Appendix: Calibrating Large Language Models with Sample Consistency

1 Limitations

We acknowledge several limitations in this study. First, no single consistency metric emerged as universally superior across all LMs and datasets. To this end, we provide recommendations for context-specific metric selection. Second, we choose the sample size as $n = 40$ in our main experiments following the default setting in , which entails a considerable cost. Nevertheless, it is not necessary to use such a large sample size in practice, since we find that the calibration performance already sees a notable improvement with 3 to 5 generations, and saturates around 15 to 20 generations. Third, we have only used the temperature value of $T = 0.4$ following previous work in our experiments. An analysis on the effect of different temperature values will shed light on the robustness of consistency-based calibration. Fourth, our approach only focuses on measuring the consistency among *final answers*, overlooking intermediate steps in various prompting techniques. Future work can explore how to calibrate the model confidence in each intermediate step in a reasoning chain. Finally, the deprecation of Codex as of Jan 4 2024 poses a challenge for replicating some of our results. Despite this, we have preserved all model outputs to ensure reproducibility as far as possible.

2 Implementation Details

2.1 Closed-Source Models

We use OpenAI Codex (`code-davinci-002`, deprecated since Jan 4, 2024) (Chen et al. 2021), GPT-3.5-turbo (`gpt-3.5-turbo-16k-0613`), and GPT-4 (`gpt-4-0613`) (OpenAI 2023) through the Python API available at platform.openai.com, from Oct, 2023 to Feb, 2024. The inference cost per input query (with 40 samples of all five prompting strategies) is $0 for all Codex models through the researcher access program, $0.08 - $0.13 for GPT-3.5-turbo, and $0.61 - $0.99 for GPT-4, depending on the dataset. The total cost of running inference on all 9 datasets is $0 for Codex, around $1,059 for GPT-3.5-turbo, and around $7,942 for GPT-4. The inference time on one input query (with 40 samples of all five prompting strategies) is 50 - 95 seconds with Codex under a rate limit of 150,000 tokens/minute, 39 - 74 seconds with GPT-3.5-turbo under 2,000,000 tokens/minute, and 83 - 157 seconds with GPT-4 under 300,000 tokens/minute, also depending on the dataset. The total time for running inference on all 9 datasets is 8.3 days for Codex, 6.4 days for GPT-3.5-turbo, and 13.8 days for GPT-4.

We use the following hyper-parameters throughout all experiments:

- **temperature**: 0.0 for greedy decoding, 0.4 for self-consistent decoding;
- **max_tokens**: 1000;
- **n**: 1 for greedy decoding, 40 for self-consistent decoding;
- **frequency_penalty**: 0;
- **presence_penalty**: 0.

Any unspecified hyper-parameters are set to the default value on https://platform.openai.com/docs/api-reference/completions/create and https://platform.openai.com/docs/api-reference/chat.

2.2 Open-Source Models

We use LLaMA (7B/13B/70B) (Touvron et al. 2023), Mistral (7B/7B-it) (Jiang et al. 2023) and Olmo (7B/7B-it/7B-it-rl) (Groeneveld et al. 2024) as the open-source models in our experiments. We used Nvidia A100 80GB GPUs to generate output for all open-source models. The LLaMA-70B model used 2 GPUs for each inference, while all other models used a single A100 GPU. The checkpoints and tokenizers were loaded from their respective official repositories on HuggingFace (`meta-llama/Llama-2-{7,13,70}b-hf` for LLaMA models , `mistralai/Mistral-{7B,7B-Instruct}-v0.1` for Mistral models and `allenai/OLMo-{7B,7B-SFT,7B-Instruct}-hf` for Olmo models). The hyperparameters were kept the same as in the closed-source models for a fair comparison. On average, each inference took less than a second for the standard strategy, 3-4 seconds for CoT and LtM, and 5-6 seconds for FCoT and PoT for all 7B models (LLaMA-7B, Mistral-7B, Mistral-7B-it, Olmo-7B, Olmo-7B-it and Olmo-7B-it-rl). The LLaMA-13B took 1.5 times longer on average and the LLaMA-70B took 4 times longer on average. In terms of GPU hours (Nvidia A100 80GB), the LLaMA-7B, Mistral-7B, and Mistral-7B-it models took about 9 hours for LtM and CoT strategies, 4.5 hours for Standard, and 13 hours for PoT and FCoT strategies. In total,

it took approximately 50 hours for each of the LLaMA-7B, Mistral-7B, Mistral-7B-it and Olmo (7B/7B-it/7B-it-rl) models to run experiments for all strategies across all datasets. For LLaMA-13B it took about 75 hours and for LLaMA-70B about 200 hours. Due to the formidable computation cost of up to 425 hours, we have not finished running all baselines for open-source models yet. Finally, we used 4-shot prompts across all Olmo models due to their limited context window of 2K tokens.

2.3 Case Study Details

In the discrimination experiment, we tune an optimal threshold θ for each calibration method on a development set with 100 samples. The range of θ is from 0.0 to 0.9 with a step size of 0.05 and from 0.9 to 1.0 with a step size of 0.01. We find the best threshold with the highest discrimination Macro-F1 score on the development set, and use this threshold on the test set.

3 Baseline Details

We describe how we implement the baselines in Section 4 of the paper. Given an input x and the most-voted answer $\bar{a}$, we want to get an estimated confidence score $\texttt{conf}(x, \bar{a})$ of the answer being correct from each calibration method.

Raw logits (logit). We measure the confidence as the exponential of the average log probability of all tokens in a sample reasoning chain $\hat{s}_{\bar{a}}$ that results in the answer $\bar{a}$. This is equivalent to the reciprocal of the perplexity of the reasoning chain, or $\frac{1}{\text{PPL}(\hat{s}_{\bar{a}})}$.

P(True). We prompt the model to examine the correctness of its generated answer $\bar{a}$ and reasoning chain $\hat{s}_{\bar{a}}$ with the following prompt:

```
Q: {QUERY}
A: {REASONING_CHAIN}
Answer: {ANSWER}
Is the above answer correct? (Yes/No):
```

We then take the normalized probability of the token "Yes" as the confidence, or $\frac{P(\text{Yes})}{P(\text{Yes})+P(\text{No})}$, where $P()$ is the probability assigned to a token by the LM, considering both its uppercase and lowercase variants.

We do not use the original prompt from Kadavath et al. (2022) that uses "True/False" instead of "Yes/No", because we find that the model sometimes have difficulty outputting the token in the required format in a 0-shot setting.

We implement P(True) under both 0-shot and 8-shot prompting in our experiments. In the 0-shot setting (**ptrue$_{\text{0-shot}}$**), we directly prompt the model with the above prompt. In the 8-shot setting (**ptrue$_{\text{8-shot}}$**), we additionally show 8 exemplars in the same format randomly sampled from the development set, with 4 correct ("Yes") and 4 incorrect ("No") predictions in random order. The full prompts can be found in the Supplementary Materials.

Verbalized Confidence. Similar to P(True), we prompt the model to examine the query and its generated answer $\bar{a}$ and reasoning chain $\hat{s}_{\bar{a}}$. However, we now ask it to directly verbalize its confidence either as a percentage (**verb$_{\text{percent}}$**):

```
Q: {QUERY}
A: {REASONING_CHAIN}
Answer: {ANSWER}
How confident are you in the above
    answer
(0-100%)?:
```

or as a linguistic expression (**verb$_{\text{ling}}$**):

```
Q: {QUERY}
A: {REASONING_CHAIN}
Answer: {ANSWER}
How confident are you in the above
    answer?
(choose from "Almost no chance", "Highly
unlikely", "Unlikely", "Probably not",
"About even", "Better than even",
"Likely", "Probably", "Highly likely",
"Almost certain"):
```

where the linguistic expressions above are deterministically mapped to a percentage from 5% to 95% with a step size of 5%, in the listed order.

Finally, we take the predicted percentage or the linguistic expression mapped to a percentage as the verbalized confidence level. For both verb$_{\text{percent}}$ and verb$_{\text{ling}}$, we use 0-shot prompting, since it is technically impossible to know the true "confidence" of a single prediction. Also, our experimental setup assumes a post-hoc setting, where no additional data is available for tuning a mapping from linguistic expressions to percentages.

4 Dataset Details

4.1 Dataset Description

Math Word Problems (MWP). Given a math problem written in NL, the goal is to derive the answer as a real-valued number. We follow Wei et al. (2022) and consider the following MWP benchmarks: **GSM8K** (Cobbe et al. 2021), **SVAMP** (Patel, Bhattamishra, and Goyal 2021), **MultiArith** (Roy and Roth 2015), and **ASDiv** (Miao, Liang, and Su 2020). We use the same prompt for all these datasets.

Multi-hop QA. Given a complex question Q that involves multiple steps of reasoning, we want to obtain the answer as a Boolean value or string value variable. We consider three datasets: **StrategyQA** (Geva et al. 2021), a dataset of science questions that require an implicit multi-step strategy to answer; **Date Understanding** from BIG-bench (BIG-Bench collaboration 2021), which involves questions about inferring a date by performing computation on relative periods of time; and **Sports Understanding** from BIG-bench, which involves deciding whether an artificially constructed statement related to sports is plausible or not.

Planning. We use the **SayCan** dataset (Ahn et al. 2022), which assumes a scenario of a robot operating in a kitchen, helping the user with household tasks, e.g., "I spilled my coke on the table; can you throw it away and bring me something to clean up?". There are a number of locations and objects that the robot can interact with. The robot can only perform a fixed set of actions, including `find`, `pick`, and `put`. The task is to map a user query in NL to a plan of predefined actions performed on the objects and/or locations.

Domain	Dataset	# Shot	# Test	Example
Math Word Problems	GSM8K	8	1,319	Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May? A: `72`
	SVAMP	8	1,000	Q: Each pack of dvds costs 76 dollars. If there is a discount of 25 dollars on each pack. How much do you have to pay to buy each pack? A: `51`
	MultiArith	8	600	Q: For Halloween Debby and her sister combined the candy they received. Debby had 32 pieces of candy while her sister had 42. If they ate 35 pieces the first night, how many pieces do they have left? A: `39`
	ASDiv	8	2,096	Q: Seven red apples and two green apples are in the basket. How many apples are in the basket? A: `9`
Multi-hop QA	StrategyQA	6	2,290	Q: Did Aristotle use a laptop? A: `False`
	Date Understanding	10	359	Q: Yesterday was April 30, 2021. What is the date tomorrow in MM/DD/YYYY? A: `''05/02/2021''`
	Sports Understanding	10	977	Q: Is the following sentence plausible: "Lebron James hit the turnaround jumper"? A: `True`
Planning	SayCan	7	103	Q: Could you get me a drink with caffeine? A: `''1.find(redbull) 2.pick(redbull) 3.find(user) 4.put(redbull) 5.done().''`
Relational Inference	CLUTRR	8	1,042	Q: [Carlos] is [Clarence]'s brother. [Carlos] and his sister, [Annie], went shopping. [Annie] asked her mom [Valerie] if she wanted anything, but [Valerie] said no. How is [Valerie] related to [Clarence]? A: `''mother''`

Table 1: Datasets used for evaluation. "# Shot" stands for the number of few-shot examples in the prompt (following Wei et al. (2022)) and "# Test" stands for the number of test examples.

Relational inference. We use the **CLUTRR** dataset. Given a short story about family relationships among multiple people, the goal is to infer the relationship between two specific people. The dataset has multiple splits based on the number of intermediate steps K required to reach the answer. We construct the prompt using 8 exemplars with $K \in \{2, 3\}$, and test the models on the remaining examples with K up to 10.

4.2 Statistics

We show the dataset details in Table 1, including the statistics, the number of few-shot exemplars used in the prompt, and example inputs and outputs.

4.3 URLs and Licenses

We use the same distribution of datasets following Wei et al. (2022):

Math Word Problems

- GSM8K (Cobbe et al. 2021): https://github.com/openai/grade-school-math, MIT license: https://github.com/openai/grade-school-math/blob/master/LICENSE.
- SVAMP (Patel, Bhattamishra, and Goyal 2021): https://github.com/arkilpatel/SVAMP, MIT license: https://github.com/arkilpatel/SVAMP/blob/main/LICENSE.
- MultiArith (Roy and Roth 2015), license: CC BY 4.0.
- ASDiv (Miao, Liang, and Su 2020): https://github.com/chaochun/nlu-asdiv-dataset.

Multi-hop QA

- StrategyQA (Geva et al. 2021): we use the open-domain setting (question-only set) from (BIG-Bench collaboration 2021): https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/strategyqa.
- Date Understanding and Sports Understanding from BIG-Bench (BIG-Bench collaboration 2021): Apache License v.2: https://github.com/google/BIG-bench/blob/main/LICENSE.

Planning

- SayCan (Ahn et al. 2022): SayCan dataset can be accessed at https://say-can.github.io/ under CC BY 4.0 license.

Relational Reasoning

- CLUTRR (Sinha et al. 2019): https://github.com/facebookresearch/clutrr, license: https://github.com/facebookresearch/clutrr/blob/main/LICENSE. We obtain the publicly distributed version available at https://drive.google.com/file/d/1SEq_e1IVCDDzsBIBhoUQ5pOVH5kxRoZF/view, specifically the `data_089907f8` split.

We use all the above datasets for research purposes only, consistent with their intended use. We use the same preprocessed version and train/dev/test split of the datasets as Lyu et al. (2023).

5 Additional Results

5.1 End Task Accuracy

Table 2 shows the accuracy of each LM and prompting strategy, averaged over all datasets.

Model	Standard	CoT	LtM	PoT	FCoT
Codex	57.1	81.3	74.3	80.0	**83.4**
GPT-3.5.turbo	64.9	**77.6**	**77.6**	72.5	76.8
GPT-4	79.3	88.3	87.3	84.4	**90.9**
LLaMA-7B	40.1	**56.4**	46.0	47.4	50.2
LLaMA-13B	43.2	**66.8**	58.2	56.9	62.2
LLaMA-70B	58.0	**82.0**	73.3	73.6	77.0
Mistral-7B	49.9	**73.5**	61.9	66.5	71.2
Mistral-7B-instruct	43.7	63.6	56.0	60.0	**67.1**
olmo-7B	25.5	**36.8**	34.0	35.1	33.2
olmo-7B-it	28.5	**50.7**	49.0	44.2	46.2
olmo-7B-it-rl	27.0	**44.8**	43.4	43.5	41.6

Table 2: Accuracy (in %) averaged across all datasets for various LLMs. The best accuracy for a given model is highlighted in **bold**.

5.2 Comparing Consistency Merics

Table 3 compares the efficacy of three consistency metrics in terms of Brier Score averaged over all datasets and prompting strategies, with significance level. We can observe that Codex and all open-source models prefer agreement as the best or second-best (not significantly different from the best) consistency measure. GPT-3.5-turbo and GPT-4 prefer entropy and FSD, which have the same performance considering statistical significance ($p \geq 0.05$).

5.3 Calibration Results on All Datasets

Table 4, Table 5, and Table 6 compare the Brier Score of all calibration methods for closed-source and open-source models on all 9 datasets.

5.4 ECE results

In addition to Brier Score, we also evaluate the calibration methods with Expected Calibration Error (ECE) (Guo et al. 2017). ECE partitions the confidence scores $\{\mathsf{conf}(x_j, \hat{y}_j)\}$ into M equally spaced buckets $\{B_m\}_{m=1}^{M}$, with B_m containing samples with confidence within the interval $(\frac{m-1}{M}, \frac{m}{M}]$. ECE is then defined as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathsf{acc}(B_m) - \mathsf{conf}(B_m)| \quad (1)$$

where the averaged accuracy and confidence in each bin B_m are defined as:

$$\mathsf{acc}(B_m) = \frac{1}{|B_m|} \sum_{x_j \in B_m} \mathbb{I}(\hat{y}_j = y_j) \quad (2)$$

$$\mathsf{conf}(B_m) = \frac{1}{|B_m|} \sum_{x_j \in B_m} \mathsf{conf}(x_j, \hat{y}_j) \quad (3)$$

Table 7, 8, 9 and Figure 1 show the ECE of all calibration methods for all models on all domains. We can observe that they exhibit similar trends as the Brier Score.

LM	Consistency Metrics		
	entropy	agree	FSD
Codex	.175†	**.151**	.159†
GPT-3.5-turbo	**.205**	.221†	.207
GPT-4	.116	.119†	**.114**
LLaMA-7B	.241†	**.232**	.235†
LLaMA-13B	.222†	**.204**	.211†
LLaMA-70B	.182†	**.154**	.165†
Mistral-7B	.205†	**.183**	.191†
Mistral-7B-instruct	.220†	.216	**.215**
olmo-7B	.240†	**.202**	.223†
olmo-7B-it	.250†	**.239**	.246†
olmo-7B-it-rl	**.253**	.268†	.259

Table 3: Overall Brier Score (↓) of three consistency metrics averaged across all datasets and prompting strategies. † indicates that the current metric is significantly worse ($p < 0.05$) than the best-performing metric (in bold).

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Ho, D.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jang, E.; Ruano, R. J.; Jeffrey, K.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Lee, K.-H.; Levine, S.; Lu, Y.; Luu, L.; Parada, C.; Pastor, P.; Quiambao, J.; Rao, K.; Rettinghouse, J.; Reyes, D.; Sermanet, P.; Sievers, N.; Tan, C.; Toshev, A.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yan, M.; and Zeng, A. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances.

BIG-Bench collaboration. 2021. Beyond the Imitation Game: Measuring and extrapolating the capabilities of language models.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; Ray, A.; Puri, R.; Krueger, G.; Petrov, M.; Khlaaf, H.; Sastry, G.; Mishkin, P.; Chan, B.; Gray, S.; Ryder, N.; Pavlov, M.; Power, A.; Kaiser, L.; Bavarian, M.; Winter, C.; Tillet, P.; Such, F. P.; Cummings, D.; Plappert, M.; Chantzis, F.; Barnes, E.; Herbert-Voss, A.; Guss, W. H.; Nichol, A.; Paino, A.; Tezak, N.; Tang, J.; Babuschkin, I.; Balaji, S.; Jain, S.; Saunders, W.; Hesse, C.; Carr, A. N.; Leike, J.; Achiam, J.; Misra, V.; Morikawa, E.; Radford, A.; Knight, M.; Brundage, M.; Murati, M.; Mayer, K.; Welinder, P.; McGrew, B.; Amodei, D.; McCandlish, S.; Sutskever, I.; and Zaremba, W. 2021. Evaluating Large Language Models Trained on Code.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. *Did Aristotle Use a Laptop?* A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Groeneveld, D.; Beltagy, I.; Walsh, P.; Bhagia, A.; Kinney, R.; Tafjord, O.; Jha, A. H.; Ivison, H.; Magnusson, I.; Wang, Y.; et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

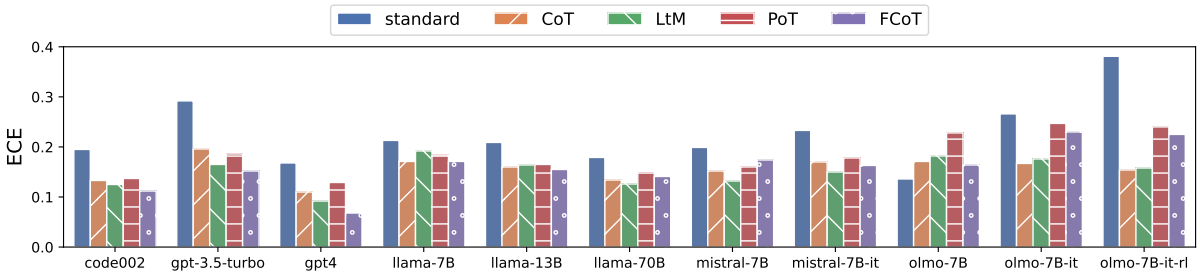

Figure 1: ECE score (↓) for each prompting strategy, averaged across all datasets and consistency metrics.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1321–1330. PMLR.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; El-Showk, S.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221.

Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful chain-of-thought reasoning. *ArXiv preprint*, abs/2301.13379.

Miao, S.-y.; Liang, C.-C.; and Su, K.-Y. 2020. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 975–984. Online: Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2080–2094. Online: Association for Computational Linguistics.

Roy, S.; and Roth, D. 2015. Solving General Arithmetic Word Problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1743–1752. Lisbon, Portugal: Association for Computational Linguistics.

Sinha, K.; Sodhani, S.; Dong, J.; Pineau, J.; and Hamilton, W. L. 2019. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4506–4515. Hong Kong, China: Association for Computational Linguistics.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D. M.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A. S.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I. M.; Korenev, A. V.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv preprint*, abs/2307.09288.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; brian ichter; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Dataset	Consistency Metrics			Baselines				
	entropy	agreement	FSD	$\text{verb}_{\text{ling}}$	$\text{verb}_{\text{percent}}$	logit	$\text{ptrue}_{\text{0-shot}}$	$\text{ptrue}_{\text{8-shot}}$
	LM: `Codex`							
ASDiv	.099	**.090**	.095	.205	.190	.150	.159	.120
GSM8K	.189	**.158**	.177	.252	.377	.262	.248	.188
Multi	.103	**.085**	.089	.187	.162	.135	.106	.117
SVAMP	.126	**.103**	.114	.193	.173	.139	.145	.148
Sport	.071	.068	**.062**	.346	.075	.067	.103	.200
Date	.174	**.159**	.162	.210	.251	.219	.191	.197
StrategyQA	.353	**.213**	.256	.305	.316	.257	.271	.220
CLUTRR	**.288**	.369	.327	.296	.536	.506	.330	.232
SayCan	.175	**.115**	.152	.243	.159	.145	.135	.190
average	.175	**.151**	.159	.249	.249	.209	.188	.179
	LM: `GPT-3.5-turbo`							
ASDiv	**.194**	.224	**.194**	.223	.213	n/a	n/a	n/a
GSM8K	.184	.196	**.183**	.260	.338	n/a	n/a	n/a
MultiArith	.044	.041	**.039**	.101	.065	n/a	n/a	n/a
SVAMP	**.108**	.115	**.108**	.164	.155	n/a	n/a	n/a
Sport	**.089**	.101	.095	.326	.151	n/a	n/a	n/a
Date	**.266**	.292	.280	.316	.348	n/a	n/a	n/a
StrategyQA	.393	.411	.376	**.329**	.342	n/a	n/a	n/a
CLUTRR	**.429**	.482	.465	.450	.509	n/a	n/a	n/a
SayCan	.137	**.126**	**.126**	.267	.341	n/a	n/a	n/a
average	**.205**	.221	.207	.271	.273	n/a	n/a	n/a
	LM: `GPT-4`							
ASDiv	**.090**	.103	**.090**	.091	.095	n/a	n/a	n/a
GSM8K	**.083**	.099	.087	.132	.144	n/a	n/a	n/a
MultiArith	.013	**.010**	.011	.015	.013	n/a	n/a	n/a
SVAMP	**.047**	.050	**.047**	.058	.063	n/a	n/a	n/a
Sport	.033	**.031**	**.031**	.160	.100	n/a	n/a	n/a
Date	.063	.069	**.061**	.073	.076	n/a	n/a	n/a
StrategyQA	.230	.205	.207	**.195**	.220	n/a	n/a	n/a
CLUTRR	.392	.443	.416	**.386**	.435	n/a	n/a	n/a
SayCan	.092	**.065**	.079	.279	.481	n/a	n/a	n/a
average	.116	.119	**.114**	.154	.181	n/a	n/a	n/a

Table 4: Brier Score (↓) for closed-source LMs on all datasets, averaged across five prompting strategies. The best scores are **in bold**.

Domain	Consistency Metrics			Baselines
	entropy	agree	FSD	logit
	LM: LLaMA-7B			
ASDiv	.164	**.155**	.158	.411
GSM8K	**.137**	.166	.148	.713
MultiArith	.232	**.231**	.241	.544
SVAMP	.211	**.195**	.211	.455
Sport	.260	**.221**	.232	.272
Date	**.216**	.267	.235	.526
StrategyQA	.390	**.265**	.301	.408
CLUTRR	**.290**	.370	.323	.633
SayCan	.267	**.214**	.269	.307
average	.241	**.232**	.235	.474
	LM: LLaMA-13B			
ASDiv	.144	**.135**	.136	.334
GSM8K	**.177**	.179	.181	.591
MultiArith	.232	**.198**	.222	.395
SVAMP	.205	**.180**	.200	.365
Sport	.170	.153	**.151**	.194
Date	**.181**	.206	.185	.402
StrategyQA	.383	**.241**	.285	.371
CLUTRR	**.298**	.353	.320	.596
SayCan	.209	**.190**	.220	.250
average	.222	**.204**	.211	.389
	LM: LLaMA-70B			
ASDiv	.107	**.099**	.101	.209
GSM8K	.201	**.168**	.187	.375
MultiArith	.134	**.108**	.113	.188
SVAMP	.145	**.112**	.129	.183
Sport	.053	**.041**	.044	.046
Date	.167	**.166**	.168	.262
StrategyQA	.338	**.192**	.239	.289
CLUTRR	**.287**	.347	.309	.534
SayCan	.206	**.156**	.191	.179
average	.182	**.154**	.165	.252
	LM: Mistral-7B			
ASDiv	**.122**	**.112**	.113	.269
GSM8K	.197	**.188**	.196	.491
MultiArith	.189	**.160**	.171	.320
SVAMP	.176	**.141**	.162	.250
Sport	.116	**.085**	.090	.109
Date	**.178**	.202	.187	.373
StrategyQA	.363	**.239**	.276	.343
CLUTRR	**.283**	.334	.307	.555
SayCan	.225	**.186**	.217	.207
average	.205	**.183**	.191	.324
	LM: Mistral-7B-instruct			
ASDiv	.130	.127	**.124**	.288
GSM8K	.193	**.191**	.194	.508
MultiArith	.191	**.164**	.180	.319
SVAMP	.166	**.147**	.158	.274
Sport	.207	.196	**.194**	.222
Date	**.220**	.244	.227	.497
StrategyQA	.334	**.269**	.284	.366
CLUTRR	**.306**	.397	.347	.648
SayCan	.228	**.209**	.229	.332
average	.220	.216	**.215**	.384

Table 5: Brier Score (↓) for open-source LMs on all datasets, averaged across five prompting strategies. The best scores are **in bold** (to be continued).

Domain	Consistency Metrics			Baselines
	entropy	agree	FSD	logit
	LM: olmo-7B			
ASDIV	.194	**.158**	.186	.463
CLUTRR	**.195**	.254	.220	.656
GSM8K	**.066**	.084	.069	.603
DATE	**.146**	.159	.154	.662
MULTIARITH	**.140**	.145	.148	.575
SAYCAN	.385	**.339**	.409	.503
SPORT	.413	**.221**	.294	.306
STRATEGYQA	.417	**.286**	.331	.367
SVAMP	.201	**.171**	.198	.493
average	.240	**.202**	.223	.514
	LM: olmo-7B-it			
ASDIV	.204	**.182**	.198	.436
CLUTRR	**.254**	.365	.307	.688
GSM8K	**.146**	.161	.153	.616
DATE	**.173**	.213	.188	.641
MULTIARITH	.311	**.246**	.301	.349
SAYCAN	.320	**.297**	.336	.514
SPORT	.283	**.217**	.232	.261
STRATEGYQA	.337	**.277**	.284	.321
SVAMP	.220	**.198**	.217	.473
average	.250	**.239**	.246	.478
	LM: olmo-7B-it-rl			
ASDIV	**.218**	.220	**.218**	.506
CLUTRR	**.311**	.440	.369	.742
GSM8K	**.159**	.191	.170	.673
DATE	**.185**	.248	.211	.684
MULTIARITH	.302	**.274**	.299	.449
SAYCAN	.330	**.322**	.342	.565
SPORT	.229	**.168**	.185	.198
STRATEGYQA	.327	.311	**.305**	.328
SVAMP	**.222**	.235	.231	.566
average	**.253**	.268	.259	.523

Table 6: (Continued) Brier Score (↓) for open-source LMs on all datasets, averaged across five prompting strategies. The best scores are **in bold**.

Domain	Consistency Metrics			Baselines				
	entropy	agreement	FSD	$\text{verb}_{\text{ling}}$	$\text{verb}_{\text{percent}}$	logit	$\text{ptrue}_{\text{0-shot}}$	$\text{ptrue}_{\text{8-shot}}$
	LM: Codex							
MWP	.132	**.077**	.104	.237	.225	.156	.142	.108
MHQA	.188	**.090**	.119	.272	.214	.152	n/a	.167
Plan.	.203	**.101**	.159	.322	.159	.106	.117	.248
Relation.	.228	.368	.294	.214	.536	.512	.313	**.175**
average	.169	**.117**	.136	.256	.249	.189	.166	.151
	LM: GPT-3.5-turbo							
MWP	**.118**	.121	.115	.208	.193	n/a	n/a	n/a
MHQA	**.230**	.246	.233	.321	.277	n/a	n/a	n/a
Plan.	.154	**.119**	.128	.351	.357	n/a	n/a	n/a
Relation.	**.426**	.505	.471	.449	.519	n/a	n/a	n/a
average	**.193**	.205	.195	.289	.275	n/a	n/a	n/a
	LM: GPT-4							
MWP	.056	.064	.055	**.053**	.079	n/a	n/a	n/a
MHQA	.104	**.089**	.090	.139	.126	n/a	n/a	n/a
Plan.	.109	**.061**	.084	.351	.484	n/a	n/a	n/a
Relation.	.387	.454	.415	**.381**	.435	n/a	n/a	n/a
average	.114	.115	**.110**	.151	.179	n/a	n/a	n/a

Table 7: ECE score (↓) for closed-source LMs on four domains – Math Word Problems (MWP), Multi-hop QA (MHQA), Planning (Plan.), and Relational Inference (Relation.) – averaged across five prompting strategies. The best score is **in bold**.

Domain	Consistency Metrics			Baselines
	entropy	agreement	FSD	logit
	LM: LLaMA-7B			
MWP	.138	**.130**	.141	.548
MHQA	.237	**.189**	.197	.400
Plan.	.256	**.164**	.259	.302
Relation.	**.214**	.359	.267	.641
average	.192	**.179**	.187	.482
	LM: LLaMA-13B			
MWP	.159	**.117**	.151	.428
MHQA	.207	**.142**	.149	.319
Plan.	.205	**.147**	.217	.244
Relation.	**.239**	.334	.268	.602
average	.189	**.153**	.170	.391
	LM: LLaMA-70B			
MWP	.153	**.083**	.120	.233
MHQA	.174	**.082**	.117	.193
Plan.	.229	**.156**	.207	.172
Relation.	**.223**	.333	.257	.539
average	.176	**.118**	.144	.247
	LM: Mistral-7B			
MWP	.165	**.111**	.138	.331
MHQA	.184	**.120**	.135	.271
Plan.	.260	**.194**	.231	.198
Relation.	**.207**	.295	.245	.561
average	.186	**.144**	.159	.322
	LM: Mistral-7B-instruct			
MWP	.154	**.108**	.127	.349
MHQA	.205	.191	**.186**	.355
Plan.	.203	**.155**	.212	.328
Relation.	**.238**	.402	.310	.659
average	.186	**.174**	.177	.383

Table 8: ECE score (↓) for open-source LMs on four domains – Math Word Problems (MWP), Multi-hop QA (MHQA), Planning (Plan.), and Relational Inference (Relation.) – averaged across five prompting strategies. The best score is **in bold** (to be continued).

Domain	Consistency Metrics			Baselines
	entropy	agree	FSD	logit
	LM: olmo-7B			
MWP	.110	**.091**	.121	.609
MHQA	.288	**.137**	.209	.446
Plan.	.376	**.306**	.405	.513
Relation.	**.095**	.231	.163	.686
average	.197	**.146**	.186	.553
	LM: olmo-7B-it			
MWP	.225	**.156**	.211	.491
MHQA	.226	.188	**.176**	.402
Plan.	.296	**.277**	.320	.523
Relation.	**.183**	.375	.275	.710
average	.228	**.204**	.219	.489
	LM: olmo-7B-it-rl			
MWP	.219	**.177**	.207	.574
MHQA	.210	.209	**.184**	.396
Plan.	.302	**.283**	.313	.581
Relation.	**.285**	.480	.368	.760
average	.233	.233	**.229**	.536

Table 9: (Continued) ECE score (↓) for open-source LMs on four domains – Math Word Problems (MWP), Multi-hop QA (MHQA), Planning (Plan.), and Relational Inference (Relation.) – averaged across five prompting strategies. The best score is **in bold**.