EXPERTQA: Expert-Curated Questions and Attributed Answers

Anonymous ACL submission

Abstract

As language models are adopted by a more sophisticated and diverse set of users, the importance of guaranteeing that they provide factually correct information supported by verifiable sources is critical across fields of study. This is especially the case for high-stakes fields, such as medicine and law, where the risk of propagating false information is high and can lead to undesirable societal consequences. Previous work studying attribution and factuality has not focused on analyzing these characteristics of language model outputs in domain-specific scenarios. In this work, we conduct human evaluation of responses from a few representative systems along various axes of attribution and factuality, by bringing domain experts in the loop. Specifically, we collect expert-curated questions from 484 participants across 32 fields of study, and then ask the same experts to evaluate generated responses to their own questions. In addition, we ask experts to improve upon responses from language models. The output of our analysis is EXPERTQA, a high-quality long-form QA dataset with 2177 questions spanning 32 fields, along with verified answers and attributions for claims in the answers.[1]

1 Introduction

As the influence of large language models (LLMs) grows beyond the computer science community, *experts* from various fields are rapidly adapting LLMs for assistance in information-seeking scenarios. For example, medical professionals are using these systems for performing differential diagnosis (Lee et al., 2023) and researchers are using them for faster literature surveys (Krenn et al., 2022; Birhane et al., 2023; Owens, 2023). While the use of LLMs in specialized domains has many potential benefits, it also carries significant risks. False or hallucinated claims that are confidently phrased can potentially mislead experts and propagate societal harms, especially in high stakes domains such as medicine or law (Evans et al., 2021; Dash et al., 2023; Volokh, 2023).

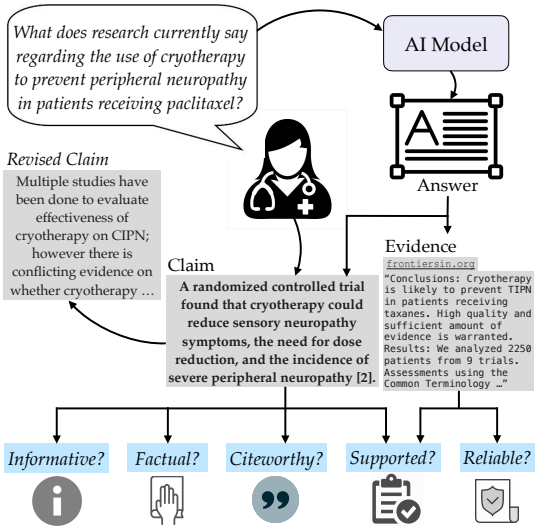

Figure 1: EXPERTQA contains 2177 information-seeking questions formulated by experts spanning 32 fields, as well as expert-verified, model-generated answers to these questions. Each claim-evidence pair in an answer is judged by experts for various properties such as the claim's informativeness, factuality, citeworthiness, whether the claim is supported by the evidence, and reliability of the evidence source. Further, experts revise the original claims to ensure they are factual and supported by trustworthy sources.

Providing citations or attributions within generated responses is a promising direction for alleviating such concerns. However, the quality of these attributions in model-generated responses, as well as the factuality of responses, is understudied in domain-specific settings. This is partly because we do not completely understand the specific information-seeking needs of experts. Although experts from different fields are naturally best suited to aid with such an evaluation, expert evaluations are rarely conducted, as bringing experts in the loop can be time-consuming and costly.

[1] Code and dataset is available at https://anonymous.

To bridge this gap, we conduct an *expert-in-the-loop* evaluation of attributed responses from a few representative systems. Having experts in the loop allows us to model a more realistic information-seeking scenario that helps us understand how people in different fields use LLMs and where their capabilities fall short. The output of our analysis is EXPERTQA, a benchmark of information-seeking questions curated by experts from 32 fields, along with verified answers from representative systems. EXPERTQA includes field-relevant questions, as well as claim-level judgements from experts along various axes of factuality and attribution.

Our evaluation is conducted by first asking qualified experts to formulate questions from their field that they are curious about or have encountered in their professional lives (§2.1). Responses to these questions are collected from a set of LLM-based systems that produce attributions for their answers (§3). These include purely generative, retrieval-augmented, and post-hoc attribution systems. We then ask experts to validate the claims and evidences found within responses to their own questions (§2.2). Experts judge each claim for its informativeness to the question, its citeworthiness, and factuality. They are also asked to judge how faithful the claim is to an accompanying evidence and rate the reliability of the evidence's source. Finally, experts revise each claim so it is faithful to reliable evidences and make a best effort attempt at ensuring the claim is factual. This overall process is described in Figure 1.

Our findings (§4) about representative systems from which responses are sampled suggest that:

1. *Retrieve-and-read systems generate more complete attributions compared to LLM prompting and post-hoc attribution, but struggle to produc citations for all cite-worthy claims.*

2. *The retrieval source significantly impacts the quality of attribution and overall factuality.*

3. *High-stakes domains such as medicine and law suffer from a large percentage of incomplete attributions (35% and 31% incomplete attributions respectively) and many attributions come from unreliable sources (51% attributions are not rated reliable by experts).*

We also measure the extent to which existing automatic methods for attribution and factuality estimation (Bohnet et al., 2022; Min et al., 2023) correlate with expert judgements (§5). We find that these metrics fall short in correlating with reference judgements of attribution and factuality. However, adapting these metrics to our data through finetuning results in improvements across domains.

The revised answers we collect can be used for improving and evaluating future models on long-form question answering. While similar datasets have been proposed (Fan et al., 2019), examples in EXPERTQA contain verified attributions and answers edited by experts. We establish several baselines and show that we can improve models by finetuning on EXPERTQA but that there is substantial room for improvement, both in terms of ROUGE and QAFactEval (§6).

2 Expert-in-the-loop Evaluation

The evaluation is conducted in multiple stages described below. In the first stage, we ask experts to write questions from their field (§2.1). In the next stage, we present responses sampled from various systems back to the same experts for analysis (§2.2). Further details about annotator backgrounds, costs and interfaces, are in Appendix A.

2.1 Stage 1: Expert-Curated Questions

Participants are recruited through Prolific and are qualified as experts if they have i) received formal education, as well as, ii) at least 3 years of work experience in their field. They are asked to write questions from their field which they have encountered in their professional life or ones they are genuinely curious about. We ask them to formulate challenging technical questions, for which it may not be possible to find a single webpage that answers them completely. We note that this question collection is aimed at closely *simulating* an information-seeking scenario with experts, since having access to real query logs is not feasible.

Each expert is asked to write 5 questions and to specify the question type(s) for each question (as shown in Table 2). These question types are formulated by adopting prior work that classifies information needs (Rose and Levinson, 2004). Because of their practical nature, at least two questions are required to be scenario-based questions (Type V, Table 2). We collect questions from 524 experts in 32 fields and manually filter them for coherence and field-relevance resulting in a dataset of 2507 questions. Examples of these questions are

Field	Question	Types
Anthropology	*Why is it that Africa's representation is still a problem in modern day times regardless of the academic writings that state otherwise?*	II,VII
Architecture	*Suppose an architect decides to reuse an existing foundation of a demolished building, what is to be considered to ensure success of the project?*	IV
Biology	*Can you explain the mechanisms by which habitat fragmentation affects biodiversity and ecosystem functioning, and provide examples of effective strategies for mitigating these impacts?*	III,VI
Chemistry	*Why does gallic acid have an affinity with trivalent iron ions?*	I
Engineering & Technology	*How different will licensing a small modular reactor be as compared to licensing traditional large nuclear power plants?*	VII
Healthcare/Medicine	*If a 48 year old woman is found to have an esophageal carcinoma that invades the muscularis propria and has regional lymph node metastases but no distant metastasis, what is her stage of cancer and what are possible recommended treatments?*	I,III
Law	*Can direct evidence in a case that has been obtained illegally be considered by the court in some cases if it directly points to the defendant's guilt?*	I
Music	*What exercises would you do in a singing class with a teenager with puberphonia?*	IV
Physics & Astronomy	*Standard Model does not contain enough CP violating phenomena in order to explain baryon asymmetry. Suppose the existence of such phenomena. Can you propose a way to experimentally observe them?*	V
Political Science	*Despite the fact that IPCC was formed in 1988, several studies have showed that argubaly more than 50% of all carbon emissions in history have been released since 1988. What does this show about IPCC and developed countries' efforts?*	VII
Visual Arts	*Tell me the step by step process of recycling a canvas.*	III

Table 1: Examples from EXPERTQA. See Table 15 for a larger list showing an example from all fields. A large percentage of examples come from high-stakes fields such as Medicine and Law.

	Question Type	Count
I	Directed question that has a single unambiguous answer	444
II	Open-ended question that is potentially ambiguous	528
III	Summarization of information on a topic	371
IV	Advice or suggestions on how to approach a problem	251
V	Question that describes a hypothetical scenario and asks a question based on this scenario	853
VI	Request for a list of resources where one can find more information	160
VII	Request for opinion on a topic	207

Table 2: Question types categorized according to various information needs that are part of EXPERTQA.

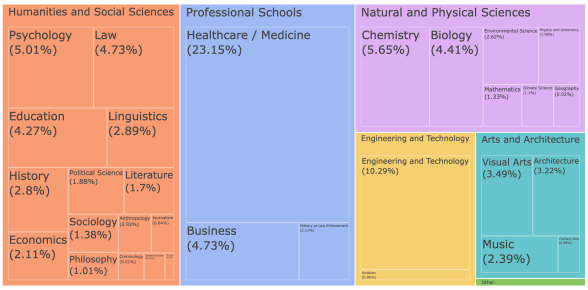

Figure 2: The distribution of questions across different fields in EXPERTQA.

presented in Table 1.

2.2 Stage 2: Answer and Claim Annotation

Next, we generate responses for the questions from stage 1 by prompting six different systems, described in §3, that provide attributions with their answers. We split each answer into claims, where claims are considered at the granularity of a sentence and extracted using the spaCy sentence tokenizer (Honnibal and Montani, 2017).[2]

In this stage of annotation, experts validate responses to their own questions on several dimensions of quality. 92% of annotators from stage 1 validated at least 1 of their own questions. The properties of answers and claims evaluated are shown in Table 3. Properties that judge answer quality are marked with A and those that judge evidence quality are marked with ❞. After labeling these claim properties, annotators edit the response to ensure that the claim is factually correct and the given references support the claim.

[2]We also considered further increasing the atomicity of claims (like Kamoi et al. (2023)) but evaluating finer-grained atomic claims incurs considerably higher annotation cost.

3 Systems Evaluated

We now describe the classes of systems from which we sampled responses to questions. All systems we evaluated produce an answer string and attributions in the form of in-line citations. Attributions are returned as URLs or passages along with URLs from where they are retrieved. Experimental details such as prompts are in Appendix B.

LLM as generator + retriever. In this paradigm, we prompt large language models in a closed-book fashion (Brown et al., 2020; OpenAI, 2023) to generate an answer with in-line citations where they provide URLs for each citation. This means that the model essentially has to generate a URL from its parametric memory. We consider GPT-4 as the LLM from which we sample responses (`gpt4`).

Post-hoc retrieval. This system differs from the above, as we only prompt LLMs to generate answers without attribution, and perform retrieval of evidence for a claim as a post-hoc step. This renders the attributions naturally unfaithful, but we believe this is still a worthwhile approach to investigate because of the strength of LLMs as generators

Property	Description	Ratings
(A) Answer Usefulness	*Is the answer useful in responding to the question?*	{*Useful*, *Partially useful*, *Not useful at all*}
(A + 👀) Attribution	*Is the claim supported by its accompanying evidence?*	{*Complete*, *Partial* or *Incomplete*, *Missing*, *N/A* (if link broken)}
(A) Informativeness	*Is the claim relevant to answering the question?*	{*Very relevant*, *A bit relevant*, *Not too important*, *Uninformative*}
(A) Factuality	*Is every word of the claim factually correct?*	{*Definitely correct*, *Probably correct*, *Unsure*, *Likely incorrect*, *Definitely incorrect*}
(👀) Source Reliability	*Is the accompanying evidence (if any) for the claim found on a website you would consider reliable?*	{*Reliable*, *Somewhat Reliable*, *Not reliable at all*}
(A) Cite-worthiness	*Is the claim necessary to be cited?*	{*Yes*, *No*}

Table 3: Properties of claims and evidences annotated in EXPERTQA.

System	Count	Abstention Rate
gpt4	174	0%
bing_chat	470	0.01%
rr_sphere_gpt4	279	37.89%
rr_gs_gpt4	452	22.69%
post_hoc_sphere_gpt4	403	0%
post_hoc_gs_gpt4	399	0%

Table 4: Number of examples sampled from different systems and the abstention rates of different systems.

and retrievers independently. The attribution corpora we consider are Sphere (Piktus et al., 2021) (post_hoc_sphere_gpt4), which is a large static dump of CommonCrawl, and Google search results (post_hoc_gs_gpt4).

Retrieve-and-read. In this class of systems, we first retrieve evidence for a question and then prompt a model to use the retrieved evidence to answer the question (Chen et al., 2017). As our attribution corpus, we again consider Sphere (Piktus et al., 2021) (rr_sphere_gpt4) and Google search results (rr_gs_gpt4). We use BM25 (Robertson et al., 2009) for retrieving from Sphere. We then generate an answer using GPT-4, providing the retrieved evidence as context. The model is instructed to generate in-line citations for each sentence, which refer to the passages in the context.

Commercial. We also consider commercial systems such as BingChat.[3] We sample responses using the balanced mode of BingChat (bing_chat).

3.1 Response Sampling

We sample uniformly from all systems but exclude abstained answers and constrain each answer to contain at most 10 claims. Attributions from gpt4 often point to broken links, so we sampled more responses from the other systems. The number of examples from each system and how frequently they abstain are in Table 4.

[3]The precise implementation of these systems is proprietary, but we can still draw conclusions about their utility.

4 Analysis

4.1 Data Statistics

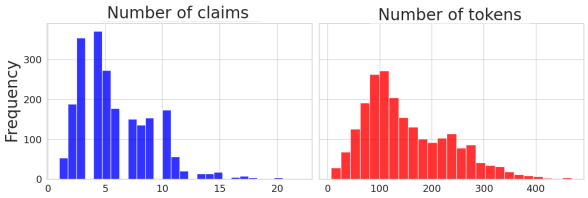

Figure 3: Histogram of the number of claims and number of tokens across all examples in EXPERTQA. The average number of claims and tokens across examples is 5.79 and 152.12 respectively

The total number of examples validated in EXPERTQA is 2177. The distribution of the number of claims and tokens is shown in Figure 3. The distribution of examples across fields and question types are presented in Figure 2 and Table 2 respectively.

4.2 Manual Analysis

To estimate the reliability of the collected human labels, we, the authors, computed our agreement with the reference labels from two fields in which the authors are experts. We sampled 60 questions each from Engineering & Technology and Medicine, sampling answers uniformly from all systems. For each claim, we label our agreement with the reference label for each property from Table 3. Our analysis, as summarized in Figure 4, shows high agreement (> 85%) for most labels in both fields considered.

4.3 Analysis of Expert Evaluations

We present the Likert distribution for claims across all systems and properties in Figure 5. Below we summarize the main conclusions from our analysis.

Majority of answers are useful, but answers from purely generative systems are considered more useful. We find that ∼87-89% of answers from gpt4 are marked useful. The retrieve-and-read systems (as well as bing_chat) are marked

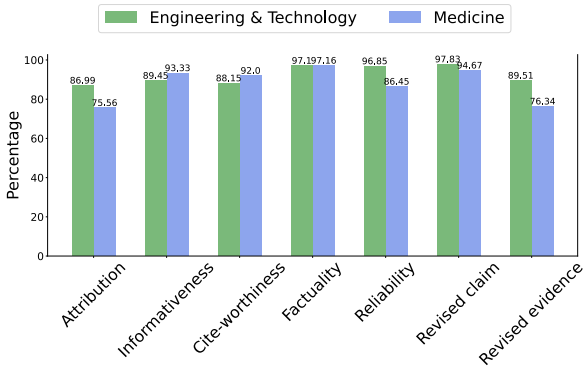

Figure 4: Percentage agreement on claim annotations based on our manual analysis.

slightly less useful (73-80%), likely because retrieved evidences are not always highly relevant. Choosing relevant evidences from the web using Google search results in more useful answers than with the smaller Sphere corpus.

Retrieve-and-read systems often generate complete attributions, but struggle to produce citations for all cite-worthy claims. While these systems have a stronger inductive bias to use the retrieved evidence to generate a response, they do not always produce attributions for cite-worthy claims (18% of these claims are missing attributions)[4]. On the other hand, post-hoc attribution systems return attributions for every single claim by design, but return more incomplete attributions. Lack of context during post-hoc retrieval can be an issue for retrieving valid attributions.

Finally, without retrieval, while `gpt4` generates citations to plausible domains (for e.g., `nasa.gov` for astronomy, `nih.gov` for medical claims), the content on these webpages is usually totally mismatched (more than 60% of the time).

Both vanilla prompting and retrieval-augmented systems generate mostly *very relevant* claims to the question. At the same time, a significant percentage of claims (30-40) are not very relevant. This includes void claims (that simply restate the question or state simplistic facts). This suggests that there is a lot of room in making answers concise and relevant.

Just over half the claims are labeled as *definitely correct* by experts. While a significant percentage of claims are labeled as correct (*probably* or *definitely*), experts do not instill high confidence in the factual correctness of claims. This might be because it is hard to judge factuality with a high degree of confidence in a short time frame. Once again, a smaller retrieval corpus (`rr_sphere_gpt4`) results in less factual claims as the model may be more likely to hallucinate.

The retrieval corpus has a significant effect on expert judgements of source reliability. Expert judgements of source reliability are directly influenced by the corpus from which evidences are retrieved. Corpora such as Sphere contain evidences that are unreliable to experts (for both `rr_sphere_gpt4` and `post_hoc_sphere_gpt4`). For example, in a question about breast cancer, evidence from a comment on a blog is retrieved and is naturally judged unreliable. Using Google search improves reliability judgements.

Majority of claims are deemed cite-worthy across systems. Only around 17-22% claims are judged not citeworthy by the experts. This suggests that most claims in responses to expert-curated questions warrant providing supporting evidence.

Domain and Question Type Trends. Figure 9 shows the distribution of labels across fields. The percentage of claims labeled factually correct is fairly high (>85%) for many fields. However, we note that across all annotated claims, **high-stakes domains such as medicine and law suffer from a significant percentage of incomplete attributions** (around 35% and 31% unsupported claims respectively). Further, **a large percentage of claims present evidences from unreliable sources** (for eg, ∼51% of medical claims have attributions from sources that are not *Reliable*). The trends across question types (Figure 10), systems clearly struggle with Type VI questions that request for a list of resources, as claims are less informative, factual, and supported by evidence.

5 Automatic Estimation of Attribution and Factuality

Prior work has proposed automatic methods to predict attribution and factuality of claims. We evaluate how reliably these methods reflect the expert labels in our collected data. We evaluate the effectiveness of these methods for claims in EXPERTQA. In both cases, we observe that **current methods show high precision but low recall when compared with human judgements.**.

[4]Figure 5 shows the Likert distribution of attribution labels on those claims deemed cite-worthy by experts.

Figure 5: The Likert distribution of labels for the different properties of answers / claims, annotated by experts. The top 3 properties (answer usefulness, claim informativeness and factuality) are judgements of answer quality and the bottom 3 (claim/evidence attribution, source reliability and claim cite-worthiness) are attribution quality.

System	AutoAIS	Num. Claims
gpt4	.156	149
bing_chat	.320	992
rr_sphere_gpt4	.689	732
rr_gs_gpt4	.778	1415
post_hoc_sphere_gpt4	.281	1158
post_hoc_gs_gpt4	.241	1500

Table 5: AutoAIS score (more attributable→1, less attributable→0) of predicted responses by the systems. Only claims annotated as *citeworthy* and *with complete support* are considered.

System	zero-shot			finetuned		
	P	R	F1	P	R	F1
gpt4	.33	.02	.05	.52	.32	.39
bing_chat	.97	.26	.41	.90	.90	.90
rr_sphere_gpt4	.89	.59	.71	.83	.90	.87
rr_gs_gpt4	.86	.74	.79	.87	.98	.92
post_hoc_sphere_gpt4	.92	.28	.43	.79	.97	.87
post_hoc_gs_gpt4	.87	.17	.29	.77	.95	.85
all	.88	.38	.53	.82	.91	.86

Table 6: Precision, Recall and F1 scores of AutoAIS labels predicted by the TRUE NLI model (0-shot vs. finetuned version on the ExpertQA train split) against human attribution judgements in ExpertQA.

5.1 Automatic Attribution Estimation

Under the *attributable to identifiable sources* (AIS) framework of Rashkin et al. (2021), previous work has found NLI models to be effective in providing automated AIS (AutoAIS) estimates (Bohnet et al., 2022). Following previous work, we use an NLI model (Honovich et al., 2022) to predict binary attribution labels of claim-evidence pairs in ExpertQA. For evidences longer than the model's sequence length (512), we use the stretching technique from Schuster et al. (2022), where we split the evidence into sentences and use the top-2 sentences with highest entailment scores as evidence.

Table 5 shows the macro-averaged AutoAIS scores for the claims annotated as having complete attributions. Compared to human judgments, the AutoAIS scores show large variance across systems. Notably, attributions from post-hoc retrieval systems receive much lower AutoAIS scores compared to retrieve-and-read systems.

We compare the per-claim AutoAIS predictions to human judgements of attribution in Table 6. The results suggest that AutoAIS estimates have high-precision yet low-recall against human judgements of attribution. To understand the discrepancy between NLI model behavior vs. human judgements, we highlight a few typical examples of attribution errors in Table 7. For NLI models, every part of the claim needs to be verifiable with the evidence, but human judgements involve more implicit world

Error Type: *Fine-grained Information Sensitivity*
Claim (post_hoc_sphere_gpt4): For water with a low pH (acidic), you can add a base or alkaline compound, such as baking soda (sodium bicarbonate) or calcium carbonate, to raise the pH [1]. **Attribution [1]**: ... To raise or lower pH, a pool custodian simply adds acids or alkalis into the water. For example, adding sodium carbonate (soda ash) or sodium bicarbonate (baking soda) will generally raise the pH and adding muriatic acid or sodium bisulfate will lower the pH. **Human**: *Cite-Worthy & Complete Support* **AutoAIS**: *0 (No or Partial Support)*.
Error Type: *Multi-Source Attributions*
Claim (bing_chat): Other radiological signs of fetal death include gas in the fetus or in the portal and umbilical vessels [1], and Deuel's halo sign [2]. **Attribution [1]**: ... Intrafetal gas is an unequivocal sign of fetal death provided it can be conclusively differentiated from maternal gas, shadows. ... **Attribution [2]**: Radiological investigation is warranted in the antenatal patient only if the findings are likely to influence future management. The major radiological signs of fetal death include overlapping of the cranial bones and Deuel's halo sign **Human**: *Cite-Worthy & Complete Support* **AutoAIS**: *0 (No or Partial Support)*.

Table 7: Examples of typical errors of AutoAIS against human judgements in EXPERTQA.

knowledge, e.g. *calcium carbonate is an alkali*. Another common mistake involves synthesizing information from multiple evidences. We observe multi-source attributions to be particularly common among bing_chat and retrieve-and-read systems.

5.2 Automatic Factuality Estimation

Prior work has proposed methods (Manakul et al., 2023; Min et al., 2023) to estimate the factuality of model generations. In particular, we use FActScore (Min et al., 2023) to estimate factuality of claims. We first break down each claim into fine-grained atomic claims using few-shot prompting with text-davinci-003. We then retrieve the top-3 relevant passages using Google search with the atomic claim as the query. The atomic claim and the evidence passages are then used to prompt gpt-3.5-turbo to say whether the atomic claim is *True* or *False*. The FActScore of a claim is the FActScore averaged across its atomic claims.

In Table 8, we report the F1 scores of the factual (T) and non-factual (F) classes and the micro-averaged overall F1 scores of the FActScore factuality scores and the reference factuality labels. FActscore scores are thresholded at 0.5 to get binary scores and reference factuality labels are 1 if the claim's factuality is labeled as *Probably correct* or *Definitely correct*, and 0 otherwise.

System	F1 (T)	F1 (F)	F1 (overall)
gpt4	0.919	0.108	0.852
bing_chat	0.912	0.134	0.841
rr_sphere_gpt4	0.884	0.106	0.795
rr_gs_gpt4	0.927	0.068	0.865
post_hoc_sphere_gpt4	0.898	0.132	0.817
post_hoc_gs_gpt4	0.939	0.158	0.886
all	0.915	0.119	0.844

Table 8: FActscore F1 scores on reference factuality labels for claims in EXPERTQA.

We find that automatic factuality estimation struggles to identify non-factual claims. In particular, predicted labels have low recall of non-factual claims. This is more often the case for retrieve-and-read systems, where the answer is generated based on retrieved evidences. The other systems use GPT-4's parametric knowledge for answer generation, which could make it easier for a similar evaluator like ChatGPT to judge factuality.

6 Long-form QA Evaluation

A beneficial output of our annotation is the revised answers produced by annotators. These answers are verified to be factual and compose a new long-form QA dataset, EXPERTQA. We consider two splits for EXPERTQA (both 80-10-10): a random split of the data and a domain-wise split, where 80% of a field's data is included in the training set and 10% is included in both validation and test sets.

6.1 Evaluation Metrics

For evaluation, we consider metrics based on similarity to a reference answer, i.e., ROUGE (Lin, 2004) and those focused on evaluating factual consistency through QA pairs generated with a reference answer, i.e., QAFactEval (Fabbri et al., 2022).

6.2 Baselines

We finetune the following open-source language models: FlanT5-11B (Chung et al., 2022), Alpaca-7B (Taori et al., 2023), Vicuna-7B (Chiang et al., 2023) and LLaMa2-7B-Chat (Touvron et al., 2023). We finetune these models with the same prompts as the ones used in their training (provided in Tables 12, 13). Further, we also report results with Llama2-70B-Chat without finetuning (marked *).

6.3 Results

Our results are shown in Table 9. We find that both Llama2-7B and Vicuna-7B outperform FlanT5-

Split	Model	R1	R2	RL	QFE
Random	FlanT5-11B	0.335	0.114	0.215	2.068
	Vicuna-7B	0.351	0.119	0.212	1.068
	Llama2-7B	0.362	0.125	0.219	1.985
	Llama2-70B*	0.320	0.101	0.181	1.050
Domain	FlanT5-11B	0.324	0.107	0.210	1.538
	Vicuna-7B	0.359	0.120	0.213	1.739
	Llama2-7B	0.363	0.124	0.219	1.726
	Llama2-70B*	0.328	0.104	0.187	0.979

Table 9: Long-form QA results after finetuning models on the random and domain splits of EXPERTQA.

11B despite the smaller model size, likely due to additional instruction finetuning for both those models. We observe that finetuning significantly improves performance (results without finetuning are in Table 14), and Llama2-70B performs worse than finetuned systems under zero-shot prompting.

7 Related Work

Attribution Generation. A few classes of systems have been proposed for generating attributions for model responses. This includes **vanilla LLM prompting** (Tay et al., 2022), where LLMs are prompted to return attributions with their answers, but the references are often hallucinated (Agrawal et al., 2023). On the other hand, **retrieve-and-read systems** (Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022) first retrieve evidence relevant for a query, and generate an answer based on the retrieved evidence. These systems are sometimes trained on human demonstrations (Nakano et al., 2021; Thoppilan et al., 2022; Menick et al., 2022). Finally, **post-hoc retrieval** (Gao et al., 2023; He et al., 2022) involves retrieving attributions after answering a query. We consider all three classes of systems for sampling responses.

Attribution Analysis Prior work has conducted analysis of system-generated attributions (Rashkin et al., 2021; Bohnet et al., 2022; Dziri et al., 2022; Chen et al., 2022; Liu et al., 2023; Muller et al., 2023; Kamoi et al., 2023; Kamalloo et al., 2023). These works suggest that systems are still far from providing precise attributions with sufficient recall for citeworthy statements. In our work, we recognize that this is problematic in specific domains where precision and recall are both critical.

Factuality Analysis. Factuality analysis of model generations has been conducted extensively in prior work (Thorne et al., 2018; Evans et al., 2021; Kryscinski et al., 2020; Maynez et al., 2020; Pagnoni et al., 2021; Lin et al., 2021; Muhlgay et al., 2023). The factuality labels we collect elicit a best-effort judgement of truthfulness of claims from experts. Prior work has also proposed methods to predict factuality of claims (Manakul et al., 2023; Kadavath et al., 2022; Agrawal et al., 2023; Azaria and Mitchell, 2023; Min et al., 2023; Feng et al., 2023; Chen et al., 2023). We use one such method (Min et al., 2023) to evaluate how well human labels in EXPERTQA correlate with automatic judgements.

Long-form QA. Existing long-form QA datasets are created using search queries (Nguyen et al., 2016; Stelmakh et al., 2022) and forums (Fan et al., 2019). Several issues have been identified with these datasets, such as vague questions and difficulty in verifying factual correctness (Krishna et al., 2021). Keeping this in mind, we construct EXPERTQA to cover practical information needs of experts along with fine-grained factuality judgements. Xu et al. (2023) conduct expert evaluation of long-form answers and emphasize the importance of evaluating multiple aspects of answers, which are also considered in our work.

Domain-specific QA. Several domain-specific QA datasets have been proposed, for domains such as medicine (Tsatsaronis et al., 2015; Pampari et al., 2018; Jin et al., 2019, 2021; Pal et al., 2022), law (Guha et al., 2023), technology (Dos Santos et al., 2015) and others (Rogers et al., 2020; Reddy et al., 2019; Hendrycks et al., 2021). However, these datasets often have limited coverage of domains. EXPERTQA contributes a unique combination of features by scaling the number of domains and providing attributions and factuality judgements.

8 Conclusion and Future Work

Our evaluation study suggests that although large language models show a lot of promise for aiding domain experts, there is large ground to cover in addressing the information needs of experts with factual and verifiable answers (Metzler et al., 2021). Experts, on the other hand, should take responses from these systems with caution, because although attributed responses can seem trustworthy, the supporting references can often be inadequate to support claims. We hope that our benchmark, EXPERTQA, can benefit the community in building improved methods for attribution & factuality estimation, and long-form question answering.

9 Limitations

Atomicity of Claims. In most cases, claims in our dataset are sentences that may not represent singular information units. This lack of atomicity in claims means that properties such as factuality and attribution need to be judged exhaustively for a claim. Collecting human judgements for finer-grained atomic claims can be significantly more expensive and is not explored in this work.

Claim Extraction. Extracting sentence-level claims from a generated answer for the purpose of evaluation is performed by using a sentence tokenizer. However, we note that existing tokenizers suffer from sentence tokenization errors (for example, when lists or tables are present in answers). This resulted in a small number of claims being excessively long and hard to evaluate.

Field Coverage. Even though we tried to cover a wide range of fields in our dataset, we missed covering questions from certain fields. Finding experts from rarer fields can be especially hard. We will consider further expanding EXPERTQA to more domains, so that it can be more broadly useful. In addition, the examples in our dataset represent the information needs of English-speaking annotators primarily based in Europe, the Americas and Africa.

Question Distribution. We elicit questions from experts by asking them to formulate questions that have come up in their professional lives or questions they are genuinely curious about. This was aimed at modeling a more realistic information-seeking scenario through our annotation. However, it is not necessary that these questions would come from a natural distribution that would be found in query logs. Since having access to such data is not possible, we attempt to match the information-seeking scenario as closely as possible.

Subjectivity of labels. Some of the properties of claims can elicit more subjective judgements, which can vary between experts from the same field. This subjectivity is not inherently captured in our data through multiple judgements, but we do estimate agreement using claims from engineering and medicine through our own labels (§4.2).

References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. 2022. Propsegment: A large-scale corpus for proposition-level segmentation and entailment recognition. *arXiv preprint arXiv:2212.10750*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi

Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Debadutta Dash, Rahul Thapa, Juan M Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H Chen, Saurabh Gombar, Lance Downing, et al. 2023. Evaluation of gpt-3.5 and gpt-4 for supporting real-world information needs in healthcare delivery. *arXiv preprint arXiv:2304.13714*.

Cicero Dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning hybrid representations to retrieve semantically equivalent questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 694–699.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *arXiv preprint arXiv:2305.08281*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *arXiv preprint arXiv:2303.01432*.

Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, Zhenpeng Yao, and Alán Aspuru-Guzik. 2022. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769.

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.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

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

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. In *Acm sigir forum*, volume 55, pages 1–27. ACM New York, NY, USA.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251v1*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Benjamin Muller, John Wieting, Jonathan H Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Baldini Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. *arXiv preprint arXiv:2305.14332*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

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. *choice*, 2640:660.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Brian Owens. 2023. How nature readers are using chatgpt. *Nature*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368,

Brussels, Belgium. Association for Computational Linguistics.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oguz, Edouard Grave, Wentau Yih, and Sebastian Riedel. 2021. The web is your oyster - knowledge-intensive NLP against a very large web corpus. *CoRR*, abs/2112.09924.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

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

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8722–8731.

Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Eugene Volokh. 2023. Large libel models? liability for ai output.

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.

A Annotation Details

Annotator backgrounds. The 484 participants involved in our study came from 26 different countries, across Europe, Africa, Oceania, North and South America. The participants were recruited through Prolific, a crowdsourcing platform[5]. To qualify as experts, participants were required to have attained a formal education in the field and have worked in the area for at least 3 years. Participants were told that their annotations will be used to evaluate the capabilities of large language models to provide truthful answers with well-supported evidences to questions from different fields. They were also informed that the data will be released publicly upon the completion of the study.

Annotator fields. The initial set of fields were listed by going through university department names, and ensuring that we cover a wide range of disciplines. Upon completing stage 1 of our annotation, we further refined these fields to represent a diverse set, for which we have enough experts.

Annotation costs. In both stage 1 and stage 2, annotators were compensated at the rate of $15 per hour with additional bonuses when annotators spent more time than we anticipated. The average time taken for stage 2 annotations was 13.83 minutes per question-answer pair. Since this task is intensive, a single annotation task was broken down into 1-3 question-answer pairs.

Annotation interface. Figures 6 and 7 show screenshots of our stage 2 annotation interface.

B Experimental Details

B.1 Hyperparameter Settings

Response collection. Across all systems, for generating responses from `gpt4`, we use a temperature of 1.0, and a maximum length of 2048 tokens. For all retrieval componenets, we use `text-embedding-ada-002` as the embedding model. The retrieve-and-read systems first retrieve top-k (k=5) evidence passages from Sphere or top-10 Google search results using the question as the retrieval query. Google search results are split into passages of 1000 tokens with 200 tokens of overlap between subsequent chunks.

On the other hand, the post-hoc citation systems simply use the claims from `gpt4` responses, but generate their own attributions by retrieving evidence for each claim in the answer. Post-hoc retrieval systems use the top-k passages (k=5) retrieved from Sphere or the top-10 Google search results with the claim as the retrieval query. Search result are split into passages the same way as retrieve-and-read systems.

Automatic attribution and factuality estimation. For automatic attribution with AutoAIS, we use the `t5_xxl_true_nli_mixture`[6] with 11B parameters by Honovich et al. (2022). For finetuning the `t5_xxl_true_nli_mixture` model on the train split of EXPERTQA, we use the DeepSpeed ZeRO optimization (Rajbhandari et al., 2020) with stage 3, a batch size of 1, a learning rate of $1e^{-4}$ and train models for 3 epochs.

Long-form QA. For finetuning FlanT5-11B, we use a batch size of 2, maximum sequence length of 512, a learning rate of 1e-4 and train models for 3 epochs. For finetuning both Llama2-7B and Vicuna-7B, we use a batch size of 4, maximum sequence length of 2048, learning rate of $2e^{-4}$ and train models for 3 epochs.

B.2 Prompts

The prompts used to generate responses from `gpt4` and `bing_chat` is provided in Table 10, while the prompt used to generate responses for retrieve-and-read systems is in Table 11.

For factuality estimation, we use the same prompts as Min et al. (2023) for both claim decomposition and atomic claim factuality prediction. Finally, for long-form QA baselines, we use the prompt in Table 12 for Llama and Table 13 for Vicuna.

C Additional Plots

Examples from all fields included in EXPERTQA are shown in Table 15. We show the distribution of all question types (from Table 2) across all fields that are part of EXPERTQA in Figure 8.

In Table 9, we summarize the label distribution of all claim properties across fields and in Table 10, we summarize the label distribution of all claim properties across question types.

In Table 14, we summarize results on long-form QA before and after finetuning models on both EXPERTQA splits.

[5] https://www.prolific.co

[6] https://huggingface.co/google/t5_xxl_true_nli_mixture

Vanilla LM QA Prompt

```
Answer the question completely and precisely in up to 500 words. You must
provide in-line citations to each statement in the answer. The citations should
appear as numbers such as [1], [2] and contain references to valid URLs on the
web. A statement may need to be supported by multiple references and should then
be cited as [1] [2].

Question: I work in the field of [FIELD]. My question is: [QUESTION]

Answer:
```

Table 10: QA Prompt for GPT4 and BingChat.

Retrieve-and-read Prompt

```
Use the following pieces of context to answer the question completely and
precisely in up to 500 words. If you don't know the answer, just say "I don't
know" and explain why the context is insufficient to answer the question.

You need to support every statement in the answer with in-line citations to
passages given in the the context. The citations should appear as numbers such
as [1], [2] that refer to the Passage IDs of the given passages. A statement may
need to be supported by multiple references and should then be cited as [1] [2].
(for example, "Paris is the capital of France [1] [2]." where "1" and "2" are
the Passage IDs of the first and second passage).

[CONTEXT]

Question: [QUESTION]
Answer:
```

Table 11: Retrieve-and-read QA prompt.

Llama2 Prompt

```
<s>[INST] «SYS»
You are a helpful, respectful and honest assistant. Always answer as helpfully
as possible, while being safe. Your answers should not include any harmful,
unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure
that your responses are socially unbiased and positive in nature.

If a question does not make any sense, or is not factually coherent, explain why
instead of answering something not correct. If you don't know the answer to a
question, please don't share false information.
«/SYS»

[QUESTION] [/INST]
```

Table 12: Llama2 prompt for long-form QA.

Vicuna Prompt

```
A chat between a curious user and an artificial intelligence assistant. The
assistant gives helpful, detailed, and polite answers to the user's questions.

USER: [QUESTION]
ASSISTANT:
```

Table 13: Vicuna prompt for long-form QA.

Split	Model	R1	R2	RL	QFE
Random	FlanT5-11B*	0.074	0.023	0.063	0.000
	FlanT5-11B	0.335	0.114	0.215	2.068
	Vicuna-7B*	0.358	0.116	0.209	0.902
	Vicuna-7B	0.351	0.119	0.212	1.068
	Llama2-7B*	0.300	0.083	0.167	1.359
	Llama2-7B	0.362	0.125	0.219	1.985
	Llama2-70B*	0.320	0.101	0.181	1.050
Domain	FlanT5-11B*	0.073	0.023	0.062	0.000
	FlanT5-11B	0.324	0.107	0.210	1.538
	Vicuna-7B*	0.352	0.114	0.203	2.596
	Vicuna-7B	0.359	0.120	0.213	1.739
	Llama2-7B*	0.303	0.087	0.169	1.799
	Llama2-7B	0.363	0.124	0.219	1.726
	Llama2-70B*	0.328	0.104	0.187	0.979

Table 14: Long-form QA results before (marked with *) and after finetuning models on the random and domain splits of EXPERTQA.

1. Login Screen

Stage 2: Detailed Instructions

Thank you for your interest in our task! We are a group of researchers conducting a study to understand how experts from various fields use AI / large language models in information-seeking scenarios. We are particularly interested in evaluating the accuracy and factual correctness of answers produced by such systems. We are inviting participants who are professionals / experts in these fields:

[Anthropology / Architecture / Biology / Business / Chemistry / Classical Studies / Criminology / Culinary Arts / Environmental Science / Economics / Education / Engineering and Technology / Geography / History / Journalism / Law / Linguistics / Literature / Mathematics / Medicine / Music / Philosophy / Physics and Astronomy / Political Science / Psychology / Theology / Sociology / Visual Arts]

The study will proceed in two stages and we would request you in both stages.

1. Question Writing: We will ask you to write a question from your domain.
2. Answer Validation and Revision: We will show you an answer produced by an AI system, and ask you to validate different aspects of this answer. We will then ask you to revise this answer to be factually correct and well-supported with citations.

The current task is the **stage 2** of the study.

Note about completion time: Note that we have made a best estimate for how long it should take to complete this task, based on a small number of participants. However, the time spent can vary across participants and across questions. If you end up spending more time than the allocated time, please feel free to let us know and we would be happy to bonus you for the extra time spent. Please prioritize **quality** and do not rush through the task. You will get a completion code after you finish annotating all 3 questions.

****Before moving on, please watch the following instruction video for this task here ****

Please enter your prolific ID down below to begin task 2:

Enter Prolific ID

Submit and start task

2. Question & Answer & Usefulness

Expert Evaluation of AI Answers: Stage 2

This task has **1 questions.**

Current Question: 1 out of 1

Make sure to **follow the instructions carefully** and submit all the questions! If an **error** occurs in the interface, just click on the link again and provide your ID.

1) Read the question, and the answer produced by the AI system.

Question:

Can direct evidence in a case that has been obtained illegally be considered by the court in some cases if it directly points to the defendant's guilt?

Answer:

No, direct evidence that has been illegally obtained cannot be used to prove a defendant's guilt. The Supreme Court of the United States has made it clear in James v. Illinois that such evidence is "inadmissible on the government's direct case, or otherwise, as substantive evidence of guilt" [1]. This statement has been reaffirmed in United States v. Havens [2] and 446 US 620 United States v. J Havens [5]. However, the Court has made exceptions to this rule for instances where prosecutors need to impeach a defendant's answers given during cross-examination, allowing them to use illegally obtained evidence for this purpose if the questions asked are within the scope of the defendant's direct examination [1][2][5]. Despite such exceptions, the overall emphasis is on safeguarding the defendant's Fourth, Fifth, or Sixth Amendment rights by not allowing violations through the use of illegally obtained evidence to hold sway in trials [4].

2) Judge whether the answer is useful to the question. Usefulness should be measured based on whether the answer is **at least partially answering the question.**

1. Useful: Answers most/all of the question.
2. Partially useful: Answers a subset of the question, or answers the question partially.
3. Not useful at all: Completely irrelevant to what the question asked for.

Usefulness *

- ○ Useful
- ○ Partially useful
- ○ Not useful at all

3. Claim & Evidence with URL + Passage

3) Following this, you will be asked to annotate the individual claims contained in the answer. Each claim is a sentence, accompanied with the evidence for the sentence returned by the system. The evidence can be presented in the form of 1) URL(s) to webpages that you may need to open, or 2) URL(s) accompanied with a relevant passage from each webpage.

If you are only given a URL, open the link to answer the questions.

If you are given a passage and a URL, you should judge support **just based on the passage.** The URL is provided simply for more context.

You are on **question 1**. This question has **5 claims.**

Current Claim: 3 out of 5

Claim:

This statement has been reaffirmed in United States v. Havens [2] and 446 US 620 United States v. J Havens [5].

Evidence:

[2] https://supreme.justia.com/cases/federal/us/446/620/
United States v. Havens :: 446 U.S. 620 (1980) :: Justia US Supreme Court Center the criminal trial. We reaffirm this assessment of the competing interests, and hold that a defendant's statements made in response to proper cross-examination reasonably suggested by the defendant's direct examination are subject to otherwise proper impeachment by the government, albeit by evidence that has been illegally obtained and that is inadmissible on the government's direct case, or otherwise, as substantive evidence of guilt. In arriving at its judgment, the Court of Appeals noted that, in response to defense counsel's objection to the impeaching evidence on the ground that the matter had not been "covered on direct," the trial court had

[5] https://m.openjurist.org/446/us/620
446 US 620 United States v. J Havens | OpenJurist permit or require that false testimony go unchallenged, with the resulting impairment of the integrity of the factfinding goals of the criminal trial. We reaffirm this assessment of the competing interests, and hold that a defendant's statements made in response to proper cross-examination reasonably suggested by the defendant's direct examination are subject to otherwise proper impeachment by the government, albeit by evidence that has been illegally obtained and that is inadmissible on the government's direct case, or otherwise, as substantive evidence of guilt. In arriving at its judgment, the Court of Appeals noted that in response to defense counsel's objection

4. Claim & Evidence with URL

3) Following this, you will be asked to annotate the individual claims contained in the answer. Each claim is a sentence, accompanied with the evidence for the sentence returned by the system. The evidence can be presented in the form of 1) URL(s) to webpages that you may need to open, or 2) URL(s) accompanied with a relevant passage from each webpage.

If you are only given a URL, open the link to answer the questions.

If you are given a passage and a URL, you should judge support **just based on the passage.** The URL is provided simply for more context.

You are on **question 1**. This question has **3 claims.**

Current Claim: 1 out of 3

Claim:

According to some sources[1] [3], AI can help architects design more sustainable buildings by using algorithms to optimize space planning, reduce waste, integrate renewable energy sources and adapt to changing environments.

Evidence:

[1] https://now.northropgrumman.com/sustainable-architecture-leans-into-artificial-intelligence/

[3] https://www.aiplusinfo.com/blog/artificial-intelligence-and-architecture/

Figure 6: Screenshots of the interface (1-4).

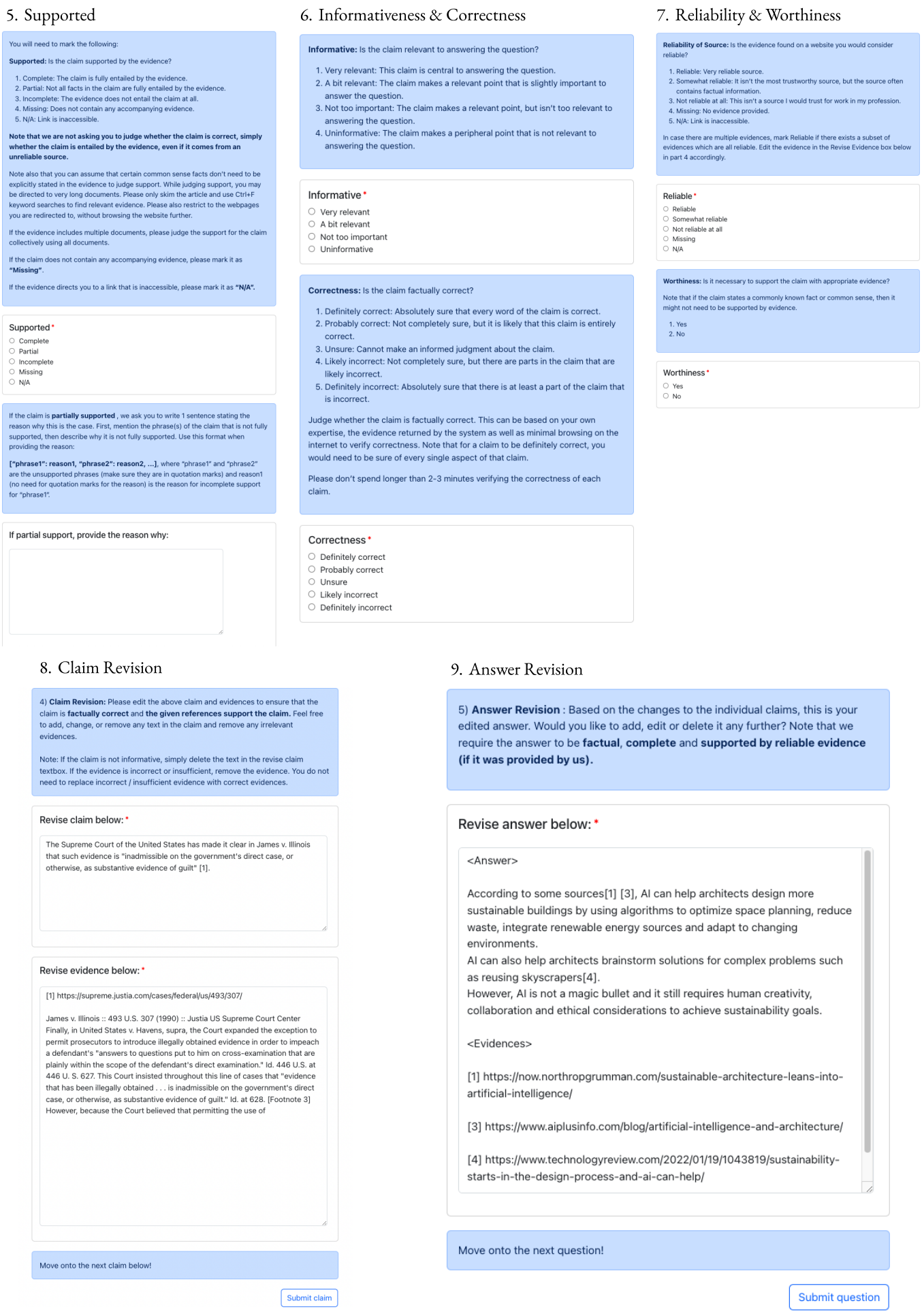

Figure 7: Screenshots of the interface (5-9).

Field	Question	Types
Anthropology	*Why is it that Africa's representation is still a problem in modern day times regardless of the academic writings that state otherwise?*	II,VII
Architecture	*Suppose an architect decides to reuse an existing foundation of a demolished building, what is to be considered to ensure success of the project?*	IV
Aviation	*Should a low value shipment take priority from a regular customer or a high value shipment from a infrequent customer?*	V
Biology	*Can you explain the mechanisms by which habitat fragmentation affects biodiversity and ecosystem functioning, and provide examples of effective strategies for mitigating these impacts?*	III,VI
Business	*If your supplier can give you a discount for a whole yearly production, how can we take this deal without affecting our budget in a critical way?*	V
Chemistry	*Why does gallic acid have an affinity with trivalent iron ions?*	I
Classical Studies	*If researchers found a new method to unroll the Herculanum papyri, would it be fair to try it on the actual papyrus, given that it could potentially destroy it?*	V
Climate Science	*If an imidazolium based ionic liquid were to be released into the environment through the aquatic compartment, what species would be affected, if any?*	II,III,V
Criminology	Mr X is an 18 year old first time offender involved in a burglary where he acted as a lookout. Which category about this information be placed under?	V
Culinary Arts	*If mezcal production in the Valley of Mexico posits the distilling of mezcal can be traced back to ancient times, how could this be attained before the arrival of the Spaniards?*	V
Economics	*Can you summarize the current economic policies and strategies of the top five global superpowers and their potential impact on the global market?*	I
Education	*Can music therapy impact a child with autism if they have noise sensory issues?*	V
Engineering and Technology	*How different will licensing a small modular reactor be as compared to licensing traditional large nuclear power plants?*	VII
Environmental Science	*Does floating solar panels minimize the risk of eutrophication or they are more trouble than their worth?*	I
Geography	*How can we overcome the limitations of remote sensing data, such as low spatial resolution and limited spectral bands?*	IV
Healthcare/Medicine	*If a 48 year old woman is found to have an esophageal carcinoma that invades the muscularis propria and has regional lymph node metastases but no distant metastasis, what is her stage of cancer and what are possible recommended treatments?*	I,III
History	*To what extent is JFK's legacy written from sympathy because of his assassination?*	II,VII
Journalism	*How many sources you must have before printing a story?*	I
Law	*Can direct evidence in a case that has been obtained illegally be considered by the court in some cases if it directly points to the defendant's guilt?*	I
Linguistics	*What are the attitudes of Received Pronunciation in the United States?*	II
Literature	*How would one go about researching the role of the mother represented in Anne Sexton's 1971 poetry volume "Transformations"?*	IV, VI
Mathematics	*Do you think there is a relation between Frobenius numbers and the Kawamata conjecture for weighted complete intersections?*	III, VII
Military or Law Enforcement	*If you get anthrax poisoning during a mission, which chemical agent should you use to neutralise the poison?*	I
Music	*What exercises would you do in a singing class with a teenager with puberphonia?*	IV
Philosophy	*How does modern neuroscience support and reject a computational theory of mind?*	III
Physics & Astronomy	*Standard Model does not contain enough CP violating phenomena in order to explain baryon asymmetry. Suppose the existence of such phenomena. Can you propose a way to experimentally observe them?*	V
Political Science	*Despite the fact that IPCC was formed in 1988, several studies have showed that argubaly more than 50% of all carbon emissions in history have been released since 1988. What does this show about IPCC and developed countries' efforts?*	VII
Psychology	*How can counselling psychologists effectively and appropriately incorporate use of self into therapy?*	III,IV,VII
Sociology	*Which factors strengthen social cohesion within societies?*	VII
Theology	*Is there any justification for the use of violence in the New Testament?*	I
Visual Arts	*Tell me the step by step process of recycling a canvas.*	III

Table 15: Examples from EXPERTQA, showing an example from every field included in the dataset.

Figure 8: The distribution of question types across all fields included in EXPERTQA.

Figure 9: Label distribution of claim properties across different fields for all systems.

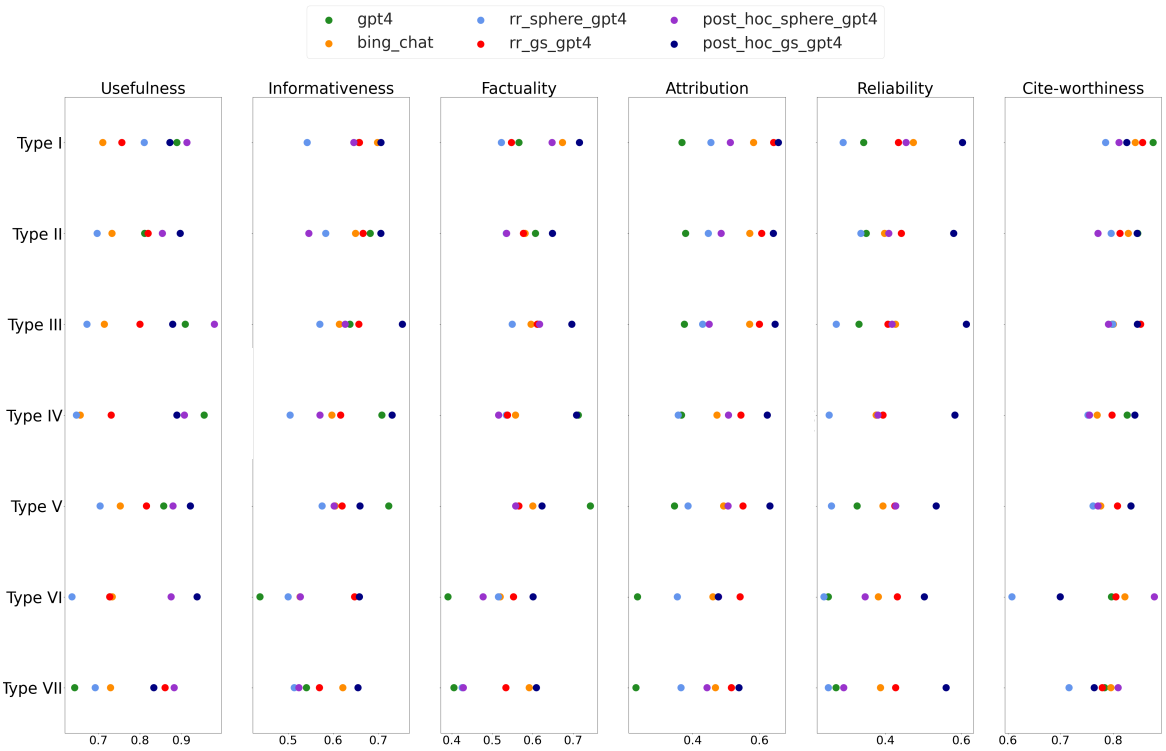

Figure 10: Label distribution of claim properties across different question types for all systems.