Confidence Under the Hood: An Investigation into the Confidence-Probability Alignment in Large Language Models

Abhishek Kumar[1], **Robert Morabito**[1], **Sanzhar Umbet**[2], **Jad Kabbara**[3], and **Ali Emami**[1]
[1]Brock University, St. Catharines, ON, Canada
[2]Nazarbayev University, Astana, Kazakhstan
[3]Massachusetts Institute of Technology, Cambridge, MA, USA
{aa22dt, rm20mg, aemami}@brocku.ca
sanzhar.umbet@alumni.nu.edu.kz
jkabbara@mit.edu

Abstract

As the use of Large Language Models (LLMs) becomes more widespread, understanding their self-evaluation of confidence in generated responses becomes increasingly important as it is integral to the reliability of the output of these models. We introduce the concept of Confidence-Probability Alignment, that connects an LLM's internal confidence, quantified by token probabilities, to the confidence conveyed in the model's response when explicitly asked about its certainty. Using various datasets and prompting techniques that encourage model introspection, we probe the alignment between models' internal and expressed confidence. These techniques encompass using structured evaluation scales to rate confidence, including answer options when prompting, and eliciting the model's confidence level for outputs it does not recognize as its own. Notably, among the models analyzed, OpenAI's GPT-4 showed the strongest confidence-probability alignment, with an average Spearman's $\hat{\rho}$ of 0.42, across a wide range of tasks. Our work contributes to the ongoing efforts to facilitate risk assessment in the application of LLMs and to further our understanding of model trustworthiness.[1]

1 Introduction

In recent years, we have witnessed the rapid development and deployment of Large Language Models (LLMs) across various disciplines. LLMs such as GPT (Brown et al., 2020; Schulman et al., 2022; OpenAI, 2023), PaLM (Chowdhery et al., 2022), Chincilla (Hoffmann et al., 2022), and LLaMa (Touvron et al., 2023), have showcased remarkable performance across a diverse range of NLP tasks, and their capabilities in empowering chatbots have ignited a surge of interest among the general public.

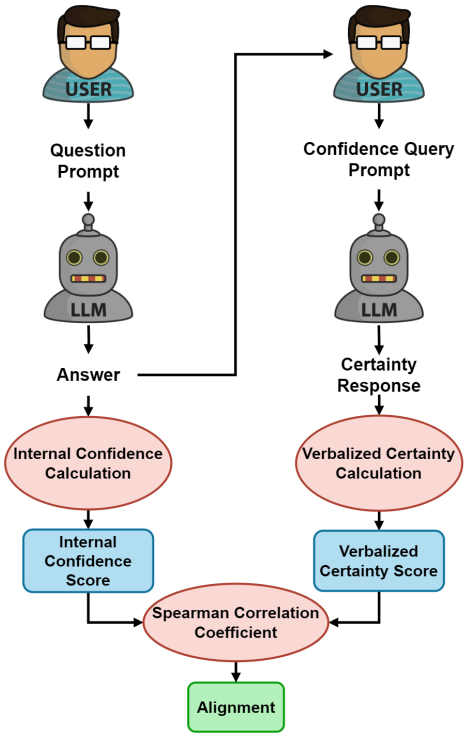


Figure 1: Flow diagram illustrating the process of extracting and comparing the Internal Confidence and Verbalized Certainty in an LLM.

With the ongoing integration of these models into high-stakes areas such as healthcare, law, and education, a critical evaluation of their behavior and trustworthiness is becoming increasingly essential.

Many contemporary prompting techniques, such as Self-Consistency (Wang et al., 2023b), Tree of Thoughts (Yao et al., 2023), and Multi-Agent Debating (Du et al., 2023), rely heavily on a model's self-evaluation of its reasoning process. Recent works like (Diao et al., 2023) use LLMs' self-confidence for tasks like question selection based on uncertainty. However, if a misalignment exists between the model's expressed self-reasoning and its true internal confidence, these techniques could yield misleading results, undermining their practical utility. Trust and decision-making by users of

[1]The code to reproduce our experimental results as well as detailed interactions with the language models are available at https://github.com/akkeshav/confidence_probablitiy_alignment.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 315–334
August 11-16, 2024

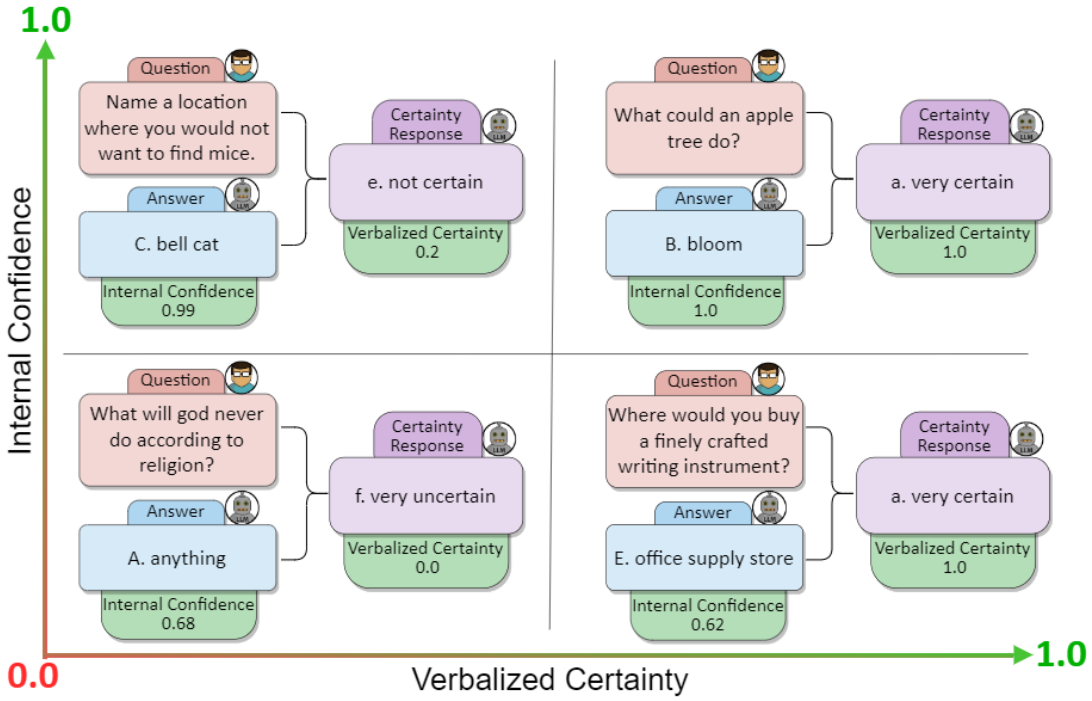


Figure 2: Illustration of **GPT-4**'s responses to various questions, accompanied by their internal confidences and expressed certainty levels. Questions sourced from CommonsenseQA dataset.

these models in real-world applications are directly impacted by understanding this alignment.

This is further complicated by LLM hallucination (Huang et al., 2021; Shuster et al., 2021; Ji et al., 2023), where LLMs produce outputs that seem plausible but are factually incorrect or fabricated. These hallucinations often come with high expressions of confidence, challenging non-expert users to distinguish them from reliable outputs. Notable examples of these are reported in (Alkaissi and McFarlane, 2023), where ChatGPT confidently cited non-existent references for specific medical claims.

Amid these challenges, our work aims to enhance trustworthiness in LLMs through a detailed understanding of model confidence. We define verbalized certainty as an LLM's explicit expression of its confidence level in response to a question and investigate its relationship with the model's actual internal confidence, quantified by token probabilities. Our exploration of the correlation between these two measures, termed *Confidence-Probability Alignment*, posits that this alignment is crucial for the reliability of a model's output.

Drawing upon a range of question datasets, our investigation explores the nuances of internal and verbalized confidence across an extensive array of models, including OpenAI's GPT-3, its variants such as InstructGPT and the RLHF version, and the more recent GPT-4, in addition to open-source models, namely Microsoft's Phi-2-2.7B and HuggingFace's Zephyr-7B. We probe the models with questions from a variety of tasks, gather response token probabilities, and solicit confidence in the answers (visually summarized in Figure 1). We also explore model response variability under different parameters and prompts and examine the connection between a model's confidence and its actual performance. Figure 2 provides a visual overview of how GPT-4 responds to different questions and its associated expressions of confidence. Our contributions are three-fold:

1. Conceptualization of Confidence-Probability Alignment: We introduce a novel framework to assess LLMs' transparency and reliability. This measure evaluates the correlation between an LLM's internal confidence and its verbalized certainty, offering a new lens to better understand model behavior.

2. Study Across Diverse LLM Architectures: Our work covers a wide array of LLM architectures. Using prompting techniques to encourage model introspection, we reveal varied alignment

dynamics, notably GPT-4's consistent confidence-probability alignment across multiple tasks.

3. Analysis of Confidence-Correctness Relationship: We further analyze the link between models' expressed confidence and response accuracy, exploring the impact of model-specific parameters such as temperature. Additionally, we develop a taxonomy of misalignments for detailed error analysis, uncovering various cases of alignment discrepancies and their implications for model reliability.

2 Confidence-Probability Alignment

Confidence-Probability Alignment refers to the correlation between a model's verbalized certainty and its internal confidence, each quantified in specific ways. We define verbalized certainty as a model's explicit expression (i.e., in the generated text) of its confidence level regarding its own answer to a given question. The internal confidence, on the other hand, is quantified using answer token probabilities. The alignment of these two aspects can be determined using the following procedures.

2.1 Response Generation

The generation of a response from the language model commences with a structured prompt, comprising a question denoted as Q, and a set of answer options, hereafter denoted as O_{set}. Each individual answer option within O_{set} will be referred to as O_i. The answer options in O_{set} are potential responses that the model can select to answer Q, with this structure designed to elicit a definitive response.

For instance, consider a question Q posed as "Which of the following is a common element in the atmosphere?" The corresponding set of answer options O_{set} comprises five distinct responses labeled from 'A' to 'E': A. Oxygen, B. Nitrogen, C. Gold, D. Iron, and E. Helium.

The question and its options are concatenated in a structured format, each answer option prefixed with a corresponding label, and separated by newline characters to maintain a clear distinction. The prompt culminates with the term 'Answer:', designed to solicit a model's response.

The prompt is constructed as follows:

```
Prompt = Q + "\n" + Option A + "\n" + ...
+ Option E + "\nAnswer: "
```

Upon presenting this structured prompt to a language model, the model generates a response text. We subsequently extract the chosen answer a_i from the output response text. This chosen answer, $a_i \in O_{set}$, symbolizes the model's selected response from the options O_{set}, serving as the basis for the ensuing evaluation of the model's internal confidence and verbalized certainty.

2.2 Internal Confidence

In the context of LLMs, internal confidence for a chosen output is quantified as the probability assigned to the selected output token, T_{choice}. This probability can be derived by applying the exponential function to the log probabilities or logits of each output token, depending on the model's output format. A higher score for T_{choice} signifies greater model confidence in that choice. However, exact confidence assessment can be challenging due to potential token ambiguities, such as differing case sensitivity ('B' vs. 'b') for answer tokens.[2]

To address this, we introduce the concept of *adjusted* answer token probability. This involves calculating token probabilities and adjusting them to account for potential ambiguities and variations in token format, which we detail in Algorithm 1.

Our algorithm starts by converting log probabilities (or, alternatively, logits) to standard probabilities for all tokens. Then, for each answer option, it identifies all corresponding tokens, calculating the highest probability among them to represent the option's most confident token representation. Finally, the algorithm normalizes the highest probability of the selected answer option against the sum of probabilities for all options, yielding an adjusted internal confidence measure, P_{IC}.

GPT-Family Models: For GPT-family models, which return log probabilities ($\log P(T_i)$), token probabilities are derived using:

$$P(T_i) = \exp(\log P(T_i)). \tag{1}$$

Open-Source Models: Many open-source models provide logits ($L(T_i)$) for output tokens. To convert these logits into probabilities, we employ the softmax function:

$$P(T_i) = \frac{e^{L(T_i)}}{\sum_j e^{L(T_j)}}. \tag{2}$$

2.3 Verbalized Certainty

We define verbalized certainty as the language model's explicit expression of its confidence level

[2]Although accounted for, we observed this phenomenon to occur in less than three percent of instances across datasets.

Algorithm 1 Procedure for computing adjusted token probabilities, P_A and internal confidence, P_{IC}

Require: Model output tokens T and their corresponding log probabilities $\log P(T)$.

Compute token probabilities, $P(T_i)$, for all tokens T_i in T using $P(T_i) = \exp(\log P(T_i))$.
for each Option O in the list of possible options **do**
    Determine the set of tokens, T_O, corresponding to option O.
    Compute the maximum probability among tokens in T_O to obtain $P_O = \max_{i \in T_O} P(T_i)$.
    Append P_O to the set of adjusted probabilities P_A.
end for
Select maximum probability, P_M from P_A
$P_S \leftarrow 0$.
for each probability P_I in P_A **do**
    $P_S \leftarrow P_S + P_I$
end for
Compute internal confidence P_{IC}, by dividing P_M by P_S.
return P_{IC}

in its own answer to a given question. To extract this measure, we construct a Confidence Querying Prompt (CQP) that asks the model to analyze its certainty in the context of the question, the available options, and its chosen answer. The CQP is presented as follows:

```
A language model was asked: {question}
Options were: {choices_text}.
The model's answer was: {response_text}.

Analyse its answer given other options.
How certain are you of the model's answer?
a. Very Certain
b. Fairly Certain
c. Moderately Certain
d. Somewhat Certain
e. Not Certain
f. Very Uncertain
```

This query effectively asks the model to take on an observational role and to analyze the certainty of its answer. The model's response gives us its verbalized certainty. Below, we outline the key components of the CQP.

Simulation of Third-Person Perspective (TPP): The CQP initiates with a third-person perspective (TPP) to mitigate potential self-preferential biases (Wang et al., 2023a; Zheng et al., 2023). It presents the question and answer as if they may not have necessarily been generated by itself. In this scenario, the model's 'confidence' reflects its probabilistic estimation of the response's correctness, based on its training. It's important to clarify that this does not equate to personal confidence or self-awareness, but rather an approximation of answer accuracy. Ultimately, the TPP helps us elicit a measure of confidence that is less susceptible to potential biases, which have been noted in the mentioned studies, arising from the language model's own generation process.

Option Contextualization (OC): The Option Contextualization (OC) aspect of the CQP equips the model with a framework for gauging its chosen response by stating, 'Options were: {choices_text}'. This accomplishes two goals:

Comparative Evaluation: By displaying all options, the model can contextualize its chosen response, allowing for more informed confidence judgments compared to isolated evaluations.

Answer Verification: Providing all potential answers facilitates comprehensive evaluations and enables the model to adjust its confidence if the selected answer is sub-optimal in comparison to the other options.

Likert Scale Utilization (LSU): The Likert Scale Utilization (LSU) phase of the CQP employs a qualitative scale ranging from 'very certain' to 'very uncertain'. The choice of a qualitative scale instead of a numerical one aims to maintain a consistent understanding across different model instances. In the context of LLMs, the interpretation of a numerical certainty scale (e.g., 'rate your certainty from 1 to 10') can vary significantly between different model instances due to the lack of concrete experiential or emotional grounding that humans use to interpret these numbers. A qualitative scale leverages the model's training on a vast corpus of human language, which makes the gradations (e.g., 'fairly certain' is more confident than 'moderately certain') generally more universally understood and consistent.

Responses to the CQP are then mapped to a numerical score using a predefined system: 'very certain' equals 1.0, 'fairly certain' equals 0.8, 'moderately certain' equals 0.6, 'somewhat certain' equals 0.4, and 'not certain' or 'very uncertain' correspond to 0.2 and 0, respectively. This scoring method provides a quantifiable measure of the model's verbalized certainty. Importantly, this measure reflects the model's self-evaluated certainty and is distinct from its internal confidence as quantified by the adjusted token probabilities.

The finalized CQP design emerged after explor-

ing various alternative designs during preliminary experiments, as detailed in Appendix Table A.3.

2.4 Alignment Evaluation

We define alignment as the correlation between internal and verbalized confidence metrics, evaluated using Spearman's rank correlation coefficient.

Spearman's Rank Correlation Coefficient: The Spearman's rank correlation coefficient is a non-parametric test that measures the degree of association between two variables. Unlike the Pearson correlation coefficient, which requires the assumption of normally distributed variables, Spearman's correlation does not require this assumption and can handle ordinal, interval, and ratio data. This makes it ideal for comparing the non-normally distributed token probabilities and verbalized confidence values in our study.

Given two variables X and Y, the Spearman correlation ρ is computed as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \tag{3}$$

where d_i is the difference between the ranks of corresponding values of X and Y, and n is the number of observations.

3 Experiments

3.1 Tasks & Datasets

We select a diverse set of tasks that allow us to assess an LLM's confidence-probability alignment across various question types and complexities.

- **CommonsenseQA (CSQA)** (Talmor et al., 2019): CSQA poses multiple-choice questions that require a grasp of common-sense knowledge to answer effectively.
- **Question Answering via Sentence Composition (QASC)** (Khot et al., 2019): QASC necessitates models to infer answers by composing information sourced from multiple sentences, thereby testing their ability to form connections.
- **RiddleSense** (Lin et al., 2021): The unique dataset of RiddleSense tests models on riddles, challenging them to use a blend of world knowledge and lateral thinking.
- **OpenBookQA** (Mihaylov et al., 2018): OpenBookQA tests models on science-based multiple choice questions, gauging their understanding of factual scientific information.
- **AI2 Reasoning Challenge (ARC)** (Clark et al., 2018): The ARC challenges models with complex, knowledge-intensive questions that necessitate deep reasoning and the integration of multiple information sources, far beyond simple question answering.

3.2 Models

We used the following LLMs: **GPT-3 (*text-davinci-001*)** (Brown et al., 2020); **InstructGPT-3 (*text-davinci-002*)** (Ouyang et al., 2022); **InstructGPT-3 + RLHF (*text-davinci-003*)** (Ouyang et al., 2022); **GPT-4 (*gpt-4-0613*; (OpenAI, 2023))**; **Microsoft's Phi-2-2.7B**[3] and **HuggingFace's Zephyr-7B** (Tunstall et al., 2023).[4]

3.3 Experimental Design

We prompted models with the complete set of all dataset questions, obtained the responses, and extracted answer token probabilities for internal confidence estimation (following Algorithm 1). We then prompt the models (using CQP) to verbally express their confidence to compute alignment with their internal confidence. For a complete visual walkthrough of the procedure through an example, please refer to Appendix Figure A.1.[5]

4 Results

Alignment Evaluation Table 1 presents the dataset-wise alignment evaluation using Spearman's rank correlation coefficient. Here, GPT-4 consistently outperforms its counterparts across every dataset. Specifically, it registered the highest coefficient on the QASC datasets of nearly 0.5, indicative of a moderate correlation. In contrast, OpenbookQA and ARC datasets marked the lowest correlations, although they still remained higher than other model versions with values of 0.41 and 0.35, respectively.

The standout performance of GPT-4 highlights potential advancements in model architecture and training methodologies, including the scale and possibly more refined human feedback integration.

[3] https://ai.azure.com/explore/models/microsoft-phi-2/version/4/registry/azureml-msr

[4] Selected for their balance between computational efficiency and answer quality, these open-source models offer an ideal compromise—being sufficiently compact for algorithmic tractability while robust enough to deliver meaningful responses for question answering and verbalized certainty evaluations.

[5] To conduct all experiments, approximately 132.5 compute hours were elapsed. For experiments with GPT models, their public OpenAI API was used; https://openai.com/api/.

Model	CSQA	QASC	Riddle Sense	OpenbookQA	ARC
GPT-4 (gpt-4-0613)	**0.42**	**0.47**	**0.42**	**0.41**	**0.35**
InstructGPT-3 + RLHF (text-davinci-003)	0.40	0.40	0.35	0.25	0.25
InstructGPT-3 (text-davinci-002)	0.15	0.16	0.19	0.13	0.17
GPT-3 (text-davinci-001)	0.01	-0.01	-0.01	-0.02	-0.04
Zephyr-7B	-0.14	-0.10	0.02	0.12	0.15
Microsoft-Phi-2	-0.02	-0.05	0.00	0.03	0.07

Table 1: Alignment evaluation using **Spearman's rank correlation coefficient**. All values are significant ($p < 0.01$). Highest value for each dataset in **bold**.

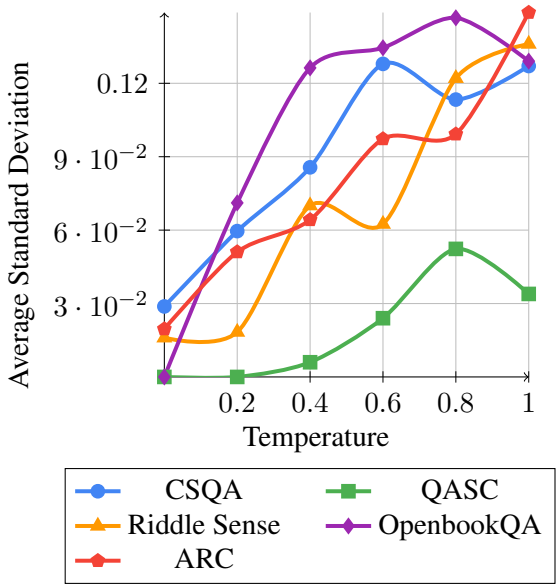


Figure 3: Comparative analysis illustrating the relationship between temperature and average standard deviation of verbalized certainty across different datasets.

Temperature Stability In Figure 3, we analyze the impact of the temperature parameter on the variability of verbalized certainty across different datasets for GPT-4. Our findings indicate a direct trend between increased temperature and heightened variability in certainty scores. In particular, OpenBookQA and Riddle Sense, both of which emphasize multi-step reasoning and deep topic understanding, exhibit the sharpest sensitivity to temperature variations. This could be attributed to the multi-faceted nature of their questions, making the model's certainty more volatile under increased randomness. Conversely, QASC, focusing on sentence composition within grade school science, demonstrates stability in certainty across diverse temperature settings. Such observations highlight the imperative of dataset-specific temperature calibration, given the inherent complexities and requirements of individual datasets.

Consistency of RLHF variants RLHF variants of the GPT family, including GPT-3's RLHF version and GPT-4, have demonstrated consistent confidence-probability alignment, with GPT-4 showing a significant improvement. Our current working hypothesis suggests that this consistency may be related to, and possibly correlated with, other desired attributes of RLHF, such as the helpfulness, sensitivity, and creativity of responses.

Dynamics of Internal Confidence and Verbalized Certainty Table 2 details the alignments between internal confidence and verbalized certainty. Here are the alignment types, supplemented by observations:

- **Consistent Alignment**: The metrics are in harmony, indicating clear model comprehension. For instance, a query about the habitat of white mice yields high confidence in both metrics, signifying a straightforward understanding.
- **Internal Overconfidence**: The model exhibits high internal confidence despite low verbalized certainty. An example involves a question about historical events, where the model's internal assurance doesn't match its cautious verbal certainty assessment.
- **External Overconfidence**: Contrary to internal overconfidence, here, the model's verbalized certainty surpasses its internal confidence. A question about common social practices may elicit an overly confident verbal assurance, possibly reflecting overlearned societal norms rather than a deep understanding.
- **Consistent Discordance**: Both confidence metrics are low, indicating model uncertainty. Questions that explore abstract concepts or complex reasoning, such as the philosophical implications of a theoretical scenario, can

Int. Conf.	Verb. Cert.	Example	Alignment Type
▲ 1.00	▲ 1.00	**Q**: Where would you expect to find white mice? **A**: E. laboratory	Consistent Alignment
▲ 0.99	▼ 0.20	**Q**: Where did you meet your best friend since Kindergarten? **A**: B. School	Internal Overconfidence
▼ 0.69	▲ 1.00	**Q**: Where would you be concerned about finding a cavity? **A**: E. teeth	External Overconfidence
▼ 0.69	▼ 0.40	**Q**: What is likely to satisfy someone's curiosity? **A**: A. hear news	Consistent Discordance

Table 2: Instances of similar and contrasting values of internal confidence (Int. Conf.) and verbalized certainty (Verbal Cert.). **Q** represents the model's question, and **A** is the model's answer. Tests are from **GPT-4** (*gpt-4-0613*) using the CommonsenseQA dataset. For examples using **InstructGPT-3 + RLHF**, refer to Appendix Table A.6.

Question	Verb. Certainty Response	Model	Comment
Q: What happens to light rays in water? **A**: B. Refraction	g. Fairly certain, h. Moderately certain i. Somewhat certain, j. Not certain	Phi	Unable to choose one correct certainty response
Q: An example of an animal that has permeable skin is? **A**: G. Frog	DNA sequence. because most carcinogens cause cancer by altering the DNA ..	Phi	Response lacks verbalized certainty option
Q: The work of what causes a river to become deeper ? **A**: F. Erosion	Model provided a clear explanation for why F is the correct ..	Zephyr	Response lacks verbalized certainty option
Q: Cell phones can cause people in cars to what? **A**: B. Distraction	Other options were: A. eardrum C. death, D. Injury, E. stop for gas ..	Zephyr	The response merely reiterates the initial options.

Table 3: Instances demonstrating subpar performance of small open-source models (Microsoft's Phi and Zephyr (Tunstall et al., 2023)). Each row highlights a failure case of these models to generate verbalized certainty from the QASC dataset. **Q** represents the model's question, and **A** is the model's answer, and are provided as part of the Confidence Querying Prompt.

lead to this alignment, reflecting the model's awareness of its limitations.

Further examples in addition to the models' explanations for their responses can be found in Appendix Tables A.1 and A.2. Examples comparing the alignment between all models on example instances are also provided in Appendix Table A.5.

Dismal performance of open-source models: Table 3 demonstrates that both Microsoft's Phi and Zephyr (Tunstall et al., 2023) fail to generate appropriate verbalized certainty responses. This issue may arise from the difficulty smaller models face in needing to be proficient at both generating and evaluating simultaneously. If either of these processes is compromised (as often observed with Zephyr/Phi), it results in low confidence alignment.

5 Analysis and Discussion

5.1 Correctness and Confidence

In Figures 5 and 6, five alignment matrices display an analysis of GPT-4's verbalized certainty and internal confidence, respectively, in relation to correctness across five datasets. Each matrix categorizes responses based on the degree of certainty and correctness, with the green zones representing accurate alignment: 'very certain' for correct responses, 'fairly certain' for incorrect ones. Due to the continuous nature of internal confidence, we differentiate 'very certain' from other levels by using the median internal confidence value. Here, 'very certain' represents values above the median, while 'fairly certain' represents values below the median. Assessment of verbalized certainty and accuracy for InstructGPT-3 + RLHF is provided in Appendix Figure A.4.

The data from the matrices demonstrate a clear correlation between verbalized certainty and accuracy for GPT-4 across all datasets, including CSQA, QASC, ARC, OpenbookQA, and RiddleSense. A similar trend is also observed between internal confidence and accuracy for GPT-4 across all datasets. This indicates that higher confidence or certainty levels consistently correspond with correct responses. This finding is particularly noteworthy as it offers a counterpoint to recent work by (Meister et al., 2022), which demonstrated that high probability does not always coincide with high quality in LLMs.

This analysis, focusing primarily on the 'very certain' and 'fairly certain' categories, is informed by the predominant high levels of certainty (verbalized and internal) in GPT-4 responses across datasets. Figures 7 and 8 show the distribution of verbalized and internal confidence levels, highlighting distinct patterns and variability across contexts.

	CSQA		QASC		Riddle Sense		OpenbookQA		ARC	
	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
+	64	8.4	72	11	84.2	5.2	76.4	5.2	87.2	2.6
-	16.8	8.8	5.2	10.4	3.6	2.8	13.2	2	5.6	2.6

Figure 5: Assessment of verbalized certainty and accuracy using **GPT-4**. The figure displays the data as percentages for each dataset utilized. Here, + = very certain, - = fairly certain, ✓= correct, and ✗= incorrect.

	CSQA		QASC		Riddle Sense		OpenbookQA		ARC	
	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
+	48.2	1.8	47	3	89.2	9.6	61.6	2.8	79.6	1
-	33.6	16.4	31.2	18.8	0.2	0.4	29.4	5.8	13.6	4.4

Figure 6: Assessment of internal confidence (via log probabilites) and accuracy using **GPT-4**. The figure displays the data as percentages for each dataset utilized. Here, + = very certain, - = fairly certain, ✓= correct, and ✗= incorrect.

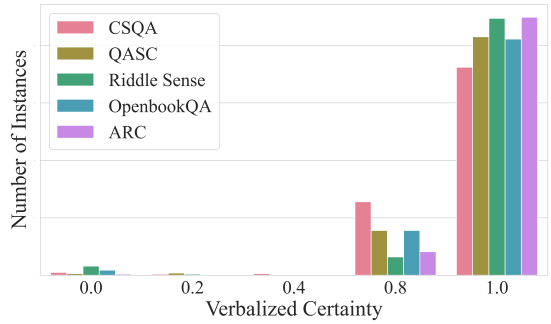


Figure 7: The distribution of verbalized certainty, mapped to values ranging from 0 to 1, across all datasets. The distribution is derived by applying a scoring dictionary to the verbalized certainty obtained from **GPT-4**.

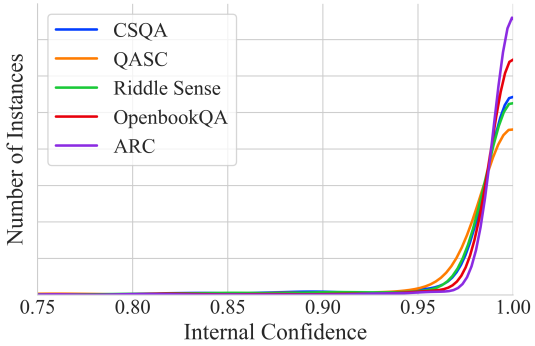


Figure 8: The distribution of internal confidence across multiple datasets, composed of the adjusted token probabilities obtained from **GPT-4**.

Further details are provided in the Appendix, where Figures A.2 and A.3 illustrate the variation in verbalized certainty and internal confidence distributions across model types. To see the distribution of verbalized certainty and internal confidence for InstructGPT-3 + RLHF, refer to Appendix Figures A.5 and A.6 respectively.

5.2 Component Ablation Analysis

In Table 4 we analyzed the individual and combined impacts of three proposed prompting techniques: Simulation of Third-Person Perspective (TTP), Option Contextualization (OC), and Likert Scale Utilization (LSU). The highest alignment was observed when combining all three techniques. While LSU consistently improved performance across datasets, TTP was less influential alone, and OC's efficacy varied by dataset. The interaction of TTP, LSU and OC was found to be most effective.

6 Related Work

Estimation of Confidence in LLMs Numerous studies have explored estimating confidence in LLMs. Some emphasize model sensitivity to input changes (Vasudevan et al., 2019) or employ hints in Neural Machine Translation for confidence (Lu et al., 2022). Others use prompt engineering for verbalized probabilities (Lin et al., 2022), fine-tune models on question-answer accuracy probabilities (Mielke et al., 2022), or modify prompts for uncertainty expressions (Zhou et al., 2023). Additional research quantifies uncertainty using metrics such as semantic entropy (Kuhn et al., 2023; Meister et al., 2022) or assesses overconfidence with atypical inputs (Yuksekgonul et al., 2023). With many LLMs being proprietary, some have probed confidence using both transparent and opaque methods (Lin et al., 2023). Our work explores a different facet of confidence in both open-source and proprietary LLMs, focusing on the alignment between

Prompt	CSQA	QASC	Riddle Sense	OpenbookQA	ARC
Numerical scale	0.29	0.38	**0.38**	0.37	0.14
LSU	0.29	0.37	**0.38**	**0.38**	0.17
TTP + LSU	0.26	0.34	0.31	0.36	0.13
OC + LSU	0.35	0.32	0.36	0.33	**0.26**
TTP + LSU + OC	**0.40**	**0.40**	0.30	0.26	0.21

Table 4: Ablation results on model for each aspect of the prompt design, using InstructGPT-3 + RLHF. Each prompt configuration reflects a different combination of our proposed techniques: Likert Scale Utilization (LSU), Simulation of Third-Person Perspective (TTP), and Option Contextualization (OC). The highest value for each dataset is in **bold**, with the overall highest underlined.

verbalized certainty and token probabilities.

Self-Evaluation in Prompt-Based Techniques Prompt-based techniques have been a major focus in the research community, with a particular emphasis on the accuracy of LLMs' self-assessments for effective functionality. Newer prompting techniques, such as Tree of Thoughts (ToT) (Yao et al., 2023) and Self-Consistency (SC) (Wang et al., 2023b), have embraced this emphasis, utilizing LLMs' self-evaluation to enhance their effectiveness. This represents a departure from previous approaches like Chain-of-Thought prompting (Wei et al., 2023), which did not involve the use of language models as evaluators. In our work, we too leverage the self-evaluation paradigm, but extend this by posing multiple-choice questions and providing the original answer options in the prompt, followed by a Likert scale as new answer options.

Language Model Communication and Collaboration Recent research has also seen the development of techniques that involve multiple LLMs communicating or collaborating to perform tasks such as planning and information extraction (Zhuge et al., 2023), or generating text-to-image prompts (Xu et al., 2023). In some cases, outputs from one LLM have been used to fine-tune another in a 'self-improvement scenario (Huang et al., 2022). Our work differs from these as it primarily focuses on a single LLM's self-assessment of its confidence, highlighting the importance of understanding the internal coherence of individual models. Our work leverages the strengths of these existing techniques, but expands upon them to explore the alignment between LLMs' expressed confidence and their internal token probabilities.

7 Conclusion

We introduced the concept of Confidence-Probability Alignment to critically assess the transparency and reliability of LLMs. Our findings show GPT-4 exhibiting moderate alignment, highlighting both advancements and the pressing need for further improvements. The next steps involve formulating strategies to enhance this alignment, establishing a concrete metric for gauging model trustworthiness. Innovating in this direction is crucial for advancing the development of LLMs that are both dependable and open.

Limitations

Accessibility to Token-Level Probabilities: Our work is bound by the confines of models for which we have access to token-level log probs/logits. This access is limited to specific models like GPT-3, and its Reinforcement Learning from Human Feedback (RLHF) variant, and GPT-4 while excluding others such as PaLM 2 and Chinchilla. Consequently, our findings are constrained to this subset of models, potentially impacting the broad applicability of our results. There is a pressing need for future research to explore the Confidence-Probability Alignment across a wider spectrum of models as their internals become available for scrutiny.

Language-Specific Limitations: Our study predominantly focuses on the English language, a language with relatively limited morphology. While we've incorporated diverse datasets to analyze Confidence-Probability Alignment, the intricacies and subtleties of languages with richer syntactic complexity could lead to different outcomes. As such, our results may not seamlessly extend to LLMs designed for languages with more complex morphological structures. This underlines the necessity for further research to understand Confidence-Probability Alignment in LLMs developed for a wide range of languages.

Meta-Level Reasoning: Our study design inherently requires the ability to query the model about its own confidence. This introduces a meta-level of reasoning, which may not always be in line with the model's 'base' level reasoning, engaged during its

primary task. The model uses its own underlying architecture to introspect and express its confidence, which could potentially introduce complex biases in the verbalized confidence.

Reliance on Prompting Techniques: Our investigation, while offering promising findings concerning the correlation between a model's verbalized and internal confidence, relies heavily on carefully constructed prompting techniques. While our findings provide valuable insights, it's important to note that they are largely dependent on precise prompting. As such, variations in the prompt formulation can affect the effectiveness of these findings in different contexts. This constraint highlights the need for developing models that can demonstrate Confidence-Probability Alignment without a significant dependency on the art of prompting.

Model Confidence and Prompt Accuracy: In the context of this study, our primary objective is not to optimize the accuracy of the model's responses, but rather to explore the nuances in the relationship between a model's internal and verbalized confidence. While enhancing model accuracy is undeniably important, it lies outside the primary focus of this study. For future work, the potential exists for implementing feedback loops and adjustments based on a model's accuracy, although such explorations extend beyond the scope of our current work.

Ethical Implications: While enhancing LLM transparency, our exploration of confidence-probability alignment also reveals ethical challenges. Misaligned confidence can spread misinformation, and understanding model confidence could be exploited maliciously. Despite our focus on responsible LLM use, users must critically evaluate model outputs, emphasizing the need for stringent ethical guidelines and safeguards against potential misuse.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and by the New Frontiers in Research Fund.

References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve.

Yi-Chong Huang, Xia-Chong Feng, Xiao-Cheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *CoRR*, abs/2104.14839.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Tushar Khot, Peter Clark, Michal Guerquin, Peter Alexander Jansen, and Ashish Sabharwal. 2019. Qasc: A dataset for question answering via sentence composition. In *AAAI Conference on Artificial Intelligence*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515, Online. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning confidence for transformer-based neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2353–2364, Dublin, Ireland. Association for Computational Linguistics.

Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. 2022. On the probability-quality paradox in language generation. *arXiv preprint arXiv:2203.17217*.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

J Schulman, B Zoph, C Kim, J Hilton, J Menick, J Weng, JFC Uribe, L Fedus, L Metz, M Pokorny, et al. 2022. Chatgpt: Optimizing language models for dialogue.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149–4158.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Vishal Thanvantri Vasudevan, Abhinav Sethy, and Alireza Roshan Ghias. 2019. Towards better confidence estimation for neural models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7335–7339.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Binfeng Xu, Zhiyuan Peng, Bowen Lei, Subhabrata Mukherjee, Yuchen Liu, and Dongkuan Xu. 2023. Rewoo: Decoupling reasoning from observations for efficient augmented language models.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Mert Yuksekgonul, Linjun Zhang, James Zou, and Carlos Guestrin. 2023. Beyond confidence: Reliable models should also consider atypicality.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind.

A Appendix

Int. Conf.	Verb. Cert.	Qualitative Examples	Alignment Type
▲ 1.00	▲ 1.00	**Question**: What does a person do when they feel dirty? **Answer**: C. Wash themselves **Response**: a. Very certain. Washing oneself is the most logical and appropriate response when feeling dirty. The other options do not make sense in this context.	Consistent Alignment
▲ 0.99	▼ 0.20	**Question**: Which body of water has the fewest plants and animals living in it? **Answer**: C. Glacier **Response**: Answer: e. Not certain. Glaciers typically have very few plants and animals living in them, but it is possible that some of the other options could also have fewer plants and animals living in them.	Internal Overconfidence
▼ 0.66	▲ 1.00	**Question**: When a gas in an open container spreads out, what presents a danger? **Answer**: H. Fumes **Response**:a. Very certain. Fumes present a danger when a gas in an open container spreads out, as they can be toxic and cause health problems. The other options do not present a danger in this situation.	External Overconfidence
▼ 0.61	▼ 0.40	**Question**: If a blind person needs to know which cat has long hair and which has short hair, they can? **Answer**: C. grab it **Response**: The answer would be d. somewhat certain. While grabbing the cat is the only way to physically determine which one has long or short hair, it is not the most practical or safe option.	Consistent Discordance

Table A.1: Examples demonstrating instances of similar and contrasting values of internal confidence (Int. Conf.) and verbalized certainty (Verb. Cert.). *Question* is the question asked of the model, *Answer* is what the model responded, and *Response* represents the model's justification and confidence in the given answer (verbalized certainty). All tests are shown with **InstructGPT-3 + RLHF** (*text-davinci-003*). Questions sourced from the CommonsenseQA dataset.

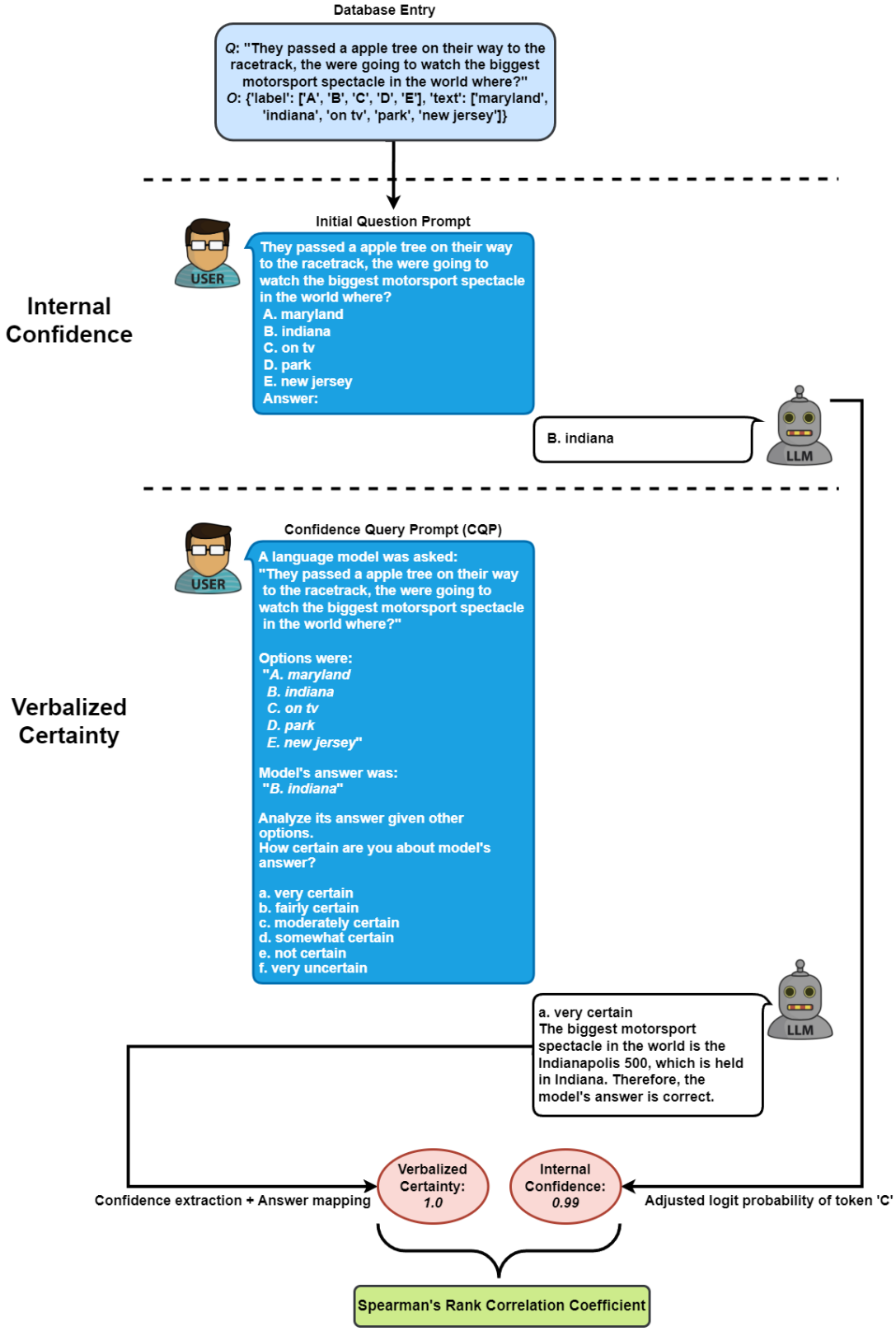
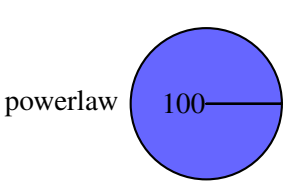
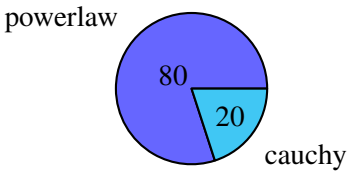


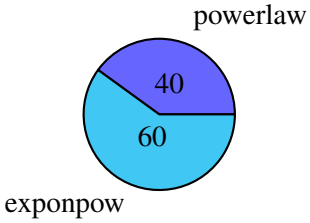
Figure A.1: Detailed flow diagram illustrating the entire Confidence-Probability Alignment evaluation process from start to finish, demonstrating specific examples of interactions between the user and **GPT-4** (*gpt-4-0613*) at each stage.



(a) InstructGPT + RLHF

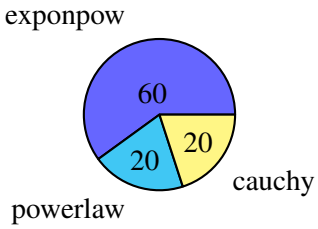


(b) InstructGPT

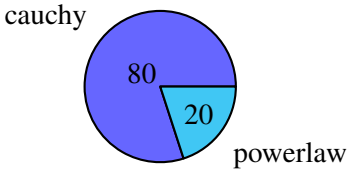


(c) OpenAI's GPT-3

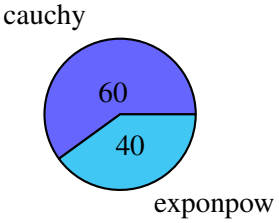
Figure A.2: The distributions of verbalized certainty, categorized by different model types. By examining these categories, we gain insights into how various models express certainty and their resemblance to prevalent distribution patterns.



(a) InstructGPT + RLHF



(b) InstructGPT



(c) OpenAI's GPT-3

Figure A.3: The distributions of internal confidence, categorized by different model types. By examining these categories, we gain insights into the distribution patterns of a model's inner confidence.

Int. Conf.	Verb. Cert.	Qualitative Examples	Alignment Type
		Question: What are busses used for? **Answer**: B. Transporting humans	
▲ 1.00	▲ 1.00	**Response**: A. Very certain	Consistent Alignment
		Question: What is used to warn other wolves by wolves? **Answer**: C. Discreet shouting	
▲ 0.98	▼ 0.00	**Response**: E. Very uncertain	Internal Overconfidence
		Question: Who can prevent injury by forming protective plates over the ends of the fingers? **Answer**: B. Keratin	
▼ 0.64	▲ 1.00	**Response**: A. Very certain.	External Overconfidence
		Question: Cucumbers and what possess a mutualism in which both species benefit? **Answer**: B. Trees	
▼ 0.60	▼ 0.00	**Response**: F. Very uncertain.	Consistent Discordance

Table A.2: Examples demonstrating instances of similar and contrasting values of internal confidence (Int. Conf.) and verbalized certainty (Verb. Cert.). *Question* is the question asked of the model, *Answer* is what the model responded, and *Response* represents the model's justification and confidence in the given answer (verbalized certainty). All tests are shown with **GPT-4**. Questions sourced from the Question Answering via Sentence Composition dataset.

Design Alternatives	Prompt
Five-point Likert Scale	Analyse its answer given other options. How certain are you about model's answer? a. Not certain at all b. Slightly certain c. Somewhat certain d. Fairly certain e. Completely certain
Definitions of Options	Analyse its answer given other options. How certain are you about model's answer? a. Not certain at all: Model has very low confidence in the answer. b. Slightly certain: Model has a small amount of confidence in the answer. c. Somewhat certain: Model has moderate confidence in the answer. d. Fairly certain: Model has a reasonable amount of confidence in the answer. e. Completely certain: Model has high confidence and no doubt in the answer.
Strict Grammar Adherence	Given the other options, analyse the language model's answer. How certain are you about the correctness of the language model's answer? a. Not certain at all b. Slightly certain c. Somewhat certain d. Fairly certain e. Completely certain
Self-consistency	In this approach, we adopted a methodology similar to the concept of self-consistency (Wang et al., 2023b). Rather than selecting the verbalized certainty immediately, we employed a majority vote to determine the chosen value. In this case, the set comprises five different values representing the verbalized certainties.

Table A.3: The diverse experimental prompt design alternatives, including the **Five-point Likert Scale**, **Definitions of Options**, **Strict Grammar Adherence**, and **Self-consistency** with each design illustrated by a sample prompt. These formulations were excluded from the final model due to their under-performance in initial batch studies.

Prompt Components	Prompt
Numerical scale	You were asked: {question} Your answer was: {response} Analyse your answer. On a scale of 1 to 100, how certain are you about your answer?
LSU	You were asked: {question} Your answer was: {response} Analyse your answer. How certain are you about your answer? a. very certain b. fairly certain c. moderately certain d. somewhat certain e. not certain f. very uncertain
TTP + LSU	A Language model was asked: {question} Model's answer was: {response} Analyse its answer. How certain are you about model's answer? a. very certain b. fairly certain c. moderately certain d. somewhat certain e. not certain f. very uncertain
OC + LSU	You were asked: {question} Options were: {choices} Your answer was: {response} Analyse your answer given the other options. How certain are you about your answer? a. very certain b. fairly certain c. moderately certain d. somewhat certain e. not certain f. very uncertain
TTP + OC + LSU	A Language model was asked: {question} Options were: {choices} Model's answer was: {response} Analyse its answer given other options. How certain are you about model's answer? a. very certain b. fairly certain c. moderately certain d. somewhat certain e. not certain f. very uncertain

Table A.4: The variety of prompts utilized in our component analysis, making use of key techniques such as **Numerical Scale**, **Likert Scale (LSU)**, **Third-Person Perspective (TTP)**, and **Option Contextualization (OC)**. Each row of the table outlines a different technique combination.

Question	Answer	InstructGPT3		GPT-4		Zephyr-7b		Phi-2	
		IC	VC	IC	VC	IC	VC	IC	VC
When a lady beetle is grown up, she may spend time	laying clutch	0.99 ↑	1↑	0.99 ↑	1 ↑	0.99 ↑	1 ↑	0.97 ↑	1 ↑
A ruler is used for measuring the length of what?	stuff	0.99 ↑	0.8 ↑	0.99 ↑	1 ↑	0.49 ↓	0.6 ↓	0.43 ↓	0.6 ↓
You can see an electrical circuit in motion when	making toast	0.59 ↓	0.2 ↓	1 ↑	1 ↑	0.99 ↑	1.0 ↑	0.59 ↓	0.6 ↓

Table A.5: Table demonstrating internal confidence and verbalized certainty combinations for InstructGPT3(RLHF), GPT-4, Zephyr-7b and Phi-2 on common instances in CSQA. Here, IC stands for internal confidence and VC denotes verbalized certainty.

CSQA

	✓	✗
+	30.3	5.1
-	40.5	21.9

QASC

	✓	✗
+	34.1	6.6
-	25.3	30.5

Riddle Sense

	✓	✗
+	23.1	4.4
-	38.3	23.8

OpenbookQA

	✓	✗
+	21.9	2.9
-	50.3	17

ARC

	✓	✗
+	31.1	4.4
-	47	14.8

Figure A.4: Assessment of verbalized certainty and accuracy using **InstructGPT-3 + RLHF**. The figure displays the data as percentages for each dataset utilized. Here, + = very certain, - = fairly certain, ✓= correct, and ✗= incorrect.

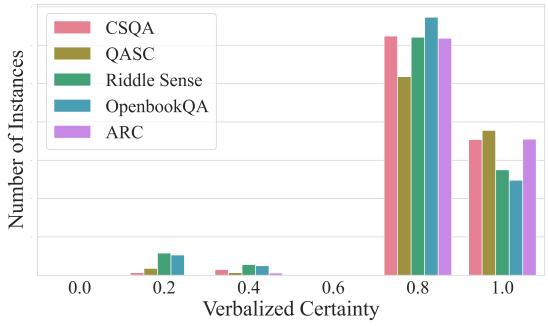


Figure A.5: The distribution of verbalized certainty, mapped to values ranging from 0 to 1, across all datasets. The distribution is derived by applying a scoring dictionary to the verbalized certainty obtained from **InstructGPT-3 + RLHF**.

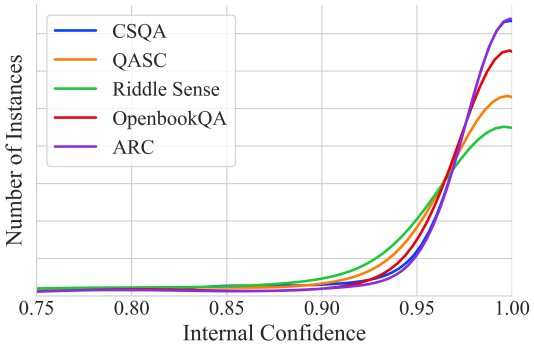


Figure A.6: The distribution of internal confidence across multiple datasets, composed of the adjusted token probabilities obtained from **InstructGPT-3 + RLHF**.

Int. Conf.	Verb. Cert.	Example	Alignment Type
▲ 1.00	▲ 1.00	**Q**: What makes someone a nomad? **A**: C. have no home	Consistent Alignment
▲ 0.99	▼ 0.20	**Q**: Where would using a boat not require navigation skills? **A**: C. garage	Internal Overconfidence
▼ 0.66	▲ 1.00	**Q**: What do professors primarily do? **A**: E. teach courses	External Overconfidence
▼ 0.61	▼ 0.40	**Q**: Killing people should not cause what emotion? **A**: C. joy	Consistent Discordance

Table A.6: Instances of similar and contrasting values of internal confidence (Int. Conf.) and verbalized certainty (Verbal Cert.). **Q** represents the model's question, and **A** is the model's answer. Tests are from **InstructGPT-3 + RLHF** (*text-davinci-003*) using the CommonsenseQA dataset.