Evaluating LLM Answers via Bidirectional Natural Language Inference

Anonymous ACL submission

Abstract

Evaluating large language model (LLM) outputs in open-ended question answering typically involves comparing a prediction to a reference answer, often by prompting another LLM to judge semantic equivalence. However, failing to mimic the reference answer does not necessarily imply that the model's response is incorrect. Both predictions and references may express information with varying levels of precision, making equivalence judgments insufficient. We introduce an evaluation framework that accounts for these variations by leveraging a Natural Language Inference (NLI) model bidirectionally. Specifically, we assess whether the prediction entails, contradicts, or is neutral with respect to the reference, and vice versa, resulting in a 3×3 matrix of interpretable outcomes. We show that these outcomes align well with human judgments across the three main interpretative categories.

1 Introduction

Evaluating open-ended question answering by large language models (LLMs) remains a challenging problem. Traditional metrics (BLEU, ROUGE, exact match) rely on lexical overlap with a reference answer, while more recent approaches leverage embedding-based similarity (e.g., BERTScore) or even use LLMs themselves as judges of answer quality. However, these methods tend to overemphasize similarity to the reference. In open-ended settings, both the prediction and the reference may convey valid information at different levels of precision, making strict equivalence-based evaluation inadequate.

We argue that evaluation should distinguish between what is factually correct and what the reference explicitly enables us to verify. In this perspective, a model may produce a cautious answer, omitting uncertain details while remaining correct, or a specialized answer, introducing additional information not present in the reference.

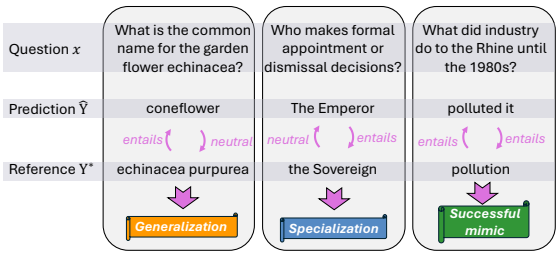


Figure 1: Illustration of the proposed bidirectional NLI-based evaluation.

To capture these nuanced relationships, we leverage Natural Language Inference (NLI). By applying an NLI model bidirectionally, from prediction to reference and from reference to prediction, we derive a set of interpretable categories that characterize the nature of the relationship between answers. This framework enables a fine-grained and automatic analysis of LLM predictive behavior beyond scalar similarity scores. We illustrate our method in Figure 1.

A key motivation for this work is to better account for cautiousness in LLMs. Recent work in NLP has emphasized the value of models that avoid committing factual errors under uncertainty (Rajpurkar et al., 2018; Lin et al., 2022; Kalai et al., 2025). Most existing approaches formalize this idea through full abstention, typically via explicit responses such as "I don't know". While useful, this binary notion captures only a limited form of cautious behavior. In contrast, we focus on a more subtle and frequently observed phenomenon: partial abstention, where a model avoids over-specific or potentially inaccurate details while still providing a broadly correct answer. For instance, a model may state a general fact while omitting uncertain dates, names, or numerical values. This notion relates to prior work in machine learning on set-valued predictions (Zaffalon et al., 2012; Yang et al., 2016; Kunitomo-Jacquin et al., 2025), where models express uncertainty by returning sets of

	Y^* *entails* $\hat{Y}$	Y^* *neutral* $\hat{Y}$	Y^* *contradicts* $\hat{Y}$
$\hat{Y}$ *entails* Y^*	**Successful mimic**	**Specialization**	**NLI inconsistency**
$\hat{Y}$ *neutral* Y^*	**Generalization**	**Compatibility**	**NLI inconsistency**
$\hat{Y}$ *contradicts* Y^*	**NLI inconsistency**	**NLI inconsistency**	**Clear mistake**

Table 1: Bidirectional NLI classification of predicted answer $\hat{Y}$ for the reference answer Y^*.

plausible outputs. Unlike these approaches, our bidirectional NLI-based framework does not prescribe how cautiousness should be rewarded; instead, it provides a necessary first step: the automatic detection and characterization of partial abstentions/generalizations, i.e., imprecision in otherwise correct predictions.

2 Related Work

NLI-inspired metrics have already been proposed to incorporate entailment signals as part of their evaluation framework. For example, the automatic evaluation metric NUBIA (Kane et al., 2020), which employs a pre-trained NLI model (RoBERTa fine-tuned to predict sentence similarity) to extract logical features reflecting entailment or contradiction between a generated text and a reference, showed competitive performance in evaluating machine translation and image captions quality. In the context of summarization, SummaC mobilised NLI for summarization consistency evaluation (Laban et al., 2022). By splitting documents and summaries into sentence pairs and feeding them into an NLI model, then aggregating entailment signals, it achieved state-of-the-art accuracy in flagging inconsistent summary content. We can also note, the evaluation metric MENLI based on NLI model predictions (Chen and Eger, 2023). Its authors showed that their NLI based metric yields good robustness to adversarial attacks.

Our work fits within this line of NLI-based evaluation approaches for generated text. But rather than reducing evaluation to a scalar score, we use bidirectional NLI relations to derive interpretative categories that offer a finer-grained characterization of model behavior.

3 Method

Let x denote a question, $\widehat{Y}$ an LLM predicted answer and Y^* one reference answer. We suppose that both $\widehat{Y}$ and Y^* may be affected by different levels of imprecision, as it is naturally the case in open-ended generation. In order to capture these different levels imprecision, we adopt the standard classification used in NLI introduced by the SNLI corpus (Bowman et al., 2015). We assume that an NLI classifier categorizes the relationship between a premise P and a hypothesis H into one of three possible classes: entailment, contradiction, or neutrality. See Appendix A, for interpretations of each class. Note that prior studies have shown that NLI model predictions can be logically inconsistent in practice (Arakelyan et al., 2024; Wu and Last, 2025) which has repercussion for our method.

In our approach, we apply NLI bidirectionally to each pair $(\widehat{Y}, Y^*)$, then depending on those results, classify the pair into six distinct state of comparison, that we will refer as to *interpretative categories*, as shown in Table 1. These interpretative categories offer a nuanced typology of the model behavior beyond correctness:

- **Successful mimic**: The prediction accurately reproduces the content of the reference. This is typically captured by traditional evaluation metrics based on similarity.

- **Generalization**: The prediction is more general or vague than the reference, yet remains fully supported by it. This reflects a form of partial abstention, where the model avoids overcommitment in the presence of uncertainty. Such cautious behavior can be desirable in high-stakes application when committing an error leads to serious consequences.

- **Specialization**: The prediction provides more specific information than the reference, potentially introducing additional details. While such answers may be correct, their validity cannot be confirmed solely based on the reference. Such specialization may be beneficial in some applications but also carry a risk of hallucination.

- **Compatibility**: the reference does not provide sufficient information to determine the correctness of the prediction. Such cases often involve plausible but unverifiable answers or

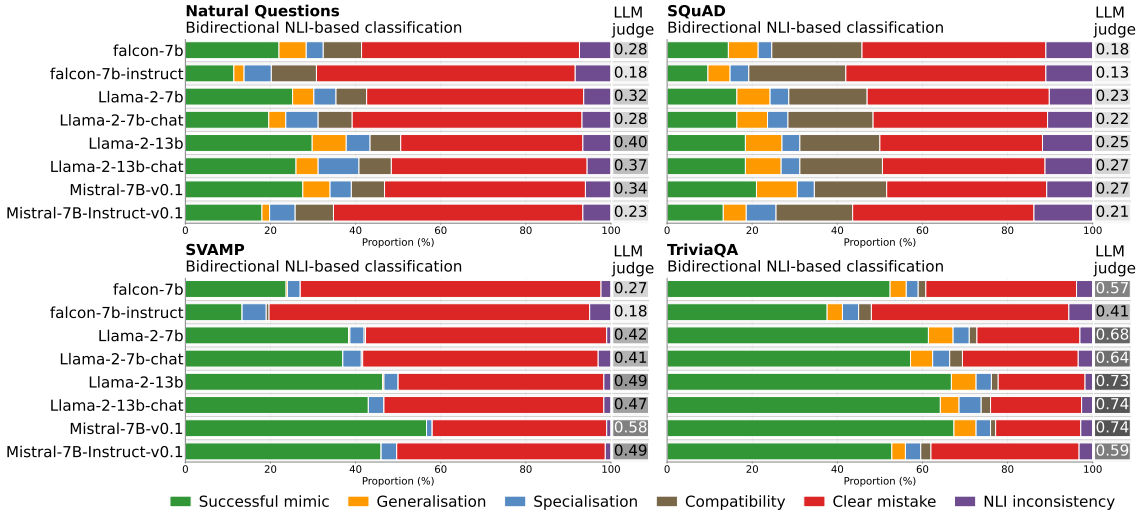


Figure 2: Results of bi-directional NLI-based classification compared with LLM as a judge baseline *gpt-4o-mini*.

alternative formulations that fall outside the scope of the reference.

- **Clear mistake**: The prediction directly contradicts the reference. This configuration corresponds to a clear factual error that can be automatically refuted based on the reference. It is a strong signal of incorrectness.
- **NLI inconsistency**: This may arise in practice due to limitations of the NLI model itself. Their presence may indicate either model instability or semantic ambiguity that challenges automatic resolution.

Note that both *Compatibility* and *Specialization* reflect situations where the factual status of the prediction cannot be conclusively determined from the reference alone. Since these responses may contain either truthful elaborations or commit factual errors, they should be prioritized for human verification to detect potential hallucinations.

4 Experiments

We provide a brief description of the considered LLMs, Question Answering datasets and evaluation baselines in Appendix B. Our analysis is conducted over 14,400 instances, aggregated from the 32 (8 LLMs, 4 datasets) experiments, each containing 500 samples, except for SVAMP dataset which includes 300 samples. In terms of NLI models, we used DeBERTa-Large-MNLI (He et al., 2020).

4.1 Bidirectional NLI-based classification

Figure 2 presents the distribution of the interpretative categories obtained with our bidirectional NLI-based classification across models and datasets compared with the LLM as-a-judge metric.

While models can exhibit similar performance according to the LLM-as-a-judge baseline, this aggregate score masks substantial differences in their underlying behaviors. For instance, Falcon-7B and Mistral-7B-Instruct achieve comparable scores on Natural Questions (0.28 vs. 0.23), yet display markedly different profiles:

> Falcon-7B produces a higher proportion of *Successful mimic* and fewer *Clear mistake*, while tending to generate **more general** answers. In contrast, Mistral-7B-Instruct tends toward **more specific** responses.

This illustrates the ability of our framework to provide a fine-grained analysis beyond surface similarity, capturing how model predictions differ from the reference in terms of behavioral patterns such as specificity or generality. To further illustrate these differences, we provide examples in Appendix C.

4.2 Correlation analysis with baseline metrics

We computed Spearman correlation coefficients ρ between one-hot representations of our interpretative categories and the baseline metrics as shown in Figure 3. The results confirm that baseline metrics strongly correlate with the *Successful mimic* category. In contrast, F1, ROUGE-L, BLEU, and the LLM judge exhibit only weak positive correlations with *Generalization* and *Specialization*. Notably, the LLM judge shows a relatively stronger correlation with *Specialization*, suggesting that it tends to accept some predictions as correct even when they appear more specific or detailed. We

provide additional insight into the comparison with the LLM-as-a-judge baseline in Appendix C.

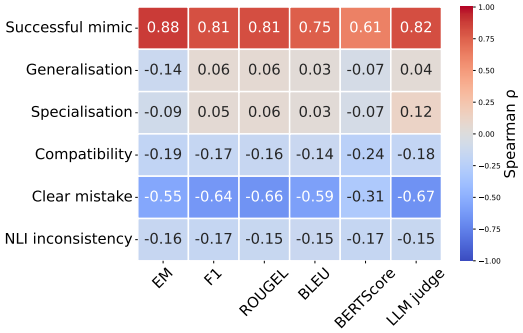
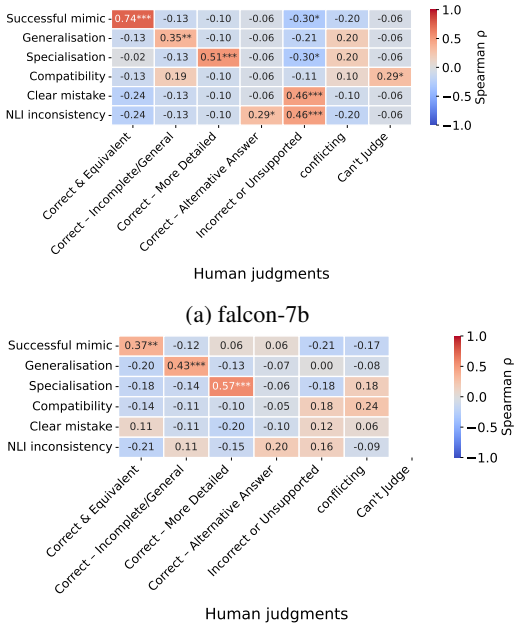


Figure 3: Spearman correlation ρ between one-hot encodings of our interpretative categories and baseline metrics (computed over 14,400 instances).



(a) falcon-7b

(b) Mistral-7B-Instruct

Figure 4: Spearman correlation ρ between one-hot encodings of our interpretative categories and human judgments, computed over 60 selected instances of dataset Natural Questions for falcon-7b 4a and Mistral-7B-Instruct 4b. Asterisks indicate statistical significance ($*$ for $p < 0.05$, $**$ for $p < 0.01$ and $***$ for $p < 0.001$).

4.3 Alignment with human judgment

We evaluated our approach on a subset of 60 instances, obtained by sampling 10 examples from each of our six interpretative categories to ensure a balanced representation of the different behaviors captured by our classification. Human judgments were collected via the crowd-sourcing platform Prolific with three annotators per instance. Each annotator had to select one of the categories: *Correct & Equivalent*, *Correct – More Detailed*, *Correct – Incomplete/General*, *Correct – Alternative Answer*, *Incorrect or Unsupported*, *Can't Judge* (see details in Appendix B). Final human labels are obtained via majority voting across annotators. In cases of complete disagreement, instances are marked as conflicting.

Results in terms of Spearman correlation ρ between one-hot encodings of our interpretative categories and human judgment categories, computed over 60 selected instances from the Natural Questions dataset for Falcon-7B and Mistral-7B-Instruct, are presented in Figure 4. Both experiments show strong and statistically significant correlations between *Successful mimic* and *Correct & Equivalent*. In addition, *Generalization* correlates with *Correct – Incomplete/General*, while *Specialization* aligns with *Correct – More Detailed*, indicating that our taxonomy captures meaningful differences in the level of specificity of correct answers. Furthermore, *Clear mistake* is positively associated with *Incorrect or Unsupported*, reinforcing the interpretability of our categories. In contrast, our *Compatibility* and *NLI inconsistency* interpretative categories appear to span a broad and heterogeneous set of scenarios, making their relationship with human judgments less clear. While some significant correlations are observed in one experimental setting, these patterns are not consistent across experiments.

5 Conclusion

The proposed evaluation framework classifies the states of comparison between prediction and reference into interpretative categories thanks to bidirectional NLI. *Successful Mimic* category correlates with the reward signal of conventional baselines. Furthermore, *Successful Mimic*, *Specialization*, and *Generalization*, exhibit meaningful alignment with human judgments.

We view this work as a simple yet necessary step toward evaluating and ultimately rewarding cautious behavior in LLMs. In future work, we plan to leverage this framework as an automatic filtering mechanism to reduce the burden of human fact-checking.

Limitations

Multiple references To simplify our analysis, we restrict our setting to a single reference answer. Extending the framework to handle multiple references, common in open-ended QA, is left for future work.

Full abstention cases Our framework is not designed to properly handle full abstention responses (e.g., "I don't know"). Such predictions do not express comparable informational content with the reference, which limits the effectiveness of NLI-based comparison. As a result, they are not reliably captured by categories such as *Generalization* and would require dedicated detection mechanisms.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of nli models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Yanran Chen and Steffen Eger. 2023. Menli: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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

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.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. Nubia: Neural based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37.

Lucie Kunitomo-Jacquin, Abdelhak Imoussaten, Sebastien Destercke, Christophe Marsala, and Ken Fukuda. 2025. Towards evaluating set-valued predictions with partial observations. In *2025 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 1–4. IEEE.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wei Wu and Mark Last. 2025. Transitive self-consistency evaluation of nli models without gold labels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22637–22653.

Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. 2016. The costs of indeterminacy: how to determine them? *IEEE transactions on cybernetics*, 47(12):4316–4327.

Marco Zaffalon, Giorgio Corani, and Denis Mauá. 2012. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301. Imprecise Probability: Theories and Applications (ISIPTA'11).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A NLI results interpretation details

We provide interpretations of each class obtained from an NLI in Table 2.

B Additional Experiments Details

To obtain prediction answers, we employed top-K sampling with $K = 50$ and nucleus sampling with $p = 0.9$ at a temperature of $T = 0.1$ and provided 5-few shots to the models.

Prompt for LLM-as-a-judge (LLM judge)

*We are assessing the quality of answers to the following question: **{question}** \n The expected answer is: **{correct_answer}**. \n The proposed answer is: **{predicted_answer}** \n Within the context of the question, does the proposed answer mean the same as the expected answer? \n Respond only with yes or no.\n Response:*

Figure 5: **placeholders** are denoted in bold.

Datasets Our experiments were performed on four datasets. Two are general knowledge datasets, TriviaQA (Joshi et al., 2017) and SQuAD (Rajpurkar et al., 2018), one contain general domain questions from Google search, Natural Questions (NQ) (Kwiatkowski et al., 2019), and one dataset contains natural language math problems (SVAMP) (Patel et al., 2021). In case of several references answers we use only the first one.

Models We select a set of varied systems allowing to observe different behavior profiles. We select 2 model sizes for each family where available, and test both the base and instruction-tuned versions. Concretely, we work with LLama 2 models (Touvron et al., 2023) with 7B and 13B parameters (*Llama-2-7b*, *Llama-2-7b-chat*; *Llama-2-13b*, *Llama-2-13b-chat*), the Falcon models (Almazrouei et al., 2023) with 7B parameters (*falcon-7b*, *falcon-7b-instr*, and Mistral models (Jiang et al., 2023) with 7B parameters (*Mistral-7B-v0.1*, *Mistral-7B-Instr*).

Baselines

- Exact Match (EM) is a binary metric indicating whether the predicted answer exactly matches the reference after standard normalization (Rajpurkar et al., 2016).

Class	Entailment	Contradiction	Neutrality
Notation	P *entails* H	P *contradicts* H	P *neutral* H
Interpretation	P being true entails H must be true.	P being true entails H must be false.	P being true does not allow to conclude whether H is true or false.

Table 2: NLI classification output results interpretations. P and H denote the premise and hypothesis.

- F1 is the harmonic mean of precision and recall computed over normalized token overlap between prediction and reference (Rajpurkar et al., 2016).
- ROUGEL (RecallOriented Understudy for Gisting Evaluation) measures recall based on the longest common subsequence between prediction and reference (Lin, 2004)
- BLEU (Bilingual Evaluation Under study) computes n-gram precision between prediction and reference(Papineni et al., 2002)
- BERTscore estimates semantic similarity using contextualized token embeddings from BERT-like models (Zhang et al., 2019)
- LLM judge: A language model here, *gpt-4o-mini* (OpenAI), is prompted to decide whether the predicted answer is semantically equivalent to the reference via a binary judgment. See prompt presented in Figure 5.

We compute EM and F1 using a SQuAD-style implementation, ROUGEL with `rouge-score`, BLEU with `SacreBLEU` (Post, 2018), and BERTScore with the `bert-score` package (Zhang* et al., 2020).

Licenses The datasets used in this paper are released under Apache 2.0 (TriviaQA), CC BY-SA 4.0 (SQuAD), MIT (SVAMP), and CC BY-SA 3.0 (NQ).

Human Annotation Details Annotators were instructed to selection one of the category bellow following instructions presented in Figure 6:

- *Correct & Equivalent:* The answer is correct and essentially the same as the reference.
- *Correct – More Detailed:* The answer is correct and provides more detail than the reference (extra info is accurate).

In this study, you will evaluate answers generated by an AI system.

For each item, you will be shown:

- a question,
- a reference answer,
- an AI-generated answer

Your task is to judge whether the AI answer is correct, complete, more specific or appropriate with respect to the reference.

You will select the most appropriate category from a predefined list.

If you are unsure, you may **briefly** (no more than 2 minutes) consult **reliable external sources** (e.g., Wikipedia or other reputable websites) to verify factual information. If you are still unsure after searching select "**can't judge**" option.

Important instructions:

- Please rely on your own understanding and judgment.
- Do NOT use AI tools (e.g., ChatGPT or similar systems) to make your judgment.
- Complete all the 60 items.
- Your responses may be reviewed for quality and consistency.
- The task requires careful reading and thoughtful evaluation, but no specialized expertise is required.

Figure 6: Annotation instruction

- *Correct – Incomplete/General:* The answer is correct but less specific or incomplete compared to the reference.
- *Correct – Alternative Answer:* The answer is different from the reference but is still a valid and correct answer to the question.
- *Incorrect or Unsupported:* The answer is incorrect (contains wrong or unsupported information, or doesn't answer the question).
- *Can't Judge:* Not enough information or unclear, so you cannot decide if the answer is correct.

C Additional results

Figure 7 provides additional insight into the comparison with the LLM-as-a-judge baseline by showing the conditional probability that the LLM judge considers an answer correct given our classification. The results show a high probability of the LLM judge predicting correctness for answers classified as *Specialization*, with $P(\text{LLM judge} = 1 \mid \textit{Specialization}) = 0.68$ suggesting that it accepts plausible additional details beyond the reference.

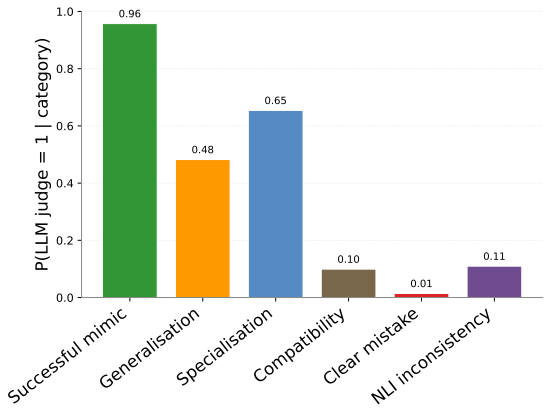


Figure 7: Conditional probability that the LLM judge considers an answer correct given our classification (computed over 14,400 instances).

We provide in Table 3, examples of question instances with their associated reference answers, predictions answer generated with Falcon-7B, bidirectional NLI classification results, LLM-judge scores, and human annotations majority vote results.

Question	Reference answer	Predicted answer	Bidirectional NLI classification	LLM judge	Human majority
who does dwyane wade play for right now?	the Miami Heat	Miami Heat	Successful mimic	1.0	Correct & Equivalent
who played john coffey in the movie the green mile?	Michael Clarke Duncan	Tom Hanks	Clear mistake	0.0	Incorrect or Unsupported
when does middle-earth shadow of mordor come out?	September 2014	September 24, 2014	Specialisation	1.0	conflicting
when was the last person executed in florida?	Eric Scott Branch	2010	NLI inconsistency	0.0	Incorrect or Unsupported
when did the great fire of london end?	5 September 1666	1666	Generalisation	0.0	conflicting
who created the convention on the rights of the child?	The UN General Assembly	the United Nations	Generalisation	1.0	conflicting
when did this season of american idol start?	March 11, 2018	2018	Generalisation	0.0	Correct – Incomplete/General
the first vice president of india who become the president letter was?	Sarvepalli Radhakrishnan	Dr. Sarvepalli Radhakrishnan	Specialisation	1.0	conflicting
who is the minister of local government in zimbabwe?	Hon July Moyo	July Moyo	Generalisation	1.0	Correct & Equivalent
created propaganda to support the war effort & encourage the sale of war bonds?	the government	the Office of War Information	Specialisation	0.0	conflicting
who discovered the atom and in what year?	early 1800s	J.J. Thompson	Compatibility	0.0	conflicting
when does the next season of wynonna earp start?	2018	June 12th	Compatibility	0.0	conflicting

Table 3: Few instances from Natural Questions dataset.