PeerQA: A Scientific Question Answering Dataset from Peer Reviews

Tim Baumgärtner,[1] **Ted Briscoe,**[2] **Iryna Gurevych**[1,2]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt
[2]Mohamed bin Zayed University of Artificial Intelligence
www.ukp.tu-darmstadt.de

Abstract

We present PeerQA, a real-world, scientific, document-level Question Answering (QA) dataset. PeerQA questions have been sourced from peer reviews, which contain questions that reviewers raised while thoroughly examining the scientific article. Answers have been annotated by the original authors of each paper. The dataset contains 579 QA pairs from 208 academic articles, with a majority from ML and NLP, as well as a subset of other scientific communities like Geoscience and Public Health. PeerQA supports three critical tasks for developing practical QA systems: Evidence retrieval, unanswerable question classification, and answer generation. We provide a detailed analysis of the collected dataset and conduct experiments establishing baseline systems for all three tasks. Our experiments and analyses reveal the need for decontextualization in document-level retrieval, where we find that even simple decontextualization approaches consistently improve retrieval performance across architectures. On answer generation, PeerQA serves as a challenging benchmark for long-context modeling, as the papers have an average size of 12k tokens.[1]

1 Introduction

The number of scientific articles is increasing exponentially (Fire and Guestrin, 2019; Bornmann et al., 2020), leading to an increase in review work and leaving researchers with an ever-expanding number of publications to read to keep up with their field. Therefore, novel tools are required to support reviewing work and enable readers to consume information from scientific articles more efficiently (Brainard, 2020; Kuznetsov et al., 2024). Automatic Question Answering (QA) systems can provide such support, allowing researchers and reviewers to productively extract information from an article, particularly if integrated directly into the reading and reviewing interface (Zyska et al., 2023; Lo et al., 2024). QA systems can also improve the quality of peer review, e.g., by avoiding questions in a review that are addressed in the article but potentially overlooked by a reviewer. However, the development of QA models is limited by the availability of high-quality and realistic datasets in the scientific domain to measure the performance of methods. Collecting scientific QA data is challenging because it requires expert annotators who are difficult to recruit. Furthermore, naturally occur-

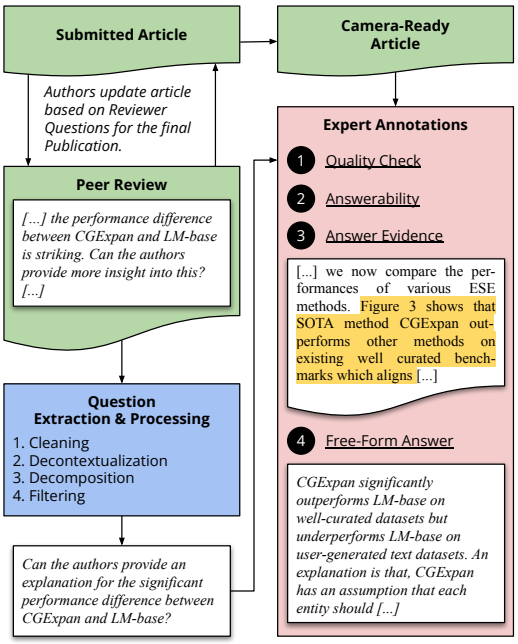


Figure 1: Overview of the PeerQA data collection process. From the peer review process (in green), we extract and process questions from the reviews. Given the published version of the article and a question, an expert (in our case, the original paper authors) (1) checks the question and modifies or discards it, (2) annotates whether it is answerable or not (i.e. if there is sufficient information in the paper), and if so (3) highlights the evidence to answer the question and finally (4) provides a free-form answer to the question.

[1]Our code and data is available at https://github.com/UKPLab/peerqa.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 508–544
April 29 - May 4, 2025

Dataset	Papers	QA	Domain	Questioner Knowledge	Question Source	Annotators	Answer Source	Answer Types
BioASQ	43011	4615	BioMed.	–	–	Experts	Abstract	Y/N, Ex, FF
QASPER	1585	5049	NLP	Abstract	Crowdsourced	Practitioners	Paper	Y/N, Ex, FF, U/A
QASA	113	1798	AI/ML	Full Paper	Crowdsourced	Practitioners	Paper	Ex, FF, U/A
PeerQA	208	579	Multi	Full Paper	Reviews	Experts	Paper	Ex, FF, U/A

Table 1: Comparison of the most relevant scientific QA datasets. In BioASQ, experts come up with questions without a document in mind. Answer types abbreviations: Y/N = Yes/No, Ex = Extractive or Evidence Retrieval, FF = Free-Form Answers, U/A = Unanswerable). The QA column reports the number of question-answer annotations.

ring questions are difficult to source compared to the general domain, where search engine logs can be used (Nguyen et al., 2016; Kwiatkowski et al., 2019). Previous work resorted to recruiting practitioners or graduate students and focused only on Machine Learning (ML) or Natural Language Processing (NLP) domains (Dasigi et al., 2021; Lee et al., 2023). Annotators of these datasets have various degrees of knowledge, e.g., having read only the abstract, skimmed the paper, or sometimes read the paper fully. Collecting questions from annotators has the downside of questions not being realistic, such as asking questions that would not be raised naturally or being generic when the questioner has superficial knowledge of the paper.

To this end, we introduce PeerQA, a real-world, scientific, document-level Question Answering dataset. PeerQA supports three crucial tasks for QA over scientific articles: Given a question and a paper, evidence sentences relevant to the question need to be retrieved. Based on these, the answerability of the question can be decided. Finally, the dataset contains free-form reference answers addressing the question. We leverage peer reviews to source questions, and answers are annotated by the authors of the respective papers. While most questions are from ML and NLP papers, 10% of questions come from other scientific domains, including Geoscience and Public Health. Figure 1 provides an overview of our data collection process. To summarize, our contributions are the following:

1. We release PeerQA, a QA dataset over scientific articles with questions sourced from peer reviews and answers annotated by authors. We release a set of 579 annotated samples (from 208 papers), as well as 12k unlabeled questions (from 2.6k papers). We show the properties of the collected data, including various statistics, question topics, and classes.

2. We establish baselines for all three tasks in PeerQA: Evidence Retrieval, Question Answerability, and Free-Form Answer Generation, and outline which factors contribute to model performance.

2 Related Work

Peer Review Many tasks and applications leverage peer reviews as a data source, including argument mining (Hua et al., 2019; Cheng et al., 2020; Kennard et al., 2022), helpfulness and score prediction (Xiong and Litman, 2011; Gao et al., 2019), review generation (Yuan et al., 2022; D'Arcy et al., 2024), tagging and linking review comments with the paper (Kuznetsov et al., 2022; D'Arcy et al., 2024), rebuttal generation (Purkayastha et al., 2023), the study and analysis of peer review (Kang et al., 2018; Ghosal et al., 2022) and more general contexts such as document revision (Ruan et al., 2024). In PeerQA, we utilize peer reviews to source a scientific QA dataset.

Scientific QA QA datasets in the scientific domain can generally be categorized as larger-scale datasets that are (semi-) automatically created and small expert-annotated datasets.

Among the larger-scale but (semi-) automatically created QA datasets are PubMedQA (Jin et al., 2019), in which questions are sourced from article titles that are phrased as questions. Answers are either yes, no, or maybe, and a subset is expert-annotated. SciDefinition (August et al., 2022) uses templates to generate questions about the definition of scientific terms. Kulshreshtha et al. (2021) create a dataset in the ML and Biomedicine domain with questions sourced from Google's "People also ask" suggestions and answers from the search engine's span extraction feature. Wan et al. (2024) generate a large-scale, scientific QA dataset by distilling a generation model from GPT-4 instructed to output QA pairs given a paper. Auer et al. (2023) develop question templates to automatically generate questions that are answerable from the Open Research Knowledge Graph (Jaradeh et al.,

Venue	Domain	Papers	Questions	Evidence	Free-Form	Ev. & FF.	Unanswerable
ICLR 23	ML	49	153	103	89	63	36
ICLR 22	ML	44	137	107	75	68	14
NeurIPS 22	ML	25	79	56	51	40	16
ARR 22	NLP	45	87	60	61	49	21
COLING 20	NLP	15	31	25	13	13	5
ACL 17	NLP	7	20	16	10	9	4
CoNLL 16	NLP	5	12	7	7	7	5
ESD 23	Geoscience	5	17	10	11	5	1
ESurf 23	Geoscience	3	16	16	9	9	0
F1000 22	Mixed	10	27	14	10	4	10
Total		**208**	**579**	**414**	**336**	**267**	**112**

Table 2: Number of collected question-answer pairs per venue in PeerQA. *Evidence* shows the number of questions that have at least one sentence annotated addressing the question. *Free-Form* reports the number of questions with an annotated free-form answer. The *Ev. & FF.* column reports the union of both. Finally, the *Unanswerable* column reports the number of questions that can not be answered due to insufficient information in the paper.

2019) covering factoid questions, e.g., about the metadata of a paper, or questions that require inference over multiple papers. The questions in PeerQA are all focused on a single publication and the content of it, and our baselines use only the unstructured text of the article. PeerQA is an expert-annotated QA resource, where questions are sourced from human-written peer reviews and answers are annotated by paper authors.

Regarding expert annotated datasets, the BioASQ challenge (Tsatsaronis et al., 2015; Krithara et al., 2023) is an open-domain QA dataset from biomedical experts. Experts come up with questions and corresponding answers (yes/no, factoid, list, and free-form), which are additionally grounded in sentences from abstracts of publications on PubMed. While this is one of the greatest available resources for biomedical QA, annotating answers only in abstracts limits the question and answer complexity. Compared with PeerQA, questions are also more general, i.e., they are not asked within the context of a specific paper, and answers can be found in various articles. Most similar to our work are the QASPER (Dasigi et al., 2021) and QASA (Lee et al., 2023) datasets. In QASPER, NLP practitioners have read the abstract of a paper and raised questions about the paper. This leads to generic questions applicable to many papers (e.g., "Which baselines did they compare?") and questions that are easy to answer from the full paper. QASA takes this a step further by giving question annotators access to the full paper, instructing them to either skim or read it in more detail. In both these datasets, annotators create questions and answers; in contrast, our questions are based on peer reviews, i.e., they have been naturally raised by a reviewer, a domain expert who has read the paper in detail. Besides the questions, the answers in PeerQA are provided by experts, i.e., the authors of the respective papers. Table 1 provides an overview of these differences. To summarize, PeerQA is the first scientific QA resource with natural questions and all QA pairs annotated by paper authors.

In concurrent work, Singh et al. (2024) also explore extracting questions from peer reviews in the ML domain. Unlike PeerQA, their approach uses the authors' responses provided during the rebuttal to obtain reference answers. To identify supporting evidence from the paper for each answer, they employ a hybrid approach that combines manual and automated mapping of the answers to relevant information in the paper.

Long-Context QA Dialogue and QA systems grounded in a document have recently gained traction (Muresan et al., 2023). In this vein, NarrativeQA (Kočiský et al., 2018) contains questions about movie scripts and books with an average length of 63k tokens. Pang et al. (2022) construct a multiple-choice dataset over books and articles with an average length of 5k tokens focusing on questions that require reading the article in detail. Reddy et al. (2024) extend FinQA (Chen et al., 2021b) to financial documents with an average of 123k words. ConditionalQA (Sun et al., 2022) is a dataset of government documents with an average length of 1.5k tokens and answers tied to certain input conditions. PeerQA serves as another resource for long-context QA, with documents having an average length of 12k tokens and 30% of questions requiring combining information from more than one location in the paper.

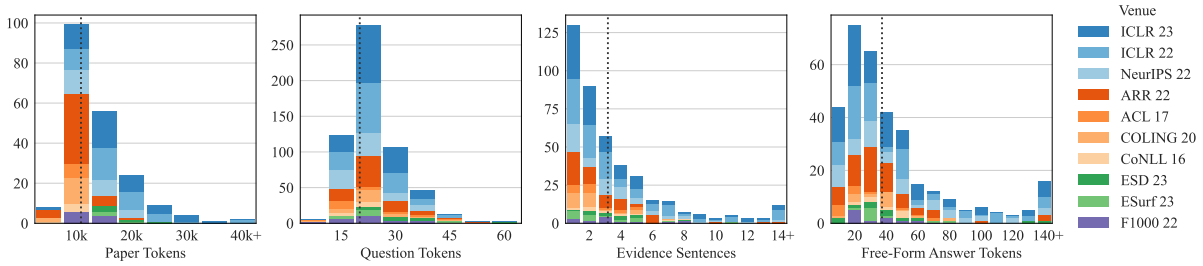


Figure 2: Statistics of the PeerQA dataset. The color coding shows the distribution per venue and by the scientific community (i.e., blue colors for ML, orange for NLP, green for Geosciences, and purple for mixed). The gray dotted line indicates the average. The leftmost histogram shows a paper distribution, while the others show a distribution of questions. We measure the number of tokens using the `Llama-3` tokenizer.

3 PeerQA

3.1 Data Collection

Figure 1 provides an overview of the data collection process. We use papers and peer reviews from NLPeer (Dycke et al., 2023) and extend this set with journals and conferences that publish peer reviews and camera-ready versions publicly. Specifically, the data from ARR 2022 (containing papers published at ACL and NAACL 2022), COLING 2020, ACL 2017, CoNLL 2016, and F1000 was curated in NLPeer, partially based on previous data collections (Kang et al., 2018; Kuznetsov et al., 2022) and published under a CC-BY-NC-SA 4.0 license. The data from the Geoscience domain is published under a CC-BY 4.0 license in two journals: Earth System Dynamics[2] (ESD) and Earth Surface Dynamics[3] (ESurf). For ICLR 2022/2023 and NeurIPS Datasets and Benchmark Track 2022, we retrieve papers and reviews from OpenReview. Since they are without any license, we do not publish them in our release but provide a download and processing script. All questions and answers in PeerQA are published under CC-BY-NC-SA 4.0.

Paper Processing. We extract the full text of the camera-ready version of a publication, including equations and captions, using GROBID 0.8 (Lopez, 2008–2024), which also groups sentences into paragraphs, which we use later in our experiments.

Question Processing. From the peer reviews of each paper, we extract an initial set of questions using all sentences ending in a question mark, resulting in 17910 questions.[4] The resulting questions comprise three problems: First, they are noisy as peer reviews often contain spelling or grammar mistakes. Second, they are contextualized into the preceding sentences of the review, i.e., their actual meaning can only be understood from the context of the review but not in isolation. Third, some questions contain compounds of multiple or follow-up questions after applying the decontextualization step. We deemed this problematic for our annotations as it would obfuscate which evidence aligns with which part of the question. To address these issues, we conduct two preprocessing steps: First, we create a clean and contextualized version of a question using InstructGPT[5] (Ouyang et al., 2022). For this, we prompt the model with the preceding three sentences of the review and the extracted question to generate a single question that is context-independent. Conveniently, due to the good fluency of Large Language Models (LLM), this also addresses the noisiness of the original question. To detect multiple or follow-up questions, we employ a constituency parser (Kitaev and Klein, 2018; Kitaev et al., 2019) and flag questions with root-level conjunctions. We then decompose these questions adopting InstructGPT again.

Finally, we manually filter all resulting questions to include only information-seeking types of questions and discard questions that contain errors due to the preprocessing steps or not being relevant for a QA dataset. Specifically, we ensure that questions address the *content* of the paper (e.g., we discard questions of rhetorical nature or about typos and layout) and are *decontextualized* correctly (i.e., we discard questions that are ambiguous, contain hallucinations or references such as line numbers that are not present in a camera-ready version).[6] In this

[2] `https://www.earth-system-dynamics.net`

[3] `https://www.earth-surface-dynamics.net`

[4] In preliminary experiments, we extracted questions based on syntax. However, this resulted in many false positives.

[5] We use `text-davinci-003`. However, when we added the Geoscience subset, `text-davinci-003` was no longer available. Thus, we resorted to `gpt-4-0125-preview`.

[6] This filtering step has largely been done by a graduate

step, we remove 30% of the questions, yielding the final set of 12546 questions.

Answer Annotation. Our questions were asked based on the submitted article. However, answers are annotated in the final publication. Hence, our annotation process relies on authors incorporating reviews into the final version for questions to be answerable. Questions might also be answerable when reviewers overlooked details in the submission that already answer their questions. For each paper, we contact paper authors via email requesting their voluntary participation in answering the questions (see §K.1).[7] We implement multiple layers to instruct authors on how to complete the task. First, we provide a high-level description of the task in the initial email and a link to the detailed annotation guideline. We updated the annotation guideline during data collection with common questions we received. Moreover, we explain the annotation interface and demonstrate the task in a short video. Finally, our annotation interface (see §K.2) also contains UI elements that provide hints to the authors explaining the task. The annotation task comprises 4 steps: First, authors can provide feedback on a question, e.g., to remove or update it. Second, the authors highlight any text in the PDF of the final paper that is relevant to answering the question, which we refer to as Answer Evidence. Third, the authors provide free-form text that directly answers the question. Alternatively, questions can also be flagged as unanswerable. Unanswerable questions can, for example, occur when a question from a reviewer has been answered in the rebuttal but was not incorporated into the final publication. While we ask authors to perform all steps, some questions only have answer evidence or a free-form answer, but not both. The annotated evidence is mapped to the text extracted from the PDF. We notice that GROBID occasionally misses paragraphs that can not be mapped to the annotated evidence. We publish the raw annotated data and the mapped data, allowing future research with access to better PDF extraction tools to use the full dataset.

Quality Control Besides manually filtering questions and removing low-quality or irrelevant ones, we also provide the experts with a way to improve the dataset's quality. In our annotation interface, authors can leave feedback for a question, e.g., if they find it imprecise and wish to correct or remove it. All feedback has been manually processed, and the questions have been updated or removed. Finally, we notice a high variance in the free-form answer quality. While some answers are clear and concise, others are more succinct and provide less detail. Although we give detailed guidelines on how to write the free-form answer to the authors, since we only engage briefly with them, it is challenging to enforce a similar quality. To counter this, we augment the collected answers with rephrases from GPT-4 (OpenAI et al., 2023).[8]

Following this process, we obtained 579 answers from 208 papers. Table 2 reports the number of annotations per venue. We also release the remaining 11967 questions from 2623 papers that have not been answered.[9]

3.2 Analysis

We report distributional statistics of the dataset in Figure 2. Notably, the average paper length is 11723 tokens, which provides an interesting benchmark for long-context generative models. Furthermore, questions are relatively long, with an average of 20.2 tokens (the average length in BioASQ, QASPER, and QASA is 13.2, 10.2, and 17.7, respectively). One reason for this is the question processing pipeline, particularly the decontextualization step. Reviewers construct questions potentially consisting of multiple sentences. During preprocessing, the question has been rephrased to contain all this information. We analyze the semantic similarity between the final and original questions, finding that 90% of questions have a similarity of more than 0.6 and 50% more than 0.82.[10] This shows that our processed questions remain highly similar to the original questions in the review. On average, questions have 3.8 annotated answer evidence sentences. Besides, 30% of questions have non-consecutive answer evidence, i.e., the evidence is distributed non-contiguously over the paper.[11]

We run a topic model to understand which questions are contained in PeerQA, specifically BERTopic (Grootendorst, 2022). We find

NLP student supported by the paper authors.

[7] For the 5 CoNLL papers, we were unsuccessful in contacting the authors. Therefore, the annotations were performed by a senior NLP professor and co-author of this paper.

[8] See §B.3 and §B.4 for prompts.

[9] The number of mapped evidence from the noisy text extraction is reported in §A. Examples are provided in §N. §D reports a breakdown by venue for the unlabeled questions.

[10] §C provides a detailed analysis of the similarities.

[11] §E reports more answer evidence statistics.

Model	Architecture	MRR				Recall@10			
		Para.	+Title	Sent.	+Title	Para.	+Title	Sent.	+Title
MiniLM-L12-v2	Cross-Encoder	**0.4723**	0.4839	**0.3644**	**0.3654**	0.6467	0.6709	0.3505	**0.3746**
Contriever	Dense	0.3494	0.3624	0.2778	0.2773	0.5567	0.5340	0.2896	0.2910
Contriever-MS	Dense	0.4095	0.4408	0.3184	0.3160	0.6160	0.6314	0.3361	0.3538
Dragon+	Dense	0.4657	**0.4845**	0.3345	0.3433	0.6563	0.6817	0.3637	0.3667
GTR-XL	Dense	0.3955	0.4142	0.3048	0.2981	0.5940	0.6122	0.3522	0.3190
ColBERTv2	Multi-Dense	0.4368	0.4122	0.3480	0.3491	0.6287	0.6371	0.3607	0.3544
BM25	Sparse	0.4288	–	0.2850	–	0.6388	–	0.3058	–
SPLADEv3	Sparse	0.4536	0.4725	0.3477	0.3419	**0.6661**	**0.6851**	**0.3757**	0.3687

Table 3: Answer evidence retrieval results on paragraph (Para.) and sentence (Sent.) level and with decontextualizing the passages by prepending the title (+Title). Top-scoring models are in bold, and runner-ups are underlined.

community-specific clusters (e.g., mentions of *language* or *annotation* for NLP; *carbon* or *soil* for Geoscience), topics about specific elements of the paper (e.g., figures, tables, or equations) or specialized clusters (e.g., adversarial attacks or fine-tuning/hyperparameter related questions).[12] While the topic analysis clusters questions semantically, we also sample 100 questions randomly and sort them into one of 8 question classes: Methods, Data, Implications, Definitions, Comparisons, Analysis, Justification, and Evaluation.[13] We find that 44% of questions aim to clarify the methods or data, followed by 12% of questions asking the authors to justify a decision.[14]

4 Experiments

4.1 Answer Evidence Retrieval

We set up the answer evidence retrieval task as an information retrieval problem: Given a query, the model computes a score for each passage in the paper, where a passage can be a paragraph or sentence. To evaluate the answer evidence retrieval task, we test models of various architectures, including cross-encoder (Nogueira and Cho, 2019), dense retrieval (Reimers and Gurevych, 2019), multi-vector dense retrieval (Khattab and Zaharia, 2020), sparse (Zamani et al., 2018) and lexical models. Specifically, a cross-encoder model concatenates the query and passage and outputs a relevance score. In contrast, dense approaches encode query and passage independently by the same or individual models, obtaining a high-dimensional representation for each. A score is computed via dot-product or cosine-similarity between the two representations. Multi-vector approaches represent a query and passage not by a single but by many representations, e.g., for each token. The relevance score is computed by taking the sum of the maximum score between each query and passage token. Lexical approaches use term matching and weighting between the query and passage. Building upon this, sparse models perform a semantic query and/or document expansion to overcome the lexical gap. Concretely, we evaluate: `MiniLM-L12-v2` (Wang et al., 2020; Thakur et al., 2021), `Contriever` (Izacard et al., 2022), `Dragon+` (Lin et al., 2023), `GTR` (Ni et al., 2022), `ColBERTv2` (Santhanam et al., 2022), `BM25` (Robertson and Zaragoza, 2009) and `SPLADEv3` (Lassance et al., 2024).

Besides various models, we investigate the impact of retrieving paragraphs or sentences. We use the paragraphs extracted by GROBID and mark any paragraph as relevant that contains a relevant sentence. Furthermore, we investigate a baseline to improve the decontextualization by prepending the title, which has been shown beneficial in cases where decontextualization is required (Wang et al., 2024).

We evaluate using Mean Reciprocal Rank (MRR) (Craswell, 2009), which considers the first relevant passage in a ranked list. While a typical question in PeerQA often has multiple answer evidence sentences (cf. Figure 2), they frequently belong to the same paragraph or are close to each other. Therefore, pointing a user to the respective paragraph in a real-world application would already be useful as further relevant information usually clusters around the same location. We also measure the quality of the entire ranking by evaluating Recall@10. We chose 10, as most questions have fewer relevant sentences (cf. Figure 6).

[12] A list of topics and their size can be found in §M. We also apply the topic model to the unlabeled questions.

[13] The annotation was performed by two graduate students, reaching a substantial agreement of 0.68 Cohens Kappa.

[14] Class definitions and the distribution can be found in §O.

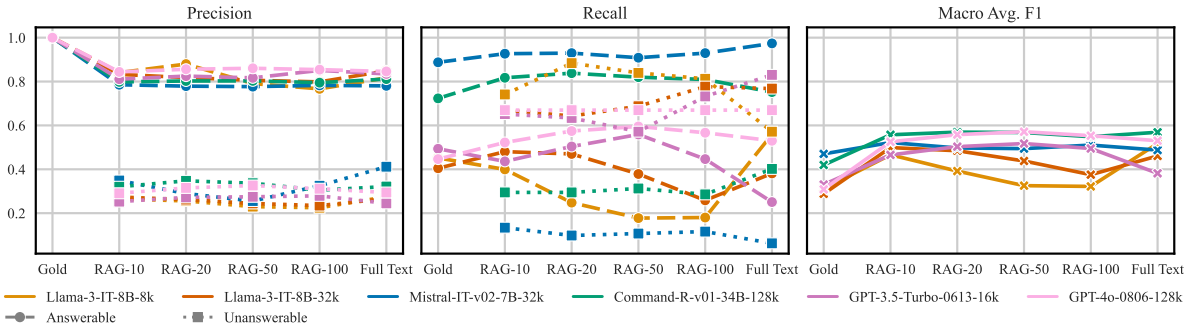


Figure 3: Answerability scores (y-axis) with different contexts (x-axis). In the *Gold* setting, the model is only provided with the annotated, relevant paragraphs (i.e., no unanswerable questions are available in this setting); in *Full Text*, the entire paper is provided in the context (and potentially truncated); otherwise, the top-scoring passages by `SPLADEv3` are provided. The Precision and Recall plots show the Answerable (- -) and Unanswerable (··) classes.

4.2 Answerability and Answer Generation

We set up the answerability task as a binary classification problem: given a question and context, a model predicts whether a question is answerable or not. We label all questions as answerable with annotated answer evidence and all as unanswerable, which the authors flagged as such. The answer generation task is set up as a sequence-to-sequence task, i.e., given the question and the context, the answer needs to be generated. For both tasks, we employ instruction-tuned LLMs. For the answerability task, we prompt the model to either answer the question if sufficient evidence is provided or to generate *No Answer*. However, to obtain generations for all answerable questions, we remove the instruction to generate *No Answer* from the prompt for the answer generation task (see §G and §H for the prompts). We experiment with providing as context the gold passages (ablating retrieval errors), the top-k retrieved paragraphs (where $k \in \{10, 20, 50, 100\}$), and the full text. This is a Retrieval Augmented Generation (RAG) (Lewis et al., 2020) setup, except we retrieve from a single, long document instead of a corpus.We truncate the paragraphs if required by the maximum context size of the models and decode greedily from the models. Specifically, we use `Llama-3-8B-Instruct` (Dubey et al., 2024), which we also extend to a 32k context size with dynamic rope-scaling, `Command-R`[15], `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023), `GPT-3.5-Turbo-0613-16k` and `GPT-4o-0806` (OpenAI et al., 2023). We evaluate the answerability task as a binary classification problem. We evaluate with macro-F1 to counter the imbalance between the number of answerable (383) and unanswerable (112) questions.

Evaluating generative AI for long-form QA is a challenging, ongoing research topic by itself (Krishna et al., 2021; Xu et al., 2023). We chose a diverse set of evaluation metrics, including Rouge-L (Lin, 2004), AlignScore (Zha et al., 2023) and Prometheus-2 (Kim et al., 2024). AlignScore is a model-based metric trained on a broad range of text alignment data, among others, on QA. AlignScore breaks the reference into passages of roughly 350 words and the generation into sentences. The model is trained to measure how much each generated sentence is aligned with the information in the reference passage. In practice, we notice that free-form answers provided by the authors can contain information that is not present in the paper. Therefore, besides using only the free-form answer as ground truth, we also compare the generation to the concatenated answer evidence paragraphs. The Prometheus-2 model is an LLM-as-a-judge model (Zheng et al., 2023) fine-tuned on feedback and judgment data generated by GPT-4 on a large set of custom score rubrics. We provide a scoring rubric that measures the correctness of the generated answer given the reference on a scale from 1-5.[16]

5 Results

5.1 Answer Evidence Retrieval

Table 3 reports the retrieval results. Across models, we find that retrieving the paragraph yields higher scores than the sentence. Appending the title to the paragraph further improves results (except `ColBERTv2`'s MRR), showing that decontextualiz-

[15] https://docs.cohere.com/docs/command-r

[16] See §J for the Prometheus prompt and score rubric.

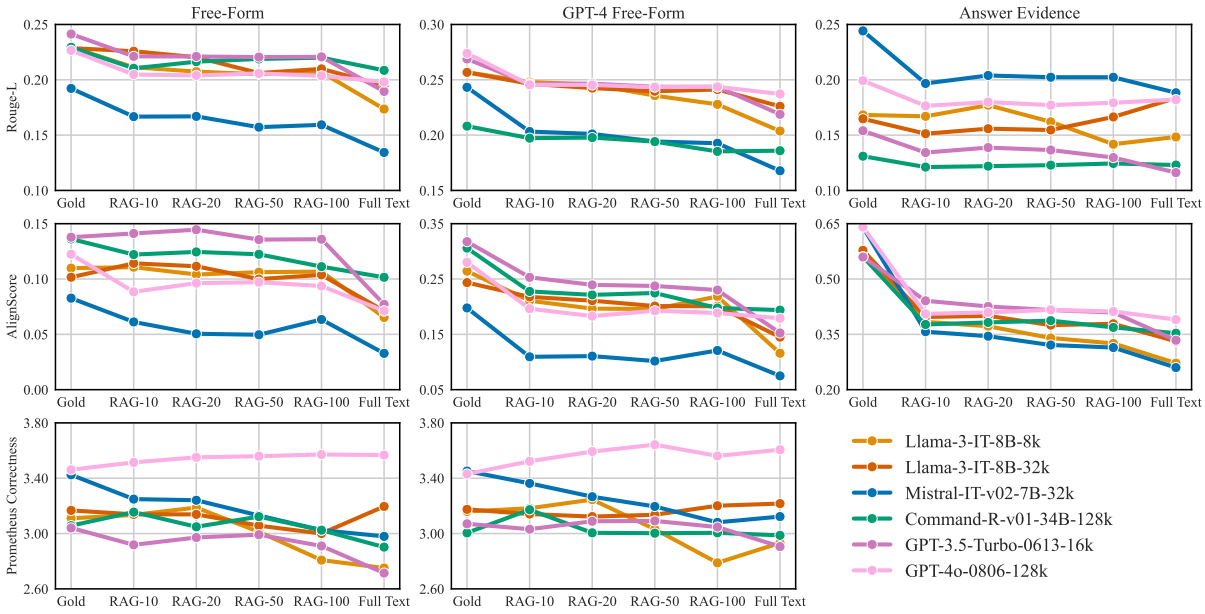


Figure 4: Rouge-L F1, AlignScore and Prometheus Correctness metrics between the annotated free-form answer (1. column), the GPT-4 augmented answer (2. column), the annotated evidence passages (3. column), and the generated answer.

ing the paragraphs from the paper helps. However, we find that MRR and Recall remain the same for most models when prepending the title on a sentence level. Since sentences are short, we conjecture that adding a title influences the overall representation too much, while on a paragraph level, the title only accounts for a fraction of the overall tokens. Overall, we find that `MiniLM-L12-v2`, `Dragon+`, and `SPLADEv3` perform the best. We proceed with `SPLADEv3` for the RAG experiments as it achieves the highest recall.

5.2 Answerability

We report precision, recall, and macro-F1 on the answerability task in Figure 3. We observe similar precision for all model and context settings. Precision for answerable questions is much higher than for unanswerable ones. When looking at recall, we find notable differences between the models. While `Mistral` and `Command-R` obtain relatively high recall on answerable questions and low recall on unanswerable questions, the `Llama` and `GPT` models obtain high recall on unanswerable questions and lower recall on answerable questions. This pattern can be explained: `Mistral` and `Command-R` tend to predict an answer more often, while `Llama` and `GPT` tend to predict the question is unanswerable, showing that all models have a bias towards one of the classes. `Command-R` and `GPT-4o` provide the best trade-off, shown by the highest macro-F1.

5.3 Answer Generation

Figure 4 reports the evaluation metrics comparing the generated answers to either the free-form reference answer, the GPT-4 augmented answer, or the gold paragraphs. Generally, models perform best with the gold answer evidence. Therefore, the annotated evidence provides a strong signal to answer the question. The scores achieved with the gold evidence represent an upper bound. However, higher scores might be possible with more context to better understand the gold answer evidence (or potentially unannotated but useful passages). Upon manual inspection of model errors, we find that lower performance is caused by evaluation failures or information in the free-form answer that is not supported by the evidence (i.e., information that is coming from the author's knowledge that might be general about the field or specific to the paper and did not make it into the camera-ready version). Generally, LLMs perform better in RAG, with fewer but relevant contexts, than in the full-text setting on PeerQA. This shows that despite LLMs' large context sizes, it is more effective to employ a retriever filtering relevant information than leaving this step to the internal workings of the LLM. A notable exception is GPT-4o, which exhibits stable performance with increasing context sizes and increasing performance on answer correctness. GPT-4o is also the most recent and

Error Class	
Correct / Evaluation Error	43.75%
Partially Correct	12.50%
Reasoning Error	10.00%
Implicit Evidence Only	7.50%
Insufficient Context	11.25%
Insufficient Evidence	12.50%
Insufficient Free-Form Answer	3.75%

Table 4: Error analysis of `GPT-3.5`'s generations with gold evidence. §R provides definitions and examples for error classes.

powerful model in our evaluation, demonstrating the improved abilities of state-of-the-art models on long-context tasks. We further analyze the answer generation performance of the RAG setting by measuring the correlation between the retriever recall and the generation metric. We find mostly positive correlations between the retrieval and generation performance across models. While the correlation is not very strong (up to $r = 0.42$), it confirms that with increased retrieval performance, the generation improves (Salemi and Zamani, 2024).[17]

Error Analysis. We analyze the lowest performing 80 generations[18] of `GPT-3.5` to better understand the errors and report them in Table 4. We find many low-scoring generations are correct despite at least one of the evaluation metrics providing a low score, for example, when the generation is more verbose or expresses the correct answer differently (*Evaluation Error*). However, we find the metric with the highest score for these generations to be above the 50th percentile in 91% of the cases. This shows that using different metrics against different ground truths is plausible and catches the alleged failures. Further, we observe the model is only *partially correct* when the free-form answer contains important additional details. In other cases, the model fails to reason correctly over the evidence, e.g., it arrives at an opposite conclusion than the correct answer. Similarly, when the evidence is only implicit or requires expert domain knowledge, the model fails. Lastly, there are also a few errors in the data. In 11.25% of cases, the gold evidence is not self-sufficient, i.e., more context from the paper would be required, e.g., to understand previously introduced concepts. These errors can likely be recovered through additional retrieval. Other times the answer by the authors is not entailed by the evidence (*Insufficient Evidence*) or the free-form answer only reports the element in the article, but not an actual answer (*Insufficient Free-Form Answer*).

[17] §S.1 reports correlations across all metrics and contexts.

[18] Specifically, we sort by the minimum performance of all metrics, considering all questions that have both evidence and free-form annotations and use the *gold* evidence as context.

6 Conclusion

We introduced the PeerQA dataset to advance and study question answering on scientific documents. We sourced PeerQA's questions from peer reviews and obtained answer annotations from the paper authors. Our dataset supports three crucial tasks for developing QA systems: evidence retrieval, answerability, and answer generation. We analyzed the collected data and established baseline systems for all three tasks. For evidence retrieval, we find that decontextualization is key to improving performance. On the answerability task, we find that models tend to either over- or under-answer, showing a bias for one of the classes. Further, although models can fit the entire paper into context in the answer generation task, providing the model with the top passages from a retriever outperforms the full-text setting. We also show that with increased retrieval performance, the answer generation improves. Finally, our error analysis highlights the need for better evaluation metrics and model reasoning abilities.

7 Limitations

Dataset Size. General domain QA datasets usually comprise up to three magnitudes more data than PeerQA (e.g., NQ has 323k samples). However, collecting high-quality data in the scientific domain is challenging due to the requirement for expert annotators. Since science has many domains, it is impractical to collect training data for each of them. Instead, models need to generalize in an unsupervised manner, at most leveraging few-shot examples. Therefore, we introduce PeerQA as an evaluation resource to test the generalizability of models. The size is in line with other recent datasets such as HumanEval (Chen et al., 2021a) (164 examples), TruthfulQA (Lin et al., 2022) (817), and GPQA (Rein et al., 2024) (448). In addition, we release the unlabeled data, comprising 12k questions from 2.6k papers, that can be used for more annotations, unsupervised learning, and further study of reviews. Small evaluation datasets also have the advantage of reduced iteration time over experimental settings, lesser use of compute resources, and a smaller environmental impact.

Science Domains. While PeerQA covers more scientific domains compared to prior work, there is a limited amount of data beyond the ML and NLP domains. A major challenge in data collection is the availability of public peer reviews with openly licensed scientific articles (Dycke et al., 2022). We call on the scientific community to further transform reviewing practices to an open format.

English-Only. PeerQA is limited to English since it is dominant in scientific writing. Nevertheless, publications in other languages exist, and our data collection framework can be applied to any language. The evaluated retrievers are English-only models (except `BM25`, which is language-agnostic). Some retrieval models have multi-lingual counterparts (e.g., `mContriever`); however, due to a lack of multi-language data, their performance remains unclear. Some of the evaluated generative models are also multilingual; the performance in other languages is likely to be different than in English.

Free-Form Annotations. While authors possess the ultimate expertise in their papers, they usually have knowledge beyond the information in their publications. Some free-form answers contain information not included in the answer evidence. For this reason, we also compare the generated answer with the annotated answer evidence, measuring if the model can produce answers entailed by the information in the paper.

Long-Form QA Evaluation. Evaluating free-form answers is challenging and an ongoing area of research. To evaluate different aspects, we use three metrics against two ground truths. Ideally, we would have multiple free-form answer references; however, even collecting a single response has proven to be challenging. In the hope of better metrics, we also publish the generated answers of our baselines to facilitate adaptation to future, improved methods.

Methods. Many LLMs and methods (Zhao et al., 2023) exist that could be applied to the tasks in PeerQA. Therefore, more sophisticated and specialized methods might exceed the reported performances. However, we focus on introducing the dataset and establishing baseline systems with widely used retrieval and generative models.

8 Ethical Considerations

All annotators in PeerQA are authors of accepted articles at conferences or in journals. We do not collect any of their personal information or who has provided the answers. By the nature of our data collection protocol, we only contact authors who have provided their email publicly along with their publication and contact each author individually. Authors have participated voluntarily in the data collection, and we try to keep their workload low by only asking few questions (on average 2.8). Furthermore, the authors have largely already answered questions during peer review (see §C), making them familiar with the questions and answers, further reducing their workload.

One objective of PeerQA is to advance the study of peer review, including developing methods and tools to facilitate the authoring and reviewing of scientific articles. Particularly, LLMs have the potential to support authors and reviewers in their work (Kuznetsov et al., 2024). However, these models also have biases and weaknesses. For example, in our question answerability task, we clearly observe that some models are biased towards one class, i.e., predicting the question as answerable or unanswerable (see §5.2). Therefore, these methods can only be used as assistants that support humans. PeerQA sheds light on these issues, raising awareness of potential weaknesses in these models and their careful application in science.

Acknowledgements

This work has been funded by the German Research Foundation (DFG) as part of the QASciInf project (grant GU 798/18-3). Further, we gratefully acknowledge the support of Microsoft with a grant for access to the OpenAI GPT models via the Azure cloud (Accelerate Foundation Model Academic Research).

We thank the anonymous reviewers for their helpful suggestions for improving this paper and Sukannya Purkayastha, Max Eichler, and Haritz Puerto for their insightful feedback throughout the paper-writing process. Our gratitude also goes to Maike Nowatzki for reviewing the questions in the Geoscience domain, to Richard Eckart de Castilho for his support with our annotation platform, and to Sebastian Alles for his assistance in establishing the compute and annotation infrastructure. Finally, we are grateful to all authors who have voluntarily participated in creating this dataset.

References

Sören Auer, Dante Augusto Couto Barone, Cassiano Bartz, E. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry I. Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *CoRR*, abs/2212.08037.

Lutz Bornmann, Rüdiger Mutz, and Robin Haunschild. 2020. Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *CoRR*, abs/2012.07675.

Jeffrey Brainard. 2020. Scientists are drowning in covid-19 papers. can new tools keep them afloat? *Science*, 13(10):1126.

Jan Buchmann, Xiao Liu, and Iryna Gurevych. 2024. Attribute or abstain: Large language models as long document assistants. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8113–8140, Miami, Florida, USA. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021a. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021b. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. Association for Computational Linguistics.

Nick Craswell. 2009. Mean reciprocal rank. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 1703–1703. Springer US, Boston, MA.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. MARG: multi-agent review generation for scientific papers. *CoRR*, abs/2401.04259.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2024. ARIES: A corpus of scientific paper edits made in response to peer reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6985–7001, Bangkok, Thailand. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon,

Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Yes-yes-yes: Proactive data collection for ACL rolling review and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 300–318, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Stefano Fiorucci. 2024. Rag Evaluation with Prometheus 2. https://haystack.deepset.ai/blog/rag-evaluation-with-prometheus-2. Accessed: 27. September, 2024.

Michael Fire and Carlos Guestrin. 2019. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *GigaScience*, 8(6):giz053.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. 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 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, 17(1):1–29.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. 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*.

Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, page 243–246, New York, NY, USA. Association for Computing Machinery.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Devang Kulshreshtha, Robert Belfer, Iulian Vlad Serban, and Siva Reddy. 2021. Back-training excels self-training at unsupervised domain adaptation of question generation and passage retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7064–7078, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, and Iryna Gurevych. 2024. What can natural language processing do for peer review? *CoRR*, abs/2405.06563.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. Splade-v3: New baselines for SPLADE. *CoRR*, abs/2403.06789.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-In Lee, and Moontae Lee. 2023. QASA: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19036–19052. PMLR.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid,

Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu B. Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2024. The semantic reader project. *Commun. ACM*, 67(10):50–61.

Patrice Lopez. 2008–2024. Grobid. `https://github.com/kermitt2/grobid`.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Smaranda Muresan, Vivian Chen, Kennington Casey, Vandyke David, Dethlefs Nina, Inoue Koji, Ekstedt Erik, and Ultes Stefan, editors. 2023. *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics, Toronto, Canada.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele,

Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14479–14495, Singapore. Association for Computational Linguistics.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Re3: A holistic framework and dataset for modeling collaborative document revision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4635–4655, Bangkok, Thailand. Association for Computational Linguistics.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2395–2400, New York, NY, USA. Association for Computing Machinery.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923, Miami, Florida, USA. Association for Computational Linguistics.

Haitian Sun, William Cohen, and Ruslan Salakhutdinov. 2022. ConditionalQA: A complex reading comprehension dataset with conditional answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3627–3637, Dublin, Ireland. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 873–876, New York, NY, USA. Association for Computing Machinery.

Yuwei Wan, Aswathy Ajith, Yixuan Liu, Ke Lu, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian T. Foster. 2024. Sciqag: A framework for auto-generated scientific question answering dataset with fine-grained evaluation. *CoRR*, abs/2405.09939.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2024. DAPR: A benchmark on document-aware passage retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4313–4330, Bangkok, Thailand. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507, Portland, Oregon, USA. Association for Computational Linguistics.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

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.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM International Conference*

on Information and Knowledge Management, CIKM '18, page 497–506, New York, NY, USA. Association for Computing Machinery.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Dennis Zyska, Nils Dycke, Jan Buchmann, Ilia Kuznetsov, and Iryna Gurevych. 2023. CARE: Collaborative AI-assisted reading environment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 291–303, Toronto, Canada. Association for Computational Linguistics.

A PDF extraction of Answer Evidence

Text extraction from PDF is not a perfect process. Unfortunately, this means that some annotated answer evidence (and therefore also answerable questions) must be discarded in our experiments since their evidence has not been extracted correctly. Table 5 shows the number of annotated answer evidence (Evidence), as well as the number of questions whose evidence has been extracted correctly (Evidence Mapped). We nevertheless make the complete dataset public so future research with better PDF processing tools can leverage more annotations.

Venue	Questions	Evidence	Ev. Mapped
ICLR 23	153	103	93
ICLR 22	137	108	99
NeurIPS 22	79	56	55
ARR 22	87	60	55
COLING 20	31	25	23
ACL 17	20	16	16
CoNLL 16	12	7	4
ESD 23	17	10	10
ESurf 23	16	16	16
F1000 22	27	14	12
Total	**579**	**414**	**383**

Table 5: Number of questions with answer evidence that could be mapped to the PDF extracted text.

B Pre- & Post-Processing Prompts

B.1 Question Clean-Up & Decontextualization

Given the extracted question and previous sentences (`context`) from the peer review, we use the following prompt to decontextualize the question:

```
This is part of a scientific peer review where the reviewer raises a question regarding the paper.
"""
{context} {question}
"""
Write the last question such that it can be comprehended independently without the context of the review. Resolve any references to the review. Respond with a single question.
```

B.2 Question Decomposition

In case the constituency parser detects a conjunction, we use the following prompt to decompose the question:

```
This is a sentence from a peer review containing two questions.
"""
{question}
"""
Write the questions such that each can be comprehended independently without the context of the other question. Resolve any references in the second question. Therefore, the fundamental question information needs to be duplicated in each question.
```

B.3 Answer Free-Form Augmentation with Evidence

To ensure a similar quality and verbosity of answers, we augment the free-form answers provided by the authors using the prompt below in case the question has annotated evidence. If it does not have annotated evidence, we use the prompt in §B.4.

```
You are a helpful scientific research assistant. Your task is to write clean answers, given noisy answers from a scientific question answering dataset. The question has been asked during a peer review of a scientific article. Given the question, background information extracted from the paper, and a noisy answer, your task is to write a clean answer. Write a concise answer that directly answers the question. Make sure all information in your answer is covered by the background. Incorporate additional information from the original answer. Write the answer neutrally, i.e., as a third person (and not the author) answering the question. For example, use "The authors" instead of "We".
Question: {question}
Background: {evidence}
Original Answer: {answer}
Rephrased Answer:
```

B.4 Answer Free-Form Augmentation without Evidence

```
You are a helpful scientific research assistant. Your task is to write clean answers, given noisy answers from a scientific question answering dataset.
```

```
The question has been asked during a peer review of a scientific article. Given the question and a noisy answer, your task is to write a clean answer. Write a concise answer that directly answers the question. Incorporate the information from the original answer. Write the answer neutrally, i.e., as a third person (and not the author) answering the question. For example, use "The authors" instead of "We".
Question: {question}
Original Answer: {answer}
Rephrased Answer:
```

C Question Grounding

Figure 5 visualizes the similarity between the processed question and original review sentences. We use `all-MiniLM-L6-v2`[19] to compute the similarity. As detailed in §3.1, we extract questions from the peer review and contextualize them with the preceding three sentences from the review. To understand whether our preprocessing has altered the original question or not, we compute the maximum similarity between the final processed question and the four sentences of the review (i.e., the question and the three proceeding questions). We find that 90% of questions have a similarity of at least 0.60, and 50% are more than 0.82 similar to the final processed question. This shows the quality of our cleaning, decontextualization, and decomposition steps: Questions are generally highly similar and, therefore, grounded in the original peer review.

D Unlabeled Data

Besides the 579 questions with answer annotations, we additionally release all preprocessed and filtered 12k questions from 2.6k papers that have not been answered. Table 6 shows the breakdown per venue.

E Answer Evidence Statistics

Figure 6 reports the number of answer evidence depending on the retrieval unit. Note that this only includes answer evidence that we could map into the text extracted from the PDF. Non-consecutive chunks are essentially the number of different locations in the paper with answer evidence.

[19] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

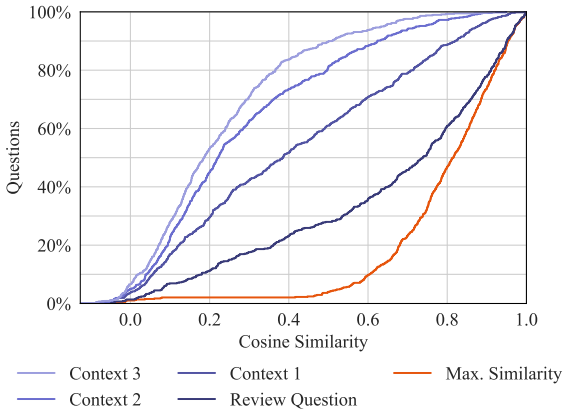


Figure 5: Empirical cumulative distribution function of the cosine similarity between the processed question and the sentences in the review. *Context n* refers to the n-th preceding sentence before the raw, unprocessed *Review Question*. *Max. Similarity* takes the max operation over these four similarity scores, i.e., reports the similarity the processed question is most similar to.

Venue	Questions	Papers
ICLR 23	5199	1188
ICLR 22	3987	824
NeurIPS 22	1186	110
ARR 22	470	188
COLING 20	70	33
ACL 17	147	54
CoNLL 16	3	3
ESurf 23	312	51
ESD 23	246	48
F1000 22	347	124
Total	11967	2623

Table 6: Number of unlabeled questions and papers per venue.

While the answer evidence for most questions comes from a single place, 30% of questions have more than one location in the paper that addresses the question. While requiring to retrieve from multiple sources is related to multi-hop question answering (Welbl et al., 2018; Yang et al., 2018), our setup is slightly different. We have also investigated the performance of questions with single vs multiple answer evidence chunks and have not found consistent differences. The information in the different chunks is not necessarily complementary, but it can also be that similar information is contained in each chunk, or a single chunk is sufficient to answer the question.

F RAG Recall@k

Figure 7 shows the recall at various cutoffs k for `SPLADEv3`, the best-performing model answer ev-

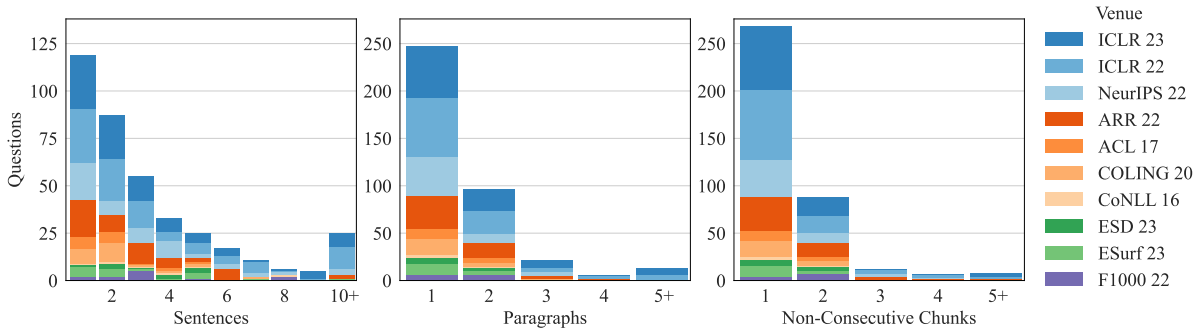


Figure 6: Number of evidence sentences (left), paragraphs (middle), and non-consecutive chunks (right) per question with annotated answer evidence.

idence retrieval task. This model is used as a retrieval model for the retrieval augmented answer generation experiments.

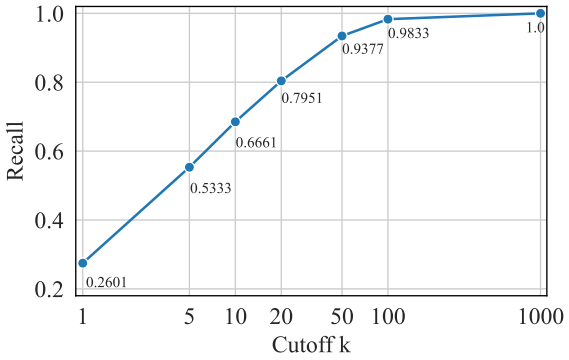


Figure 7: Recall@k of SPLADEv3 on the answer evidence retrieval task in the paragraph setting. The paragraphs retrieved by SPLADEv3 are used in the RAG experiments.

G Answerability Prompts

We use the following prompts to determine whether a question is answerable or not in the setting where we provide the full text (§G.1), the gold or retrieved paragraphs (§G.2).

G.1 Full-Text

```
Read the following paper and answer the question. If the paper does not answer the question, answer with "No Answer".
Question: {question}
Paper: {paper}
Answer:
```

G.2 RAG

```
Read the following paragraphs of a paper and answer the question. If the paragraphs do not provide any information to answer the question, answer with "No Answer".
Question: {question}
Paragraphs: {paragraphs}
Answer:
```

H Answer Generation Prompts

We use the following prompts to generate answers in the full-text (§H.1) and RAG (§H.2) setting.

H.1 Full-Text

```
Read the following paper and answer the question.
Question: {question}
Paper: {paper}
Answer:
```

H.2 RAG

```
Read the following paragraphs of a paper and answer the question.
Question: {question}
Paragraphs: {paragraphs}
Answer:
```

I Model Sizes and Computational Resources

Answer Retrieval The number of parameters for each retrieval model is reported in Table 7. The retrieval experiments have been conducted on a Titan RTX 24GB.

Answerability & Answer Generation Sizes for the models used in the answerability and answer generation task are reported with the model names. The number of parameters for the proprietary GPT-3.5 and GPT-4o models are unknown, and we use it via the Azure API. We deploy the other models on a single A100 80GB GPU, except

Model	Parameters (M)
MiniLM-L12-v2	33
Contriever	110
Dragon+	110
GTR-XL	1240
ColBERTv2	110
BM25	–
SPLADEv3	110

Table 7: Number of parameters in Millions of the evaluated retrieval models.

`Command-R` for which we require 2 A100 GPUs to fit also the longest paper fully into memory. All generation experiments use greedy decoding and use the `vllm` library (Kwon et al., 2023) in version 0.4.2.

J Evaluation Metric Details

Answer Evidence Retrieval To evaluate the answer evidence retrieval task, we use the mean reciprocal rank and recall implemented by the `pytrec_eval` (Van Gysel and de Rijke, 2018) package in version 0.5.

Un/Answerability To compute the precision, recall, accuracy and F1 scores of the question answerability task, we use the classification report provided by `scikit-learn` (Pedregosa et al., 2011) version 1.4.0.

Free-Form Answer Generation The generated answers are evaluated with Rouge (Lin, 2004), AlignScore (Zha et al., 2023) and Prometheus (Kim et al., 2024). For Rouge, we use the longest common subsequence (Rouge-L) between the generated answer and the reference answer. We use the `rouge-score` package in version 0.1.2 via Hugging Face's `datasets` package (Lhoest et al., 2021). We also stem the generated and reference answer before computing the metric with the Porter Stemmer. All reported Rouge-L scores are F1 metrics. For AlignScore, we use the fine-tuned checkpoint based on RoBERTa-large (Liu et al., 2019) and use the `nli_sp` mode, which splits the generation into sentences and uses a 3-way classification head to obtain scores. We use the original implementation in version 0.1.3.[20] For Prometheus, we use the `prometheus-eval` package with version 0.1.20[21] and the 7B-v2.0 model[22] and instruct the model to evaluate the correctness with respect to the reference answer. Following Kim et al. (2024) and the score rubric proposed by Fiorucci (2024) we use the following prompt:

```
###Task Description:
An instruction (might include an Input inside it), a response to evaluate, a reference answer that gets a score of 5, and a score rubric representing a evaluation criteria are given.
1. Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
3. The output format should look as follows: "(write a feedback for criteria) [RESULT] (an integer number between 1 and 5)"
4. Please do not generate any other opening, closing, and explanations.
###The instruction to evaluate:
Your task is to evaluate the generated answer against the reference answer for the question: {question}
###Response to evaluate:
{generation}
###Reference Answer (Score 5):
{reference answer}
###Score Rubrics:
Correctness
Score 1: The answer is not relevant to the question and does not align with the reference answer.
Score 2: The answer is relevant to the question but deviates significantly from the reference answer.
Score 3: The answer is relevant to the question and generally aligns with the reference answer but has errors or omissions.
Score 4: The answer is relevant to the question and closely matches the reference answer but is less concise or clear.
```

[20] https://github.com/yuh-zha/AlignScore/commit/a0936d5afee642a46b22f6c02a163478447aa493

[21] https://github.com/prometheus-eval/prometheus-eval/releases/tag/v0.1.20

[22] https://huggingface.co/prometheus-eval/prometheus-7b-v2.0

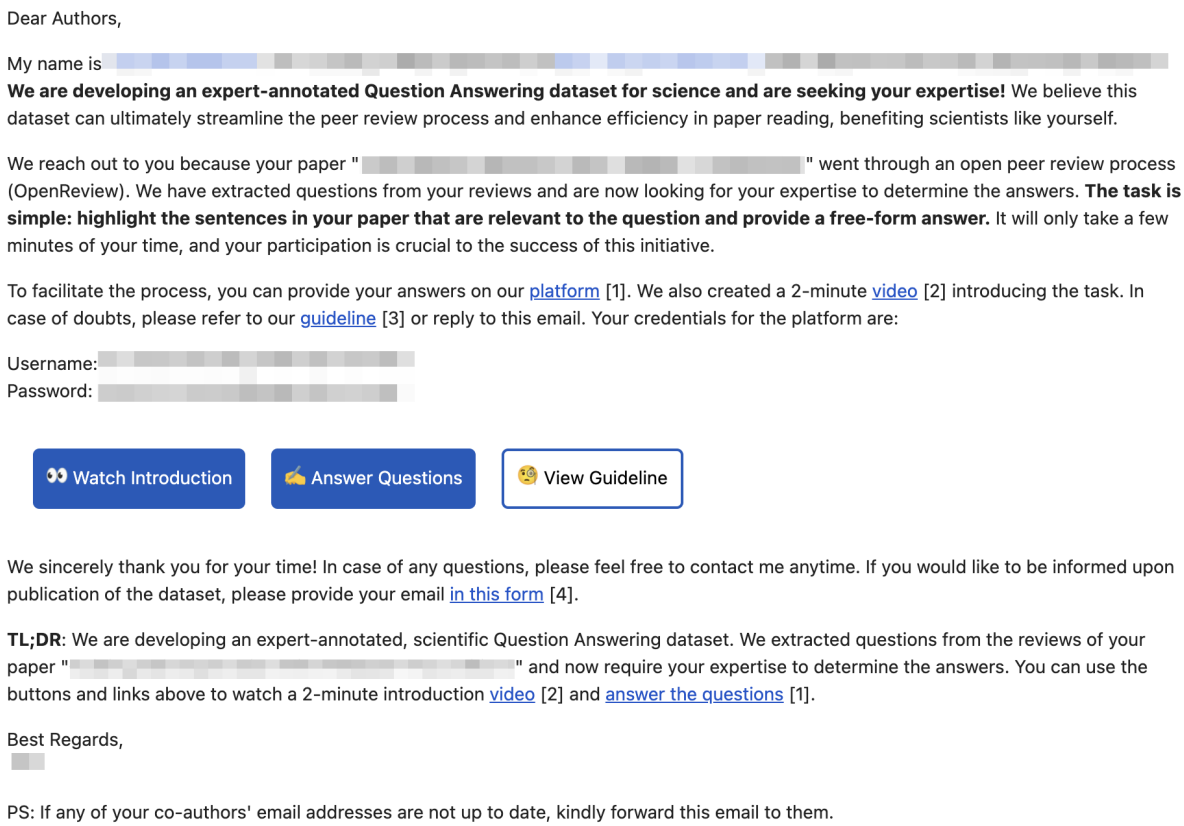
```
Score 5: The answer is highly relevant,
fully accurate, and matches the reference
answer in both content and clarity.
###Feedback:
```

Human Evaluation of AlignScore Zha et al. (2023) has evaluated correlation with human judgments extensively, particularly on factual consistency datasets. To show the reliability of AlignScore on our data, we manually label 100 randomly generated answers. Specifically, we compare the free-form and generated answer on a 1-5 Likert scale, evaluating whether the generation matches the free-form answer. This yields a significant ($p<0.01$) Spearman correlation of 0.449, indicating a moderate alignment between Human evaluation and the AlignScore metric.

K Annotation Instructions and Interface

K.1 Contact Email

Figure 8 shows an instance of an email that invited the authors to participate in the data collection. Email addresses have been extracted from papers or, in the case of EGU and F1000, addresses to corresponding authors are provided online. To prevent spamming authors, we have ensured that no author received more than 3 emails (e.g., when they were listed as authors on multiple papers of a venue). Email addresses were only used to contact authors and are not part of the dataset. We also do not publicize the code to extract email addresses from papers.



Dear Authors,

My name is

We are developing an expert-annotated Question Answering dataset for science and are seeking your expertise! We believe this dataset can ultimately streamline the peer review process and enhance efficiency in paper reading, benefiting scientists like yourself.

We reach out to you because your paper " " went through an open peer review process (OpenReview). We have extracted questions from your reviews and are now looking for your expertise to determine the answers. **The task is simple: highlight the sentences in your paper that are relevant to the question and provide a free-form answer.** It will only take a few minutes of your time, and your participation is crucial to the success of this initiative.

To facilitate the process, you can provide your answers on our platform [1]. We also created a 2-minute video [2] introducing the task. In case of doubts, please refer to our guideline [3] or reply to this email. Your credentials for the platform are:

Username:
Password:

👀 Watch Introduction | ✍️ Answer Questions | 🧐 View Guideline

We sincerely thank you for your time! In case of any questions, please feel free to contact me anytime. If you would like to be informed upon publication of the dataset, please provide your email in this form [4].

TL;DR: We are developing an expert-annotated, scientific Question Answering dataset. We extracted questions from the reviews of your paper " " and now require your expertise to determine the answers. You can use the buttons and links above to watch a 2-minute introduction video [2] and answer the questions [1].

Best Regards,

PS: If any of your co-authors' email addresses are not up to date, kindly forward this email to them.

Figure 8: Exemplary contact email that has been sent to authors requesting their participation in answering the questions.

K.2 Annotation Interface

The annotation interface for providing answers is shown in Figure 9. The camera-ready PDF of the publication is shown on the right-hand side, while answer annotations can be provided on the left side. In *1.2 Question Feedback*, authors can leave free-form feedback about the question, e.g., if it should be removed or modified. By clicking on the *Add* button in *2.1 Answer Evidence*, text spans in the PDF can be highlighted. One highlight can span over several sentences or even pages. Multiple spans can be added by clicking the *Add* button again. In *2.2. Answer Free Text*, the free-form answer to the question can be given. Finally, in *3.1.*, the authors can also mark the question as unanswerable or provide further feedback to the question. If none of the categories apply, feedback on why the question is unanswerable can also be provided in *No Answer Reason Free Text*.

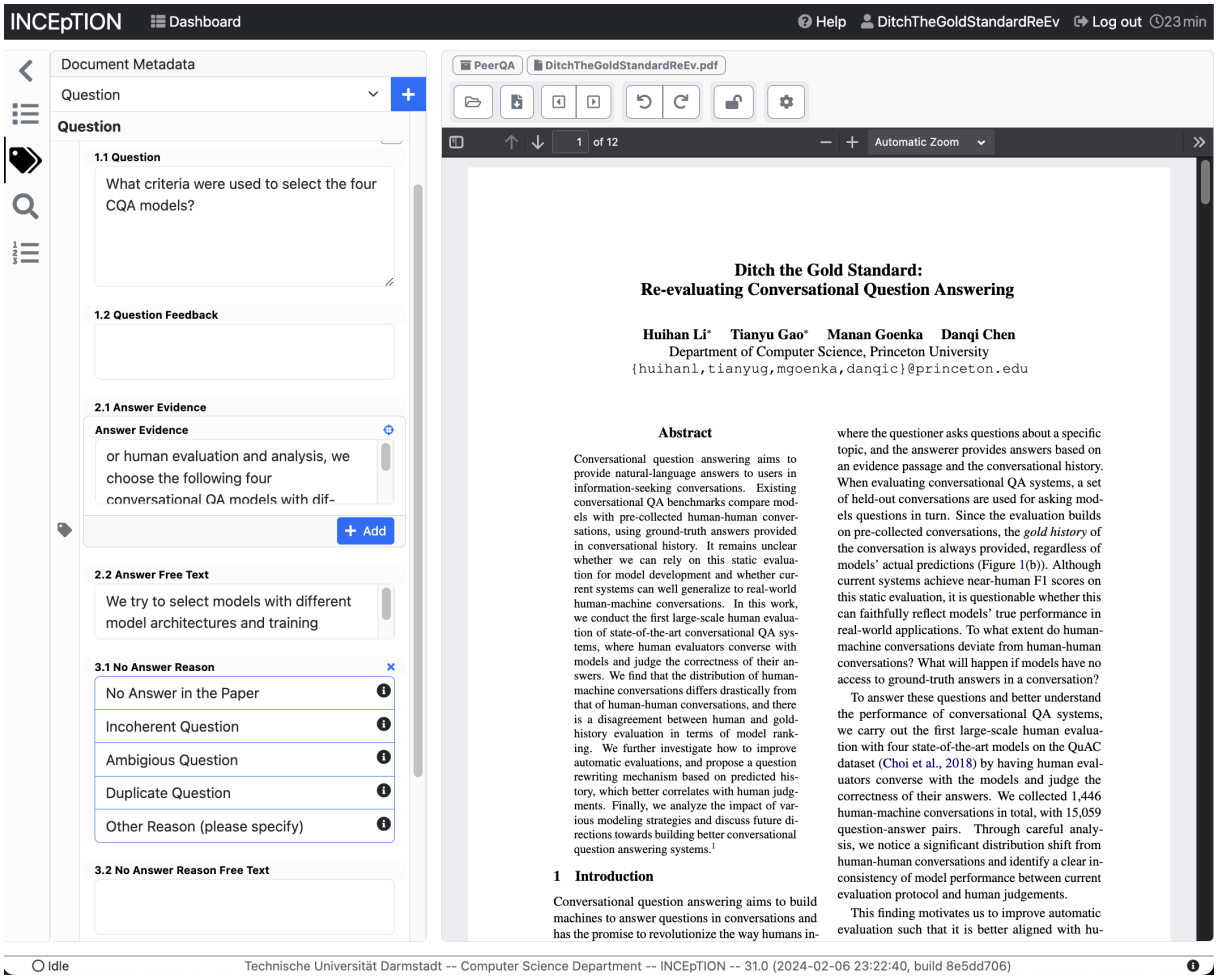


Figure 9: Screenshot of the annotation interface. The annotation consists of four parts. First, the annotator can provide feedback to the question, e.g., to correct its meaning or provide their interpretation. Second, answer evidence is annotated by highlighting sentences in the PDF. Third, a free-form answer can be provided, directly answering the question. Lastly, if a question is unanswerable or is of low quality, the interface provides an option to flag the question.

L Question Sunburst

We visualize the starting 4-grams of all questions in Figure 10. All words have been lowercased and lemmatized, and rare n-grams have been discarded.

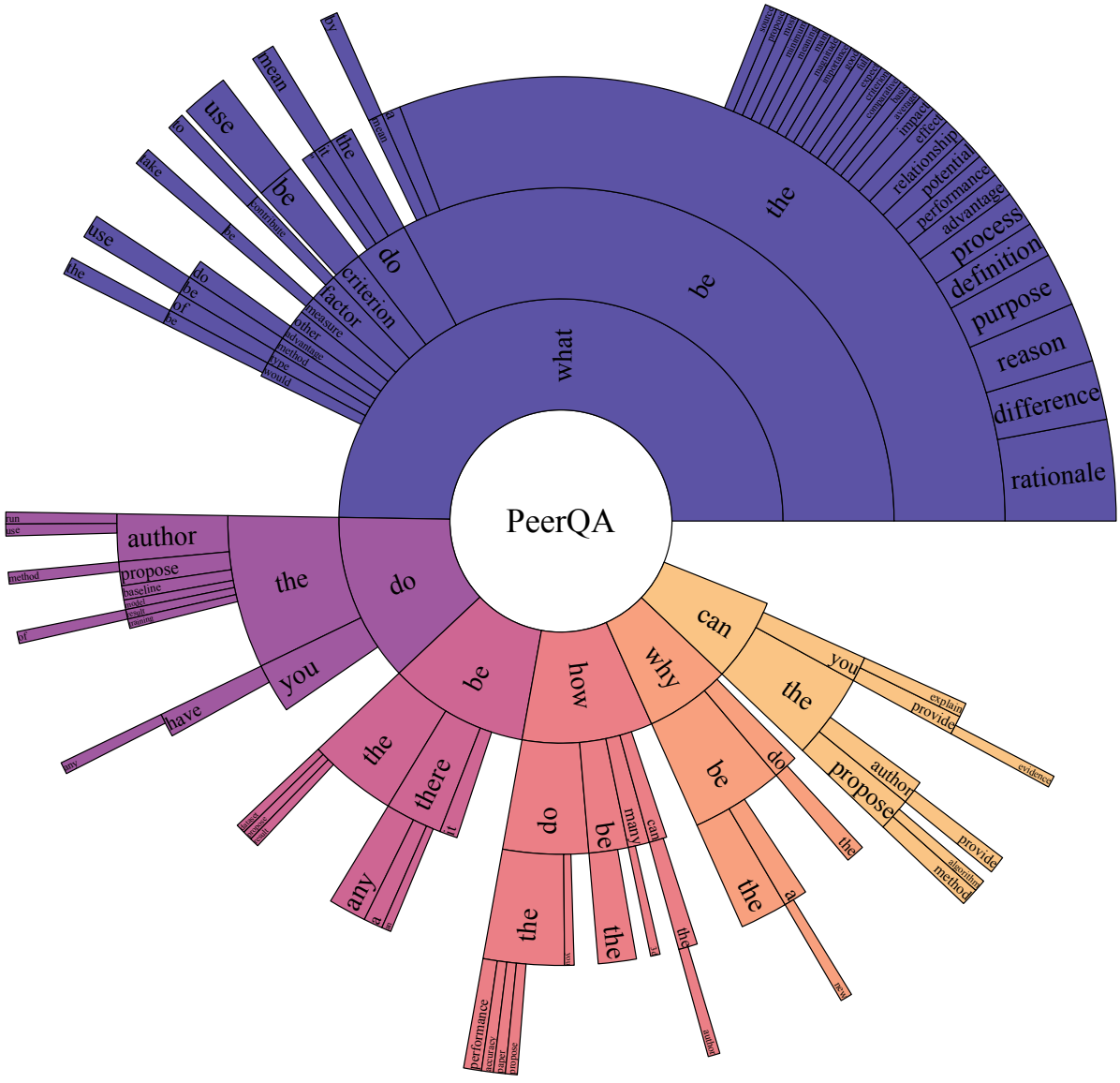
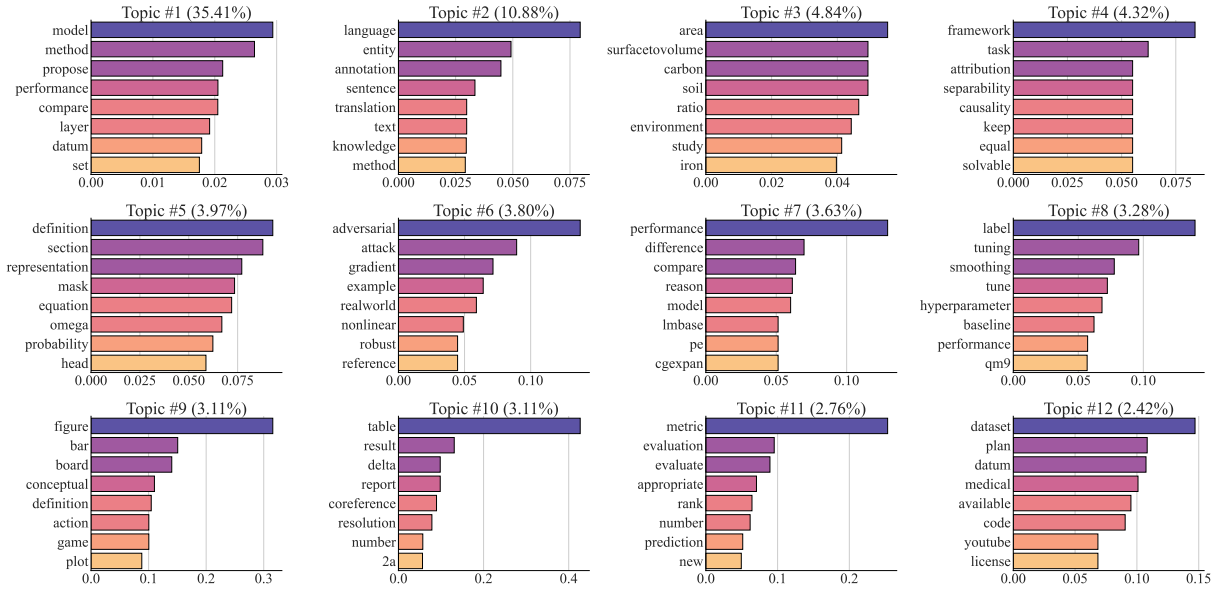


Figure 10: Sunburst diagram of the 4-grams in the PeerQA questions.

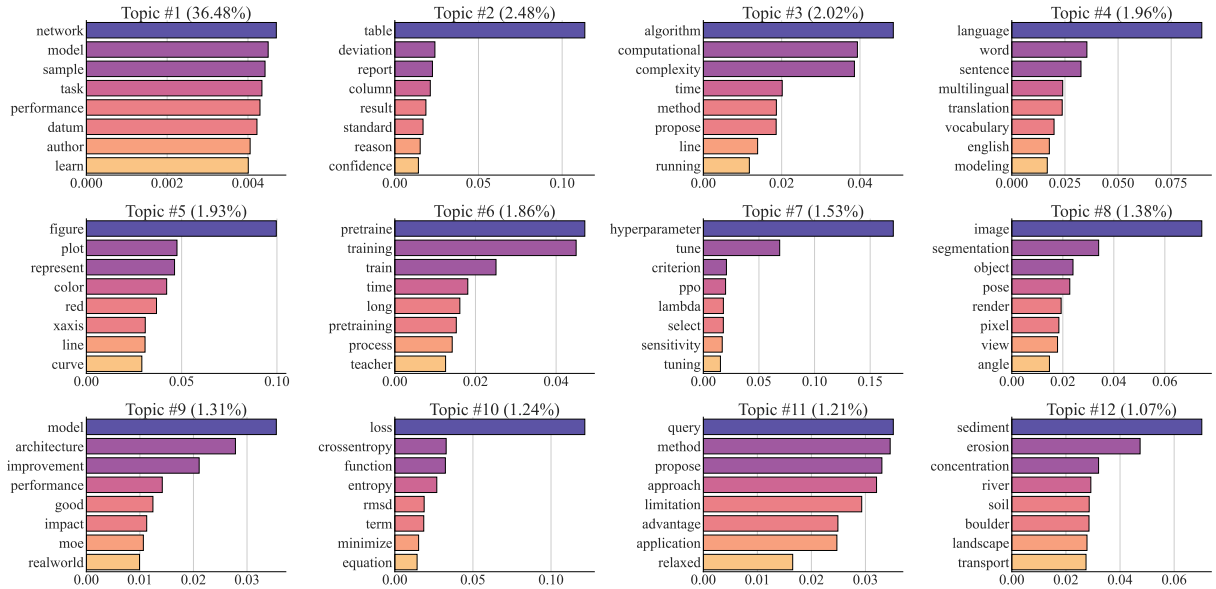
M Question Topics

Figure 11 reports the result of applying BERTopic (Grootendorst, 2022) on the labeled (11a) and unlabeled (11b) questions. Table 8 additionally shows representative questions for the topics. We use the standard `all-MiniLM-L6-v2` sentence transformer model to compute embeddings. After embedding, stopwords have been removed, and the words have been lemmatized using spacy (Honnibal et al., 2020) to improve the keyword extraction.

In §3.2, we analyzed the topics of the labeled questions cluster. We found them to be focused on the scientific community and its subtopics or related to elements in the paper. We observe similar clusters in the unlabeled data (e.g., topic #4 for NLP, topic #12 for Geoscience, #6 focusing on (pre-)training, #5 on figures and plots).



(a) Topics of Labeled Questions



(b) Topics of Unlabeled Questions

Figure 11: Top 8 keywords for the top 12 question topics. The first topic contains all questions that could not be assigned during clustering. The bar shows the keyword's c-TF-IDF score. The top figure shows the topics for the labeled questions and the bottom for the unlabelled.

#	Size	Representative Questions
1	35.4%	Why does the baseline have significantly better performance on ACE 2004 and ACE 2005 compared to Yu et al. (2020), but similar performance on OntoNotes 5 and CoNLL 2003? Does the proposed alternate way to use linear by computing the mean absolute value of the weights associated with it differ from the original linear model proposed by Dalvi et al. (2019)? Is the design of the proposed method arbitrary for all layers of a given VIT model, or are some layers fixed?
2	10.9%	How does your method differ from existing methods for visual and language understanding with multilinguality, such as VQA, captioning, and retrieval? What other downstream tasks, such as natural language inference, question answering, and semantic role labeling, have been tested using an encoder that has been transferred from language 1 to language 2 without any parameter updates? How can the authors ensure that the natural language sentences produced from the "ground truth" activity graphs accurately describe the scene?
3	4.8%	Is the authors' conclusion about the accuracy of the CMIP6 climate models in simulating the processes based on the agreement between the observed data and the models' predictions in terms of the residual variability? Do the authors assume that iron is sourced from the platform when considering the feasibility of coastal seaweeds, which have a very low surface-to-volume ratio, competing for iron against the typically small and specialized open ocean phytoplankton that have a high surface-to-volume ratio? Can we assume that coastal seaweeds, which have a very low surface-to-volume ratio, would be competitive in iron uptake against the mostly small and specialized open ocean phytoplankton that have a high surface-to-volume ratio, especially in iron-limited areas?
4	4.3%	Can the authors provide a justification for why only four datasets were used to evaluate the visual search models, rather than a more diverse collection of datasets? Why should associations that are solvable by AI be kept in the framework, when the purpose of the framework is to collect associations that are difficult for models but solvable by humans? Does the proposed approach address issues related to assigning different attributions to features that have the same effect on the model or assigning positive attributions to features with no effect?
5	4.0%	How does the paper incorporate section titles into the BOS representation? What is the purpose of multiplying the scalar sc(omega, q) by the inner product of omega and q in equation 5? What is the difference between the `\odot` and `\cdot` symbols in the equation for computing the overall source mask from the k masks?
6	3.8%	Is it possible to consistently find perturbations to empirically robust adversarial examples that result in a correctly classified image? How does the paper define the concept of an "adversarial L2 ball" when it appears to suggest that every sample should have the same classification as `\tilde {x}`, contrary to the expectation that each sample within the ball should have a different classification compared to x? Could the authors provide further justification for their claim that the gradient-based attack is responsible for the shift between test and training data observed in the adversarial attack?
7	3.6%	What is the performance of larger GLM models compared to state-of-the-art results, given that hardware resources do not appear to be a constraint? What is the expected relationship between the performance of the algorithms and the number of updates per sample, memory size, and batch size? What could explain the difference in performance between the DICTA test set and the new test set, particularly the difference between the cha and wor scores?
8	3.3%	What protocol did you use to decide when to stop training and to select hyperparameters for each dataset when no labeled target data is available? Does label smoothing always improve performance, or are there cases where it can degrade performance? Does label smoothing always improve the performance of the hyperparameter-fine tuning procedure?
9	3.1%	What are the vertical uncertainty bars in Figure 13? What would be the correct classification for the image in Figure 1 where the space bar is hidden? What is the reason for the sudden change in the green and blue curves in Figure 2 at epoch 90?
10	3.1%	What is the impact of adjusting δ on the results of Table 1? Is the coreference resolution pipeline depicted in Table 1 universally accepted in the field of coreference resolution? What is the difference between the results in Table 1 and Table 2(a)?
11	2.8%	Is there an optimal number of MSD points to use in order to minimize the error on the estimated parameters, and is there an option to automatically determine this number? Why is a new metric, concept purity, introduced instead of using the same set of metrics provided in Yuan et al. (2020)? What benefits does improving the upper bound for the Information Gain evaluation metric provide in practice?
12	2.4%	Are all participants in the trial pregnant women who are less than 36 weeks gestation? What is the rationale for changing the data distribution if the KB was compiled by medical papers? What ethical considerations were taken into account when selecting the data for the dataset?

Table 8: Questions clustered into the top 12 topics by BERTopic. *Representative Questions* are automatically identified by BERTopic.

N Example Annotations

	Question	Answer Evidence	Free-Form Answer	GPT-4 Augmented Answer
ICLR 23	What benefits does improving the upper bound for the Information Gain evaluation metric provide in practice?	to measure the quality of such models An interpretable alternative evaluation metric for likelihood-based models is Normalized Information Gain (NInGa)	interpretable and robust evaluation of full likelihood models	Improving the upper bound for the Information Gain evaluation metric provides the benefit of enabling a more interpretable and robust evaluation of full likelihood models in practice.
ICLR 22	Does the expected label always match the most probable label given the noisy data?	This makes it clear that, overall, the pattern zig-zags, first moving towards the unknown true label before eventually turning to memorize the wrong label	No, that is why zig-zag might happen for the hard labels.	No, the expected label does not always match the most probable label given the noisy data. This is indicated by the observed zig-zag pattern, where the process initially moves towards the true label but eventually turns to memorize the incorrect label.
NeurIPS 22	What criteria were used to select the videos and classes for the study?	This is inspired by the fact that complex human activities in real-world settings are usually hierarchical and compositional across space and time.	We selected videos and classes that contain complex human activities that are hierarchical and compositional.	The criteria for selecting the videos and classes for the study were based on the presence of complex human activities that are hierarchical and compositional in nature.
ARR 22	What impact does the layer norm approximation have on the performance of the tasks discussed in the paper?	Another challenge of THE-X is the attention score input of layer normalization. In most cases, the scale of multi-head attention output is very dense around [-1,1]. However, before normalization, we also observe the attention scores are scarily sparse, with some extreme values reaching 1e4, which is difficult for our LN-distill stage. To prevent the overflow attention scores, we use the weight decay of Adam optimizer as regularization.	Layernorm approximation may lead to attention overflow, which dramatically affect the performance of downstream tasks	The approximation of layer normalization can potentially lead to attention overflow. This can significantly impact the performance of the downstream tasks discussed in the paper.
COLING 20	What measures were taken to ensure that paths were not shared between the training and test sets.?	To avoid giving an unfair advantage to the path-based models, we filtered both validation and test sets to only include hyponym queries that are unseen anywhere in the full taxonomy paths of the training data. We also exclude hyponyms from the test and validation sets which appear as hyponyms in the training set7 to prevent the models from merely copying.	All queries observed during path training and all hyponyms in the training set that also appear as hyponyms in the test set are eliminated.	The authors implemented measures to ensure that paths were not shared between the training and test sets by filtering both validation and test sets to only include hyponym queries that were not present in the full taxonomy paths of the training data. Additionally, any hyponyms that appeared in the training set were excluded from the test and validation sets to prevent the models from simply replicating the data.
ACL 17	What would be the impact of not having access to pre-trained embeddings on the performance of this method?	Our main contribution is to show that the context sensitive representation captured in the LM embeddings is useful in the supervised sequence tagging setting. When we include the LM embeddings in our system overall performance increases from 90.87% to 91.93% F 1 for the CoNLL 2003 NER task, a more then 1% absolute F1 increase, and a substantial improvement over the previous state of the art. We also establish a new state of the art result (96.37% F 1) for the CoNLL 2000 Chunking task. Importantly, the LM embeddings amounts to an average absolute improvement of 1.06 and 1.37 F 1 in the NER and Chunking tasks, respectively.	Performance decreases by about 1% on the CoNLL 2003 NER task, and by 1.4% on the CoNLL 2000 Chunking task when removing pretrained language model embeddings.	The absence of pre-trained language model embeddings would result in a decrease in performance by approximately 1% on the CoNLL 2003 NER task, and by around 1.4% on the CoNLL 2000 Chunking task.
CoNLL 16	Why were post-editing rates chosen over prediction (h)ter for intrinsic uncertainty evaluation?	Our decision to focus on post-editing time was based on the fact that time is a more complete measure of post-editing effort, capturing not only technical effort like HTER, but also cognitive effort (Koponen et al., 2012). Additionally, time is more directly applicable in real translation environments – where uncertainty estimates could be useful, as it relates directly to productivity measures.	The authors state that normalised post-editing times (i.e. post-editing rates) are a more realistic measure with respect to operational effort, and can be adapted to different application scenarios.	The authors opted for post-editing rates over prediction (h)ter for intrinsic uncertainty evaluation because they believed that time is a more comprehensive measure of post-editing effort. This measure not only captures technical effort, like HTER, but also cognitive effort. Furthermore, time is more directly applicable in real translation environments, where uncertainty estimates could be beneficial, as it correlates directly to productivity measures. Thus, post-editing rates provide a more realistic measure of operational effort and can be adapted to various application scenarios.
ESD 23	What is the impact on detection or signal-to-noise ratio (SNR) if RR-based fingerprint construction is not applied?	robust detection and attribution (D&A) of forced change in the water cycle are impaired by low signal-to-noise ratios. The concept of detection and attribution is to use climate model simulations in which the applied forcings are known and internal variability can be reduced by averaging multiple realisations, to estimate a so-called fingerprint that represents the effect of the applied forcings on climate variables of interest.	detection won't be possible if no SNR-maximising method/process is applied to the noise observational data	If the RR-based fingerprint construction is not applied, it could potentially impair the detection process due to the low signal-to-noise ratios. This could make detection impossible without the application of a method or process that maximizes the signal-to-noise ratio in the observational data.
ESurf 23	Does the orientation of the grid, with potentially more north-facing slopes in the Elwha area than in the Quinault area, affect the absorption of solar radiation, thereby influencing glacier distribution?	Our flowline model captures the variability in the direction of the steepest precipitation gradients but neglects variability across the valleys that may be important, especially for spatially variable drivers of ablation, such as topographic shading.	We do not include topographic shading as a factor influencing melting in the model - so the answer to this question is no - not in our model. In the real world, the north-facing slopes of the Elwha headwaters may have been important in the mass balance of past glaciers.	The authors' model does not account for topographic shading as a factor influencing melting, thus the orientation of the grid does not affect the absorption of solar radiation in their model. However, in reality, the north-facing slopes of the Elwha headwaters may have played a significant role in the mass balance of past glaciers.
F1000 22	How was the conceptual model in Figure 2 developed?	Participants were recruited from their schools and colleges through the activity recruitment poster and " Word-of-Mouth " from the teachers and lecturers. Then, their crafts will be evaluated by art teachers and ranked on a leaderboard. The participants who were ranked on the leaderboard will receive certificates and prizes. After experiencing two-level EGL, 29 students were purposively selected as FGD participants. Figure 2 shows the overview of the EGL activity flow.	It is based on the procedures of the EGL activity from participation to focus group discussion.	The conceptual model depicted in Figure 2 was developed based on the procedures of the Experiential Group Learning (EGL) activity. This process ranged from participant recruitment to the focus group discussion.

Table 9: Exemplary questions, answer evidence, and free-form answers of the PeerQA dataset from all venues.

O Question Classes

Table 10 reports the number of questions per class and representative questions for that class. The definitions for each class are the following:

Method Clarification Questions to better understand a specific detail (e.g., a parameter) or inner workings of a proposed or used method, including methods used for obtaining data or details about the experiment setup/process.

Data Clarification Questions to understand the process of obtaining data or properties of the used data for an experiment, however, excluding questions about a method to obtain data.

Justification/Rationale Questions that challenge an assumption, ask the authors to motivate a decision's reasoning or are critical towards a process/finding.

Analysis Questions asking for a better understanding of a result, e.g., why a method works or questions asking about what factors contribute to a result/finding.

Implication Questions about potential real-world applications, transfers of the data/method/findings to other applications/domains/tasks, or wider-scoped consequences of the findings.

Definition Questions about the (intended) meaning of a certain phrase or term used in the paper.

Comparison Questions asking for comparisons or differences between methods/data or different studies.

Evaluation/Evidence Questions asking for details about a result (excluding analysis of results), details of the evaluation process, or evidence to support a certain claim.

Class	Size	Representative Questions
Method Clarification	31%	How was the fine tuning done for the step sizes in the experiments?, Did the baselines in both experiments 1 and 2 only use a single seed? What is the set of signed input gradients in the second paragraph of section 4.2?
Data Clarification	13%	Do the experts who annotated the dataset have expertise in linguistics or in the domain of the dataset? What is the time resolution of the forcing data used in the study, specifically, is it daily? Do the vocabulary items of the templates used in the paper have adequate representation in the training data?
Justification/Rationale	12%	What motivated the authors to theoretically analyze the dense case and then empirically evaluate the sparse case? Are ten locations sufficient to represent the variety of surfaces in urban environments? Why is the chosen metric appropriate for evaluating the results?
Comparison	11.5%	How does the proposed method compare to other types of vision transformers, such as Swin Transformer or Multiscale Vision Transformers? What is the difference between the MOMA dataset and the MOMA-LRG dataset? How does the performance of the filter-kd model compare to models trained using label smoothing and knowledge distillation with the optimum temperature?
Analysis	9%	What factors influence the degree of separability when adapting a model to a task? Is it clear what the source of the improvements of Histruct+ (Roberta-base) over Bertsumext are? What factors were responsible for the success of the path-based model?
Implications	8%	What are the potential applications of the data presented in this paper? Can the proposed data augmentation be applied to other tasks besides ILA? Do you think that the same framework on variance of ensembles would work equally well in the semantic feature space as in the space of logits?
Evaluation/Evidence	8%	What is the evidence that the generative model is successful in synthesizing new molecules? Do you evaluate playing strength of agents by restricting them by MCTS iteration counts or by time limits? Did the authors run multiple trials to evaluate the performance of the graph-based neural network?
Definition	7.5%	What is the definition of difficulty used in the paper to analyze the learning path of the network's predicted distribution? What is the variational approximation of c given by the query and support sets? What is the definition of f_{i+1}?

Table 10: Distribution of question classes based on 100 questions randomly sampled from PeerQA. *Representative Questions* shows manually picked questions that best correspond to the definition of the class.

P Answerability Evaluation

Table 11 shows detailed evaluation metrics for the answerability task, and Figure 3 visualizes them. We report Precision, Recall, and F1-Score on both the answerable and unanswerable questions, as well as the average accuracy, weighted, and macro F1-Score.

		Answerable ($N = 383$)			Unanswerable ($N = 112$)			Average		
Model	Ctx.	Prec.	Recall	F1	Prec.	Recall	F1	Acc.	W-F1	M-F1
Llama-3 IT-8B-8k	G	**1.0000**	0.4517	0.6223	–	–	–	0.4517	0.6223	0.3112
	10	0.8407	0.3995	0.5416	0.2652	**0.7411**	0.3906	0.4768	0.5074	0.4661
	20	**0.8796**	0.2480	0.3870	0.2558	**0.8839**	0.3968	0.3919	0.3892	0.3919
	50	0.7907	0.1775	0.2900	0.2298	**0.8393**	0.3608	0.3273	0.3060	0.3254
	100	0.7667	0.1802	0.2918	0.2247	**0.8125**	0.3520	0.3232	0.3054	0.3219
	FT	0.8168	0.5587	0.6636	0.2747	0.5714	0.3710	0.5616	0.5974	0.5173
Llama-3 IT-8B-32k	G	**1.0000**	0.4047	0.5762	–	–	–	0.4047	0.5762	0.2881
	10	0.8326	0.4804	0.6093	0.2737	0.6696	0.3886	0.5232	0.5593	0.4989
	20	0.8182	0.4700	0.5970	0.2618	0.6429	0.3721	0.5091	0.5461	0.4846
	50	0.8056	0.3786	0.5151	0.2444	0.6875	0.3607	0.4485	0.4802	0.4379
	100	0.7984	0.2585	0.3905	0.2345	0.7768	0.3602	0.3758	0.3837	0.3754
	FT	0.8488	0.3812	0.5261	0.2663	0.7679	0.3954	0.4687	0.4965	0.4608
Mistral IT-v02-7B-32k	G	**1.0000**	**0.8877**	**0.9405**	–	–	–	**0.8877**	**0.9405**	**0.4703**
	10	0.7854	**0.9269**	**0.8503**	**0.3488**	0.1339	0.1935	**0.7475**	**0.7017**	0.5219
	20	0.7790	**0.9295**	**0.8476**	0.2895	0.0982	0.1467	**0.7414**	0.6890	0.4971
	50	0.7768	**0.9086**	**0.8375**	0.2553	0.1071	0.1509	**0.7273**	0.6822	0.4942
	100	0.7824	**0.9295**	**0.8496**	**0.3250**	0.1161	0.1711	**0.7455**	**0.6961**	0.5103
	FT	0.7803	**0.9739**	**0.8664**	**0.4118**	0.0625	0.1085	**0.7677**	**0.6949**	0.4875
Command-R v01-34B-128k	G	**1.0000**	0.7232	0.8394	–	–	–	0.7232	0.8394	0.4197
	10	0.7985	0.8172	0.8077	0.3204	0.2946	0.3070	0.6990	0.6944	**0.5574**
	20	0.8025	0.8381	0.8199	**0.3474**	0.2946	0.3188	0.7152	**0.7065**	**0.5694**
	50	0.8031	0.8198	0.8114	**0.3365**	0.3125	0.3241	0.7051	**0.7011**	0.5677
	100	0.7949	0.8094	0.8021	0.3048	0.2857	0.2949	0.6909	0.6873	0.5485
	FT	0.8113	0.7520	0.7805	0.3214	0.4018	0.3571	0.6727	0.6847	**0.5688**
GPT-3.5 Turbo-0613-16k	G	**1.0000**	0.4935	0.6608	–	–	–	0.4935	0.6608	0.3304
	10	0.8107	0.4360	0.5671	0.2526	0.6518	0.3641	0.4848	0.5211	0.4656
	20	0.8248	0.5039	0.6256	0.2720	0.6339	0.3807	0.5333	0.5702	0.5032
	50	0.8168	0.5587	0.6636	0.2747	0.5714	0.3710	0.5616	0.5974	0.5173
	100	0.8507	0.4465	0.5856	0.2789	0.7321	0.4039	0.5111	0.5445	0.4948
	FT	0.8348	0.2507	0.3855	0.2447	**0.8304**	0.3780	0.3818	0.3838	0.3818
GPT-4o 0806-128k	G	**1.0000**	0.4465	0.6173	–	–	–	0.4465	0.6173	0.3087
	10	**0.8439**	0.5222	0.6452	0.2907	0.6696	**0.4054**	0.5556	0.5909	0.5253
	20	0.8560	0.5744	0.6875	0.3151	0.6696	**0.4286**	0.5960	0.6289	0.5580
	50	**0.8604**	0.5953	0.7037	0.3261	0.6696	**0.4386**	0.6121	0.6437	**0.5712**
	100	**0.8543**	0.5666	0.6813	0.3112	0.6696	**0.4249**	0.5899	0.6233	**0.5531**
	FT	**0.8458**	0.5300	0.6517	0.2941	0.6696	**0.4087**	0.5616	0.5967	0.5302

Table 11: Evaluation results on the answerability task of various LLMs, with different context settings (G = Gold Evidence, FT = Full-Text, 10/20/50/100 = Top-k passages). Note that the class distribution is imbalanced. There are a total of 383 answerable and 112 unanswerable questions. W-F1 is Weighted F1, M-F1 is Macro F1.

Q Answer Generation Evaluation

Table 12 reports the exact numbers of the free-form answer generation experiment for all models and contexts, corresponding to Figure 4.

		Rouge-L			AlignScore			Prometheus	
Model	Ctx.	AE	FF	GPT-4 FF	AE	FF	GPT-4 FF	FF	GPT-4 FF
Llama-3 IT-8B-8k	G	0.1683	0.2295	0.2569	0.5731	0.1098	0.2643	3.1102	3.1593
	10	0.1670	0.2113	**0.2479**	0.3839	0.1107	0.2107	3.1347	3.1828
	20	0.1771	0.2074	0.2458	0.3719	0.1041	0.1965	3.1878	3.2454
	50	0.1621	0.2050	0.2357	0.3402	0.1062	0.1958	3.0122	3.0313
	100	0.1418	0.2069	0.2278	0.3255	0.1067	0.2184	2.8082	2.7885
	FT	0.1484	0.1736	0.2037	0.2719	0.0653	0.1159	2.7510	2.9321
Llama-3 IT-8B-32k	G	0.1648	0.2286	0.2567	0.5778	0.1016	0.2436	3.1673	3.1749
	10	0.1513	**0.2258**	0.2464	0.3970	0.1142	0.2177	3.1388	3.1410
	20	0.1558	0.2204	0.2425	0.4001	0.1115	0.2109	3.1388	3.1227
	50	0.1546	0.2061	0.2397	0.3750	0.0999	0.2011	3.0571	3.1358
	100	0.1664	0.2099	0.2412	0.3785	0.1037	0.2008	3.0000	3.2010
	FT	0.1835	0.1948	0.2260	0.3311	0.0711	0.1450	3.1959	3.2167
Mistral v02-7B-32k	G	**0.2442**	0.1922	0.2432	0.6407	0.0827	0.1977	3.4245	**3.4517**
	10	**0.1967**	0.1667	0.2032	0.3573	0.0612	0.1094	3.2490	3.3629
	20	**0.2039**	0.1670	0.2011	0.3449	0.0505	0.1107	3.2408	3.2663
	50	**0.2023**	0.1572	0.1943	0.3211	0.0496	0.1017	3.1306	3.1958
	100	**0.2023**	0.1593	0.1927	0.3142	0.0634	0.1209	3.0245	3.0809
	FT	**0.1883**	0.1344	0.1678	0.2599	0.0328	0.0750	2.9796	3.1227
Command-R v01-34B-128k	G	0.1310	0.2294	0.2081	0.5604	0.1362	0.3059	3.0571	3.0052
	10	0.1211	0.2104	0.1973	0.3767	0.1221	0.2275	3.1551	3.1723
	20	0.1220	0.2164	0.1978	0.3823	0.1245	0.2213	3.0490	3.0052
	50	0.1229	0.2188	0.1941	0.3872	0.1223	0.2247	3.1224	3.0026
	100	0.1244	0.2200	0.1853	0.3688	0.1112	0.1976	3.0245	3.0052
	FT	0.1230	**0.2085**	0.1859	0.3530	**0.1015**	**0.1939**	2.9020	2.9869
GPT-3.5 Turbo-0613-16k	G	0.1540	**0.2414**	0.2688	0.5596	**0.1378**	**0.3175**	3.0408	3.0705
	10	0.1342	0.2212	0.2462	**0.4410**	**0.1412**	**0.2531**	2.9184	3.0313
	20	0.1388	**0.2211**	**0.2465**	**0.4255**	**0.1446**	**0.2394**	2.9714	3.0888
	50	0.1365	**0.2205**	**0.2437**	0.4159	**0.1356**	**0.2374**	2.9918	3.0914
	100	0.1297	**0.2207**	**0.2437**	0.4092	**0.1360**	**0.2301**	2.9102	3.0470
	FT	0.1162	0.1895	0.2188	0.3341	0.0771	0.1524	2.7143	2.9060
GPT-4o 0806-128k	G	0.1992	0.2266	**0.2739**	**0.6410**	0.1224	0.2802	**3.4612**	3.4308
	10	0.1765	0.2048	0.2455	0.4055	0.0884	0.1963	**3.5143**	**3.5222**
	20	0.1798	0.2039	0.2453	0.4094	0.0963	0.1830	**3.5510**	**3.5927**
	50	0.1771	0.2058	0.2433	**0.4164**	0.0971	0.1926	**3.5592**	**3.6423**
	100	0.1793	0.2036	0.2436	**0.4120**	0.0936	0.1886	**3.5714**	**3.5614**
	FT	0.1821	0.1981	**0.2372**	**0.3900**	0.0713	0.1790	**3.5673**	**3.6057**

Table 12: Evaluation results on the answer generation task of various LLMs, with different context settings (G = Gold Evidence, FT = Full-Text, 10/20/50/100 = Top-k passages) and the metric computed against different ground truths (AE = Answer Evidence Paragraph, FF = Free-Form Answer, GPT-4 FF = GPT-4 rephrased Free-Form Answer). Rouge-L measures lexical overlap; AlignScore measures factual consistency; Prometheus measures answer correctness using an LLM-as-a-judge approach between the generation and the annotated Free-Form Answer or Answer Evidence.

R Answer Generation Error Analysis

As outlined in §5.3, we conducted an error analysis on GPT-3.5's generations. Table 13 defines each error class, and Table 14 provides an example for each class.

R.1 Error Classes

Error Class	Definition
Evaluation Error	The generated answer is correct; however, at least one of the metrics provides a low score.
Partially Correct	The generated answer is correct; however, the free-form answer provides additional details that are not covered by the generation.
Reasoning Error	The generation is incorrect. The model fails to arrive at the same conclusion as the free-form answer.
Implicit Evidence Only	The generation is incorrect. The evidence only implies the correct answer, making it challenging for the model to infer it.
Insufficient Context	The answer is incorrect because further context is required to interpret the evidence correctly (e.g., abbreviations in the context are not resolved, or information established earlier in the paper is missing from the evidence).
Insufficient Free-Form Answer	The author's free-form answer points only to an answer in the paper but does not contain an answer (e.g., "The architecture is depicted in Figures D.3 and D.4").
Insufficient Evidence	The highlighted context by the authors does not provide sufficient information to answer the question well.

Table 13: Error classes definitions for analyzing the failure modes of the generation models.

R.2 Error Examples

Table 14: Examples for each error class.

Error Class	Question	Evidence	Reference Answer	Generated Answer
Evaluation Error	What does "but train the detector head with $r(\cdot)$ online in a single stage" mean?	To obtain the features for a region b, we apply the VLM pooling layer P(•) on the cropped backbone output features F(I) (see Sec. 3.2 for notations). Because the pooling layer requires fixed-size inputs, e.g. 7x7 for R50 (Radford et al., 2021), we crop and resize the region features with ROI-Align R(•) (He et al., 2017) (see Fig. 2b). Unlike existing works (Gu et al., 2022;Du et al., 2022), we do not crop and resize the RGB image regions and cache their embeddings in a separate offline process, but train the detector head in one stage. This is simpler and more space-efficient. In addition, we do not crop VLM region features with R(•) during training because the backbone features are frozen.	It means that we train the detection coordinates and the class labels for all boxes simlutanously and jointly, similar to Mark-RCNN and related algorithms, as opposed to cropping individual object boxes and feeding their features to a separate model.	"But train the detector head with $r(\cdot)$ online in a single stage" means that the detector head is trained during the same stage as the rest of the training process, rather than in a separate offline process.
Partially Correct	Does the term "aes" refer to the phenomenon that smiling faces are estimated to be older than neutral faces in direct age estimations, or does it also refer to the phenomenon that smiling face group is retrospectively estimated to be younger?	Interestingly, several studies have reported that humans have a counterintuitive bias regarding age. We associate smiling with youth, that is, it is generally believed that when people see a smiling person, they feel that person is younger. Indeed, previous research has provided evidence that individuals with a smile appear younger than those with other facial expressions (Hass et al., 2016;Voelkle et al., 2012). However, contrary to the commonly held association between smiling and youth, Ganel (2015) showed that a smiling face is estimated to be older than a neutral face. This phenomenon, in which smiling faces are evaluated as being older than neutral faces, is called the aging effect of smiling (AES; Ganel & Goodale, 2021). AES is attributed to wrinkles around the eyes caused by smiling (Ganel, 2015;Ganel & Goodale, 2021). In contrast, when participants were asked to retrospectively estimate the mean age of several faces (i.e., face group), they estimated that the smiling face group was younger than the neutral face group (Ganel & Goodale, 2018). These studies indicate that the effect of emotional expressions on age estimation depends on the method of estimation (i.e., directly or retrospectively).	The term AES refers specifically to the phenomenon that a smiling face makes one look older.	The term "aes" refers to both the phenomenon that smiling faces are estimated to be older than neutral faces in direct age estimations and the phenomenon that the smiling face group is retrospectively estimated to be younger.
Reasoning Error	How are the depth and normal maps combined for training the proposal network in phase 1?	There are two possible ways to ensemble geometric cues: (1) Stack the two geometric cues together and train a single object proposal network on these stacked inputs in Phase-I; (2) Train two object proposal networks and extract pseudo boxes separately, then merge them into a single pseudo box pool for Phase-II training. The details of the merging process is described in Appendix A. We conduct ablation studies on these two methods. From Figure 6a, we demonstrate that empirically, ensembling pseudo labels is slightly better than using stacked inputs for Phase-I training. Throughout the paper, we use the pseudo label ensembling for GOOD-Both.	They are combined by merging their pseudo labels.	The depth and normal maps are combined by stacking them together and training a single object proposal network on these stacked inputs in Phase 1.
Implicit Evidence Only	Does combining RGB, depth, and normals improve the performance of the model for smaller classes?	More comparison of GOOD-Both and GOOD-All. For GOOD-All, the performance gains in detecting small objects (AR s) are too small to compensate for the losses in detecting larger objects (AR m and (AR l)), leading to overall inferior performance.	The combination of RGB, depth, and normals does not significantly improve the model's performance for smaller classes. The gains in detecting small objects are too small to offset the losses in detecting larger objects, leading to an overall inferior performance.	The combination of RGB, depth, and normals does not significantly improve the model's performance for smaller classes. The gains in detecting small objects are too small to offset the losses in detecting larger objects, leading to an overall inferior performance.
Insufficient Context	Do the training languages used for the non-baseline models in the cross-lingual experiments have any overlap with German?	In our experiments we make use of the following datasets: The English Nancy Krebs dataset (16h) from the Blizzard challenge 2011 (Wilhelms-Tricarico et al., 2011;King and Karaiskos, 2011); The German dataset of the speaker Karlsson (29h) from the HUI-Audio-Corpus-German (Puchtler et al., 2021); The Greek (4h), Spanish (24h), Finnish (11h), Russian (21h), Hungarian (10h), Dutch (14h) and French (19h) subsets of the CSS10 dataset (Park and Mulc, 2019). In order to investigate the effectiveness of our proposed LAML procedure, we train a Tacotron 2 model and a FastSpeech 2 model on the full Karlsson dataset as a strong baseline. We also train another Tacotron 2 model and another FastSpeech 2 model on speech in 8 languages with one speaker per language (Nancy dataset and CSS10 dataset) and fine-tune those models on a randomly chosen 30 minute subset from the Karlsson dataset. To our surprise, we did not only match, but even outperform the model trained on 29 hours with the model fine-tuned on just 30 minutes in multiple metrics.	The pretraining data does not overlap with German. Only the data used for fine-tuning is in German.	Yes, the training languages used for the non-baseline models in the cross-lingual experiments have overlap with German. The German dataset of the speaker Karlsson is used for training, along with the English Nancy Krebs dataset and subsets of the CSS10 dataset which include Dutch and French.
Insufficient Evidence	Can the parameters of the BLOSUM matrix be estimated by the model instead of being pre-computed?	In the first ablation study, we investigate the influence of the BLOSUM embeddings by removing them as input to the GRU. Overall, the absence of the BLOSUM embeddings slows down convergence and sometimes make the learning process unstable, but ultimately does not strongly affect accuracy (see Figure 5). 3). The results without the BLOSUM embeddings are shown in pink (MAP) and purple (marginal).	The Blosum matrix is used to serve as a very simple to compute encoding technique to stabilize and accelerate the learning process. With larger datasets it is not necessary to use the Blosum encodings or averaged blosum embedding.	No, the parameters of the BLOSUM matrix cannot be estimated by the model instead of being pre-computed.
Insufficient Free-Form Answer	What is the architecture of the system being discussed in the paper?	*<Omitted for brevity>*	The architecture is depicted in Figures D.3 and D.4	The architecture is depicted in Figures D.3 and D.4

S Answer Generation Correlation Analysis

S.1 Recall

Figure 12 visualizes the relationship between the recall of the retrieval model (in this case `SPLADEv3`) at different cutoffs and the answer generation performance measured by different metrics.

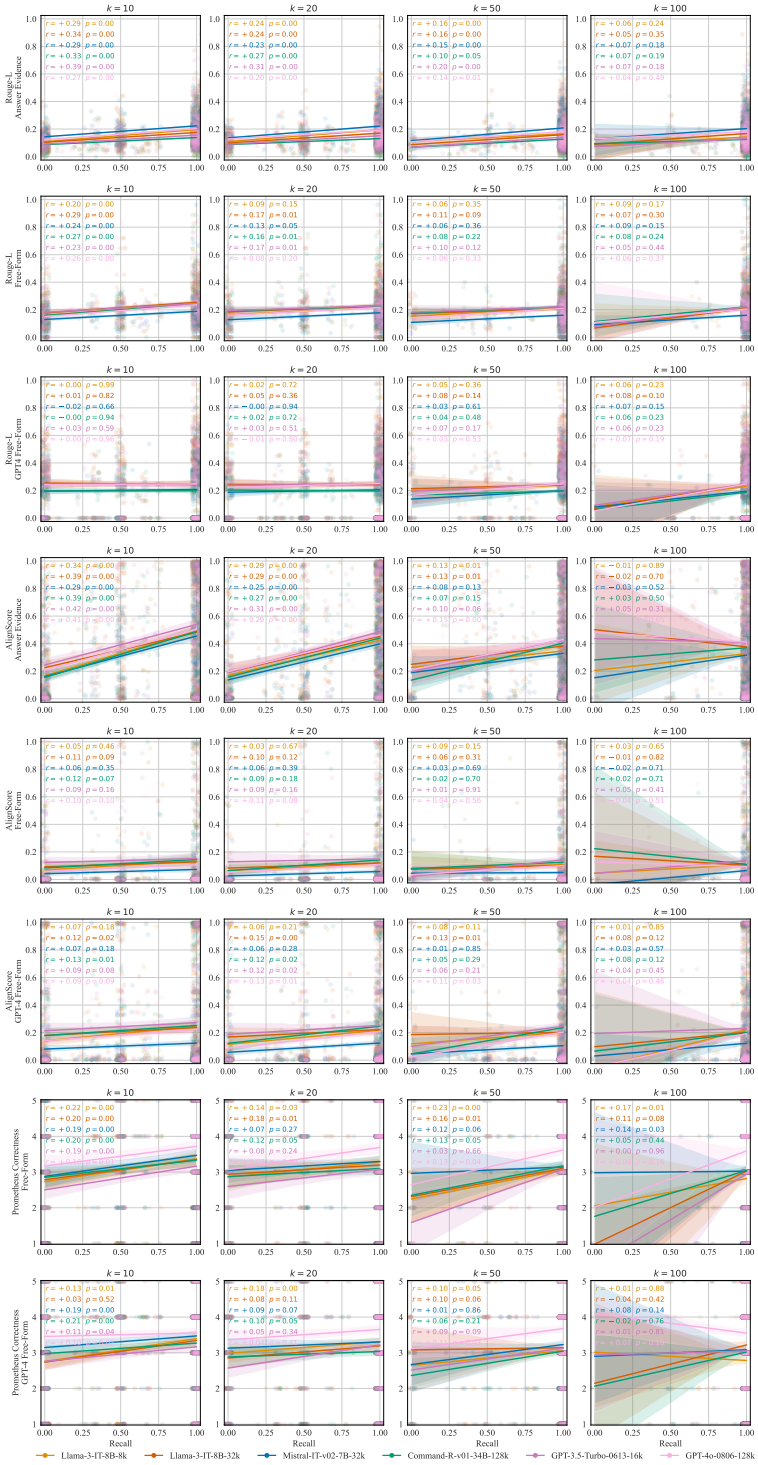


Figure 12: Pearson correlation (r) with the corresponding p-value between the recall (x-axis) at k (columns) and the answer generation performance (y-axis) according to different metrics (rows). Therefore, each circle represents a single QA pair of a specific model. We added 0.03 x-jitter to the markers to improve visibility.

S.2 Mean Evidence Position

Figure 13 visualizes the Pearson correlation between the answer generation metric (Rouge-L, AlignScore, or Prometheus-2 compared to either the answer evidence, the annotated free-form answer or the GPT-4 augmented free-form answer as ground truth) and the mean token position of the answer evidence. All generations are taken from the full-text setting, i.e., where the entire paper text was given as input to the model. To compute the mean token position for each answer evidence, we compute the number of tokens in the paper before the evidence sentence. If a question has multiple answer evidence, we take the average position. We only find a weak relationship that is statistically insignificant in many cases. Nevertheless, some p-values show statistical significance, indicating that for some settings, the generation performance declines when the answer evidence is relatively towards the end of the paper. This finding is also consistent with related work such as Buchmann et al. (2024).

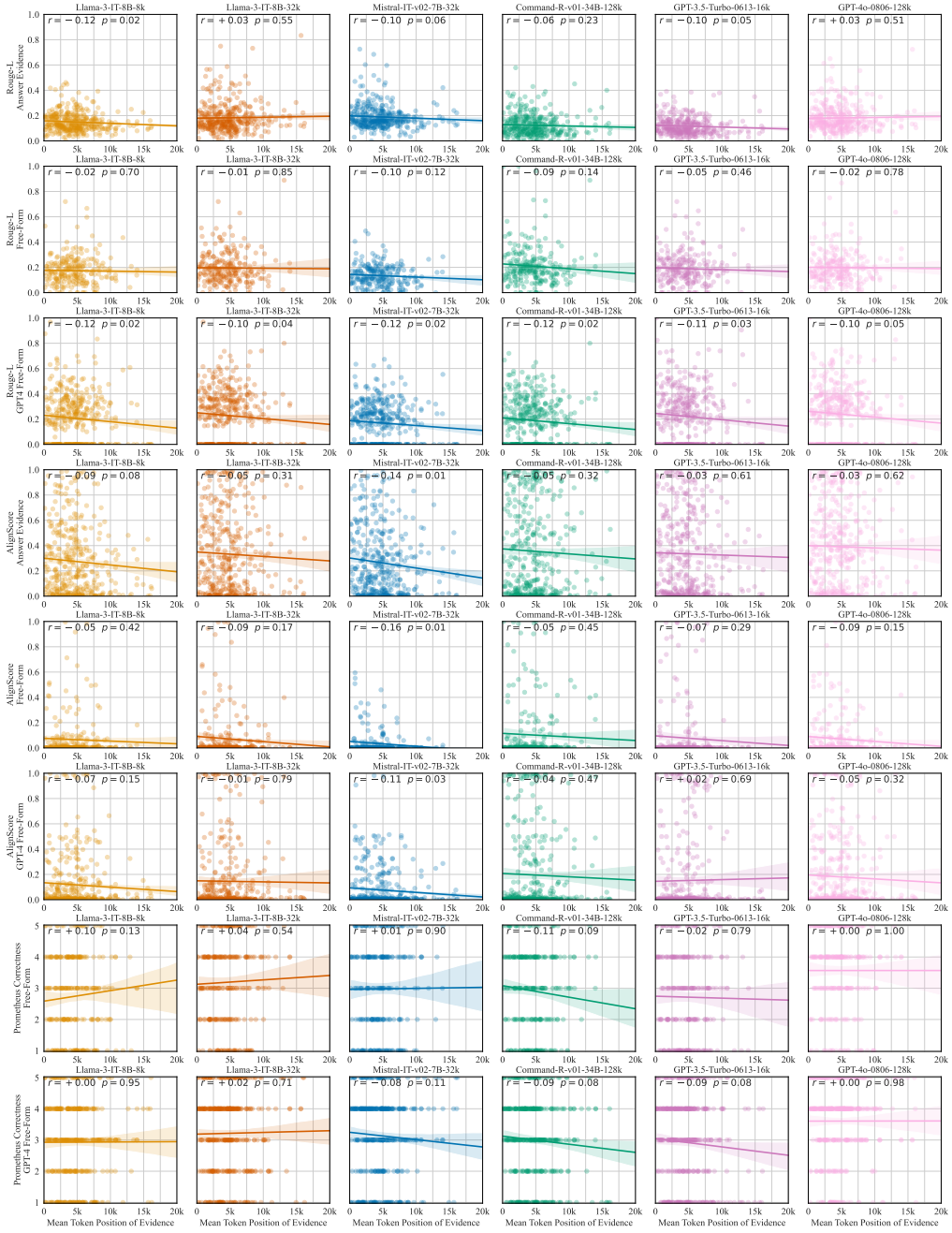


Figure 13: Pearson correlation (r) with the corresponding p-value between the answer generation evaluation metric (y-axis) and the mean token position of the annotated answer evidence (x-axis).

T Answer Generation Similarities

We compute the average similarity of the generated answers between all models. We embed the generated answers with `all-MiniLM-L6-v2` and compute the cosine similarity between the generations of the models. Figure 14 visualizes the similarities with the gold and retrieved evidence and full-text settings. We find that all models produce fairly similar outputs for the gold setting, i.e., where the annotated answer evidence is provided. With increasing retrieved evidence as context (i.e., RAG-10 - RAG-100), the similarity between the model outputs decreases but remains relatively high.

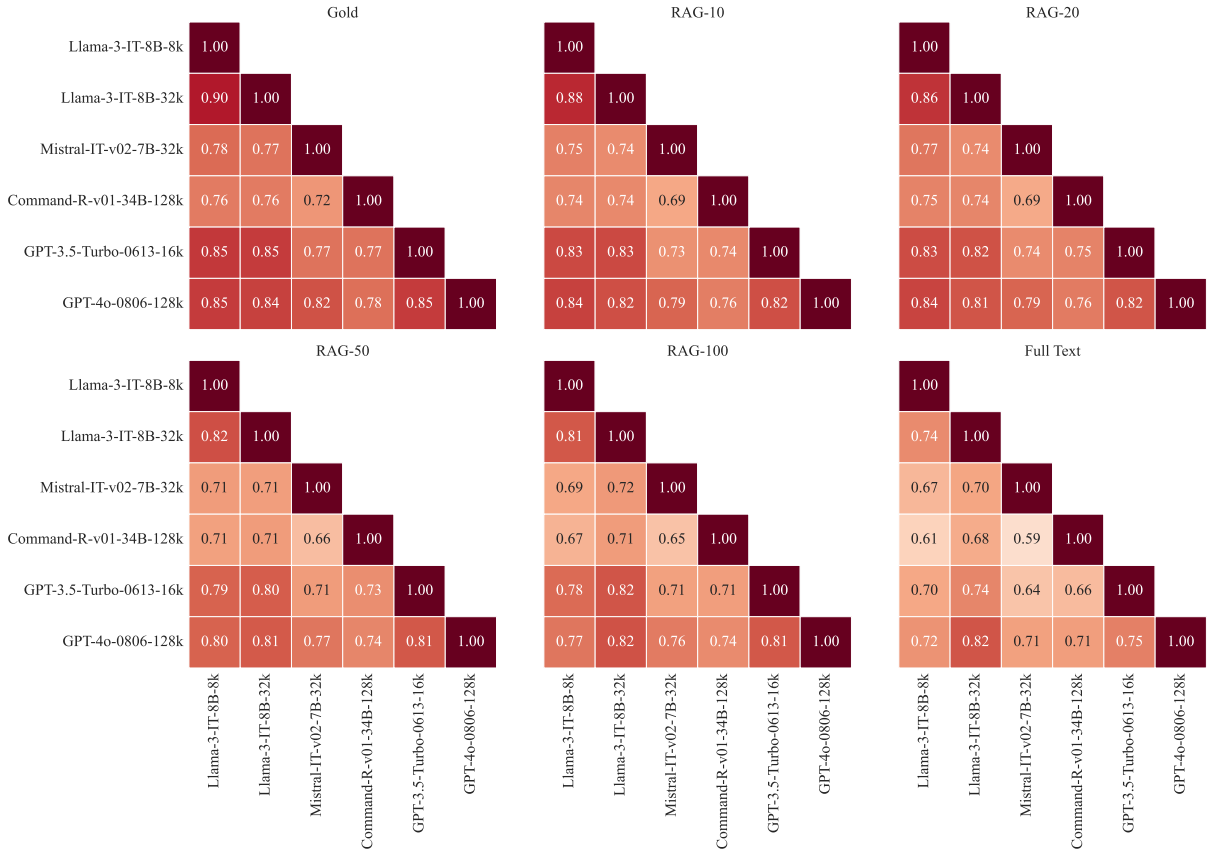


Figure 14: Semantic similarity of the generated answers between models with different context settings.

U Attributable Question Answering

	MRR	Recall@10
SPLADEv3	0.4536	0.6661
GPT-3.5-Turbo-0613-16k	0.2440	0.2762
GPT-4o-0806-128k	0.5429	0.5339

Table 15: Evidence retrieval scores in the attributable question answering setting.

We have considered answer generation based on the top retrieved paragraphs (RAG) or using the full context (§4.2). In the RAG setup, the answer generation can generally be attributed to the retrieved passages (assuming the model is faithful to the context). However, when using the full text as context, attribution to the passage level is not trivial. Recently, attributable question answering has gained momentum (Bohnet et al., 2022; Gao et al., 2023; Malaviya et al., 2024), where in addition to generating an answer, the model is supposed to cite evidence supporting it. Therefore, we also conduct an experiment where the model is conditioned on the full text of the paper and is tasked to "cite" any paragraphs on which the generated answer is based. We prepend an id before each paragraph and include an instruction on how to cite. Specifically, we use the following prompt:

`Read the following paper and answer the question. Provide one or several evidence paragraphs that can be used to verify the answer. Give as few paragraphs`

```
as possible, but as many that provide evidence to the answer. Your answer must have the following format: "<answer> [X] [Y]". In your reply, replace <answer> with your answer to the question and add any references in square brackets. Your answer must be followed by the ids of the relevant segments from the document.
Question: {question}
Paper: {paper}
Answer:
```

This setting has the challenge that the model does not provide a ranked list of all paragraphs but an unordered list of what it considers relevant. Therefore, we rank the cited paragraphs in the order in which the LLM generates them.

Table 15 reports the results of the evidence retrieval with the attributable question answering setup. We find that for GPT-3.5, the scores fall far behind the performance of a dedicated retrieval model (e.g., SPLADEv3). For GPT-4o, the MRR outperforms SPLADEv3, however, the Recall@10 is inferior.

We further investigate the answer generation performance of the attributable QA setup, reporting the results in Table 16. Compared with the RAG setting using the top 20 paragraphs retrieved by SPLADEv3, the attributable QA setup performs worse. A RAG setup is also significantly more cost and compute-efficient, particularly considering the long context of papers. Specifically, the average paragraph in PeerQA has 94 tokens, leading to an average of 1880 tokens to encode in the RAG-20 setting. In contrast, on average, a paper has 11723 tokens. Therefore, the full-text setup is 6.24 times more expensive than the RAG-20 setting.

		Rouge-L			AlignScore			Prometheus	
Model	Ctx.	AE	FF	GPT-4 FF	AE	FF	GPT-4 FF	FF	GPT-4 FF
GPT-3.5 Turbo-0613-16k	20	**0.1388**	**0.2211**	**0.2465**	**0.4255**	**0.1446**	**0.2394**	**2.9714**	**3.0888**
	FT	0.1162	0.1895	0.2188	0.3341	0.0771	0.1524	2.7143	2.9060
	FT Cite	0.1099	0.1846	0.2057	0.2453	0.1128	0.1564	2.4340	2.4837
GPT-4o 0806-128k	20	0.1798	**0.2039**	**0.2453**	**0.4094**	0.0963	**0.1830**	3.5510	3.5927
	FT	**0.1821**	0.1981	0.2372	0.3900	0.0713	0.1790	**3.5673**	**3.6057**
	FT Cite	0.1262	0.1857	0.1602	0.2678	**0.1177**	0.1622	2.7143	2.5614

Table 16: Answer generation scores in the attributable question answering setting ("FT Cite") and two baselines for comparisons. In bold the best performing setup per metric.