Fine-grained Hallucination Detection and Mitigation in Long-form Question Answering

Anonymous EMNLP submission

Abstract

Long-form question answering (LFQA) aims to provide thorough and in-depth answers to complex questions, enhancing comprehension. However, such detailed responses are prone to hallucinations and factual inconsistencies, challenging their faithful evaluation. This work introduces *HaluQuestQA*, the first hallucination dataset with localized error annotations for human-written and model-generated LFQA answers. HaluQuestQA comprises 698 QA pairs with 4.7k span-level error annotations for five different error types by expert annotators, along with preference judgments. Using our collected data, we thoroughly analyze the shortcomings of long-form answers and find that they lack comprehensiveness and provide unhelpful references. We train an automatic feedback model on this dataset that predicts error spans with incomplete information and provides associated explanations. Finally, we propose a prompt-based approach, *Error-informed refinement*, that uses signals from the learned feedback model to refine generated answers, which we show reduces hallucination and improves answer quality. Furthermore, humans find answers generated by our approach comprehensive and highly prefer them (84%) over the baseline answers.[1]

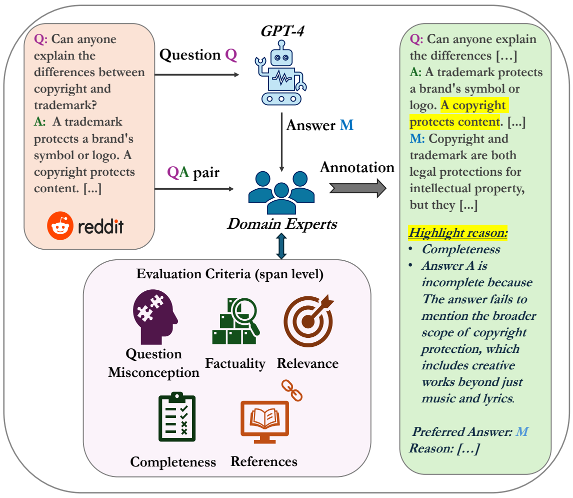

Figure 1: An overview of our data collection process. Based on our defined aspects, we collect expert human judgments for question-answer pairs on the Reddit platform and their corresponding answers from GPT-4.

1 Introduction

Long-form question answering (LFQA) provides comprehensive, user-friendly, and in-depth responses to complex questions by leveraging state-of-the-art large language models (LLMs) and retriever components (Krishna et al., 2021; Nakano et al., 2021). While LLMs generate plausible and convincing answers, they also frequently produce factually inconsistent, irrelevant, and incomplete content (Goyal and Durrett, 2020; Laban et al., 2022; Menick et al., 2022; Ji et al., 2022), which limits their applicability in real-world applications.

Simplistic evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) do not align with human experts' judgments on long-form answers (Wang et al., 2022). There are many aspects of LFQA – factuality, completeness, and relevance – that require evaluation, motivating us to focus on *span-level fine-grained* error detection. While previous studies have focussed on evaluating factual errors in long-form text generation (Lee et al., 2022; Min et al., 2023; Li et al., 2023; Muhlgay et al., 2023), other aspects of evaluation, such as response completeness and relevance – which can potentially mislead and confuse users – have been largely overlooked.

LLMs make many errors for LFQA, which require expert human annotations to detect (Gillick and Liu, 2010; Iskender et al., 2020; Wang et al., 2022). Recent work from Xu et al. (2023a) reports that aspects such as *factuality, relevance, completeness, structure, references, and accessibility* are essential for evaluating long-form answers.

[1] We make our code and data available at: github.com/lfqa-hallucination

There are no prior studies for LFQA that examine these errors at the span level. Span-level error annotation and categorization have been important for evaluating and improving systems in other generation tasks such as machine translation (Freitag et al., 2021). We fill this gap by collecting *HaluQuestQA*, a dataset of LFQA answers annotated at the span level with five different error types: *question misconception, factuality, completeness, relevance, and helpful references*; by expert annotators, in addition to preference judgments, as shown in Figure 1.

Next, we train an automatic feedback model on this dataset that predicts erroneous answer spans with incomplete information and provides associated explanations. The feedback model provides fine-grained feedback in the form of error location (sentence level), error reason, and confidence score without the aid of a reference text (Xu et al., 2023b). Finally, we propose ERROR-INFORMED REFINEMENT, a prompt-based approach that uses signals from the feedback model to refine generated answers (Madaan et al., 2023), which we show reduces hallucination and improves answer quality.

Our contributions can be summarized as follows: (1) We release *HaluQuestQA*, a dataset of span-level error annotations on pairs of human-written and model-generated answers. Our data analysis shows that long-form answers lack comprehensiveness and provide unhelpful references; (2) We train a feedback model to detect span-level errors aligned with expert human judgments; (3) We propose Error-informed refinement, an approach to refine LLM-generated answers with fine-grained feedback provided by our learned model. Our approach consistently outperforms baselines utilizing coarse-grained feedback (lacking detailed error justifications), reducing hallucinations.

2 Related Work

Human evaluation. Prior work (Krishna et al., 2021) has shown that human evaluation for LFQA tasks is challenging due to long answer lengths, and expert annotators are required to evaluate them effectively. Xu et al. (2023a) hire (non-)expert annotators and identify nine multi-faceted aspects for meaningful LFQA evaluation. While some of these fine-grained aspects, such as factuality (Goyal and Durrett, 2020; Laban et al., 2022), coherence (Goyal et al., 2022), and completeness (Tang et al., 2024), have been studied to investigate hallucinations in dialogue summarization tasks, ours is amongst the first works to study LFQA-centric properties such as *question misconception, factuality, relevance, completeness, and helpful references*, at the span-level.

Detecting and Mitigating Hallucinations in LLMs. Increasing focus on the reliability of LLMs has led to the development of explainable evaluation metrics (Zhong et al., 2022; Fu et al., 2023) to detect errors in LLM generations. Xu et al. (2023b) present InstructScore, an explainable metric based on LLaMA (Touvron et al., 2023a), to obtain detailed error analysis for LLM-generated text. However, most of the current evaluation metrics require hard-to-obtain gold references. Recent work proposes a reference-free evaluation metric, TIGERSCORE (Jiang et al., 2023b) that can locate, categorize, and explain errors across various text generation tasks, including summarization, translation, and LFQA. While LLM-based metrics can detect diverse errors, it is not always plausible to have an external evaluator during real-time inference; hence, sampling-based approaches (Chen et al., 2023; Manakul et al., 2023; Malon and Zhu, 2024) have been proposed, wherein consistency across multiple sampled model outputs is used as a measure of factuality.

Reinforcement learning with human feedback (RLHF) (Ziegler et al., 2019), a framework to incorporate human feedback to align LMs, has been used to reduce undesirable LLM generations (Ouyang et al., 2022; Bai et al., 2022a,b). Wu et al. (2023b) propose fine-grained RLHF, a framework that enables learning reward models associated with span-level human feedback on different error types. However, training multiple reward models is complex and compute-intensive. A recent alignment technique, direct preference optimization (DPO) (Rafailov et al., 2023) bypasses the reward modeling step in RLHF and has been used to fine-tune LMs for factuality using preference ranking over model responses (Tian et al., 2023). Human feedback has also been used to train feedback models (Wang et al., 2023; Xu et al., 2024) to guide the refinement of LLM outputs (Madaan et al., 2023; Welleck et al., 2023), improving answer quality. However, these feedback models are either not trained to provide fine-grained error feedback or rely on the ground truth passage to detect errors, which may not always be accessible for open-domain QA tasks. Our work aims to annotate

Category (# samples)	Preference		Krippendorff's α
	Human	Model	
Physics (94)	33%	67%	0.01
Chemistry (96)	22%	78%	0.20
Biology (110)	25%	75%	0.36
Technology (110)	16%	84%	0.53
Economics (110)	14%	86%	0.31
History (92)	9%	91%	0.52
Law (86)	16%	84%	0.59
Average	19.29%	80.71%	0.36

Table 1: Overview of HaluQuestQA and expert answer preferences, with experts' agreement on a smaller subset (~ 15%) calculated using Krippendorff's alpha.

fine-grained errors in LFQA, using this data to train a reference-free feedback model for sentence-level error detection with justifications. We further propose a prompt-based approach to refine answers with feedback, enhancing their comprehensiveness.

3 HaluQuestQA (HQ^{2}A)

Prior LFQA evaluations with non-expert (Nakano et al., 2021) and expert (Xu et al., 2023a) annotators collect preference judgments over model responses. However, overall preference is not indicative of fine-grained errors in LFQA. As a first step, we annotate span-level errors in long-form answers, with explanations from domain experts.

3.1 Hiring Annotators

We recruit domain experts on Prolific's academic annotation platform for seven domains shown in Table 1. The expert selection is based on age (22-32), demographics (US and UK), education (undergraduate or graduate degree in the target domain), and native language (English). For each target domain, we first conduct a small pilot comprising ten samples, where given a question and two candidate answers, the experts evaluate the answers and mark the incorrect spans based on our defined evaluation criteria (§3.2). Based on the pilot results, we choose three experts per domain and give them each a large-scale study containing 35-50 question-answer pairs. We collect expert judgments for 698 questions.

3.2 Task Setup

We evaluate two answers (human and model-generated) to the same question. This setting enables us to identify errors made by humans and state-of-the-art LFQA systems. We chose GPT-4 (`gpt-4-0314`) as the LFQA model to evaluate since previous work (Bhat et al., 2023) has shown it to outperform existing open-source LLMs (LLaMA and Alpaca (Taori et al., 2023)) in reasoning and inferring from long context. Since this model has likely seen training data up to September 2021, it may have already seen the ELI5 dataset released by Fan et al. (2019) during its pre-training. Thus, we scrape more recent questions from the *r/explainlikeimfive* subreddits posted between November 2022 to March 2023. The questions on the ELI5 are classified into domains via the FLAIR label (tag containing post information), which lets us perform domain-specific analysis. For unclassified categories (like History and Law), we cluster the OTHER category questions (not in pre-defined ELI5 domains), using K-means clustering (Selim and Ismail, 1984) and identify the domain-specific questions. For each domain, we sample between 100-200 questions with their highest voted answer of length ranging between 50-500 words (more details in Appendix A).

To obtain the model-generated answers, we zero-shot prompt the GPT-4 model (Appendix B.1). We host the annotation task on the INCEpTION platform (Klie et al., 2018) and evaluate the following:[2]

1. **Question misconception:** False assumptions made within the given question.
2. **Factuality:** Accuracy and correctness of the answer as per verifiable facts.
3. **Relevance:** Specificity and meaningfulness of the answer.
4. **Completeness:** Answer comprehensiveness ensuring all question aspects are addressed.
5. **References:** (Un)helpful examples, analogies, and external references (websites or links) in the answer.

Based on the defined evaluation criteria, annotators identify and highlight mistakes in the question or answers with free-form justifications, in addition to overall answer preference. Examples from HaluQuestQA are shown in Table 2.

3.3 Quantitative Analysis

Experts prefer GPT-4 answers over human answers. As shown in Table 1, experts display a

[2]We provide detailed annotation guidelines in Appendix F.

Aspect	Question	Answer A	Answer B	Expert Justification
Question Misconception	How do people in prison get so jacked with terrible nutrition and no supplements?	*(human)* [...] While prison food generally doesn't taste great, it does have calories and protein. That can also be supplemented with the stuff they can buy on commissary (fish packs, chicken packs, cheese, [...]	*(model)* [...] In prison, inmates often have a lot of free time, which they can use to work out and develop a consistent exercise routine. They may not have access to a fully-equipped gym, but [...]	This is an assumption, not every prison offers terrible food without nutritional value, also lots of "easy" food choices (lentils, eggs, bread, meat,..) offer high amounts of protein which is needed for weight gain.
Factuality	Why can you sleep through the night without going to the bathroom?	*(human)* Your body produces a hormone called (vasopressin) antidiuretic hormone, or adh, which causes the body to produce less urine at night. It's the reason [...]	*(model)* During sleep, your body produces less urine [...] Additionally, the hormone vasopressin helps to reduce urine production at night. This combination of decreased [...]	The urine production is decreased due to the hormone, not additionally. The more concentrated urine is because you cannot drink while sleeping.
Relevance	Why is it taboo to talk about people's wages?	*(human)* Because companies used to be able to fire people [...] discussing your wages is a federally protected right now in the USA, so by all means discuss discuss discuss! [...]	*(model)* Talking about wages is considered taboo because it can create tension and jealousy among coworkers [...]. Employers often discourage discussions to prevent potential [...]	This does not contribute to the answer.
Completeness	Can anyone explain the differences between copyright and trademark?	*(human)* A trademark protects a brand's symbol or logo. A copyright protects content. So the ac/dc logo with the lightning bolt would be trademarked. [...]	*(model)* [...] Copyright protects creative works like books, music, and art, [...]. Trademark, on the other hand, protects brand identifiers like logos, [...]	The answer fails to mention the broader scope of copyright protection, which includes creative works beyond just music and lyrics.
References	How does software that can detect text generated by ChatGPT work?	*(human)* [...] output has a high degree of regularity [...] It's like comparing a handknit scarf to an industrial machine's knit scarf: one might have small mistakes that show it was made by a human hand, whereas the other is almost perfect because the machine follows an exact pattern.	*(model)* ChatGPT detection software works by analyzing textual patterns and language use to determine if the text was generated by an AI like ChatGPT. Essentially, these programs compare the input text with known characteristics of AI-generated text [...]	This example is well suited for the explanation in the paragraph. It gives a clear representation of how the software detects ChatGPT text in my head.

Table 2: Examples of expert annotated errors in long-form answers based on the defined evaluation criteria.

high preference (80.7%) for *model-generated* answers from GPT-4 compared to human answers. Potentially, humans prefer fluent answers, and LLMs are known to optimize for fluency (Wu et al., 2023a; Coyne and Sakaguchi, 2023). Moreover, the preference of our annotators is corroborated by similar findings in summarization (Liu et al., 2023) and LFQA (Xu et al., 2023a), who show that GPT-3 answers score higher than human answers.

Science questions are challenging for LLMs. Model-generated answers are strongly preferred by experts in history, law, technology, and economics (>80%). In contrast, the science domains are more challenging, with a preference for model answers ranging between 60%-80%.

Expert (dis)agreement. In Table 1, we report Krippendorf's alpha (Hayes and Krippendorff, 2007) as a measure of agreement for experts' overall answer preference. Our expert annotators achieve moderate agreement in technology, history, and law, fair agreement in biology and economics, and slight agreement in physics and chemistry.[3] We emphasize that the disagreement between experts is *not* a failure of our evaluation. Instead, it highlights the challenges of identifying fine-grained errors in answers, affecting overall preference. Moreover, prior work has similar findings for human disagreement in LFQA evaluation (Xu et al., 2023a).

[3]Interpretation of agreement follows Wong et al. (2021)

Answer scoring. We score human and model answers on our defined evaluation criteria to understand how experts' answer preferences diverge across different domains. For each of *question misconception* and *reference* aspects, the score $\mathcal{S} = 1$ when the question has no misconceptions and the references, if provided, help answer the question; otherwise, $\mathcal{S} = 0$. For aspects of *factuality*, *relevance*, and *completeness*, we calculate $\mathcal{S}$ as:

$$\mathcal{S} = 1 - \left(\frac{\text{\# Error sentences}}{\text{Total \# of sentences}} \right)$$

For calculating the overall answer scores, we leave out the question misconception scores because this aspect pertains to the question. We sum the other aspect scores and include the overall answer preference scores ($\mathcal{S} = 1$ if preferred) to get the final score. Finally, we normalize this score between 0 and 1. In Figure 2, we report the fine-grained aspect scores for human and model answers across different domains and discuss our findings below.

Questions from technology and economics are biased. Ambiguous and misinformed questions

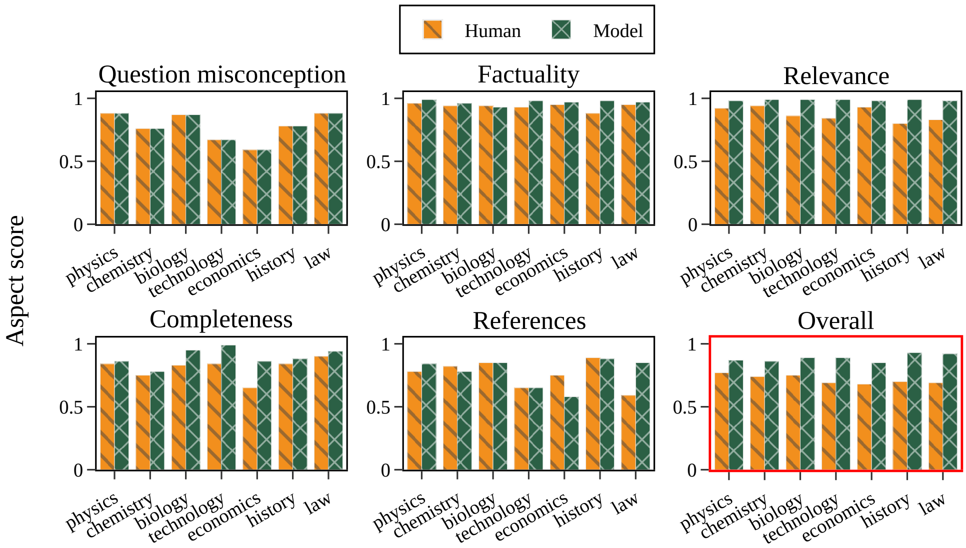

Figure 2: Comparison of fine-grained scores of the human-written and model-generated answers for different evaluation criteria. The last figure (with red boundary) shows the averaged and normalized overall scores. A higher score represents fewer errors in the answers.

can lead to undesirable answers (Cole et al., 2023; Kim et al., 2023). Therefore, fair answer scoring requires prior estimation of question quality. For this, we utilize the question misconception aspect and find that questions from all evaluated domains consist of misconceptions arising from the user's bias or misinformation. This is especially prominent in technology and economics, where $\sim 40\%$ of the questions are misinformed – users have low domain knowledge to ask the right questions.

Answers lack comprehensiveness and provide unhelpful references. We observe that human-written and model-generated answers score high on *factuality* and *relevance* aspects, meaning most of the information provided in the answers is verifiable, trustworthy, and related to the question. Interestingly, the answers score low on the *completeness* and *references* aspects, lacking important information and providing web references and examples that are not useful, as per the experts' judgments. Specifically, models hallucinate and provide incorrect or made-up web links. In contrast, human answers digress from the topic, providing irrelevant information that leads to undesirable conclusions.

Overall, model answers score better than the human answers in all the evaluated domains. While this is due to their better performance over humans on the considered aspects, we believe that the persuasive nature of model answers (Salvi et al., 2024) also plays a crucial role in their higher preference.

4 Hallucination Mitigation

In §3.3, we have shown that LFQA answers lack comprehensiveness and omit helpful information. Therefore, we train a feedback model to identify erroneous answer spans with *incomplete information* and provide free-form error justifications. Our approach, ERROR-INFORMED REFINEMENT, uses this feedback to refine answers and improve their overall quality without human intervention.

4.1 Error Feedback Model

Given an input question and an LFQA response, the feedback model generates a label *[Complete]* or *[Incomplete]* for every sentence $1 \ldots n$ in the response and gives associated reasons for the incomplete sentences (see Figure 3). We model this as a sequence-to-sequence task and finetune a LLaMA2-13B model (Touvron et al., 2023b).

Fine-tuning. Training the feedback model requires high-quality error annotations with justifications. To this end, we utilize our HQ^2A dataset and extract QA pairs with errors in the completeness aspect. For every extracted sample, we segment the answer into sentences and mark every sentence with the *[Complete]* or *[Incomplete]* tag along with the expert's justifications. The final dataset consists of 509 samples split into train (90%) and test (10%) sets. We train the model with batch size 4, learning rate $2e-5$, and sequence length 1024 for 5 epochs. We list the prompts used in Appendix B.2.

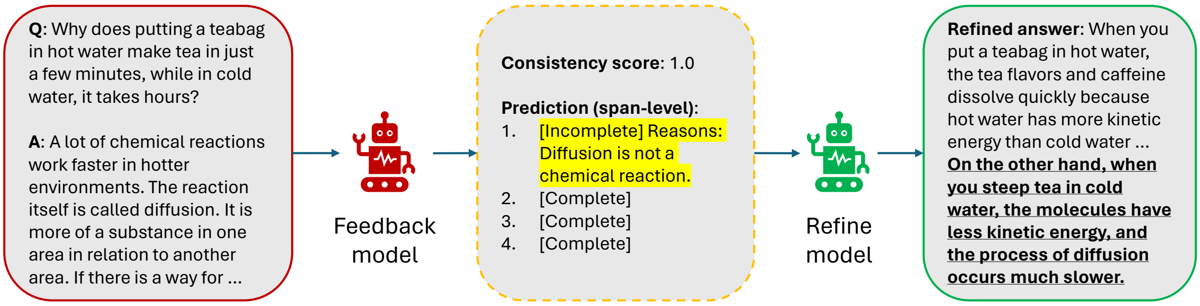

Figure 3: A pictorial view of our Error-informed refinement approach. The feedback model takes as input a question-answer pair and outputs span level error with justifications and a consistency score. The refine model uses this feedback to improve the original answer.

Inference. The trained feedback model hallucinates web references in about 20% of test samples. This likely occurs because the training data includes web references in expert error justifications, which the model struggles to replicate coherently. To combat this, we opt for a sampling-based approach (Malon and Zhu, 2024) to provide more consistent feedback. The intuition is that trustworthy details and references should appear in many other generated samples. Hence, during the decoding step, we use nucleus sampling (Holtzman et al., 2020) with p=0.9 and sample 20 responses from the feedback model and check their consistency in two stages: 1) TAG CONSISTENCY: This pertains to the consistency of span-level tag predictions, *complete* or *incomplete*, for each sampled response. The tag consistency score is calculated by counting the number of other sampled responses that match the tag sequence of each sampled output and averaging over the total number of samples. Formally, if the sampled tag predictions $p_1, ..., p_n$ consist of tag sequences $t_1, ..., t_n$ where t_i is a list of tag predictions for every span, the score for sample i is

$$\mathcal{S}_{\mathcal{TC}} = \frac{1}{n} \sum_{s=1}^{n} 1_{t_i = t_s} \tag{1}$$

where $1_{t_i=t_s}$ is 1 if the tag sequence t_i is the same as tag sequence t_s and 0 if not. The samples with the highest score are selected for the next stage. 2) REASON CONSISTENCY: We assess the consistency of justifications given for the incomplete spans from the remaining samples. Specifically, we count the number of other sampled justifications from the LLM that matched each token of each sampled output and score each justification by the average count per token. Formally, if the sampled justifications $j_1, ..., j_n$ consist of words $w_i^k, k = 1...m_i$, the score of sample i is

$$\mathcal{S}_{\mathcal{RC}} = \frac{1}{m_i} \sum_{k=1}^{m_i} \sum_{s=1}^{n} 1_{w_i^k \in j_s} \tag{2}$$

where $1_{w_i^k \in j_s}$ is 1 if token w_i^k is in the justification j_s and 0 if not. Finally, we select the sample output with the highest score as the feedback for the refinement model. After sampling, we notice a 50% reduction in reference hallucinations, down to $\sim 5-10\%$ test set samples.

4.2 Error-Informed Refinement (EIR)

Our approach is shown in Figure 3 and consists of two main components: an error feedback model (§4.1), and a refinement model. Given an input prompt x_i and a corresponding human-written or model-generated response y_i, the feedback model $\mathcal{E}$ generates a targeted feedback f_i that represents the quality of y_i in free-form natural language. Finally, the refinement model uses x_i, y_i, and f_i, generating a refined and improved output response $\hat{y}_i$. The following sections describe our approach in more detail.

Refinement Model. Our experiments use the LLaMA2-13B chat LLM and its DPO optimized version (see Appendix C) as the refinement models. In each case, the model is 0-shot prompted with the fine-grained error feedback received from the error detection model. We also experiment with two strong baseline feedback models, 1) IMPROVE: The refinement model is 0-shot prompted to improve the answer without any feedback provided. 2) GENERIC: The refinement model is 0-shot prompted to improve the answer with a generic error feedback that asks the model to provide a more complete and accurate answer. We list the prompts used in Appendix B.3.

Datasets & Evaluation Metrics. We test our error-informed refinement approach on three datasets: HQ^2A with span-level error annotations for answer completeness, ASQA (Stelmakh et al., 2022), and ELI5 (Fan et al., 2019). The ASQA dataset consists of 6K ambiguous factoid questions with long-form answers synthesized from multiple sources to resolve the ambiguities. ELI5 consists of 270K long-form answers covering general topics from the subreddits "explainlikeimfive", "askscience", and "AskHistorians" on Reddit.

We evaluate the refined answers using TigerScore, a trained reference-free metric that identifies errors in LLM-generated text and assigns a score based on error severity. Specifically, we use the LLaMA-7B trained version of TigerScore, which highly correlates with humans for error detection in LFQA tasks (Jiang et al., 2023b) while being much less expensive than human evaluation. Furthermore, we evaluate the error correction capabilities of our refinement approach using precision, recall, and F1. Lastly, we conduct a human evaluation to evaluate the comprehensiveness and preference of the refined answers compared to gold answers.

5 Results

We explore several research questions: 1) Can our learned feedback model detect errors in LFQA systems and help in downstream answer refinement task? 2) Does fine-grained feedback produce better quality LFQA answers than coarse-grained feedback? 3) Does fine-grained feedback help mitigate hallucinations and improve the comprehensiveness of LFQA answers? 4) Are comprehensive answers from our approach preferred by humans?

5.1 Detecting Errors via Feedback Model

Since detecting erroneous spans in long-form answers is hard, we measure the accuracy of our feedback model in three different settings; model-detected erroneous spans are entirely different (DIFFERENT), adjacent (ADJACENT), and exactly similar (EXACT) to the human-annotated spans. In Table 3, we show the sentence-level error detection accuracy of the feedback model as compared to the strong human baseline. The feedback model detects the *exact* and *adjacent* error spans with a combined accuracy of 61%. However, it is important to note that the model gives high consistency scores when confident in its predictions. A consistency score less than 0.80 means that the model is unsure in its error prediction feedback, while a score above 0.85 shows that the prediction highly aligns with humans.

Dataset	Error span	Accuracy (↑)	Consistency Score (↑)
	Different	**38.56** ± 0.93 %	0.71 ± 0.02
HQ^2A	Adjacent	24.18 ± 0.92%	0.82 ± 0.01
	Exact	37.25 ± 0.00 %	**0.86** ± 0.01

Table 3: Accuracy of our feedback model in detecting sentence-level errors compared to the expert error annotations. The feedback model predictions closely align with humans at consistency scores above 0.80.

We further evaluate our error feedback model by comparing the gap in the downstream LFQA refinement task when we use human-annotated error feedback. This evaluation measures the effectiveness of our feedback model in guiding the refinement of long-form answers and reducing hallucinations. In Table 4, we present the refinement performance of our feedback model as compared to the expert human feedback on HQ^2A. We find that our feedback model's performance is very competitive, reducing hallucinated samples by 2% and improving F1 score by 4% compared to the expert human feedback. This result validates the effectiveness of our feedback model in refining LFQA answers.

5.2 Fine- vs. Coarse-grained Feedback

Table 4 shows the quality of answers refined using different forms of feedback plus the baseline quality of answers from the datasets. We observe that inadequate feedback deteriorates the quality of generation. While directly prompting the refinement model (IMPROVE) performs better than the baseline, prompting with more targeted feedback (GENERIC) consistently outperforms the IMPROVE approach and generates better quality LFQA answers. This highlights the importance of providing detailed feedback to the refinement model.

In contrast, providing fine-grained feedback from our error detection model (EIR) outperforms coarse-grained feedback and even fine-grained human feedback (on HQ^2A), delivering consistent improvements in reducing hallucinated samples and hallucination scores by $\sim 3\%$ and $\Delta \sim 38\%$, respectively, and improving F1 scores by $\sim 5\%$ over all the evaluated datasets. Using our DPO-aligned refinement model does not reduce the hallucinated samples. However, it achieves the best hallucination score on ASQA and ELI5, showing that opti-

Dataset	Approach	Tigerscore		Error Correction		
		% Hallucinated samples (↓)	Hallucination score (↓)	Precision (↑)	Recall (↑)	F1 (↑)
HQ^2A	Human feedback	2.61 ± 0.92	0.09 ± 0.01	0.86 ± 0.04	**1.00** ± 0.00	0.94 ± 0.02
	Baseline	19.61	0.63	-	-	-
	Improve	1.31 ± 0.92	0.05 ± 0.04	**1.00** ± 0.00	0.93 ± 0.05	0.97 ± 0.02
	Generic	1.31 ± 0.92	0.05 ± 0.03	0.97 ± 0.04	0.97 ± 0.05	0.97 ± 0.02
	EIR	**0.65** ± 0.92	**0.03** ± 0.04	0.97 ± 0.04	**1.00** ± 0.00	**0.98** ± 0.02
	EIR w/ DPO	4.57 ± 2.44	0.07 ± 0.02	0.90 ± 0.08	0.87 ± 0.05	0.88 ± 0.06
ASQA	Baseline	34.81	1.20	-	-	-
	Improve	20.85 ± 1.00	0.68 ± 0.03	0.70 ± 0.02	0.71 ± 0.01	0.70 ± 0.01
	Generic	18.67 ± 0.52	0.61 ± 0.01	0.72 ± 0.01	0.75 ± 0.01	0.74 ± 0.00
	EIR	**16.63** ± 0.41	0.51 ± 0.02	**0.73** ± 0.00	**0.82** ± 0.02	**0.77** ± 0.01
	EIR w/ DPO	22.61 ± 0.26	**0.45** ± 0.01	0.64 ± 0.00	0.77 ± 0.01	0.71 ± 0.00
ELI5	Baseline	22.93	0.82	-	-	-
	Improve	10.05 ± 0.18	0.36 ± 0.02	0.75 ± 0.00	0.86 ± 0.00	0.80 ± 0.00
	Generic	6.06 ± 0.23	0.22 ± 0.01	0.84 ± 0.01	0.91 ± 0.00	0.87 ± 0.00
	EIR	**3.81** ± 0.30	**0.13** ± 0.01	**0.88** ± 0.01	**0.96** ± 0.01	**0.92** ± 0.01
	EIR w/ DPO	5.71 ± 0.25	**0.13** ± 0.00	0.83 ± 0.00	0.94 ± 0.01	0.88 ± 0.00

Table 4: Results of the quality of answers refined through coarse-grained and fine-grained feedback. We include two baselines using coarse-grained feedback: IMPROVE and GENERIC for all the datasets. Additionally, we include the results for expert human feedback on our collected test set.

mization helps correct major errors in the answers. We show further evidence of the role of alignment in reducing hallucinations in Appendix E.1.

5.3 Human Evaluation

To test the comprehensiveness and overall quality of the answers generated using our refinement approach, we hire three annotators and perform a human evaluation on a subset of 50 samples each from HQ^2A, ASQA, and ELI5 datasets.

Table 5 shows the results of our human evaluation of the original and refined answers. Annotators find the answers produced by our approach comprehensive, meaning all the questions are answered thoroughly without omitting important information. However, a comprehensive answer does not necessarily mean a better answer. Therefore, we also evaluate the overall preference of our answers, incorporating factors such as factuality and relevance compared to the baseline answers. We observe that annotators significantly prefer the refined answers ($\sim 84\%$) across all the datasets, indicating their factual correctness and relevance. We provide details on the human agreement in Appendix E.2.

Dataset	App.	Comprehensiveness(↑)	Preference(↑)
HQ^2A	Baseline	0.00%	7.84 %
	Refined	**100** %	**92.16** %
ASQA	Baseline	82.00 %	40.00 %
	Refined	**100** %	**60.00** %
ELI5	Baseline	38.00 %	0.00 %
	Refined	**100** %	**100** %

Table 5: Human evaluation results on the comprehensiveness and preference of refined answers over the baseline answers from three datasets.

6 Conclusion

In this work, we introduce HALUQUESTQA, a dataset of expert human judgments on fine-grained errors (question misconception, factuality, relevance, completeness, and references) in LFQA. Using our dataset, we analyze the pitfalls of human and model long-form answers, identifying issues with comprehensiveness and unhelpful references. To address these, we propose ERROR-INFORMED REFINEMENT, an approach that uses signals from our learned feedback model to refine LLM responses. Our feedback model outperforms baseline feedback models and expert human feedback in guiding answer refinement and reducing hallucinations. A human evaluation confirms the effectiveness of our approach, with participants finding our refined answers more comprehensive and preferable to baseline outputs.

Limitations

Despite providing an in-depth analysis on hallucinations in human and model generated responses, our work only focusses on the LFQA task. Thus, we encourage future work to apply our findings to different tasks such as summarization, translation, etc. We study a diverse but limited scope of long-form answers drawn from online community platforms. More diverse questions from different domains such as education or commercial may have different issues and might be to be evaluated in a different way.

Our trained error detection model shows high correlation with human annotations but relies on a high consistency of model outputs. The model may hallucinate if the consistency score is low (< 0.80). Training larger models with more high quality data might be an interesting future work to get better results. Lastly, in our refinement approach, we only experiment with the intstruction-tuned variant of LLaMA2. Models with better or worse instruction following capabilities may give different results and improving the refinement process can be a great future direction to mitigate hallucinations.

Ethics and Broader Impact Statement

The expert annotation data collection protocol has been determined to be exempt from review by an IRB board. All the collected data will be publicly available under the CC BY-SA 4.0 license. We hire annotators on the academic annotation platform Prolific and gather no sensitive user information except demographics and annotator performance data. We examined the collected data and ascertained that it contains no toxic or harmful content.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Meghana Moorthy Bhat, Rui Meng, Ye Liu, Yingbo Zhou, and Semih Yavuz. 2023. Investigating answerability of llms for long-form question answering. *CoRR*, abs/2309.08210.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *CoRR*, abs/2311.17311.

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 530–543. Association for Computational Linguistics.

Steven Coyne and Keisuke Sakaguchi. 2023. An analysis of gpt-3's performance in grammatical error correction. *CoRR*, abs/2303.14342.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *CoRR*, abs/2104.14478.

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

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. *CoRR*, abs/2010.05478.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. SNaC: Coherence error detection for narrative summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629.

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. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023b. Tigerscore: Towards building explainable metric for all text generation tasks.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 996–1009. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *CoRR*, abs/2305.11747.

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.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4140–4170. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in*

Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Christopher Malon and Xiaodan Zhu. 2024. Self-consistent decoding for more factual open responses. *CoRR*, abs/2403.00696.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin J. Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *CoRR*, abs/2203.11147.

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. *CoRR*, abs/2305.14251.

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. *CoRR*, abs/2307.06908.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

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*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. On the conversational persuasiveness of large language models: A randomized controlled trial. *CoRR*, abs/2403.14380.

Shokri Z. Selim and M. A. Ismail. 1984. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(1):81–87.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wing mei Wong, Jon Burnsky, Jake W. Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization.

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.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *CoRR*, abs/2311.08401.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan

Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. Modeling exemplification in long-form question answering via retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2079–2092, Seattle, United States. Association for Computational Linguistics.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Shepherd: A critic for language model generation. *CoRR*, abs/2308.04592.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ka Wong, Praveen K. Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7053–7065. Association for Computational Linguistics.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael R. Lyu. 2023a. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *CoRR*, abs/2303.13648.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023b. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023a. 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.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. Llmrefine: Pinpointing and refining large language models via fine-grained actionable feedback.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023b. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5967–5994. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2023–2038. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A Data Analysis

This section presents additional insights on our HaluQuestQA (HQ^{2}A) dataset.

A.1 Answer Length Distribution

Figure 4 compares the length distribution of human-written and model-generated answers. We observe that the length of human and model answers is comparable, resulting in a fair evaluation. Across all domains, the length of collected answers ranges between 50-500 words with an average length of 100 words.

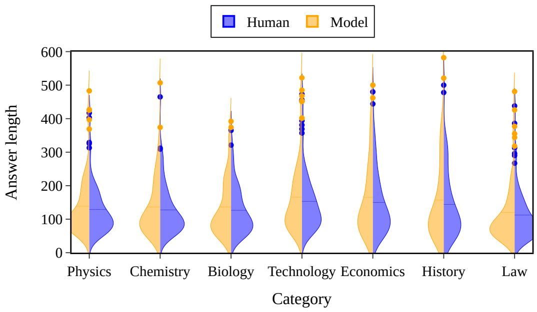

Figure 4: Answer length distribution of human-written and model-generated answers (H/M) in our expert-annotated dataset.

A.2 Overall Answer Preference

In Figure 5, we plot the word frequency distribution of the free-form answer justifications provided by our expert annotators. Apart from our considered evaluation aspects, we observe that the annotators also find answers *clarity*, *conciseness*, and *ease of understanding* helpful in deciding the overall best answer. We encourage future LFQA research to consider these aspects in their evaluation.

B Prompts

This section lists the prompts for data collection, training the error detection model, and refining answers using our Error-informed approach.

B.1 Data Collection

We prompt GPT-4 in a zero-shot manner to generate responses to questions asked on the Reddit platform, as shown in Listing 1. We use the default generation parameters in OpenAI API with temperature=0.1 and max_tokens=1.5*(human_answer_length). We specifically instruct the model to generate a response of length similar to the corresponding human response on Reddit to compare model-generated and human-written answers fairly on our defined evaluation criteria.

```
f"""Your task is to answer a question
↪ by providing a clear and concise
↪ explanation of a complex concept in
↪ a way that is accessible for
↪ laypeople. The question was posted
↪ on the Reddit forum Explain Like
↪ I'm Five (r/explainlikeimfive).
↪ Please keep in mind that the
↪ question is not literally meant for
↪ 5-year-olds, so you should not
↪ answer the question in a way that
↪ you are talking to a child. Your
↪ answer should be around
↪ {human_answer_length} words and
↪ should break down the concept into
↪ understandable parts, providing
↪ relevant examples or analogies
↪ where appropriate. You should also
↪ aim to make your explanation easy
↪ to follow, using clear and concise
↪ language throughout. Your answer
↪ should maintain accuracy and
↪ clarity. When appropriate, you can
↪ start with one sentence summarizing
↪ the main idea of the answer.

Question: {question}

Answer (around {human_answer_length}
↪ words):
"""
```

Listing 1: Zero-shot prompt for GPT-4 to generate long-form answers to questions asked on the ELI5 subreddit on the reddit platform.

B.2 Feedback Model

We use expert error annotations for the *completeness* aspect from our HQ^{2}A dataset to train our feedback model. In Listing 2, we show an example prompt used to train our feedback model. Given an instruction and input question-answer, the output is a sentence-level prediction of answer completeness with detailed justifications.

B.3 Refinement Model

As detailed in §4.2, the refinement model uses coarse-grained feedback (IMPROVE and GENERIC) and fine-grained feedback from the learned error detection model to refine input answers. We list the prompts used for IMPROVE, GENERIC and incorporating fine-grained feedback in Listing 3, Listing 4 and Listing 5, respectively.

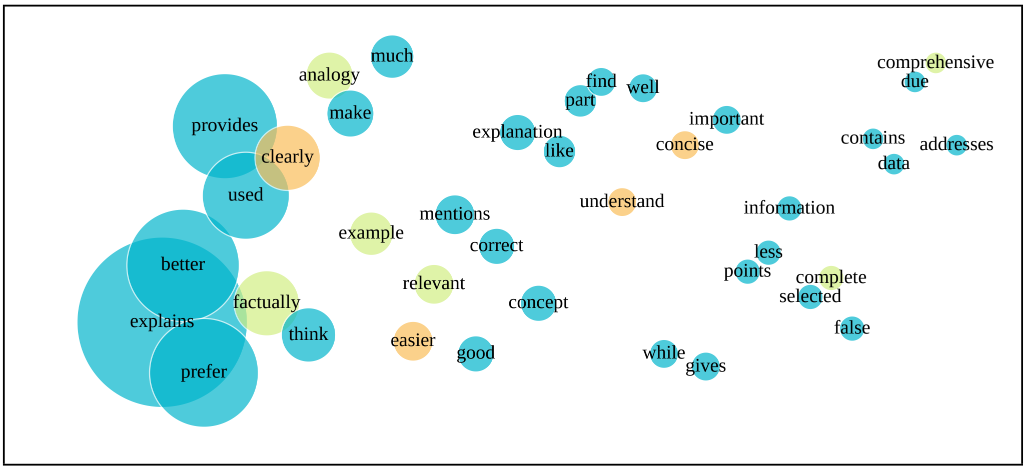

Figure 5: Distribution of the top 50 most common words mentioned by our expert annotators in their overall answer justifications. The size and color of the bubble represent the word frequency and importance, respectively. The green and orange colors denote the important evaluated and non-evaluated aspects, respectively, while blue depicts the generic terms used in answer justifications.

C Mitigating Hallucinations with Preference Optimization

While language models acquire large amounts of world knowledge and strong reasoning skills from unsupervised training over massive web corpora, aligning them with human expectations is often hard. Model alignment techniques like DPO allow us to directly use preference data to optimize the language model by casting the RL-based objective used by existing RLHF methods to an objective that can be directly optimized via a simple binary cross-entropy loss. This simplifies the process of refining LLMs greatly. The following paragraphs detail how we use DPO to reduce LLM hallucinations.

Implementation details. We model data from HQ^2A as a preference dataset where every question has a chosen and a rejected response selected by expert annotators based on the given evaluation criteria. Using this dataset, we fine-tune the LLaMA2-7B-chat (Touvron et al., 2023b) and Mistral-7B-instruct-v0.1 (Jiang et al., 2023a) models with the DPO algorithm. We use $batch\_size = 16$, $warmup\_ratio = 0.1$, $learning\_rate = 2e-5$, $num\_epochs = 5$, $beta = 0.1$, and $max\_length = 1024$ for training the models.

Due to compute limitations, we train Llama2-13B-chat model on our preference dataset using LoRA (Hu et al., 2022). We use the following training parameters: $r = 256$, $alpha = 128$, $lora\_dropout = 0.05$, $learning\_rate = 5e-5$, $beta = 0.1$, $max\_length = 1024$ and train the model for 5 epochs.

Datasets & Evaluation Metrics. We experiment with three datasets: HQ^2A, ASQA (Stelmakh et al., 2022), and ELI5 (Fan et al., 2019). HQ^2A dataset consists of 698 high-quality long-form question-answer pairs split into train (80%), dev (10%), and test (10%) sets. The ASQA dataset consists of 6K ambiguous factoid questions with long-form answers synthesized from multiple sources to resolve the ambiguities. ELI5 consists of 270K long-form answers covering general topics from the subreddits "explainlikeimfive", "askscience", and "AskHistorians" on the Reddit platform.

We report the quality of the generated long-form answers using TigerScore (Jiang et al., 2023b), a trained reference-free evaluation metric to pinpoint mistakes in the LLM-generated text. TigerScore detects hallucinations in the input text and assigns a hallucination score based on the severity of the error detected. Specifically, we use the LLaMA-7B trained version of TigerScore, which highly correlates with humans for error detection in LFQA tasks (Jiang et al., 2023b). We also measure the factual correctness of the generated answers using sample-based consistency metrics (**?**). Following their approach, we zero-shot prompt a LLaMA-13B-chat model to check if i^{th} sentence in the original answer is supported by the sampled answer S^n and return a score x_i^n using the mapping: {"Yes: 1.0", "No: 0.0", "N/A: 0.5"}. The final consistency score is then calculated as:

```
f"""### Instruction:
When given a question and answer
↪ statements, evaluate whether each
↪ given statement provides sufficient
↪ information for answering the
↪ question.
Use the '[Incomplete]' tag to indicate
↪ answer incompleteness, and
↪ '[Complete]' tag to indicate
↪ completeness, with reasons.
Please note that the answer can have
↪ single, multiple or no incomplete
↪ statements.

### Input:
Question: Can anyone explain the
↪ differences between copyright and
↪ trademark?
Answer: 1. A trademark protects a
↪ brand's symbol or logo.
2. A copyright protects content.
3. So the ac/dc logo with the lightning
↪ bolt would be trademarked.
4. The music and lyrics to
↪ thunderstruck would be copyrighted.
5. Edit: eli10 addendum: just to be
↪ clear, the content of a copyright
↪ can also be an image.
6. So the album cover to
↪ thunderstruck's album, razor's
↪ edge, would be copyrighted because
↪ it is artistic content owned by
↪ someone, but doesn't identify ac/dc
↪ as a whole.

### Response: 1. [Complete]
2. [Incomplete] Reasons: The answer
↪ fails to mention the broader scope
↪ of copyright protection, which
↪ includes creative works beyond just
↪ music and lyrics.
3. [Complete]
4. [Complete]
5. [Complete]
6. [Complete]
"""
```

Listing 2: An example prompt used for training LLaMA2-13B model for error feedback.

```
f"""
Answer the following question:
↪ "{question}"
Your answer is: "{answer}".
Please improve your answer.
Your improved answer:

"""
```

Listing 3: Zero-shot prompt for LLaMA2-13B-chat model to refine long-form answers without feedback from the error detection model (IMPROVE).

```
f"""
Answer the following question:
↪ "{question}"
Your answer is: "{answer}".
The answer is not complete.
Please improve your answer.
Your improved answer:

"""
```

Listing 4: Zero-shot prompt for LLaMA2-13B-chat model to refine long-form answers with generic feedback (GENERIC).

```
f"""
Answer the following question:
↪ "{question}"
Your answer is: "{answer}".
The answer is not complete because:
"{reason}".
Please improve your answer.
Your improved answer:

"""

# reasons are given as:
# 1. Reason 1
# 2. Reason 2
# ...
```

Listing 5: Zero-shot prompt for LLaMA2-13B-chat model to refine long-form answers with error feedback from the error detection model.

$$S_{Prompt}(i) = \frac{1}{N}\sum_{n=1}^{N} x_i^n$$

D Training, Infrastructure and Runtime

We use a server with 8 NVIDIA A100 Tensor Core GPUs, each with 80GB VRAM, to run all our experiments. Each experiment required, at most, two A100 GPUs. Fine-tuning the LLaMA2-13B feedback model took 6 hours on 2 A100 GPUs using our HQ^2A dataset. LoRA fine-tuning of the LLaMA2-13B-chat refinement model took 2 hours on a single A100 GPU using the preference data from HQ^2A. Refining answers with our ERROR-INFORMED REFINEMENT approach took 0.5, 3, and 23 hours for the HQ^2A, ASQA, and ELI5 datasets, respectively, on a single A100 GPU. The evaluation of the refined answers with TigerScore (LLaMA-7B) utilized the VLLM inference library (Kwon et al., 2023) and took approximately 1, 15, and 30 minutes for HQ^2A, ASQA, and ELI5 datasets, respectively, on a single A100 GPU.

Dataset (# samples)	Instruct Model	Tigerscore		SelfCheck Consistency (↓)
		% Hallucinated samples (↓)	Hallucination score (↓)	
HQ^2A (70)	LLaMA2-7B	18.57 ± 0.00	**0.60** ± 0.00	0.166 ± 0.014
	LLaMA2-7B + DPO	**15.71** ± 0.00	0.66 ± 0.00	**0.162** ± 0.015
	Mistral-7B	20.00 ± 0.00	0.57 ± 0.00	**0.266** ± 0.011
	Mistral-7B + DPO	**17.14** ± 0.00	**0.54** ± 0.00	0.285 ± 0.011
ASQA (948)	LLaMA2-7B	**26.58** ± 1.49	**0.86** ± 0.06	0.187 ± 0.014
	LLaMA2-7B + DPO	28.41 ± 1.06	0.89 ± 0.02	**0.178** ± 0.006
	Mistral-7B	62.09 ± 0.35	2.08 ± 0.01	0.578 ± 0.003
	Mistral-7B + DPO	**60.80** ± 0.56	**2.03** ± 0.01	**0.555** ± 0.008
ELI5_GENERAL (1000)	LLaMA2-7B	9.93 ± 1.05	0.32 ± 0.04	0.133 ± 0.001
	LLaMA2-7B + DPO	**9.33** ± 0.66	**0.29** ± 0.03	**0.130** ± 0.004
	Mistral-7B	29.97 ± 0.97	0.90 ± 0.04	0.327 ± 0.003
	Mistral-7B + DPO	**22.77** ± 1.03	**0.72** ± 0.03	**0.319** ± 0.011
ELI5_SCIENCE (1000)	LLaMA2-7B	**9.47** ± 0.47	0.31 ± 0.02	**0.137** ± 0.003
	LLaMA2-7B + DPO	**9.47** ± 0.76	**0.30** ± 0.00	0.139 ± 0.004
	Mistral-7B	34.10 ± 0.94	1.07 ± 0.02	0.320 ± 0.004
	Mistral-7B + DPO	**29.03** ± 1.51	**0.95** ± 0.04	**0.297** ± 0.010
ELI5_HISTORY (1000)	LLaMA2-7B	9.63 ± 0.59	0.30 ± 0.02	**0.188** ± 0.005
	LLaMA2-7B + DPO	**7.60** ± 0.08	**0.22** ± 0.01	0.189 ± 0.005
	Mistral-7B	26.23 ± 0.38	0.79 ± 0.02	0.363 ± 0.016
	Mistral-7B + DPO	**22.17** ± 1.31	**0.69** ± 0.04	**0.345** ± 0.013

Table 6: Results of aligning LLMs with DPO using our collected answer preference data. We measure the hallucinations using Tigerscore and the consistency of model outputs using SelfCheckGPT.

E Additional Results

E.1 Aligning LLMs

Table 6 shows the results for training language models with DPO using our collected preference annotations. Our preference-tuned models outperform the strong baseline models and reduce hallucinated generations in all the evaluation settings except the LLaMA model on the ASQA dataset. We hypothesize that this is due to the ambiguous nature of questions in the ASQA dataset that can have multiple correct answers.

We also observe that the models become more robust and generate more consistent responses after preference-tuning. The only exception is the Mistral model on our held-out test set, which has lower response consistency. We believe this is likely due to the conservative nature of DPO-trained models wherein, during sampling, it can refrain from answering a question in some cases and not in others, leading to a lower consistency score.

E.2 Human Evaluation

This section presents additional details of our human evaluation of the answers refined with our Error-informed feedback approach. In Table 7, we present the agreement of our annotators on two evaluation metrics: comprehensiveness and overall answer preference. The annotators strongly agree that the refined answers are comprehensive, i.e., the answer contains all the required information as asked by the question. For the overall answer preference compared to the baseline, we observe weak agreement between annotators, primarily due to the low agreement value on the ASQA dataset. We hypothesize that the annotators struggle to align on ASQA due to the ambiguous nature of the questions in this dataset, which may have multiple correct answers, and choosing between two answers is difficult.

Dataset	Comprehensiveness (↑)	Preference (↑)
HQ^2A	0.70	0.31
ASQA	0.86	0.02
ELI5	0.92	0.61
Average	0.83	0.31

Table 7: Agreement of annotators on the comprehensiveness and preference of refined answers over the baseline answers from three datasets.

F Annotation Guidelines

We have previously described our data collection setup in §3.3. This section provides additional details on the annotation interface, detailed task instructions, and annotation procedure.

F.1 Annotation Interface

In Figure 6, we show the interface for collecting expert error annotations on LFQA answers. For every question, experts see a human-written and model-generated answer (randomized order). Our expert annotators must select the evaluation layer (top right) and highlight the error span in the question or answer, giving justifications with web references, wherever applicable. After annotating for all the evaluation criteria, experts judge the better answer and mark it in the left pane, giving reasons for their preference.

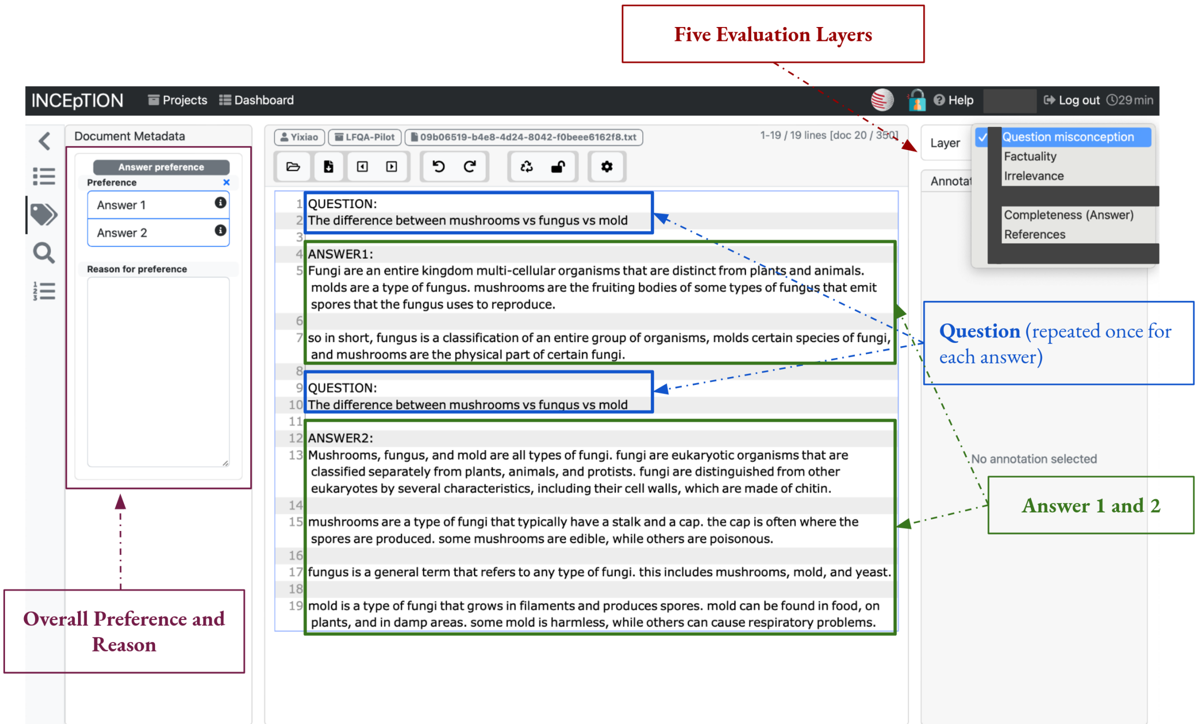

Figure 6: Screenshot of annotation interface for collecting expert error annotations on LFQA answers.

F.2 Task Instructions

We provide experts with detailed task instructions for evaluating answers according to the defined evaluation criteria. We go through every evaluation aspect in depth, defining it and giving annotation examples for clarification, as detailed in the next paragraphs.

1) Question Misconception. You should select a span of text in the question that contains a misconception or false assumption. The question is repeated twice. You only need to select the span in one repetition. If you select such spans, we would like you to indicate in your reason (obligatorily):

- whether the answers reject or correct the misconception/false assumption,
- if no answer rejects/corrects it, please explain in your reason why that is a misconception/false assumption (preferably with references).

Example:
Question: Why is it so important for humans to have a balanced nutrition but not for animals? Most animals have a fairly simple diet, carnivores eat only meat their whole life, cows eat exclusively grass etc.

So why are human bodies so picky and need a balance of protein, fat, carbs etc from different sources to perform well?

2) Factuality. You should select a span of text in the answers that is factually incorrect. If you select such spans, we would like you to (obligatorily):

- preferably give references (e.g., credible websites, academic papers, or books) that show the content is factually wrong, or
- give examples that show the content is factually wrong.

Example:
Question: Why is it so important for humans to have a balanced nutrition but not for animals? Most animals have a fairly simple diet, carnivores eat only meat their whole life, cows eat exclusively grass etc. So why are human bodies so picky and need a balance of protein, fat, carbs etc from different sources to perform well?
Answer: Animals generally have a simpler diet than humans. For example, carnivores only eat meat, while cows only eat grass...
Reason: This is a reductionist view of animal nutrition as it doesn't consider how animals have evolved and the complexities of the food chain. For example, lions are carnivores that only eat meat but they eat the stomach of zebras that contain grass/plants and are able to digest it.

3) Relevance. You should select a span of text in the answers that is irrelevant to answering the question. Removing such content should not affect the overall quality of an answer. If you select such spans, we would like you to (obligatorily):

- explain why the selected text is not relevant to answering the question.

Example:
Question: What is happening when you get migraines that cause you to lose part of your vision for a short time?
Answer: My wife gets these. An ocular migraine is a condition where the blood vessels in the optic nerve at the back of your eye tighten and swell, resulting in various visual distortions. While classic migraines usually result in intense headaches, sensitivity to light and loud sounds, or nausea, ocular migraines are not necessarily painful.
Reason: Answer contains irrelevant information (writer's wife having them, migraine may not be painful). The person's wife's personal health condition doesn't provide useful information to the question, and the question doesn't ask about whether ocular migraines are painful or not.

4) Completeness. You should: (a) select a span of text in the answer that does not offer enough details, or (b) select the label Answer 1 or Answer 2 if some relevant information that should be included in the answer is missing. If you select such spans, we would like you to (obligatorily):

- offer the details or relevant information that you think should be included. References from credible sources is encouraged.

Example:
Question: Why does alcohol make your throat or stomach feel warm when drinking?
Answer: There are a few reasons why alcohol might make your throat or stomach feel warm. first, alcohol is a solvent, which means it can dissolve other substances. when you drink alcohol, it can break down the barrier between your stomach and your throat, causing a burning sensation. second, alcohol can irritate the lining of your stomach, causing inflammation. this can also lead to a burning feeling. finally, when you drink alcohol, your blood vessels dilate, or widen. this increased blood flow can cause a feeling of warmth.
Reason: The answer should give more details on what the barrier is. [Offer relevant information based on your knowledge, for example, what the barrier is.]

5) References/example/analogy. You should select a span of text in the answer that is examples/analogies or a reference to a external credible source (e.g., websites, papers, or books). If you select such spans, we would like you to (obligatorily):

- indicate whether the example/analogy/credible source is helpful, and
- give a reason on why it is helpful/not helpful.

Example:
Question: How does a speaker vibrate at multiple frequencies simultaneously to deliver sounds to our ears? I'll try and explain what I mean.. say a song is playing - how do the vocals, drums, guitars etc which all have different frequencies (aka a plate is moving back & forth at a rate to deliver these sound waves) play at the same time? Surely the plate can't vibrate to deliver two or more frequencies at once?
Answer: Imagine an ocean with a consistent wave. It flows up and down, with equal distance between the two waves at any time. Now imagine I push a larger, shorter wave into this ocean. The two waves will collide, resulting in some new wave pattern. This new wave pattern is a combination of those two waves. Speakers work similarly. If I combine two sound waves, I get a new combination wave that sounds different.
Reason: I like the analogy with the ocean waves, and due to how visual the explanation is it is easier to understand in my opinion.

Answer Preference. Based on the five criteria you have went through, please weight all good and bad points of each answer and choose the one that is better **overall**. Writing a motivation to support your choice is obligatory.

1. When writing your motivation, please refer to the five layers of evaluation.
2. If there are other aspects you used to make your decision but are not in the evaluation layers, please mention them in the reason
3. If you have quotations from the answers, please indicate which answer are the quotations from.
4. Here are some aspects for you to consider (not obligatorily):
    - Nice example/analogy, to the point, generic, concise, informative, useful, well structured, easy to follow ...

Overall Requirement. The overall task requirements are summarized below. Please read them carefully to avoid redoing the task.

1. You have to highlight spans in **both** question answers for these aspects and **give reason** why you highlight a span for an aspect.
2. Mark as many spans as necessary.
3. Please be **objective** in your reasons and avoid using phrases like "I believe" or "I think".
4. Your reasons should be informative and succinct.
5. Please **use declarative sentences** and avoid using questions in your reasons.
6. Products like ChatGPT or BARD are absolutely not allowed.

F.3 Annotation Procedure

The expert annotators spend around 15-20 minutes per question, highlighting the demanding nature of this task. We accordingly pay £10/hour and provide a bonus of £10 for good-quality annotations, resulting in a total cost of £3000 to collect expert judgments for 698 questions. The annotators understand that we will use their annotated data for research purposes. We show a screenshot of an expert annotated answer in Figure 7.

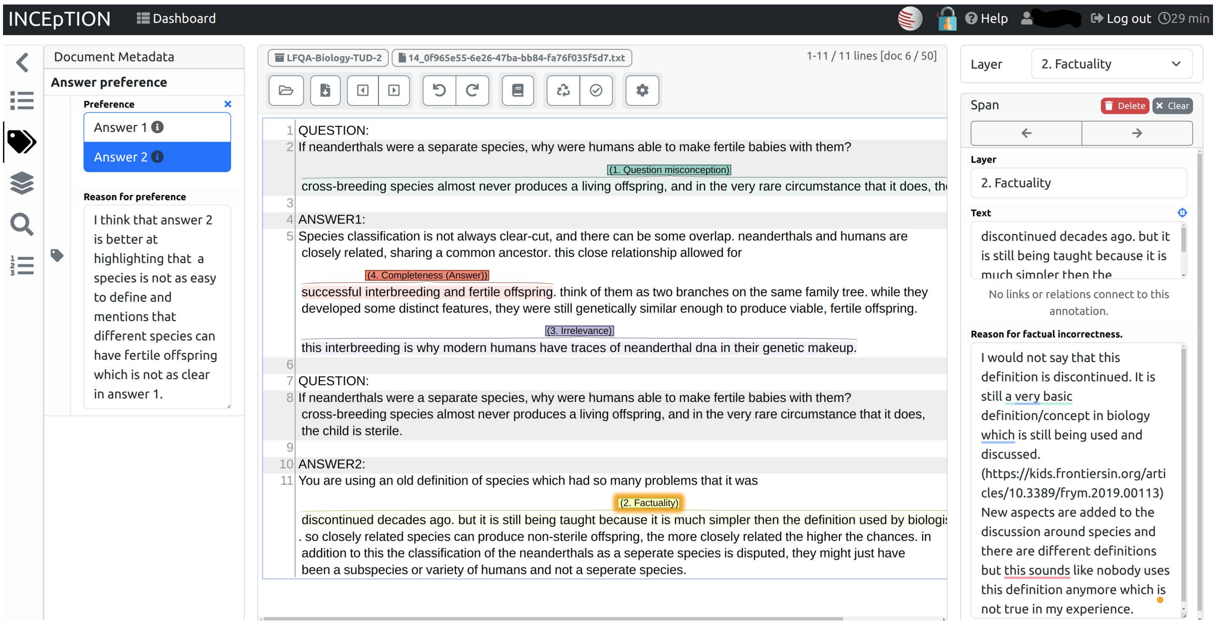

Figure 7: Screenshot of an expert annotated answer on the INCEpTION platform.