Evaluating Open-Domain Question Answering in the Era of Large Language Models

Ehsan Kamalloo ♢♣ **Nouha Dziri** ♠ **Charles L. A. Clarke** ♣ **Davood Rafiei** ♢

♢ University of Alberta ♣ University of Waterloo
♠ Allen Institute for Artificial Intelligence

ekamalloo@uwaterloo.ca

Abstract

Lexical matching remains the *de facto* evaluation method for open-domain question answering (QA). Unfortunately, lexical matching fails completely when a plausible candidate answer does not appear in the list of gold answers, which is increasingly the case as we shift from extractive to generative models. The recent success of large language models (LLMs) for QA aggravates lexical matching failures since candidate answers become longer, thereby making matching with the gold answers even more challenging. Without accurate evaluation, the true progress in open-domain QA remains unknown. In this paper, we conduct a thorough analysis of various open-domain QA models, including LLMs, by manually evaluating their answers on a subset of NQ-OPEN, a popular benchmark. Our assessments reveal that while the true performance of all models is significantly underestimated, the performance of the InstructGPT (zero-shot) LLM increases by nearly +60%, making it on par with existing top models, and the InstructGPT (few-shot) model actually achieves a new state-of-the-art on NQ-OPEN. We also find that more than 50% of lexical matching failures are attributed to semantically equivalent answers. We further demonstrate that regex matching ranks QA models consistent with human judgments, although still suffering from unnecessary strictness. Finally, we demonstrate that automated evaluation models are a reasonable surrogate for lexical matching in some circumstances, but not for long-form answers generated by LLMs. The automated models struggle in detecting hallucinations in LLM answers and are thus unable to evaluate LLMs. At this time, there appears to be no substitute for human evaluation.[1]

1 Introduction

Reliable benchmarks have been a bedrock to measuring progress in open-domain QA, the task of answering information-seeking questions over a massive text corpus. In recent years, we have seen great strides in open-domain QA by novel models (Chen et al. 2017; Wang et al. 2018; Clark and Gardner 2018; Lee et al. 2019; Asai et al. 2020; Izacard and Grave 2021b,a; Khattab et al. 2021; Singh et al. 2021; Asai et al. 2022; *inter alia*) that continue to raise state-of-the-art on well-established benchmarks such as Natural Questions-OPEN (Lee et al., 2019). The standard procedure for evaluating open-domain QA models, borrowed from reading comprehension (Rajpurkar et al., 2016), is to perform lexical matching between gold answers provided in the benchmark and models' predictions. However, as the performance of open-domain QA approaches that of humans,[2] these classic evaluation methods begin to fail. Such failures largely stem from the incomplete list of gold answers that do not fully cover all plausible answers. For example, in Figure 1, "*Jicheng*" is a correct answer to *what was the city of Beijing previously known as?* while not annotated as a gold answer in Natural Questions-

Figure 1: Examples of failures in open-domain QA evaluation. **Top:** *Jicheng* is a credible answer although not present in the list of gold answers. Existing automated evaluation mechanisms fail to identify it as correct. **Bottom:** A seemingly correct but *unattributable* answer from InstructGPT (Ouyang et al., 2022) for which automatic evaluation goes astray.

[1] Code and data are released at https://github.com/ehsk/OpenQA-eval.

[2] typically equipped with a search engine

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5591–5606

OPEN (NQ-OPEN; Lee et al. 2019).

With the recent success of generative QA systems in the open-domain setting (Izacard and Grave, 2021b; Roberts et al., 2020), it becomes harder for lexical matching to recognize correct answers, and in turn for us, to recognize performance differences between models. The problem is exacerbated by a tendency of Large Language Models(LLM)-based systems (Brown et al. 2020; Chowdhery et al. 2022; Zhang et al. 2022; Black et al. 2022; *inter alia*) to occasionally hallucinate plausible but incorrect answers (Dziri et al., 2022; Ye and Durrett, 2022). For instance, in Figure 1, InstructGPT (Ouyang et al., 2022) generates "*Jack Nicholson*" in great details to answer *who won the oscar for best actor in 1975?* but although looks natural, the answer is not factually correct (he won in 1976). Therefore, human confirmation of answer correctness demands additional effort and care due to the ability of LLMs to formulate these answers as complete and seemingly authoritative.

While it might be assumed that improved performance under lexical matching would reflect improved performance in an absolute sense, even if some correct answers are missed, we show this assumption does not hold. For this purpose, we manually re-evaluate several open-domain QA models on a random subset of NQ-OPEN (Lee et al., 2019), an established benchmark. Not only is true performance substantially underestimated by this benchmark, but the relative performance of the models alters after re-evaluation: InstructGPT (zero-shot) achieves an accuracy of 12.6% on our NQ-OPEN subset, but our human judgment reveals its true performance to be 71.4%, a nearly +60% improvement. Our linguistic analysis of the failure cases of lexical matching, an extension of a similar study by Min et al. (2021), shows that the mismatches are mostly linguistically shallow and could be captured by simple patterns, such as regular expressions.

In contrast, automated evaluation mechanisms such as BEM (Bulian et al., 2022) based on semantic matching between the gold answers and generated answers produce a relative performance that is mostly consistent with human evaluation, although the absolute improvements are lower. However, long-form answers, generated by LLMs, introduce a new challenge that did not occur on prior models; they are prone to carry unattributable information (Rashkin et al., 2021). Automated evaluation models often deem the hallucinated responses correct, which is why, InstructGPT (zero-shot) is overestimated under these models, compared to human judgment.

We repeated this experiment with the 20-year-old CuratedTREC dataset (Voorhees, 2003) that provides its gold answers in the form of regular expressions. We observe that the relative performance of models remains mostly consistent under all three evaluation mechanisms, i.e., regular expressions, human evaluation, and semantic matching, with only slight differences in absolute performance. However, the ranking discrepancy still persists between the two LLMs, i.e., InstructGPT (zero-shot) and InstructGPT (few-shot). Also, only under human judgment does the absolute performance of LLMs exceed that of the heavily engineered statistical NLP systems from 20 years ago on this collection. Until recently, the best of these classical systems has been substantially superior to even the best of the modern neural models. In light of our observations, we highlight that while semantic matching against exact answers would have been sufficient for QA evaluation prior to LLMs, they cannot accurately evaluate LLMs.

2 Related Work

Answer Equivalence in QA. One way to tackle this task is through the automatic collection of alternative plausible answers from auxiliary knowledge sources such as a knowledge base (Si et al., 2021). However, the effectiveness of this approach is heavily contingent on the presence of answers in the knowledge source, which is often not the case. For instance, numerical answers or common phrases are unlikely to be found in a knowledge base. Moreover, matching gold answers with knowledge base entries can also be problematic as their surface forms may not be identical. Thus, these approaches fail to scale for various types of answers. Another line of work focuses on building models to perform semantic similarity between candidate answers and gold answers, which can supersede lexical matching for verifying answers (Chen et al., 2019, 2020; Risch et al., 2021; Bulian et al., 2022). These methods indeed work well in reading comprehension because the presence of an input context often curtails the possibilities of models' generated answers. However, they are susceptible to failure in open-domain QA where questions should be answered without any additional context. Similarly, unsupervised semantic similarity-based evaluation met-

rics such as BERTScore (Zhang et al., 2020) that rely on token-level matching of contextualized representations exhibit poor correlation with human judgment in QA evaluation (Chen et al., 2019) and lack the ability to capture attributability (Maynez et al., 2020).

Human Judgment in QA. Many works (Roberts et al., 2020; Min et al., 2021) resort to human evaluation to assess QA models. Although using humans for evaluation is expensive and not scalable, Min et al. (2021) find that the performance of QA systems bumps up 23% on average using human judgment. The substantial gap between the true performance and token-based metrics showcases the long known strictness problem of lexical matching.

3 Open-domain QA Evaluation

The task of open-domain QA is referred to finding answers for information-seeking questions given a massive knowledge source such as Wikipedia (Voorhees and Tice, 2000). The questions are typically factoid with short answers and acontextual (Rogers et al., 2022). Open-domain QA datasets encompass questions with their annotated gold answers that serve as a reference for evaluation. Following reading comprehension (Rajpurkar et al., 2016), evaluation is carried out via lexical matching using the following two widely used metrics to measure the performance of models:

- **Exact-Match accuracy (EM)**: A candidate answer is deemed correct iff it can be found in the set of gold answers. The ratio of correct answers in the test collection is reported as EM accuracy.
- **F_1 score**: Considering answers as bags of tokens, a candidate answer receives a partial score (F_1) iff its tokens overlap with those of a gold answer. The maximum F_1 score over a set of gold answers is assigned to the candidate answer. The final metric at corpus-level is measured via averaging F_1 scores over the test collection.

Based on the implementation of Rajpurkar et al. (2016), answers are normalized (i.e., case-folded, and punctuation and articles are discarded) to compute these metrics.

3.1 Models

We select open-domain QA models with publicly available codebase and reproduce their reported results. For all models, the "base" flavors are chosen for the experiments. In total, we use 12 models.

Retriever-Reader Models. DPR (Karpukhin et al., 2020) is a well-known open-domain QA model that consists of a bi-encoder retriever and leverages an extractive reader. In addition to DPR, we pair several retrievers with Fusion-In-Decoder (FiD; Izacard and Grave 2021b), a prominent generative model that condition generating an answer on a list of passages: ANCE (Xiong et al., 2021), Contriever[3] (Izacard et al., 2022) RocketQAv2 (Ren et al., 2021), and FiD-KD (Izacard and Grave, 2021a). Further, we leverage GAR (Mao et al., 2021), a sparse retrieval model that augments questions with relevant contextual information generated by a fine-tuned T5 (Raffel et al., 2020). We fuse ANCE and GAR results with BM25, namely ANCE+ and GAR+, as they led to better results. We also use R2-D2 (Fajcik et al., 2021) that combines extractive and generative readers.

End-to-End Models. EMDR^2 (Singh et al., 2021) is an end-to-end model that jointly trains a dense retriever with a FiD-style reader. We also use EviGen (Asai et al., 2022) that jointly learns to predict the evidentiality of passages and to generate the final answer in a multi-task fashion.

Closed-book Models. We use InstructGPT[4] (Ouyang et al., 2022) in two settings, following Brown et al. (2020): zero-shot and few-shot where the prompt includes 64 question/answer pairs, randomly sampled from the NQ-OPEN training data.

3.2 Dataset

We select questions from NQ-OPEN (Lee et al., 2019), a popular open-domain QA benchmark, that consists of 3610 questions in the test set. We randomly sample 301 questions from NQ-OPEN. Answers are generated via the prominent open-domain QA models, described in §3.1, for the selected questions. In total, the number of unique answers generated by the 12 models for 301 questions amounts to 1490 question/answer pairs. Our experiments are done on Wikipedia, following the same settings provided by Karpukhin et al. (2020).

[3] https://huggingface.co/facebook/contriever-msmarco

[4] text-davinci-003, the details about this model are available at https://beta.openai.com/docs/model-index-for-researchers.

4 Strategies for Evaluating Open-domain QA Models

Our goal is to shed light on the discrepancies between the actual and the measured accuracy of open-domain QA models. To this end, we adopt three evaluation mechanisms in addition to lexical matching to assess 12 open-domain QA models and draw a comparison between their estimated accuracy and the token-based performance.

4.1 Supervised Evaluation via Semantic Similarity

A common paradigm to evaluate QA systems is to cast evaluation as a classification task where the goal is to decide whether gold answers and candidate answers are semantically equivalent or not (Risch et al., 2021; Bulian et al., 2022). To this end, we use a recent BERT-based model, namely BEM (Bulian et al., 2022), that is trained on a human-annotated collection of answer pairs given a question, derived from SQuAD (Rajpurkar et al., 2016). For evaluation, we feed a question along with a gold answer and a candidate answer to BEM and take its prediction. For questions with multiple gold answers, each gold answer is independently tested with a candidate answer. Once matched with either of the gold answers, a candidate answer is deemed correct.

4.2 Zero-shot Evaluation via Prompting

We also test the ability of LLMs for evaluating QA models. In open-domain QA, the task of answer equivalence requires supplementary information in the absence of a given context, e.g., matching "*Jicheng*" with "*Peking*" in Figure 1; therefore, LLMs are a reasonable choice here because they are equipped with an implicit memory that encompass knowledge (Roberts et al., 2020), serving thus as an auxiliary information. To use LLMs for evaluating models, we elicit the following prompt through InstructGPT (Ouyang et al., 2022):

```
Question: what was the city of Beijing
previously known as?
Answer: Peking
Candidate: Jicheng
Is candidate correct?
```

We include the gold answer along with the candidate answer in the prompt, akin to the semantic similarity mechanism, as the objective here is to verify the correctness of the candidate. We call this evaluation method, InstructGPT-eval. We also test GPT-4 (OpenAI, 2023) using the same evaluation method, namely GPT4-eval, and observe that its results, reported in §A, closely resemble to those obtained from InstructGPT-eval.

4.3 Human Evaluation

Human evaluation reflects the true performance of a model and serves as a basis for checking the feasibility of other evaluation mechanisms. For this purpose, we ask two human annotators[5] to judge whether a given answer to a question is correct or not. We present only question/answer pairs to human annotators to avoid any inadvertent biases, i.e., the annotators do not know which answers correspond to which model nor do they know if an answer is a gold answer. Annotators are allowed to use a search engine to find evidence that supports or rejects a candidate answer. Our annotation procedure is specifically geared towards open-domain QA unlike those of Risch et al. (2021) and Bulian et al. (2022) that are designed for reading comprehension where annotators decide equivalence between a pair of answers given a question and a context.

The Fleiss' Kappa score between the two annotators is 72.8%, i.e., 202 disagreements out of 1490 cases (13.6%), indicating substantial agreement. Most disagreements arise from questions that are more likely to possess subjective answers. They mainly fall into three categories: ambiguous (e.g., "*what is the corporate tax rate in great britain*"), list-style (e.g. "*who dies in the lost city of z*"), and time-dependent (e.g. "*latest series of keeping up with the kardashians*") questions. We ask a third annotator to judge the 202 cases where the two annotators diverged and take a majority vote to determine the correctness. The accepted answers by the annotators are then added to the set of gold answers for the selected questions. We compute the accuracy of the 12 models after amending the gold answers and compare it with the original accuracy that is computed via lexical matching.

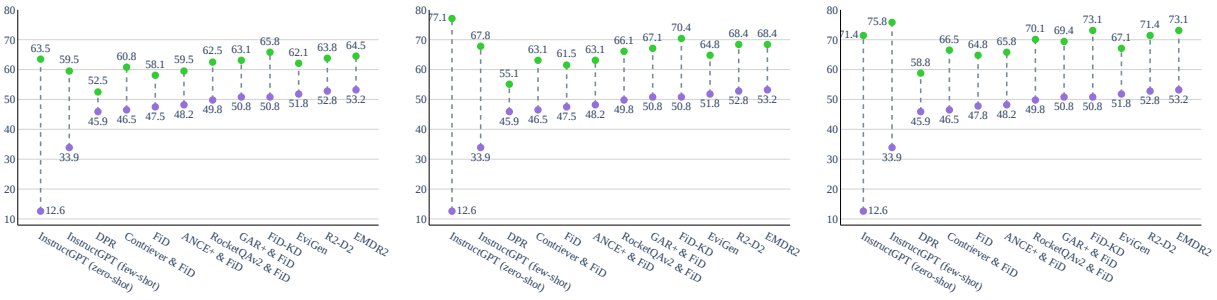
4.4 Results and Discussion

Table 1 presents the accuracy of the open-domain QA models, computed using the three evaluation mechanisms, BEM, InstructGPT-eval, and Human, compared to the de facto EM accuracy. The accuracy of all models consistently surges across all

[5]The human annotators are the authors of this paper.

Model	K	*Entire Data* (3.6K) EM	F_1	*Sampled* (301) EM	F_1	*BEM* Acc	Δ	*InstructGPT-eval* Acc	Δ	*Human* Acc	Δ
InstructGPT (zero-shot)	-	14.6	-	12.6	27.5	63.5	**+50.9**	**77.1**	**+64.5**	71.4	**+58.8**
InstructGPT (few-shot)	-	29.9	-	33.9	50.5	59.5	+25.6	67.8	+33.9	**75.8**	+41.9
DPR	50	40.9	47.8	45.9	52.3	52.5	+6.6	55.1	+9.2	58.8	+12.9
FiD	100	46.5	53.7	47.8	55.4	58.1	+10.3	61.5	+13.7	64.8	+17.0
ANCE+ & FiD	50	47.3	54.8	48.2	55.9	59.5	+11.3	63.1	+14.9	65.8	+17.6
RocketQAv2 & FiD	100	47.7	55.6	49.8	58.7	62.5	+12.7	66.1	+16.3	70.1	+20.3
Contriever & FiD	100	47.9	55.4	46.5	55.9	60.8	+14.3	63.1	+16.6	66.5	+20.0
FiD-KD	100	49.6	57.4	50.8	61.2	**65.8**	+15.0	70.4	+19.6	73.1	+22.3
GAR+ & FiD	100	49.8	57.4	50.8	59.7	63.1	+12.3	67.1	+16.3	69.4	+18.2
EviGen	20	49.8	57.0	51.8	59.5	62.1	+10.3	64.8	+13.0	67.1	+15.3
$EMDR^2$	50	51.5	**59.5**	**53.2**	**62.6**	64.5	+11.3	68.4	+15.2	73.1	+19.9
R2-D2	25	**52.4**	59.0	52.8	61.4	63.8	+11.0	68.4	+15.6	71.4	+18.6

Table 1: Accuracy of several open-domain QA models on a randomly sampled subset of 301 questions from NQ-OPEN using lexical matching and the three evaluation mechanisms along with the reported results of these models on the entire dataset. **K** refers to the number of passages fed to a model to generate an answer. InstructGPT (zero-shot) and InstructGPT (few-shot) achieve the highest raise in accuracy across all three additional evaluation methods. Only under human assessment does InstructGPT (few shot) outperform all other models.



(a) BEM (b) InstructGPT-eval (c) Human

Figure 2: Accuracy of 12 open-domain QA models on the NQ-OPEN subset of 301 questions using EM (purple points) and the three evaluation mechanisms (green points). For LLMs, the ranking of models under BEM and InstructGPT-eval is not consistent with human evaluation, while the rest of the models are ranked similarly under the two evaluation method. InstructGPT (few shot) outperforms other models only under human assessment.

three evaluation mechanisms, i.e., 16%, 21%, and 24% on average for BEM, InstructGPT-eval, and Human, respectively. InstructGPT (zero-shot) and InstructGPT (few-shot) are the top 2 models with the highest raise in accuracy across the evaluation mechanisms, whereas the amended result of DPR achieves the lowest increase. Moreover, the accuracy reported using BEM and InstructGPT-eval are yet lower than that of human judgment, i.e., trailing 7.6% and 2.9% on average across all open-domain QA models, respectively.

More importantly, the ranking of models is readjusted by applying the three evaluation mechanisms. Figure 2 visualizes the accuracy of the open-domain QA models before (using only EM) and after our evaluation. $EMDR^2$, originally the best performing model, loses the top spot to InstructGPT (few-shot) by a nearly +3% margin using human evaluation. BEM picks FiD-KD as the best model, whereas the LLM-based evaluation method estimates the highest accuracy for InstructGPT (zero-shot). Also, the Kendall's τ correlation of InstructGPT-eval, and BEM with human evaluation is 0.75, and 0.70, respectively, whereas EM and F_1 show a significantly weaker correlation of 0.23 and 0.37.

In contrast to human evaluation, BEM and InstructGPT-eval show that InstructGPT (zero-shot) has 4%, and 9% advantage, respectively, over InstructGPT (few-shot). To further investigate this phenomenon, we manually examine the InstructGPT (zero-shot) generated answers that are deemed incorrect by humans. We identify 47 unattributable answers out of 86 answers. The generated answers of InstructGPT (zero-shot) tend to be long statements that offer supplementary information, which raises the risk of containing hallucinated content. InstructGPT-eval accepts 30 of those answers

(∼10% error over the 301 questions), whereas BEM incorrectly predicts 18 (∼6% error) answers as correct. Interestingly, GPT4-eval performs better and misidentifies only 9 cases (∼3% error). Yet, these results highlight that the automated methods are prone to misjudging hallucinated long answers, essentially rendering them unreliable against answers generated by LLMs.

5 Linguistic Analysis of Correct Answers

In this section, we aim to examine model answers that are not considered correct based on EM, but are in fact acceptable according to our assessment. Min et al. (2021) conducted a similar analysis on 50 questions for the participating models in the EfficientQA competition at NeurIPS 2020. In line with this work, we provide an in-depth analysis on a broader scale using more recent models to emphasize the drawbacks of widely used lexical-based evaluation metrics and semantic similarity methods. We further dissect the categories presented by Min et al. (2021) into more detailed sub-categories. Specifically, we group the 493 question/answer pairs that are deemed correct by humans while cannot be matched with gold answers into hierarchical categories as follows:[6]

Semantic Equivalence: Model predictions and gold answers convey the same meaning while not matching verbatim:

(i) Multinominal entities, e.g., "*Bhimrao Ramji Ambedkar*" and "*B. R. Ambedkar*."

(ii) Synonymous answers, e.g., "*a virtual reality simulator*" and "*a virtual reality world*."

(iii) More elaborate answers, e.g., "*Typically , no*" and "*not required in all jurisdictions*."

(iv) Exact-Match in explanatory answers, e.g., "*1995*" and "*Michael Jordan returned to the NBA in 1995.*"

(v) Bridging/Abridging, e.g., "*citizens*" vs. "*ordinary citizens*" or "*in the Gospel of Luke*" vs. "*Gospel of Luke*."

(vi) Tokenization mismatches, especially in the presence of punctuation marks, e.g., "*s-block*" and "*s - block*."

[6]Long answers, generated by LLMs, are annotated based solely on the parts that candidate answers are mentioned.

Symbolic Equivalence: In case of numeric answers, gold answers and predicted ones can be symbolically identical either exactly or approximately while their surface text differs, e.g., "*about 3.99 degrees*" vs. "*3.97 degrees*" or the year "*1524*" vs. "*the 16th century*."

Intrinsic Ambiguity in Questions: Ambiguous questions have several interpretations, each of which can lead to different answers. Min et al. (2020) found that ambiguity is prevalent in NQ-OPEN. Unlike other categories, mismatches that stem from ambiguity are not rooted in answers and instead, arise from questions themselves. For instance, "*when does the next episode of iZombie air?*" presupposes a reference point in time that can only be clarified within a context. Thus, both "*May 07, 2018*" and "*February 26, 2018*" are correct, depending on when the question is asked.

Granularity Discrepancies: Predicted answers may appear at different granularity levels than the gold answers. This case often arises for answers indicating spatial or temporal references. Indeed, under different presuppositions, some granularity levels are more preferable than others. Nonetheless, both predictions and gold answers are valid. We further categorize this discrepancy into:

(i) Temporal granularity discrepancy, e.g., "*when was the 50th star added to the united states flag?*" can be answered by both "*1960*" and "*July 4, 1960*."

(ii) Spatial granularity discrepancy, e.g., both "*Camping World Stadium*" and "*Orlando, Florida*" answer the question "*where is the citrus bowl held this year?*"

List-style Questions: Actual answers to these kinds of questions encompass a set of plausible answers that is not fully specified in gold answers. For these questions, model answers are deemed correct if they are among at least one gold answer. We broke this group down into:

(i) List questions, e.g., gold answers to "*list of strict nature reserve in the Philippines*" consist of six locations that is by no means comprehensive.

(ii) Open-ended questions such as "*what is an example of a government monopoly in the United States?*" where "*the United States*

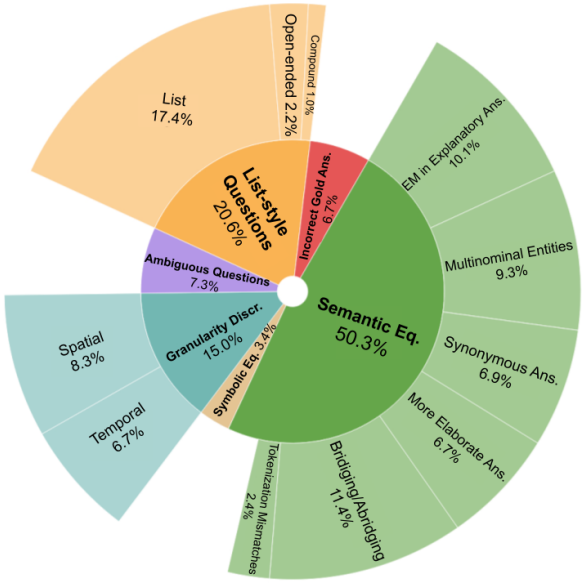


Figure 3: Statistics of exact-match failure modes determined via our linguistic analysis

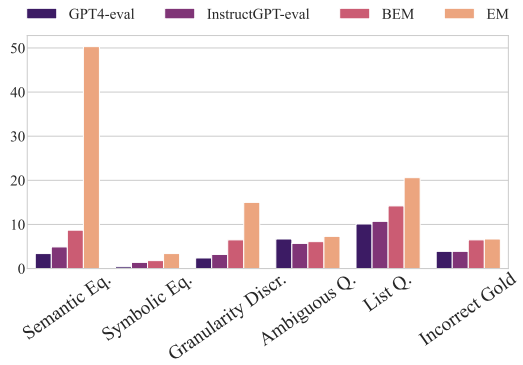


Figure 4: Percentage of high-level failure modes for each evaluation method on NQ-OPEN.

Postal Service," not listed among gold answers, is a correct answer.

(iii) Compound questions ask about multiple pieces of information in one question. They are a special case of multi-hop questions (Yang et al., 2018), e.g., "*when was the canadian pacific railway started and finished?*" where the gold answer is "*between 1881 and 1885*" vs. "*Started in 1881 and finished in 1885.*" that is a correct answer.

Incorrect Gold Answers: Models produce correct answers, but gold annotations are incorrect. Mismatches in this category are a byproduct of data quality issues. For example, the answer to "*what is the largest ethnic group in Mexico today?*" is annotated "*K'iche'*", whereas the correct answer is "*Mestizos*."

5.1 Discussion

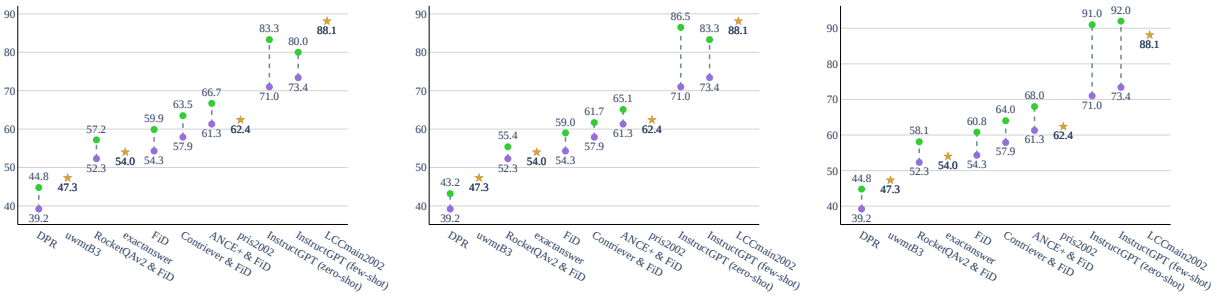
The statistics for each category are presented in Figure 3. Semantic equivalence (50.3%) is the most common failure mode of exact matching. The most frequent subcategories within this category are bridging/abridging (11.4%), EM in explanatory answers (10.1%), and multinominal entities (9.3%). Other top frequent failure modes are list-style questions (20.6%) and granularity discrepancy (15.0%). Interestingly, most of these failure cases are related to syntactical variations of answers, which is why specifying gold answers via regular expressions can be useful in capturing these variations. Moreover, 14% of EM failures are attributed to data quality issues, i.e., ambiguity and incorrect gold answers.

Error Analysis of Automated Evaluation Methods. The answers that InstructGPT-eval and BEM reject but humans consider correct are a subset of EM failures.[7] More precisely, InstructGPT-eval and BEM reduce the 493 failure cases of EM to 149 (70% ↓) and 217 (56% ↓), respectively. For GPT4-eval, the number of failure cases is 137 (72% ↓), only slightly lower than InstructGPT-eval. The breakdown of the high-level failure categories for each evaluation method is shown in Figure 4. The three automated evaluation methods are able to fix most of the failures corresponding to semantic equivalence, granularity discrepancy, and symbolic equivalence. However, they do not perform that well on list-style questions where InstructGPT-eval and GPT4-eval still fail on more than 10% of the EM failures, and BEM falls short on 14%. They also perform nearly on par with EM on data quality-related failure cases, i.e., incorrect gold answers and ambiguous questions.

6 Regex Matching on CuratedTREC

An alternative to lexical matching between gold answers and predicted answers during evaluation is to specify gold answers as regular expression patterns. Regex matching allows for capturing syntactical answer variations where exact-match falls short. In this section, our main goal is to highlight

[7] With only 3 exceptions: InstructGPT-eval rejects only 2 actually correct answers matching with gold answers that correspond to list questions where candidate answers appear in the middle of the gold answers. Moving the candidate answer to the top of the gold answer list would fix the issue. Similarly, BEM rejects only 1 exactly matched correct answer, i.e., "*P-A-D-A-W-A-N.*" while the gold answer is "*Padawan*".



(a) BEM (b) InstructGPT-eval (c) Human

Figure 5: Accuracy of several open-domain QA models on CuratedTREC 2002, computed via regex matching, along with the results of three evaluation mechanisms. Purple points represent the EM accuracy, and green points depict accuracy achieved via BEM, InstructGPT-eval, and human judgment. Classic statistical models from TREC QA 2002 are shown as orange stars. InstructGPT (few shot) outperforms the best of these classic models *only* under human assessment.

the advantages and pitfalls of using answer patterns in QA evaluation by comparing its results with our three evaluation mechanisms, described in §3.1.

Dataset. We make a comparison across open-domain QA models on CuratedTREC 2002 (Baudiš and Šedivỳ, 2015), a dataset whose gold answers are specified by regular expressions. The questions in CuratedTREC are derived from the dataset in the QA tracks (Voorhees, 2003) of TREC 2001 to 2003 after a manual review to discard ambiguous or outdated questions. The knowledge source for TREC QA is originally English news text, namely AQUAINT, from three news sources (AP, NYTimes, and Xinhua), dating back to the late 90s. Here, we opt for the original knowledge source to replicate the same environment as TREC QA 2002 so as to quantitatively gauge progress over two decade by comparing recent models with the models that took part in the QA track in 2002. This experiment is an out-of-distribution test for the neural models to check whether they are actually capable of using the knowledge source to answer questions or they answer from memory because the old news articles is less likely to have appeared in the pretraining corpus. However, LLMs inevitably do not use the knowledge source as they perform the task from their memory in a closed-book fashion. CuratedTREC 2002 consists of 444 questions whose answers are looked up in the AQAUINT corpus, comprising around 1M news articles. We follow Karpukhin et al. (2020) to split the articles into non-overlapping passages of 100 words, which amounts to over 4M passages in total.

Models. Out of the 12 models, we keep the ones that do not require further training on CuratedTREC 2002, leaving us with 7 models. These models produce 1872 unique answers on CuratedTREC 2002. We also obtained the submitted run files of the participants in the TREC QA 2002 track from TREC organizers to compute their accuracy on CuratedTREC 2002. We include top 4 teams as baselines: LCCmain2002 (88.1%; Pasca and Harabagiu 2001), pris2002 (62.4%), exactanswer (54.0%), and uwmtB3 (47.3%).

Similar to NQ-OPEN, we ask two annotators to judge 1872 question/answer pairs, followed by a third annotator who evaluates the diverging cases. The Fleiss' Kappa score between the two annotators is 83.5%, i.e., 150 disagreements (8.0%), indicating an almost perfect agreement.

The results are shown in Figure 5. Interestingly, the ranking of models via regex matching is left unchanged by all three evaluation mechanisms, except for InstructGPT (zero-shot) and InstructGPT (few-shot). Consistent with our observation on NQ-OPEN, both BEM and InstructGPT-eval assign a higher accuracy to InstructGPT (zero-shot) over InstructGPT (few-shot). However, in contrast to NQ-OPEN, they do not overestimate InstructGPT (zero-shot). Human evaluation shows that InstructGPT (few-shot), by scoring 92%, is the best performing model, analogous to NQ-OPEN. Among the non-LLM models, ANCE+ and Contriever consistently surpass other models. Similar to EM, regex matching is too rigid albeit to a lesser extent. In particular, the accuracy is underestimated by 6.6%, 6.4%, and 9.9% on average via BEM, InstructGPT-eval, and human evaluation, respectively.

We note that LCCmain2002, an original TREC run, outperforms all models prior to our assessment. Human evaluation highlights that both InstructGPT

models are superior to LCCmain2002 by +1.9% (for zero-shot) and +2.9% (for few-shot). However, BEM and InstructGPT-eval fail to reflect this result. For other non-LLM models, ANCE+ and Contriever surpass pris2002 via all three evaluation methods (with the exception of Contriever using InstructGPT-eval). An interesting finding here is that although neural open-domain QA models are repeatedly proven to be powerful in accomplishing state-of-the-art, LCCmain2002, a heavily engineered statistical method from 20 years ago, ruffles their feathers by a substantial margin of 20%. Only under human judgment does the absolute performance of LLMs surpass LCCmain2002.

7 Conclusion

Despite the simplicity and ubiquity of lexical matching as an evaluation metric in open-domain QA, it is unnecessarily rigid because plausible candidate answers are likely not to appear in the list of gold answers. This flaw has been long known, but the efforts to circumvent it have been mostly artisanal. In this paper, we report a systematic study of lexical matching by manually judging answers generated by several prominent open-domain QA models. We found that LLMs achieve state-of-the-art on NQ-OPEN. The accuracy of models is severely underestimated, with most EM failure cases stemming from syntactical variations of answers. Moreover, a zero-shot prompting method can be a reasonable substitute for human evaluation although it cannot detect unattributability in long-form answers. Our insights and analysis in this paper will hopefully underpin the development of solid evaluation techniques in open-domain QA.

Limitations

Our main focus in this work is limited to factoid information-seeking questions that typically prompt short answers. However, lexical matching is adopted by more complicated forms of QA that require complex reasoning. More precisely, QA tasks such as multi-hop reasoning (Yang et al., 2018), discrete reasoning (Dua et al., 2019), and causal relations (Lin et al., 2019) also warrant similar systematic analysis as studied in this paper.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback.

References

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-guided generation for knowledge-intensive NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2226–2243, Seattle, United States. Association for Computational Linguistics.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Petr Baudiš and Jan Šedivỳ. 2015. Modeling of the question answering task in the YodaQA system. In *International Conference of the cross-language evaluation Forum for European languages*, CLEF'15, pages 222–228. Springer-Verlag.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Martin Fajcik, Martin Docekal, Karel Ondrej, and Pavel Smrz. 2021. R2-D2: A modular baseline for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 854–870, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2021b. 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.

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.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for OpenQA with ColBERT. *Transactions of the Association for Computational Linguistics*, 9:929–944.

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.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen-tau Yih. 2021. NeurIPS 2020 EfficientQA competition: Systems, analyses and lessons learned. volume 133 of *Proceedings of Machine Learning Research*, pages 86–111. PMLR.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. Technical report.

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

Marius A. Pasca and Sandra M. Harabagiu. 2001. High performance question/answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 366–374, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. 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.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2022. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

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.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems*, volume 34, pages 25968–25981.

Ellen M. Voorhees. 2003. Overview of the TREC 2002 question answering track. In *TREC*.

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'00)*, Athens, Greece. European Language Resources Association (ELRA).

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R^3: Reinforced ranker-reader for open-domain question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30378–30392. Curran Associates, Inc.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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 Zero-shot Evaluation using GPT-4

For the sake of completeness, we test the ability of GPT-4 (OpenAI, 2023) for evaluating QA models as explained in §4.2. We find that GPT4-eval results aligns with the trends observed in InstructGPT-eval, albeit displaying marginal improvements. Following the Table 1 layout, Table 2 presents the accuracy of the open-domain QA models, computed using GPT4-eval in conjunction with lexical matching, InstructGPT-eval, and human judgment as reference points. The accuracy of all models consistently increases by an average of 20% using GPT4-eval, which is similar to the increase level observed in InstructGPT-eval. Moreover, analogous to InstructGPT-eval, the GPT4-eval accuracies are, on average, 3.3% lower than those of human judgment.

Figure 6 visualizes the accuracy of the open-domain QA models on NQ-OPEN using EM and GPT4-eval, similar to Figure 2. Unlike InstructGPT-eval, GPT4-eval estimates the highest accuracy for FiD-KD, followed by InstructGPT (zero-shot), InstructGPT (few-shot), and EMDR^2. Also, the Kendall's τ correlation of GPT4-eval with human judgment is 0.79, slightly higher than 0.75 of InstructGPT-eval.

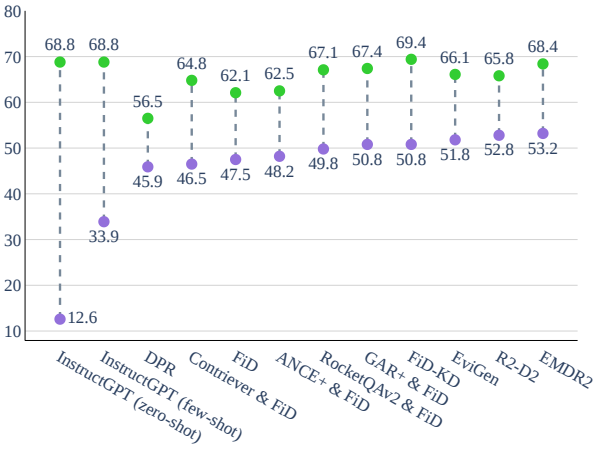


Figure 6: Accuracy of 12 open-domain QA models on the NQ-OPEN subset of 301 questions using EM (purple points) and GPT4-eval (green points).

Error Analysis: As illustrated in Figure 4, GPT4-eval errors closely resemble the errors found in InstructGPT-eval. However, for a small number of cases, GPT4-eval demonstrates unique erratic behaviours. First, for 2 cases, the model exhibits overconfidence in its internal memory and disregards gold answers that can be simply matched using EM. For example, GPT4-eval incorrectly rejects the candidate answer "*Jermaine Jackson*" (that is also a gold answer) to the question "*Who sings Somebody's Watching Me with Michael Jackson?*" We also observe the contradictory response of "*No, the candidate is correct*" for 2 candidate answers that are correct, but are not included in the gold answers. Moreover, GPT4-eval incorrectly abstains from evaluating 2 candidate answers because it thinks more context is needed. For instance, it falsely utters

> "*I cannot determine if the candidate is correct, as there is not enough information provided about the show "Fall" and the character Rose. Valene Kane is an actress, but without more context, it is unclear if she is related to this specific show or character.*"

as a response to the question "*Who is Rose in the Fall season 2?*" and the candidate answer "*Rose is a new character introduced in the second season of the show Fall. She is a mysterious woman who is connected to the supernatural events occurring in the town.*" that is entirely fabricated.

Results on CuratedTREC 2002: As shown in Figure 7, GPT4-eval follows closely InstructGPT-eval on CuratedTREC 2002. Specifically, it indicates a higher accuracy for InstructGPT (zero-shot) compared to InstructGPT (few-shot) and ranks LCCmain2002 ahead of both InstructGPT models despite human evaluation suggesting otherwise.

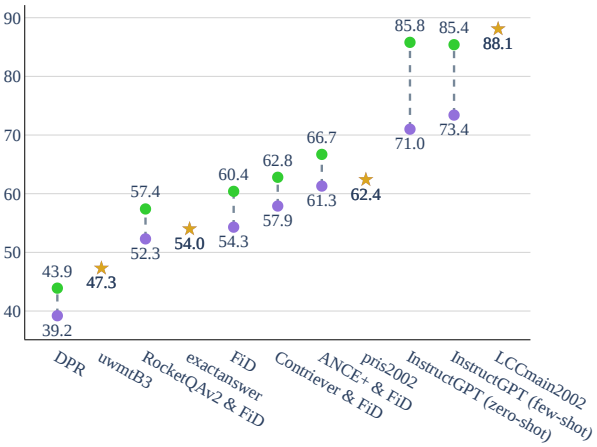


Figure 7: Accuracy of several open-domain QA models on CuratedTREC 2002, computed via regex matching (purple points), along with the results of GPT4-eval (green points), similar to Figure 5. Classic statistical models from TREC QA 2002 are shown as orange stars.

Model	*Sampled* (301)		*InstructGPT-eval*		*GPT4-eval*		*Human*	
	EM	**F_1**	**Acc**	**Δ**	**Acc**	**Δ**	**Acc**	**Δ**
InstructGPT (zero-shot)	12.6	27.5	**77.1**	**+64.5**	68.8	**+56.2**	71.4	**+58.8**
InstructGPT (few-shot)	33.9	50.5	67.8	+33.9	68.8	+34.9	**75.8**	+41.9
DPR	45.9	52.3	55.1	+9.2	56.5	+10.6	58.8	+12.9
FiD	47.8	55.5	61.5	+13.7	61.8	+14.0	64.8	+17.0
ANCE+ & FiD	48.2	55.9	63.1	+14.9	62.5	+14.3	65.8	+17.6
RocketQAv2 & FiD	49.8	58.7	66.1	+16.3	67.1	+17.3	70.1	+20.3
Contriever & FiD	46.5	55.9	63.1	+16.6	64.8	+18.3	66.5	+20.0
FiD-KD	51.2	61.6	70.4	+19.6	**69.4**	+18.6	73.1	+22.3
GAR+ & FiD	50.8	59.7	67.1	+16.3	67.4	+16.6	69.4	+18.2
EviGen	51.8	59.5	64.8	+13.0	66.1	+14.3	67.1	+15.3
$EMDR^2$	**53.2**	**62.6**	68.4	+15.2	68.4	+15.2	73.1	+19.9
R2-D2	52.8	61.4	68.4	+15.6	65.8	+13.0	71.4	+18.6

Table 2: Accuracy of several open-domain QA models on a randomly sampled subset of 301 questions from NQ-OPEN using lexical matching, GPT4-eval, human evaluation. Only GPT4-eval results are new here. The rest of the results are already reported in Table 1 and copied here solely as a reference. GPT4-eval demonstrates approximately similar behaviour as InstructGPT-eval when ranking the models.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In an unnumbered section after Section 7

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Our findings are summarized in the abstract, Section 1 (Introduction), and Section 7 (Conclusion).

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes, we used Natural Questions-open (Section 3.2) and CuratedTREC 2002 (Section 6).

☒ B1. Did you cite the creators of artifacts you used?
Section 3.2 and Section 6

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Natural Questions-open is a subset of Natural Questions that is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0) license. CuratedTREC 2002 is a subset of TREC QA 2002 whose access is governed by TREC organizers. Both datasets are public and freely available for research purposes.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3.2 and Section 6

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We only worked with the test dataset; the details are stated in Section 3.2 and Section 6.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

We used publicly available code and pre-trained models from previous work to reproduce their results. We didn't train or fine-tune any models ourselves.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 3.1, 3.2 and 6

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We reproduced previous work, cited in Section 3.1. The metrics are described in Section 3.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

We, the authors, did the annotations, described in the paper.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 4.3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.3 (footnote): the authors were the annotators

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.