TRUSTSCORE: REFERENCE-FREE EVALUATION OF LLM RESPONSE TRUSTWORTHINESS

Danna Zheng[1], Danyang Liu[1], Mirella Lapata[1], Jeff Z. Pan[1,2]
[1] School of Informatics, University of Edinburgh, UK
[2] Huawei Edinburgh Research Centre, CSI, UK
{dzheng, danyang.liu}@ed.ac.uk, mlap@inf.ed.ac.uk
http://knowledge-representation.org/j.z.pan/

ABSTRACT

Large Language Models (LLMs) have demonstrated impressive capabilities across various domains, prompting a surge in their practical applications. However, concerns have arisen regarding the trustworthiness of LLMs' outputs, particularly in closed-book question-answering tasks, where non-experts may struggle to identify inaccuracies due to the absence of contextual or ground truth information. This paper introduces `TrustScore`, a framework based on the concept of Behavioral Consistency, which evaluates whether an LLM's response aligns with its intrinsic knowledge. Additionally, `TrustScore` can seamlessly integrate with fact-checking methods, which assesses alignment with external knowledge sources. The experimental results show that `TrustScore` achieves strong correlations with human judgments, surpassing existing reference-free metrics, and achieving results on par with reference-based metrics.

1 INTRODUCTION

Large-scale language models (LLMs) have recently been in the spotlight due to their impressive performance in various NLP tasks, sparking enthusiasm for potential applications (Kaddour et al., 2023; Bubeck et al., 2023). However, a notable concern has emerged regarding the ability of LLMs to generate plausible yet incorrect responses (Tam et al., 2022; Liu et al., 2023; Devaraj et al., 2022), particularly challenging for users without specialized expertise. Consequently, users are often advised to employ LLMs in scenarios where they can confidently assess the information provided.

This concern is particularly salient in closed-book question-answering tasks (Roberts et al., 2020; Su et al., 2023), where the absence of contextual cues complicates the evaluation process. Unlike tasks such as reading comprehension (Yao et al., 2023), summarization (Aharoni et al., 2023), and natural language inference (Chen & Eger, 2023), where answers can be verified against provided context, assessing the trustworthiness of an LLM's response in a closed-book setting poses a greater challenge. In such settings, LLMs rely solely on their parametric knowledge (Pan et al., 2023) to generate answers, making it difficult to determine whether these responses truly align with their underlying knowledge due to the black-box nature of LLMs. Traditional approaches like fact-checking, which rely on evidence from external knowledge bases (Pan et al., 2017a;b), may be hindered by inaccessible or insufficient information, rendering them ineffectively.

To address these challenges, this paper introduces `TrustScore`, a framework centered around the concept of *Behavioral Consistency*, which evaluates whether an LLM's response aligns with its parametric knowledge by assessing whether the LLM makes consient choices among its responses and other distractors. Behavioral consistency serves as a necessary condition for ensuring an LLM's responses are in alignment with its parametric knowledge, enhancing the likelihood of trustworthiness. Furthermore, `TrustScore` can seamlessly integrate with fact-checking methods, offering a more holistic evaluation of an LLM's response. We evaluate the effectiveness of `TrustScore` on FLAN-T5 (Chung et al., 2022), LLaMA (Touvron et al., 2023), and GPT-3.5, showing strong correlation with human judgments and outperforming existing reference-free [1] metrics while achieving performance comparable to reference-based metrics.

---

[1]Reference in this paper refers to gold answers.

2 PROPOSED METHOD

Given an LLM denoted as M, a context-free question represented as q, and the corresponding response generated by LLM M for question q denoted as a, the objective of `TrustScore` is to provide a relative score to show to what extent can the response be trusted. The resulting `TrustScore` falls within the range [0, 1]. A higher score means the response is more likely to be trusted.

The key component of `TrustScore` is a Behavioral Consistency (Trust$_{BC}$) evaluator as illustrated in Figure 1, which assesses whether the responses provided by the LLM are consistent with its intrinsic parametric knowledge. This evaluator is designed to function independently, particularly useful in situations where external knowledge bases are not accessible, or there is not enough information available. Moreover, Trust$_{BC}$ can integrate with a Fact-Checking (Trust$_{FC}$) module to further assess trustworthiness using an external knowledge base when one is available. This dual approach allows for a comprehensive assessment of an LLM's response, ensuring it is evaluated both for internal consistency and external factual consistency.
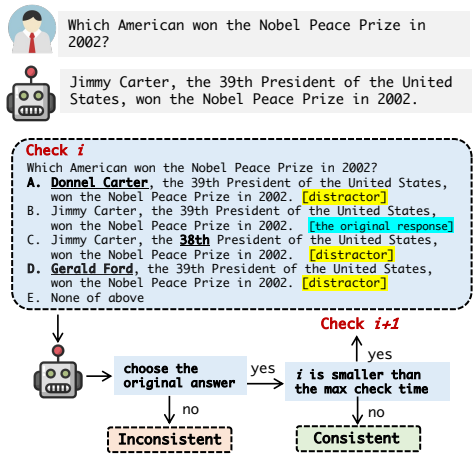

Figure 1: An illustration of Behavioral Consistency Evaluator (Trust$_{BC}$). LLM responses pass through Trust$_{BC}$ to determine if the model selects its original answer consistently across multiple iterations.

2.1 TRUST$_{BC}$ STAND-ALONE

The Behavior Consistency Evaluator aims to determine whether the LLM M consistently selects response r when presented with distractors $R = \{r_1, r_2, r_3, ...\}$. We set a maximum check limit of n times, where LLM M is tasked with choosing the correct option from five choices (including response r, three random distracting options, and 'None of the above') in each attempt. Note that the distracting options are different across different attempts. If LLM M does not choose response r in any of the attempts, it implies that it does not consistently provide the response to question q in line with its parametric knowledge. Ideally, if we conduct an infinite number of checks and LLM M consistently selects response r, it implies that response r is genuinely rooted in the LLM's underlying knowledge. The behavior check process terminates when inconsistent behavior occurs (LLM M does not select response r) or when it reaches the maximum check limit n. If inconsistent behavior occurs, Trust$_{BC}$ return 0; otherwise, if inconsistent behavior does not occur up to the maximum check limit n, Trust$_{BC}$ return 1.

To generate high-quality distractors, we propose a priority-based substitution algorithm:

Words necessitating substitution are prioritized based on their informativeness. (1) Words exclusively present in response r are given higher priority than those occurring in both response r and question q; (2) Entities take precedence over other words; (3) Nouns and numbers are prioritized over verbs and adjectives, which, in turn, take precedence over words of other parts of speech.

Substitute words are prioritized based on their sources. (1) DBpedia entities filtered by types are top priority; (2) words filtered by embedding similarity have the second-highest priority; (3) words randomly sampled from the vocabulary filtered by POS tags are the lowest priority.

Details of the priority-based substitution algorithm are in Appendix A.5.1. Examples of the generated multi-choice questions are provided in A.6.

2.2 INTEGRATION OF TRUST$_{BC}$ AND TRUST$_{FC}$

When external knowledge bases are available, Trust$_{BC}$ can collaborate with Trust$_{FC}$ for a thorough evaluation of responses. The role of Trust$_{FC}$ is to verify if a response r to a question q aligns with facts from external knowledge bases. This project does not seek to pioneer new fact-checking methods. We adopt a widely accepted approach of extracting relevant evidence from knowledge bases and assessing its support for the response r through an entailment model. Trust$_{OV}$ combines

scores from Trust_{BC} and Trust_{FC} based on criteria detailed in Appendix A.5.3. This integration is guided by several principles: (1) Responses supported by external knowledge are deemed more trustworthy than those contradicted by it. (2) A response insisted upon by LLM is more trustworthy than the response not insisted upon by LLM. (3) Factual consistency is prioritized over behavioral consistency.

3 EXPERIMENT SETUP

Data Collection We have created a composite dataset, MixedQA, for the purpose of analyzing the correlation between `TrustScore` and human evaluations. It consists of 1,000 open-ended questions randomly selected from the test sets (or development sets when test sets are unavailable) of five datasets: Natural Questions (Kwiatkowski et al., 2019), WebQuestions (Berant et al., 2013), TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), and PopQA (Mallen et al., 2022). We ignore the support context in these datasets to fit the closed-book setting. The detailed selection criteria are in the Appendix A.3.

LLMs We test our `TrustScore` on the answer generated by FLAN-T5-XL, LLaMA-7B, and GPT-3.5-Turbo. LLMs perform the question-answering task in a few-shot setting. To fully exploit `TrustScore`, we prompt LLM to output response in long sentence instead of short phrase. The implementation detail including the prompts and hyper-prameters are provided in Appendix A.5.

Baseline Metrics We employ the following baselines: reference-based metrics, reference-free metrics with and without evidence retrieval. Reference-based metrics include popular lexical matches like Exact Match, BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee & Lavie, 2005), and pretrained evaluators such as BERTScore (Zhang et al., 2019), BARTScore (Yuan et al., 2021), and GPTScore[2] (Fu et al., 2023). Reference-free metrics include BARTScore and GPTScore, both with and without evidence retrieval, for a fair comparison with our `TrustScore`.

Evaluation To validate the effectiveness of `TrustScore`, we calculate the correlation between `TrustScore` and human judgments on the MixedQA dataset. Humans annotators applied binary judgments to categorize the 3,000 answers generated by LLMs as "correct" or "incorrect." We initiated this with a pilot annotation of fifty examples, achieving a strong agreement score of 0.89. Detailed information about the annotation process is provided in Appendix A.4.

4 RESULTS

Table 1 shows Pearson's r correlations between various metric scores and human judgments for answers produced by LLMs.

Behavioral Consistency Analysis In experiments where retrieved evidence is unavailable, our Trust_{BC} demonstrates the highest correlation with human judgments, outperforming other reference-free metrics. Remarkably, Trust_{BC} also outperforms some reference-based metrics when evaluating responses from GPT-3.5. This

Table 1: Pearson's r correlation for baseline metrics and `TrustScore` against human judgments in MixedQA. *NAN* indicate uncomputable correlations when all exact match scores are 0. Unless underlined, all reported correlations are statistically significant at $p < 0.01$.

Metrics	FLAN-T5	LLaMA	GPT-3.5
Reference-based Metrics			
Exact Match	0.635	*NAN*	*NAN*
BLEU-1	0.655	0.435	0.289
BLEU-4	0.536	0.376	0.228
ROUGE-L	0.607	0.533	0.382
METEOR	0.603	0.536	0.106
BERTScore	0.573	0.266	0.087
BARTScore-ref	0.570	0.474	0.368
GPTScore-ref	0.610	0.575	0.453
Reference-free Metrics (w/o Evidence Retrieval)			
BARTScore-src	-0.075	-0.099	-0.151
GPTScore-src	-0.055	-0.081	-0.139
Trust_{BC} (Ours)	0.313	0.133	0.353
Reference-free Metrics (w/ Evidence Retrieval)			
BARTScore-evid	0.070	0.091	-0.013
GPTScore-evid	0.075	0.186	0.024
Trust_{FC}	0.593	0.533	0.449
Trust_{OV} (Ours)	0.613	0.539	0.484

is particularly significant given that, in many real-world scenarios, relevant evidence is unavailable or insufficient, making reference-based or retrieval-based metrics inapplicable. In such cases, our

---

[2]GPTScore includes several backbones, and we chose FLAN-T5-base in this work.

$Trust_{BC}$ can effectively assess response trustworthiness. Comparing the $Trust_{BC}$ scores across different LLMs, we observe that the $Trust_{BC}$ score for LLaMA is lower than those for FLAN-T5 and GPT3.5. This difference could be attributed to LLaMA being a base model, while FLAN-T5 is fine-tuned with instructions and GPT-3.5 is fine-tuned with both instructions and human feedback. These fine-tuning processes may enhance the behavioral consistency of the latter models.

Integration of $Trust_{BC}$ and $Trust_{FC}$ In experiments where retrieved evidence is available, $Trust_{OV}$ demonstrates human correlations that surpass those of existing reference-free metrics and even outperform most reference-based metrics. We note that the $Trust_{OV}$ score consistently outperforms $Trust_{FC}$ across various LLMs. This observation substantiates our assertion that $Trust_{BC}$ is not only an effective reference-free metric on its own but can also complement $Trust_{FC}$ to achieve better performance in a reference-based setting.

Robustness to Diverse Answers   From Table 1, we observe that traditional lexical overlapping-based metrics (e.g., Exact Match, BLEU, and ROUGE, etc.) achieve higher correlation scores for responses generated by FLAN-T5, but lower scores for answers generated by GPT3.5. This discrepancy can be attributed to FLAN-T5's tendency to produce concise answers. Some of these answers match reference answers exactly, making them easier to evaluate using reference-based metrics. Conversely, GPT-3.5 tends to generate longer and more diverse answers, which may deviate from reference answers, making lexical-based metrics less effective. In contrast to these lexical overlapping-based metrics, `TrustScore` excels in evaluating GPT-3.5-generated responses and also achieves comparable results for FLAN-T5-generated responses. This highlights its robustness across different types of answers.

5   RELATED WORK

QA Evaluation Metrics   Traditional metrics such as Exact Match, F1, BLEU (Papineni et al., 2002), Rouge (Lin, 2004), and METEOR (Banerjee & Lavie, 2005), rely on lexical matching, often overlooking semantic content. BERTScore (Zhang et al., 2019), BARTScore (Yuan et al., 2021) and GPTScore(Fu et al., 2023) addresses this limitation by leveraging contextualized embeddings to capture semantic similarity. However, these methods are predominantly designed for reference-based evaluation. While BARTScore and GPTScore can also be applied in a reference-free manner by directly comparing generated text to source sentences, their effectiveness is noticeably context-sensitive especially in QA evaluation due to substantial divergence between answers and questions.

Behavior Consistency   Jang et al. (2022) introduced the BECEL benchmark, evaluating language models across tasks like natural language inference and semantic analysis, highlighting the importance of behavioral consistency. Asai & Hajishirzi (2020) and Ribeiro et al. (2019) improved consistency in question-answering by enriching training data with logical and linguistic insights. Cohen et al. (2023) proposed the LM vs LM framework for factuality evaluation through cross-examination between LMs, using multi-turn interactions to uncover inconsistencies. However, this model relies on the subjective judgment of the examiner LM, introducing potential bias. Our approach contrasts by offering an objective mechanism for identifying inconsistencies without the constraints of LM vs LM, such as needing LMs with over 10 billion parameters and reliance on instruction-folowing and long-context reasoning ability. Additionally, unlike LM vs LM, which depends solely on consistency for correctness, our method integrates external knowledge for a more comprehensive evaluation.

6   CONCLUSION

We introduce `TrustScore`, a novel reference-free evaluation framework that assesses the trustworthiness of LLM responses. Our experiments, conducted on our newly curated MixedQA dataset, demonstrate effectiveness of the `TrustScore` framework. `TrustScore` exhibits a strong correlation with human judgments, surpassing existing reference-free metrics, and achieves results comparable to reference-based metrics. As for future work, we might look into how to combine `TrustScore` with model editing methods (Han et al., 2023).

REFERENCES

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3562–3591, 2023.

Akari Asai and Hannaneh Hajishirzi. Logic-guided data augmentation and regularization for consistent question answering. *arXiv preprint arXiv:2004.10157*, 2020.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*, 2023.

Ashwin Devaraj, William Sheffield, Byron C Wallace, and Junyi Jessy Li. Evaluating factuality in text simplification. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2022, pp. 7331. NIH Public Access, 2022.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

Xiaoqi Han, Ru Li, Hongye Tan, Wang Yuanlong, Qinghua Chai, and Jeff Z. Pan. Improving Sequential Model Editing with Fact Retrieval. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP 2023)*, pp. 11209–11224, 2023.

Zhiwei Hu, Victor Basulto, Zhiliang Xiang, Ru Li, and Jeff Z. Pan. Transformer-based Entity Typing in Knowledge Graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pp. 5988–6001, 2022.

Zhiwei Hu, Victor Basulto, Zhiliang Xiang, Ru Li, and Jeff Z. Pan. Multi-view Contrastive Learning for Entity Typing over Knowledge Graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 12950–12963, 2023.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. Becel: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3680–3696, 2022.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022.

J. Z. Pan, G. Vetere, J.M. Gomez-Perez, and H. Wu (eds.). *Exploiting Linked Data and Knowledge Graphs for Large Organisations*. Springer, 2017a.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, , and Damien Graux. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Special Issue on Trends in Graph Data and Knowledge. Transactions on Graph Data and Knowledge (TGDK)*, 1:1–38, 2023.

J.Z. Pan, D. Calvanese, T. Eiter, I. Horrocks, M. Kifer, F. Lin, and Y. Zhao. *Reasoning Web: Logical Foundation of Knowledge Graph Construction and Querying Answering*. Springer, 2017b.

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

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6174–6184, 2019.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Dan Su, Mostofa Patwary, Shrimai Prabhumoye, Peng Xu, Ryan Prenger, Mohammad Shoeybi, Pascale Fung, Animashree Anandkumar, and Bryan Catanzaro. Context generation improves open domain question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 781–796, 2023.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. Evaluating the factual consistency of large language models through summarization. *arXiv preprint arXiv:2211.08412*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

Zijun Yao, Yantao Liu, Xin Lv, Shulin Cao, Jifan Yu, Juanzi Li, and Lei Hou. Korc: Knowledge oriented reading comprehension benchmark for deep text understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11689–11707, 2023.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

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

A APPENDIX

A.1 LIMITATIONS

Due to the high cost of collecting human evaluation as gold labels and executing LLMs like GPT-3.5, which we include in our experiments, we do not test `TrustScore` on large-scale data. Instead, we test it on MixedQA, which consists of 1000 open-ended questions randomly collected from five question-answering datasets.

In this work, we do not test `TrustScore` on long-form questions that require elaborate and in-depth answers. We contend that our general framework can be adapted to the long-form question-answer task. We will leave it in our future work.

A.2 ETHICS STATEMENT

The MixedQA dataset is collected from several datasets that are public for academic use and do not contain sensitive information. We will make the dataset and the codebase of the project freely available online for academic use.

Our human evaluation involves human participation. Two annotators were recruited and trained. They were asked to annotate the correctness of answers and filter out samples that might cause ethical problems. No sensitive personal information is involved in the process.

A.3 DATA SELECTION CRITERIA FOR MIXEDQA

We collected MixedQA from a diverse set of five datasets:

Natural Questions (Kwiatkowski et al., 2019)   This dataset consists of questions extracted from web queries, with each question accompanied by a corresponding Wikipedia article containing the answer.

WebQuestions (Berant et al., 2013)   This dataset comprises questions originating from web queries, meticulously matched with entries in FreeBase.

TriviaQA (Joshi et al., 2017)   This dataset incorporated a collection of questions sourced from Quiz League websites. Each question in this dataset is complemented by web pages and Wikipedia searches that may potentially contain the answer.

HotpotQA (Yang et al., 2018)   This dataset primarily contains Wikipedia-based multi-hop questions, with sentence-level supporting facts annotation.

PopQA (Mallen et al., 2022)   This dataset is composed of questions aimed at covering factual information in the less frequently explored areas, essentially addressing the long tail of information.

We extracted questions exclusively from the test or development sets, discarding the supporting context to adapt to a closed-book setting. Our data collection process involved the following steps:

1. Open-Ended Questions Extraction: Initially, we utilized an automated script to filter out comparison questions (e.g.,"Who was born earlier, Emma Bull or Virginia Woolf?") and "yes or no" questions (e.g., "Is Ferocactus a type of plant?") in order to focus on open-ended questions.

2. Input to Language Models: We input the filtered questions, devoid of any contextual information, to three language models: FLAN-T5-XL, LLaMA-7B, and GPT-3.5-Turbo.

3. Answerable Questions Extraction: We observed that FLAN-T5 occasionally generated sub-par answers, characterized by repetitive words or even blank responses. LLaMA-7B, on the other hand, sometimes failed to provide a relevant answer and instead repeated the original question. ChatGPT's responses, meanwhile, often fell into the category of inconclusive answers, with statements like "I am not sure," "I don't know," or "The information is not

provided in the question." Such responses lacked the necessary informative content required for trustworthiness verification. To address this issue, we employed an automated script to filter out questions with such responses, retaining only those questions that could be answered by the three LLMs.

4. Random Sampling: Finally, to ensure a balanced and representative dataset, we randomly selected 200 questions from each of the five datasets for `TrustScore` evaluation.

A.4 HUMAN EVALUATION DETAILS

The 1,000 questions are evenly distributed between two annotators. Each question accompanied by answers from three LLMs: FLAN-T5, LLaMA, and GPT-3.5. These questions were also paired with ground truth answers from the original dataset. During our pilot study, we discovered a number of incorrect ground truth answers. As a result, annotators were advised to use these answers as a guide and to verify the correctness of answers via manual internet searches.

System	Cohen's κ
FLAN-T5	0.88
LLaMA	0.82
GPT	0.92
Overall	0.89

Table 2: Inter-annotator agreement for the double annotations on fifty examples.

We evaluated the reliability of human annotations by measuring the agreement between annotators. Table 2 shows the results of the Cohen's κ Cohen (1960), which accounts for agreement that is expected by chance. The annotations on answers of FLAN-T5, LLaMA, and GPT have Cohen's κ values of 0.88, 0.82, and 0.92, respectively, indicating a high level of agreement. The overall Cohen's κ value is 0.89, confirming the quality of the annotations in the MixedQA dataset.

A.5 IMPLEMENTATION DETAILS

A.5.1 DETAILS OF THE SUBSTITUTION ALGORITHM

To generate high-quality distractors, we propose a substitution algorithm that produces a dictionary of substitute candidates D, where keys represent words to be substituted and the values are the corresponding potential substitute words. As shown in Algorithm1, the substitution algorithm involves three functions to search for substitute candidates for a given entity or word: *find_dbpedia_ents*, *find_candidates_by_embedding*, and *find_candidate_by_pos*. The *find_dbpedia_ents* function retrieves DBpedia entities with the same DBpedia types (Hu et al., 2022; 2023) as the entities to be substituted. We employ the `DBpedia Spotlight for Spacy` to detect DBpedia entities in response r and obtain their corresponding DBpedia types. For example, the DBpedia types for an entity like "Kobe" might include "DBpedia:Person," "DBpedia:Athlete," and "DBpedia:BasketballPlayer." Subsequently, we use SPARQL to search for DBpedia entities that satisfy these types, starting with entities that match all types and gradually lowering the number of required matching types if necessary. The *find_candidates_by_embedding* function identifies the most similar words by computing cosine similarity between the mean of the projection weight vectors of the words to be substituted and the vectors for each word in a pretrained word2vec model (`word2vec-google-news-300`). The *find_candidate_by_pos* function randomly selects words with the same POS tag from the `word2vec-google-news-300` corpus.

Once we obtain the substitute candidate dictionary D, we generate a set of distractors by replacing one word or entity in response r with its candidates in each iteration. This approach minimizes the edit distance between the distractors and response r, making it more challenging for LLM M to distinguish between the options.

Algorithm 1: Priority-based Substitution Algorithm

Input: Question q, answer a, candidate minimum number n
Output: Substitute candidate dictionary D

```
1  Function Find_candidates_for_entities(entities_list, num):
2      candidate_dict = defaultdict(list)
       /* search candidates for DBpedia entities (source:  DBperdia)    */
3      for each entity in entities_list do
4          if the entity is a DBpedia entity then
5              candidate_dict[entity] = find_dbpedia_ents(entity, num)

6      delete the item whose value is an empty list in candidate_dict
       /* search candidates for entities (source:  embedding similarity)
          */
7      if the candidate number in candidate_dict ¡ num then
8          for each entity in entities_list do
9              candidate_dict[entity] += find_candidates_by_embedding(entity, num)

10     delete the item whose value is an empty list in candidate_dict
       /* search candidates for each word in entities (source:  embedding
          similarity)                                                   */
11     if the candidate number in candidate_dict ¡ num and there exits multi-word entity in entities_list then
12         for each entity in the multi-word entities do
13             candidate_dict[entity] += [replace the non-high frequency word in the entity with the candidates searched through
                  find_candidates_by_embedding(word, num)]

14     delete the item whose value is an empty list in candidate_dict
       /* search candidates for entities (source:  randomly selected from
          corpus with same pos)                                         */
15     if the candidate number in candidate_dict ¡ num then
16         for each entity in entities_list do
17             candidate_dict[entity] += find_candidates_by_pos(entity, num)

18     return candidate_dict

19
20  Function Find_candidates_for_words(words_list, num):
21     candidate_dict = defaultdict(list)
       /* search candidates for words (source:  embedding similarity)    */
22     for word in words_list do
23         candidate_dict[word] = find_candidates_by_embedding(word, num)

24     delete the item whose value is an empty list in candidate_dict
       /* search candidates for words (source:  randomly selected from
          corpus with same pos)                                         */
25     if the candidate number in candidate_dict ¡ num then
26         for ent in words_list do
27             candidate_dict[word] = find_candidates_by_pos(word, num)

28     return candidate_dict

29
30  Function Main:
31     ents₁, ents₂ = get entities in answer a        // 1 only in answer, 2 also in question
32     nncds₁, vbjjs₁, nncds₂, others = split words in answer a with their pos tag
33     level₁, level₂, level₃, level₄ = ents₁, nncds₁, vbjjs₁ + ents₂ + nncds₂, others

34
35     D = defaultdict(list)
36     if level₁ then
37         D = Find_candidates_for_entities(level₁, n)

38     else if level₂ then
39         D= Find_candidates_for_words(level₂, n)

40     else if level₃ then
41         if vbjjs₁ then
42             D = Find_candidates_for_words(vbjjs₁, n)

43         if the candidate number in D ¡ n and ents₂ exits then
44             D = merge(D, Find_candidates_for_entities(ents₂, n))

45         if the candidate number in D ¡ n and nncds₂ exits then
46             D = merge(D, Find_candidates_for_words(nncds₂, n))

47     else if level₄ then
48         D=Find_candidates_for_words(level₄, n)

49     else
50         print("I did not find the token that needs to be replaced.")

51     return D
```

A.5.2 TRUST$_{FC}$

We retrieve the top 10 related passages as evidence e for the query *"answer ar" correctly answers the question of "question q"* using a BM25-based Robertson et al. (2009) method from the `wikipedia-dpr` corpus Karpukhin et al. (2020). Following Qin et al. (2023), we employ `GPT-Turbo-3.5` as an entailment model to judge whether the evidence e can validate the response r generated for question q.

A.5.3 SCORING CRITERION

The scoring criterion is shown in Table 3.

Table 3: The scoring criterion for Trust$_{OV}$ which integrate the Trust$_{BC}$ and Trust$_{FC}$.

Trust$_{BC}$	Trust$_{FC}$	Trust$_{OV}$
Consistent	Support	1
Inconsistent	Support	0.8
Consistent	Neutral	0.6
Inconsistent	Neutral	0.4
Consistent	Contradict	0.2
Inconsistent	Contradict	0

A.5.4 HYPER-PARAMETERS

FLAN-T5 and LLaMA employ a greedy decoding strategy with a maximum length of 1042 tokens. GPT-3.5 uses a temperature setting of 0 with a maximum length of 1042 tokens. The maximum check limit of behavior consistency evaluator is set as 10.

A.5.5 PROMPTS

Table 4 presents the prompt used in closed-book question answering task.

Table 4: The prompt for Close-book Question Answering

Prompt for Close-book Question Answering
INSTRUCTION: Please give answers to the following questions about knowledge.
Question: who has been ranked no. 1 in the latest football rankings announced by fifa?
Answer: Germany has been ranked no. 1 in the latest football rankings announced by fifa.
Question: who sings i just want to use your love tonight?
Answer: English rock band the Outfield sings i just want to use your love tonight.
Question: where was the movie the glass castle filmed?
Answer: The movie the glass castle was filmed in Welch, West Virginia.
Question: who was the first lady nominated member of the rajya sabha?
Answer: Mary Kom was the first lady nominated member of the rajya sabha.
Question: what is the tiger's name in life of pi?
Answer: Richard Parker is the tiger's name in life of pi.
Question: {Q}
Answer:

Table 5 presents the prompt used in behavior consistency evaluator.

Table 5: The prompt for Behavior Consistency Evaluation

Prompt for Behavior Consistency Evaluation
INSTRUCTION: Please give answers to the following multi-choice questions about knowledge. Question: who has been ranked no. 1 in the latest football rankings announced by fifa? A) Germany has been ranked no. 1 in the latest football rankings announced by fifa. B) India has been ranked no. 1 in the latest football rankings announced by fifa. C) Canada has been ranked no. 1 in the latest football rankings announced by fifa. D) Austria has been ranked no. 1 in the latest football rankings announced by fifa. E) None of above. Answer: E Question: who sings i just want to use your love tonight? A) Latin rock band the Outfield sings i just want to use your love tonight. B) English Power pop band the Outfield sings i just want to use your love tonight. C) English rock band the Outfield sings i just want to use your love tonight. D) English melodic sensibility band the Outfield sings i just want to use your love tonight. E) None of above. Answer: C Question: where was the movie the glass castle filmed? A) The movie the glass castle was filmed in London. B) The movie the glass castle was filmed in Welch, West Virginia. C) The movie the glass castle was filmed in Philadelphia. D) The movie the glass castle was filmed in Budapest. E) None of above. Answer: B Question: who was the first lady nominated member of the rajya sabha? A) William Randolph Hearst was the first lady nominated member of the rajya sabha. B) Jesse Speight was the first lady nominated member of the rajya sabha. C) Thurlow Weed was the first lady nominated member of the rajya sabha. D) Mary Kom was the first lady nominated member of the rajya sabha. E) None of above. Answer: D Question: what is on a mcchicken sandwich from mcdonalds? A) A breaded chicken patty is on a mcchicken sandwich from mcdonalds. B) A Hot dog chicken patty is on a mcchicken sandwich from mcdonalds. C) A breaded Bacon is on a mcchicken sandwich from mcdonalds. D) A breaded Teriyaki chicken is on a mcchicken sandwich from mcdonalds. E) None of above. Answer: A Question: {Q} Answer:

Table 6 presents the prompt used in factual consistency evaluator.

Table 6: Prompt for Fact Checking

Prompt for Fact Checking
INSTRUCTION: Assess whether the provided evidence aligns with or contradicts the given question-answer pair, and categorize the relationship as either 'support', 'contradict', or 'neutral'. Evidence: [evidence e] Question: [question q] Answer: [answer a] Relation:

A.6 EXAMPLES OF MULTI-CHOICE QUESTIONS GENERATED FOR BEHAVIOR CONSISTENCY EVALUATION

Table 7 presents the examples of the multi-choice questions used in Behavior Consistency Evaluator

Table 7: Examples of Multi-Choice Questions Generated for Behavior Consistency Evaluation.

Examples of Multi-Choice Questions Generated for Behavior Consistency Evaluation
When did all night long come out lionel richie?   A) All night long came out in 1975. [distractor]   B) All night long came out in 1986. [distractor]   C) All night long came out in 1983. [original response]   D) All night long came out in 1999. [distractor]   E) None of the above.
Who built the first temple for god in jerusalem?   A) Gopala III built the first temple for god in jerusalem. [distractor]   B) King Solomon built the first temple for god in jerusalem. [original response]   C) Mohammad Hatta built the first temple for god in jerusalem. [distractor]   D) Alexander the Great built the first temple for god in jerusalem. [distractor]   E) None of the above.
which state was returned to spain after the revolutionary war?   A) Illinois was returned to spain after the revolutionary war. [distractor]   B) Louisiana was returned to spain after the revolutionary war. [distractor]   C) Florida was returned to spain after the revolutionary war. [original response]   D) Lincolnshire was returned to spain after the revolutionary war. [distractor]   E) None of the above.