🎯 Localizing and Mitigating Errors in Long-form Question Answering

Rachneet Sachdeva[♠], Yixiao Song[♡], Mohit Iyyer[♡], Iryna Gurevych[♠]

[♠]Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt
[♡]University of Massachusetts Amherst

www.ukp.tu-darmstadt.de

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 1.8k 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 errors and improves the quality of the answers across multiple models. Furthermore, humans find the answers generated by our approach comprehensive and highly prefer them (84%) over the baseline answers.[1]

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 hallucinate and produce factually inconsistent, irrelevant, and incomplete content (Goyal and Durrett, 2020; Laban et al., 2022; Menick et al., 2022; Ji et al., 2022), which are difficult to detect for both humans and machines.

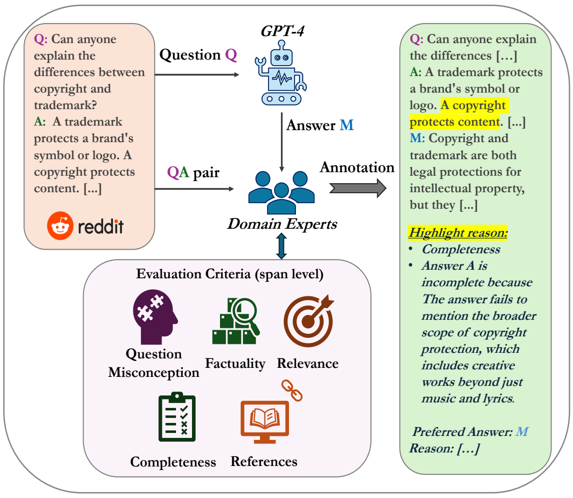

Figure 1: Overview of our data collection process. Using five fine-grained evaluation criteria, we collect *span-level* expert human judgments on question-answer pairs from the Reddit platform, as well as on corresponding answers generated by GPT-4.

Traditional evaluation metrics of answer quality, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020) yield only a single score, obscuring the error *type*, *severity*, and *location* in the answer. We take inspiration from machine translation, which moved beyond this simplistic evaluation paradigm by localizing and categorizing errors (Freitag et al., 2021; Kocmi et al., 2024), resulting in higher-quality and more interpretable evaluations. We make a similar contribution to the field of LFQA by asking human annotators to identify spans from the answers that correspond to errors and categorize each span into an error schema that we design. Our work is the first to explore error localization in LFQA, offering a more detailed and interpretable evaluation of answers.

LLMs make many errors for LFQA that require targeted evaluation. Xu et al. (2023a) highlight that key aspects such as *factuality, relevance, completeness, structure, references, and accessibility*

[1]Code and data available at: github.com/lfqa-errors

Findings of the Association for Computational Linguistics: ACL 2025, pages 20437–20469

are essential to evaluate long-form answers. Other recent studies have also validated the importance of aspects such as question misconception (Krishna et al., 2021), factuality (Wu et al., 2023b; Jiang et al., 2023b; Wang et al., 2024b), relevance (Tang et al., 2024), and completeness (Samarinas et al., 2025) in determining answer quality. While prior work has mainly focused on evaluating factuality (Lee et al., 2022; Min et al., 2023; Li et al., 2023; Muhlgay et al., 2023) and faithfulness (Su et al., 2022) in long-form text generation, other aspects of evaluation, such as response completeness and relevance (which can particularly mislead users), have received less attention.

Our work addresses this gap by introducing *HaluQuestQA*, a dataset of long-form answers annotated at the span level with five error types: *question misconception, factuality, completeness, relevance, and references*. Expert annotators provide these annotations and preference judgments, as shown in Figure 1.

Next, we train an automatic feedback model on this dataset to predict erroneous answer spans that lack key details to address the question comprehensively. The feedback model provides fine-grained feedback, identifying error locations (sentence level), error justification, and a confidence score, all without relying on reference texts (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), reducing errors and improving answer quality across multiple LLMs.

Our contributions are summarized as follows:

- We release *HaluQuestQA*, a dataset with span-level error annotations on pairs of human-written and model-generated answers. Our analysis reveals that long-form answers often lack comprehensiveness and provide unhelpful references.
- We train a feedback model to identify erroneous answer spans with *incomplete information*, aligned with expert human judgments. Although our dataset encompasses multiple errors, our feedback model focuses on completeness errors, which are identified as the most critical issue in the LFQA answers.
- We propose Error-Informed Refinement, an approach that applies fine-grained feedback from our learned model to improve the quality of human-written and LLM-generated answers.

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 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 used to investigate errors in summarization tasks, ours is amongst the first works to study LFQA-centric errors at the span level. To this end, we collect span-level annotations of LFQA errors, enabling high-quality and interpretable evaluations that can be used to improve answer quality. While this has been done for machine translation (Freitag et al., 2021; Kocmi et al., 2024), it has not yet been applied to long-form question answering.

Automatic evaluation. 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. Jiang et al. (2023b) propose a reference-free evaluation metric, TIGERSCORE, that can locate, categorize, and explain errors across various text generation tasks, including LFQA. While LLM-based metrics can detect diverse errors, they are prone to hallucinations due to training data quality. In this study, we collect expert-annotated data on fine-grained LFQA errors and train a feedback model for accurate error detection.

Mitigating errors with human feedback. Reinforcement learning with human feedback (RLHF) (Ziegler et al., 2019) incorporates human feedback to train reward models and align LLMs, reducing undesirable generations (Ouyang et al., 2022; Bai et al., 2022a,b; Wu et al., 2023b). A recent alignment technique, direct preference optimization (DPO) (Rafailov et al., 2023), bypasses the computationally expensive reward modeling step and has been used to fine-tune LMs for factuality (Tian et al., 2023). Human feedback has also been used to fine-tune feedback models (Wang et al., 2023; Xu et al., 2024) to guide the refinement

Category (# samples)	Preference		Krippendorf'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 Krippendorf's alpha (Hayes and Krippendorff, 2007) (Appendix A.5).

of LLM outputs (Madaan et al., 2023; Welleck et al., 2023; Ji et al., 2023), improving answer quality. However, these feedback models either lack fine-grained error feedback or depend on ground truth passages, which may not always be available in open-domain QA. In our study, we develop a reference-free feedback model to refine LFQA answers with detailed error feedback.

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 all erroneous or problematic parts (phrase, sentence, or multiple sentences) based on our defined evaluation criteria (§3.2). After carefully evaluating the pilot results for relevance, clarity, and factuality, we choose three experts per domain and give them each a large-scale study containing 35-50 QA pairs. We collect expert judgments for 698 questions.

3.2 Task Setup

We evaluate two answers (human and model-generated) to the same questions. 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 outperforming existing open-source LLMs (LLaMA and Alpaca (Taori et al., 2023)) in reasoning and inferring from long context. Since GPT-4's training data extends 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 with their highest-voted answers from the *r/explainlikeimfive* subreddits posted between November 2022 and March 2023, following Xu et al. (2023a). We provide further details of the setup 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 links in the answer.

Annotators highlight errors in questions or answers based on evaluation criteria, provide free-form justifications, and indicate overall answer preferences.

3.3 Data Curation

To ensure high-quality annotations, we implement several checks before, during, and after the annotation process. Before annotation, we ensure the quality of scraped questions by removing duplicates, discarding questions under five words, and verifying grammar and fluency. For the corresponding answers, we select responses that receive the highest upvotes on Reddit and have a minimum length of 50 words. We also verify that the responses are not generated by AI-assisted tools (e.g., ChatGPT[3])

[2] We provide detailed annotation guidelines in Appendix F.
[3] `https://chat.openai.com`

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. [...] can buy on commissary (fish [...]	*(model)* [...] In prison, inmates often have a lot of free time, which they can use to work out and develop a consistent exercise routine. [...]	This is an assumption, not every prison offers terrible food without nutritional value, also lots of "easy" food choices (lentils, eggs, [...]
Factuality	What is difficult about developing nuclear weapons that makes some countries incapable of making them?	*(human)* The main problem is the nuclear fuel [...] Uranium is a fairly rare element on its own, but to make a bomb you need lots of a very rare isotope of uranium (u-235) [...]	*(model)* Developing nuclear weapons is a highly complex and resource-intensive [...] To create a nuclear bomb, a nation must have a deep understanding of [...]	The claim is false. According to the World Nuclear Association, Uranium is not a rare element and occurs in potentially recoverable concentrations in many types [...]
Relevance	Why is it taboo to talk about people's wages?	*(human)* Because companies used to be able to fire [...] discussing your wages is a federally [...] so by all means discuss discuss discuss! [...]	*(model)* Talking about wages is considered taboo because it can create tension and jealousy [...]. Employers often discourage discussions [...]	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 [...]	*(model)* [...] Copyright protects creative works like books, music, and art, [...]. Trademark, on the other hand, protects brand identifiers [...]	The answer fails to mention the broader scope of copyright protection, which includes creative works beyond just music and lyrics.
References	Why is 1000mbps internet only guaranteed at 900mbps and what's the equipment needed to get the 1000mbps advertised?	*(human)* [...] It's similar to asking why postal mail weighs more than just the paper you are sending... Everything that tells it how to get to the other side encapsulates (contains) it, and makes up that extra weight.	*(model)* Internet speeds are often advertised as "up to" a certain speed because various factors can affect the actual performance. For instance, network congestion... To get the full 1000mbps advertised, [...]	The example is terrible and does not give any idea of what the writer is talking about.

Table 2: Examples of expert annotated errors in long-form answers based on the defined evaluation criteria. The answers in *green* are expert-preferred answers. Additional examples are provided in Table 13 in the Appendix.

using a commercial AI detector (GPTZero[4]). We find that none of the responses are flagged as AI-generated, though we acknowledge the inherent limitations of AI detection tools. During annotation, annotators are encouraged to contact the authors anonymously through Prolific for clarification to reduce potential errors. After the study, we manually review error spans, justifications, references, and preference judgments, verifying their quality and ensuring no AI involvement. Iterative feedback and bonus payments further incentivize high-quality work. Examples are shown in Table 2.

3.4 Quantitative Analysis

As shown in Table 1, experts display a high preference (80.7%) for GPT-4 answers compared to human answers. We hypothesize that 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., 2023b) and LFQA (Xu et al., 2023a), who show that GPT-3 answers score higher than human answers.

Comparing different domains, we observe that experts strongly prefer GPT-4 answers in history, law, technology, and economics (>80%). However, in science domains like physics, biology, and chemistry, model preference drops to 60-80%. GPT-4's strong performance in history and law can be attributed to its ability to learn facts that remain consistent over time, during its training process. However, it struggles with college-level scientific questions requiring advanced reasoning (Sun et al., 2024; Wang et al., 2024a) – and our dataset includes complex, real-world scientific problems that surpass college-level difficulty, likely contributing to its lower performance in scientific domains.

3.5 Fine-grained Answer Scoring

We score human and model answers on our defined evaluation criteria to understand how experts' answer preferences diverge across different domains. For the *question misconception* aspect, the score $\mathcal{S} = 1$ when the question has no misconceptions; otherwise, $\mathcal{S} = 0$. For aspects of *factuality*, *relevance*, and *completeness*, $\mathcal{S} = 1 - \left(\frac{\text{\# Error sentences}}{\text{Total \# of sentences}}\right)$, while the score for *reference* is calculated as $\mathcal{S} = 1 - \left(\frac{\text{\# Error references}}{\text{Total \# of references}}\right)$.
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.

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

[4] https://app.gptzero.me/

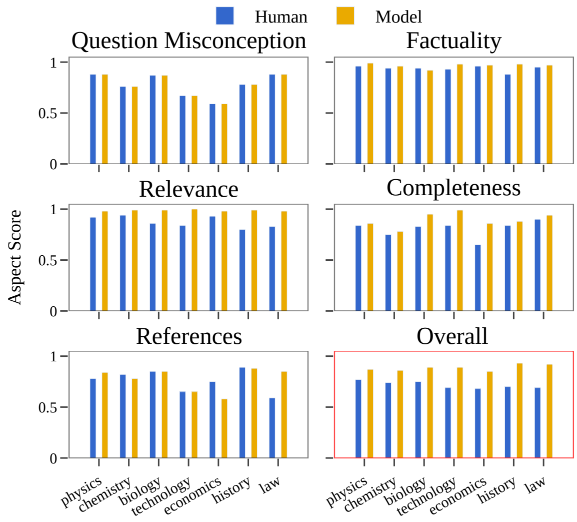

Figure 2: Comparison of fine-grained scores of the human-written and model-generated answers for different evaluation criteria. The last figure (with red boundary) shows the averaged and normalized overall scores. A higher score represents fewer errors in the answers.

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 ~40% of the questions are misinformed – users have low domain knowledge to ask the right questions. Thus, we encourage future research to assess the capability of LLMs to rebut misconceived questions.

2) Answers lack comprehensiveness and provide unhelpful references. We observe that human-written and model-generated answers score high on *factuality* and *relevance*, meaning most of the information provided is verifiable, trustworthy and relevant to the question. However, the answers score low on *completeness* and *references* aspects, lacking important information and providing web references and examples that are not helpful (Liu et al., 2023a), according to expert judgments. Specifically, GPT-4 hallucinates and provides incorrect or fabricated web links, while human answers digress from the topic and include irrelevant information.

Overall, GPT-4 answers score better than the human answers in all evaluated domains. While this is due to its better performance over humans in the considered aspects, the persuasive nature of the model responses (Salvi et al., 2024) also plays a crucial role in their higher preference.

4 Error Mitigation

In §3.4, we have shown that the LFQA answers lack completeness and omit useful 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 the answers and improve their overall quality without human intervention.

4.1 Error Feedback Model

Given an input question and an LFQA response, we fine-tune the LLaMA2-13B model (Touvron et al., 2023b) to generate a label *[Complete]* or *[Incomplete]* for every sentence $1 \ldots n$ in the response and provide associated reasons for the incomplete sentences (see Figure 3).

Dataset & Training. Training the feedback model requires high-quality error annotations with justifications. To support this, we extract QA pairs with annotated completeness errors from our dataset, which includes both sentence- and phrase-level annotations. Since ~65% of annotated completeness errors occur at the sentence level (see Table 6), we adopt sentence-level granularity as both representative and practical for training and evaluating our feedback model. To facilitate this, we convert phrase-level errors to sentence-level annotations by assigning the phrase's error label and justification to its containing sentence. Furthermore, in cases where annotators mark the entire answer as incomplete (Appendix F.2), we assign the same error label and justification to each sentence in the answer. We provide illustrative examples of sentence-level annotations in Table 12 in the Appendix.

Finally, after preprocessing, we segment each human- or model-generated answer into sentences and label each sentence as *[Complete]* or *[Incomplete]*, with the corresponding expert-provided justification. The final dataset consists of 509 samples with a 90/10 train/test split. Fine-tuning on this dataset is essential for accurate error feedback generation, as general-purpose LLMs used in a zero-shot setting are not well-suited to detecting completeness errors (§5.1). We provide the training details in Appendix B.2.

Inference. The trained feedback model hallucinates web references in about 20% of test sam-

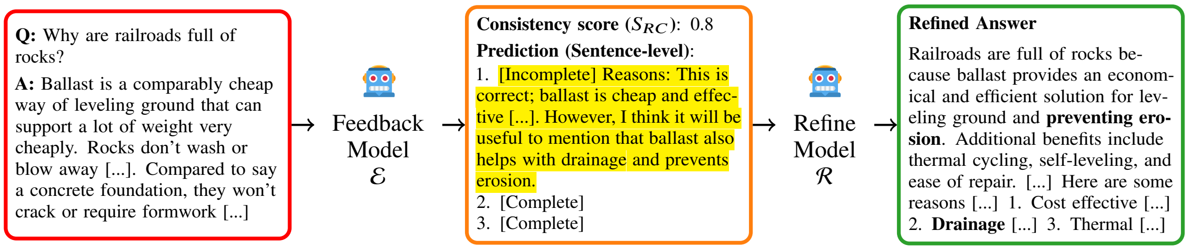

Figure 3: A pictorial view of our Error-Informed Refinement approach. The FEEDBACK model takes a question-answer pair as input and outputs sentence-level error with justifications and a consistency score. The REFINE model uses this feedback to improve the original answer. Additional refined examples are in Table 14 (Appendix).

ples. 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 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 $p_1, \ldots, p_n$ consist of tag sequences $t_1, \ldots, t_n$ where t_i is a list of tag predictions for every span, the score for sample i is

$$\boxed{\mathcal{S}_{\mathcal{T}C} = \frac{1}{n} \sum_{s=1}^{n} 1_{t_i = t_s}}$$

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 $j_1, \ldots, j_n$ consist of words $w_i^k, k = 1 \ldots m_i$, the score of sample i is

$$\boxed{\mathcal{S}_{\mathcal{R}C} = \frac{1}{m_i} \sum_{k=1}^{m_i} \sum_{s=1}^{n} 1_{w_i^k \in j_s}}$$

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 highest scoring output as feedback for the refinement model. After sampling, **reference hallucinations reduce by 50% (from 20% to ~10% of the test set).**

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 $\mathcal{R}$ uses x_i, y_i, and f_i to generate a refined and improved output $\hat{y}_i$. The following sections describe our approach.

Refinement Model & Baselines. Our experiments use the LLaMA2-13B chat LLM and its DPO optimized version (see Appendix C) as the refinement model. 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 error 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^{2}A 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

Approach	Model	Accuracy (%)			Weighted Accuracy (%) (↑)	Consistency Score (S_{RC}) (↑)
		Exact (↑)	Adjacent (↓)	Different (↓)		
Zero-shot	LLaMA2-13B	23.53 ± 1.60	**7.84** ± 0.00	68.63 ± 1.60	34.31 ± 1.44	0.52 ± 0.02
Zero-shot	GPT-3.5-Turbo	25.49	11.76	62.75	37.65	**0.99**
Fine-tuning w/ HQ^2A	LLaMA2-13B	**37.25** ± 0.00	24.18 ± 0.92	**38.56** ± 0.93	**53.20** ± 0.37	0.80 ± 0.01

Table 3: Accuracy and Consistency Score (S_{RC}) of zero-shot and fine-tuned models in detecting sentence-level errors on HQ^2A, averaged over three runs with standard deviations (except for GPT-3.5). Best scores are in **bold**.

from the Reddit forum "Explain Like I'm Five".

We evaluate the refined answers using TIGERScore, a trained reference-free metric that identifies errors in LLM-generated text and assigns an *error 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 cost-effective. We also measure how well our refinement approach corrects errors identified by TIGERScore using precision, recall, and F1 score metrics. Finally, we conduct a human evaluation to assess the comprehensiveness and preference of 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 errors and improve the comprehensiveness of LFQA answers? (4) Are comprehensive answers from our approach preferred by humans?

5.1 Detecting Errors via Feedback Model

To measure the error detection accuracy of our feedback model, we propose an evaluation across three fine-grained categories: (1) **Exact**: Erroneous sentences identified by the model exactly match the human-annotated erroneous sentences. This category represents the most stringent evaluation of model performance. (2) **Adjacent**: Erroneous sentences identified by the model are adjacent to, or closely related to, human-annotated erroneous sentences. Here, "adjacent" refers to a sentence preceding or following the human-annotated error sentence. These near-misses may still aid in understanding or resolving the error due to the contextual relation between the preceding/following sentence and the actual error sentence. (3) **Different**: Erroneous sentences identified by the model do *not* match, precede or follow human-annotated error sentences, capturing instances where the model detects completely unrelated error spans.

To capture the overall error detection performance across the defined evaluation categories, we introduce a **weighted accuracy** metric:

$$\text{Accuracy}_{\text{wt}} = \begin{aligned} & w_{\text{Exact}} \cdot \frac{\#\,\text{Exact matches}}{\text{Total errors}} + w_{\text{Adj}} \cdot \frac{\#\,\text{Adjacent matches}}{\text{Total errors}} \\ & + w_{\text{Diff}} \cdot \frac{\#\,\text{Different matches}}{\text{Total errors}}. \end{aligned}$$

where w_{Exact}, w_{Adj}, and w_{Diff} represent the weights assigned to each category according to its relative importance. We assign $w_{\text{Exact}} = 1.0$ to reward the model's capability of correctly detecting errors and $w_{\text{Adj}} = 0.5$ to quantify the importance of near-misses which may still provide insight on the actual errors. Consequently, $w_{\text{Diff}} = 0.1$ to penalize the model for its incorrect error detection.

In Table 3, we show the sentence-level error detection accuracy of the zero-shot LLaMA2-13B and GPT-3.5-Turbo and our fine-tuned feedback models compared to the strong human baseline. Our fine-tuned feedback model improves the detection of correct error spans (*exact*) by ~14% and ~12% and reduces the detection of incorrect error spans (*different*) by ~30% and ~24% compared to the zero-shot feedback models LLaMA2-13B and GPT-3.5-Turbo, respectively. Specifically, our feedback model outperforms GPT-3.5-Turbo by ~16% on our weighted accuracy metric while maintaining a high consistency score of 0.80. This shows that the model effectively learns to identify completeness errors, even when fine-tuned on *limited but high-quality* HQ^2A samples, aligning with recent findings (Zhou et al., 2023; Xia et al., 2024) on fine-tuning with small but carefully curated datasets. In Appendix B.2, we discuss systematic patterns of errors learned by the feedback model.

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 refine-

Dataset	Approach	TIGERScore		Error Correction		
		% Error samples (↓)	Error 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	-	-	-
	Zero-shot	15.69 ± 0.00	0.34 ± 0.00	0.56 ± 0.00	0.90 ± 0.00	0.69 ± 0.00
	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 (*Ours*)	**0.65** ± 0.92	**0.03** ± 0.04	0.97 ± 0.04	**1.00** ± 0.00	**0.98** ± 0.02
ASQA	Baseline	34.81	1.20	-	-	-
	Zero-shot	35.02 ± 0.00	1.08 ± 0.00	0.50 ± 0.00	0.62 ± 0.00	0.55 ± 0.00
	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 (*Ours*)	**16.63** ± 0.41	**0.51** ± 0.02	**0.73** ± 0.00	**0.82** ± 0.02	**0.77** ± 0.01
ELI5	Baseline	22.93	0.82	-	-	-
	Zero-shot	9.61 ± 0.00	0.27 ± 0.00	0.74 ± 0.00	0.89 ± 0.00	0.81 ± 0.00
	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 (*Ours*)	**3.81** ± 0.30	**0.13** ± 0.01	**0.88** ± 0.01	**0.96** ± 0.01	**0.92** ± 0.01

Table 4: Results on the quality of original answers from the datasets (BASELINE); answers from zero-shot prompting LLaMA2-13B-chat (ZERO-SHOT); answers refined with coarse-grained feedback (IMPROVE and GENERIC), fine-grained feedback (EIR) and human feedback on HQ^2A. Reported results are averages over three iterations with standard deviations. Best results are in **bold green**, and the second-best results are in orange. We report results with LLaMA3-8B-Instruct (Dubey et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a) models in Appendix E.3.

ment of long-form answers and reducing errors. In Table 4, we present the refinement performance with our feedback model compared to the expert human feedback on HQ^2A. Our feedback model reduces error samples by 2% and improves the F1 score by 4% over expert human feedback, validating its effectiveness in refining LFQA answers.

5.2 Fine- vs. Coarse-Grained Feedback

Table 4 presents the quality of BASELINE answers (original dataset instances) refined using coarse- and fine-grained feedback. We also evaluate answers generated through zero-shot prompting LLaMA2-13B-chat for comparison.

Our results show that inadequate feedback can deteriorate generation quality. While directly prompting the refinement model to generate answers (ZERO-SHOT) or improve answers without detailed feedback (IMPROVE) performs better than the baseline, using more targeted feedback, such as asking the model to complete the answer (GENERIC), consistently leads to higher-quality LFQA answers. In contrast, fine-grained feedback from our error detection model (EIR) outperforms coarse-grained feedback and fine-grained human feedback (on HQ^2A), reducing error samples and error scores by ~3% and ~Δ38%, respectively, and improving F1 scores by ~5%, on average.

We also investigated the impact of aligning the refinement model with human preferences from HQ^2A with DPO. Despite promising initial results in reducing LFQA errors (Appendix E.1), the resulting refinement model ultimately did not outperform the vanilla refinement model (Appendix E.2).

5.3 Human Evaluation

We conduct a human evaluation with three annotators to test the completeness and overall quality of the answers generated using our refinement approach. For 50 questions each from the HQ^2A, ASQA, and ELI5 datasets, we present annotators with a pair of answers—one from the dataset (baseline) and one refined by our method.

To evaluate completeness, we adopt a comparative *comprehensiveness* metric: annotators judge which answer more fully addresses all parts of the question, based on our defined criteria for identifying completeness errors (see Appendix F.2). To assess the overall answer quality, annotators consider broader factors, such as the factual precision and relevance (Appendix F.2), when selecting their preferred answer.

Table 5 shows the results of our human evaluation of the BASELINE and REFINED answers.

Dataset	Pref.	Comprehensiveness(↑)	Overall(↑)
HQ^{2}A	Baseline	0.00%	7.84%
	Refined	**100**%	**92.16**%
	Tie	0.00%	-
ASQA	Baseline	0.00%	40.00%
	Refined	18.00%	**60.00**%
	Tie	**82.00**%	-
ELI5	Baseline	0.00%	0.00%
	Refined	**62.00**%	**100**%
	Tie	38.00%	-

Table 5: Human evaluation results on the comprehensiveness and preference of answers refined with EIR over the original answers from the datasets (BASELINE). Details on the human agreement are in Appendix E.4.

We observe that refined answers are considered more comprehensive in ~60% of cases and preferred overall in ~84% of comparisons on average across all evaluated datasets, demonstrating improved completeness and quality over the baseline answers.

6 Conclusion

We introduce HALUQUESTQA, a dataset of expert judgments on fine-grained errors 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 errors. 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 errors 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 have experimented with the instruction-tuned variants of the LLaMA2, LLaMA3, and Mistral models. 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 errors.

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.

Acknowledgements

This research work is funded by the German Research Foundation (DFG) as part of the UKP-SQuARE project (grant GU 798/29-1) and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. Yixiao Song and Mohit Iyyer are supported by the award IIS-2312949 from the National Science Foundation (NSF).

We thank Sukannya Purkayastha and Haritz Puerto for their insightful feedback on the paper and Manika Arvind Arora for the valuable feedback on the annotation setup. Lastly, we are grateful to our dedicated annotators who helped create the HaluQuestQA dataset.

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.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

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.

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.

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.

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.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. *CoRR*, abs/2310.06271.

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.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popovic, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *CoRR*, abs/2406.11580.

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.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7001–7025. 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. 2023b. 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.

Chris Samarinas, Alexander Krubner, Alireza Salemi, Youngwoo Kim, and Hamed Zamani. 2025. Beyond factual accuracy: Evaluating coverage of diverse factual information in long-form text generation. *CoRR*, abs/2501.03545.

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.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 744–756, Dublin, Ireland. Association for Computational Linguistics.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19053–19061. AAAI Press.

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.

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.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024a. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yuqi Wang, Lyuhao Chen, Songcheng Cai, Zhijian Xu, and Yilun Zhao. 2024b. Revisiting automated evaluation for long-form table question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14696–14706, Miami, Florida, USA. Association for Computational Linguistics.

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*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. 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*.

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 Collection and Analysis

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

A.1 Domain Classification

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 answers.

A.2 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.

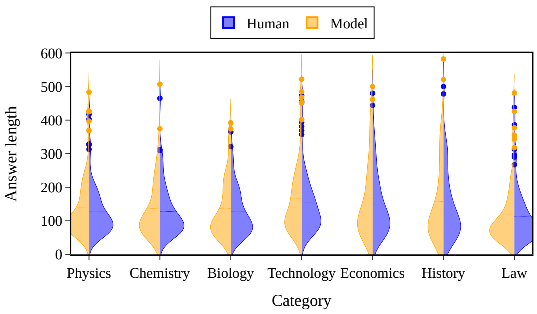

Figure 4: Answer length distribution of human-written and model-generated answers (H/M) in our expert-annotated dataset.

A.3 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.

A.4 Span-level Annotations

In Table 6, we present the distribution of errors annotated at different span-levels (phrase, sentence and multi-sentence). Our findings show that experts identify phrase-level errors in approximately 35% of the cases, indicating that a substantial portion of errors are nuanced and cannot be effectively captured at the sentence level. This highlights the need to employ fine-grained span-level annotations to enhance the evaluation process, as they provide deeper insight into the nature and exact location of errors, ultimately leading to improved answer quality by targeting specific errors.

Error Type	**Annotated Spans**		
	Phrase-level	**Sentence-level**	**Multi-sentence-level**
Question Misconception	38.89%	52.47%	8.64%
Factuality	42.40%	44.88%	12.72%
Relevance	25.00%	39.13%	35.87%
Completeness	35.81%	34.63%	29.56%
References	31.24%	30.77%	38.00%
Average	34.67%	40.38%	24.96%

Table 6: Overview of error types and the corresponding annotation distribution across phrase-level, sentence-level, and multi-sentence-level errors.

A.5 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.[5] 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).

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.

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

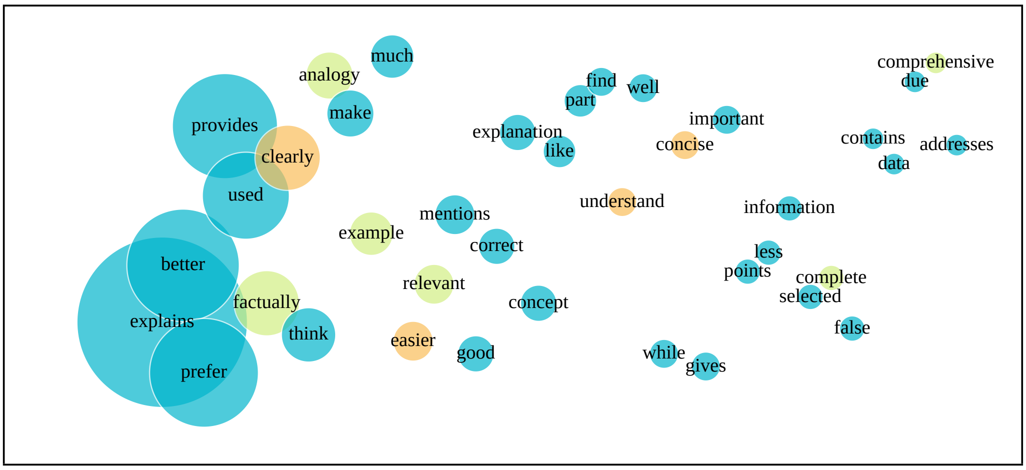

Figure 5: Distribution of the top 50 most common words mentioned by our expert annotators in their overall answer justifications. The size and color of the bubble represent the word frequency and importance, respectively. The green and orange colors denote the important evaluated and non-evaluated aspects,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

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.

B.2 Feedback Model

We use expert error annotations for the *completeness* aspect from our HQ^2A dataset and train the feedback model for 5 epochs with a learning rate $2e-5$ and a sequence length of 1024. 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.

There are clear patterns of how humans assess incomplete answers, which the model can learn from our training data. We explain them below, along with examples from our annotated dataset:

1. **Missing explanation of key concepts in answer.** These refer to concepts that are introduced in the answer but fail to explain it properly, assuming the reader already knows about it.

   Example:

   Q: How does rendering a video game resolution above your monitor resolution make the picture more crisp? You would think the resolution is already "maxed out", how does this improve things?

```
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.

A: It's called supersampling and basically gives the computer more options for what to render for any given pixel. If you "render" at a higher resolution so that each final pixel is 4 pixels in the supersample then the computer gets 4 reference points for what color that pixel could be instead of just one. [...]

Comment: Supersampling is not explained clearly in this answer.

2. **Failure to Address All Parts of the Question.** The answer answers only a part of the question.

   Example:

   Q: Why can't spam call centers be automatically shut down? Additionally, why can't spam calls be automatically blocked, and why is nobody really doing a whole lot about it? It seems like this is a problem that they would have come up with a solution for by now.

   A: Spam call centers can't be automatically shut down because they are often located in countries with weak regulation and they use advanced techniques to hide their identity, like caller id spoofing. Blocking spam calls is challenging because these callers use ever-changing phone numbers, making it difficult to maintain an updated block list. While there are some efforts to combat spam calls, such as telecom companies implementing call-filtering systems and governments enacting regulations, the issue is deeply rooted in the economic incentive for spam callers and the complexities in tracking them down. Ultimately, tackling spam calls requires international cooperation and continuous technology advancements.

   Comment: This answer is incomplete because it does not fully answer parts of the question but only hints at them, such as "Why is nobody really doing a whole lot about it?" and "Why hasn't a solution been found by now?".

3. **Unfinished sentences or the lack of logical flow.** Sometimes, a sentence in the answer may be left unfinished or lacks a clear, logical flow, making it hard to understand.

   Example:

   Q: How can a collision cause a loss of energy? Even if friction and energy loss is ignored?

   A: There isn't any energy loss. It's conserved. There is a kinetic energy loss because, in most collisions, there's heat, sound, and (often) distortion of the colliding objects. The energy changed forms to some that aren't kinetic, so naturally you've got less kinetic energy after than before. Your second equation (vf2=...) is just conservation of momentum. This doesn't looks like it's an energy question.

   Comment: The first sentence says "There isn't any energy loss. It's conserved." But immediately after, the answer states "There is a kinetic energy loss..." without making a clear distinction between total energy conservation and kinetic energy loss.

4. **Lack of examples to provide more clarity.** Incomplete answers may lack references or examples to provide clarity to the reader.

   Example:

   Q: How come a CPU can function without any driver updates whereas a GPU needs one every month or so for stability?

   A: The eli5 answer is that you do run updates for your CPU, just don't notice that because they are packaged into stuff like iOS updates or windows updates. Operating systems have a thing called "kernel" that can be seen as the driver of your CPU (yes yes I know is not a perfect analogy but this is eli5) and that is updated regularly by windows/mac/linux operating systems.

   Comment: It doesn't mention any example of such an update or list any type of such update to give a clearer picture of the event it is describing.

5. **Insufficient justification for a claim.** Ideas/-claims without backing them up with evidence or explanations.

   Example:

   Q: If most electronic appliances' efficiency losses are through heat, does that mean that electric heaters are 100% efficient?

   A: Yes, electric heaters convert 100% of the power that they consume into heat. So they have an efficiency of 100%. Heat pumps move heat from one area (outside your house) to another area (inside your house) the amount of heat they move is typically about 3 times more than the power they consume. So the in terms of energy-to-heat efficiency, they are 300%+ efficient. But thermodynamically they are not "creating" heat from nothing. So heat pumps are not perpetual motion machines, they don't break any of the laws of thermodynamics.

Comment: The answer lacks sufficient justification for the claim that electric heaters are 100% efficient. More information about the mechanisms it uses to achieve that 100% efficiency would make the answer more complete.

```
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.

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.

```
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).

C Mitigating Errors 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 direct preference optimization (DPO) (Rafailov et al., 2023) 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 errors.

Implementation details. We model data from HQ^2A as a preference dataset where every question

```
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.

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 errors in the input text and assigns an error 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 (Manakul et al., 2023). 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:

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

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.

E Additional Results

E.1 Aligning LLMs

Table 7 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 error 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

Dataset (# samples)	Instruct Model	TIGERScore		SelfCheck Consistency (↓)
		% Error samples (↓)	Error 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 7: Results of aligning LLMs with DPO using our collected answer preference data. We measure the errors using Tigerscore and the consistency of model outputs using SelfCheckGPT. Reported results are averages over three iterations with standard deviations. The best scores are marked in **bold**.

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 EIR with DPO

In Table 8, we present the quality of answers refined using different types of feedback (coarse- and fine-grained), alongside the BASELINE answers. Additionally, we include the results for answers refined with the DPO-aligned model. While the DPO-aligned refinement model does not outperform the vanilla refinement model in reducing the overall number of error samples, it achieves the best error scores on ASQA and ELI5. This suggests that the DPO optimization is still effective in correcting major errors to some extent.

E.3 Fine- vs. Coarse-grained Feedback

In Table 9 and Table 10, we show the results on the quality of answers generated with zero-shot prompting (ZERO-SHOT) as well as answers refined using coarse (IMPROVE and GENERIC) and fine-grained (EIR) feedback, using the LLaMA3-8B-Instruct and Mistral-7B-Instruct-v0.3 models, respectively. Similar to the observations in §5.2, we notice that inadequate feedback deteriorates the quality of generation.

When using LLaMA3-8B-Instruct as the refinement model, the direct prompting (ZERO-SHOT) and refining without detailed feedback (IMPROVE) approaches improve answer quality over the BASELINE (original answers from the dataset) on all the datasets, except ASQA, where the ZERO-SHOT approach generates lower quality answers than BASELINE, likely due to the ambiguous nature of the questions in the ASQA dataset. On the contrary, prompting with more targeted feedback (GENERIC) consistently outperforms the BASELINE, ZERO-SHOT, and IMPROVE approaches, generating better quality LFQA answers and giving the best scores on HQ^2A. Furthermore, providing fine-grained feedback from our error detection model (EIR) outperforms coarse-grained feedback on ASQA and ELI5 datasets, reducing error samples and error scores by ~8% and ~Δ68%, respectively, and improving F1 scores by ~11% on average.

When using Mistral-7B-Instruct-v0.3 as the refinement model, the approach to refine answers without detailed feedback (IMPROVE) improves answer quality over the BASELINE, ZERO-SHOT, and

Dataset	Approach	TIGERScore		Error Correction		
		% Error samples (↓)	Error 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 (*Ours*)	**0.65** ± 0.92	**0.03** ± 0.04	0.97 ± 0.04	**1.00** ± 0.00	**0.98** ± 0.02
	EIR w/ DPO (*Ours*)	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 (*Ours*)	**16.63** ± 0.41	0.51 ± 0.02	**0.73** ± 0.00	**0.82** ± 0.02	**0.77** ± 0.01
	EIR w/ DPO (*Ours*)	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 (*Ours*)	**3.81** ± 0.30	**0.13** ± 0.01	**0.88** ± 0.01	**0.96** ± 0.01	**0.92** ± 0.01
	EIR w/ DPO (*Ours*)	5.71 ± 0.25	**0.13** ± 0.00	0.83 ± 0.00	0.94 ± 0.01	0.88 ± 0.00

Table 8: Results on the quality of original answers from the datasets (BASELINE), answers refined with coarse-grained feedback (IMPROVE and GENERIC), and fine-grained feedback (EIR). Additionally, we include the results of refinement with expert human feedback on our collected data. Reported results are averages over three iterations with standard deviations. The best scores are marked in **bold**.

even the GENERIC approach, achieving the best scores on HQ^2A. We hypothesize that this is due to the capability of the model to understand simplistic feedback instructions to improve answers, leading to a better performance than the GENERIC approach. In contrast, providing fine-grained feedback from our error detection model (EIR) outperforms coarse-grained feedback on ASQA and ELI5 datasets, reducing error samples and error scores by ~2% and ~Δ27%, respectively, and improving F1 scores by ~4% on average.

E.4 Human Evaluation

This section presents additional details of our human evaluation of the answers refined with our Error-informed feedback approach. In Table 11, 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	Approach	TIGERScore		Error Correction		
		% Error samples (↓)	**Error score (↓)**	**Precision (↑)**	**Recall (↑)**	**F1 (↑)**
HQ^{2}A	Human feedback	1.96 ± 0.00	0.07 ± 0.01	0.90 ± 0.01	**1.00** ± 0.00	0.95 ± 0.00
	Baseline	19.61	0.63	-	-	-
	Zero-shot$_{\text{LLaMA3}}$	17.65 ± 0.00	0.46 ± 0.00	0.53 ± 0.00	0.80 ± 0.00	0.64 ± 0.00
	Improve	2.61 ± 0.92	0.04 ± 0.04	0.88 ± 0.04	**1.00** ± 0.00	0.93 ± 0.02
	Generic	**0.00** ± 0.00	**0.00** ± 0.00	**1.00** ± 0.00	**1.00** ± 0.00	**1.00** ± 0.00
	EIR (*Ours*)	1.30 ± 0.92	0.03 ± 0.03	0.96 ± 0.04	0.96 ± 0.05	0.96 ± 0.02
ASQA	Baseline	34.81	1.20	-	-	-
	Zero-shot$_{\text{LLaMA3}}$	42.83 ± 0.00	1.39 ± 0.00	0.41 ± 0.00	0.55 ± 0.00	0.47 ± 0.00
	Improve	30.09 ± 0.53	0.82 ± 0.01	0.55 ± 0.01	0.72 ± 0.01	0.62 ± 0.01
	Generic	20.92 ± 0.62	0.51 ± 0.03	0.66 ± 0.01	0.81 ± 0.01	0.72 ± 0.01
	EIR (*Ours*)	**10.16** ± 0.65	**0.23** ± 0.02	**0.82** ± 0.02	**0.89** ± 0.01	**0.85** ± 0.01
ELI5	Baseline	22.93	0.82	-	-	-
	Zero-shot$_{\text{LLaMA3}}$	3.22 ± 0.00	0.10 ± 0.00	0.91 ± 0.00	0.96 ± 0.00	0.93 ± 0.00
	Improve	3.05 ± 0.14	0.09 ± 0.01	0.90 ± 0.01	0.97 ± 0.00	0.93 ± 0.01
	Generic	2.70 ± 0.18	0.06 ± 0.01	0.91 ± 0.01	0.97 ± 0.00	0.94 ± 0.00
	EIR (*Ours*)	**0.99** ± 0.06	**0.02** ± 0.01	**0.96** ± 0.01	**0.99** ± 0.00	**0.97** ± 0.01

Table 9: Results on the quality of original answers from the datasets (BASELINE), answers from 0-shot prompting LLaMA3-8B-Instruct (ZERO-SHOT), answers refined with coarse-grained feedback (IMPROVE and GENERIC), and fine-grained feedback (EIR) using LLaMA3-8B-Instruct refinement model. Additionally, we include the results of refinement with expert human feedback on our collected data. Reported results are averages over three iterations with standard deviations. he best results are in **bold green** and the second-best results are in orange .

Dataset	Approach	TIGERScore		Error Correction		
		% Error samples (↓)	**Error score (↓)**	**Precision (↑)**	**Recall (↑)**	**F1 (↑)**
HQ^{2}A	Human feedback	1.96 ± 0.00	0.07 ± 0.01	0.90 ± 0.01	**1.00** ± 0.00	0.95 ± 0.00
	Baseline	19.61	0.63	-	-	-
	Zero-shot$_{\text{Mistral}}$	3.92 ± 0.00	0.16 ± 0.00	0.83 ± 0.00	**1.00** ± 0.00	0.91 ± 0.00
	Improve	**1.30** ± 1.85	**0.03** ± 0.05	**0.96** ± 0.05	0.96 ± 0.05	**0.96** ± 0.05
	Generic	1.96 ± 0.00	0.05 ± 0.03	0.90 ± 0.01	**1.00** ± 0.00	0.95 ± 0.00
	EIR (*Ours*)	4.57 ± 1.85	0.15 ± 0.05	0.85 ± 0.06	0.93 ± 0.09	0.88 ± 0.05
ASQA	Baseline	34.81	1.20	-	-	-
	Zero-shot$_{\text{Mistral}}$	39.35 ± 0.00	1.24 ± 0.00	0.45 ± 0.00	0.58 ± 0.00	0.51 ± 0.00
	Improve	13.53 ± 0.44	0.32 ± 0.03	0.77 ± 0.01	0.86 ± 0.01	0.81 ± 0.01
	Generic	15.85 ± 1.05	0.40 ± 0.03	0.74 ± 0.02	0.83 ± 0.01	0.78 ± 0.01
	EIR (*Ours*)	**10.72** ± 0.96	**0.23** ± 0.02	**0.81** ± 0.02	**0.90** ± 0.01	**0.85** ± 0.01
ELI5	Baseline	22.93	0.82	-	-	-
	Zero-shot$_{\text{Mistral}}$	7.91 ± 0.00	0.25 ± 0.00	0.79 ± 0.00	0.90 ± 0.00	0.84 ± 0.00
	Improve	3.22 ± 0.16	0.09 ± 0.01	0.89 ± 0.01	0.96 ± 0.01	0.93 ± 0.00
	Generic	3.81 ± 0.15	0.11 ± 0.01	0.88 ± 0.01	0.96 ± 0.01	0.92 ± 0.00
	EIR (*Ours*)	**3.02** ± 0.19	**0.08** ± 0.01	**0.90** ± 0.01	**0.97** ± 0.01	**0.94** ± 0.00

Table 10: Results on the quality of original answers from the datasets (BASELINE), answers from 0-shot prompting Mistral-7B-Instruct-v0.3 (ZERO-SHOT), answers refined with coarse-grained feedback (IMPROVE and GENERIC), and fine-grained feedback (EIR) using Mistral-7B-Instruct-v0.3 refinement model. Additionally, we include the results of refinement with expert human feedback on our collected data. Reported results are averages over three iterations with standard deviations. he best results are in **bold green** and the second-best results are in orange .

Dataset	**Comprehensiveness** ($\uparrow$)	**Preference** ($\uparrow$)
HQ^{2}A	0.70	0.31
ASQA	0.86	0.02
ELI5	0.92	0.61
Average	0.83	0.31

Table 11: Agreement of annotators on the comprehensiveness and preference of refined answers over the baseline answers from three datasets.

Aspect Question	Answer	Original Annotation	Expert Justification	Sentence-level Annotation
Completeness	What are the pros/cons of borrowing money from the bank vs paying cash when buying a house? And why would it ever be a good idea to borrow if you can pay cash and pay no interest to the bank?	The main advantage of borrowing money from a bank (a mortgage) to buy a house is that it allows you to purchase a property without having all the cash up front. This enables people to become homeowners earlier in life and potentially benefit from the appreciation of the home's value. Paying a mortgage also helps build credit history, which can be beneficial for obtaining future loans or credit cards. On the other hand, paying cash for a house has the advantage of no interest costs, which can save a substantial amount of money over time. Cash buyers may also have more negotiating power, as they are not reliant on bank approval and can close the deal quickly. In certain situations, it can be better to borrow money even if you have the cash. This is because investing the cash into diversified investments (stocks, bonds, etc.) may provide better returns than paying off the mortgage early, although it comes with some level of risk. Additionally, having cash reserves can provide financial flexibility for emergencies or future opportunities.	This answer is missing the disadvantages of getting a mortgage and paying in cash. It could mention how getting a mortgage means you will have to pay interest, as well as any mortgage fees (such as arrangement fees or valuation fees). It could also discuss how your home could be put at risk if you are unable to keep up with the repayments since a mortgage is secured against the value of your property. In terms of paying in cash, the answer could explain that this will narrow your investment portfolio, thereby missing the opportunity to invest in other assets that could increase your wealth over time. Source: (Heming,2022) https://www.moneysupermarket.com/mortgages/advantages-and-disadvantages/	1. [Incomplete] Reasons: This answer is missing [...] 2. [Incomplete] Reasons: This answer is missing [...] ... 8. [Incomplete] Reasons: This answer is missing [...]
Completeness	Why do people often say diet sodas are just as bad or worse than regular sodas?	Diet sodas are often considered just as bad or worse than regular sodas because they contain artificial sweeteners instead of sugar. These sweeteners, like aspartame or sucralose, can have negative effects on the body. For example, they may confuse the body's natural ability to regulate calorie intake by tricking it into thinking it's consuming real sugar. This can lead to overeating and weight gain. Additionally, some studies have suggested that artificial sweeteners might negatively impact gut bacteria, potentially leading to health issues. While diet sodas have fewer calories than regular sodas, they can still contribute to poor overall health due to these artificial ingredients. Therefore, many people believe they are no better than, or even worse than, regular sodas.	Overall health is a combination of nutrition and exercise, artificial ingredients do not necessarily contribute to poor health.	1. [Complete] 2. [Complete] ... 6. [Incomplete] Reasons: Overall health is a [...] 7. [Complete]

Table 12: Examples of transforming expert annotations for completeness to sentence-level annotations for training the feedback model.

F Annotation Guidelines

We have previously described our data collection setup in §3.5. 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.

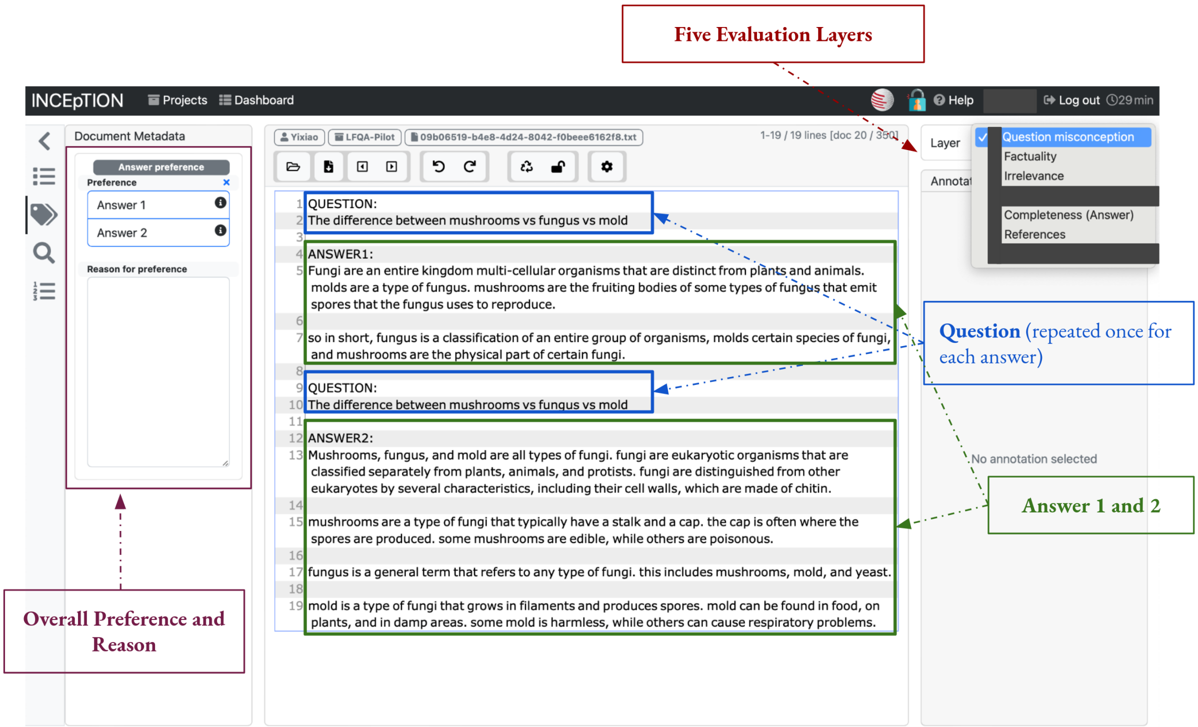

Figure 6: Screenshot of annotation interface for collecting expert error annotations on LFQA answers.

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.

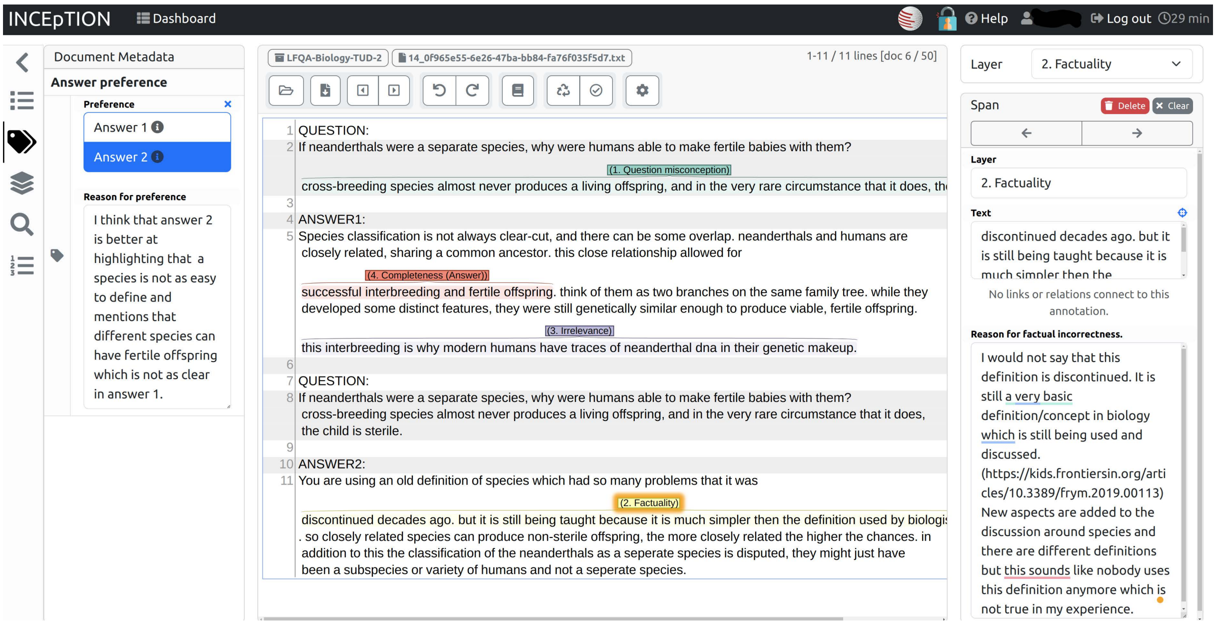

Figure 7: Screenshot of an expert annotated answer on the INCEpTION platform.

Table 13: Additional examples of expert annotated errors in long-form answers based on the defined evaluation criteria. The answers in *green* are expert-preferred answers. The overall answer preference justifications are not shown here.

Aspect	**Question**	**Answer A**	**Answer B**	**Expert Justification**
Question Misconception	How is GPS free? GPS has made a major impact on our world. How is it a free service that anyone with a phone can access? How is it profitable for companies to offer services like navigation without subscription fees or ads?	*(human)* GPS "receivers" are just radios that listen to broadcasts by specific satellites, and then use some math to calculate the location. There is no back and forth data transmission at all. So there is no capacity limit and no extra costs for more users. The us government created GPS as a military program. It would likely still be maintained even if no civilians used it, just like it was before it was opened up to the public.	*(model)* GPS is free because it's a public service provided by the US government through a network of satellites. Your phone or device receives signals from these satellites to determine its location. Companies make profits by offering additional services or applications that use GPS data, such as navigational apps or ads based on your location. These companies don't need to pay for access to GPS data because the US government maintains and funds this infrastructure.	Although GPS is a free service, made accessible to civilians for free by Ronald Reagan, America's former president at the time, it is not entirely without cost, as U.S. taxpayers fund it. Therefore, if you live and work in the U.S. and use a smartphone, you do not receive GPS on your phone for free, even though you are not directly paying for it; you are paying indirectly through taxes. On the other hand, for people living outside of the U.S., GPS might be available on their smartphones at no additional charge, but they often need to pay indirectly through internet service fees before they can use it. Source1: https://www.quora.com/How-is-GPS-sustained-as-a-free-service Source2: https://www.gps.gov/policy/funding/
Question Misconception	If movies run at 24 fps, why do games look so bad unless they run at at least 60 fps?	*(human)* U/dragoniel is right about motion blur, but the more important answer is that movies very rarely move the camera, they certainly don't spin it 180 degrees in a fraction of a second. If they did everything in the shot would be a blurry mess. A whole load of thought goes into how you get the shot you want without too much camera motion and all the bullet time stuff from the matrix was done with many still cameras taking a single photo, not moving a camera on a rocket sled.	*(model)* Movies appear smooth at 24 frames per second (fps) due to motion blur created by cameras capturing the natural blur of moving objects. Video games lack this natural motion blur as they consist of sharp, static images. To compensate, games need higher frame rates, like 60 fps, to create smoother transitions between static images, and reduce distractions caused by choppiness or sudden changes while playing. Additionally, games involve frequent interactions with controllers, requiring faster feedback to feel more responsive, which makes higher frame rates even more essential for a good gaming experience.	This is a misconception. Not all games look bad on FPS below 60. In fact, most games run smoothly between 30 and 60 FPS, although some games will not work smoothly on fps below 60. Source: https://www.avg.com/en/signal/9-ways-boost-gaming-rig
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.
Factuality	Why hasn't there been anyone to come out as a competitor to ticketmaster?	*(human)* There are lots of different ticketing companies. Dozens, possibly hundreds. Ticketmaster however is locked in to the biggest venues that feature the biggest artists, either because they have exclusive contracts or they outright own the venue. If you look at smaller venues than stadiums and arenas they often use different ticketing companies.	*(model)* Ticketmaster's dominance is due to exclusive contracts with venues and promoters, giving them control over ticket sales. Competitors struggle to break into the market because they lack access to these popular events. However, some smaller companies like eventbrite and axs have been growing, targeting niche markets and smaller venues to challenge ticketmaster's monopoly.	Ticketmaster's market share increased from 66% to 70% between 2021 and 2022. Therefore, smaller ticketing companies have so far been unable to challenge Ticketmaster's monopoly. Source: https://www.usesignhouse.com/blog/ticketmaster-stats

Continued on next page

Aspect	Question	Answer A	Answer B	Expert Justification
Relevance	How do publishers of ebooks prevent piracy?	*(human)* Some publishers try to only publish on platforms that take complete control of the distribution channels, file types, software and hardware to make it hard for people to copy the ebook without permission. Most famous of these distributors is amazon with their kindle ebook services. But others are doing the same. This does however make it harder for the customers to read the books unless they have the required hardware. And it does not completely prevent piracy. Other publishers make sure that it is easier and better service to buy the book outright instead of pirating it. There is not actually a very big market for pirated ebooks as there is for movies and TV series. So it can be hard to find pirated versions of the books you want, especially high quality versions. That makes the $3-10 for an ebook directly from the publisher a preferable deal to a lot of people. The people who do not want to pay this probably would not want to buy the book even if they did not find any pirated versions. So there is no loss in revenue to not implement any digital rights management.	*(model)* Ebook publishers use several methods to prevent piracy, which mainly involve digital rights management (DRM) and watermarking. DRM is a technology used to restrict access to and usage of copyrighted material, such as ebooks. In simple terms, DRM is like a lock that prevents unauthorized copying and sharing of ebooks. When you purchase an ebook, you're usually given a key (in the form of an account or device authorization) to unlock and read the book. This prevents users from easily sharing the ebook file with others, as they will not have the necessary key to open it. Some popular drm systems used for ebooks include adobe's adept and amazon's kindle drm. Watermarking is another technique ebook publishers use to deter piracy. Instead of locking the content like drm, a digital watermark embeds identifiable information into the ebook file. This information can be visible (like a logo or text) or invisible (embedded in the file's data). [...]	This does not tell us how publishers prevent piracy; it only provides a scenario explaining why piracy would not occur in books as it would in visual media.
Relevance	Why do companies like Meta have so many employees, what do they do?	*(human)* In my experience software companies lean heavily on product and project managers. To stay competitive in tech you have to constantly be trying to innovate (at least that's the idea), which takes a lot of people to come up with and plan out these projects, even if many never see the light of day. Beyond that, maintaining a tech infrastructure like that is very complex,, goes well beyond devs. [...] Facebook also has a huge footprint in the open source community, in that they actively maintain many sophisticated projects they built for their products that are now used by millions of other products, like their react frameworks. Not to mention they heavily use other open source softwares and dedicate some resources to maintaining those as well [...] Successful tech companies are incredibly convoluted and complex in my experience. Many times there are whole teams working on "nothing" because their part of the business is on its way out, or they were part of another company that was acquired, and in either case, someone in the middle hasn't gotten around to trimming the fat/reallocating those resources. Source: have worked for a faang company, other huge tech companies, now work as a private contractor bc of how ** annoying these conglomerates are to work with.	*(model)* Meta, formerly known as Facebook, is a massive tech company that requires a large number of employees to manage and maintain its various services, products, and operations. [...] let's break it down into main areas:1. Software development: Meta operates numerous services and applications, including Facebook, Instagram, Whatsapp, and the oculus virtual reality platform. Each of these services requires teams of software engineers, product managers, and designers to develop, maintain, and improve the platforms. [...] 2. Infrastructure and operations: with billions of active users, Meta must maintain and develop a vast infrastructure of servers, networks, and data centers to handle the enormous amount of data being generated, processed, and stored. [...] 3. Research and development: companies like meta are always innovating and creating new technologies. They require extensive r&d teams that include experts in fields like artificial intelligence, virtual reality, and machine learning. These teams drive the development of cutting-edge technologies and solutions for Meta's products. [...] 4. Sales and marketing: meta generates a significant portion of its revenue from advertising. [...] 5. User support and moderation: with billions of users worldwide, [...]	Adds nothing to the explanation is a rant by the person.

Continued on next page

Aspect	Question	Answer A	Answer B	Expert Justification
Completeness	How do pharmacies work? Do they just have every kind of medication at all time? How is a prescription ready within an hour? Eta thank you everyone for taking the time to respond! I know it seems super obvious as to how it works, this was a late night thought I had and needed to know?	*(human)* General yes, pharmacies will keep enough of the more commonly requested medications in stock so that most prescriptions can be filled the same day. When that's not possible, depending on various factors, many pharmacies may have multiple deliveries from suppliers each week, making it possible to order anything that isn't in stock within a few days if it's available from the supplier.	*(model)* Pharmacies keep a stock of commonly prescribed medications on hand. When you provide your prescription, the pharmacist checks their inventory, measures and packages the drug, and puts the proper labeling and instructions on it. If the medication isn't in stock, they can usually order it and have it ready in a short time. Efficient organization and well-trained staff help facilitate the quick turnaround for prescriptions.	The answer should specify factors that may affect the availability of medications, such as manufacturing issues, disruptions in supply, and stockouts of essential medicines (World Health Organization). For more information, you can visit: https://www.who.int/our-work/access-to-medicines-and-health-products
Completeness	What are the pros/cons of borrowing money from the bank vs paying cash when buying a house? And why would it ever be a good idea to borrow if you can pay cash and pay no interest to the bank?	*(human)* If you can make more money doing something else with your cash than using it to pay down your mortgage, then it's beneficial to take out a loan. It depends on your credit rating what kind of interest you'll get on your loan, and what the long term stock market trends are, or if you want to buy multiple properties at the same time, for example. But making money in the stock market is not a given, so there is risk involved in going that route. Also, you want to have cash on hand for possible emergency situations, like big unexpected medical bills, your car dies, you lose your job, etc. On the other hand, if you lose your job and you have a mortgage, then you might end up defaulting on your loan, and lose the home and any of the equity you had in it.	*(model)* The main advantage of borrowing money from a bank (a mortgage) to buy a house is that it allows you to purchase a property without having all the cash up front. This enables people to become homeowners earlier in life and potentially benefit from the appreciation of the home's value. Paying a mortgage also helps build credit history, which can be beneficial for obtaining future loans or credit cards. On the other hand, paying cash for a house has the advantage of no interest costs, which can save a substantial amount of money over time. Cash buyers may also have more negotiating power, as they are not reliant on bank approval and can close the deal quickly. In certain situations, it can be better to borrow money even if you have the cash. This is because investing the cash into diversified investments (stocks, bonds, etc.) may provide better returns than paying off the mortgage early, although it comes with some level of risk. Additionally, having cash reserves can provide financial flexibility for emergencies or future opportunities.	This answer is missing the disadvantages of getting a mortgage and paying in cash. It could mention how getting a mortgage means you will have to pay interest, as well as any mortgage fees (such as arrangement fees or valuation fees). It could also discuss how your home could be put at risk if you are unable to keep up with the repayments since a mortgage is secured against the value of your property. In terms of paying in cash, the answer could explain that this will narrow your investment portfolio, thereby missing the opportunity to invest in other assets that could increase your wealth over time. Source: (Heming,2022) https://www.moneysupermarket.com/mortgages/advantages-and-disadvantages/
References	Why can't spam call centers be automatically shut down? Additionally, why can't spam calls be automatically blocked, and why is nobody really doing a whole lot about it? It seems like this is a problem that they would have come up with a solution for by now.	*(human)* Most of the spam call centers originate in countries where the governments don't _really_ care all that much. For example, a good number of them are in smaller cities in India, where the politicians are willing to turn a blind eye for the right price (and so long as they don't go around scamming Indians). From there, the call centers are just another overseas call. There are a number of _legitimate_ call centers in foreign countries, so looking to block the scam ones is a needle in a stack of needles. Its a ton of effort for the phone company with little actual benefit to them.	*(model)* Spam call centers can't be automatically shut down because they are often located in countries with weak regulation and they use advanced techniques to hide their identity, like caller id spoofing. Blocking spam calls is challenging because these callers use ever-changing phone numbers, making it difficult to maintain an updated block list. While there are some efforts to combat spam calls, such as telecom companies implementing call-filtering systems and governments enacting regulations, the issue is deeply rooted in the economic incentive for spam callers and the complexities in tracking them down. Ultimately, tackling spam calls requires international cooperation and continuous technology advancements.	The example is not helpful in addressing the query. It is a controversial example that will stir up the emotions of readers, and therefore distract them.

Table 14: Examples of answers refined using our Error-informed Refinement approach. We show the feedback from the error-feedback model and the corresponding refined answers from the refinement model incorporating the feedback. The highlighted portion in the refined answer depicts the feedback incorporated in the answer.

Question	**Answer**	**Feedback**	**Refined Answer**
[ELI5] Why do cars sound different in reverse?	Reverse gears often are straight-cut instead of cross-cut because they get used far less and are cheaper to make.[...] The gears on the top are cross-cut and make far less noise and last a lot longer. Straight cut makes more noise because the teeth sort of slap against each other as the gears mesh. Cross cut are quiet because the more gently mesh with each other over a diagonal.	Consistency score: 1.0 Prediction: 1. [Incomplete] Reasons: This is correct but could be improved by mentioning that straight-cut gears are louder because there is more friction between the teeth when they meet. This causes more heat buildup and wear on the teeth, which contributes to the noise. [...] 2. [Complete] 3. [Complete] 4. [Complete] 5. [Complete]	Cars can sound different in reverse due to the type of gears used in the transmission. Reverse gears are often straight-cut instead of cross-cut because they are used less frequently and are therefore cheaper to manufacture. Straight-cut gears have teeth that meet end-to-end, causing more friction and heat buildup, which results in a louder noise. In contrast, cross-cut gears have teeth that mesh at an angle, reducing the amount of friction and heat generated, resulting in a quieter noise. Additionally, straight-cut gears tend to have more play between the teeth, which can also contribute to the noise. To hear the difference for yourself, you can listen to a video of a car reversing with both straight-cut and cross-cut gears.[...]
[ELI5] How have they brought air inside the International Space Station? In liquid form?	In liquid form? In a way, yeah. Usually in the form of water, which is safe to transport, doesn't have to be pressurized and you only have to vent the excess hydrogen.	Consistency score: 1.0 1. [Incomplete] Reasons: The answer does not completely describe how this is possible. It mentions that water is used as a source of oxygen, but it fails to explain how this works. Additionally, it does not address the issue of bringing air into the ISS from an external source. 2. [Complete] 3. [Complete]	To bring air inside the International Space Station (ISS), the astronauts use a process called öxygen generation,ẅhere they convert carbon dioxide and other gases present in the space station's atmosphere into breathable oxygen. This process involves several steps: 1. Carbon Dioxide Removal: [...] 2. Oxygen Generation: [...] 3. Air Revitalization: [...] 4. Humidification: [...] 5. Filtering: [...] In terms of bringing air into the ISS from an external source, there are actually two methods used: 1. Atmospheric Replenishment: [...] 2. Cargo Vehicle Supply: [...] So, to summarize, the ISS has systems in place to generate oxygen from carbon dioxide and other gases present in the space station's atmosphere, as well as methods for bringing in air from external sources when necessary.
[ELI5] Why are chairs slightly incurved instead of the other way around?	Some busses and trains in Sweden actually have chairs that make you sit with a straight back, but i think It's cause people find those types of chairs more comfortable.	Consistency score: 1.0 Prediction: 1. [Incomplete] Reasons: The answer does not explain why most chairs are slightly inclined.	Chairs are designed with a slight incline rather than being perfectly straight for several reasons. Firstly, an angled seat allows users to maintain better posture while sitting. When the backrest is at a slight angle, it supports the natural curvature of the spine, promoting good posture and reducing strain on the back and neck muscles. This can help prevent long-term health issues associated with poor sitting positions, such as lower back pain and herniated discs. Additionally, the inclination of the chair helps to distribute weight evenly across the seat and backrest, reducing pressure points and providing a more comfortable sitting experience. This is particularly important for people who spend extended periods of time sitting, such as office workers or students. While some buses and trains in Sweden may have straight-backed seats, this design is not typical of most chairs found in everyday use. In fact, many ergonomic experts recommend that chairs be designed with a slight incline to promote proper posture and reduce discomfort.[...]

Continued on next page

Question	Answer	Feedback	Refined Answer
[ASQA] When did the first cellular phone come out?	The first handheld cellular mobile phone was demonstrated by John F. Mitchell and Martin Cooper of Motorola in 1973. In 1983, the DynaTAC 8000x was the first commercially available handheld mobile phone. The phone had a talk time of just thirty minutes and took ten hours to charge.	Consistency score: 0.55 Prediction: 1. [Incomplete] Reasons: This is correct but it would be useful to mention that this demonstration didn't lead to the development of an actual product. Instead, it showed the potential of such a device. 2. [Complete] 3. [Complete]	The first handheld cellular mobile phone was demonstrated by John F. Mitchell and Martin Cooper of Motorola in 1973. This demonstration showcased the potential of a portable device that could communicate wirelessly over long distances, revolutionizing the way people communicated while on-the-go. However, it wasn't until 1983 that the first commercially available handheld mobile phone, the DynaTAC 8000x, became available to consumers. With a talk time of just thirty minutes and a charging time of ten hours, this early model paved the way for future advancements in mobile technology. Today, smartphones have become an essential part of modern life, offering a wide range of features and capabilities beyond simply making calls and sending texts.
[ASQA] Where is the danube river located in europe?	The Danube river is the second longest river in Europe, running through Central and Eastern Europe. Starting in Germany the river flows through Austria, Slovakia, Hungary, Croatia, Serbia, Romania, Bulgaria, Moldova and Ukraine before draining into the Black Sea. The river flows through many cities including Pančevo, Drobeta-Turnu Severin, Bratislava, Ruse, Novi Sad, Ulm, Vienna, Zemun, Vidin, Regensburg, Ingolstadt, Linz, Győr, Dunaújváros, Smederevo, Izmail, Belgrade, Apatin, Galați, Vukovar and Budapest.	Consistency score: 0.70 Prediction: 1. [Complete] 2. [Complete] 3. [Incomplete] Reasons: These are just names of cities; it would be useful to know where they are located (which countries) so you can understand how the river flows through different nations.	The Danube River is the second longest river in Europe, flowing through Central and Eastern Europe for a total length of approximately 2,857 kilometers (1,777 miles). It originates in the Black Forest region of Germany, then passes through or along the borders of several other European countries, including Austria, Slovakia, Hungary, Croatia, Serbia, Romania, Bulgaria, Moldova, and Ukraine before emptying into the Black Sea. Along its journey, the Danube River flows through numerous cities and towns, some of which include: * In Germany: Ulm, Ingolstadt, and Regensburg * In Austria: Linz and Vienna * In Slovakia: Bratislava * In Hungary: Budapest, Gyor, and Szeged * In Croatia: Vukovar and Osijek * In Serbia: Belgrade and Smederevo * In Romania: Bucharest and Galati * In Bulgaria: Ruse and Vidin * In Moldova: Kishinev * In Ukraine: Odessa and Reni Each of these cities has a unique history and culture, [...].
[HQ^{2}A] How did human trafficking become a thing? Why do people kidnap and sell others?	Human trafficking became a thing likely in pre-history and hasn't stopped at all. There is always a need for people to do something and being sent from one place to another. Various forms of slavery is obviously a prominent activity and has been present in many societies since pre-history and still exists today.	Consistency score: 0.55 Prediction: 1. [Incomplete] Reasons: This is correct; it mentions that human trafficking has occurred since prehistoric times and continues today. However, it would be useful to trace the historical emergence of modern slave trade and human trafficking. For example, it could explain how during the transatlantic slave trade, African slaves were transported across the Atlantic Ocean to the Americas (https://www.britannica.com/topic/transatlantic-slave-trade). This would help to provide context and give further insight into the history of human trafficking. 2. [Complete] 3. [Complete]	Human trafficking, unfortunately, has a long and complex history that dates back to pre-historic times. The practice of capturing and enslaving individuals for various purposes, including forced labor and sexual exploitation, can be traced back to ancient civilizations such as Greece, Rome, and Egypt. However, the modern form of human trafficking, which involves the use of force, fraud, or coercion to exploit people for labor or commercial sex, gained momentum during the transatlantic slave trade era. During this period, millions of Africans were forcibly taken from their homes and sold into slavery in the Americas, where they were subjected to brutal working conditions and treated as property rather than human beings. This tragic chapter in human history laid the foundation for the modern human trafficking industry, which has evolved over time to include a wide range of activities, such as child labor, domestic servitude, [...] So, why do people engage in human trafficking? The reasons are complex and varied, but some common motivations include financial gain, power and control, and a lack of empathy for the victims. Traffickers often target vulnerable individuals, such as those living in poverty or fleeing conflict, and use manipulation, coercion, and violence to keep them [...]

Continued on next page

Question	Answer	Feedback	Refined Answer
[HQ^{2}A] How are e-fits constructed in the pursuit of a suspect?	E-fits, or electronic facial composites, are digital images created to help identify suspects in criminal investigations. They are constructed by combining different facial features based on the eyewitness's description of the suspect. The process starts with an interview between the eyewitness and a trained police officer or forensic artist. The officer uses specialized software, which contains a database of various facial features, such as eyes, noses, mouths, and hairstyles. The eyewitness selects the features that best match their memory of the suspect's appearance. Once the initial facial composite is created, the officer and the eyewitness work together to refine the image. They may adjust the size, shape, and position of features to improve accuracy. The final e-fit is then used to aid police investigations, typically by sharing with the public and other law enforcement agencies, in the hopes of identifying the suspect.	Consistency score: 1.0 Prediction: 1. [Complete] 2. [Complete] 3. [Complete] 4. [Complete] 5. [Complete] 6. [Complete] 7. [Complete] 8. [Incomplete] Reasons: It would be helpful if this section elaborated more on how e-fits are shared and why it is beneficial to do so.	E-fits, or electronic facial composites, are digital images created to help identify suspects in criminal investigations. They are constructed by combining different facial features based on the eyewitness's description of the suspect. The process starts with an interview between the eyewitness and a trained police officer or forensic artist. The officer uses specialized software, which contains a database of various facial features, such as eyes, noses, mouths, and hairstyles. The eyewitness selects the features that best match their memory of the suspect's appearance. Once the initial facial composite is created, the officer and the eyewitness work together to refine the image. They may adjust the size, shape, and position of features to improve accuracy. Once the e-fit is complete, it can be shared with the public and other law enforcement agencies through various channels, such as social media, local news outlets, and online crime reporting platforms. Sharing the e-fit helps to increase the chances of identifying the suspect, as it allows a larger pool of people to view and recognize the individual. Additionally, sharing the e-fit can also serve as a deterrent to the suspect, as they may become aware that the authorities are actively searching for them. In addition to being shared with the public, e-fits can also be distributed to other law enforcement agencies, such as federal authorities, neighboring jurisdictions, and international partners. This helps to expand the search area and increase the likelihood of identifying the suspect. Furthermore, e-fits can be updated and revised as new information becomes available, allowing the investigation to continue until the suspect is apprehended. [...]