Accurate and Nuanced Open-QA Evaluation Through Textual Entailment

Peiran Yao and **Denilson Barbosa**
Department of Computing Science
University of Alberta
{peiran, denilson}@ualberta.ca

Abstract

Open-domain question answering (Open-QA) is a common task for evaluating large language models (LLMs). However, current Open-QA evaluations are criticized for the ambiguity in questions and the lack of semantic understanding in evaluators. Complex evaluators, powered by foundation models or LLMs and pertaining to semantic equivalence, still deviate from human judgments by a large margin. We propose to study the entailment relations of answers to identify more informative and more general system answers, offering a much closer evaluation to human judgment on both NaturalQuestions and TriviaQA while being learning-free. The entailment-based evaluation we propose allows the assignment of bonus or partial marks by quantifying the inference gap between answers, enabling a nuanced ranking of answer correctness that has higher AUC than current methods.

1 Introduction

Open-domain question answering (Open-QA) is a long-established task requiring systems to generate precise answers to factual questions on any topic, from information in a large corpus of text (Voorhees and Tice, 2000; Zhang et al., 2023). A more restricted form of open-domain QA where answers are short is still regarded as challenging and as a reasonable test for the capabilities of recent large language models (LLMs) (Anil et al., 2023; Touvron et al., 2023), particularly when it comes to the assessment of LLM honesty (Yang et al., 2023), calibration (Tian et al., 2023). Open-QA benchmarks (Joshi et al., 2017; Kwiatkowski et al., 2019; Lee et al., 2019; *inter alia*), consisting of pairs of curated questions and manually-annotated gold answers, have been under intense scrutiny because current automated evaluations have been found primitive, flawed, and insufficient to capture the true capabilities of Open-QA systems (Chen et al., 2019; Boyd-Graber and Börschinger, 2020; Kamalloo et al., 2023; Wang et al., 2023).

Open-QA Evaluators. Let $\mathcal{S}$ be the set of finite strings. Given a question $q \in \mathcal{Q} \subset \mathcal{S}$, an Open-QA system generates a free-text *system answer* $a \in \mathcal{A} \subset \mathcal{S}$, while reference *gold answer*(s) $a^* \in \mathcal{A}^* \subset \mathcal{S}$ are provided by humans. In the typical setting, an evaluator $f : \mathcal{Q} \times \mathcal{A} \times \mathcal{A}^* \mapsto \{0, 1\}$ compares the system answer a with the gold answer(s) a^* to provide an *evaluator judgment* of whether the system correctly answered the question q.

While a wide variety of evaluators would be possible, current Open-QA benchmarks resort to fairly strict and primitive evaluators, which do a poor job with under-specified questions or when the system provides an answer that is either more general or more specific than the gold standard, and are believed to have hindered the understanding of LLM's "emergent" abilities (Schaeffer et al., 2023).

Ambiguity in Open-QA benchmarks. Questions from Open-QA benchmarks are often ambiguous and under-specified (Boyd-Graber and Börschinger, 2020), leading to multiple possible answers that are not always covered by the gold answers (Si et al., 2021). Figure 1 presents an example from the NaturalQuestions benchmark (Kwiatkowski et al., 2019) where *"Oak Island"* is the sole gold answer to the question *"Where is the TV show The Curse of Oak Island filmed?"*. However, due to the lack of specificity, a case can be made that *more specific* answers such as *"on Oak Island, a small island off the coast of Nova Scotia, Canada"*, or *more general* answers such as *"Nova Scotia, Canada"* should be accepted. The former covers the gold answer and provides more details, while the latter has a lower level of specificity than the gold answer. Exact word matching, a commonly used evaluator, fails with both answers. More advanced automated evaluations, including

Code and data of the work are available at https://github.com/U-Alberta/QA-partial-marks.

Findings of the Association for Computational Linguistics: ACL 2024, pages 2575–2587
August 11-16, 2024 ©2024 Association for Computational Linguistics

semantic similarity models and LLM in-context learning, are also shown to be incapable of capturing such intricacies and deviate from human judgment (Kamalloo et al., 2023; Wang et al., 2023).

Our contributions. We study semantic relations between system answers and gold answers that consider whether system answers cover the gold answer while providing more details, or vice versa. Naturally, textual entailment, the task of determining whether a piece of text entails or contradicts another, is a suitable and training-free tool for this categorization. We propose to use textual entailment for the evaluation of Open-QA systems, and show that entailment-based evaluation metrics, even when used without finetuning, are consistent with human judgments and are more effective in capturing the true capabilities of a range of Open-QA systems when evaluating system answers on both NaturalQuestions (NQ) and TriviaQA (TQ; Joshi et al., 2017). We also propose to use entailment-based evaluation metrics to assign bonus or partial marks to system answers by quantifying the inference gap between system answers and gold answers. We argue that our metric offers a more informative and fairer alternative to current binary evaluation metrics. Such a more accurate and nuanced QA evaluation scheme is valuable in solidifying the large body of concurrent studies that build on short-answer QA evaluations.

2 Related Work

Typical Open-QA evaluators rely on exact word match accuracy (lexical match), F_1 score over word matches (formally defined in Bulian et al. (2022)), some semantic similarity model such as BERTScore (Zhang et al., 2020) or BLEURT (Sellam et al., 2020), or zero-shot or in-context learning using an LLM (Chen et al., 2023b; Kamalloo et al., 2023). These approaches have been under scrutiny, however. Bulian et al. (2022), Kamalloo et al. (2023) and Wang et al. (2023) have looked at quantitatively assessing the correctness of automated evaluators by comparing the judgments they produce against *gold judgments* produced by human experts. They find that unsupervised automated evaluators, including those powered by pre-trained foundation models and LLMs, are not consistent with human judgments.

Wang et al. (2023) released the EVOUNA dataset for the evaluation of automated evaluators, with 3,020 questions from NQ and 1,938 from TQ. Questions are filtered to exclude those with outdated gold answers, and system answers generated by state-of-the-art Open-QA systems, namely DPR (Karpukhin et al., 2020) + FiD (Izacard and Grave, 2021), InstructGPT and ChatGPT (Ouyang et al., 2022), GPT-4 (OpenAI, 2023), and BingChat, are annotated with gold judgments. An ideal automated evaluator should produce judgments that are consistent with human judgments, and thus, have higher F_1 scores and accuracies when evaluated against human judgments. We base our work on EVOUNA and study the relation between gold answers and system answers, and subsequently derive an entailment-based evaluator $\hat{f}$ that is more consistent with human judgments. We also extend the range of $\hat{f}$ from a $\{0, 1\}$ binary prediction to $\mathbb{R}$ to achieve a more informative and fairer evaluation.

To overcome traditional metrics' lack of semantic understanding and, henceforth, underestimation of performance, recent efforts have focused on understanding the semantic equivalence of answers and developing evaluators accordingly (Si et al., 2021; Bulian et al., 2022; Kamalloo et al., 2023). We argue that the semantic equivalence is not the only relation between valid system answers and gold answers. A valid answer can range from a vague, less informative one (*e.g.*, a range, time period, or region) to a very specific and detailed answer (*e.g.*, a precise number, time, or location).

Bulian et al. (2022) proposed to accept all answers that contain at least all relevant content of the gold answer and no misleading content, while making no explicit distinction between semantic equivalence and entailment. We extend this idea and propose to assign a partial order to system answers with regard to how much relevant information the answer contains relative to the gold answer.

3 The Answer Hierarchy

Let A^* denote the set of gold answers for an Open-QA benchmark. We define two other sets of answers: A_{sup} is the set of *superior* answers that provide more information than what is in the gold standard, and A_{inf} is the set of *inferior* answers that only address the question partially. Given a system answer a and the corresponding gold answer a^*, we say that $a \in A_{sup}$ if and only if a entails a^* within the context of the question. (This naturally extends to the case where multiple gold answers are given). Similarly, we say that $a \in A_{inf}$ if and only if it is entailed by a^*. Finally, an answer would be

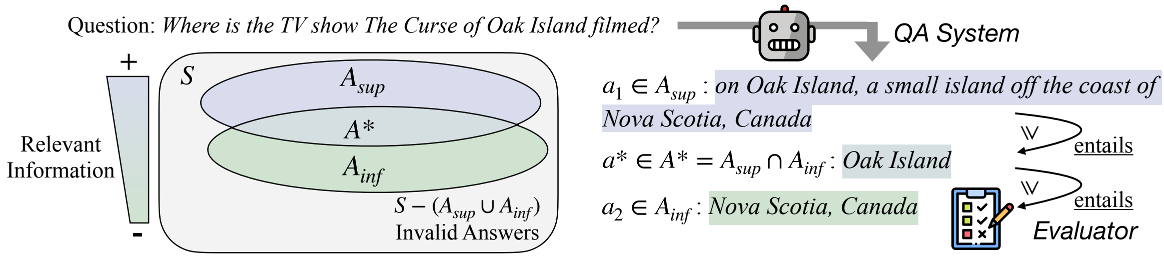

Figure 1: QA systems may generate a variety of correct answers that are neither exact matches nor semantic equivalents of the gold answer. Judging by the amount of information relevant to the gold answer that the system answers provide, we obtain a partial order of system answers with respect to the gold answer using textual entailment, and group answers into a hierarchy of subsets.

	DPR-FiD		**InstructGPT**		**ChatGPT**		**GPT-4**		**BingChat**	
Evaluator	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
Lexical Match†	92.0	89.7	86.9	84.8	85.0	80.3	87.6	82.5	87.8	82.3
BERTScore†	83.5	75.1	77.6	69.5	81.2	72.8	84.3	76.0	77.5	67.5
GPT-3.5†	**95.3**	**93.6**	87.2	84.1	86.9	82.2	86.8	80.9	77.3	69.5
Entailment	94.8	92.5	**92.7**	**90.2**	**92.6**	**88.9**	**93.8**	**90.1**	**92.6**	**88.1**
Entailment (small)	91.5	88.5	88.0	85.4	87.7	83.2	89.9	85.0	87.8	82.0
GPT-3.5 (best prompting)	95.5	93.9	88.3	84.5	89.4	84.5	91.2	86.0	87.1	80.4
Another Human†	97.4	96.3	97.8	96.8	96.5	95.6	97.9	96.6	97.2	95.5
on EVOUNA-NaturalQuestions										
Lexical Match†	91.8	94.7	94.8	92.3	95.2	92.3	94.8	91.1	94.1	89.8
BERTScore†	75.1	65.5	84.1	75.7	88.4	80.8	90.5	93.5	88.3	80.4
GPT-3.5†	**97.3**	**95.7**	94.2	91.2	95.5	92.5	95.7	92.4	88.2	80.9
Entailment	96.8	94.7	**96.6**	**94.2**	**96.6**	**94.2**	**97.4**	**95.3**	**95.9**	**92.5**
Another Human†	100	100	99.6	99.4	99.2	98.8	99.2	99.8	99.9	95.5
on EVOUNA-TriviaQA										

Table 1: Using human judgments as the gold standard, entailment-based evaluation of Open-QA systems on both NQ and TQ yields higher F_1 scores and accuracies than lexical match, BERTScore, and GPT-3.5 when evaluating the judgments against gold judgments in EVOUNA. Metrics chosen following Wang et al. (2023). Higher scores and accuracies indicate that evaluator judgments are more consistent with human judgment. Judgments from another human are included as a reference of the upper bounds induced by ambiguity and inconsistencies in creating the gold judgments. Top performing evaluators are in bold. †: scores reported by Wang et al. (2023).

incorrect if and only if a and a^* are not entailed by each other; in this case $a \in S - (A_{sup} \cup A_{inf})$. As a special case, it follows that answer a is equivalent to the gold answer a^* *iff.* $a \in A_{sup} \cap A_{inf}$.

Before textual entailment, question-answer pairs are rewritten as declarative statements (introduced by Demszky et al. (2018) as QA2D) using GPT-3.5, as the question and context are important for assessing answers (Kamalloo et al., 2023). For example, the gold answer in Figure 1 is converted into the statement "*The TV show The Curse of Oak Island is filmed on Oak Island*". Inspired by LLM's strong performance in natural language inference tasks (Qin et al., 2023), we use GPT-3.5 to conduct textual entailment tests. In the finalized approach, two steps, converting a question-answer pair to a declarative statement, and performing textual entailment test on declarative statements, are perform by a LLM in a few-shot manner. We validate that for both steps GPT-3.5 as the LLM achieves high statistical reliability (Appendix A.2) in terms of the agreement across different seeds, and high validity (Appendix A.3) in terms of the alignment with human labels. Implementation details and a worked example are provided in Appendix A.1.

System answers deserve partial credits and bonus credits. The above-described entailment-based evaluator reveals that $A_{sup} \oplus A_{inf}$ (the disjoint union) represents a considerable amount of valid system answers that would otherwise be disregarded (Table 9, 10). Assuming the hierarchy holds, we would see a 10.1% and a 6.5% increase in accuracy for NQ and TQ, respectively, which account for the reported underestimation of QA

Method	F_1	Acc
Llama-2 (SFT)	94.6	92.3
Llama-2 + NLI (SFT)	94.8	92.6
CVI	84.7	73.5
Entailment (0-shot)	93.5	90.2

Table 2: Without doing supervised finetuning (SFT), entailment-based evaluation yields comparable performance to data-driven approaches like finetuned Llama-2-7B and CVI when evaluating system answers on NQ.

performance (Bulian et al., 2022; Wang et al., 2023). We validate the hierarchy by demonstrating that higher positions correspond to better answers judged by humans. This is supported by one-tailed Fisher's exact tests, all yielding significant results with $p < 0.01$ with the exception of DPR-FiD and $A_{sup} - A^{\star}$ in TQ. Details of the statistical tests can be found in Appendix B.2.

The answer hierarchy is a superior automated evaluator. Treating $A_{sup} \cup A_{inf}$ as correct answers[1] bring the evaluation results of various systems closer to human performance, as shown in Table 1, and confirms the observation that current evaluators[2] unfairly misrepresent the capabilities of those systems. However, unlike previous studies which resorted to manual inspections of answers, our evaluator allows the same observation in an automated way. Moreover, our evaluator can be used for any benchmark. The system where entailment does not improve the performance is DPR-FiD, which is an extractive model that outputs a span of text that requires less semantic understanding to evaluate than complete sentences. Nevertheless, the entailment evaluator assessed that system very closely to the numbers reported in the literature.

Although learning-free, entailment is comparable to finetuned evaluators. Bulian et al. (2022) and Kamalloo et al. (2023) advocate for learned evaluators to close the gap between automated and human evaluation. For comparison, we partition EVOUNA-NQ by questions into 50:50 train/test splits, and finetune a Llama-2 (Touvron et al., 2023) model which only performs slightly better than our method. We show that explicitly including entailment as a feature improves the finetuned model (*+NLI* in Table 2). Moreover, we finetune another Llama-2 model with the same training data, but with system answers as contexts to predict gold answers, in order to use conditional $\mathcal{V}$-information (CVI; Hewitt et al., 2021; Chen et al., 2023a) as another evaluator that builds on usable information from the system answers. The results are shown in Table 2. Entailment also yields better results than in-context learning with four examples (*best prompting* in Table 1). Details of the finetuning process and the learned baselines are in Appendix C.2.

Out-of-the-box entailment outperforms prompt engineering. Since our method is implemented solely with GPT-3.5, it can be seen as a prompt/flow-engineering method that outperforms the best prompt engineering technique among those Kamalloo et al. (2023) extensively explored (*best prompting* in Table 1), such as Chain-of-Thought (Wei et al., 2022) and in-context learning. Meanwhile, the pre-processing and entailment tests can be implemented independent of LLMs, for example using DeBERTa (He et al., 2023) as the NLI model and Llama-2-7B as the question to statement conversion model (*small* in Table 1), while still achieving improved results.

4 Towards Partial Marks

Bulian et al. (2022) demonstrated by examples that the seemingly continuous F_1 score is not indicative of how close the system answer is to the gold answer. Going beyond directly using the classification probability from the NLI model (Chen et al., 2021), we hypothesize that measuring the inference gap, *i.e.* how many steps, assumptions, and additional pieces of information are needed to derive a system answer a from a gold answer a^*, can be used to assign partial marks in a way that reflects semantic closeness. Inspired by Chain-of-Thought prompting, explainable natural language inference (Camburu et al., 2018), and LLM-based decompositions of implicit content (Hoyle et al., 2023), we propose to use LLM (GPT-3.5 in our experiments) to explain step-by-step the inference process behind how a^* entails a, along with assumptions and additional knowledge required (*"Inference"*). Based on the explanation, a score of inference difficulty is directly produced by the LLM (*"LLM Score"*), or the number of steps is counted (*"#Steps"*) as partial marks. Details of the scoring schemes, worked examples, and alternatives are discussed in Appendix D.

We have examined the inter-set ranks in the answer hierarchy in §3. When it comes to the intra-set ranking of partial answers in $A_{inf} - A_{sup}$,

[1] Alternative choices discussed in Appendix B.1.

[2] Details about baselines in Appendix C.1.

Method	AUC
Inference + LLM Score	0.91
Inference + #Steps	0.91
LLM Score	0.88
F_1 Score	0.78

Table 3: Using LLM to explain the inference process behind how gold answers entail the system answers leads to higher AUROC in predicting human judgements on NQ, making it a good candidate for partial marks.

Brunner-Munzel tests (Brunner and Munzel, 2000) show that both LLM Score and #Steps, as well as other baselines in Table 3, assign higher scores to human-accepted answers ($p < 0.001$). Quantitatively, Table 3 shows that the inference-process-based scores have higher AUROC on EVOUNA-NQ than F_1 score or using GPT-3.5 to directly assess system answers on a 5-point scale ("*LLM Score*"), indicating that partial marks assigned by our method are suitable for capturing the nuanced goodness differences between system answers.

5 Conclusion

In theory, textual entailment is considered AI-Complete (Dagan et al., 2009) - an embodiment of general AI that solves all AI tasks, Open-QA evaluation included. In practice, we showed that state-of-the-art textual entailment provides a simple yet powerful replacement for Open-QA evaluation, and it offers the prospect of soft and partial marks.

Limitations

Our current work studies the potential of textual entailment as a fairer and finer-grained replacement for Open-QA evaluation. We only explored using the method for benchmarking QA systems. However, it remains a highly interesting topic to investigate how it can be used as a softer signal for training QA systems with the potential of improvements, given the success of smoothed labels (Hinton et al., 2015; Szegedy et al., 2016). We only studied QA benchmarks consisting of mostly factoid questions and relatively short and simple answers. For QA tasks that require more complex, and potentially multi-passage multi-facet answers, it is unclear how well the original entailment method can be directly applied. Future work is required to investigate the entailment relations and the matching between multiple units of meaning, such as in Laban et al. (2022), to extend our work to more complex QA tasks.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). This work is also supported in part by a gift from Scotiabank. Icons in Figure 1 are designed by OpenMoji under CC BY-SA 4.0 license, and by Flaticon-Freepik.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

Edgar Brunner and Ullrich Munzel. 2000. The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42(1):17–25.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop* *on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023a. REV: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030, Toronto, Canada. Association for Computational Linguistics.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023b. Phoenix: Democratizing ChatGPT across languages. *CoRR*, abs/2304.10453.

Ido Dagan, Bill Dolan, Bernado Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *CoRR*, abs/1809.02922.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1626–1639. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. 2021. What's in a name? answer equivalence for open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9623–9629, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece*. European Language Resources Association.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-qa evaluation. In *Advances in Neural Information Processing Systems*, volume 36, pages 77013–77042. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *CoRR*, abs/2312.07000.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A survey for efficient open domain question answering. In *Proceedings of the 61st Annual Meeting* of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? A comparative study on chatgpt and fine-tuned BERT. *CoRR*, abs/2302.10198.

A Entailment Test Implementation

A.1 Detailed Settings

When using entailment to obtain the answer hierarchy in §3, we use `gpt-3.5-turbo-1106`. The gold answer-question and system answer-question pairs are converted to two declarative statements using the first prompt in Table 14. The two examples in the first prompt are chosen from EVOUNA-NQ, and as the dataset size is large enough, we do not need to use a separate dataset for prompt engineering. The two declarative statements are then used as the premise and hypothesis and vice versa in the second prompt in Table 14 to obtain the entailment classification in two directions. For all GPT-3.5 API calls, we set `seed=42, temperature=0.0` to ensure reproducible results, and set `max_tokens=300`.

Here is a working example of the entailment test on a system answer generated by InstructGPT and a gold answer from NQ:

> Question: where is fe best absorbed in the body
> Gold answer: in the duodenum
> System answer: Iron is best absorbed in the small intestine.
>
> Gold statement: Fe is best absorbed in the body in the duodenum.
> System statement: Iron is best absorbed in the small intestine.
>
> Entailment test: Gold statement entails system statement, but not the other way around. Therefore, the system answer belongs to $A_{inf} - A_{sup}$. Meanwhile, human annotator judged the system answer as correct in EVOUNA.

A.2 Assessment of Reliability

To assess the statistical reliability (consistency) of our method, we measure the agreement across different random seeds, and the potential impact on the overall performance. We repeat the LLM-based steps on EVOUNA-NQ and EVOUNA-TQ subsets of size 2,000, each using random seeds 0,1,2,3 for GPT3.5 calls while keeping the rest of the settings controlled.

Reliability of question-answer to statement conversion. We calculate the consistency of generated statements from the same question-answer pairs across different seeds using BLEU and exact sentence matching, as shown in Table 4. These results indicate that the generated statements are fairly consistent across different seeds with almost all statements being identical or very similar.

Dataset	BLEU	Exact Match
NQ	93.9 ± 1.6	$86.7\% \pm 2.1$
TQ	94.7 ± 0.1	$83.6\% \pm 0.2$

Table 4: Reliability of question-answer to statement conversion, measured by average pairwise BLEU scores and percentages of exact matches across three runs.

Reliability of textual entailment test. We measure the agreement of textual entailment predictions across different pairs of seeds for the same golden-system answer pairs using Cohen's Kappa, as in Table 5. The results are interpreted as *almost perfect* agreement according to Landis and Koch's (1977) guideline.

Dataset	0 vs 1	0 vs 2	0 vs 3	1 vs 2	1 vs 3	2 vs 3
NQ	0.902	0.900	0.902	0.922	0.917	0.920
TQ	0.873	0.882	0.870	0.872	0.870	0.865

Table 5: Reliability of textual entailment test, measured by pairwise Cohen's Kappa across three runs.

Reliability of hierarchy construction. The result from textual entailment is used to categorize a system answer into one of the sets in the hierarchy (Table 9 and 10). Again we measure the agreement of the categorization across different pairs of seeds using Cohen's Kappa. The results in Table 6 show even better agreement than the textual entailment test as multiple candidate golden answers are considered in this step.

Reliability of QA evaluation. We assess whether different seeds lead to different QA system evalu-

Dataset	0 vs 1	0 vs 2	0 vs 3	1 vs 2	1 vs 3	2 vs 3
NQ	0.906	0.912	0.907	0.932	0.928	0.929
TQ	0.921	0.925	0.924	0.927	0.917	0.917

Table 6: Reliability of answer hierarchy construction, measured by pairwise Cohen's Kappa across three runs.

ation results (Table 1) as reflected by the variance of F1 scores and accuracy. On the two subsets, different seeds have virtually no impact on the overall QA evaluation as seen in Table 7.

Dataset	F_1	Accuracy
NQ	0.918 ± 0.002	0.876 ± 0.003
TQ	0.962 ± 0.001	0.934 ± 0.001

Table 7: Reliability of QA evaluation, measured by the variance of F1 scores and accuracy across three runs.

A.3 Assessment of Validity

Converting a question-answer pair to a declarative statement (known as QA2D) is a well-established task. Demszky et al. (2018) provided a dataset where the dev set has 10,344 question-answer pairs with a human-written declarative statement (test set unavailable). We compare our 2-shot LLM generated statements with the human-written statements using BLEU and ROGUE (Table 8), and the generations are very similar to human-written statements and a fine-tuned T5 baseline[3].

Model	BLEU	ROGUE-1	ROGUE-2	ROGUE-L
GPT-3.5	72.5	92.5	83.5	85.8
T5	72.7	90.1	82.4	85.8

Table 8: Comparison of generated declarative statements with human-written statements on QA2D dataset.

Our zero-shot prompt for textual entailment (Table 14) is adapted from Qin et al. (2023). They have tested the validity of this textual entailment test method on RTE and CB datasets and reported a high accuracy of 0.86 and 0.89 respectively. Zhong et al. (2023) used a slightly different prompt for the same task and reported GPT-3.5 NLI accuracy to be higher than finetuned BERT-large and RoBERTa-large on both MNLI-m and RTE.

Finally, the validity of final QA evaluation is confirmed by the high correlation with human judgment (Table 1).

[3] `domenicrosati/QA2D-t5-base` on Huggingface.

B Inter-set Order Validation

B.1 Hierarchy of Answer Sets

The entailment test organizes system answers into a hierarchy of sets: in Table 9 and 10, the four rows corresponds to the four sets at different levels of the hierarchy: (1) $A_{sup} - A_{inf}$, (2) $A_{sup} \cap A_{inf}$, (3) $A_{inf} - A_{sup}$, and (4) $S - (A_{sup} \cup A_{inf})$. The size of the sets are shown in the *Count* column.

Rank	in A_{sup}	in A_{inf}	Count
(1)	Yes	No	514
(2)	Yes	Yes	10,061
(3)	No	Yes	1,000
(4)	No	No	3,470

Table 9: Distribution of system answers in different sets of the answer hierarchy for EVOUNA-NQ.

Rank	in A_{sup}	in A_{inf}	Count
(1)	Yes	No	168
(2)	Yes	Yes	7,890
(3)	No	Yes	460
(4)	No	No	1,172

Table 10: Distribution of system answers in different sets of the answer hierarchy for EVOUNA-TQ.

We propose to treat the union (denoted as $\cup$) of A_{inf} and A_{sup} as the correct system answers. We chose to include $A_{inf} - A_{sup}$ as we follow the human annotation guideline of EVOUNA that considers the lack of specificity in questions and accepts answers of all levels of specificity. Meanwhile, alternative choices like excluding $A_{inf} - A_{sup}$ (denoted as $-$) have negligible impact on the discussion, as the $A_{inf} - A_{sup}$ sets have a small size for both NQ and TQ, as shown in Table 9 and 10. We report the evaluation results of excluding $A_{inf} - A_{sup}$ in Table 11.

B.2 Statistical Tests

As summarized in §3, we conduct statistical tests to verify if the four sets do have a order. We hypothesize that the higher the rank, the more likely the system answer is correct. We use one-tailed Fisher's exact test do a pairwise comparison the distribution of human judgements in the four sets with DPR-FiD, a method with extractive nature that makes semantic understanding excessive and lexical matching sufficient, excluded. The results are shown in Table 12.

	DPR-FiD		InstructGPT		ChatGPT		GPT-4		BingChat	
Evaluator	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
Entailment (∪)	94.8	92.5	92.7	90.2	92.6	88.9	93.8	90.1	92.6	88.1
Entailment (−)	95.1	93.1	92.5	90.5	91.6	88.0	93.6	90.1	92.3	87.6
on EVOUNA-NaturalQuestions										
Entailment (∪)	96.8	94.7	96.6	94.2	96.6	94.2	97.4	95.3	95.9	92.9
Entailment (−)	96.2	93.9	95.0	92.3	96.5	94.2	96.7	94.2	94.8	90.9
on EVOUNA-TriviaQA										

Table 11: When $A_{inf} - A_{sup}$ is excluded from the judged correct answers (denoted as −), the evaluation results of various systems do not change significantly compared to when $A_{inf} - A_{sup}$ is included (denoted as ∪). Our discussion in §3 is not affected by the choice of including $A_{inf} - A_{sup}$.

Dataset	Test	odds ratio	p
NQ	(1)>(2)	1.35	0.008
	(2)>(3)	2.59	2e-40
	(3)>(4)	6.25	8e-108
TQ	(1)>(2)	0.17	N/A
	(2)>(3)	5.22	3e-46
	(3)>(4)	7.88	8e-54

Table 12: Results of Fisher's exact test for the answer hierarchy in EVOUNA.

C Baseline Method Details

C.1 Unsupervised Evaluators

Wang et al. (2023) evaluated multiple unsupervised evaluators, including lexical match, BERTScore, and GPT-3.5, on both EVOUNA-NQ and EVOUNA-TQ. We make the comparisons with the numbers reported in their paper and refer the readers to Wang et al. (2023) for the detailed settings of those baseline evaluators. They also explored four additional prompting methods for the GPT-3.5 evaluator: Ignoring Background Information, Giving Reasons, Chain-of-Thought, and In-Context Learning, with the exact prompts provided in their paper. For each category in Table 1, we choose the best performing method among the four for comparison as in the *GPT-3.5 (best prompting)*. This represents the upper bound performance of their prompt engineering efforts that is only achievable if an oracle exists that knows the best prompt for each QA system.

For *Entailment (small)*, we use the same prompt as in Table 14 row 1, but with 4-bit quantized `Llama-2-7B-GPTQ`[4] instead of GPT-3.5 as the model for question to statement conversion. We use a finetuned `DeBERTa-v3-large`[5] by Reimers and Gurevych (2019) as the NLI model.

[4] `TheBloke/Llama-2-7B-GPTQ` on Huggingface.

[5] `cross-encoder/nli-deberta-v3-large` on Huggingface.

C.2 Learned Evaluators

We perform a half-half partition of the EVOUNA-NQ dataset by question type to create a training set and a test set, where no question-answer pairs with the same question falls in the same split. A `Llama-2-7b-chat-hf` model is finetuned on the training set by inserting the question, gold answer, system answer, and human judgment into the templates in Table 15. During inference, the same templates are used with human judgment left empty. Finetuning is done with the Huggingface `PEFT` (Mangrulkar et al., 2022) and `TRL` (von Werra et al., 2020) libraries. For CIV, two models with and without system answers as rationales are finetuned on the training set in the same fashion using templates in Table 16.

D Partial Mark Scoring

If the entailment test shows that the system answer is in $A_{inf} - A_{sup}$, we use GPT-3.5 and the prompt in Table 17 row 1 to generate an explanation of what inference process is required to deduce the system answer from the gold answer (*Inference*). The example system answer a_2 in Figure 1 is in $A_{inf} - A_{sup}$, and the explanation generated is as follows:

1. The TV show the Curse of Oak Island is filmed on Oak Island. (Given in S1)
2. Oak Island is located in Nova Scotia, Canada. [[INFO]]
3. Therefore, the TV show the Curse of Oak Island is filmed in Nova Scotia, Canada. (Combining steps 1 and 2)

Given the inference process explanation, we manually design the following partial mark scoring heuristics:

1. **CIA**: `-#Step*10-#INFO*3-#ASSUMPTION*5`
2. **C**: `-#Step*10`
3. **IA**: `-#INFO*3-#ASSUMPTION*5`

As an alternative, we use GPT-3.5 to score the difficulty of the inference process in the 5-point scale by providing the prompt in Table 17 row 2 as an additional message after the explanation step (*Inference+LLM Score*). The *LLM Score* baseline skips the explanation step and directly use GPT-3.5 to provide a 5-point-scale score using the prompt in Table 17 row 3.

The three manually designed scoring scheme are not significantly different from each other or from the automated *Inference+LLM Score* as shown in Table 13.

Method	**AUC**
Inference + LLM Score	0.9119
Inference + CIA	0.9120
Inference + IA	0.9118
Inference + C	0.9118
LLM Score	0.8827
F_1 Score	0.7770

Table 13: Area under the receiver operating characteristic curve (AUROC) in predicting human judgements on NQ system answers for more scoring schemes.

Description	Prompt
Convert question-answer pair to a declarative statement	Convert a question answer pair to a declarative statement, following these two examples: Q: where is the tv show the curse of oak island filmed A: Oak Island S: The TV show the Curse of Oak Island is filmed on Oak Island. Q: who wrote the first declaration of human rights A: Cyrus S: Cyrus wrote the first declaration of human rights Do not provide explanations. Provide the statement only. Follow the above examples and convert this pair: Q: {question} A: {answer} S:
Entailment test	Please identify whether the premise entails or contradicts the hypothesis in the following premise and hypothesis. The answer should be exact “entailment”, “contradiction”, or “neutral”. Provide only the answer from the three options. Do not provide explanations. Premise: {premise} Hypothesis: {hypothesis} Is it entailment, contradiction, or neutral?

Table 14: Prompts for the entailment test. The second prompt adapted from Qin et al. (2023).

Description	Prompt
Template for fine-tuning Llama-2	<s> [INST] Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers, simply answer Yes or No. Question: {question} Golden answers: {golden answer} AI answer: {system} [/INST] {system answer} </s>
Template for fine-tuning Llama-2 with NLI as a feature	<s> [INST] Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers, simply answer Yes or No. Question: {question} Golden answers: {golden answer} AI answer: {system} Can golden answers be inferred from AI answer: {yes or no} Can AI answer be inferred from golden answers: {yes or no} [/INST] {system answer} </s>

Table 15: Prompts for finetuned Llama-2-7B evaluators.

Description	Prompt
Template for training the QA model with system answer as the rationale	<s> [INST] Given the fact: {system answer}, answer this question: {question} [/INST] {golden answer} </s>
Template for training the QA model without rationales	<s> [INST] Answer this question: {question} [/INST] {golden answer} </s>

Table 16: Prompts for training the QA model with and without system answers as rationales (for CVI) by finetuning Llama-2.

Description	Prompt
Inference: Explain the inference process	We have two statements S1 (the premise) and S2 (the hypothesis). S1 entails S2. S1: {s1} S2: {s2} Now, list the reasoning process step by step to show how S2 can be deduced from S1. List the steps as numbered statements, starting from 1. If a step involves information not mentioned in S1 and S2, append [[INFO]] after the step. If an assumption must be made to deduce S2 from S1, append [[ASSUMPTION]] after the step. Provide the reasoning steps only. Do not include any other information.
Inference + LLM Score: Rate the inference difficulty based on the explanation	Based on the reasoning steps, rate how hard it is to deduce S2 from S1. 1: Very easy 2: Easy 3: Neither easy nor hard 4: Hard 5: Very hard Consider how many assumptions are needed, how much information is needed, and how much reasoning is needed. Return a number from 1 to 5 only. Do not include any other information.
LLM Score: Directly use LLM to provide a score of answer closeness	Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers? Simply give a score from 1 to 5. 1: The AI-generated answer is completely wrong. 2: The AI-generated answer is mostly wrong. 3: The AI-generated answer is neither wrong nor right. 4: The AI-generated answer is mostly right. 5: The AI-generated answer is completely right. Question: {question} Golden answers: {golden answer} AI answer: {system answer}

Table 17: Prompts for generating the inference explanation and scoring the inference difficulty.