Detecting correct answers to open questions and its impact on language models' confidence scores

Guido Ivetta
Fundación Vía Libre
Universidad Nacional de Córdoba, Argentina

Hernán J. Maina
CONICET
Universidad Nacional de Córdoba, Argentina

Luciana Benotti
CONICET
Universidad Nacional de Córdoba, Argentina

Abstract

In this paper, we study how different methods for classifying correct answers to open questions impact calibration scores in language models. We compare seven different techniques to perform this task. In this setting, we find that even though answer classification techniques have up to 21% differences between each other, calibration scores are not affected significantly.

We find these results show evidence of unreliability on commonly used metrics in this field.

1 Introduction

Evaluating question answering in natural language is a challenging task because different answers can be correct. In this paper we extend the paper (Cole et al., 2023) by addressing a limitation in the research evaluation methodology, which considers an answer correct only if it exactly matches the gold standard answer. Let us illustrate the limitation. Suppose the model is prompted with the question *'What voltage is the common zinc-carbon or alkaline AAA battery?'* and it generates the answer *'1.5 volts'* which is classified as erroneous because the gold standard answers are *'1.5v'* and *'1,5'*.

We employ seven distinct techniques to assess correct answers, comparing them with each other and with human annotation, providing a more comprehensive evaluation framework. This approach allows for a nuanced understanding of model performance, considering variations in acceptable responses and discovering its impact on results.

2 Previous work

It is usual to find a list of accepted correct answers in question answering (QA) datasets, this poses a challenge when evaluating outputs of generative models. This problem is acknowledged in (Voorhees and Tice, 2000): *"it is quite difficult to determine automatically whether the difference between a new string and a judged string is significant with respect to the correctness of the answer."*

Automatic evaluation of question answering frequently use two metrics at the token level: Exact Match (EM) and Token F1 (F1). (Bulian et al., 2022) describe the ways they both fall short of capturing the difference between *significant and insignificant span differences*. Both techniques imperfectly capture the answer equality and can over or underestimate the performance of models, this can be seen in (Kocmi et al., 2021), (Gehrmann et al., 2021), (Chen et al., 2019) and (Chen et al., 2020). Bulian et al., notes that one obvious limitation of token-level measures is their direct dependence on the diversity of the reference answers collected for the dataset (Chen et al., 2019). Bulian claims this could be addressed by extending the annotations, but this is both expensive and has diminishing returns as the true collection of all correct answers might be large.

Experiments comparing the accuracy of different answer classification techniques can be found in (Risch et al., 2021), using the SQuAD Dataset (Rajpurkar et al., 2016), GermanQuAD (Möller et al., 2021), and NQ-open (Kwiatkowski et al., 2019). More information on evaluation, specially in open questions can be found in (Honovich et al., 2021), (Honovich et al., 2022), (Eyal et al., 2019), (Fabbri et al., 2022), (Schuster et al., 2021).

3 Methodology

In all of our experiments, we worked with 4-bit-quantized Falcon-7b (Almazrouei et al., 2023), an open-source lightweight language model. We opted to use this model due to its popularity and availability. The full 40B parameter version of the Falcon model topped the OpenLLM Leaderboard in HuggingFace in June of 2023 (Delangue, 2023).

```
Question: What does a manometer measure?
Answer: Pressure.

Question: Who was Pope during World War Two?
Answer: Pius XII.

...

Question: Who is the director of Scarface?
Answer:
```

Figure 1: Section of a 4-Shot prompt used to perform the experiments. The example QA pairs are selected randomly so that each input to the model has a different version of the prompt.

3.1 Confidence Scores

The confidence scores computed in our experiments are the ones proposed in (Cole et al., 2023) and were validated by one of the authors. Following this, a brief description of each is provided.

• **Likelihood**: This is the product of the log probabilities for the generated sequence.

• **Sampling Repetition**: The fraction of times that the sampled outputs match the greedy output

• **Sampling Diversity**: This score is inversely proportional to the number of distinct samples and is estimated as 0 if all samples are different, computed by Formula 1.

$$1 - \frac{num\_unique}{num\_samples} \quad (1)$$

3.2 Evaluation Metrics

For evaluation metrics, we again replicated (Cole et al., 2023) calculating *Expected Calibration Error (ECE)*, *ROC-AUC* and *Cov@Acc*. More details on the section "Evaluation Setup" of the that paper.

3.3 Correct Answer Classification

We compare seven different techniques to classify correct answers. We selected these options due to their widespread adoption and its use in a similar experiment in (Risch et al., 2021). For the **human annotation**, an annotator was given the question, possible answers and the model's answer and was asked to classify the answer as correct or incorrect. A second annotator also performed this task on 10% of the dataset to calculate inter-annotator agreement. The annotations had 97% of agreement, resulting in a Cohen's Kappa coefficient of 0.94. In this metric's terms, this is classified as almost

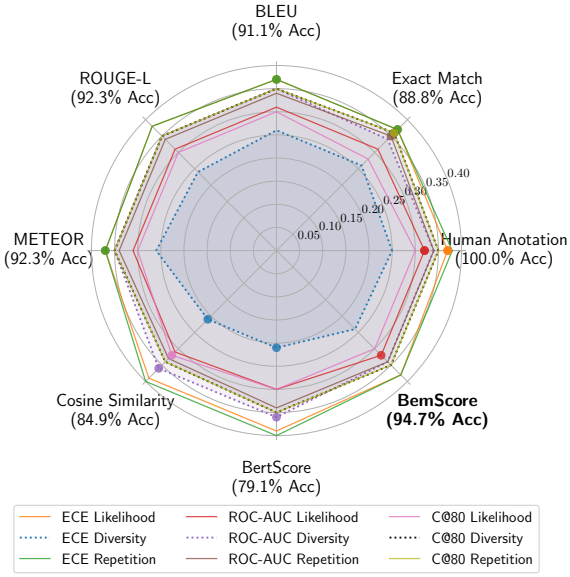

Figure 2: Radar chart of all considered techniques to classify correct answers. The accuracy value displayed under each technique was computed comparing it to human annotation. All evaluation metrics (ECE, ROC-AUC and C@80) are displayed across all confidence score methods (Likelihood, Diversity, Repetition). ECE is better when smaller, ROC-AUC and C@80 are better when higher. The vertices with the best performance in each metric are highlighted with a circle. Almost all metrics' graphs resemble a regular octagon, meaning their performance do not vary significantly across answer classification technique.

perfect agreement. The automatic techniques are: **BLEU**, **ROUGE-L**, **METEOR**, **Cosine Similarity**, **BertScore** and **BemScore** (Bulian et al., 2022). BemScore is a fine-tuned BERT model to classify answer equivalence using the SQuAD Dataset (Rajpurkar et al., 2016).

4 Datasets

Due to computing power limitations, in this work we focused on a subset of the datasets used on (Cole et al., 2023). We randomly sampled 1000 question-answer pairs from the TriviaQA dataset (Joshi et al., 2017).

5 Experiment Setup

As the setup of (Cole et al., 2023) described, we focus on few-shot in-context learning. Specifically, the prompt is composed of four question and answer pairs from the training set of TriviaQA. To reduce variance across experiments, the example QA pairs are selected randomly so that each input to the model has a different version of the prompt. In Figure 1, a partial example is shown.

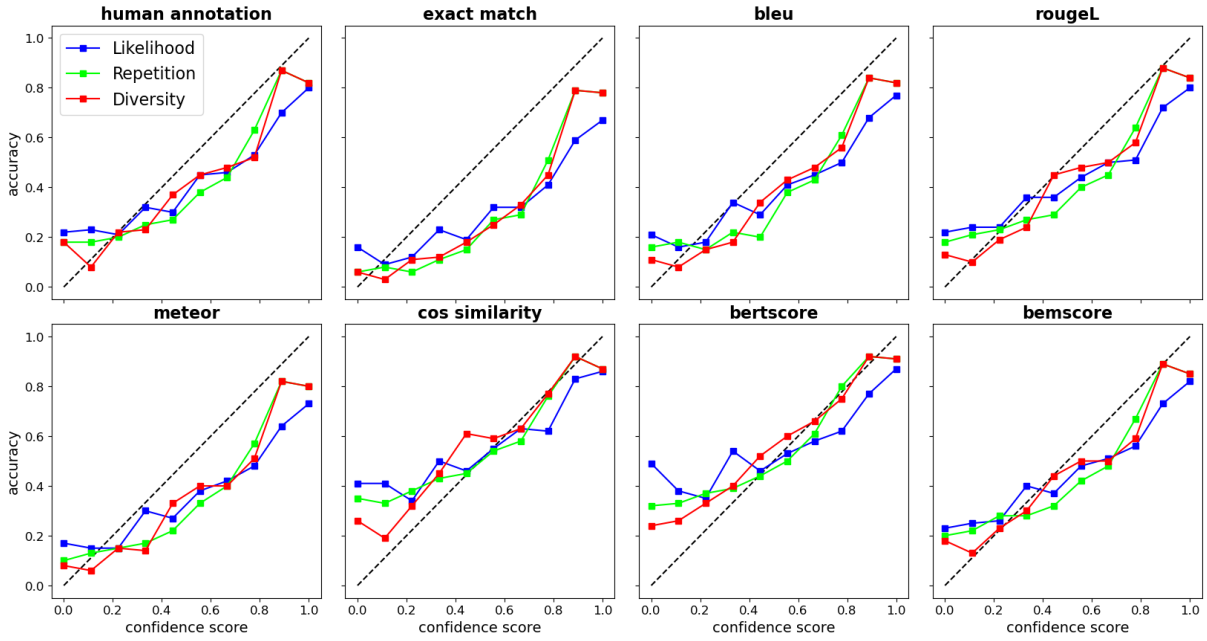

Figure 3: Plot of calibration error by comparing bucketed accuracy to bucketed confidence scores across methods and techniques, using TriviaQA as dataset. Each figure represents a technique analyzed to classify correct answers. In every category the results of the Likelihood, Repetition and Diversity methods are tested, where following the diagonal is perfect performance.

6 Results

In Figure 2 we can observe accuracies compared to human performance for each technique. *Exact-match*, the only one used in (Cole et al., 2023), achieved an accuracy of 88.8%. *BemScore*, the top scorer reached an accuracy of 94.7%. *BertScore* performed the worst, at 79.1% of human accuracy. The top and lowest scorer have a relative difference of 19.7%. The standard deviation between accuracies across techniques is 0.054.

Calibration metrics results can be found in Figure 2. Almost all metrics' graphs resemble a regular octagon, meaning their performance do not vary significantly across techniques. On average, the difference between the lowest score and the highest score on each metric-method pair differ 9%. The average standard deviation was 0.014.

In Figure 3, bucketed accuracy to bucketed confidence can be observed. A perfect score is represented by the diagonal, where wrong answers have low confidence scores and correct answers have high confidence scores. We find some differences between answer classification techniques: **Exact Match**, **METEOR**, and **BLEU** lie mostly below the diagonal, meaning that confidence scores usually are higher than the proportion of correct answers according to this method. This means confidence scores usually overstate their value. On the other hand, **Cos Similarity** and **BertScore** lie mostly above the diagonal, understating their value.

7 Conclusions

We present an analysis of techniques of classifying correct answers in a QA context, and their impact on calibration scores. More specifically using the TriviaQA dataset (Joshi et al., 2017).

Differences in accuracy related to human classification were found. However, we found consistency across almost all tested calibration scores. As discussed in Section 6, techniques mostly over- or undervalue their calibration score, however our metrics gave similar results for both. This could mean that the metrics used don't differentiate between both errors. These results shed light on the unreliability of the accepted metrics in the area. If an experiment was conducted using **BertScore** or **BemScore** as the only answer classification technique, similar conclusions would be reached on confidence metrics by the researchers while having 19.7% of relative error between them. We recommend not to take task of classifying correct answers lightly, and employ human annotation to assess the errors of the automatic technique used.

We hope this study will contribute to further development of better metrics and improve the evaluation and usability of QA systems.

Limitations

Reproducing research conducted with large, resource-intensive language models using limited computing power presents a significant challenge. Our available hardware constrained the scale and complexity of our experiments. This constraint affected the size of the model, the training data, and the efficiency of training times. While we aimed for faithful replication, these resource limitations influenced the extent to which we could emulate the conditions of (Cole et al., 2023).

Additionally, the closed-source nature of the original language model posed inherent limitations. This lack of transparency hindered our ability to fully understand the architecture of the model originally used, potentially restricting our capacity to address certain research aspects comprehensively.

We ran our experiments on only one dataset, a more in-depth study could be performed when analyzing different contexts and answer types. All questions were non-ambiguous, one of the main focuses of (Cole et al., 2023) was the impact of ambiguity on calibration. Adding this layer to our study could improve our understanding on the subject.

The development of a metric that would be able to mitigate the challenges discussed in this paper was not discussed and is left as future work.

Acknowledgments

We thank Julian Martín Eisenschlos for his valuable comments and feedback while discussing this work. We are also very grateful to Latinx in AI (LXAI) and the reviewers for the opportunity to share our ideas in this Workshop

This work used computational resources from CCAD – Universidad Nacional de Córdoba (https://ccad.unc.edu.ar/), which are part of SNCAD – MinCyT, República Argentina. it was also supported by the computing power of Nodo de Cómputo IA, from Ministerio de Ciencia y Tecnología de la Provincia de Córdoba in San Francisco - Córdoba, Argentina.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

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.

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.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Jeremy Cole, Michael 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*, pages 530–543, Singapore. Association for Computational Linguistics.

Clément Delangue. 2023. Open llm leaderboard - a hugging face space by huggingfaceh4.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv

Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. 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*, Vancouver, Canada. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, 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*, 7:453–466.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. 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, Austin, Texas. Association for Computational Linguistics.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 200–207, New York, NY, USA. Association for Computing Machinery.