A Appendix

1. Submission introducing new datasets must include the following in the supplementary materials:
   (a) Dataset documentation and intended uses. Recommended documentation frameworks include datasheets for datasets, dataset nutrition labels, data statements for NLP, and accountability frameworks.
   (b) URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers.
   (c) Author statement that they bear all responsibility in case of violation of rights, etc., and confirmation of the data license.
   (d) Hosting, licensing, and maintenance plan. The choice of hosting platform is yours, as long as you ensure access to the data (possibly through a curated interface) and will provide the necessary maintenance.
2. To ensure accessibility, the supplementary materials for datasets must include the following:
   (a) Links to access the dataset and its metadata. This can be hidden upon submission if the dataset is not yet publicly available but must be added in the camera-ready version. In select cases, e.g when the data can only be released at a later date, this can be added afterward. Simulation environments should link to (open source) code repositories.
   (b) The dataset itself should ideally use an open and widely used data format. Provide a detailed explanation on how the dataset can be read. For simulation environments, use existing frameworks or explain how they can be used.
   (c) Long-term preservation: It must be clear that the dataset will be available for a long time, either by uploading to a data repository or by explaining how the authors themselves will ensure this.
   (d) Explicit license: Authors must choose a license, ideally a CC license for datasets, or an open source license for code (e.g. RL environments).
   (e) Add structured metadata to a dataset's meta-data page using Web standards (like schema.org and DCAT): This allows it to be discovered and organized by anyone. If you use an existing data repository, this is often done automatically.
   (f) Highly recommended: a persistent dereferenceable identifier (e.g. a DOI minted by a data repository or a prefix on identifiers.org) for datasets, or a code repository (e.g. GitHub, GitLab,...) for code. If this is not possible or useful, please explain why.
3. For benchmarks, the supplementary materials must ensure that all results are easily reproducible. Where possible, use a reproducibility framework such as the ML reproducibility checklist, or otherwise guarantee that all results can be easily reproduced, i.e. all necessary datasets, code, and evaluation procedures must be accessible and documented.
4. For papers introducing best practices in creating or curating datasets and benchmarks, the above supplementary materials are not required.

B Data License and Maintenance Plan

The EVOUNA data we create is open sourced at `https://github.com/wangcunxiang/QA-Eval`. The test data is under Apache License 2.0. We plan to collect more data from various datasets, including Natural Questions, Trivia and WebQuestions, and open source it for future research.

C Data

C.1 An example of Processing BingChat Answer

Raw Answer:

The revolution period of Venus around the sun is **224.7 Earth days** [1]. Is there anything else you would like to know about Venus?

Source: Conversation with Bing, 2023/3/31(1) Planet Venus: 20 interesting facts about the scorching world. https://www.space.com/15988-venus-planet-weird-facts.html Accessed 2023/3/31. (2) What Is The Revolution Period Of Venus In Earth Years? Here's The .... https://www.odysseymagazine.com/what-is-the-revolution-period-of-venus-in-earth-years/ Accessed 2023/3/31. (3) What Is Venus's Revolution Period in Earth Days? | Sciencing. https://sciencing.com/venuss-revolution-period-earth-days-2411.html Accessed 2023/3/31. (4) Top 7 what is the revolution period of venus in earth years 2022. https://entornoit.com/top-7-what-is-the-revolution-period-of-venus-in-earth-years-2022/ Accessed 2023/3/31. (5) The length of a day on Venus keeps changing | Astronomy.com. https://astronomy.com/news/2020/01/the-length-of-a-day-on-venus-keeps-changing Accessed 2023/3/31. (6) Your Age on Other Worlds | Exploratorium. https://www.exploratorium.edu/ronh/age/ Accessed 2023/3/31.

Processed Answer:

The revolution period of Venus around the sun is 224. 7 Earth days.

C.2 Human Annotation Guidelines

Here is a question, a set of golden answers (split with /), an AI-generated answer. You are required to judge (1) whether the question have answers that change over time, simply annotate Yes or No; (2) whether the golden answer contain severe errors; (3) whether the AI-generated answer is correct according to the question and golden answers, simply annotate Yes or No.

Here is a set of guidelines for task (1) whether the question have answers that change over time:

- If the question is clearly time-sensitive, then it is Yes.
- If there are words closely related to the current time node such as "this year", "last year", "next time" and "last time" in this question, then it is Yes.
- If the question contains values that change over decades, such as "who is the player with the most goals in the World Cup so far", then it is Yes.
- If the question contains values that do not change in decades, such as "what is the tallest mountain in the world", then it is No.

If the answer to task (1) is Yes, skip to the next.

Here is a set of guidelines for task (2) whether the golden answer contain severe errors:

- If the golden answer has structure errors, then it is Yes.
  Example: Question: the south west wind blows across nigeria between? Golden: till September
- If the golden answer is obviously not what is asked, then it is Yes.
- If the golden answer has format errors, then it is Yes.
  Example: Question: what season does bart bass die in gossip girl? Golden: (
- If the golden answer has only factual errors, then it is No.

(We also present some examples shown in Section C.3.1.) If the answer to task (2) is Yes, skip to the next.

Here is a set of guidelines for task (3):

- If the question specifies a number (e.g., names of four people), and the response does not meet this requirement (e.g., provides only one name), the answer is deemed incorrect.
- Spelling errors in the responses are considered mistakes. For example, if "golden answer" is misspelled as "gloden answer," the response is marked as incorrect.

- For questions related to specific times, such as "When was the term social justice first used?" a response of "1840s" would be considered correct. However, if the answer needs to be precise to a specific day, month, and year, each time component needs to be factually accurate for the response to be marked as correct.
- For location-based queries, like "Where was Oak Island filmed?", a response of "Canada" would be deemed correct. But, if the answer requires specific details like state, city, or county, each geographical component must be accurate for the answer to be considered correct.
- If there is a direct answer and subsequent explanation in the response, then only focus on whether the direct answer is correct, not whether the subsequent explanation is correct

These guidelines were strictly followed to maintain the reliability and validity of the evaluation process.

C.3 Supplements to the Annotation

AI-generated answers.

For local-deployed models (DPR+FiD) and models can be accessed with APIs (text-davinci-003 for GPT-3.5 and gpt-3.5-turbo for ChatGPT-3.5), we generate the answers locally. For models that can only be interacted within he webpage, including ChatGPT-4 (we do not have API permissions) and BingChat, we ask the annotators to get the answer by interacting in the webpage and make judgement for the three tasks.

Data assignment.

We ask one annotator to judge samples with answers generated by DPR+FiD, GPT-3.5 and ChatGPT-3.5; one for samples by ChatGPT-4; one for samples by Bing Chat, for convenience.

Improper questions or goldens. If a sample has an improper question or improper goldens, we mark the sample as improper. Since we have three different annotators to judge improper questions and goldens, if at least two annotators mark the improper result as True, we mark it as True, then we ask the left annotator (if there exists) to re-annotate the sample.

C.3.1 Golden Error Examples

Here some examples whose golden answer has obvious mistake. The first two have factual errors while the next one has the structure error and the last one has format error.

Question: was star wars a book or a movie first? Golden: film

Question: what is the democracy of the united states? Golden: federal republic

Question: the south west wind blows across nigeria between? Golden: till September

Question: what season does bart bass die in gossip girl? Golden: (

D Methods

D.1 DPR and FiD

The DPR model retrieves relevant documents from all given documents to answer a specific question. Given a question q and a database D with each document denoted as d, the DPR model comprises two main components: the question encoder Q_{enc} and the document encoder D_{enc}. Both typically rely on neural networks, such as BERT [Devlin et al., 2019].

The question encoder Q_{enc} maps a question q to a dense vector representation $q_{emb} = Q_{enc}(q)$, and the document encoder D_{enc} maps each document d in the database D to a dense vector representation $d_{emb} = D_{enc}(d)$.

We compute the similarity between the question embedding q_{emb} and each document embedding d_{emb} using the dot product:

$$s(d, q) = q_{emb} \cdot d_{emb} \tag{3}$$

Documents in the database D are ranked based on their similarity scores, and the top k most relevant documents D_k are retrieved. These documents are then used as input for the reader model $\mathcal{M}_{reader}$ to generate an answer $\hat{a}$ to the question q:

$$\hat{a} = \mathcal{M}_{reader}(q, D_k) \tag{4}$$

D.2 BERT-Score

Given a reference $r = A$ and a hypothesis $h = \hat{a}$, we first obtain their contextualized word embeddings using a pre-trained BERT model:

$$\begin{aligned} E_r &= \text{BERT}(r), \\ E_h &= \text{BERT}(h) \end{aligned} \tag{5}$$

Next, we compute the cosine similarity between each token in the reference and each token in the hypothesis:

$$S_{i,j} = \frac{E_{r_i} \cdot E_{h_j}}{|E_{r_i}||E_{h_j}|} \tag{6}$$

We then find the optimal token matchings using the maximum cosine similarity:

$$\begin{aligned} P_r &= \frac{1}{|r|} \sum_{i=1}^{|r|} \max_{j=1}^{|h|} S_{i,j}, \\ P_h &= \frac{1}{|h|} \sum_{j=1}^{|h|} \max_{i=1}^{|r|} S_{i,j} \end{aligned} \tag{7}$$

Finally, the BERT-score is calculated as the F1 score between the reference and hypothesis:

$$\text{BERT-score} = \frac{2 \cdot P_r \cdot P_h}{P_r + P_h} \tag{8}$$

To decide whether the AI-generated answer is positive or not, we set a threshold τ and classify the prediction $\hat{y}$ as positive if the BERT-score is above the threshold and as negative otherwise:

$$\hat{y} = \begin{cases} \text{Positive}, & \text{BERT-score} >= \tau \\ \text{Negative}, & \text{BERT-score} < \tau \end{cases} \tag{9}$$

E Analysis

E.1 Additional Analysis for Open-QA

From the Table 4, we have several additional observations:

All models perform better on TriviaQA compared to Natural Questions. This might suggest that the TriviaQA dataset, which is known for its trivia-style questions, is more aligned with the kind of diverse and general knowledge these models have been trained on. In contrast, the Natural Questions dataset, which is derived from real Google search queries, might contain more complex or niche questions that are challenging for the models.

Table 7: Performance of Eval-Models on EVOUNA. In each cell, the left is the precision while the right is the recall.

	NQ-FiD	NQ-GPT35	NQ-ChatGPT35	NQ-ChatGPT4	NQ-BingChat
Lexical Matching	99.8/81.2	99.5/77.1	96.0/76.2	99.6/79.8	97.6/79.8
BERT-Score	76.7/91.7	74.7/80.8	81.9/80.6	89.4/81.4	86.7/70.2
GPT-3.5	96.3/94.3	92.2/82.5	93.7/81.1	96.6/82.7	95.6/64.8
Another Human	98.5/96.3	97.8/97.8	97.8/95.3	99.0/96.8	98.7/95.8
on EVOUNA-NaturalQuestions					
	TQ-FiD	TQ-GPT35	TQ-ChatGPT35	TQ-ChatGPT4	TQ-BingChat
Lexical Matching	100/87.7	99.0/88.0	100/89.7	97.9/88.8	98.4/87.9
BERT-Score	86.2/62.4	83.1/82.9	85.0/82.4	90.8/86.9	93.5/77.2
GPT-3.5	98.9/95.8	98.1/89.5	98.3/93.2	98.2/93.3	97.8/84.5
Another Human	100/100	99.4/99.7	98.9/99.5	99.8/100	99.8/100
on EVOUNA-TriviaQA					

Table 8: The Proportions of Evaluation Outcomes Across Three Evaluators on the EVOUNA-NQ Dataset.

	True Positive	True Negative	False Positive	False Negative
Lexical Matching	57.5	26.8	1.1	14.7
BERT-Score	57.7	13.6	14.8	14.0
GPT-3.5 Evaluator	57.8	24.5	3.3	14.3
GPT-3.5 Evaluator without NQ-BingChat	59.8	26.7	3.6	9.9

GPT-3.5 vs ChatGPT-3.5 : These two models have very similar performance, both achieving approximately 65% accuracy on NQ and 72-76% on TQ. This similarity is expected, as they are versions of the same base model, with the main difference being that ChatGPT is fine-tuned specifically for conversational contexts.

GPT-4 vs GPT-3.5 and ChatGPT-3.5 : The newer model GPT-4 significantly outperforms both GPT-3.5 and ChatGPT-3.5 on both datasets. This suggests that the improvements incorporated into GPT-4, likely including a larger model size and potentially refined training techniques, have resulted in substantial gains in question answering performance.

ChatGPT-4 vs BingChat : These two models exhibit the highest performance on both datasets. Their performance is remarkably similar, with GPT-4 outperforming Bing Chat by only a small margin on both datasets. This suggests that the two models, despite potentially having quite different architectures and training procedures, have reached similar levels of proficiency in question answering.

LLMs vs. Retrieval-based Methods : The DPR+FiD model, a representative of traditional retrieval-based methods, performs comparably to the earlier language models (GPT-3.5 and ChatGPT-3.5), but falls behind the newer ones (ChatGPT-4 and Bing Chat). This indicates that while retrieval-based methods remain competitive, the newer generation of language models have surpassed them in terms of question answering capability. This could be due to the ability of these large models to better understand and generate natural language, enabling them to generate more accurate and contextually appropriate answers.

E.2 Supplemental Analysis for QA-Eval

Table 7 showcases the performance of various evaluation models on EVOUNA-NaturalQuestions and EVOUNA-TriviaQA datasets. The reported metrics are precision and recall.

Table 9: Distribution of error types across different generative models on the NQ-test dataset. Each cell represents the proportion of the respective error type to ***all responses*** generated by the model.

	InAcc	InCom	IrrA	OutInf	MisQs	Others
DPR + FiD	25.0	3.0	0.9	1.2	1.7	0.0
GPT-3.5	25.3	5.4	0.3	2.1	1.8	0.1
ChatGPT-3.5	23.2	7.9	0.5	1.4	2.4	0.2
GPT-4	13.3	2.8	0.3	1.2	1.3	0.0
Bing Chat	9.5	7.6	1.3	1.3	0.8	0.5

Looking at the EVOUNA-NaturalQuestions results, we observe that Lexical Matching and GPT-3.5 evaluation models achieve high precision across all QA models. However, the Lexical Matching model tends to have lower recall compared to GPT-3.5. BERT-Score has relatively lower precision but delivers better recall, indicating its ability to identify relevant answers but with a higher false positive rate. Human evaluation, as expected, provides near-perfect precision and recall scores.

For the EVOUNA-TriviaQA results, a similar pattern is observed. Lexical Matching, GPT-3.5, and human evaluation maintain high precision across all QA models. BERT-Score sees a drop in precision but has comparable recall, especially with the TQ-ChatGPT35 and TQ-ChatGPT4. Again, human evaluation shows nearly perfect performance.

The results underscore the different strengths of the evaluation models: Lexical Matching for precision, BERT-Score for recall, and GPT-3.5 and human evaluation for both. However, all models' performance varies with the dataset and QA model, emphasizing the importance of multiple evaluation methods for comprehensive assessment.

E.3 Error Analysis in Open-QA

We classify the errors in the Open-QA scenario into several distinct categories:

- Inaccurate Information (InAcc): These errors occur when the model's response, while relevant to the question, contains inaccuracies.
- Incomplete Answer (InCom): This type of error is characterized by the model providing pertinent information but failing to fully address the question.
- Irrelevant Answer (IrrA): The model's response bears no relevance to the posed question.
- Outdated Information (OutInf): These errors occur when the model provides information that was correct at some point in the past but is no longer valid or applicable.
- Misinterpretation of the Question (MisQs): This category includes errors where the model misinterprets the question's intent or context.
- Other Errors: This catch-all category includes any errors that don't fit into the above classifications.

To perform this error classification, we initially used ChatGPT-4 to conduct a preliminary categorization of the Open-QA error data. Subsequently, human annotators were engaged to review and correct the classification results. The finalized results are represented in Table 9.

Analyzing the data reveals several interesting patterns. Notably, Bing Chat appears to have the highest rate of 'Incomplete Answer' errors, suggesting that while it generally understands the question, it often fails to provide a comprehensive answer. However, it also has the lowest rate of 'Inaccurate Information' errors, implying that the quality of the information it provides is usually high.

Conversely, DPR + FiD, GPT-3.5, and ChatGPT-3.5 all have similar rates of 'Inaccurate Information' errors, indicating a potential challenge in maintaining accuracy for these models. GPT-4 seems to outperform the other models in both 'Inaccurate Information' and 'Incomplete Answer' errors, suggesting an overall improvement in the quality and completeness of its responses.

It's also worth noting the relatively low incidence of 'Outdated Information' and 'Misinterpretation of the Question' errors across all models, suggesting that these areas are less problematic in current models.

This error analysis is helpful in identifying the strengths and weaknesses of different models and provides valuable insights into the areas that need further improvements.

E.4 Error Analysis in QA-Eval

E.4.1 Limitations of Each Evaluator

Based on our theoretical analysis and observations of erroneous cases, we identified the following issues with each type of evaluator:

Lexical Matching:

- Lack of Semantic Understanding: The exact match metric doesn't take into account the semantic meaning of the answers. It only checks if the predicted answer is exactly the same as the ground truth, even if the predicted answer is semantically correct but phrased differently.
- Inability to Handle Synonyms: The exact match metric cannot handle synonyms. If the predicted answer uses a different word that has the same meaning as the word in the ground truth answer, the exact match metric will consider it as a wrong answer.
- Inability to Handle Paraphrasing: Similar to the point above, the exact match metric cannot handle paraphrasing. If the predicted answer is a paraphrase of the ground truth answer, the exact match metric will consider it as a wrong answer.
- Inability to Handle Partially Correct Answers: The exact match metric cannot handle partially correct answers. If the predicted answer is partially correct, the exact match metric will consider it as a wrong answer.
- Inability to Handle Reordered Words: The exact match metric cannot handle reordered words. If the predicted answer has the same words as the ground truth answer but in a different order, the exact match metric will consider it as a wrong answer.
- Inability to Handle Different Levels of Detail: The exact match metric cannot handle different levels of detail. If the predicted answer provides more or less detail than the ground truth answer but is still correct, the exact match metric will consider it as a wrong answer.
- Inability to Handle Different Formats: The exact match metric cannot handle different formats. If the predicted answer is in a different format than the ground truth answer (for example, dates or numbers), the exact match metric will consider it as a wrong answer.

These limitations highlight the need for more sophisticated evaluation metrics that can understand the semantic meaning of the answers and handle synonyms, paraphrasing, partially correct answers, reordered words, different levels of detail, and different formats.

Neural Evaluation: The limitations of neural evaluation methods, such as BERT-Score and BLEURT, are evident. Most crucially, many neural evaluations are primarily designed to measure the similarity between two phrases or sentences. They are not tailored for binary tasks, especially those assessing the factual correctness of answers. Instead, they provide a continuous score that gauges the similarity between the generated text and the reference text, rendering them directly unsuitable for this particular task. In our study, we employed BERT-score and BLEURT for this task by setting a threshold. However, the performance of both BERT-score and BLEURT was suboptimal. The primary shortcoming of neural evaluations for this task is their misalignment with its requirements.

Furthermore, BERT-score has the following limitations:

-
- Sensitivity to Verbosity: BERT-score may penalize verbose answers even if they contain the correct information. If the AI-generated answer provides a detailed explanation while the golden answer is concise, the score might be lower than expected.
- Mismatched Focus: If the AI-generated answer is correct but emphasizes different aspects or details than the golden answer, BERT-score might not recognize the similarity, leading to a lower score.
- Lack of Contextual Understanding: BERT-score measures the similarity between embeddings but might not fully capture the contextual nuances of certain answers, especially when there are multiple valid ways to answer a question.
- Synonym and Paraphrasing Issues: BERT-score might not always recognize synonyms or paraphrased answers as being equivalent to the golden answer, leading to potential discrepancies in scoring.
- Threshold Limitations: Setting a fixed threshold (e.g., 0.5) for determining correctness can be arbitrary. Some answers might be just below the threshold but still be correct, while others might be just above but incorrect.
- Doesn't Account for Minor Details: BERT-score might not be sensitive enough to minor inaccuracies in the AI-generated answer, especially if the overall semantic content is similar to the golden answer.
- Lack of Absolute Truth Measure: BERT-score is a relative measure of similarity between two pieces of text. It doesn't provide an absolute measure of the truthfulness or correctness of an answer.
- Influence of Sentence Structure: The structure or order of sentences in the AI-generated answer compared to the golden answer might affect the score, even if the content is the same.
- Generalization Issues: BERT-score is based on pre-trained embeddings. It might not generalize well to niche topics or questions that require specialized knowledge outside of its training data.
- Over-reliance on Embeddings: While embeddings capture semantic information, they might not always capture the nuanced differences between two pieces of text, especially in a QA setting where precision is crucial.

In summary, while BERT-score is a powerful metric for evaluating text similarity, its application in a QA-eval task has limitations.

GPT-3.5 has its own set of limitations:

- Literal Interpretation: One of the limitations is the model's tendency to interpret questions or golden answers too literally. This can lead to situations where the evaluator fails to recognize correct answers that provide a broader context or a different interpretation that still addresses the core of the question.
- Overgeneralization: Another challenge is the model's propensity to overgeneralize based on its vast training data. This can result in the evaluator deeming an answer as correct even if it doesn't align specifically with the nuances of the question at hand.
- Misleading Emphasis: The evaluator might sometimes be swayed by partial correctness in an answer. If an answer emphasizes certain correct elements, the evaluator might overlook primary claims that are factually incorrect, leading to a misleading evaluation.
- Unknowable Reasoning: There are instances where the evaluator's judgment is puzzling, even to human experts. The model might deem an answer as correct that has no discernible correlation with the golden answer. This limitation underscores the "black-box" nature of deep learning models, where their internal reasoning processes remain opaque.

- Lack of Feedback Mechanism: Especially with closed-source models, there's a lack of a feedback loop to correct or fine-tune the model based on its evaluation errors. This can lead to repeated mistakes or biases in evaluation.
- Sensitivity to Prompt Engineering: Both closed-source and open-source LLMs can be sensitive to the way questions are framed or prompts are constructed. This can introduce variability in the evaluation, where slight rephrasings might lead to different judgments.
- Potential Bias: All LLMs, whether closed or open source, can inherit biases from their training data. In the context of QA-Eval, this might manifest as favoring certain types of answers or being biased against certain topics or contexts.

E.4.2 Error Categories

Based on the aforementioned limitations, we have designed a set of Evaluator Error categories. This includes two common errors found across all evaluators as well as specific errors unique to each type of evaluator.

General Error Categories for All Evaluators

- **Paraphrasing Error**: The evaluator fails to recognize answers that paraphrase the golden answer correctly but do not contain the exact substring.

  Example: Question: "What is the process by which plants convert sunlight into energy?" Golden Answer: "Photosynthesis" Generated Answer: "The mechanism plants use to transform light into energy is termed the photosynthetic process."

  Explanation: the generated answer is a paraphrase of the "Photosynthesis" but does not contain the word directly.
- **Synonym Error**: The evaluator fails to recognize answers that use synonyms or alternative phrasing to convey the same meaning as the golden answer.

  Example: Question: "What's another term for a doctor?" Golden Answer: "Physician" Generated Answer: "A medical practitioner."

  Explanation: "medical practitioner" is a synonym for "physician" but isn't a direct substring.

Specific Error Categories for Lexical Matching

- **Partial Match Error**: The evaluator fails to recognize answers that contain a part of the golden answer but not the entire substring.

  Example: Question: "Who painted the Mona Lisa?" Golden Answer: "Leonardo da Vinci" Generated Answer: "The Mona Lisa was painted by Leonardo."

  Explanation: only "Leonardo" is mentioned, not the full "Leonardo da Vinci".
- **Structure Variation Error**: The evaluator fails to recognize answers that essentially convey the same information as the golden answer but there's a variation in how it's structured.

  Example: Question: "When did 'Amnesia: The Dark Descent' come out?" Golden Answer: "8 September 2010" Generated Answer: "Amnesia: The Dark Descent was released on September 8, 2010."

  Explanation: the date format in the generated answer has an extra comma than the golden answer, even though the information is the same.
- **Overall Misleading Error**: The evaluator mistakenly recognizes the answer as correct because it contains a substring from the golden answer, even if the overall context of the answer is misleading.

  Example: Question: "Who wrote 'The Great Gatsby'?" Golden Answer: "F. Scott Fitzgerald" Generated Answer: "Ernest Hemingway and F. Scott Fitzgerald were close friends, but Hemingway wrote 'The Old Man and the Sea'."

Explanation: The generated answer contains the substring "F. Scott Fitzgerald", which might lead the Lexical Matching Evaluator to judge it as correct. However, the overall context of the answer is misleading, suggesting a relationship between Hemingway and "The Great Gatsby", which is incorrect.

Specific Error Categories for Neural Evaluation

- **Contextual Misunderstanding Error**: The evaluator might misjudge answers based on word embeddings and fail to capture the context in which certain words or phrases are used.

  Example: Question: "Who wrote 'Romeo and Juliet'?" Golden Answer: "William Shakespeare" AI-generated Answer: "Shakespeare wrote many plays."

  Explanation: Even though the AI answer mentions Shakespeare, it doesn't directly answer the question.

- **Threshold Sensitivity**: Answers that are just below the threshold might be correct but are judged as incorrect, and vice versa.

  Example: Question: "What's the capital of France?" Golden Answer: "Paris" AI-generated Answer: "The capital city of France is Paris."

  Explanation: The AI answer is correct but might score just below the threshold due to added context.

- **Extended Answer Error**: The evaluator might penalize answers that provide more context or details than the golden answer, even if they are correct, because the BERT-score only considers the similarities of the candidates and references.

  Example: Question: "Who painted the Mona Lisa?" Golden Answer: "Leonardo da Vinci" AI-generated Answer: "Leonardo da Vinci, a renowned Italian artist, painted the Mona Lisa."

  Explanation: The AI answer provides more context but is still correct.

Specific Error Categories for LLM-evaluator

- **Literal Interpretation Error**: The evaluator might take the question or golden answer too literally and fail to recognize correct answers that provide a broader context or interpretation.

  Example: Question: "Which bird is known for its beautiful tail?" Golden Answer: "Peacock" Generated Answer: "Many birds have beautiful tails."

  Explanation: The evaluator might take a literal approach and accept the general statement as correct without focusing on the specific bird in question.

- **Overgeneralization Error**: The evaluator might generalize based on its training data and judge an answer as correct even if it's not specific to the question.

  Example: Question: "Who wrote 'Pride and Prejudice'?" Golden Answer: "Jane Austen" Generated Answer: "An English author."

  Explanation: The evaluator might accept the general answer as it's not technically wrong, even though it lacks specificity.

- **Misleading Emphasis Error**: The evaluator might judge an answer as correct if it includes some correct information and put emphasis on it, and overlook the incorrect primary claim.

  Example: Question: "What's the primary gas in Earth's atmosphere?" Golden Answer: "Nitrogen" Generated Answer: "Oxygen, which makes up about 78% of the atmosphere."

  Explanation: GPT-3.5 might focus on the correct percentage and overlook incorrect mention of "Oxygen" as a primary gas.

- **Unknowable Reasons**: The evaluator makes an incorrect judgment for an unknowable reason. Even humans cannot figure out why the LLM thinks the generated answer is correct since it has no correlation with the golden answer.

Table 10: The error results for Lexical Matching evaluator, BERT-Score evaluator and GPT-3.5 evaluator. Each kind evaluator has common error types and specific error types. General error rate indicates the error proportion of this evaluator on this subset.

	NQ-FiD	NQ-GPT35	NQ-ChatGPT35	NQ-ChatGPT4	NQ-BingChat
Paraphrasing Error	29%	37%	29%	60%	49%
Synonym Error	18%	12%	37%	12%	19%
Partial Match Error	48%	30%	13%	10%	20%
Structure Variation Error	4%	16%	15%	12%	7%
Overall Misleading Error	1%	5%	6%	6%	5%
Lexical Matching: General error rate	11.75	15.2	19.7	16.8	17.7
	NQ-FiD	NQ-GPT35	NQ-ChatGPT35	NQ-ChatGPT4	NQ-BingChat
Paraphrasing Error	4%	24%	29%	39%	39%
Synonym Error	4%	7%	4%	5%	5%
Contextual Misunderstanding Error	63%	22%	23%	20%	15%
Threshold Sensitivity Error	25%	33%	20%	18%	15%
Extended Answer Error	4%	14%	24%	18%	26%
BERT-Score: General error rate	25.0	30.5	27.2	23.2	32.4
	NQ-FiD	NQ-GPT35	NQ-ChatGPT35	NQ-ChatGPT4	NQ-BingChat
Paraphrasing Error	16%	52%	36%	52%	47%
Synonym Error	22%	12%	21%	18%	17%
Literal Interpretation Error	21%	4%	11%	6%	13%
Overgeneralization Error	17%	13%	8%	8%	6%
Misleading Emphasis Error	7%	2%	5%	3%	6%
Unknowable Reasons Error	17%	8%	19%	13%	11%
GPT3.5: General error rate	6.4	16.0	17.8	16.6	30.5

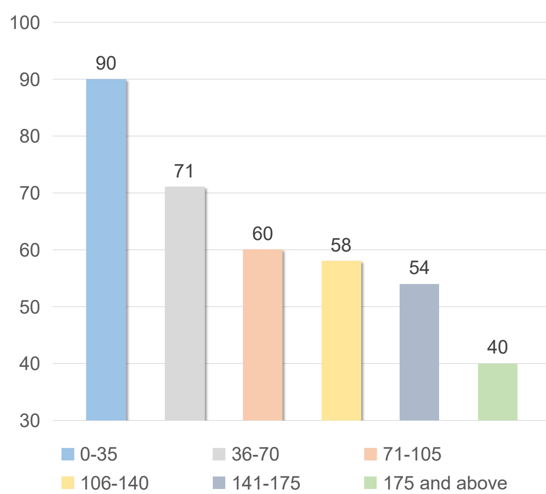

Figure 3: Correlation between the evaluation accuracy of GPT-3.5 and the answer length in tokens across all models.

Example: Question: "Who was the first chief minister of West Bengal?" Golden Answer: "Prafulla Chandra Ghosh" Generated Answer: "The first Chief Minister of West Bengal was Dr. Bidhan Chandra Roy."

Explanation: GPT-3.5 takes the generated answer as correct, but Dr. Bidhan Chandra Roy is apparently not Prafulla Chandra Ghosh.

Table 11: GPT-3.5 evaluator performance with different prompt strategies on the EVOUNA-NQ set. Each cell displays accuracy (left) and F1 score (right).

	NQ-FiD	NQ-GPT35	NQ-ChatGPT35	NQ-ChatGPT4	NQ-BingChat
Original	93.6/95.3	83.7/86.8	82.2/86.9	84.5/89.7	69.7/77.2
Ignoring Background	93.7/95.3	82.8/85.9	80.8/85.5	81.1/87.1	65.7/73.4
Giving Reasons	89.6/91.9	76.3/78.5	73.2/78.2	67.9/75.8	55.6/62.2
Chain-of-Thoughts	84.4/88.1	**84.9/88.4**	**84.2/89.0**	**88.7/93.0**	**80.2/86.9**
In-Context-Learning	93.2/95.0	84.5/88.3	83.3/88.0	86.3/91.2	75.1/82.3

E.4.3 Length Analysis on QA-Eval

Figure 3 depicts the relationship between GPT-3.5's evaluation accuracy and the number of tokens present in the answers produced by all models. The token count is segmented into six distinct categories: 0-35, 36-70, 71-105, 106-140, 141-175, and 175 and above. The corresponding accuracy for these ranges are 90, 71, 60, 58, 54, and 40 respectively. Additionally, the average token counts for the answers by each model are as follows: FiD (4.8 tokens), GPT-3.5 (31.4 tokens), ChatGPT (41.9 tokens), GPT-4 (39.9 tokens), and BingChat (49.7 tokens).

We can draw several observations: 1. GPT-3.5's evaluation accuracy exhibits an inverse correlation with the length of the answer. As the number of tokens in the answer escalates, the evaluation accuracy diminishes. This could indicate that GPT-3.5 may struggle to accurately evaluate more extended responses, potentially due to challenges in retaining context or comprehending intricate or unfamiliar constructs in longer text spans. 2. Considering the average token counts, FiD, the model that generates the shortest responses on average (4.8 tokens), would predominantly fall into the 0-35 token range where GPT-3.5 has its peak accuracy (90). This observation could imply that GPT-3.5 would exhibit optimal evaluation performance with responses generated by the FiD model. 3. Conversely, models like Bing Chat, which on average yield longer responses (49.7 tokens), would generally fall into the token ranges where GPT-3.5's evaluation accuracy is lower. This can partially explain why GPT-3.5 performs worse than Lexical Matching in NQ-BingChat and TQ-BingChat.

E.5 Enhancing QA-Eval through Prompt Engineering

We also examine strategies to improve LLM' (specifically, GPT-3.5) performance in QA-Eval via prompt engineering. Four distinct methods were explored: Ignoring Background Information; Providing Reasons for Judgments; Chain of Thoughts [Wei et al., 2022]; In-Context Learning [Dong et al., 2023].

Table 12 outlines the specific prompts used for each method with GPT-3.5 in QA-Eval. The prompts are designed to elicit different model behaviors or responses.

We adopt an approach from Auto-Cot [Zhang et al., 2023] using K-Means clustering [Hartigan and Wong, 1979] to select representative examples for in-context learning. To avoid data leakage, we employ cross-domain clustering; we cluster NQ sets for TQ experiments and vice versa. For example, we select representative examples from NQ-ChatGPT4 for experiments on TQ-ChatGPT4. Four representative examples are chosen for each dataset.

Table 11 presents the performance of GPT-3.5 evaluator with different prompts on the EVOUNA-NQ dataset. Here are the insights: Directing GPT-3.5 to ignore the background information degrades performance on four datasets with long answers (NQ-GPT35/ChatGPT35/ChatGPT4/BingChat). Requiring the model to reason its judgments negatively impacts performance across all datasets. The effects of Chain-of-Thoughts and In-Context-Learning vary. For instance, both methods significantly improve performance on four datasets with long answers, but Chain-of-Thoughts shows a substantial decline on the NQ-FiD. This variability suggests that the influence of these techniques depends on the data distribution.

Table 12: Specific prompts used in each method for GPT-3.5 on QA-Eval.

Methods	Prompts
Original	Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers, simply answer Yes or No
Ignoring Background	Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers, please only consider the answer itself, ignore the background information. Simply answer Yes or No.
Giving Reasons	Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers. Please make a judgment and give the reason. Your answer must be <Yes or No>|<Reason>
Chain-of-Thoughts	Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers. Please think step by step and make a judgment in the end. You must give your chain of thoughts. Your answer must be <your chain of thoughts>|<Yes or No>. (chain of thoughts and final judgment must be split with '|')
In-Context-Learning	Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers, simply answer Yes or No. Here are some examples: Example 1: AAA; Example 2: BBB; Example 3: CCC; Example 4: DDD.

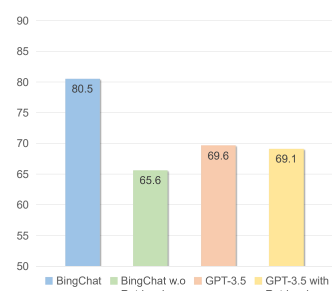

Figure 4: The performance of Bing Chat and GPT-3.5 on NQ set with or without retrieval.

E.6 Does retrieval Help in LLM?

In our quest to determine the impact of retrieval on Large Language Models (LLMs) in an Open-QA setting, we investigate two distinct scenarios. Firstly, we assess the performance of Bing Chat when retrieval is disabled. Secondly, we augment GPT-3.5 with a retrieval mechanism and gauge its effectiveness.

Performance of Bing Chat Without Retrieval In this experiment, we modify the standard prompt fed to Bing Chat by preceding the question q with the instruction "Please do not search, answer the

Table 13: Performance of BERT-Score and BLEURT on the EVOUNA. In each cell, the left is the accuracy while the right is the Macro-F1.

	NQ-FiD	NQ-GPT35	NQ-ChatGPT35	NQ-ChatGPT4	NQ-BingChat
Lexical Matching	86.9/86.0 89.6/88.8	84.8/84.3	80.3/78.2	83.2/78.1	82.3/77.7
BERT-Score	75.0/66.0	69.5/64.8	72.8/66.0	76.8/65.8	67.6/59.5
BELURT	84.4/79.9	74.1/63.9	78.0/64.9	85.0/66.3	82.8/65.0
GPT-3.5	93.6/92.6	84.0/83.0	82.2/79.5	83.4/77.2	69.5/65.5
Another Human	**96.3/95.6**	**96.8/96.2**	**95.6/95.2**	**96.6/94.4**	**95.5/93.2**
on EVOUNA-NaturalQuestions					
	TQ-FiD	TQ-GPT35	TQ-ChatGPT35	TQ-ChatGPT4	TQ-BingChat
Lexical Matching	90.0/86.0 91.8/88.2	92.3/89.6	92.3/87.7	91.1/81.3	89.8/79.3
BERT-Score	65.4/59.6	75.7/66.6	80.7/65.4	83.4/62.7	80.4/63.9
BELURT	88.1/77.8	82.9/66.6	85.2/66.1	88.8/66.2	90.8/64.7
GPT-3.5	95.7/93.2	91.2/88.3	92.7/87.2	92.5/82.2	81.2/69.0
Another Human	**99.7/99.8**	**99.4/99.6**	**98.8/99.2**	**99.8/99.9**	**99.8/99.9**
on EVOUNA-TriviaQA					

following question directly:". We choose a sample of 500 questions from the NQ test dataset, filtering out those unsuitable for this setting. The results of this experiment are depicted in the left section of Figure 4.

The data suggests a significant decline in Bing Chat's performance when retrieval is disabled, dropping approximately 15 percentage points from 80.5 to 65.6. This is comparable to the performance of GPT-3.5 (65.0), which lacks a retrieval mechanism. This substantial decline implies that the retrieval component significantly boosts the performance of the LLM underpinning Bing Chat in an Open-QA context.

Augmenting GPT-3.5 with a Retrieval Mechanism For the second scenario, we employ the same Dense Retriever used in the DPR+FiD model (referenced in Section3.2) to fetch relevant passages from the database for a given question. We then integrate these passages into the prompt supplied to GPT-3.5. The prompt reads: "We have a question here: QUESTION. Now, we have the following relevant passages: PASSAGE 1; PASSAGE 2; PASSAGE 3; PASSAGE 4; PASSAGE 5. Please answer the question referring to the above passages."

The results of this experiment, shown in the right section of Figure 4, reveal a slight decrease in performance with the addition of retrieval, falling from 69.6 to 69.1. This suggests that simply injecting retrieved passages into the prompts, without any form of thoughtful adaptation, does not contribute positively to the LLM's performance in an Open-QA setting.

E.7 BLEURT Evaluator

We also conducted a QA-Eval analysis on a more recent Neural-Evaluation model, BLEURT [Sellam et al., 2020]. Similar to BERT-Score, we applied a threshold to BLEURT to make it suitable for QA-Eval. In this work, we set the threshold at 0.2 based on observed distributions. The results are shown in the Table 13. Although BLEURT outperforms BERT-Score on most datasets, it still lags significantly behind the performance of Lexical Matching, GPT-3.5 and human, especially in terms of Macro-F1.

E.8 Additional Open-QA Models

We have conducted experiments on more transparent Open-QA models, including Atlas [Izacard et al., 2022], Llama-2 [Touvron et al., 2023], Chat-Llama-2 [Touvron et al., 2023] on 500 samples on NQ test subset. During our experiments, we notice that the base version of LLaMa-2 occasionally

Table 14: Open-QA and QA-Eval results of Atlas and Chat-Llama2 on 500 samples of NQ. In each cell, the left is the accuracy while the right is the Macro-F1.

	NQ-Atlas	NQ-ChatLlama2
Lexical Matching	92.6/92.5	89.5/88.2
BERT-Score	67.1/65.8	68.2/68.0
GPT-3.5	64.7/63.9	66.1/53.6
Human Score on NQ-Atlas: 47.9; Human Score on NQ-ChatLlama2: 29.7		

Table 15: Error results of Eval-Models on the EVOUNA. In each cell, the left is the error rates while the right is the times compared with another human results.

	NQ-FiD	NQ-GPT35	NQ-ChatGPT35	NQ-ChatGPT4	NQ-BingChat
Lexical Matching	13.1/3.5x 10.4/2.8x	15.2/4.8x	19.7/4.5x	16.8/4.9x	17.7/3.9x
BERT-Score	25.0/6.8x	30.5/9.5x	27.2/6.2x	23.2/6.8x	32.4/7.2x
GPT-3.5	6.4/1.7x	16.0/5.0x	17.8/4.0x	16.6/4.9x	30.5/6.8x
Another Human	**3.7/1.0x**	**3.2/1.0x**	**4.4/1.0x**	**3.4/1.0x**	**4.5/1.0x**
on EVOUNA-NaturalQuestions					
	TQ-FiD	TQ-GPT35	TQ-ChatGPT35	TQ-ChatGPT4	TQ-BingChat
Lexical Matching	10.0/33.3x 8.2/27.3	7.7/12.8x	7.7/6.4x	8.9/44.5x	10.2/51.0x
BERT-Score	34.6/115.3x	24.3/40.5x	19.3/16.1x	16.6/83.0x	19.6/98.0x
GPT-3.5	4.3/14.3x	8.8/14.7x	7.3/6.1x	7.5/37.5x	18.8/94.0x
Another Human	**0.3/1.0x**	**0.6/1.0x**	**1.2/1.0x**	**0.2/1.0x**	**0.2/1.0x**
on EVOUNA-TriviaQA					

deviated from our instructions. As a result, we chose to proceed with Chat-Llama-2 for a more consistent evaluation. The results are shown in Table 14.

It's evident from the results that the performance of ATLAS and Chat-Llama2 is somewhat below the models discussed in our paper. Moreover, the evaluators' performance on NQ-Atlas and NQ-ChatLlama2 is consistent with the trends observed for the models we initially discussed.

F Additional Related Work

Hashimoto et al. [2019] have also studied the correlations between human evaluation and automated metrics in NLP. However, there are key differences that set our research apart. First, We only discuss the Open-QA task, underscoring the nuances and challenges specific to this domain, while their research casts a wider net, aiming to bridge the gap between human and automated evaluation methods across various natural language generation tasks. Second, there are different emphasis on Human Evaluation, We introduce the EVOUNA dataset, which is enriched with human-annotated results, providing a fresh perspective on evaluation in the Open-QA domain, while They advocate for a unified framework that correlates human judgments with statistical metrics, offering a holistic approach to evaluation in NLP. Last, we present the QA-Eval task and the EVOUNA dataset, tailored specifically for evaluating Open-QA systems, while heir research offers a comprehensive framework designed for a broader spectrum of natural language generation tasks.