Task type	Topic-based summarization train	 test	Other train	 test
question generation	2221	264	9011	1089
base	6235	770	43285	5250
critiqueability	31279	3983	55042	6988
critique	15277	1944	19194	2532
helpfulness	41724	5096	0	0
refinement	14323	1823	19194	2532
corroboration	0	0	42058	5273
corroboration quotes	6235	770	0	0
critique quotes	14234	1814	0	0

Table 4: Number of tasks of each type in our training and test sets, split by topic-based summarization and other (a mix of question-answering and summarization tasks). During training, 50% of the refinement tasks are converted to direct refinement tasks, and 50% of the corroboration quotes are converted to "answer quotes"

A Additional dataset details

A.1 Labelers

Our labelers are contractors hired by OpenAI and paid hourly. Labelers are fluent in English and the majority are native speakers. We communicate with our labelers via Slack, where we send instructions, gather feedback, and discuss tasks.

We occasionally check labeler quality using a variety of techniques: looking at critique likelihood (by other labelers) of their demonstrations, looking at agreement rates on rankings (we generally share 5% of tasks between 10 labelers).

A.2 Collection details

We collect data in three rounds, roughly corresponding to the base task, the critique task, and the helpfulness task. Thus we have three distinct web interfaces for data collection, each of which went through multiple iterations throughout the project.

A.2.1 Base task

When collecting data for the base task, we show labelers a passage and ask them to come up with a number of questions, as well as answers to each question. For topic-based summarization, we ask them to have at least one question for which there is no relevant information in the passage and the answer is trivial. Some variants:

1. We sometimes also collected misleading answers that should be clearly wrong, but take readers a long time to determine as wrong. We asked for labelers to aim to have answers with different kinds of flaws, e.g. accuracy flaws contradicting part of the text that are hard to find or not stated explicitly and coverage flaws leaving out important details that are easy to overlook. We also ask labelers to aim for the flaws to be severe. Finally, labelers wrote critiques of the misleading answer (typically only one, as per the initial requirement that it be hard to spot a flaw).
2. We sometimes asked for lists of "quote corroborations". For each quote corroboration, the labeler highlighted a set of spans in the answer, and a set of corroborating spans in the passage

A.2.2 Critique task

When collecting data for the critique task, we show labelers a passage, multiple questions about the passage, and multiple model-generated answers for each question.

We always ask for a Likert rating for each answer and a ranking of the answers.

Critiques We then ask for a series of critiques for each answer, roughly in descending order of importance or severity. Critiques are instructed to be relatively atomic, so they should not point out multiple unrelated issues. We also asked for critiques to be as specific as possible, avoiding broad critiques like "This answer could be better".

Each critique was given a severity, one of "Minor", "Medium", "Major" and "Critical", each intended to be about twice as severe as the previous. Labelers were able to skip critiquing answers that were very similar to another answer.

Refinements When we collected refinements, it was done so jointly with critiques, with a corresponding refinement for each critique. Some answers were too poor to be meaningfully refined, in which case labelers marked the answer to be completely rewritten instead.

Since we collect multiple critiques, we collect a series of refinements as well, with each refinement being an improvement over the previous refinement (or the original answer). All critiques were expected to apply to the original answer as well as the refinement. (Early on, we had them mark for each critique whether it applied, but we abandoned this later.)

Note that this means that for training, all refinement demonstrations were using human-written critiques for input. Furthermore, refinement demonstrations are of model-written answers about half the time, and on (partially) human-written refinements the other half.

Critiqueability In collecting critiques, we are also implicitly collecting critiqueability labels. We assume the original answer to be uncritiqueable if and only if no critique is given. We enforce that there are critiques whenever Likert rating is below a 7. Similarly, when refining, the final refinement is assumed to be uncritiqueable, and all previous refinements are assumed to be critiqueable.

Variants in data collection that we we explored throughout the project:

1. Collecting a set of "corroborations" for each answer, of natural language explanations that support the answer.
2. No refinements
3. For topic-based summarization, we asked for a category for each critique, one of:
    - Coverage: summary missing relevant information from passage
    - Accuracy: summary giving incorrect information
    - Coherence: summary is poorly written, confusing or nonsensical
    - Other: a catch-all bucket for everything else
4. For topic-based summarization, we also explored collecting quotes. For each critique, we asked them to give "passage highlights", required for Coverage critiques, and "answer highlights", required for Accuracy and Coherence critiques. The answer highlights would be spans in either the original answer or a refinement being critiqued.

A.2.3 Helpfulness task

When collecting data for the helpfulness task, we show labelers a passage, multiple questions about the passage, and one model-generated answer for each question. We then generate between 8 and 16 model critiques per answer.

For each answer, if no model critiques are helpful, we ask labelers to judge whether there exist any helpful critiques. If some model critiques are helpful, we ask if the labeler has a substantively different and better critique. In either case, they may choose to write a new critique, and mark its severity and category.

We also asked labelers to rank the helpful critiques, though we did not end up using this data.

Variants we explored:

1. We sometimes asked labelers to mark when critiques were "clearly helpful", meaning they were unambiguously helpful rather than nit-picky.
2. We sometimes asked labelers to mark severity and category of all model-generated critiques marked as helpful.

A.3 Base tasks

Early in the project, we asked labelers to create question-answering and summarization tasks. However, we later switched to topic-based summarization and used that for the majority of the project. As noted, our results are reported on topic-based summarization tasks only. However, we left the original question-answering and summarization tasks in the training set.

For topic-based summarization, we asked that the topics be chosen such that writing summaries required more than keyword searching the article. We also asked that the topics require including some significant information that would not be included in a non-topical paragraph-long summary of the original passage.

A.4 Auxiliary tasks

Based on the various data we collected throughout the project, we included a number of auxiliary tasks in the supervised training data. Aside from those mentioned in Table 1, the ones which were included in the final models were:

1. **Question creation** Our labelers were asked to write 1-8 questions based on a passage and give demonstrations of answers to those questions (topic-based summarization or otherwise) at the same time. During model training, we include the auxiliary task of creating a slate of questions given a passage.
2. **Corroborations** We explored collecting corroborations of answers, which explain why aspects of an answer are correct. In general, it is most interesting to critique things that are explanation-like, as opposed to short answers with no explanation (e.g. a mathematical proof rather than just a statement). With topic-based summarization, this was less important, as the answers are somewhat self-explanatory, simplifying our research setup.
3. **Corroboration quotes** We include the task of retrieving relevant quotes from the passage which corroborate an answer. We also include a variant which conditions on the span of the answer being corroborated.
4. **Question quotes** We include the task of retrieving relevant quotes from the passage, based only on the question.

At training time, we sometimes convert between various tasks as a form of data augmentation. The most important one is that we convert conditional refinement tasks to direct refinement tasks 50% of the time. We also convert corroboration quotes to question quotes 50% of the time.

We also experimented with various other tasks which were used in some models during the project, but were removed for the final version. For example, we experimented with the task of generating a slate of critiques, rather than a single critique. This has the benefit that during assistance, the model might be less inclined to produce duplicates. However, we switched to writing single critiques to simplify our setup.

A.5 Formatting details

We use a simple formatting scheme which always orders things as: passage, question, answer, critiqueability, critique, helpfulness, refinement.

For example, critique tasks look like

```
{passage}

Question: {question}

Answer: {answer}

Are there any critiques for the above answer? If so, write one
{binary critiqueability}
{critique}
```