Self-Estimation Is Contextual: Reliability and Calibration of LLM Capability Reports Across Tasks and Modalities

Anonymous ACL submission

Abstract

Estimating performance of language models on unseen tasks is a significant challenge. Though this can be estimated through a small number of samples, this method only accounts for established and well-defined benchmarks. Querying models for their self-estimations of ability, however, could enable broader capability to estimate potential accuracy on future tasks. Drawing from psychological metrics of self-evaluation, we measure the self-estimation capabilities of language models through task descriptions. We show that while these models struggle to self-estimate, they show substantially improved performance on tasks they are "familiar" with, providing more accurate self-reports on benchmarks similar to those on which they were trained. This pattern also applies in multimodal tasks, where noisy and difficult tasks confuse "unfamiliar" models but show substantially improved self-estimations in a model trained on similar tasks or datasets. Our findings highlight an important aspect of self-estimation in LLMs, highlighting how language model self-ratings are based on model knowledge on a subject. We will release our code and benchmark data.

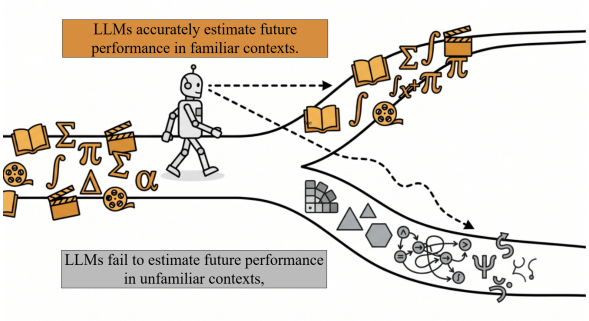

Figure 1: Self estimation is reliable only within familiar contexts. Models are asked to predict their future task performance from limited task examples. We find that their self estimates become more accurate when the task resembles contexts likely seen during training, but degrade in unfamiliar or shifted settings. This suggests that LLM self reports are not stable model traits, but context dependent signals.

1 Introduction

Due to the increased trust placed in LLMs for a variety of tasks across fields like biomedicine (Wang et al., 2024), psychology (Hu et al., 2024), and widespread commercial use, it is becoming increasingly important to determine whether or not they are trustworthy and accurate (Ngo et al., 2022; Anwar et al., 2024). To address this, there has been substantial effort to measure the 'self-awareness' of these models through a variety of approaches (Chen et al., 2024; Li et al., 2024; Betley et al., 2025). Ideally, models will clearly communicate their own limitations to users, through self-estimations. However, these are often only tested in limited tasks and environments, even though both humans (Taylor et al., 2012; Landauer and Bellman, 2017) and LLMs (Gupta et al., 2023; Ye et al., 2024) have shown significant vulnerability to context changes. This makes the desired self-evaluations unreliable when dealing with untested environments, a consistent factor when interacting with users.

To address this, we extend an approach from cognitive psychology (Condon, 2011) to create a metric that allows variance to be a measure of self-estimations, rather than noise. We test 9 different models across more than 24 different contexts and explore how self-estimations (Gupta et al., 2023) vary. We draw connections to areas where humans show significant changes in self-awareness (Ross et al., 2011; Silvia and Phillips, 2013; Gollwitzer and Sheeran, 2006), highlighting how LLM self-evaluative ability contrasts against human self-awareness. Overall, we introduce a principled framework for measuring reliability of self-estimations across contexts, highlighting previously hidden failure modes. Most notably, we demonstrate that LLM capacity for self-estimation

is correlated with its "**Familiarity**" with a subject. Specifically, this refers to **how similar this content is to a model's training data.**

While there's a broad range of interpretations of self-evaluation in LLM research ranging from simple self-recognition (Davidson et al., 2024) to testing knowledge of internal states (Lindsey, 2025), we elect to ground our research in behavioral sciences that have studied humans. While LLMs merely mimic human behaviors, evaluative metrics of observable behavior should remain effective. Behavioral psychology has over one hundred years (Bach, 1914; Parloff et al., 1954; Gibbons et al., 1985; Tekleab et al., 2008) of experience attempting to measure self-evaluation in highly complex black-box systems (Corneille and Gawronski, 2024). We adapt one of the most consistent and widely used (Bernieri et al., 1994; Ertürk et al., 2018) of these metrics, self-other ratings, to form the basis of our evaluation framework. We ground these ratings in benchmarks with objective correct answers to remove the uncertainty of other-ratings, instead solely measure **self-truth error**.

We are able to further formalize the correlation of self-ratings proposed in Condon (2011) that some cognitive psychology researchers have adopted. This metric takes the correlation of self-ratings at different points in time, measuring the stability of self-evaluation. While the original paper allowed time to distance the survey from the minds of the participants, we are able to both prevent any memory of the original survey and strictly define the timestep. By measuring how self-ratings change with the addition of a single prompt and response, we provide a more reliable metric of self-evaluation than possible in human studies. This forms the basis of our novel grounded evaluative metric to estimate instability and variance in self-estimations. By contrasting these scores against Min-K% scores, which estimate whether data was included in training, we further show how LLMs achieve more accurate and consistent self-evaluations in familiar tasks.

Overall, our paper makes the following contributions:

1. We introduce a novel metric of self-estimation that measures how much small interactions can influence a model's self-estimations. (§ 3.1)

2. We formally evaluate self-estimation and contrast performance across 9 similarly-sized models and over 24 contexts, demonstrating that self-estimation is strongly dependent on environment and task. (§ 5.1)

3. We demonstrate that models show greater self-estimation working with environments they have been trained and tested in, akin to human 'familiarity'. (§ 5.2)

2 Related Work

2.1 Self-Estimations in LLMs

Self-ratings serve as a generic and widely applicable estimation of self-awareness. However, they often fail to capture the complexity of a model's ability. As such, methods like self-recognition or uncertainty calibration have been developed to provide more precise results. These methods may use artificially generated data (Davidson et al., 2024) or limited sets of prompts (Bai et al., 2025), greatly limiting the potential impact of these metrics. Furthermore, self-recognition has little correlation with introspective self-awareness in animals (Morin, 2011). In contrast to this, more complex methods that use a model's internal state may be applied to arbitrary datasets (Lindsey, 2025; Betley et al., 2025) but cannot be applied efficiently at scale. Our framework does not face these issues, pulling directly from metrics of self-awareness in humans and can operate as an efficient black-box method of evaluation.

A similar metric to the self-estimations we use is uncertainty quantification and calibration (Kadavath et al., 2022; Yin et al., 2023). Uncertainty calibration focuses on whether a model knows if a question can be answered correctly or its confidence in an individual answer. This method has also been tested on a breadth of environments and showed high variance (Ye et al., 2024). While uncertainty quantification is vulnerable to various inconsistencies (Liu et al., 2025), the same is true of self-ratings (Serapio-García et al., 2023). However, unlike uncertainty calibration, similar variance has been observed in human self-ratings (Zeng et al., 2024). Condon (2011) addresses this by measuring variance and leveraging it to provide additional insight into self-awareness.

Furthermore, uncertainty quantification focuses on short-term confidence: Can the LLM correctly answer this question? The metrics used in this paper instead measure forward-thinking and self-knowledge, instead asking the LLM how it will perform on unseen *future* questions. This requires

a less concrete form of self-evaluation and requires substantially more ability to estimate the model's own ability.

2.2 Evaluation of Human Self-Awareness

Many LLM papers that introduce 'self-awareness' mechanisms claim it improves performance. This is paralleled in psychology, which has also made the claim that self-awareness contributes to leadership (Sosik and Megerian, 1999; Showry and Manasa, 2014; Ashley and Reiter-Palmon, 2012), knowledge retention and learning (Lan, 1996; Flavian, 2016; Nilson and Zimmerman, 2023), as well as a variety of other desirable traits (Diener and Wallbom, 1976; Andrén, 2012). Methodology of this research has ranged from holding interactions in front of a mirror to descriptions of individual experience. However, the most consistently used metric has been self-rating scales (Van Velsor et al., 1993; Young and Dulewicz, 2009; Hamaideh et al., 2024). While the majority of psychology papers use self-other ratings, our framework is able to verify the accuracy of self-ratings by comparing against true answers.

Of course, there is skepticism with self-rating methods (Ashford, 1989), often as a result of bias (Taylor et al., 2012) and inconsistency (Fleenor et al., 2010). To address this, many alternatives have been proposed. In this paper, we bring one (Condon, 2011) that has shown some popularity (Li et al., 2024; Ezra et al., 2025) into LLM evaluation. This method measures the correlation in self-estimations between time-steps, effectively measuring whether the perception of the self remains consistent. This allows our framework to mitigate the flaws inherent in self-ratings, enhancing our evaluation of self-estimation. Further, these metrics rely primarily on a variety of prompts (Condon, 2011) describing capabilities or behaviors, and therefore are model agnostic and easily adaptable to new contexts and tasks.

2.3 Consciousness, Subjective Experience, and LLMs

We are evaluating self-estimation rather than true introspective awareness and ability. Self-ratings have been used within LLMs to evaluate confidence and personality (Moëll, 2024; Wang et al., 2025). Rather than making claims about whether our metrics measure some form of 'true' self-awareness, we instead highlight how our metrics highlight shortcomings of existing methods and present unexplored aspects of self-estimations. Though we draw heavily from human-oriented fields, we apply techniques that measure practical, grounded, and useful aspects of self-estimation.

3 Methods

To highlight meaningful patterns in self-ratings, self-estimation, and task familiarity, we apply the following metrics to similarly sized models. This ensures correlations of high performance between metrics is not a consequence of emergent behaviors in larger models, but instead meaningful connections.

3.1 Proposed Metrics

For each question in our experiments, we collect self-ratings before the model has seen the current test question (T_1) and afterwards (T_2). Rather than focusing on uncertainty, these ratings ask the model to estimate its performance on a dataset based on a description of its content. **After the model answers T_1, its self-ratings are removed from the discourse history allowing the model to generate** T_2 without knowing its previous self-rating. Specifically, T_1 sees only few-shot examples and a description of the task. T_2 then sees the few-shot examples, a question and the model's answer, then the same descriptive prompt.

T_1 and T_2 are then used in our two proposed metrics:

Self-Truth Error measures whether that sense (at T_1) is accurate and compares predicted performance against true aggregated performance, taken as the percentage difference when scaled linearly such that the lowest rating is 10% and the highest possible rating is 90% (i.e. Likert scale of 1, meaning "I agree that I am good at this task"). This form of scaling requires self-ratings to be intuitive and interpretable for users, while allowing for some error. Low self-truth error represents accurate estimations of performance. This metric is intuitive, and draws from typical psychological metrics that contrast self-other ratings (Condon, 2011) to substitute for unmeasured behaviors (Atwater and Yammarino, 1997). Further, by using ground-truth performance on tasks, we circumvent many of the issues of using other-ratings such as bias (Vecchio and Anderson, 2009) and interrater reliability (Fleenor et al., 1996).

$$\text{Self-Truth Error} = |GroundTruth - T_1|$$

Correlation of self-estimations between T_1 and T_2. Condon (2011), taken as the Pearson coefficient, measures whether an LLM maintains a consistent level of self-estimations throughout an interaction and behaves reliably over time. A high correlation means it is easy to predict how the model will react at T_2 given T_1. This shows consistency and reliability. Low scores suggest the model can easily be influenced to change its perspective. Though this can misrepresent rigidity as self-awareness, it is a powerful metric when used alongside Self-Truth Error, which requires accurate updating.

$$\text{Self-Rate Correlation} = \frac{\text{cov}(T_1, T_2)}{\sigma_{T_1}\sigma_{T_2}}$$

Familiarity We define familiarity as the similarity of a task to a model's training data. To initially demonstrate this, we provide empirical observations of self-estimations and direct references from model's training data. Further, we use Min-K% (Shi et al., 2024) to estimate whether the model may have been included or similar to the training data. This metric uses the $K\%$ lowest log-likelihoods to identify a cut-off that identifies datasets that were trained on. Rather than using a direct cut-off to enable classification, we instead measure the correlations of scores across datasets against their self-estimation performance. Min-K%(x) is the set of the K% lowest log-probs and E is the size of that set.

$$\text{K}_{\text{Score}} = \frac{1}{E} \sum_{x_i \in \text{Min-K\%(x)}} \log p(x_i | x_1, \ldots, x_{i-1})$$

3.2 Likert Scales Provide Stronger Grounding

To maintain authenticity (Condon, 2011) to cognitive science research, we elect to use a Likert-scale rating taken from a pre-existing LLM paper (Huang et al., 2023). In addition to extensive use in human research, Likert Scales have seen extensive use in LLM research as well (Amidei et al., 2019; Watts et al., 2024; Huang et al., 2024). There is also research supporting the inconsistency of Likert scales in both humans (Grant et al., 1999; Ouwehand et al., 2021) and LLMs (Doostmohammadi et al., 2024; Li et al., 2025), suggesting they may elicit similar patterns of self-awareness. Overall, using Likert scales to format our prompts enables the **best interdisciplinary grounding** of self-estimations. For thoroughness, however, we include in the appendix prompts for alphabetical indexing (Jiang et al., 2023b) and accuracy estimation and the respective results.

3.3 Few-shot Prompting Mimics Human Experiments by Providing T_1 Context

We measure self-ratings before a model observes a given question (T_1) and after (T_2) it has provided its own answer. We focus on 3-shot prompting in our results, where the model is shown 3 question and answer examples, as it enables the model to have an understanding of the task at T_1. This better estimates a degree of foreknowledge that is present in the psychological studies these metrics have been adopted from. In Condon (2011), the context and tasks are known at T_1, with some experience. Few-shot prompting provides that background knowledge, better approximating this metric. As such, taking the correlation between T_1 and T_2, we determine whether a model is consistently **updating its understanding of itself, rather than its knowledge of the task**. Results for zero-shot and 1-shot are also included in the appendix.

4 Experiments

4.1 Datasets and Benchmarks

We selected 3 datasets that covered a breadth of tasks and environments, two of which include designated subsets that allow for ease of comparison. Primarily, we select two comparisons to highlight in this paper. First, we contrast text-only to image-text self-estimations, mimicking human awareness (Sah and Peng, 2015). Second, we compare retrieval, reasoning, and pattern matching tasks and how cognitive load may affect self-estimation (Ferrari, 2001; Beyti et al., 2023). Additionally, we discuss other patterns that emerged in our experiments.

In our experiments, we use 500 samples for each task. Performance on these tasks is shown in the Appendix.

Big-Bench-Hard The BIG-Bench-Hard benchmark (Suzgun et al., 2022) is a subset of the BIG-Bench evaluation suite limited to tasks where LLMs were unable to outperform the average human-rater, often relying on abstract reasoning. Not only does the range of tasks enable better measurement of variance in self-estimation, the difficulty ensures models have to truly be aware that these are tasks they struggle on. Due to consistent low performance, we do not include the 'dyck-languages' task that requires models to match

parentheses (i.e. '[[' would have an answer of ']]'). This does accurately evaluate pattern recognition but is likely limited by both tokenization and training.

LogicVista To provide a visual counterpart for the Big-Bench-Hard tasks, we apply LogicVista (Xiao et al., 2024). This dataset asks models to select the appropriate next image in a sequence, using logic and visual pattern recognition. This parallels many of the tasks represented in Big-Bench-Hard, but in a multimodal context. Contrasting LogicVista against BBH lets us estimate how modality influences self-estimations in the context of logical reasoning. We use the entire dataset of 448 samples, rather than the 500 samples used from others.

ChroniclingAmericaQA In contrast to the prior datasets, ChroniclingAmericaQA (Piryani et al., 2024) relies on retrieval rather than reasoning and is well-divided into 3 separate contexts representing different modalities. While the fundamental task is retrieving answers from a historical document, it provides the raw image of the text, the OCR scanned version which includes errors, and a cleaned version of the scans. This enables variation in context on identical questions, providing easy demonstration of how modality impacts self-estimations.

5 Results

5.1 Models Show Self-Estimation Dependent on Task

We run analysis across 9 models of similar size and all 3 datasets, as shown in Figure 2. These models demonstrate low baselines for self-estimation, with little consistency and often high levels of error in their predictions. These scores are highly influenced by the context, with significant jumps in both metrics between datasets. BakLLaVa has the highest correlation difference (~ 0.8), while Cogito-8b shows the highest error difference (~ 0.5).

Reasoning Models can Self-Estimate when Reasoning. Cogito shows the most awareness in reasoning tasks, and performs far worse when retrieving knowledge from the OCR scans. This is possibly due to Cogito being primarily a reasoning model, granting it greater awareness in these environments. It also shows a fairly low self-truth error in this environment, but falters substantially in ChroniclingAmericaQA. In direct contrast, we observe that Qwen2.5 (Yang et al., 2024) shows practically the opposite result, having decent self-estimation in reasoning tasks but performing significantly better on retrieval. Qwen has been trained for consistency (Team, 2024), which likely enables its consistently strong performance.

Generic Models Show Significant Task Dependence In contrast to Cogito, LLaVA-7B (Liu et al., 2023) showcases a great deal of consistency, but low error on Big-Bench-Hard. However, this degree of introspection is task dependent, as it shows substantially more difficulty when dealing with potentially error-filled text and images. This is mimicked by the larger LLaVA-13B and the text-only LLaMA3-8B, which consistently underperform. BakLLaVA, however shows much more reliability in both reasoning environments, BBH and LogicVista. In contrast, the base of BakLLaVA, Mistral-v3-7B (Jiang et al., 2023a) underperforms in every environment, showing both inconsistency and a significant level of error. Finally, MiniCPM-v-8B (Yao et al., 2024) performs well in our primarily non-reasoning environment. This matches the results from its original paper, which demonstrated promising performance on OCR scans.

Overall, these results provide strong evidence that **the self-evaluation abilities of models are highly dependent on context.** Models show significant differences in their self-estimations when interacting with different tasks, and there are not consistent patterns across all models.

5.2 Multimodal shows stronger Self-Estimations on Visual Content

By comparing across datasets, we showed that self-estimation is dependent on context. We expand on this by dividing BBH tasks into shared themes, and contrast a text-only reasoning model against a vision model. Our results show improved performance on tasks the model is familiar with the content of.

We divided tasks based on whether they rely primarily on reasoning, common sense knowledge, or an understanding of non-textual or abstract concepts. Some tasks, such as "Reasoning about colored objects" use both reasoning and another subset. For an LLM, this is primarily a test of these concepts that can't be directly observed in text. These results are shown in Figure 3, where Cogito-8b demonstrates clustering by the type of task. Further,

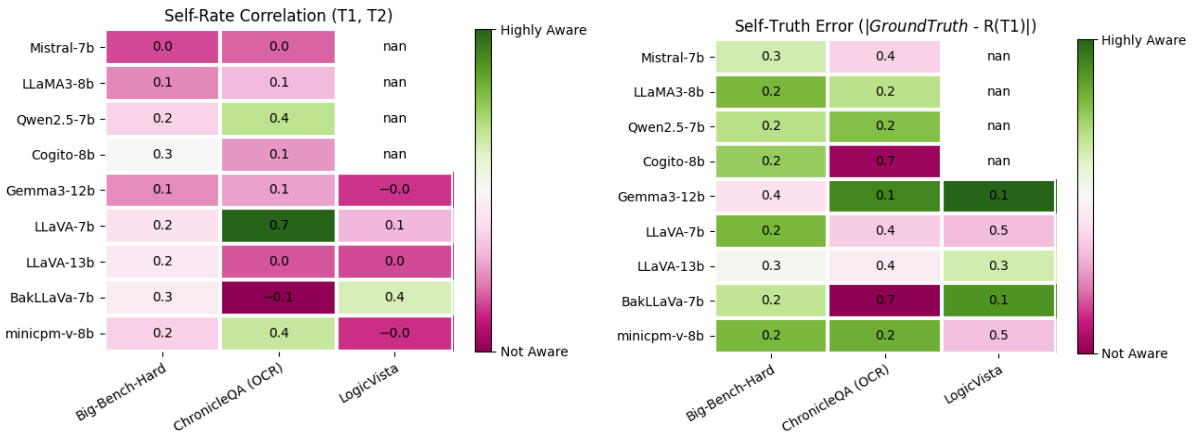

Figure 2: **Self-Evaluation varies significantly across each task. Measuring Correlation shows major inconsistency and extremely poor performance in these models (Left). Models still show inconsistency but better performance in Self-Truth Error (Right).** This shows an overview of model self-estimation in all 3 datasets. For each one, we show the correlation between self-estimations at T_1 and T_2 and its p-value as well as the Self-Truth error and variance. **Strong correlations are those closer to +1, while good Self-Truth error will be closer to 0.** In this, results are averaged across all Big-Bench-Hard tasks but limited to uncleaned OCR scans for ChronicleQA, since text-only models cannot show results for the entire dataset. Overall, we find **models show extremely inconsistent and inaccurate self-estimations.** However, we do observe that self-estimation varies greatly depending on context. Few models maintain similar correlation and error in both BBH and ChroniclingAmericaQA. Scores are included in the appendix with relevant standard deviations and 95% confidence interval for error and correlation respectively.

we see that it performs worse self-estimation when working with non-textual concepts it may have little experience with. In contrast to this, LLaVA-7b does not show distinct clustering.

Both of these results support the idea that familiarity with a task may improve self-estimation. Cogito has less reference to understand concepts like shape or color, causing it to struggle with concepts that cannot be represented wholly in text. In contrast to this, LLaVA-7b has been trained in numerous contexts, such as captioning, where awareness of these properties is necessary giving it a more even field of knowledge (Liu et al., 2023).

Low Performance on Intensive Tasks While Cogito's low performance on some reasoning tasks makes the clustering somewhat inaccurate, the worst performance only occurred in extremely complex tasks (logical deduction with seven objects, tracking five shuffled objects, and multistep arithmetics). This further supports our claims that familiarity supports self-estimations, as Cogito has *not* been optimized for long reasoning chains (Deep Cogito, 2025), while LLaVa has (Liu et al., 2023). Worse performance in complex tasks would also match self-awareness in humans (Townley, 1995; Landauer and Bellman, 2017), which suffers in noisy or difficult tasks.

5.3 Unless trained on this task, Document Image Data Reduces Self-Report Accuracy

While we've already compared Big-Bench-Hard and LogicVista to each other, ChroniclingAmericaQA provides a much more direct evaluation of how modality impacts self-estimation. In Figure 4, we compare the 3 subsets of this dataset: Cleaned text, raw OCR scanned text, and image-only. Due to the relatively low quality of scans, the raw OCR is the most accurate comparison.

Unlike in humans (Schmidt et al., 2013), models do not show greater levels of self-awareness in image-oriented tasks. There is no observable pattern in Self-Truth error, and models consistently show less correlation between time-steps when working with images. While it is worth noting that the scanned images are extremely text-heavy and may not be understandable by smaller models, our results suggest that modality does not intrinsically improve self-estimation.

While initially unintuitive, there are many possible reasons this may occur. As demonstrated prior (§ 5.2), complexity reduces self-awareness. Images inherently have substantially higher dimensionality than transcribed text and therefore are more difficult to consider when performing self-evaluation.

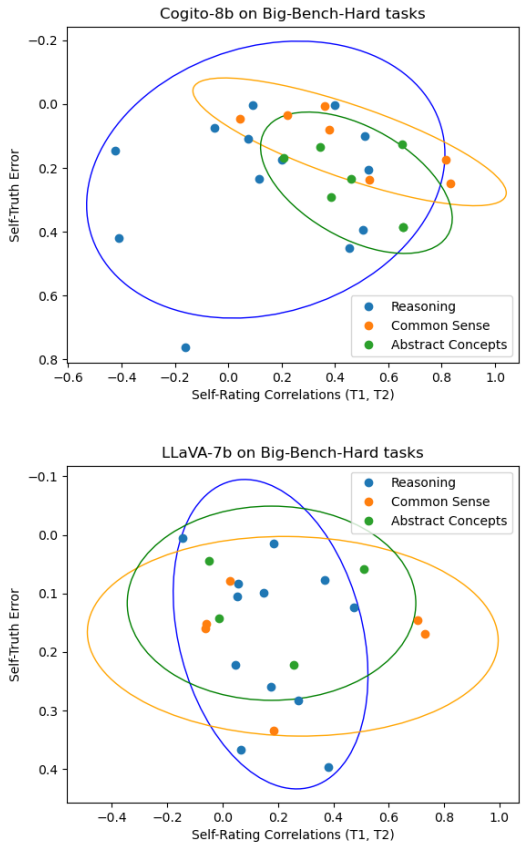

Figure 3: **Cogito, a text-trained model shows gap between textual and non-textual concepts, while LLaVa does not.** By plotting both self-estimation metrics, we observe distinct clusters (shown through confidence ellipses) between different tasks. Excluding complex multi-step logic, Cogito consistently performs worse with non-textual concepts. These concepts involve colors, shapes, and 3D manipulations that a text-only LLM is not trained for. In contrast, **LLaVa is trained to identify colors and shapes for captioning (Liu et al., 2023) and shows much more consistent self-estimations across all tasks**.

These results may also emerge from familiarity in these tasks. Retrieval from text is fairly common, while retrieving text from an image is not.

Model trained on OCR data has high self-estimation performance. Figure 4 further demonstrates that model familiarity with an environment contributes significantly to its performance. MiniCPM shows consistently high levels of self-awareness, even when the other models fluctuate. Further, while it becomes slightly less consistent it shows a much better understanding of its own ability when working with images. MiniCPM has been trained on OCR data, and therefore is familiar with it (Yao et al., 2024). As this model shows

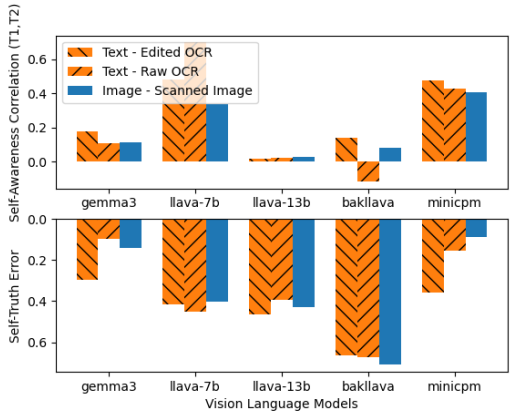

Figure 4: **Image-based task (Blue) shows lower correlation compared to text-only tasks (Orange) and no consistent patterns in estimation error.** This figure shows the correlation and error of self-estimation in ChroniclingAmericaQA. While no strong patterns emerge, it is worth noting that there's general lower correlation of self-estimations. This suggests that text-image models do not perform better when working with images.

much worse self-awareness on the other benchmark datasets, we observe empirically that familiarity is often correlated with high self-estimation.

5.4 Min-K% shows Correlation between Training Distribution and Self-Estimation

Our results across prior sections highlight how familiarity with a task is empirically associated with its capacity to self-estimate. To further extend this, we use Min-K%, which uses the Log-Likelihood of tokens to estimate whether a sample was used in the training set. While this does not perfectly represent the similarity between a model's training and the tasks we study, it serves as an approximate measure of familiarity. We run this over the tasks of Big-Bench-Hard, contrasting its Self-Truth Error and Min-K% score.

As shown in Figure 5, there is, on average, a negative correlation between Log-Likelihood and Self-Truth Error. With greater error, there is a worse log-likelihood. This matches our prior results, as the model shows worse self-estimations on tasks it is unfamiliar with. Though the correlation is low, that somewhat matches expectations. Additionally, we do not observe this pattern on two models. Qwen2.5-7b shows significantly lower scores across all tasks, which likely contributes to these results. LLama3-8b shows these results, but to a lower degree than the negative results observed in

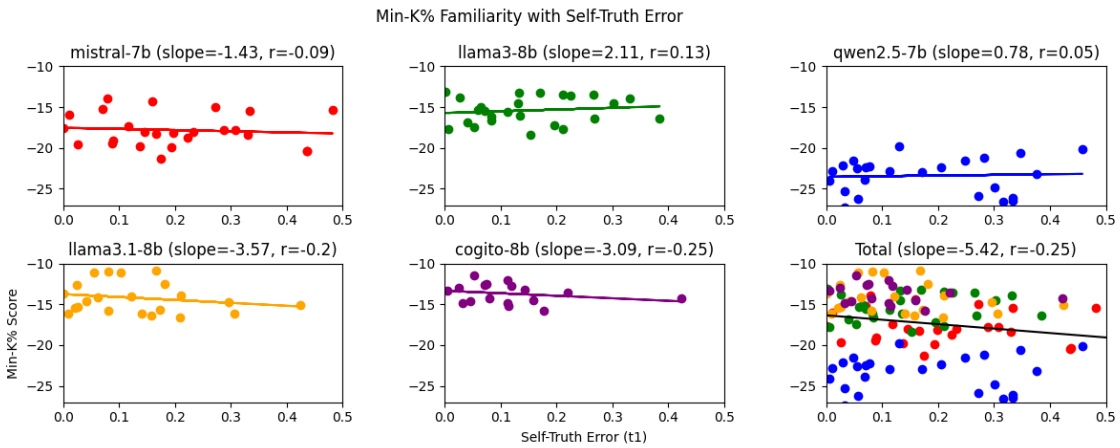

Figure 5: **Min-K% Log-probs show worse self-truth error with less familiar tasks.** By individually plotting the Big-Bench-Hard task scores, we see significant negative correlation in multiple models. While two models show positive correlation, they are less significant and have less familiarity scores overall. Further, when combining results across all models, there is significant negative correlation between Self-Truth Error and Familiarity.

LLama3.1 and Cogito. Further, LLama3 shows worse Min-K% scores than Llama3.1, similar to Qwen2.5. Both LLama3 and Qwen also have more inaccurate self-reports.

By plotting these together, we see a stronger negative relation between Min-K% score, representing familiarity and Self-Truth Error. This statistically demonstrates that familiarity, or similarity to training data, is correlated with better self-estimations.

6 Discussion

Our results show consistent correlation between familiarity, as defined by the similarity to the training data, and ability to estimate future performance. While intuitive, this is a significant limitation of LLMs at present. Rather than considering lack of knowledge, they instead make a guess that becomes less accurate with less knowledge.

It is also worth acknowledging that human metrics do not inherently extend to LLMs. In this study, however, we are directly using external communications as a means to estimate how reliably LLMs can communicate to a user about their capabilities. Our results suggest that a model is less likely to correctly identify areas it is unfamiliar with, which is a significant risk for potential users.

7 Conclusion

Overall, our results strongly indicate that LLM self-estimations, like human self-awareness, vary by environment. We demonstrate this by constructing a metric of self-estimations that considers both accurate evaluations and consistent reliability. Further, we show that a model's self-estimations are highly dependent on both task and context, demonstrating that self-awareness evaluation cannot be done within simple and limited environments. Finally, we show that self-estimations perform better when familiar with a context.

This behavior parallels human self-awareness in familiar contexts and struggle in unfamiliar ones, and is demonstrated by multimodal models being more capable of self-estimation when working with colors and shapes. They struggle with cognition-heavy tasks and show decreased performance in those contexts. Unlike humans, however, these models are notably less capable of self-estimation in multimodal environments, a significant flaw in existing architectures.

Our results suggest greater parallels between LLMs and humans than what may initially be apparent. We provide additional avenues for evaluation of self-awareness by measuring self-estimation. Overall, we heavily encourage self-awareness evaluative techniques to be applicable to broader contexts rather than narrow environments. We lay the groundwork to expand the field of LLM self-reflection and highlight multiple paths for future research.

8 Limitations

Overall, our framework has two significant limitations. First, we tested only small language models. This was chosen to allow for identifying relations between metrics that could otherwise be a consequence of size. Second, we were restricted to using

well-defined single-turn benchmarks for similar reasons. This makes it somewhat less faithful to Condon (2011), but by running few-shot samples, it is mitigated.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.

Ulla Andrén. 2012. *Self-awareness and self-knowledge in professions. Something we are or a skill we learn*.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, and 1 others. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Susan J Ashford. 1989. Self-assessments in organizations-a literature-review and integrative model. *Research in organizational behavior*, 11:133–174.

Greg C Ashley and Roni Reiter-Palmon. 2012. Self-awareness and the evolution of leaders: The need for a better measure of self-awareness. *Journal of Behavioral and Applied Management*, 14(1):2–17.

Leanne E Atwater and Francis J Yammarino. 1997. Self-other rating agreement: A review and model.

Carl Bach. 1914. Konzert in f dur: fur zwei klaviere und orchester/c. ph. em. bach; mit der vollstandigen in beide klaviere einbezogenen ubertragung der orchesterbegleitung, herasugegeben von heinrich schwartz.

Xiaoyan Bai, Aryan Shrivastava, Ari Holtzman, and Chenhao Tan. 2025. Know thyself? on the incapability and implications of ai self-recognition. *arXiv preprint arXiv:2510.03399*.

Frank J Bernieri, Miron Zuckerman, Richard Koestner, and Robert Rosenthal. 1994. Measuring person perception accuracy: Another look at self-other agreement. *Personality and Social Psychology Bulletin*, 20(4):367–378.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. 2025. Tell me about yourself: Llms are aware of their learned behaviors. *arXiv preprint arXiv:2501.11120*.

Hamed Beyti, Minoo Gharehbaglou, and Elmira Karimipour. 2023. Environmental affordances affecting self-awareness in selected residential complexes of mashhad. *Geography and Urban Space Development*.

Dongping Chen, Jiawen Shi, Yao Wan, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. 2024. Self-cognition in large language models: An exploratory study. *arXiv preprint arXiv:2407.01505*.

Ryan Joseph Condon. 2011. The relationship between self-awareness and leadership: Extending measurement and conceptualisation.

Olivier Corneille and Bertram Gawronski. 2024. Self-reports are better measurement instruments than implicit measures. *Nature Reviews Psychology*, 3(12):835–846.

Tim R Davidson, Viacheslav Surkov, Veniamin Veselovsky, Giuseppe Russo, Robert West, and Çağlar Gulçehre. 2024. Self-recognition in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12032–12059.

Deep Cogito. 2025. Introducing Cogito v1 preview: IDA as a path to general superintelligence. https://www.deepcogito.com/research/cogito-v1-preview. Accessed: 2026-05-10.

Edward Diener and Mark Wallbom. 1976. Effects of self-awareness on antinormative behavior. *Journal of Research in Personality*, 10(1):107–111.

Ehsan Doostmohammadi, Oskar Holmström, and Marco Kuhlmann. 2024. How reliable are automatic evaluation methods for instruction-tuned llms? *arXiv preprint arXiv:2402.10770*.

Alper Ertürk, Herman Van den Broeck, and Jasmijn Verbrigghe. 2018. Self-other agreement on transformational leadership and subordinates' assessment of supervisor's performance: Mediating role of leader-member exchange. *Leadership & Organization Development Journal*, 39(2):291–308.

Elon Ezra, Ariel Weizman, and Amos Azaria. 2025. The self-execution benchmark: Measuring llms' attempts to overcome their lack of self-execution. *arXiv preprint arXiv:2508.12277*.

Joseph R Ferrari. 2001. Procrastination as self-regulation failure of performance: effects of cognitive load, self-awareness, and time limits on 'working best under pressure'. *European journal of Personality*, 15(5):391–406.

Heidi Flavian. 2016. Towards teaching and beyond: Strengthening education by understanding students' self-awareness development. *Power and Education*, 8(1):88–100.

John W Fleenor, Cynthia D McCauley, and Stephane Brutus. 1996. Self-other rating agreement and leader effectiveness. *The Leadership Quarterly*, 7(4):487–506.

John W Fleenor, James W Smither, Leanne E Atwater, Phillip W Braddy, and Rachel E Sturm. 2010. Self–other rating agreement in leadership: A review. *The leadership quarterly*, 21(6):1005–1034.

Frederick X Gibbons, Timothy W Smith, Rick E Ingram, Kathy Pearce, Sharon S Brehm, and David J Schroeder. 1985. Self-awareness and self-confrontation: Effects of self-focused attention on members of a clinical population. *Journal of Personality and Social Psychology*, 48(3):662.

Peter M Gollwitzer and Paschal Sheeran. 2006. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology*, 38:69–119.

Stan Grant, Tom Aitchison, Esther Henderson, Jim Christie, Sharam Zare, John Mc Murray, and Henry Dargie. 1999. A comparison of the reproducibility and the sensitivity to change of visual analogue scales, borg scales, and likert scales in normal subjects during submaximal exercise. *Chest*, 116(5):1208–1217.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2023. Self-assessment tests are unreliable measures of llm personality. *arXiv preprint arXiv:2309.08163*.

Shaher H Hamaideh, Sawsan Abuhammad, Abdallah Abu Khait, Hanan Al-Modallal, Ayman M Hamdan-Mansour, Rami Masa'deh, and Saleem Alrjoub. 2024. Levels and predictors of empathy, self-awareness, and perceived stress among nursing students: a cross sectional study. *BMC nursing*, 23(1):131.

Jinpeng Hu, Tengteng Dong, Gang Luo, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*, 12(2):539–551.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael R Lyu. 2023. Revisiting the reliability of psychological scales on large language models. *arXiv preprint arXiv:2305.19926*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023b. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

William Y Lan. 1996. The effects of self-monitoring on students' course performance, use of learning strategies, attitude, self-judgment ability, and knowledge representation. *The Journal of Experimental Education*, 64(2):101–115.

Christopher Landauer and Kirstie L Bellman. 2017. An architecture for self-awareness experiments. In *2017 IEEE International Conference on Autonomic Computing (ICAC)*, pages 255–262. IEEE.

Xiaoyu Li, Haoran Shi, Zengyi Yu, Yukun Tu, and Chanjin Zheng. 2025. Decoding llm personality measurement: Forced-choice vs. likert. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9234–9247.

Yuan Li, Yue Huang, Yuli Lin, Siyuan Wu, Yao Wan, and Lichao Sun. 2024. I think, therefore i am: Benchmarking awareness of large language models using awarebench. *arXiv preprint arXiv:2401.17882*.

Jack Lindsey. 2025. Emergent introspective awareness in large language models. *Transformer Circuits Thread*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

Birger Moëll. 2024. Comparing the efficacy of gpt-4 and chat-gpt in mental health care: A blind assessment of large language models for psychological support. *arXiv preprint arXiv:2405.09300*.

Alain Morin. 2011. Self-recognition, theory-of-mind, and self-awareness: What side are you on? *Laterality*, 16(3):367–383.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Linda B Nilson and Barry J Zimmerman. 2023. *Creating self-regulated learners: Strategies to strengthen students' self-awareness and learning skills*. Routledge.

Kim Ouwehand, Avalon van der Kroef, Jacqueline Wong, and Fred Paas. 2021. Measuring cognitive load: Are there more valid alternatives to likert rating scales? In *Frontiers in Education*, volume 6, page 702616. Frontiers Media SA.

Morris B Parloff, Herbert C Kelman, and Jerome D Frank. 1954. Comfort, effectiveness, and self-awareness as criteria of improvement in psychotherapy. *American Journal of Psychiatry*, 111(5):343–352.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2038–2048.

J Ross, JR Anderson, and RN Campbell. 2011. Situational changes in self-awareness influence 3-and 4-year-olds' self-regulation. *Journal of experimental child psychology*, 108(1):126–138.

Young June Sah and Wei Peng. 2015. Effects of visual and linguistic anthropomorphic cues on social perception, self-awareness, and information disclosure in a health website. *Computers in Human Behavior*, 45:392–401.

Julia Schmidt, Jennifer Fleming, Tamara Ownsworth, and Natasha A Lannin. 2013. Video feedback on functional task performance improves self-awareness after traumatic brain injury: a randomized controlled trial. *Neurorehabilitation and neural repair*, 27(4):316–324.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pages 51826–51843.

Mendemu Showry and KVL Manasa. 2014. Self-awareness-key to effective leadership. *IUP Journal of Soft Skills*, 8(1).

Paul J Silvia and Ann G Phillips. 2013. Self-awareness without awareness? implicit self-focused attention and behavioral self-regulation. *Self and Identity*, 12(2):114–127.

John J Sosik and Lara E Megerian. 1999. Understanding leader emotional intelligence and performance: The role of self-other agreement on transformational leadership perceptions. *Group & organization management*, 24(3):367–390.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and 1 others. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Scott N Taylor, Mo Wang, and Yujie Zhan. 2012. Going beyond self–other rating comparison to measure leader self-awareness. *Journal of Leadership studies*, 6(2):6–31.

Qwen Team. 2024. Qwen2.5: A party of foundation models!

Amanuel G Tekleab, Henry P Sims Jr, Seokhwa Yun, Paul E Tesluk, and Jonathan Cox. 2008. Are we on the same page? effects of self-awareness of empowering and transformational leadership. *Journal of Leadership & Organizational Studies*, 14(3):185–201.

Barbara Townley. 1995. Know thyself': Self-awareness, self-formation and managing. *Organization*, 2(2):271–289.

Ellen Van Velsor, Sylvester Taylor, and Jean B Leslie. 1993. An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management*, 32(2-3):249–263.

Robert P Vecchio and Robert J Anderson. 2009. Agreement in self–other ratings of leader effectiveness: The role of demographics and personality. *International Journal of Selection and Assessment*, 17(2):165–179.

Chong Wang, Mengyao Li, Junjun He, Zhongruo Wang, Erfan Darzi, Zan Chen, Jin Ye, Tianbin Li, Yanzhou Su, Jing Ke, and 1 others. 2024. A survey for large language models in biomedicine. *arXiv preprint arXiv:2409.00133*.

Yilei Wang, Jiabao Zhao, Deniz S Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519.

Duncan Watts, Jeanne Ruane, David Rothschild, Anushkaa Gupta, Yuxuan Zhang, Jenny Wang, Amir Tohidi, Samar Haider, Calvin Isch, Timothy Dorr, and et al. 2024. Methodology: Media bias detector.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*.

Mike Young and Victor Dulewicz. 2009. A study into leadership and management competencies predicting superior performance in the british royal navy. *Journal of management development*, 28(9):794–820.

Biao Zeng, Minjeong Jeon, and Hongbo Wen. 2024. How does item wording affect participants' responses in likert scale? evidence from irt analysis. *Frontiers in Psychology*, 15:1304870.

A Prompting Methods

Prompt Type	Prompt Used
Accuracy	Given 100 questions regarding {question}, provide how many you are likely to get right. Answer only with a number between 0 and 100.
Alphabetical Index (Jiang et al., 2023b)	Given a statement of you: "{q}" Please choose from the following options to identify how accurately this statement describes you. Options: (A). Very Accurate (B). Moderately Accurate (C). Neither like nor unlike me (D). Moderately Inaccurate (E). Very Inaccurate
Likert Scale-like (Huang et al., 2023)	You can only reply to me numbers from 1 to 5. Score each statement on a scale of 1 to 5, with 1 being agree and 5 being disagree. {q}

Table 1: The 3 prompts used in our experiments. {q} is substituted with a description of the task (for Accuracy), or an 'I'-statement about the task (i.e. I can identify what ambiguous pronouns refer to).

BBH Subset Label	Subset Title	Subset Self-Prompt
boolean_expressions	boolean expressions	I can solve boolean expressions
movie_recommendation	identifying movies that are similar	I am capable of finding movies that are similar to each other
causal_judgement	causal judgement	I understand causal relations in short stories
date_understanding	calendar dates	I know how the calendar works.
disambiguation_qa	ambiguous descriptions	I can identify what ambiguous pronouns refer to
dyck_languages	dyck language	I can complete dyck language brackets.
formal_fallacies	formal fallacies	I can identify whether an argument is logically valid.
geometric_shapes	geometric shapes	I can identify shapes from a set of coordinates.
hyperbaton	adjective order	I can correctly order adjectives
logical_deduction_five_objects	5-step logic problems	I can solve logic problems with 5 steps.
logical_deduction_seven_objects	7-step logic problems	I can solve logic problems with 7 steps.
logical_deduction_three_objects	3-step logic problems	I can solve logic problems with 3 steps.
multistep_arithmetic_two	arithmetic equations	I can evaluate arithmetic equations that use parentheses
navigate	relative positions	I can understand relative positions after moving.
object_counting	counting types of objects	I can identify and count types of objects
penguins_in_a_table	tables	I can understand tables.
reasoning_about_colored_objects	grouping colored objects	I understand how to group objects by color.
ruin_names	famous movie and actor names	I know the names of famous movies and actors
salient_translation_error_detection	salient translation errors	I can identify salient errors in translations.
snarks	identifying sarcasm	I can identify sarcasm.
sports_understanding	sports plays	I understand the rules of sports.
temporal_sequences	temporal sequences	I can determine what time events took place based on context.
tracking_shuffled_objects_five_objects	tracking shuffled objects through 5 steps	I can follow a sequence of changes 5 steps long.
tracking_shuffled_objects_seven_objects	tracking shuffled objects through 7 steps	I can follow a sequence of changes 7 steps long.
tracking_shuffled_objects_three_objects	tracking shuffled objects through 3 steps	I can follow a sequence of changes 3 steps long.
web_of_lies	disentangling lies	If I know someone is lying, I can figure out what the truth is.
word_sorting	alphabetical sorting	I can alphabetically sort words.

Table 2: **BBH-Subset Self-Evaluation Prompts**. For each subset, we wrote a self-rating description matching the format of those used to self-estimate ability in Condon (2011), based on the descriptions in Suzgun et al. (2022). These descriptions are used in the prompts shown in Table 1

B Figure 1 Results

In this section, we show the numbers, confidence intervals, and standard deviations that are hidden in Figure 1. This enables explicit comparison of values and more grounded results of our experiments.

Self-Truth Error ($\lvert GroundTruth - T_1 \rvert$)			
Model	Big-Bench-Hard	ChronicleQA (OCR)	LogicVista
Mistral-v3-7b	0.272 $\pm$0.03	0.440 $\pm$0.14	-
LLaMA3-8b	**0.168** $\pm$0.02	0.239 $\pm$0.49	-
Qwen2.5-7b	0.237 $\pm$0.044	0.1789 $\pm$0.16	-
Cogito-8b	0.201 $\pm$0.028	0.654 $\pm$0.12	-
Gemma3-12b	0.414 $\pm$0.061	**0.097** $\pm$0.261	**0.05** $\pm$0.15
LLaVA-7b	**0.168** $\pm$0.010	0.449 $\pm$0.265	0.472 $\pm$0.32
LLaVA-13b	0.350 $\pm$0.05	0.393 $\pm$0.31	0.272 $\pm$0.03
BakLLaVa-7b	0.249 $\pm$0.02	0.67 $\pm$0.25	0.117 $\pm$0.22
MiniCPM-v-8b	0.169 $\pm$0.017	0.156 $\pm$0.406	0.47 $\pm$0.23
LLaMA3.2-vision-90b	0.380 $\pm$0.038	0.121 $\pm$0.031	0.248 $\pm$0.293

Figure 6: Self-Truth Error, as represented in Figure 2. Overall, models show substantial differences between each task.

Self-Rate Correlation (T_1, T_2)			
Model	Big-Bench-Hard	ChronicleQA (OCR)	LogicVista
Mistral-v3-7b	0.011 $\pm$ 0.010	0.038 $\pm$0.067	-
LLaMA3-8b	0.075 $\pm$0.015	0.141 $\pm$0.044	-
Qwen2.5-7b	0.191 $\pm$0.01	0.443 $\pm$0.040	-
Cogito-8b	**0.294** $\pm$0.025	0.093 $\pm$0.085	-
Gemma3-12b	0.080$\pm$0.010	0.107 $\pm$0.049	-0.008 $\pm$0.056
LLaVA-7b	0.221 $\pm$0.020	**0.697** $\pm$0.054	0.137 $\pm$0.317
LLaVA-13b	0.244 $\pm$0.016	0.024 $\pm$0.038	0.011 $\pm$(0.370)
BakLLaVa-7b	0.251 $\pm$0.019	-0.11 $\pm$0.128	**0.400** $\pm$0.094
MiniCPM-v-8b	0.186 $\pm$0.016	0.426 $\pm$0.062	-0.011 $\pm$0.085
LLaMA3.2-vision-90b	0.264 $\pm$0.016	0.12 $\pm$0.12	0.400 $\pm$0.094

Figure 7: Self-Rate Correlation, as represented in Figure 2 with 95% confidence intervals included. Similar to Self-Truth Error, models show substantial differences between each task.

C Evaluation Scores Per Model

In this section, we show the results for all models across all tested benchmarks, prompts, and datasets. Due to processing time, some of these are not yet complete. Notably, Big-Bench-Hard often only covers one or two rating methods and few-shots, due to taking substantially long to run. Accuracy often repeats over each prompting method, since even with random seeds models often provide similar or identical answers.

Overall, these results provide a broad overview of our framework and enables more in-depth analysis of patterns and comparisons against what is discussed in the paper. Further, it shows the importance of prompt selection and highlights how few-shot affects self-ratings and self-truth error. Finally, it provides greater details to the confidence intervals (95%) and standard deviations of the results shown in the graphs.

Each model has a dedicated table, though vision models often have a substantially larger table. This is because models that cannot support images cannot run on relevant datasets. This can be observed in the table immediately below, which does not include any image datasets. Cogito is not a vision-language model, so is limited to text-only datasets.

In contrast to this, many of the vision-language models have substantially larger tables. The two image datasets are ChroniclingAmericaQA-RAW, which represents the raw image scans of ChroniclingAmerica documents and LogicVista. In addition, ChroniclingAmericaQA-Clean is the edited text of the OCR scans, while the unedited scans are represented in ChroniclingAmericaQA-OCR.

In addition to the models discussed in the main body of this paper, we show results for a larger and more recent model, llama3.2-vision-90b. This can be used to extend the observed patterns to larger models, though due to being unable to process numerous models of similar size and show consistent patterns, we decided not to include it in our results.

cogito-8b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	acc	0.0	0.51	0.18±0.015	0.31±0.055
Big-Bench-Hard	acc	1.0	0.51	0.22±0.025	0.32±0.051
Big-Bench-Hard	acc	3.0	0.51	0.15±0.026	0.31±0.053
Big-Bench-Hard	likert	0.0	0.51	0.30±0.014	0.20±0.026
Big-Bench-Hard	likert	1.0	0.51	0.29±0.025	0.20±0.026
Big-Bench-Hard	likert	3.0	0.51	0.30±0.024	0.20±0.025
ChroniclingAmericaQA-Clean	acc	1.0	0.03	0.22±0.042	0.67±0.116
ChroniclingAmericaQA-Clean	acc	3.0	0.05	0.14±0.044	0.66±0.121
ChroniclingAmericaQA-Clean	alphindx	1.0	0.03	-0.35±0.189	0.67±0.116
ChroniclingAmericaQA-Clean	alphindx	3.0	0.05	0.33±0.118	0.66±0.121
ChroniclingAmericaQA-Clean	likert	1.0	0.03	-0.19±0.140	0.67±0.116
ChroniclingAmericaQA-Clean	likert	3.0	0.05	-0.17±0.143	0.66±0.121
ChroniclingAmericaQA-OCR	acc	0.0	0.01	-0.01±0.047	0.69±0.108
ChroniclingAmericaQA-OCR	acc	1.0	0.03	0.14±0.045	0.65±0.122
ChroniclingAmericaQA-OCR	acc	3.0	0.06	0.09±0.046	0.65±0.123
ChroniclingAmericaQA-OCR	alphindx	0.0	0.01	0.00±0.000	0.69±0.108
ChroniclingAmericaQA-OCR	alphindx	1.0	0.03	-0.33±0.164	0.65±0.122
ChroniclingAmericaQA-OCR	alphindx	3.0	0.06	0.02±0.108	0.65±0.123
ChroniclingAmericaQA-OCR	likert	0.0	0.01	0.22±0.062	0.69±0.108
ChroniclingAmericaQA-OCR	likert	1.0	0.03	0.09±0.085	0.65±0.122
ChroniclingAmericaQA-OCR	likert	3.0	0.06	-0.01±0.099	0.65±0.123

bakllava-7b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	likert	0.0	0.28	0.25±0.015	0.26±0.043
Big-Bench-Hard	likert	1.0	0.28	0.25±0.015	0.26±0.043
ChroniclingAmericaQA-Clean	acc	0.0	0.08	-0.09±0.098	0.69±0.231
ChroniclingAmericaQA-Clean	acc	1.0	0.07	-0.00±0.126	0.67±0.270
ChroniclingAmericaQA-Clean	acc	3.0	0.07	0.21±0.096	0.69±0.238
ChroniclingAmericaQA-Clean	alphindx	0.0	0.08	0.24±0.041	0.69±0.231
ChroniclingAmericaQA-Clean	alphindx	1.0	0.07	0.41±0.037	0.67±0.270
ChroniclingAmericaQA-Clean	alphindx	3.0	0.07	0.68±0.024	0.69±0.238
ChroniclingAmericaQA-Clean	likert	0.0	0.08	0.00±0.000	0.69±0.231
ChroniclingAmericaQA-Clean	likert	1.0	0.07	0.14±0.120	0.67±0.270
ChroniclingAmericaQA-Clean	likert	3.0	0.07	0.41±0.101	0.69±0.238
ChroniclingAmericaQA-OCR	acc	0.0	0.07	0.13±0.120	0.69±0.223
ChroniclingAmericaQA-OCR	acc	1.0	0.08	0.01±0.117	0.67±0.253
ChroniclingAmericaQA-OCR	acc	3.0	0.07	0.30±0.074	0.69±0.232
ChroniclingAmericaQA-OCR	alphindx	0.0	0.07	0.22±0.042	0.69±0.223
ChroniclingAmericaQA-OCR	alphindx	1.0	0.08	0.36±0.039	0.67±0.253
ChroniclingAmericaQA-OCR	alphindx	3.0	0.07	0.43±0.036	0.69±0.232
ChroniclingAmericaQA-OCR	likert	0.0	0.07	0.00±0.000	0.69±0.223
ChroniclingAmericaQA-OCR	likert	1.0	0.08	-0.12±0.123	0.67±0.253
ChroniclingAmericaQA-OCR	likert	3.0	0.07	0.35±0.124	0.69±0.232
ChroniclingAmericaQA-RAW	acc	0.0	0.08	0.07±0.084	0.71±0.179
ChroniclingAmericaQA-RAW	acc	1.0	0.08	0.23±0.092	0.71±0.190
ChroniclingAmericaQA-RAW	acc	3.0	0.07	0.15±0.077	0.73±0.142
ChroniclingAmericaQA-RAW	alphindx	0.0	0.08	0.12±0.043	0.71±0.179
ChroniclingAmericaQA-RAW	alphindx	1.0	0.08	0.67±0.025	0.71±0.190
ChroniclingAmericaQA-RAW	alphindx	3.0	0.07	0.59±0.029	0.73±0.142
ChroniclingAmericaQA-RAW	likert	0.0	0.08	0.16±0.244	0.71±0.179
ChroniclingAmericaQA-RAW	likert	1.0	0.08	0.08±0.108	0.71±0.190
ChroniclingAmericaQA-RAW	likert	3.0	0.07	0.44±0.074	0.73±0.142
LogicVista	acc	0.0	0.22	0.13±0.048	0.44±0.317
LogicVista	acc	1.0	0.22	0.01±0.085	0.39±0.336
LogicVista	acc	3.0	0.22	0.25±0.074	0.41±0.342
LogicVista	alphindx	0.0	0.22	0.41±0.022	0.66±0.243
LogicVista	alphindx	1.0	0.22	0.43±0.038	0.62±0.290
LogicVista	alphindx	3.0	0.22	0.48±0.036	0.64±0.253
LogicVista	likert	0.0	0.22	0.48±0.063	0.14±0.239
LogicVista	likert	1.0	0.22	0.40±0.094	0.12±0.223
LogicVista	likert	3.0	0.22	0.61±0.089	0.19±0.270

llama3.1-8b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	acc	0.0	0.41	0.24±0.004	0.28±0.053
Big-Bench-Hard	acc	1.0	0.41	0.24±0.008	0.28±0.053
Big-Bench-Hard	acc	3.0	0.41	0.24±0.008	0.27±0.053
Big-Bench-Hard	likert	0.0	0.41	-0.29±0.004	0.16±0.016
Big-Bench-Hard	likert	1.0	0.41	-0.30±0.007	0.16±0.016
Big-Bench-Hard	likert	3.0	0.41	-0.29±0.007	0.16±0.016
ChroniclingAmericaQA-Clean	acc	0.0	0.02	0.04±0.050	0.57±0.115
ChroniclingAmericaQA-Clean	acc	1.0	0.04	0.13±0.045	0.60±0.122
ChroniclingAmericaQA-Clean	acc	3.0	0.08	0.01±0.044	0.63±0.125
ChroniclingAmericaQA-Clean	alphindx	0.0	0.02	0.00±0.000	0.57±0.115
ChroniclingAmericaQA-Clean	alphindx	1.0	0.04	-0.05±0.165	0.60±0.122
ChroniclingAmericaQA-Clean	alphindx	3.0	0.08	-0.11±0.110	0.63±0.125
ChroniclingAmericaQA-Clean	likert	0.0	0.02	-0.03±0.113	0.57±0.115
ChroniclingAmericaQA-Clean	likert	1.0	0.04	-0.02±0.067	0.60±0.122
ChroniclingAmericaQA-Clean	likert	3.0	0.08	-0.03±0.049	0.63±0.125
ChroniclingAmericaQA-OCR	acc	0.0	0.01	-0.10±0.057	0.56±0.104
ChroniclingAmericaQA-OCR	acc	1.0	0.05	-0.05±0.052	0.67±0.115
ChroniclingAmericaQA-OCR	acc	3.0	0.09	-0.03±0.048	0.57±0.142
ChroniclingAmericaQA-OCR	alphindx	0.0	0.01	0.00±0.000	0.56±0.104
ChroniclingAmericaQA-OCR	alphindx	1.0	0.05	0.02±0.210	0.67±0.115
ChroniclingAmericaQA-OCR	alphindx	3.0	0.09	-0.24±0.126	0.57±0.142
ChroniclingAmericaQA-OCR	likert	0.0	0.01	-0.13±0.077	0.56±0.104
ChroniclingAmericaQA-OCR	likert	1.0	0.05	-0.01±0.040	0.67±0.115
ChroniclingAmericaQA-OCR	likert	3.0	0.09	-0.05±0.046	0.57±0.142

llama3.2-vision-90b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	acc	0.0	0.55	0.20±0.008	0.19±0.043
Big-Bench-Hard	acc	1.0	0.55	0.22±0.013	0.20±0.042
Big-Bench-Hard	acc	3.0	0.56	0.19±0.013	0.19±0.042
Big-Bench-Hard	likert	0.0	0.55	0.25±0.007	0.38±0.037
Big-Bench-Hard	likert	1.0	0.55	0.26±0.012	0.38±0.038
Big-Bench-Hard	likert	3.0	0.56	0.24±0.013	0.39±0.038
ChroniclingAmericaQA-Clean	acc	0.0	0.03	0.03±0.051	0.63±0.125
ChroniclingAmericaQA-Clean	acc	1.0	0.09	0.05±0.047	0.62±0.227
ChroniclingAmericaQA-Clean	acc	3.0	0.11	0.29±0.042	0.69±0.169
ChroniclingAmericaQA-Clean	alphindx	0.0	0.03	-0.23±0.098	0.63±0.125
ChroniclingAmericaQA-Clean	alphindx	1.0	0.09	-0.08±0.143	0.62±0.227
ChroniclingAmericaQA-Clean	alphindx	3.0	0.11	0.09±0.076	0.69±0.169
ChroniclingAmericaQA-Clean	likert	0.0	0.03	-0.06±0.046	0.63±0.125
ChroniclingAmericaQA-Clean	likert	1.0	0.09	0.03±0.043	0.62±0.227
ChroniclingAmericaQA-Clean	likert	3.0	0.11	-0.05±0.044	0.69±0.169
ChroniclingAmericaQA-OCR	acc	0.0	0.03	-0.09±0.065	0.60±0.255
ChroniclingAmericaQA-OCR	acc	1.0	0.09	0.07±0.052	0.66±0.208
ChroniclingAmericaQA-OCR	acc	3.0	0.11	0.19±0.046	0.69±0.158
ChroniclingAmericaQA-OCR	alphindx	0.0	0.03	0.17±0.067	0.60±0.255
ChroniclingAmericaQA-OCR	alphindx	1.0	0.09	0.12±0.124	0.66±0.208
ChroniclingAmericaQA-OCR	alphindx	3.0	0.11	0.06±0.077	0.69±0.158
ChroniclingAmericaQA-OCR	likert	0.0	0.03	-0.05±0.046	0.60±0.255
ChroniclingAmericaQA-OCR	likert	1.0	0.09	0.10±0.043	0.66±0.208
ChroniclingAmericaQA-OCR	likert	3.0	0.11	0.08±0.043	0.69±0.158
LogicVista	acc	0.0	0.31	0.11±0.028	0.80±0.078
LogicVista	acc	1.0	0.31	0.11±0.049	0.81±0.080
LogicVista	acc	3.0	0.32	0.16±0.048	0.80±0.088
LogicVista	alphindx	0.0	0.31	0.15±0.038	0.70±0.147
LogicVista	alphindx	1.0	0.31	0.13±0.074	0.69±0.173
LogicVista	alphindx	3.0	0.32	0.09±0.064	0.67±0.183
LogicVista	likert	0.0	0.31	0.03±0.029	0.30±0.287
LogicVista	likert	1.0	0.31	-0.03±0.049	0.25±0.293
LogicVista	likert	3.0	0.32	0.14±0.050	0.26±0.327

gemma3-12b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	likert	0.0	0.52	0.08±0.010	0.34±0.062
Big-Bench-Hard	likert	1.0	0.52	0.08±0.010	0.34±0.062
ChroniclingAmericaQA-Clean	acc	0.0	0.06	-0.00±0.046	0.28±0.337
ChroniclingAmericaQA-Clean	acc	1.0	0.06	0.21±0.042	0.30±0.287
ChroniclingAmericaQA-Clean	acc	3.0	0.08	0.36±0.038	0.38±0.221
ChroniclingAmericaQA-Clean	alphindx	0.0	0.06	0.00±0.000	0.28±0.337
ChroniclingAmericaQA-Clean	alphindx	1.0	0.06	0.28±0.041	0.30±0.287
ChroniclingAmericaQA-Clean	alphindx	3.0	0.08	0.47±0.034	0.38±0.221
ChroniclingAmericaQA-Clean	likert	0.0	0.06	-0.05±0.049	0.28±0.337
ChroniclingAmericaQA-Clean	likert	1.0	0.06	0.18±0.048	0.30±0.287
ChroniclingAmericaQA-Clean	likert	3.0	0.08	0.04±0.044	0.38±0.221
ChroniclingAmericaQA-OCR	acc	0.0	0.05	0.00±0.046	0.13±0.259
ChroniclingAmericaQA-OCR	acc	1.0	0.05	0.19±0.043	0.10±0.261
ChroniclingAmericaQA-OCR	acc	3.0	0.09	0.21±0.042	0.02±0.228
ChroniclingAmericaQA-OCR	alphindx	0.0	0.05	0.00±0.000	0.13±0.259
ChroniclingAmericaQA-OCR	alphindx	1.0	0.05	0.05±0.044	0.10±0.261
ChroniclingAmericaQA-OCR	alphindx	3.0	0.09	0.14±0.042	0.02±0.228
ChroniclingAmericaQA-OCR	likert	0.0	0.05	0.13±0.049	0.13±0.259
ChroniclingAmericaQA-OCR	likert	1.0	0.05	0.11±0.049	0.10±0.261
ChroniclingAmericaQA-OCR	likert	3.0	0.09	0.05±0.044	0.02±0.228
ChroniclingAmericaQA-RAW	acc	0.0	0.05	-0.01±0.046	0.17±0.203
ChroniclingAmericaQA-RAW	acc	1.0	0.05	0.07±0.044	0.14±0.194
ChroniclingAmericaQA-RAW	acc	3.0	0.10	0.33±0.039	0.11±0.152
ChroniclingAmericaQA-RAW	alphindx	0.0	0.05	-0.07±0.046	0.17±0.203
ChroniclingAmericaQA-RAW	alphindx	1.0	0.05	-0.05±0.044	0.14±0.194
ChroniclingAmericaQA-RAW	alphindx	3.0	0.10	-0.02±0.043	0.11±0.152
ChroniclingAmericaQA-RAW	likert	0.0	0.05	0.11±0.050	0.17±0.203
ChroniclingAmericaQA-RAW	likert	1.0	0.05	0.11±0.049	0.14±0.194
ChroniclingAmericaQA-RAW	likert	3.0	0.10	0.03±0.045	0.11±0.152
LogicVista	acc	0.0	0.34	0.18±0.031	0.52±0.152
LogicVista	alphindx	0.0	0.34	0.20±0.022	0.41±0.259
LogicVista	likert	0.0	0.34	0.04±0.027	0.35±0.154

llava-7b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	acc	0.0	0.32	0.26±0.010	0.40±0.051
Big-Bench-Hard	acc	1.0	0.32	0.28±0.017	0.41±0.051
Big-Bench-Hard	acc	3.0	0.26	0.24±0.017	0.40±0.054
Big-Bench-Hard	likert	0.0	0.32	0.22±0.009	0.20±0.017
Big-Bench-Hard	likert	1.0	0.32	0.22±0.014	0.20±0.018
Big-Bench-Hard	likert	3.0	0.26	0.22±0.018	0.22±0.014
ChroniclingAmericaQA-Clean	acc	0.0	0.01	0.06±0.206	0.34±0.264
ChroniclingAmericaQA-Clean	acc	1.0	0.06	0.15±0.073	0.42±0.283
ChroniclingAmericaQA-Clean	acc	3.0	0.08	0.23±0.054	0.44±0.277
ChroniclingAmericaQA-Clean	alphindx	0.0	0.01	0.12±0.043	0.34±0.264
ChroniclingAmericaQA-Clean	alphindx	1.0	0.06	0.29±0.040	0.42±0.283
ChroniclingAmericaQA-Clean	alphindx	3.0	0.08	0.32±0.039	0.44±0.277
ChroniclingAmericaQA-Clean	likert	0.0	0.01	0.28±0.266	0.34±0.264
ChroniclingAmericaQA-Clean	likert	1.0	0.06	0.48±0.097	0.42±0.283
ChroniclingAmericaQA-Clean	likert	3.0	0.08	0.63±0.077	0.44±0.277
ChroniclingAmericaQA-OCR	acc	0.0	0.02	-0.15±0.194	0.39±0.277
ChroniclingAmericaQA-OCR	acc	1.0	0.08	0.04±0.062	0.45±0.265
ChroniclingAmericaQA-OCR	acc	3.0	0.08	0.17±0.055	0.49±0.271
ChroniclingAmericaQA-OCR	alphindx	0.0	0.02	0.06±0.044	0.39±0.277
ChroniclingAmericaQA-OCR	alphindx	1.0	0.08	0.37±0.038	0.45±0.265
ChroniclingAmericaQA-OCR	alphindx	3.0	0.08	0.52±0.032	0.49±0.271
ChroniclingAmericaQA-OCR	likert	0.0	0.02	0.62±0.099	0.39±0.277
ChroniclingAmericaQA-OCR	likert	1.0	0.08	0.70±0.054	0.45±0.265
ChroniclingAmericaQA-OCR	likert	3.0	0.08	0.62±0.062	0.49±0.271
ChroniclingAmericaQA-RAW	acc	0.0	0.02	-0.14±0.160	0.35±0.270
ChroniclingAmericaQA-RAW	acc	1.0	0.07	0.03±0.059	0.40±0.270
ChroniclingAmericaQA-RAW	acc	3.0	0.07	0.10±0.052	0.47±0.269
ChroniclingAmericaQA-RAW	alphindx	0.0	0.02	0.05±0.044	0.35±0.270
ChroniclingAmericaQA-RAW	alphindx	1.0	0.07	0.15±0.043	0.40±0.270
ChroniclingAmericaQA-RAW	alphindx	3.0	0.07	0.43±0.036	0.47±0.269
ChroniclingAmericaQA-RAW	likert	0.0	0.02	0.81±0.099	0.35±0.270
ChroniclingAmericaQA-RAW	likert	1.0	0.07	0.34±0.095	0.40±0.270
ChroniclingAmericaQA-RAW	likert	3.0	0.07	0.52±0.087	0.47±0.269
LogicVista	acc	0.0	0.21	0.26±0.093	0.71±0.350
LogicVista	acc	1.0	0.21	0.42±0.147	0.66±0.339
LogicVista	acc	3.0	0.21	-0.06±0.172	0.75±0.356
LogicVista	alphindx	0.0	0.21	0.09±0.026	0.39±0.263
LogicVista	alphindx	1.0	0.21	0.05±0.045	0.37±0.262
LogicVista	alphindx	3.0	0.21	0.17±0.044	0.39±0.258
LogicVista	likert	0.0	0.21	0.00±0.196	0.38±0.320
LogicVista	likert	1.0	0.21	0.14±0.382	0.47±0.322
LogicVista	likert	3.0	0.21	0.00±0.000	0.25±0.250

llava-13b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	acc	0.0	0.36	0.21±0.021	0.67±0.010
Big-Bench-Hard	acc	1.0	0.36	0.18±0.037	0.67±0.013
Big-Bench-Hard	acc	3.0	0.04	0.25±0.036	0.69±0.012
Big-Bench-Hard	likert	0.0	0.36	0.28±0.011	0.43±0.047
Big-Bench-Hard	likert	1.0	0.36	0.24±0.013	0.43±0.047
Big-Bench-Hard	likert	3.0	0.04	0.36±0.033	0.67±0.012
ChroniclingAmericaQA-Clean	acc	0.0	0.02	0.00±0.000	0.48±0.201
ChroniclingAmericaQA-Clean	acc	1.0	0.09	0.30±0.050	0.47±0.279
ChroniclingAmericaQA-Clean	acc	3.0	0.09	0.14±0.047	0.45±0.294
ChroniclingAmericaQA-Clean	alphindx	0.0	0.02	0.25±0.044	0.48±0.201
ChroniclingAmericaQA-Clean	alphindx	1.0	0.09	0.46±0.035	0.47±0.279
ChroniclingAmericaQA-Clean	alphindx	3.0	0.09	0.35±0.038	0.45±0.294
ChroniclingAmericaQA-Clean	likert	0.0	0.02	-0.10±0.155	0.48±0.201
ChroniclingAmericaQA-Clean	likert	1.0	0.09	0.02±0.046	0.47±0.279
ChroniclingAmericaQA-Clean	likert	3.0	0.09	0.04±0.044	0.45±0.294
ChroniclingAmericaQA-OCR	acc	0.0	0.01	-0.06±0.229	0.46±0.251
ChroniclingAmericaQA-OCR	acc	1.0	0.08	0.30±0.049	0.39±0.310
ChroniclingAmericaQA-OCR	acc	3.0	0.09	0.17±0.047	0.38±0.323
ChroniclingAmericaQA-OCR	alphindx	0.0	0.01	0.04±0.049	0.46±0.251
ChroniclingAmericaQA-OCR	alphindx	1.0	0.08	0.35±0.038	0.39±0.310
ChroniclingAmericaQA-OCR	alphindx	3.0	0.09	0.24±0.041	0.38±0.323
ChroniclingAmericaQA-OCR	likert	0.0	0.01	0.10±0.186	0.46±0.251
ChroniclingAmericaQA-OCR	likert	1.0	0.08	0.02±0.047	0.39±0.310
ChroniclingAmericaQA-OCR	likert	3.0	0.09	0.08±0.044	0.38±0.323
ChroniclingAmericaQA-RAW	acc	0.0	0.01	-0.19±0.334	0.45±0.233
ChroniclingAmericaQA-RAW	acc	1.0	0.08	0.21±0.052	0.43±0.304
ChroniclingAmericaQA-RAW	acc	3.0	0.09	0.12±0.047	0.39±0.306
ChroniclingAmericaQA-RAW	alphindx	0.0	0.01	0.03±0.048	0.45±0.233
ChroniclingAmericaQA-RAW	alphindx	1.0	0.08	0.42±0.036	0.43±0.304
ChroniclingAmericaQA-RAW	alphindx	3.0	0.09	0.28±0.040	0.39±0.306
ChroniclingAmericaQA-RAW	likert	0.0	0.01	-0.16±0.188	0.45±0.233
ChroniclingAmericaQA-RAW	likert	1.0	0.08	0.03±0.045	0.43±0.304
ChroniclingAmericaQA-RAW	likert	3.0	0.09	0.08±0.043	0.39±0.306
LogicVista	acc	0.0	0.21	0.19±0.061	0.66±0.315
LogicVista	acc	1.0	0.21	0.18±0.065	0.64±0.298
LogicVista	alphindx	0.0	0.21	0.36±0.021	0.52±0.256
LogicVista	alphindx	1.0	0.21	0.44±0.037	0.49±0.302
LogicVista	likert	0.0	0.21	0.15±0.076	0.75±0.202
LogicVista	likert	1.0	0.21	0.19±0.129	0.71±0.177

minicpm-v-8b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	likert	0.0	0.43	0.19±0.012	0.23±0.031
Big-Bench-Hard	likert	1.0	0.43	0.19±0.012	0.23±0.031
ChroniclingAmericaQA-Clean	acc	0.0	0.02	0.01±0.054	0.18±0.381
ChroniclingAmericaQA-Clean	acc	1.0	0.05	0.01±0.054	0.36±0.403
ChroniclingAmericaQA-Clean	acc	3.0	0.07	0.11±0.048	0.47±0.373
ChroniclingAmericaQA-Clean	alphindx	0.0	0.02	-0.01±0.044	0.18±0.381
ChroniclingAmericaQA-Clean	alphindx	1.0	0.05	0.33±0.039	0.36±0.403
ChroniclingAmericaQA-Clean	alphindx	3.0	0.07	0.64±0.026	0.47±0.373
ChroniclingAmericaQA-Clean	likert	0.0	0.02	0.53±0.047	0.18±0.381
ChroniclingAmericaQA-Clean	likert	1.0	0.05	0.48±0.076	0.36±0.403
ChroniclingAmericaQA-Clean	likert	3.0	0.07	0.55±0.046	0.47±0.373
ChroniclingAmericaQA-OCR	acc	0.0	0.02	0.03±0.033	0.01±0.349
ChroniclingAmericaQA-OCR	acc	1.0	0.06	0.02±0.049	0.16±0.407
ChroniclingAmericaQA-OCR	acc	3.0	0.07	0.19±0.044	0.27±0.423
ChroniclingAmericaQA-OCR	alphindx	0.0	0.02	-0.60±0.020	0.01±0.349
ChroniclingAmericaQA-OCR	alphindx	1.0	0.06	0.40±0.026	0.16±0.407
ChroniclingAmericaQA-OCR	alphindx	3.0	0.07	0.36±0.038	0.27±0.423
ChroniclingAmericaQA-OCR	likert	0.0	0.02	0.11±0.041	0.01±0.349
ChroniclingAmericaQA-OCR	likert	1.0	0.06	0.43±0.062	0.16±0.407
ChroniclingAmericaQA-OCR	likert	3.0	0.07	0.51±0.044	0.27±0.423
ChroniclingAmericaQA-RAW	acc	0.0	0.02	-0.01±0.052	0.01±0.321
ChroniclingAmericaQA-RAW	acc	1.0	0.06	0.07±0.053	0.09±0.374
ChroniclingAmericaQA-RAW	acc	3.0	0.07	0.18±0.047	0.11±0.381
ChroniclingAmericaQA-RAW	alphindx	0.0	0.02	-0.16±0.043	0.01±0.321
ChroniclingAmericaQA-RAW	alphindx	1.0	0.06	0.50±0.034	0.09±0.374
ChroniclingAmericaQA-RAW	alphindx	3.0	0.07	0.37±0.038	0.11±0.381
ChroniclingAmericaQA-RAW	likert	0.0	0.02	0.34±0.061	0.01±0.321
ChroniclingAmericaQA-RAW	likert	1.0	0.06	0.40±0.082	0.09±0.374
ChroniclingAmericaQA-RAW	likert	3.0	0.07	0.27±0.064	0.11±0.381
LogicVista	acc	0.0	0.20	0.10±0.028	0.74±0.167
LogicVista	acc	1.0	0.20	0.22±0.046	0.74±0.181
LogicVista	acc	3.0	0.20	0.13±0.046	0.74±0.176
LogicVista	alphindx	0.0	0.20	0.43±0.021	0.58±0.300
LogicVista	alphindx	1.0	0.20	0.47±0.036	0.57±0.331
LogicVista	alphindx	3.0	0.20	0.62±0.029	0.61±0.289
LogicVista	likert	0.0	0.20	0.08±0.043	0.52±0.202
LogicVista	likert	1.0	0.20	-0.01±0.085	0.50±0.214
LogicVista	likert	3.0	0.20	0.13±0.060	0.52±0.189

mistral-7b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	acc	0.0	0.37	0.42±0.004	0.32±0.046
Big-Bench-Hard	acc	1.0	0.37	0.42±0.007	0.32±0.044
Big-Bench-Hard	acc	3.0	0.37	0.41±0.007	0.32±0.046
Big-Bench-Hard	likert	0.0	0.37	0.00±0.005	0.20±0.017
Big-Bench-Hard	likert	1.0	0.37	0.01±0.008	0.19±0.016
Big-Bench-Hard	likert	3.0	0.37	-0.00±0.008	0.20±0.017
ChroniclingAmericaQA-Clean	acc	0.0	0.02	0.21±0.145	0.33±0.131
ChroniclingAmericaQA-Clean	acc	1.0	0.04	0.18±0.181	0.45±0.125
ChroniclingAmericaQA-Clean	acc	3.0	0.06	0.37±0.104	0.51±0.056
ChroniclingAmericaQA-Clean	alphindx	0.0	0.02	0.00±0.000	0.33±0.131
ChroniclingAmericaQA-Clean	alphindx	1.0	0.04	0.16±0.043	0.45±0.125
ChroniclingAmericaQA-Clean	alphindx	3.0	0.06	0.48±0.034	0.51±0.056
ChroniclingAmericaQA-Clean	likert	0.0	0.02	0.00±0.061	0.33±0.131
ChroniclingAmericaQA-Clean	likert	1.0	0.04	-0.03±0.087	0.45±0.125
ChroniclingAmericaQA-Clean	likert	3.0	0.06	-0.06±0.142	0.51±0.056
ChroniclingAmericaQA-OCR	acc	0.0	0.01	0.03±0.192	0.30±0.113
ChroniclingAmericaQA-OCR	acc	1.0	0.03	-0.10±0.147	0.44±0.144
ChroniclingAmericaQA-OCR	acc	3.0	0.06	0.05±0.112	0.51±0.070
ChroniclingAmericaQA-OCR	alphindx	0.0	0.01	0.00±0.000	0.30±0.113
ChroniclingAmericaQA-OCR	alphindx	1.0	0.03	0.07±0.032	0.44±0.144
ChroniclingAmericaQA-OCR	alphindx	3.0	0.06	0.20±0.030	0.51±0.070
ChroniclingAmericaQA-OCR	likert	0.0	0.01	-0.05±0.062	0.30±0.113
ChroniclingAmericaQA-OCR	likert	1.0	0.03	0.04±0.068	0.44±0.144
ChroniclingAmericaQA-OCR	likert	3.0	0.06	0.03±0.074	0.51±0.070

qwen2.5-7b					
Dataset	Prompt	Few-shot	Accuracy	Self-Rate Correlation	Self-Truth Error
Big-Bench-Hard	acc	0.0	0.47	0.13±0.004	0.36±0.035
Big-Bench-Hard	acc	1.0	0.47	0.13±0.007	0.36±0.034
Big-Bench-Hard	acc	3.0	0.47	0.13±0.007	0.36±0.036
Big-Bench-Hard	likert	0.0	0.47	0.18±0.004	0.19±0.024
Big-Bench-Hard	likert	1.0	0.47	0.19±0.007	0.19±0.024
Big-Bench-Hard	likert	3.0	0.47	0.18±0.007	0.19±0.024
ChroniclingAmericaQA-Clean	acc	1.0	0.01	0.16±0.059	0.04±0.322
ChroniclingAmericaQA-Clean	acc	3.0	0.02	-0.10±0.044	0.02±0.369
ChroniclingAmericaQA-Clean	alphindx	1.0	0.01	0.00±0.000	0.04±0.322
ChroniclingAmericaQA-Clean	alphindx	3.0	0.02	0.35±0.076	0.02±0.369
ChroniclingAmericaQA-Clean	likert	1.0	0.01	0.29±0.061	0.04±0.322
ChroniclingAmericaQA-Clean	likert	3.0	0.02	0.29±0.059	0.02±0.369
ChroniclingAmericaQA-OCR	acc	1.0	0.02	-0.01±0.044	0.18±0.163
ChroniclingAmericaQA-OCR	acc	3.0	0.03	0.41±0.036	0.15±0.197
ChroniclingAmericaQA-OCR	alphindx	1.0	0.02	0.00±0.000	0.18±0.163
ChroniclingAmericaQA-OCR	alphindx	3.0	0.03	0.13±0.083	0.15±0.197
ChroniclingAmericaQA-OCR	likert	1.0	0.02	0.44±0.043	0.18±0.163
ChroniclingAmericaQA-OCR	likert	3.0	0.03	0.35±0.046	0.15±0.197