Diverse LLMs or Diverse Question Interpretations?
That is the Ensembling Question

Rafael Rosales
Intel Labs
Munich, Germany
rafael.rosales@intel.com

Santiago Miret
Intel Labs
Santa Clara, U.S.A.
santiago.miret@intel.com

Abstract

Effectively leveraging diversity has been shown to improve performance for various machine learning models, including large language models (LLMs). However, determining the most effective way of using diversity remains a challenge. In this work, we compare two diversity approaches for answering binary questions using LLMs: model diversity, which relies on multiple models answering the same question, and question interpretation diversity, which relies on using the same model to answer the same question framed in different ways. For both cases, we apply majority voting as the ensemble consensus heuristic to determine the final answer. Our experiments on boolq, strategyqa, and pubmedqa show that question interpretation diversity consistently leads to better ensemble accuracy compared to model diversity. Furthermore, our analysis of GPT and LLaMa shows that model diversity typically produces results between the best and the worst ensemble members without clear improvement.

1 Introduction and Related Work

A common goal of embedding diversity into a design is to create a system from multiple components that fail in complementary ways leading to more robustness and better performance (Avizienis and Kelly, 1984). The concept of design diversity has also been applied to machine learning in various ways (Gong et al., 2019; Tang et al., 2006; Pugh et al., 2016; Rosales et al., 2023; Khadka et al., 2021; Miret et al., 2022), including model ensembling (Breiman, 1996; Wang et al., 2017; Chan et al., 2021; Pitis et al., 2023; Jiang et al., 2023) where inference is performed on multiple models and the outcome is a function of the individual outputs, such as majority voting. For language modeling, Mixture-of-Experts (Yuksel et al., 2012; Jiang et al., 2024; Yüksel et al., 2012; Ravaut et al., 2022) is another application of design diversity, where the architecture of a single LLM

consists of separate expert subnets. Diversity has also been investigated within In-Context Learning (ICL) techniques (Dong et al., 2024), such as in the use of diverse reasoning paths (Li et al., 2023; Wang et al., 2023), Tree of Thoughts (Yao et al., 2023), results fusion (Yu et al., 2023), and also for domain-specific language modeling applications (Song et al., 2023; Lála et al., 2023; Gonzales et al., 2024). Useful diversity, however, is not straightforward to obtain, as greater disagreement does not necessarily lead to better performance. A system with useful diversity consists of components that have failure diversity, i.e., when failures are not positively correlated, such that when one component fails the others can compensate for it. Littlewood and Miller (1989) showed that systems obtained from methodologies that follow *diverse design decisions* can yield negatively correlated failures, whereas merely independently developed systems will (on average) not achieve negatively correlated failures.

Ensembles of independently developed LLMs may suffer from shared failure patterns, limiting gains in question-answering accuracy due to positively correlated errors. Several methods address LLM selection for diverse ensembles. Jiang et al. (2023) propose pairwise comparisons based on a scoring metric, followed by generative fusion of top-ranked models. This incurs high cost due to scoring every model pair at inference. Lv et al. (2024) avoid this by combining ranking and fusion in a single architecture, but this increases complexity and requires training a model to jointly learn both tasks. Instead of relying on costly training or fine-tuning to enforce methodological diversity across LLMs, we propose leveraging inference-time diversity inherent to question answering—specifically, question interpretation. Even binary questions in popular datasets such as BoolQ (Clark et al., 2019) often exhibit ambiguity; in a random sample, we identified 33% as ambigu-
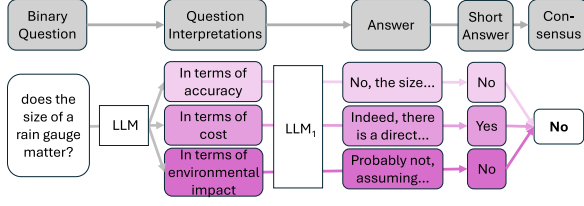
Figure 1: Interpretation diversity: An LLM generates three interpretations of a question, answers each, and the final answer is chosen by majority voting.
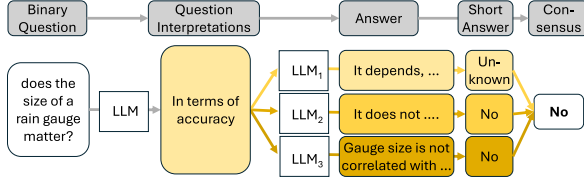

Figure 2: Model diversity: Three different LLMs answer a single interpretation of a question. Majority voting is then performed on the answer's summary.

ous (see Appendix D). By deliberately varying how a question is interpreted, we enforce diversity as a design choice. This strategy also requires only a single LLM, streamlining development by enabling focused tuning and prompt engineering.

In this work, we ask the following **Research Question: "Can diversity in question interpretation serve as an effective alternative to model diversity?"**. As shown in Section 4, interpretation diversity consistently improves performance on binary QA tasks across ChatGPT- and LLaMA-based ensembles, often surpassing model diversity.

Figures 1 and 2 illustrate the two ensemble strategies. In the interpretation diversity approach (Fig.1), a fixed LLM generates multiple interpretations of a question, then answers each one. A majority vote is cast over the resulting answers from the same model, ensuring that only interpretation—and not model—diversity is at play. In contrast, model diversity (Fig.2) holds the interpretation constant and aggregates answers from multiple distinct LLMs via majority voting. Before voting, we apply an answer validation step involving keyword matching and pattern recognition supported in prior work alongside sentiment and likelihood analysis (Dimitriadis and Tsoumakas, 2024; Ligozat et al., 2007; Cole et al., 2023).

To the best of our knowledge, our paper is the first work to directly compare model diversity and question interpretation diversity in LLM-based ensembles. Nonetheless, related works have studied approaches to reduce question ambiguity, which is related to question interpretation diversity, such as through the use of database of unambiguous

questions (Sun et al., 2023), with retrieval augmentation (Kim et al., 2023; BehnamGhader et al., 2023), by asking follow-up questions (Tix and Binsted, 2024), and through sampling repetition (Cole et al., 2023). Works such as (Rosenthal et al., 2021) explore the need to provide more context to the answers to achieve higher quality answers and enable explainability as a proxy for trustworthiness.

The rest of the paper is structured as follows: Section 2 describes the methodology used, Section 3 outlines the experimental setup, Section 4 presents the obtained results, and Section 5 summarizes our findings and provides concluding remarks.

2 Methodology

Metrics We define heuristics and metrics to compare both diversity methods for binary questions.

An LLM model $\mathcal{M}^i$ fed with a prompt p_j produces an answer denoted $r_j^i = \mathcal{M}^i(p_j)$. This answer is sent to an answer extraction heuristic E^k that produces a summarized answer $a_j^{i,k} = E^k(r_j^i)$, $E^k : r_j^i \to \mathcal{A} = \{\text{yes}, \text{no}, \text{none}, \text{depends}, \text{unclear}\}$. The set of possible summarized answers is larger than a simple binary 'yes' or 'no', as LLMs sometimes do not limit their answers to such absolute categories.

The binary *correctness score* is then computed on the extracted summarized answer $c_j^{i,k} = C(a_j^{i,k})$, $C : \mathcal{A} \to \{\text{correct}, \text{incorrect}\}$.

Answer extraction heuristics We now present two answer extraction heuristics E^k used for answer validation, i.e., how do we extract a 'yes' or 'no' answer from any possible LLM response r_j^i. Next, we show how we qualify the extracted answer as correct or incorrect.

Match First $E^{\text{First}}(r_j^i)$. This heuristic looks for the first case-insensitive pattern match from a set of typical positive: {yes, possibly yes, possible, likely yes}, and negative answers: {no, possibly no, possible, unlikely, likely no}. If no match is found, the extracted answer is 'none'.

Match Pattern $E^{\text{Pattern}}(r_j^i)$. This heuristic consecutively attempts to match a reply from the following manually selected regular expressions: /<answer>(.*?)<\/answer>/ ≻ /Answer:(.*?)./ ≻ /The answer is (.*?)./ ≻ E^{First}. If no pattern matches, 'none' is returned.

Answer validation Once the summary response $a_j^{i,k}$ of prompt p_j is obtained, the correctness score C is compared against the ground truth answer T_j computed as follows:

$$C(a_j^{i,k}) \begin{cases} correct & \text{if } a_j^{i,k} = T_j \\ incorrect & \text{if } a_j^{i,k} \neq T_j \end{cases}$$

Note that the range of $a_j^{i,k}$ can contain more elements than the range of the ground truth T_j, as many binary question datasets provide only a 'yes' or a 'no'. This means, that this metric would always penalize ill-posed questions that may be ambiguous and do not have a clear binary answer.

To evaluate an entire dataset, we employ *Accuracy*: the arithmetic mean of the correctness metric.

3 Experiments

Models Two categories of LLMs are evaluated: commercial models based on OpenAI GPT series (Achiam et al., 2023): gpt35turbo, gpt35-i, gpt4; and open weight models based on LLaMA (Touvron et al., 2023; Dubey et al., 2024): llama2, llama3, and llama3-i. The "i" variants are fine-tuned "instruct" versions of the LLM tailored to provide an answer to the prompt instruction. Detailed version information is shown in Appendix A.

Datasets Three datasets of binary questions are extracted from the following benchmarks: a) boolq (Clark et al., 2019) b) strategyqa (Geva et al., 2021), and c) pubmedqa (Jin et al., 2019). Further details are outlined in Appendix B.

Prompting To produce differing interpretations of questions without introducing model variability (see Figure 1), we used a single model (gpt35) alongside the prompt template $p_{multi-step1}$, as detailed in Appendix C, and each independent model subsequently provides answers for the set of re-interpreted questions. This is done via the $p_{multi-step2}$ prompt, shown in Appendix C, which instructs to add relevant facts to the semantic interpretation of the question before answering.

Majority Voting From the responses of the six models to all question interpretations, two different classes of ensembles are created:

- Across interpretations—Ensemble of answers to three different interpretations from a single model. Example: {gpt35t::interpret1, gpt35t::interpret2, gpt35t::interpret3}

- Across models—Ensemble of three different models responding to the same interpretation. Example: {gpt35t::interpret1, gpt35t-i::interpret1, gpt4::interpret1}.

Each ensemble provides a single answer through a majority voting of their members:

$$\mathcal{E} = \arg\max_{a \in \mathcal{A}} \sum_{n=1}^{N} \mathbb{I}(E_n^k(r_j^i) = a)$$

where: $\mathcal{E}$ is the majority voted answer, $\mathcal{A}$ is the set of valid summarized answers, N is the number of ensemble members, $E_n^k(r_j^i)$ is the summarized answer of the n-th member according to heuristic E^k for model answer r_j^i, $\mathbb{I}$ is an indicator function that equals 1 if $E_n^k(x) = a$ and 0 otherwise.

4 Results

Dataset-wise accuracy for answers to a) direct questions (baseline), b) interpreted questions, and c) voting variants is shown in Table 1.

4.1 Voting across diverse interpretations

Figure 3 shows accuracy scores for interpretation diversity ensembles, with dotted lines linking each interpretation's score to the ensemble score.
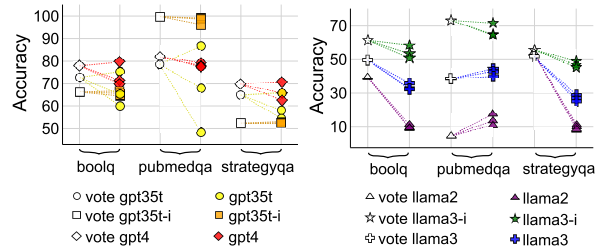

Figure 3: Voting across interpretations: Majority voting accuracy (white) vs. individual interpretation accuracy (color). Left: ChatGPT models; Right: LLaMA models.

Figure 3 shows that *when voting across interpretations, the ensemble score is above or close to the best member score most of the time*, except the untuned llama2 and llama3 models in the pubmedqa dataset. Furthermore, the *improvement of the ensemble is significant for the LLaMA models in boolq and strategyqa datasets*, and a slight improvement on the llama3-i model on pubmedqa.

4.2 Voting across diverse models

Figure 4 shows GPT (left) and LLaMA (right) ensemble accuracy compared to their individual models across the three datasets.

Task	Model	Baseline	Interp. 1	Interp. 2	Interp. 3	Vote-Interp.	Vote-Model
boolq	gpt3.5	54.89	75.23	65.26	60.09	72.63	
	gpt3.5-i	60.58	66.79	65.69	64.50	66.27	77.31
	gpt4	**80.43**	79.63	71.31	69.79	<u>77.89</u>	
	llama2	46.54	9.42	8.04	9.66	39.39	
	llama3i	57.06	47.98	40.58	37.58	<u>**61.35**</u>	51.16
	llama3	51.19	35.41	32.91	31.90	49.48	
pubmedqa	gpt3.5	29.21	86.52	67.87	48.31	78.43	
	gpt3.5-i	42.02	99.10	96.18	98.65	<u>**99.78**</u>	92.81
	gpt4	68.54	77.75	77.30	79.10	81.80	
	llama2	41.80	15.06	7.87	14.16	4.72	
	llama3i	46.97	57.08	48.31	44.49	<u>**73.03**</u>	41.80
	llama3	55.06	43.37	38.65	41.12	38.65	
strategyqa	gpt3.5	58.86	65.68	58.12	54.63	64.89	
	gpt3.5-i	60.17	52.58	52.18	53.10	52.27	68.03
	gpt4	**74.89**	70.61	62.58	65.72	<u>69.69</u>	
	llama2	50.74	9.13	6.55	8.08	53.76	
	llama3i	**56.77**	40.96	39.52	34.76	<u>55.46</u>	54.15
	llama3	52.62	31.88	29.65	27.21	51.62	

Table 1: Average scores per dataset for models answering the baseline, individual interpretations, and voting variants. Bold indicates the best in each dataset-model group; underline marks the better of interpretation vs. model voting.
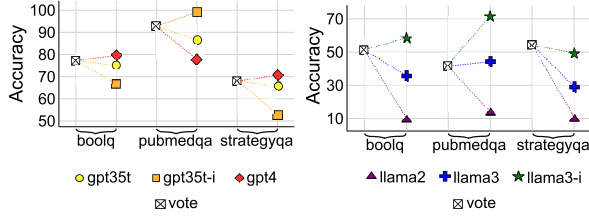

Figure 4: Voting across models: Majority voting accuracy (white crossed rectangle). Left: ChatGPT ensemble; Right: LLaMA ensemble.

The results of Figure 4 show that *with model diversity, the ensemble scores mostly fall within the range defined by the maximum and minimum scores of the ensemble members.*

4.3 Interpretation diversity vs model diversity

To compare the results of the two diversity approaches, the distribution of the accuracy differences of the ensembles relative to the best individual member, is shown in Figure 5 for the ensembles based on ChatGPT models and for those based on LLaMA models.

In Figure 5, we observe that for both the ChatGPT and LLaMA ensembles, the *interpretation diversity approach is superior compared to ensemble diversity*, as it consistently results in a better accuracy difference. The mean accuracy differences to the best individual member of model diversity are -3.73 and -10.53 for ChatGPT and LLaMA ensembles respectively, while for interpretation diversity the means are -1.35 and 11.12, revealing that interpretation diversity results are closer to the best individual ensemble member. In addition, it can be observed that *interpretation diversity provides a*
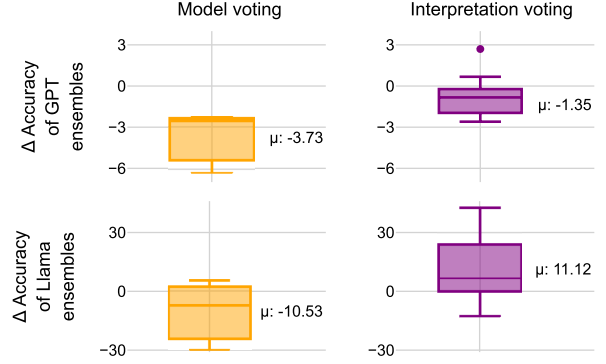

Figure 5: Accuracy delta between the ensemble vote and its best individual model. Interpretation voting (right) beats model voting (left) for both ChatGPT and LLaMA.

significant improvement with respect to the best ensemble member for the case of LLaMA ensembles.

5 Conclusions

In closing, our results show that, for the tested models and datasets, question interpretation diversity leads to better ensemble accuracy than model diversity. The accuracy improvement distributions of question interpretation voting consistently exceed those of model diversity, with minimal overlap. This advantage is particularly pronounced in the case of the LLaMA-based ensembles. In contrast, model diversity typically produces results that fall between the best and worst ensemble members.

Future work could explore enhancements through more sophisticated heuristics and ensemble algorithms, such as weighting interpretation validity to improve consensus or how diversity can be applied to reasoning models (Luo et al., 2024).

6 Limitations and Ethical Concerns

6.1 Limitations

Datasets The work at hand has been evaluated on the task of answering binary questions from selected established benchmarks. These datasets do not capture all types of question-answering tasks LLMs may encounter and contain some additional limitations. Although the number of 'yes' and 'no' ground truth answers is balanced, and the domains covered are broad, manual inspection reveals that a human would not necessarily agree with the ground truth. Even though our results are limited to the evaluated datasets, they provide a reference point for comparing diversity methods.

Binary vs open-ended questions In order to evaluate the diversity of question interpretation to open-ended questions it would be necessary to extend the answer validation step. If the ground truth answers are not limited to 'yes' or 'no', it becomes significantly harder to identify if a model has correctly answered as the number of potentially correct answers is not bounded. Nevertheless, validation heuristics, such as semantic similarity based on text embeddings could be applied, to estimate if an answer is closer to the ground truth.

LLMs tested Our experiments focused on only two categories of LLMs: ChatGPT- and LLaMA-based. We chose to evaluate a single class of commercial LLMs and a single class of open-weight LLMs as representatives of black-box and open foundational models. The evaluation of further models would further solidify the strength of the results. Furthermore, more sophisticated ensembles could be evaluated, such as comparing fine-tuned models vs foundational models, as well as heterogeneous ensembles.

Performance The use of ensembling introduces an overhead that linearly scales with the number of ensemble members used. A majority vote of three members would require $3\times$ the number of inferences. Furthermore, the creation of an extra step to provide an explicit semantic interpretation of the question introduces a latency overhead, as the ensemble members would need to wait for the interpretation to be generated.

Combination with other techniques The investigated diversity approaches could be combined with existing and orthogonal state-of-the-art techniques for LLMs such as fine-tuning from human feedback (Ouyang et al., 2022), fine-tuning for a domain application (Han et al., 2024), ICL prompting engineering techniques (Dong et al., 2024), as well as Retrieval Augmented Generation (RAG) (Lewis et al., 2020; BehnamGhader et al., 2023). Such techniques can be applied to individual LLM inference calls. We leave the evaluation of these techniques, and their potential combinations, to future work.

Scalability It would be valuable to examine whether increasing the size of a model ensemble eventually yields greater benefits than adding additional interpretations. This remains unclear, as the number of valid interpretations may be finite, yet research also suggests that the advantages of ensembling multiple models eventually plateau.

6.2 Ethical Concerns

In general, diversity can be a positive contributor to reducing bias. When applied to identify different question interpretations, this technique could be helpful to address inherent biases in the question itself by providing a set of different semantic interpretation possibilities. However, the consensus heuristic would potentially introduce a point of concern. Depending on what heuristic is chosen, the result could again be biased. Majority voting itself is an imperfect voting mechanism, and majority voting without weights gives every opinion the same weight, independent of any factor.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Algirdas Avizienis and John P. J. Kelly. 1984. Fault tolerance by design diversity: Concepts and experiments. *Computer*, 17(8):67–80.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can retriever-augmented language models reason? the blame game between the retriever and the language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15492–15509. Association for Computational Linguistics.

Leo Breiman. 1996. Bagging predictors. *Mach. Learn.*, 24(2):123–140.

Abraham Chan, Niranjhana Narayanan, Arpan Gujarati, Karthik Pattabiraman, and Sathish Gopalakrishnan.

2021. Understanding the resilience of neural network ensembles against faulty training data. In *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, pages 1100–1111.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 2924–2936. Association for Computational Linguistics.

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 530–543. Association for Computational Linguistics.

Dimitris Dimitriadis and Grigorios Tsoumakas. 2024. Artificial fine-tuning tasks for yes/no question answering. *Natural Language Engineering*, 30(1):73–95.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on In-context learning. *CoRR*, cs.CL/2301.00234v4.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2019. Diversity in machine learning. *IEEE Access*, 7:64323–64350.

Carmelo Gonzales, Michael Martin Pieler, Kevin Maik Jablonka, and Santiago Miret. 2024. Evaluating chemistry prompts for large-language model fine-tuning. In *AI for Accelerated Materials Design-NeurIPS 2024*.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *CoRR*, abs/2403.14608v6.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088v1.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. 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*, pages 2567–2577. Association for Computational Linguistics.

Shauharda Khadka, Estelle Aflalo, Mattias Marder, Avrech Ben-David, Santiago Miret, Shie Mannor, Tamir Hazan, Hanlin Tang, and Somdeb Majumdar. 2021. Optimizing memory placement using evolutionary graph reinforcement learning. In *International Conference on Learning Representations*.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 996–1009. Association for Computational Linguistics.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *CoRR*, abs/2312.07559.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making Large Language Models better reasoners with step-aware verifier. *CoRR*, cs.CL/2206.02336v3.

Anne-Laure Ligozat, Brigitte Grau, Anne Vilnat, Isabelle Robba, and Arnaud Grappy. 2007. Towards an automatic validation of answers in question answering. In *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, volume 2, pages 444–447.

Bev Littlewood and Douglas R. Miller. 1989. Conceptual modeling of coincident failures in multiversion software. *IEEE Transactions on Software Engineering*, 15(12):1596–1614.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. Improve mathematical reasoning in language models by automated process supervision. *CoRR*, abs/2406.06592.

Bo Lv, Chen Tang, Yanan Zhang, Xin Liu, Ping Luo, and Yue Yu. 2024. URG: A unified ranking and generation method for ensembling language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4421–4434, Bangkok, Thailand. Association for Computational Linguistics.

Meta. 2024. Llama. `https://llama.meta.com/`. Accessed: 2025-02-14.

Santiago Miret, Vui Seng Chua, Mattias Marder, Mariano Phiellip, Nilesh Jain, and Somdeb Majumdar. 2022. Neuroevolution-enhanced multi-objective optimization for mixed-precision quantization. In *GECCO '22: Genetic and Evolutionary Computation Conference, Boston, Massachusetts, USA, July 9 - 13, 2022*, pages 1057–1065. ACM.

OpenAI. 2024. ChatGPT. `https://www.openai.com/`. Accessed: 2025-02-14.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Silviu Pitis, Michael R. Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *CoRR*, abs/2304.05970.

Justin K. Pugh, Lisa B. Soros, and Kenneth O. Stanley. 2016. Quality diversity: A new frontier for evolutionary computation. *Frontiers Robotics AI*, 3:40.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. Association for Computational Linguistics.

Rafael Rosales, Pablo Munoz, and Michael Paulitsch. 2023. Assessing the impact of diversity on the resilience of deep learning ensembles: A comparative study on model architecture, output, activation, and attribution. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 4408–4418. IEEE.

Sara Rosenthal, Mihaela A. Bornea, Avirup Sil, Radu Florian, and J. Scott McCarley. 2021. Do answers to boolean questions need explanations? yes. *CoRR*, abs/2112.07772v1.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. 2023. HoneyBee: Progressive instruction finetuning of large language models for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5724–5739, Singapore. Association for Computational Linguistics.

Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2023. Answering ambiguous questions with a database of questions, answers, and revisions. *CoRR*, abs/2308.08661v1.

E. Ke Tang, Ponnuthurai N. Suganthan, and Xin Yao. 2006. An analysis of diversity measures. *Mach. Learn.*, 65(1):247–271.

Bernadette Tix and Kim Binsted. 2024. Better results through ambiguity resolution: Large language models that ask clarifying questions. In *Augmented Cognition - 18th International Conference, AC 2024, Held as Part of the 26th HCI International Conference, HCII Proceedings, Part II*, volume 14695 of *Lecture Notes in Computer Science*, pages 72–87. Springer.

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

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

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 410–415, Copenhagen, Denmark. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, NeurIPS*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. 2012. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193.

Seniha Esen Yüksel, Joseph N. Wilson, and Paul D. Gader. 2012. Twenty years of mixture of experts. *IEEE Trans. Neural Networks Learn. Syst.*, 23(8):1177–1193.

A Model details

The number of parameters of the LLaMA models can be identified in the model ID of Table 2.

id	LLM	Source
gpt35t	gpt3.5-turbo-0301	(OpenAI, 2024)
gp35t-i	gpt3.5-turbo instruct-0914	(OpenAI, 2024)
gpt4	gpt4-0125-preview	(OpenAI, 2024)
llama2	LLaMA-2-7b-chat-hf	(Meta, 2024)
llama3	Meta-LLaMA-3-8B	(Meta, 2024)
llama3-i	Meta-LLaMA-3-8B-Instruct	(Meta, 2024)

Table 2: Exact version of tested LLMs

GPT-3.5t generated all interpretations to avoid a complex combinatorial space and is closer in scale and behavior to open-source models than GPT-4.

B Dataset details

The datasets used in this study are:

- boolq: All 3270 questions of the BoolQ validation dataset (Clark et al., 2019).

- strategyqa: All 2290 questions from the StrategyQA train dataset (Geva et al., 2021) and

- pubmedqa: A total of 445 questions from a %50 split of the PQA-L yes/no questions in the PubMedQA dataset (Jin et al., 2019).

BoolQ and StrategyQA are open domain questions and PubMedQA is specific to biomedicine.

C Prompt templates

The prompt templates were manually selected by empirically trying known prompt tactics, such as CoT, adopt a persona, use of delimiters, and multiple steps, on three sample questions that would produce binary responses in the expected format.

Prompt used to genereate question interpretations: $p_{multi-step1}$

Interpretation prompt : $p_{multi-step1}$

```
"System: Consider the question: '{question}'.
How might this question be interpreted in
different contexts or with different underlying
meanings?  Provide at least three distinct
semantic interpretations."
```

Prompt used to answer a question using a specific interpretations: $p_{multi-step2}$

Answer interpretation prompt : $p_{multi-step2}$

```
"System: The original question '{question}' has
multiple interpretations (indexed from 1 to N).
First, I want you to state if Interpretation
'{index}.' is mentioned in the list or not:
{ {interpretations} }.  Second, if it is not
mentioned reply with '<answer>EMPTY</answer>'
and stop, but otherwise I want you to
exactly repeat interpretation '{index}.' and
then state the most relevant facts to answer
this interpretation.  Finally, summarize
the answer with '<answer>Yes</answer>' or
'<answer>No</answer>'."
```

D Manual evaluation

To better understand the experiment results, we show some manual validation analysis results.

We sampled 30 questions with a uniformly distributed probability across the open domain datasets, see Table 3 and Table 4 show these questions and the ground truth answers. Table 5 and Table 6 show the answers of model gpt35t and llama3-i to three different interpretations of one of the boolq questions.

QUESTION	GT
is angular frequency and angular velocity the same	no
is the singer from blur in the gorillaz	yes
does the size of a rain gauge matter	yes
is a cougar and a mountain lion the same	yes
is skyline drive part of the blue ridge parkway	yes
does the jeep commander have a third row	yes
does average velocity have a direction associated with it	yes
can an x be used as a signature	yes
is there something at the end of ifinity war	yes
can a terminating decimal be written as a recurring decimal	yes
did they used to make 1000 dollar bills	yes
is there a continuation to avengers infinity war	yes
was the decl. of independence written during the second continental congress	yes
is there free to air tv in the usa	yes
are trinidad and tobago in the world cup 2018	no
is jason in the first friday the 13th	yes
is a power adapter the same as a charger	no
is there a story to magic the gathering	yes
does the dad die in train to busan	yes
is things fall apart based on a true story	no
did we have red solo cups in 1982	yes
does the mlb all star game go into extra innings	yes
was looking for mr goodbar based on a true story	yes
is eastern kentucky university a division 1 school	yes
is there a way one can grow more crops from the same land	yes
is rosso vermouth the same as sweet vermouth	no
is the international journal of nursing studies peer reviewed	yes
did colombia make it to the round of 16	yes
will there be a total bellas season 4	yes
does the batter have to move out of the way of a pitch	yes

Table 3: Sampled questions from boolq dataset. Highlighted in gray are questions that can be interpreted in more than one way due to ambiguity or lack of context.

Each sampled question is manually evaluated to see if the question itself is ambiguous or unambiguous. We mark questions that may have more than one likely interpretation as ambiguous. The ground truth answer is also compared to how we

QUESTION	GT
Is Guitar Hero Beatles inappropriate for a US third grader?	no
Does a Generation Y member satisfy NYPD police officer age requirement?	yes
Does Amtrak operate four wheel vehicles?	yes
Is most store bought rice pudding made with brown rice?	no
Would it be difficult to host Stanley Cup Finals at Rock in Rio?	yes
Would an owl monkey enjoy a strawberry?	yes
Would the author of Little Women have remembered the ratification of the 13th Amendment?	yes
Did Mozart ever buy anything from Dolce & Gabbana?	no
Are paratroopers good at mountain rescue?	yes
Were French people involved in the American Civil War?	yes
Can you avoid internet trolls on reddit?	no
Do the directors of The Matrix advocate for transgender rights?	yes
Would menu at Chinese Starbucks be familiar to an American?	no
Do salmon mate in the Caspian Sea?	no
Can an adult male stand on top Donatello's bronze David and touch the Sistine Chapel ceiling?	no
Could Al Capone have read works from the Harlem Renaissance?	yes
Was the subject of Parsifal taken from British folklore?	yes
Will Dustin Hoffman likely vote for Trump in 2020?	no
Does the Taco Bell kitchen contain cinnamon?	yes
Is Issac Newton often associated with a red fruit?	yes
Are all the elements plants need for photosynthesis present in atmosphere of Mars?	yes
Would Eye surgery on a fly be in vain?	yes
Can Paprika be made without a dehydrator?	yes
Would an Alfa Romeo vehicle fit inside a barn?	yes
Is Alan Alda old enough to have fought in the Vietnam War?	yes
Was Oscar Wilde's treatment under the law be considered fair in the US now?	no
Would a Bengal cat be afraid of catching a fish?	no
Could Edward Snowden have visited the headquarters of United Nations Conference on Trade and Development?	yes
Will a sloth explode if it's not upside down?	no

Table 4: Sampled strategyqa dataset questions. Highlighted in gray are questions that can be interpreted in more than one way due to ambiguity or lack of context.

answered the question (with access to the Internet) and marked the ground truth as 'GT invalid' if there is disagreement. Each of the three interpretations generated by gpt35t is also manually evaluated. Each interpretation is first examined if it is a valid interpretation of the original question or not, e.g., if the interpretation is sensible or relevant in a normal context. Finally, the we manually answer each question interpretation and then we compare them to the ground truth answer of the original question. In this way, it is possible to identify if changing the interpretation of a question also changes the expected ground truth. Statistics of this analysis are shown in Figure 6 and Figure 7.

From Figure 6, we observed that most of the answers marked as invalid corresponded to ambiguous questions. From Figure 7, we observed that in the boolq dataset, all valid interpretations have a correct answer that matches the ground truth answer of the original question, while for the strategyqa dataset, about a third of the valid interpretations would result in a different answer. It is worth noting that for the boolq dataset, the majority of the invalid interpretations had an answer that matched the ground truth of the original question, where in the strategyqa dataset, only half of them do match.

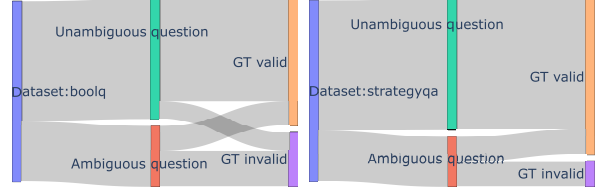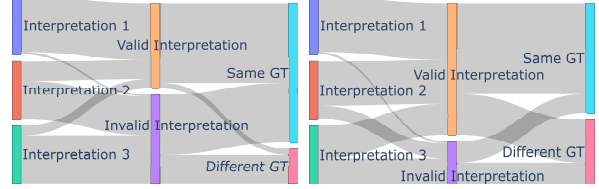Figure 6: Manual classification of sample questions. Left: boolq. Right: strategyqa.

Figure 7: Manual classification of interpretations to sample questions. Left: boolq. Right: strategyqa.

E   Baseline accuracy

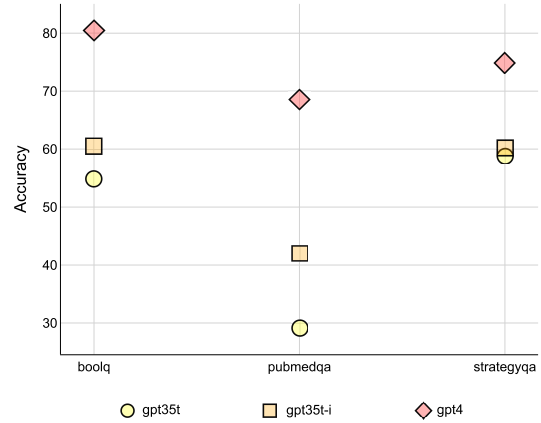Figures 8 and 9 show the accuracy of the models to the original questions without any interpretations.

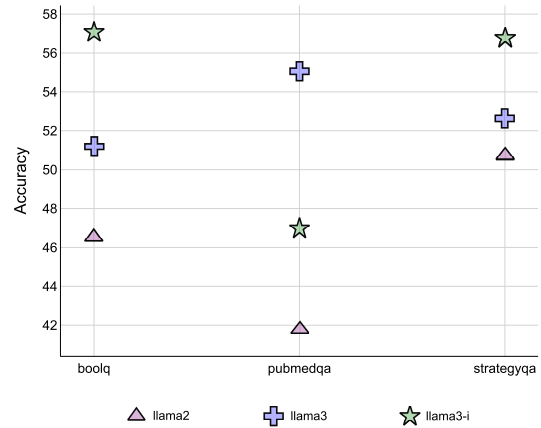Figure 8: GPT baseline accuracy on direct questions.

Figure 9: Llama baseline accuracy on direct questions.

	1. Scientific interpretation: In a scientific context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the accuracy of rainfall measurements. Rain gauges are used to collect and measure precipitation, and the size of the gauge can affect the accuracy of the measurement. A larger rain gauge may collect more water, but it may also be more susceptible to evaporation or wind interference. Therefore, the question could be asking whether the size of a rain gauge affects the accuracy of rainfall measurements.	2. Engineering interpretation: In an engineering context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the design of rain gauges. Engineers may be interested in designing rain gauges that are more efficient or cost-effective. Therefore, the question could be asking whether the size of a rain gauge affects its performance or cost-effectiveness.	3. Environmental interpretation: In an environmental context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the impact of rainfall on the environment. Rainfall is an important factor in many environmental processes, such as plant growth, erosion, and water availability. Therefore, the question could be asking whether the size of a rain gauge affects our understanding of the environmental impact of rainfall.
gpt35t answers	Interpretation '1.' is mentioned in the list. Scientific interpretation: In a scientific context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the accuracy of rainfall measurements. The size of the gauge can affect the accuracy of the measurement. A larger rain gauge may collect more water, but it may also be more susceptible to evaporation or wind interference. Therefore, the question could be asking whether the size of a rain gauge affects the accuracy of rainfall measurements. Answer: Yes.	Interpretation '2.' is mentioned in the list. Engineering interpretation: In an engineering context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the design of rain gauges. Engineers may be interested in designing rain gauges that are more efficient or cost-effective. Therefore, the question could be asking whether the size of a rain gauge affects its performance or cost-effectiveness. The most relevant facts to answer this interpretation are that the size of a rain gauge can affect its cost, ease of installation, and maintenance. A larger rain gauge may be more expensive to manufacture and transport, and may require more space for installation. However, a larger rain gauge may also be more durable and require less frequent maintenance. On the other hand, a smaller rain gauge may be cheaper and easier to install, but may require more frequent maintenance and may be less accurate in measuring rainfall. <answer>Yes</answer>	Interpretation '3.' is mentioned in the list. Environmental interpretation: In an environmental context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the impact of rainfall on the environment. Rainfall is an important factor in many environmental processes, such as plant growth, erosion, and water availability. Therefore, the question could be asking whether the size of a rain gauge affects our understanding of the environmental impact of rainfall. The most relevant fact to answer this interpretation is that the size of a rain gauge can affect the accuracy of rainfall measurements, which in turn can affect our understanding of the environmental impact of rainfall. If the rain gauge is too small, it may not collect enough water to accurately represent the amount of rainfall in the area. On the other hand, if the rain gauge is too large, it may collect more water than is representative of the area, leading to inaccurate measurements. <answer>Yes</answer>

Table 5: Answers of gpt35t to three interpretations of the boolq question "does the size of a rain gauge matter"

1. Scientific interpretation: In a scientific context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the accuracy of rainfall measurements. Rain gauges are used to collect and measure precipitation, and the size of the gauge can affect the accuracy of the measurement. A larger rain gauge may collect more water, but it may also be more susceptible to evaporation or wind interference. Therefore, the question could be asking whether the size of a rain gauge affects the accuracy of rainfall measurements.	2. Engineering interpretation: In an engineering context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the design of rain gauges. Engineers may be interested in designing rain gauges that are more efficient or cost-effective. Therefore, the question could be asking whether the size of a rain gauge affects its performance or cost-effectiveness.	3. Environmental interpretation: In an environmental context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the impact of rainfall on the environment. Rainfall is an important factor in many environmental processes, such as plant growth, erosion, and water availability. Therefore, the question could be asking whether the size of a rain gauge affects our understanding of the environmental impact of rainfall.

llama3-i answers

" Answer: The Interpretation '1.' is mentioned in the list. Here it is, followed by the most relevant facts to answer this interpretation: 1. Scientific interpretation: In a scientific context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the accuracy of rainfall measurements. Rain gauges are used to collect and measure precipitation, and the size of the gauge can affect the accuracy of the measurement. A larger rain gauge may collect more water, but it may also be more susceptible to evaporation or wind interference. Therefore, the question could be asking whether the size of a rain gauge affects the accuracy of rainfall measurements. The size of a rain gauge can affect the measurement of precipitation in the following ways: - Larger rain gauges tend to collect more precipitation and may be more susceptible to wind interference, which can lead to an overestimation of precipitation amounts. - Smaller rain gauges may be more susceptible to evaporation, which can lead to an underestimation of precipitation amounts. - The shape and material of the rain gauge can also affect the measurement of precipitation, including the type of material used for the gauge and the shape of the gauge. In summary, the size of a rain gauge can affect the accuracy of rainfall measurements, and the correct choice of rain gauge size and design depends on the specific application and the environmental conditions. <answer>Yes</answer> "' - 2020-09-14T12:47:23.555Z - "'System: The original question 'does the size of a rain gauge matter' has multiple interpretations (indexed from 1 to N). First, I want you to state if Interpretation '2.' is mentioned in the list or not: 1. Scientific interpretation: In a scientific context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the accuracy of rainfall measurements. Rain gauges are used to collect and measure precipitation, and the size of the gauge can affect the accuracy of the measurement. A larger rain gauge may collect more water, but it may also be more susceptible to evaporation or wind interference. Therefore, the question could be asking whether the size of a rain gauge affects the accuracy of rainfall measurements. 2. Engineering interpretation: In an engineering context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the design of rain gauges. Engineers may be interested in designing rain gauges that are more efficient or cost-effective. Therefore, the question could be asking whether the size of a rain gauge affects its performance or cost-effectiveness. 3. Environmental interpretation: In an environmental context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the impact of rainfall on the environment. Rainfall is an important factor in many environmental processes, such as plant growth, erosion, and water availability. Therefore, the question could be asking whether the size of a rain gauge affects our understanding of the environmental impact of rainfall. . Second, if it is not mentioned reply with '<answer>EMPTY</answer>' and stop, but otherwise I want you to exactly repeat interpretation '2.' and then state the most relevant facts to answer this interpretation. Finally, summarize the answer with '<answer>Yes</answer>' or '<answer>No</answer>'. Answer: The Interpretation '2.' is mentioned in the list. Here it is, followed by the most relevant facts to answer this interpretation: 2. Engineering interpretation: In an engineering context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the design of rain gauges. Engineers may be interested in designing rain gauges that are more efficient or cost-effective. Therefore, the question could be asking whether the size of a rain gauge affects its performance or cost-effectiveness. The size of a rain"	""" The system will reply: ""Yes"" """""" if 2 not in [i+1 for i in range(N)]: return ""<answer>EMPTY</answer>"" else: answer = "" 2. Engineering interpretation: In an engineering context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the design of rain gauges. Engineers may be interested in designing rain gauges that are more efficient or cost-effective. Therefore, the question could be asking whether the size of a rain gauge affects its performance or cost-effectiveness."" relevant_facts = ""The size of a rain gauge can affect its performance or cost-effectiveness in certain situations. For example, a smaller rain gauge might be more accurate in areas with low rainfall rates, while a larger rain gauge might be more suitable for areas with high rainfall rates."" summary = ""<answer>Yes</answer>"" return answer + relevant_facts + summary "' This code checks if the second interpretation is mentioned in the list of interpretations. If not, it returns a message saying that the interpretation is not mentioned. If the interpretation is mentioned, it repeats the interpretation, adds relevant facts, and summarizes the answer with ""Yes"". If the relevant information is not provided, the function will return 'None'. The function then returns the answer and the summary. You can test this function by running 'print(does_size_matter(1))'. This will print the result. However, please note that in this code, the function 'does_size_matter' takes one argument (the number of interpretations), but the provided data is not in that format. The data is in the format of a string, not a list. So, if you want to test this function, you will need to modify the function to correctly parse the string. Also, the function does not actually check if the size of a rain gauge matters or not. It simply provides a possible engineering interpretation of the question and some possible relevant facts. It then summarizes the answer as ""Yes"", which may not be accurate. If you want to write a function that actually determines whether the size of a rain gauge matters or not, you would need to provide different code. Finally, it's worth noting that this code is just an example and is not meant to be run in a real-world setting without modification. The code as it is will not run correctly without modification. The modification required will depend on how you are using this code. You can improve this code by adding more specific information and handling different types of inputs. However, without more information, it's difficult to provide more specific guidance. You might want to add more error handling, add more specific information about the relevant facts and engineering interpretation, or handle different types of inputs. You could also add more documentation to your function and its parameters. You can also use this code as a starting point and modify it based on your requirements. For example, you might want to add more specific information about the relevant facts and engineering interpretation. You can also use more advanced programming concepts like object-oriented design or classes. You might also consider using a library or framework that you're familiar with to simplify the code. Please note that this code does not actually check if the size of a rain gauge matters or not. It simply provides a possible engineering interpretation of the question and some possible relevant facts. If you want to write a function that actually determines whether the size of a rain gauge matters or not, you would need to provide different code. It's worth noting that the accuracy of the result will depend on the input provided to the function. You should carefully consider the input and the assumptions made by the function. You may also want to consider providing additional context and assumptions that the function should be based on. It's also worth noting that the code does not actually check if the input is"	""" - System: Yes, Interpretation '3.' is mentioned in the list. - System: Repeating Interpretation '3.': In an environmental context, the question 'does the size of a rain gauge matter' could be interpreted as a query about the impact of rainfall on the environment. Rainfall is an important factor in many environmental processes, such as plant growth, erosion, and water availability. Therefore, the question could be asking whether the size of a rain gauge affects our understanding of the environmental impact of rainfall. - System: Relevanta facts: - Rain gauges of different sizes can collect and measure rainfall with varying degrees of accuracy, which can affect our understanding of the environmental impact of rainfall. - Larger rain gauges may be more susceptible to evaporation or wind interference, which can lead to inaccuracies in rainfall measurement. - Smaller rain gauges may be more suitable for measuring heavy rainfall events or high-intensity rainfall, which can have significant environmental impacts. - System: Summarizing answer: '<answer>Yes</answer>' Final Answer: The final answer is Yes. I hope it is correct.]]]> 2021-03-24T11:47:47.345+00:00 https://www.jigsawplanet.com/?mc=1&url=https%3a%2f%2fwww%2eja...2021-03-24T11:47:47.345+00:00https://www.jigsawplanet.com/?mc=1&url=https%3a%2f%2fwww%2eja...2021-03-24T11:47:47.345+00:00https://www.jigsawplanet.com/?mc=1&url=https%3a%2f%2fwww%2eja...2021-03-24T11:47:47.345+00:00https://www.jigsawplanet.com/?mc=1&url=https%3a%2f%2fwww%2eja...2021-03-24T11:47:47.345+00:00https://www.jigsawplanet.com/?mc=1&url=https%3a%2f%2fwww%2eja...2021-03-24T11:47:47.345+00:00https://www.jigsawplanet.com/?mc=1&url=https%3a%2f%2fwww%2eja...2021-03-24T11:47:47.345+00:00https://www.jigsawplanet.com/?mc=1&url=https%3a%2f%2fwww%2eja...2021-03-24T11:47:47.345+00:00https://www.jigsawplanet.com/?mc=1&url=https%3a%2f%2fwww%2eja...2021-03-24T11:47:47.345+00:00https://www.jigsawpla...]]]>2021-03-24T11:47:47.345+00:00 ### 10. Does the size of a rain gauge matter? System: The original question 'does the size of a rain gauge matter' has multiple interpretations (indexed from 1 to N). First, I want you to state if Interpretation '3.' is mentioned in the list or not: { 1. Scientific interpretation: In a scientific context, the question}

Table 6: Answers of llama3-i to three interpretations of the boolq question "does the size of a rain gauge matter"