLM should be able to faithfully report the probability that it is correct, conditional on its own limitations and vulnerability to error. This would allow users to rely on an LLM's stated confidence. To do so, users must trust that confidence indicates accuracy. This trust is essential to enable autonomous systems to know when they are too uncertain to to take action, among many other uses. Without well-calibrated confidence, users might trust faulty outputs or doubt accurate outputs. Hallucination and miscalibration are therefore epistemic risks that cut to the very heart of the usefulness of AI.

This motivates our tests of the confidence calibration of commercially available LLMs in a variety of contexts. This work presents an analysis of 11 popular open- and closed-source LLMs on a variety of reasoning tasks. We find that:

- LLMs are, on average, overconfident.
- Models are more overconfident on hard tasks and underconfident on the easiest tasks.
- Reasoning models provide more nuanced confidence estimates.

Moreover, we add to the current literature by proposing a new test for measuring model calibration on Bayesian-inference tasks: *LifeEval*. This framework allows for:

- A continuous measure of task difficulty grounded in empirical probabilities.
- Monotonic scaling of task difficulty.
- Evaluation of model performance based off of quantitative elements of the problem at hand rather than qualitative ones.

2 Related Work

Human judgment is vulnerable to many biases, of which overconfidence may be the most consequential (Kahneman, 2011). Well-calibrated confidence is foundational to effective decision making, since committing to a course of action requires sufficient confidence in its consequences. Yet the calibration of human confidence judgments is notoriously poor. People are overconfident and confidence

judgments exhibit a "hard-easy" effect: overconfidence increases with difficulty, while underconfidence emerges on easier tasks (Lichtenstein and Fischhoff, 1977).

The most parsimonious explanation for the hard-easy effect is that it can be explained as a regression-to-the-mean artifact, a byproduct of the noisy relationship between confidence and accuracy (Boundy-Singer et al., 2023; Krueger and Mueller, 2002). Changes in difficulty have a more direct influence on accuracy than on confidence (Erev et al., 1994). As task difficulty increases, performance drops, but if confidence is imperfectly responsive to this drop in accuracy, overconfidence must grow. Conversely, as a task becomes easier and performance increases, noisy confidence judgments produce underconfidence.

Other explanations for confidence biases emphasize motivational factors (Brown, 2012; Kruger and Dunning, 1999). We might hope that artificially intelligent agents might be less biased by motivational factors and would therefore exhibit better-calibrated confidence. On the other hand, if LLMs' confidence is, as with people, a noisy signal of accuracy, then we might expect to see similar confidence biases. Evidence suggests deep neural networks are routinely more certain than they are accurate (Oelrich et al., 2020; Abdar et al., 2021), and are often poorly calibrated (Guo et al., 2017; Xu et al., 2025a). Nevertheless, recent work has suggested that large language models might overcome these weaknesses through their increasing sophistication (Kadavath et al., 2022b; Xiao et al., 2025; Leng et al., 2025; Chhikara, 2025; Li et al., 2025).

Some prior work has documented variation in confidence calibration, accompanied by variation in difficulty. As our own work will show, models are generally more overconfident on more difficult tasks. Investigating model calibration therefore requires variation in task difficulty. Prior methods have sought to assess difficulty of tasks through one of four approaches: (1) intuitive human assessment of that difficulty, (2) LLM as a judge (Hwang et al., 2025; Gobara et al., 2024), (3) scaling the context provided (Sung et al., 2025), or (4) a mixture of these methods. Unfortunately, these approaches rely on subjectivity of the annotator, model, or question author respectively. Tasks that are difficult for humans can be quite easy for frontier LLMs (Luong et al., 2025) while models may struggle with mundane human tasks (Philip and Hemang, 2024). Additionally, like humans, models are subject to their own biases which may influence their rating of task difficulty (Tabib and Deedar, 2025). Scaling the amount of context provided can mitigate some of these issues; however, it is not clear how each piece of context may impact overall difficulty. Because of this, simply adding or removing more context may not reflect the true intellectual difficulty. Furthermore, in almost all cases, evaluations rely on a coarse measure of difficulty rather than a continuous one. This leads to artificial jumps in results as a question which is barely ranked as a "medium" is treated the same as one which is almost ranked as "hard." In contrast, a continuous measure of difficulty allows for a more precise analysis of model calibration and a greater understanding of how difficulty relates to overall calibration.

We contribute to this literature by first systematically studying confidence calibration in 11 large language models across five different tests. Some of these tests are more aligned with models' abilities than others, affording a post-hoc analysis of the hard-easy effect. To isolate the effect of difficulty from other task features, we develop a new task, *LifeEval*, that affords a bias-free manipulation of difficulty while holding constant other task characteristics. LifeEval asks for probabilistic confidence judgments that we then compare to empirical probabilities. This method incorporates the benefits of moderating difficulty while sidestepping the aforementioned constraints of previous approaches.

3 Method

Our plan used six English-based question sets (see Table 1) testing 11 large language models. Five of these are marketed as reasoning models: DeepSeek-R1 (DeepSeek, 2025), Gemini 2.5 Pro (Google, 2025b), GPT-o3 (OpenAI, 2025b), Claude Sonnet 4 (Anthropic, 2025b), and Claude Sonnet 3.7 (Anthropic, 2025a).[1] We compared these models to six "chat" models: DeepSeek-V3 (DeepSeek, 2024), Gemini 2.5 Flash (Google, 2025a), GPT-4o (OpenAI, 2024) and Claude Haiku 3 (Anthropic, 2024) as well as two locally-run, instruction-tuned, versions of Llama 3.1 (8B and 70B) (Meta, 2024a,b).

Following Kadavath et al. (2022a), we asked models to state the confidence they were correct,

[1] In the interest of increasing the credibility of our results we preregistered our research plans. This preregistration precommited us to conducting and reporting a set of planned analyses. Appendix A explains deviations from our preregistered plans.

Question Set	Length	After Cleaning	Answer Fields	Context	Task Description
BoolQ	3270	2503	True, False	None	True or False trivia questions
SciQ	1000	995	A, B, C, D	None	Scientific trivia questions
LSAT-AR	230	86	A, B, C, D, E	Premise for logical question	Find optimal solution for complex situation
SAT-EN	206	173	A, B, C, D	Passage	Answering questions about a reading passage
HaluEval-QA	2000	1790	N/A	Short passage	Pre-generated Model responses to questions
LifeEval	808	751	Predicted age at death	Sex, Minimum Age	Estimate age at death given minimum age and sex

Table 1: The six question sets.

and also recorded token-level probabilities where available. For multiple choice questions sets, we had models state the probability that each answer option was correct. Open-weight models such as Llama 3.1-8B and 70B[2], allowed us to analyze both stated confidence and token probabilities from final-layer logits. Moreover, for GPT-4o we were able to get the top 5 token probabilities and used that in place of a full token distribution. We fixed temperatures at 0 with deterministic, greedy-decoding, except where constrained by API limitations (e.g., GPT-o3 enforced temperature = 1.0 and did not provide logprobs).

Each model/question-set pairing yields confidence distributions, accuracy averages, and calibration metrics. We counted a response as correct if the answer option assigned the highest probability matched the ground truth (with proportional scoring for ties). We compared confidence to observed accuracy to compute calibration statistics, most centrally Expected Calibration Error (ECE) (Naeini et al., 2015) and overconfidence. We evaluated all models under identical conditions so that observed differences in calibration or overconfidence could be attributed to the model.

We designed our prompts to give our models the best chance at good calibration (Yang et al., 2024; Zhou et al., 2025; Xu et al., 2025b). For each question set, we employed one-shot prompting that instructed the model to return its output in JSON format. Except for HaluEval, we also incorporated a chain-of-thought prompting strategy to encourage more faithful, step-by-step reasoning. We repeated the system prompt within the input to reinforce adherence to the formatting rules. In the case of multiple choice questions (MCQ), we prompted models to select an answer and state the likelihood that each option is correct. This allows us to not only observe the confidence assigned to the response but also the distribution of confidence in other answer options.

[2]Hardware constraints forced us to run Llama-70B in quantized form. We used 4-bit quantization and upscaled the weights to 16 bits for computation.

4 Question Sets

We selected a variety of different question types intended to capture a spectrum of conditions under which calibration can succeed or fail. By examining calibration across these types of questions, we seek a comprehensive understanding of model calibration.

Some questions, like true/false items, entail a two-alternative forced choice (so-called 2AFC formats). Peak scoring focuses on the favored option and the agent's confidence that it is correct. It is standard practice to assign responses to bins that subdivide the range of confidence (Moore et al., 2015; Keren, 1988). This affords the calculation of overconfidence and computed ECE over bins. Table 1 contains a brief description of each question set used in our analysis.

4.1 BoolQ and SciQ

To measure model calibration in general knowledge, we used 1000 multiple choice questions (MCQ) from the SciQ dataset (Johannes Welbl, 2017) as well as 3,270 True/False questions from the BoolQ dataset (Clark et al., 2019). We scored models against ground truth for each question.

4.2 LSAT-AR

To evaluate calibration in logical reasoning, we used 230 questions from the LSAT Analytical Reasoning section (Zhong et al., 2021). Each question contained five multiple choice answer options. These tasks required multi-step reasoning, rule application, and inference, making them well-suited for testing whether models' confidence appropriately degrades as logical complexity increases. We compared stated confidence and token-based probabilities against correctness to compute ECE and overconfidence for each model.

4.3 SAT-EN

For contextual understanding, we evaluated models on 1000 passage-based inference questions drawn from the SAT English section (Zhong et al., 2023). Each passage was accompanied by multiple choice comprehension questions requiring information extraction, inference, and reasoning about nuanced textual details. Our measure of calibration compares model confidence with actual accuracy across varying levels of passage complexity. This allowed us to test whether models maintain appropriate confidence when the answer depends on subtle contextual cues.

4.4 HaluEval

To assess confidence where LLMs are prone to hallucinating, we drew on the HaluEval question set (Li et al., 2023): 2000 question-answer pairs, 1000 truthful answers and 1000 hallucinated answers. Here, we prompted models not to produce an answer, but instead to state their confidence in the given answer's correctness. Because half of these answers were deliberately hallucinated, this setting provided a direct test of whether models could recognize and signal their own fallibility. Calibration compared stated confidence with correctness.

4.5 LifeEval

We developed a new question set, which we call LifeEval, to manipulate difficulty while holding constant the nature of the question. LifeEval asks models to predict the lifespan of a person given their age and sex. Models were then asked to report the probability that their estimate would fall within one of several radii (1, 5, 10, or 20 years) of the true lifespan. We assessed actual probability using U.S. Social Security Administration Period Life Tables (Social Security Administration, Office of the Chief Actuary, 2025). Manipulating radius, age, and sex enabled us to vary task difficulty holding all else constant. Unlike other question sets currently available, LifeEval distinguishes itself from existing benchmarks by providing a gradient of difficulty that the model can actively detect. For instance, if the model is told that a male has already lived 80 years, it can be confident in its guess landing within a 20-year radius of the truth. However, if it only knows the sex and must get within 1 year, it should be clear to the model that the actual probability is low. We use the probability of success given the optimal answer as a measure of task difficulty for our analysis. If there exists an answer to a question that can theoretically capture 100% of the mass of the conditional distribution, we can consider that question easier than one where the theoretical maximum is only 20%. By leveraging the empirical nature of LifeEval, we can understand the difficulty of a question as 1 minus its Maximum Achievable Score (MAS) as seen in Figure 5. This approach affords us another lens through which to understand model calibration: how does a model react to changing task difficulty? While there exist several methods for quantifying task difficulty they all come with their own drawbacks as discussed in Section 2. For LifeEval we computed accuracy differently from the other question sets, because we knew the true probabilities against which we could compare the LLMs' responses. For a given question, let a be the minimum age (e.g. 25), s be the sex, and r be the radius around the model's guess. Suppose a model guessed $\hat{y}(a, s)$ and has confidence $c(a, s, r)$ that $\hat{y}(a, s)$ is within a radius r of the correct outcome. We justify why our approach fits into the framework of the other question sets as follows: Imagine for a moment we had a large set of people Q where person $i \in Q$ died at age y_i. Focusing on the subset $Q_{as} = \{i : y_i \geq a, \text{sex}(i) = s\}$ of people of sex s who lived until at least the age of a, we can think of this as a binary question of asking about whether the true y_i falls in the interval with

$$\text{accuracy}(Q_{as}) = \frac{1}{|Q_{as}|} \sum_{i \in Q_{as}} \mathbb{I}\{y_i \in [\hat{r}^-, \hat{r}^+]\} \tag{1}$$

where $\mathbb{I}$ is the indicator function, $\hat{r}^- = \hat{y}(a, s) - r$, and $\hat{r}^+ = \hat{y}(a, s) + r$. Imagining $|Q_{as}| \to \infty$, we have $\text{accuracy}(Q_{as}) \to p(\hat{y}(a, s), r | a, s)$, where $p(k, r|a, s)$ is defined as

$$\mathbb{P}\big(y \in [k - r, k + r] \big| y \geq a, s\big) \tag{2}$$

By taking advantage of the actuarial life tables from the social security administration (Social Security Administration, Office of the Chief Actuary, 2025), we compute $p(k, r|a, s)$ as:

$$p(k, r|a, s) = \sum_{i=k-r}^{k+r} S_i(a, s) \cdot q_i(s) \tag{3}$$

Where $q_i(s)$ is the probability of death for a person of a given sex to die at age i once they

become i years old and $S_i(a, s)$ is the conditional probability that someone lives at least to age i given sex and minimum age. While $q_i(s)$ is provided by the life tables, we can compute

$$\begin{aligned} S_i(a, s) &= \mathbb{P}(\text{live to age } i | a, s) \\ &= \prod_{j=a}^{i-1} \big(1 - q_j(s)\big). \end{aligned} \tag{4}$$

5 Confidence Scoring

At its core, confidence calibration quantifies the alignment between subjective probability and objective accuracy. When a model assigns a subjective probability of 80%, good calibration demands it is correct 80% of the time (Dawid, 1982).

We test models' confidence in two ways: what a model *says* and what it *thinks*. The latter measure comes from final layer token probabilities, which are only available for open-source models or models that provide access. We concede that next token probabilities are not necessarily a model's certainty of a given answer being correct and may be impacted by the reasoning that precedes the answer token generation. However, we evaluated both to investigate which provides a better calibrated estimate of uncertainty.

5.1 Stated Confidence

For all question sets, we prompted models to provide a numerical confidence score from 0 to 1.0 representing the probability they are correct. In the case of multiple choice questions, models assigned probabilities to each of the answer options. In the instances where the provided probabilities did not sum to 1 we obtained normalized probabilities P_i as follows:

$$P_i = \frac{s_i}{\sum_{j \in S} s_j}, \tag{5}$$

where S is the set of options and each s_i represents a stated confidence for a given option.

5.2 Confidence via Token Probability

For models that return token probabilities in the form of either raw logit scores or log-probabilities over the top k tokens, we normalize over the viable tokens to compute a proper probability distribution. For example, if a model returns log-probabilities for only the top k tokens and the correct answer is among them, we exponentiate the log-probabilities, restrict to the subset of relevant tokens, and then normalize as follows:

$$P_i = \frac{p_i}{\sum_{j \in T} p_j}$$

where $p_i = e^{\ell_i}$ is the exponentiated log-probability for token i, and T is the set of target tokens in the output (e.g. ['A','B','C', ...] for multiple-choice questions). This ensures that the resulting probabilities sum to 1 over the restricted set and are comparable across examples.

6 Metrics

For any multiple choice question set Q, we let $y_i, \hat{y}_i \in \{1, \ldots, K\}$ denote the correct option and the model's chosen option, respectively, for question $i \in Q$ (where K denotes the number of choices). We let $C_i(k) \in [0, 1]$ denote the model's confidence that option k is correct, where $\sum_{k=1}^{K} C_i(k) = 1$. The confidence $C_i(k)$ can refer to either stated confidence or token-probability-based confidence, depending on context.

6.1 Expected Calibration Error (ECE)

Expected Calibration Error (ECE) quantifies the misalignment between a model's predicted confidence and its empirical accuracy (Pavlovic, 2025; Naeini et al., 2015). We first partition a question set Q into M disjoint bins by confidence:

$$Q_m = \big\{ i \in Q : \frac{m-1}{M} < C_i(\hat{y}_i) \leq \frac{m}{M} \big\}$$

for $m = 1, \ldots, M$. We then compute

$$\text{ECE}(Q) = \frac{1}{|Q|} \sum_{m=1}^{M} n_m \cdot \big| \text{acc}(Q_m) - \text{conf}(Q_m) \big| \tag{6}$$

where n_m is the number of questions in Q_m. Probabilities were grouped into ten equally spaced intervals from [0, 1), with an additional bin dedicated to the value 1.0. This eleventh bin identifies those distinctive instances in which a model reports absolute certainty by assigning a probability of exactly 1.

6.2 Overconfidence

Since ECE does not reveal whether miscalibration is due to over- or underconfidence, we needed a separate measure of overconfidence. We borrowed

from previous works in psychology (Klayman et al., 1999) to define overconfidence over an entire question set Q as

$$\text{overconfidence}(Q) = \text{conf}(Q) - \text{acc}(Q). \quad (7)$$

7 Results

Models report **88%** confidence on average in their favored answer option being correct. They are, in fact, correct for **79%** of questions. The calibration plot shown in Figure 1 reveals that there is a strong positive relationship between confidence and accuracy. The diagonal identity line reflects perfect calibration. Observations to the southeast of the identity line, where confidence exceeds accuracy, indicate overconfidence.

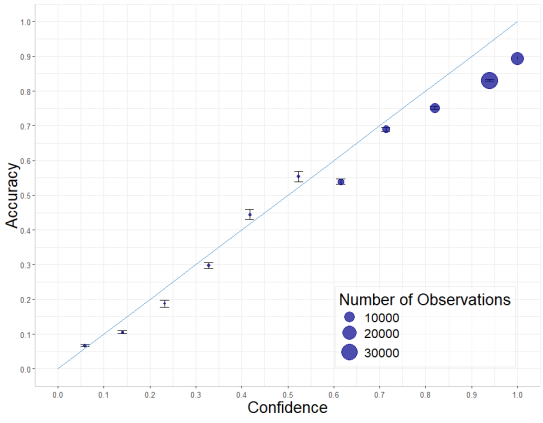

Figure 1: Calibration plot, showing accuracy conditional on confidence; observations are averaged within confidence bins, [0,0.1),[0.1,0.2)...[0.9,1),[1].

We find that models' tendency toward overconfidence varies by question set. Figure 2 shows that models tended to be overconfident on some tasks like logical reasoning and hallucination detection (*LSAT-AR* and *HaluEval* respectively). Models struggle to think through complex tasks and fail to adequately detect when they have gone astray. While accuracy was fixed for HaluEval at 52.12% and LifeEval had an upper bound of 56.8% we found that LSAT-AR proved to be the most difficult of all the unbounded question sets with an average accuracy of **58.6%**. By contrast, while models excelled at SciQ and SAT-EN, they remained consistently underconfident. The sole outlier was Llama-3.1-8B, likely due to its significantly lower parameter count.

LifeEval's four levels of difficulty afforded insight into how task difficulty affected confidence

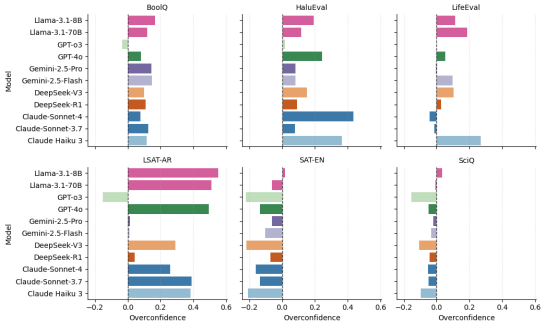

Figure 2: Overconfidence by question set and Model.

calibration. For the lowest radius (most difficult) tasks, models reported average confidence of **34.2%** but actual probability was only **9.6%**. Yet, like humans, models displayed a tendency towards underconfidence when the task got easier (i.e. the radius increased). Models reported **80.5%** confidence but actual probability was **92.0%** for their 20-year radius responses.

Comparing overconfidence by radius in Figure 3 reveals that as task difficulty increases (i.e., radius decreases), overconfidence increases. This suggests that models' reported confidence was insufficiently sensitive to variation in task difficulty.

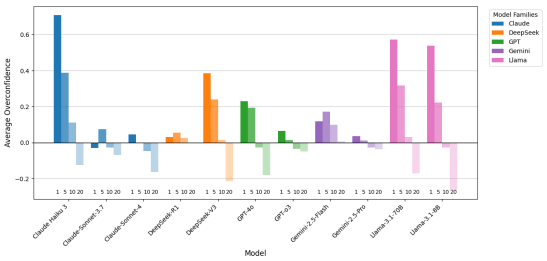

Figure 3: Overconfidence as a function of model and radius; difficulty decreases with larger accuracy radius.

When analyzing the stated confidence values for LifeEval, we found a disparity between larger reasoning models like DeepSeek-R1, which tended to provide more nuanced estimates while their smaller siblings provided less nuanced reports, as seen in Table 2. We found that models tended to imitate human confidence reporting by rounding to the nearest 5%. Stated confidence was a multiple of 5% for **91.4%** of reports by non-reasoning chat models, with some reporting a multiple of 5% for all of their responses. By contrast, only **61.1%** of stated confidence was a multiple of 5% for reasoning models. This highlights a key flaw in uncertainty estimation via model reporting. By design, models imitate human behavior; this extends to confidence reporting

Model	Type	Score (%)	ECE	Conf. (%)	% Rnd	Regression Coef.	N
Claude-Sonnet-3.7	Reasoning	54.5	0.040	53.1	90.1	0.180	808
Claude-Sonnet-4	Reasoning	54.0	0.063	49.8	98.8	0.327	808
DeepSeek-R1	Reasoning	54.4	0.031	57.2	29.0	0.053	808
Gemini-2.5-Pro	Reasoning	53.8	0.025	53.4	18.0	0.092	808
GPT-o3	Reasoning	54.2	0.029	54.1	69.8	0.189	761
Aggregate		54.2	0.037	53.5	61.1	0.168	751
Claude Haiku 3	Chat	53.0	0.267	79.8	100	0.996	808
DeepSeek-V3	Chat	53.3	0.124	63.7	100	0.782	808
Gemini-2.5-Flash	Chat	53.8	0.098	63.6	48.9	0.192	808
GPT-4o	Chat	54.5	0.085	59.8	100	0.604	808
Llama-3.1-70B	Chat	53.5	0.185	72.0	99.5	0.874	807
Llama-3.1-8B	Chat	48.4	0.142	59.9	100	0.941	800
Aggregate		52.8	0.150	66.5	91.4	0.732	751

Table 2: Performance metrics across various models on LifeEval split by model type. We report Mean Score, Expected Calibration Error (ECE), Mean Confidence, Percentage of Rounded outputs, the Regression Coefficient between difficulty and overconfidence, and number of completions (N). LifeEval has a mean Maximum Achievable Score (MAS) of 56.80%. We ran a regression comparing Overconfidence and question difficulty $(1 - MAS_{question})$. A higher regression coefficient implies the increased prevalence of the Hard-Easy effect. Aggregate rows take mean of all columns except for N which is the size of the subset of questions answered by all models (Reasoning & Chat).

where, like humans (Wallsten et al., 1993), models tend to avoid precision.

We observe stronger correlations between stated confidence and actual probability of being correct for reasoning models. This correlation is, on average, 0.94 for the five reasoning models (GPT-o3, Deepseek-R1, Gemini 2.5-Pro and 2.5-Flash, and Claude Sonnet-3.7 and Sonnet-4) and only 0.48 for the other models.

Finally, for GPT-4o and both versions of Llama-3.1 we were able to compare the stated confidence to the token probability of the answer for each question. We found that, in general, ECE for stated confidence is fairly aligned with token probabilities, with stated confidence being slightly better calibrated than token probabilities. This is likely due to token probabilities not being intended to capture the probability of correctness while stated confidence takes into account a much more holistic view. In contrast to stated probabilities that are often a multiple of 5%, less than 1% of token probabilities are multiples of 5%. We found that token probability tended to be higher than the stated confidence value. For a comparison by question set, see Figure 8 in the Appendix.

8 Discussion

We document successful metacognition in large language models. LLMs can faithfully report their confidence in the accuracy of their judgments. Nevertheless, we find 9% overconfidence (the difference between models' 81% stated confidence and their 72% accuracy). This result fits in the range of human overconfidence documented across a variety of tasks and settings: 30% overconfidence on the economic forecasts of the Survey of Professional Forecasters (Campbell and Moore, 2024); 15% overconfidence among laypeople answering general knowledge questions (Lichtenstein and Fischhoff, 1977); and just 3% overconfidence among trained forecasters scored with accuracy-compatible incentives predicting geopolitical events (Moore et al., 2017).

Confidence corresponds more closely to the actual probability of being correct for self-reflective reasoning models. Nevertheless, confidence reported by the models in our study is less responsive than is accuracy to variations in difficulty. Consequently, we observe the "hard-easy" effect documented in human confidence judgments: Overconfidence rises with difficulty (Erev et al., 1994). This is particularly evident in LifeEval, a test we devised to provide exogenous variation in difficulty holding other task characteristics constant.

The hard-easy effect arises because confidence

is a noisy signal of accuracy. Confidence is less responsive to difficulty than is accuracy. So the overall overconfidence we observe is partly a result of test items sufficiently difficult to produce overconfidence. However, there is more to it than that. Finite fallible agents will sometimes be wrong because they do not know everything (Moore, 2023). For AI systems, out-of-sample judgments may include these unknown unknowns–that is, relevant knowledge the agent lacks but is not aware of.

It is possible to imagine a model refining its confidence in light of feedback. Commercial LLMs are refined through reinforcement learning with human feedback (RLHF). However, RLHF might actually increase the overconfidence models report (Tian et al., 2023). If human users prefer models that express confident assurance, RLHF may train the LLM to report greater confidence.

LLMs' competence and confidence have led many users to rely on them heavily and unquestioningly (Hou et al., 2025). This trust might be misplaced if models express greater confidence than their accuracy justifies. When models are faced with tasks for which they cannot perform as well, they maintain the same high level of confidence. As we see in LifeEval, models fail to sufficiently reduce their confidence as performance declines with task difficulty. For LLMs to deserve users' trust, they must be able to reliably report their limitations. Many users are aware of LLMs' impressive capabilities but are wary of adoption because of the unpredictable nature of hallucination. If users do not know when they need to seek additional resources they are forced to either constantly watch over the model or remove it from their workflow entirely.

9 Future Work

Psychological research distinguishes three forms of overconfidence in humans: overplacement is the exaggerated belief that you are better than others; overestimation is thinking you are better than you are; overprecision is the excessive certainty that you know the truth (Moore and Healy, 2008). We employed single-item confidence measures that ask, "How sure are you that this answer is correct?" These sorts of item-confidence measures perfectly confound overestimation and overprecision, since being too certain of your answer is the same as overestimating your chance of being correct. However, it is possible to unconfound these two using higher-order measures. One approach elicits an estimate of the respondent's score on some test, and their certainty about that estimate. This affords the possibility of being excessively certain of an underestimate, such as the student who is convinced she failed an exam when in reality she passed. Future research should distinguish between these different forms of overconfidence in LLMs.

Future research should further examine how language models perform on Bayesian inference tasks. As LifeEval demonstrates, models consistently struggle to appropriately reduce their confidence as tasks become more difficult. Investigating this limitation across different domains may help illuminate the underlying causes and potential remedies.

In humans, one of the most useful general-purpose debiasing strategies is getting people to reflect on why it is they might be wrong (Lord et al., 1984). More specifically, inviting people to consider what information they lack helps them moderate their tendency toward overconfidence (Walters et al., 2017). The better performance of the reasoning models we examine constitutes a striking parallel. It is possible that prompts or training regimens that encourage models to engage in more reflection and self-criticism could further improve calibration and reduce overconfidence.

Limitations

Some of the question sets we used contained errors. Some of these errors were typos that made the question more ambiguous. Other errors cut off text or omitted figures referenced in the questions. We chose to keep these questions for two reasons. First, rewriting or labeling questions was problematic because it would have introduced differences from the source and reduced comparability to other studies using the same question sets. Second, we felt that the existence of such questions did not detract from our work as models ought to be well-calibrated regardless of a user's prompt. If a model does not understand a question, it should express a lower confidence in its response. Therefore, when considering key metrics such as ECE, a model's overall calibration should be largely independent from the existence of a few unclear questions.

In their responses to LifeEval questions, some models provided reasoning that referenced SSA tables. We felt that this was immaterial as regardless of having access to the data, its confidence should

be well calibrated. The over- and under-estimates we observe show that even when models have sufficient context they may not be able to properly provide well-calibrated confidence estimates

For GPT-4o we were constrained to only the top 5 tokens. This meant that there were many instances where we did not have access to a certain answer option. Without this, we opted to assign a value of 0. While the true value is invariably higher, as this is for the least desired options we don't expect this practice to impact our results in any way.

Finally, the size of our question sets varied, as seen in Table 1. This does not undermine the validity of our model-to-model comparisons. However, our overall results (such as the aggregate calibration plot in Figure 1) does overweight BoolQ. See Figure 6 in the Appendix for a more granular view of model calibration on each question set.

Ethical Considerations

One of the authors is a Visiting Faculty Researcher at Google, which created some of the LLMs analyzed in this work; however, this manuscript's work was conducted as part of their employment at a university, not at Google.

While the misuse of generative AI has become a growing issue in recent years, we do not see a way in which our work greatly exacerbates this issue. LifeEval only utilizes two demographic identifiers for a person: sex and a minimum age level. Although it would be ill-advised, if an inference provider chose to incorporate LifeEval or a similarly structured question set into their training data there would a possibility of model bias arising along these two axes. We encourage future researchers to exercise caution and transparency regarding the inclusion of such demographic markers in training pipelines.

Acknowledgments

To maintain anonymity during review, acknowledgments have been omitted.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297. ArXiv: 2011.06225.

Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. 2024. Trust in ai: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1):1568.

Anthropic. 2024. Claude haiku 3.

Anthropic. 2025a. Claude sonnet 3.7.

Anthropic. 2025b. Claude sonnet 4.

Alexander Bick, Adam Blandin, and David J Deming. 2024. The rapid adoption of generative ai. Working Paper 32966, National Bureau of Economic Research.

Zoe M. Boundy-Singer, Corey M. Ziemba, and Robbe L. T. Goris. 2023. Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, 7(1):142–154.

Jonathon D Brown. 2012. Understanding the better than average effect: Motives (still) matter. *Personality Social Psychology Bulletin*, 38(2):209–219. Citation Key: Brown2011.

Sandy Campbell and Don A Moore. 2024. Overprecision in the survey of professional forecasters. *Collabra:Psychology*, 10(1):92953.

Prateek Chhikara. 2025. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *Preprint*, arXiv:2502.11028.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *Preprint*, arXiv:1905.10044.

A. P. Dawid. 1982. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

DeepSeek. 2024. Deepseek v3.

DeepSeek. 2025. Deepseek r1.

Ido Erev, Thomas S Wallsten, and David V Budescu. 1994. Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3):519–527. Citation Key: Erev1994.

Seiji Gobara, Hidetaka Kamigaito, and Taro Watanabe. 2024. Do llms implicitly determine the suitable text difficulty for users? *Preprint*, arXiv:2402.14453.

Google. 2025a. Gemini 2.5 flash.

Google. 2025b. Gemini 2.5 pro.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *Preprint*, arXiv:1706.04599.

Irene Hou, Hannah Vy Nguyen, Owen Man, and Stephen MacNeil. 2025. The evolving usage of genai by computing students. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*, SIGCSE TS 2025, page 1481–1482. ACM.

Seonjeong Hwang, Hyounghun Kim, and Gary Geunbae Lee. 2025. Can llms estimate cognitive complexity of reading comprehension items? *Preprint*, arXiv:2510.25064.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022a. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022b. Language models (mostly) know what they know. (arXiv:2207.05221). ArXiv:2207.05221 [cs].

Daniel Kahneman. 2011. *Thinking fast and slow*. Farrar, Straus and Giroux, New York. Citation Key: Kahneman2011.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Gideon Keren. 1988. On the ability of monitoring nonveridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, 67(2):95–119. Citation Key: Keren1988.

Joshua Klayman, Jack B. Soll, Claudia González-Vallejo, and Sema Barlas. 1999. Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79(3):216–247.

Joachim I Krueger and Ross A Mueller. 2002. Unskilled, unaware, or both? the better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2):180–188.

Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134. Citation Key: Kruger1999a.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2025. Taming overconfidence in llms: Reward calibration in rlhf. (arXiv:2410.09724). ArXiv:2410.09724 [cs].

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. 2025. Conftuner: Training large language models to express their confidence verbally. *Preprint*, arXiv:2508.18847.

Sarah Lichtenstein and Baruch Fischhoff. 1977. Do those who know more also know more about how much they know? *Organizational Behavior and Human Decision Processes*, 20(2):159–183. Citation Key: Lichtenstein1977.

Charles G Lord, Mark R Lepper, and Elizabeth Preston. 1984. Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, 47(6):1231–1243. Citation Key: Lord1984a.

Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. 2025. Towards robust mathematical reasoning. *Preprint*, arXiv:2511.01846.

Meta. 2024a. Llama 3.1 70b instruct.

Meta. 2024b. Llama 3.1 8b instruct.

Don A. Moore. 2023. Overprecision is a property of thinking systems. *Psychological Review*, 130(5):1339–1350.

Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological Review*, 115(2):502–517.

Don A Moore, Samuel A Swift, Angela Minster, Barbara A Mellers, Lyle Ungar, Philip E Tetlock, Heather H J Yang, and Elizabeth R Tenney. 2017. Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, 63(11). Citation Key: Moore2017b.

Don A Moore, Elizabeth R Tenney, and Uriel Haran. 2015. *Overprecision in judgment*, page 182–212. Wiley, New York. Citation Key: Moore2014.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. volume 29.

Oscar Oelrich, Shutong Ding, Måns Magnusson, Aki Vehtari, and Mattias Villani. 2020. When are bayesian model probabilities overconfident? *arXiv:2003.04026*. ArXiv: 2003.04026.

OpenAI. 2024. Gpt-4o.

OpenAI. 2025a. Chatgpt.

OpenAI. 2025b. Gpt-o3.

Maja Pavlovic. 2025. Understanding model calibration–a gentle introduction and visual exploration of calibration and the expected calibration error (ece). *arXiv preprint arXiv:2501.19047*.

Philip and Hemang. 2024. Simplebench: The text benchmark in which unspecialized human performance exceeds that of current frontier models.

Social Security Administration, Office of the Chief Actuary. 2025. Period life table, 2022 (used in the 2025 trustees report). Web page. Presented by the Office of the Chief Actuary; accessed via SSA website.

Yoo Yeon Sung, Eve Fleisig, Yu Hou, Ishan Upadhyay, and Jordan Lee Boyd-Graber. 2025. Grace: A granular benchmark for evaluating model calibration against human calibration. *Preprint*, arXiv:2502.19684.

H. M. Shadman Tabib and Jaber Ahmed Deedar. 2025. Toward trustworthy difficulty assessments: Large language models as judges in programming and synthetic tasks. *Preprint*, arXiv:2511.18597.

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

Sahil Tripathi, Md Tabrez Nafis, Imran Hussain, and Jiechao Gao. 2025. The confidence paradox: Can llm know when it's wrong. (arXiv:2506.23464). ArXiv:2506.23464 [cs].

Thomas S Wallsten, David V Budescu, and Rami Zwick. 1993. Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, 39(2):176–190. Citation Key: Wallsten1993.

Daniel J. Walters, Philip M. Fernbach, Craig R. Fox, and Steven A. Sloman. 2017. Known unknowns: A critical determinant of confidence and calibration. *Management Science*, 63(12):4298–4307.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J. Su, and Li Shen. 2025. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. (arXiv:2505.01997). ArXiv:2505.01997 [cs].

Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2025a. Do language models mirror human confidence? exploring psychological insights to address overconfidence in llms. *Preprint*, arXiv:2506.00582.

Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2025b. Do language models mirror human confidence? exploring psychological insights to address overconfidence in llms. (arXiv:2506.00582). ArXiv:2506.00582 [cs].

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms. *Preprint*, arXiv:2412.14737.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2021. Ar-lsat: Investigating analytical reasoning of text. *Preprint*, arXiv:2104.06598.

Ziang Zhou, Tianyuan Jin, Jieming Shi, and Qing Li. 2025. Steerconf: Steering llms for confidence elicitation. *Preprint*, arXiv:2503.02863.

A Deviations from Pre-Registration

- **DeepSeek log probabilities.** DeepSeek did not provide usable token-level log probabilities, so logprob-based analyses for this model were omitted.
- **LifeEval scoring rule.** For *LifeEval*, we scored answers using the conditional true (actuarial) probability of age at death falling within a radius R around the point estimate (rather than a binary "within true range" indicator). This choice reduces threshold sensitivity and aligns the metric with probabilistic calibration.
- **GPT-o3 inference settings** For *GPT-o3*, we were constrained to a temperature of 1.0 and increased the total token budget to accommodate longer responses on complex question sets like LSAT-AR. We were also not able to obtain logprobs from model responses.

B Data Cleaning and Exclusion Criteria

Any response that we could not parse, whether due to improper formatting or incomplete output, was omitted from our analysis. A few responses to multiple-choice questions provided all zeroes for their stated confidence scores. As we did not define a procedure for handling such cases in our pre-registration, we chose to drop them from our analysis. In some cases (n = 54), models attempted to hedge their responses by saying "maybe" or "I'm not sure". Although we find these cases promising from the standpoint of human computer interaction, we did not assign a scoring rubric for such responses and felt it improper to do so post-hoc in order to measure calibration from these questions. Because of this, we chose to omit these cases from our analysis. To keep questions balanced across models, we further restricted evaluation to the subset of questions that every model answered successfully.

C Scoring

We define accuracy as

$$\text{accuracy}(Q) = \frac{1}{|Q|} \sum_{i \in Q} \mathbb{I}\{\hat{y}_i = y_i\}, \qquad (8)$$

where $\mathbb{I}$ is the indicator function, and confidence as

$$\text{confidence}(Q) = \frac{1}{|Q|} \sum_{i \in Q} C_i(\hat{y}_i). \qquad (9)$$

When $C_i(k)$ represents token probabilities and the LLM's temperature is 0, we have

$$\hat{y}_i = \arg\max_k [C_i(k)], \qquad (10)$$

in which case $C(\hat{y}_i) = \max_k C_i(k)$, but in general this is not true when other decoding strategies are used.

C.1 Scoring for HaluEval

For HaluEval, we determined whether a provided answer was correct ahead of time. Therefore, we scored each question based on the provided label such that

$$\text{accuracy}(Q) = \frac{1}{|Q|} \sum_{i \in Q} y_i. \qquad (11)$$

D AI Use Disclaimer

Generatve AI (ChatGPT, Gemini, RooCode) was used, in part, throughout this research project to aid the researchers in background research, generating and debugging code snippets, document formatting, and improving the readability of this text. The methodology, analysis, and findings presented are entirely the intellectual property of the researchers and had no origin from generative AI.

E LifeEval Plots

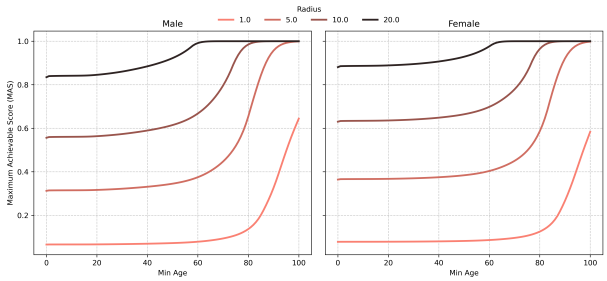

Figure 4: LifeEval allows for the monotonic decrease in task difficulty given age, sex, and radius. As the Maximum Achievable Score (MAS) increases, the task difficulty decreases.

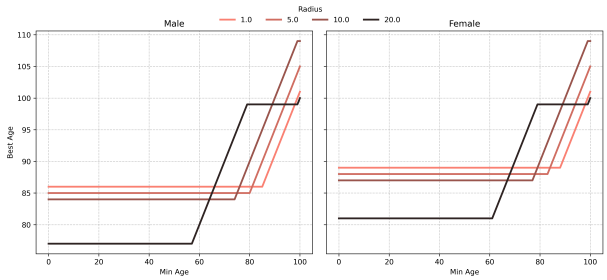

Figure 5: Best age to guess as a function of minimum age, gender, and radii. We see that the optimal age is constant until a certain minimum age is reached. Additionally, we see Female's have slightly higher overall life expectancy.

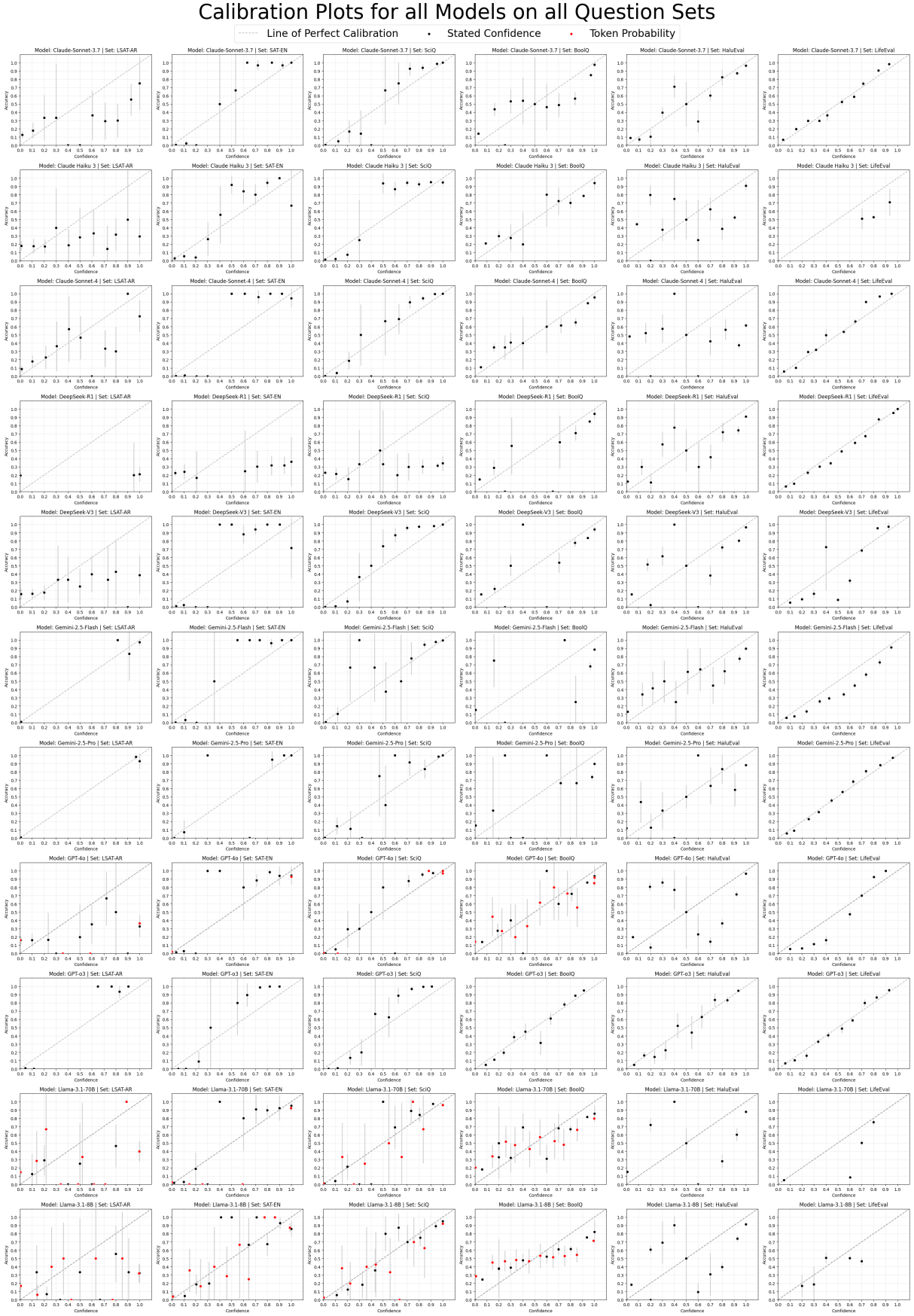

Figure 6: Side by side plots of all models (rows) and all question sets (columns). GPT-4o, Llama-3.1-70B, and Llama-3.1-8B all display their token probabilities in red.

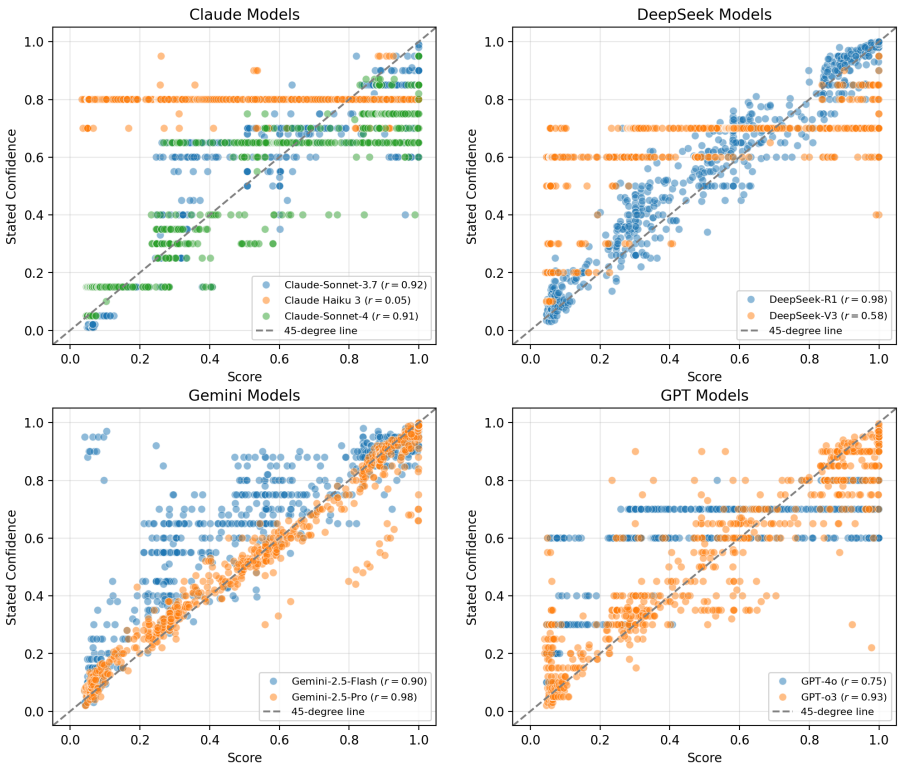

Figure 7: Each plot displays the relationship between Stated Confidence and actual Score for various model families. Each scatter plot illustrates how accurately a given family of models estimates their own performance to their true score. The prevalence of horizontal lines show the tendency for certain models to round their probability estimates.

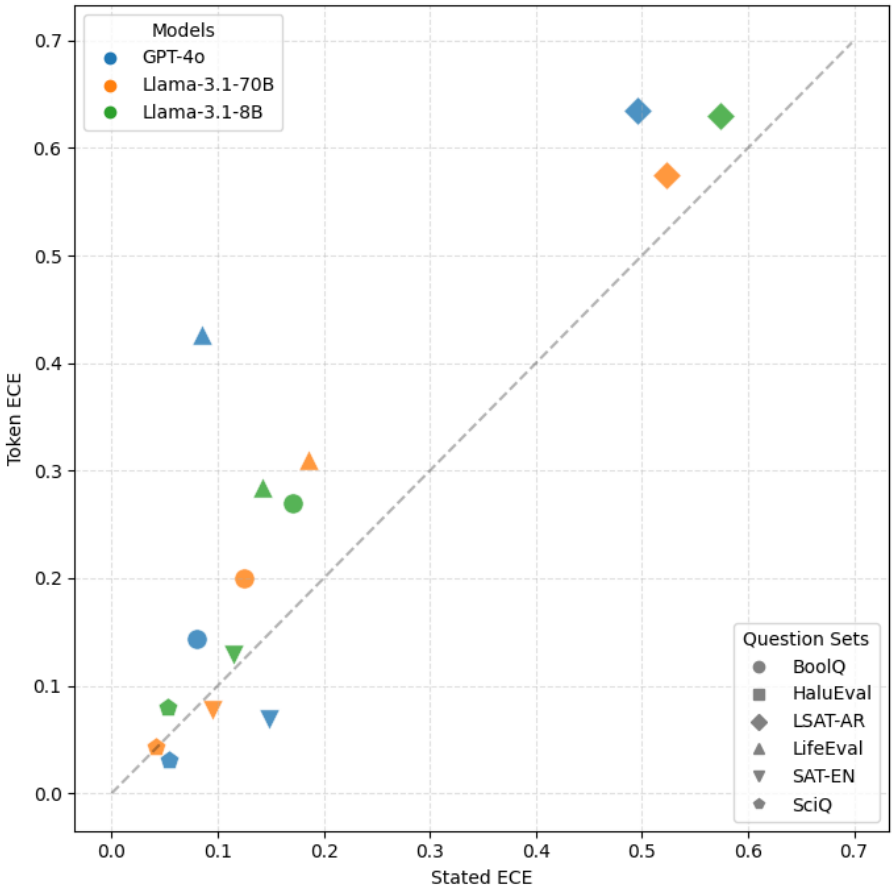

Figure 8: Comparison between the calibration error of stated confidence versus token probability for models when available. In most cases, Stated ECE was lower then Token ECE. Given the nature of HaluEval, we did not get token probabilities from the models. Because of this, HaluEval is left out from this analysis.

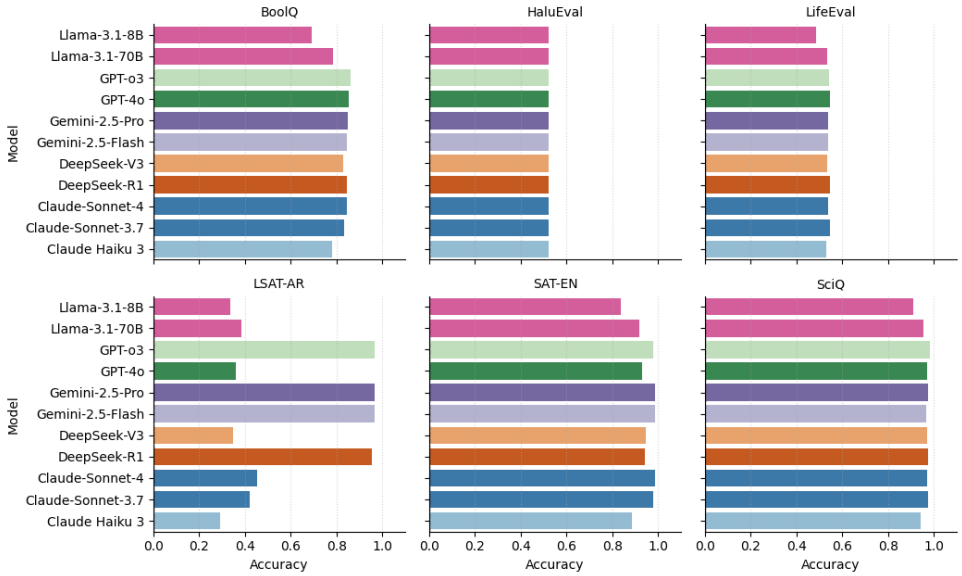

Figure 9: Accuracy for each model on all question sets. SAT-EN and SciQ had the highest performance while LSAT-AR saw the biggest variation between models.

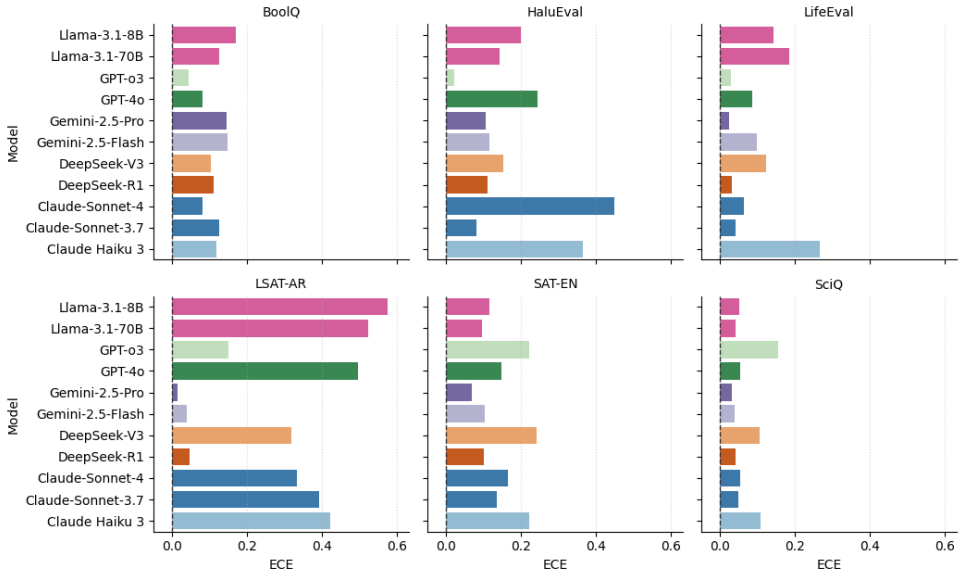

Figure 10: ECE for each model on all question sets. Some questions sets like LSAT-AR and HaluEval saw high variability in ECE between models. Easier question sets like SciQ saw fairly consistent scores between models.

Question Set	**Metrics**	Claude-Sonnet-3.7	Claude-Sonnet-4	DeepSeek-R1	GPT-o3	Gemini-2.5-Pro
BoolQ	Accuracy	83.14	84.30	84.50	86.06	84.90
	Confidence	95.57	91.97	95.61	82.59	99.35
	ECE	0.12	0.08	0.11	0.04	0.15
	(%) Rounded	46.06	96.20	68.08	50.38	73.95
HaluEval	Accuracy	52.12	52.12	52.12	52.12	52.12
	Confidence	59.91	95.68	61.07	53.84	60.14
	ECE	0.08	0.45	0.11	0.02	0.11
	(%) Rounded	99.94	100.00	99.27	55.08	100.00
LSAT-AR	Accuracy	41.86	45.35	95.35	96.51	96.51
	Confidence	80.93	71.40	99.72	81.52	97.99
	ECE	0.39	0.33	0.05	0.15	0.01
	(%) Rounded	95.35	94.19	96.51	72.09	37.21
LifeEval	Accuracy	54.49	53.98	54.43	54.25	53.84
	Confidence	53.11	49.79	57.23	54.13	53.42
	ECE	0.04	0.06	0.03	0.03	0.03
	(%) Rounded	90.15	98.80	29.03	69.77	17.98
SAT-EN	Accuracy	97.69	98.84	94.22	97.69	98.84
	Confidence	84.16	82.63	87.06	75.55	92.74
	ECE	0.14	0.17	0.10	0.22	0.07
	(%) Rounded	97.11	100.00	94.80	83.82	63.01
SciQ	Accuracy	97.29	96.88	97.59	98.09	97.59
	Confidence	92.49	91.72	93.41	82.65	95.44
	ECE	0.05	0.05	0.04	0.15	0.03
	(%) Rounded	73.37	99.60	66.23	77.19	29.75

Table 3: Performance of reasoning models across all question sets.

Question Set	Model Metrics	Claude Haiku 3	DeepSeek-V3	GPT-4o	Gemini-2.5-Flash	Llama-3.1-70B	Llama-3.1-8B
BoolQ	Accuracy	78.15	82.90	85.30	84.54	78.59	69.32
	Confidence	89.76	92.94	93.26	99.30	90.60	85.98
	ECE	0.12	0.10	0.08	0.15	0.12	0.17
	(%) Rounded	99.88	91.33	98.36	89.93	88.17	96.28
HaluEval	Accuracy	52.12	52.12	52.12	52.12	52.12	52.12
	Confidence	88.65	67.16	76.49	60.29	63.73	71.21
	ECE	0.37	0.15	0.24	0.11	0.14	0.20
	(%) Rounded	100.00	100.00	100.00	98.60	99.11	98.72
LSAT-AR	Accuracy	29.07	34.88	36.05	96.51	38.37	33.72
	Confidence	67.39	63.84	85.58	97.48	89.53	88.80
	ECE	0.42	0.32	0.50	0.04	0.52	0.57
	(%) Rounded	89.53	90.70	90.70	98.84	100.00	94.19
LifeEval	Accuracy	53.02	53.26	54.55	53.79	53.52	48.39
	Confidence	79.76	63.72	59.77	63.58	72.03	59.85
	ECE	0.27	0.12	0.09	0.10	0.19	0.14
	(%) Rounded	100.00	100.00	100.00	48.87	99.47	100.00
SAT-EN	Accuracy	88.44	94.80	93.06	98.84	91.91	83.82
	Confidence	67.51	72.86	79.39	88.53	85.82	85.52
	ECE	0.22	0.24	0.15	0.10	0.10	0.12
	(%) Rounded	98.84	100.00	96.53	98.84	99.42	94.80
SciQ	Accuracy	94.07	97.09	96.88	96.68	95.38	91.06
	Confidence	84.35	86.55	91.95	93.44	94.68	94.51
	ECE	0.11	0.11	0.05	0.04	0.04	0.05
	(%) Rounded	99.80	98.99	99.40	78.39	97.59	97.29

Table 4: Performance of chat models across all question sets.

Provider	Google	OpenAI	Anthropic	Lambda	DeepSeek	Total
Spend (USD)	228.59	234.17	94.53	941.78	7.41*	1,506.48

Table 5: Total spend by provider. New users on Google's platform receive $300 in compute credits which we did not surpass. Lambda generously provided us with a $5,000 research grant to cover our compute costs on their services. We ran both Llama models using one NVIDIA H100 GPU over a combined 283 hours. We did not keep track of our total spend on DeepSeek but we estimate the price based on publicly available pricing and our input and output token counts.

Model	Publisher	Type	Model Card
Claude Haiku 3	Anthropic	Chat	https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf
Claude Sonnet 3.7	Anthropic	Reasoning	https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf
Claude Sonnet 4	Anthropic	Reasoning	https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf
DeepSeek V3	DeepSeek	Chat	https://huggingface.co/deepseek-ai/DeepSeek-V3
DeepSeek R1	DeepSeek	Reasoning	https://huggingface.co/deepseek-ai/DeepSeek-R1
Gemini 2.5 Flash	Google	Chat	https://modelcards.withgoogle.com/assets/documents/gemini-2.5-flash.pdf
Gemini 2.5 Pro	Google	Reasoning	https://modelcards.withgoogle.com/assets/documents/gemini-2.5-pro.pdf
Llama 3.1 8B	Meta	Chat	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Llama 3.1 70B	Meta	Reasoning	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
GPT 4o	OpenAI	Chat	https://openai.com/index/gpt-4o-system-card/
GPT o3	OpenAI	Reasoning	https://openai.com/index/o3-o4-mini-system-card/

Table 6: Information about each model used.