Ask and Ye Shall Seek: Generating Clarifying Questions in Closed-Domain Multi-turn Dialogues

Aman Bansal
amanbansal@umass.edu

Rahul Seetharaman
rseetharaman@umass.edu

Rohan Lekhwani
rlekhwani@umass.edu

Shreeya Patil
shreeyadeepa@umass.edu

1 Introduction

With the advent of large language models (LLMs) like GPT3.5, BARD and YouChat among many others, people have become more accustomed to ask open-domain questions through these interfaces instead of resorting to a typical search engine. During their conversations with these agents, it has been found that users' behavior towards them is similar to their behavior towards other humans (Nass and Moon, 2000). Users tend to exhibit overlearned social behaviors such as politeness and reciprocity toward computers. This similarity leads to user queries which might sometimes be ambiguous or incomplete.

Another stark difference between using a search engine for question-answering (QA) against a language model is that the latter is limited to returning only a single answer in response to a query instead of a diverse set of results. For current LLMs it becomes imperative to get it right in their first attempt to a user query. Asking clarifying questions before providing a final response has shown users to be more forgiving of a blunder made by an agent. *The main problem to solve is what kind of clarifying questions help reduce ambiguity in the user's query to reach the final answer in minimum effort.*

The recent advancements in NLP with the Transformers architecture and transfer learning, have brought about significant improvements in performance of various downstream tasks. While there has been considerable work on adapting latest techniques for question answering, current LLMs do not always ask the *correct* clarifying questions. Secondly, there are *few datasets* with sufficient context and user turns to facilitate more research in this area. Thirdly, while the most popular LLMs like GPT3 are closed-source, the capability of *open-sourced language models* for the task has not yet been evaluated to the best of our knowledge.

In this project, we experimented with multiple strategies to ask clarifying questions to ambiguous user queries. We synthetically augment a dataset and fine-tune 7 *open-source language models* with changes to their training objectives. In addition, we identified *four* qualitative metrics that a clarification question can be evaluated upon. A substantial number of evaluations including *human and automated* evaluations for generations on the test set have been performed.

2 Proposed Outcomes

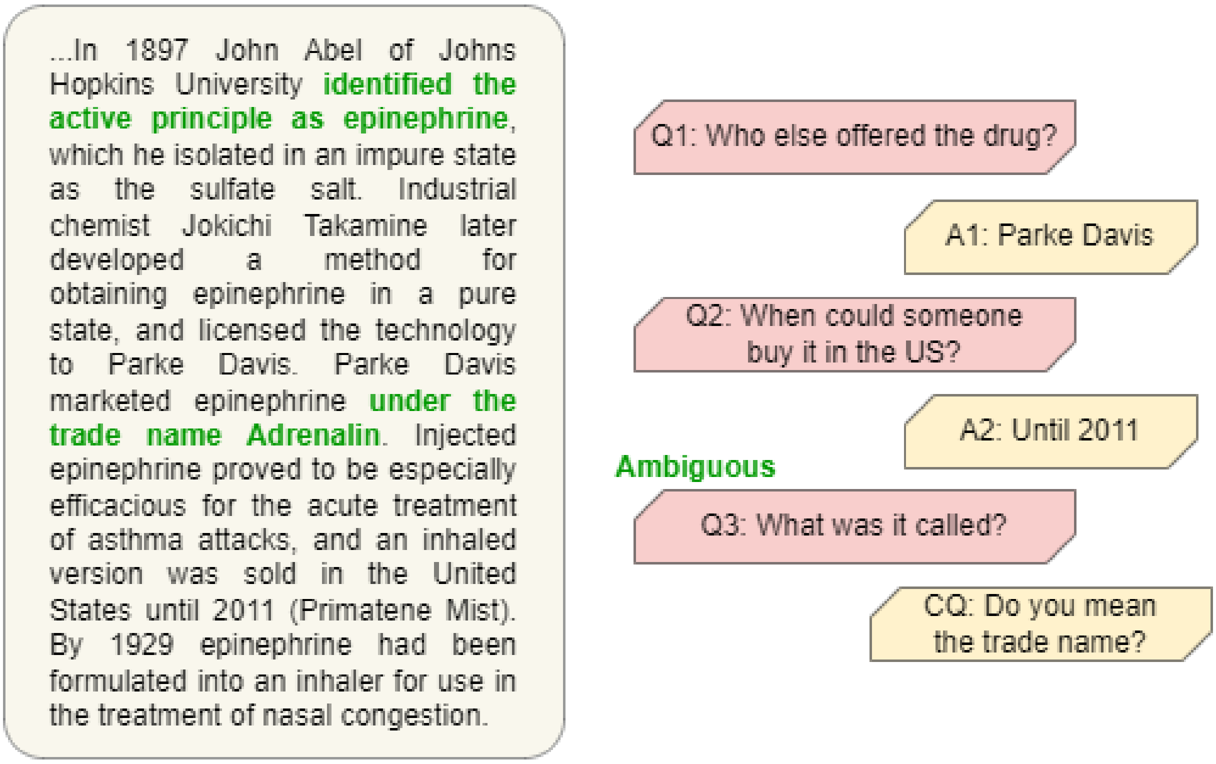
The following list represents the exhaustive list of proposed outcomes from both the initial proposal and the current report.

- ~~Multimodal Domain-Specific Dataset~~ → Text-based closed-domain dataset: Unavailability of domain-specific long-form data with correlated images. Scraping Wikipedia was not a robust solution. Additionally, we realized that training a model on image data would require much more compute resources. Based on instructor suggestions, we therefore created a closed-domain text-only dataset.

- Fine-tuning new models: Instead of only replicating the two baselines as proposed earlier, we also fine-tuned 5 new open-sourced models in addition to those, totalling to evaluations on 7 models.

- Evaluations: We perform human evaluations of 560 clarifying questions for 40 passages. We also perform GPT evaluation of 1722

The source code, annotations and evaluations can be found at https://github.com/rahulsee/ask-and-ye-shall-seek.git



...In 1897 John Abel of Johns Hopkins University identified the active principle as epinephrine, which he isolated in an impure state as the sulfate salt. Industrial chemist Jokichi Takamine later developed a method for obtaining epinephrine in a pure state, and licensed the technology to Parke Davis. Parke Davis marketed epinephrine under the trade name Adrenalin. Injected epinephrine proved to be especially efficacious for the acute treatment of asthma attacks, and an inhaled version was sold in the United States until 2011 (Primatene Mist). By 1929 epinephrine had been formulated into an inhaler for use in the treatment of nasal congestion.

Q1: Who else offered the drug?

A1: Parke Davis

Q2: When could someone buy it in the US?

A2: Until 2011

Q3: What was it called?

CQ: Do you mean the trade name?

Figure 1: A conversation with a story, multiple history turns, an ambiguous question and a clarifying question.

clarifying questions across 123 passages. In addition, we make use of popular metrics for the task to further evaluate our results.

- ~~Contextual Digression Handling~~: Based on instructor and TA recommendations we concluded that the adversarial approach proposed initially would require more clarity in terms of how to avoid topic digressions and can be explored in the future.

- Error Analysis: We perform an analysis on what common flaws are exhibited by trained models and what kind of data they occur for.

3 Related work

3.1 Open Domain Clarification

Information-seeking always involves some kind of interaction between a user and an agent. Typically, user queries are ambiguous and users do not give all the information they need to an agent in one go. This elicits the need for clarifications (Zamani et al., 2020). Previous work on mitigating ambiguities in clarifications in an open-domain setting involved diversifying results, identifying multiple possible user intents, and presenting the results accordingly. In line with this research, several datasets and approaches have been proposed. (Aliannejadi et al., 2020) propose a dataset - ClariQ, which consists of users' ambiguous queries, clarification questions, and clarification answers. In Qulac, (Aliannejadi et al., 2019) proposed a dataset, which consists of user ambiguous user queries and their clarifications in 3 to 4 facets on average. Our work is different from the ones mentioned above, in that it proposes methods to tackle ambiguity when the agent is given the context, in this case, a story, and strives to generate clarifications that strictly adhere to the context, the conversational history turns, and the user's current ambiguous question.

(Wang and Li, 2021) explore clarification question generation as a template ranking and slot filling. They use a Guided Transformer architecture with a multi-task objective of ranking a set of question templates and a simple classification head that can predict the vocabulary given the user query, and search engine page results. Approaches that use templates and slot filling are limiting in that they do not allow the model to generate creative text, in this case, clarifications. Although these approaches work really well on automatic metrics including accuracy, exact Match, precision and recall, they produce generations that are dull and lacking in engagement. Moreover, they use encoder-only models and do not fully utilize the text generation capabilities of encoder-decoder or decoder-only language models.

(Shao et al., 2022) explore the use of a self-supervised training approach to generate clarification questions on the CLAQUA dataset (Xu et al., 2019).

3.2 Multimodal clarification

In their work (White et al., 2021) explore the possibility of ambiguous questions in a multi-modal setting. They make use of image captioning datasets and use a generative model to ask questions for which the answer is typically a yes/no. The yes/no answer is given by the user. The final task of the model is to pick one image from a list of possible images given all the clarification questions and their yes/no user-provided answers.

3.3 Question generation in closed domain settings

(Du et al., 2017) explore the task of ”Learning To Ask” a paradigm where a neural seqeunce to sequence text generation model learns to ask questions based off of a reading comprehension passage. Similarly, (Rathod et al., 2022) explore multi question generation, a task that aims at generating lexically diverse but semantically similar questions.

3.4 Clarifications in closed domain settings

Answering ambiguity in a closed domain setting requires a language model to understand the context, the user turns, and ambiguous questions. There is a need to steer clear of hallucinations and generate coherent, fluent text that is extremely relevant to the context. In their work, (Guo et al., 2021) propose ABG-CoQA, a variant of the CoQA dataset, that has ambiguous questions. Specifically, the dataset consists of ambiguous questions, conversation history, and also non-ambiguous question turns, which are accordingly assigned binary labels to indicate their ambiguity. In our work, we use only ambiguous questions, which is a small subset of their dataset. In their paper, baselines are run on BART, the code for which has not been made publicly available. We replicate their baselines in addition to evaluating other larger models.

4 Dataset

For our experiments we have used Abg-COQA (Guo et al., 2021) dataset. This dataset exhibits several key characteristics that make it a valuable resource for research and evaluation. These characteristics are as follows:

- Rich Passage and Question Set: This dataset comprises of 3968 passages extracted from five distinct domains out of which 741 are ambiguous.
- Comprehensive Ambiguity Coverage: Abg-CoQA encompasses four distinct types of ambiguity. In most instances, the ambiguity becomes apparent only upon considering the conversation history and investigating all potential answers within the story. This aspect reflects the realistic nature of the dataset, as it captures the intricacies of conversational contexts where ambiguity often emerges.
- Clarification Questions and Multiple Possible Replies: To address the inherent ambiguity in the dataset, each ambiguous question in Abg-CoQA is accompanied by a clarification question and several possible replies. These clarification questions and potential replies serve the purpose of elucidating the intended meaning of the originally ambiguous question. Moreover, the provision of multiple possible replies emphasizes the diverse range of answers that can arise from the initially ambiguous question.

4.1 Data Preprocessing

There was some prepossessing required to format the data so that it can be fed to different models like BART and flan-t5. The preprocessing can be accessed within the *get_data.py* file at the root of the repository. Evaluation with GPT3.5 required constructing a single string comprising of the story, history turns and the ambiguous questions as the features and the clarifying question predicted by a model as the label.

4.2 Data Augmentation

In the original dataset (Guo et al., 2021), we encountered a substantial disparity in the distribution of instances. While we had over 5727 instances of non-ambiguous questions that required no clarifying question, we found only approximately 741

Model	Grammatical Correctness	Diversity	Clarity Index	Relevance
bart data aug	3.008	2.471	2.760	**3.424**
bart no data aug	2.876	3.170	2.943	2.75
t5_kldiv	3.213	**3.804**	**2.991**	2.983
t5_canard	3.095	3.168	2.920	2.691
ft5-base	2.930	2.930	2.808	2.682
ft5-small	3.369	3.215	2.955	2.727
ft5-large	**3.556**	3.341	3.548	2.752
ft5-xl	3.406	3.182	2.939	3.365

Table 1: The table represents the averaged GPT3.5 generated score for different models which forms a part of GPTEval.

instances of ambiguous and clarifying questions. This significant imbalance posed a challenge in effectively fine-tuning a model that could handle the complexities of ambiguity.

Methodology

GPT3,5, a powerful language model to generate additional ambiguous and clarifying questions is capable of synthetic data generation given an appropriate prompt. To achieve this, we employed two prominent techniques: Zero-shot and Few-shot prompting. These techniques allowed us to leverage the capabilities of the GPT3,5 language model to generate high-quality questions. We ensured the quality of the generated questions was good by sampling and manually checking the quality of 1 out of every 20 clarifying questions.

For zero shot generation, we utilized a specific prompt that guided the model's behavior. The prompt instructed the model to generate three clarifying questions based on the answers to the ambiguous question, drawing from the information provided in the given passage. The objective of incorporating the few-shot prompting technique was to leverage a limited number of examples to enhance the model's ability to generate improved clarifying questions. This approach aimed to capitalize on the model's capacity to generalize from a small set of instances and produce high-quality questions.

To implement the few-shot prompting technique, we formulated a specific prompt for GPT3.5. The prompt was structured as follows: ”As shown in the following example, you are given a story, question, and a few potential answers to the question. Generate three clarifying questions for the answers to the ambiguous question based on the given passage.” This prompt aimed to guide the model's behavior by providing the necessary context and instructing it to generate three clarifying questions based on the given passage and the potential answers provided.

The following table summarizes the size of the original dataset (training) along with the augmented ones.

Setting	# Clarifying Questions
Original	741
Zero-Shot	2223
Few-Shot	2223

Ensuring Diversity

While generating new questions using GPT3.5 works well, attention needs to be paid whether the model incorrectly generates questions based on the template provided to it. LLMs are prone to exhibiting patterns in their data generation behavior. Few shot examples starting with ”What” have a high probability of generating similar ”What”-prefixed questions.

To ensure, high quality and diverse few shot examples were provided, we incorporated cosine similarity along with the few-shot prompting techniques. We used nearest neighbour few shot prompting with cosine similarity to find instances of two most similar examples in comparison to the given example from the training set and then used them as examples in the few-shot prompt. We then compared the accuracy of the produced clarifying question with the given default clarifying question. (Reimers and Gurevych, 2023)

The term ”answers” in zero-shot and few-shot techniques refers to the potential responses to a clarifying question that are already present in our dataset. By incorporating these potential answers into the prompt, we aimed to improve the overall quality of the generated questions. During

Model	Grammatical Correctness	Diversity	Clarity Index	Relevance
bart	**4.7125**	2.55	3.1625	3.6375
t5_kldiv	4.5875	2.7	3.5875	3.7375
t5_inv_entropy	4.4432	2.5	3.1578	3.7765
t5_canard	4.6375	2.6	**3.6125**	3.8
ft5-base	4.4	3.75	2.9875	3.525
ft5-small	3.8875	**3.925**	2.3625	2.75
ft5-large	4.45	3.65	3.2	3.3625
ft5-xl	4.675	3.75	3.45	**3.9875**

Table 2: Averaged human evaluations on 560 instances

our experimentation, we explored various types of prompts to understand their impact on the quality of the clarifying questions. We found that when we explicitly included the possible answers to the ambiguous questions in the original dataset within the prompt, it led to a significant enhancement in the generated questions' quality. This behavior was expected, as it allowed the GPT3.5 model to benefit from the crucial contextual information related to clarifying questions. By providing the model with this context, it was better equipped to generate relevant and meaningful clarifying questions.

5 Baselines

For our baselines, we make use of the BART models mentioned in (Guo et al., 2021). We finetune a BART-large model with the inputs as $S_i, H_{i-k}..H_i, A_i$ where S_i is the story/context, H_i is a history turn and A_i is an ambiguous question with the output being C_i which is the clarification question. Each conversation has upto 3 history turns, setting to value of k to be at most 3. The code for baselines is not publicly available. We implement the baseline only on the conversations which have ambiguous turns. To overcome the problem of a small dataset, we perform data augmentation using GPT3.5 to produce more clarification questions per ambiguous questions and finetune the same architecture. The results with and without data augmentation are compared in Table 1.

6 Our Approaches

In this project, we explore several methods and models on the ABG-CoQA dataset. Specifcally, we want to answer the following questions.

1. RQ1: How does scaling improve the quality of clarification questions ? To this end, we explore the impact of model scaling on the generation process by exploring base, small, large and xl variants of the FlanT5 architecture.

2. RQ2: How does the pretraining task that was used to train the model affect the performance on downstream generation task ? To this end, we explore the impact of using - a BART model pretrained on dialogue corpus, a T5 model pretrained for a question rewriting task and the Flan T5 family of models which are pretrained for a variety of instruction following tasks.

3. RQ3: How do decoding strategies affect the diversity and quality of generations? To explore this, we have experimented with nucleus sampling and ancestral sampling.

4. RQ4: How do regularization methods help in diversifying the text generation methods, and do they have any impact on the quality of text produced? To this end, we explore two regularization strategies.
   (a) We use the inverse entropy of the model's vocabulary distribution as a regularizer to penalize extremely low entropy distributions. The inverse entropy is calculated as $I = \frac{1}{-\Sigma P(x) \cdot \log(P(x))}$, where $P(x)$ represents the probability assigned to specific token by the LM and Σ represents summation over all possible tokens.
   (b) We use a KL divergence regularization, to make sure the current model's vocabulary distribution does not deviate too much from the pre-trained model. This is to prevent catastrophic forgetting, an issue that plagues most pre-trained and

Model	Coherence	Naturalness	Groundedness	Understandability
bart no data aug	0.6843	**0.9973**	0.9408	0.**9973**
bart data aug	0.6147	0.9972	0.9684	0.9972
t5_kldiv	0.6687	0.9889	0.9800	0.9963
t5_canard	0.6859	0.9976	**0.9950**	0.9971
ft5-small	0.2339	0.7888	0.8965	0.8165
ft5-base	0.6554	0.8221	0.9923	0.9376
ft5-large	0.7002	0.9917	0.9855	0.9892
ft5-xl	**0.7504**	0.9877	0.9810	0.9956

Table 3: UniEval Scores for all our models. Language features like Coherence, Naturalness, Groundedness and Understandability are measured.

fine-tuned language models. Most pre-trained LLMs have the innate ability to produce diverse text and we investigate the impact of regularization to preserve this ability while also being able to perform well on asking clarification questions. The KL divergence loss is defined as $KL(P||Q) = \Sigma P(x) \cdot \log(\frac{P(x)}{Q(x)})$, where $P(x)$ is the probability of observing a token according to the true distribution and $Q(x)$ is the probability of observing it according to a learned distribution. The Σ denotes summation over all possible tokens in the vocabulary.

Custom loss functions for diversifying text have been researched before in the past. For example, in Frequency Aware Cross Entropy, (Jiang et al., 2019) explore frequency weighting of the cross entropy logits for each vocabulary term, in a bid to promote lexical diversity. Similarly, (Ueyama and Kano, 2020) propose incorporating inverse n-gram frequency into the model's softmax layer. Most methods, explore this problem from a *lexical* diversity standpoint and and not from a *semantic* diversity standpoint. We investigate the impact of KL-divergence regularization on one of our models for producing semantically diverse text. The KL divergence based method is directly inspired from RLHF tuning (Ziegler et al., 2020), where the RLHF tuned model is regularized using KL divergence to make sure it does not deviate too much from its original behaviour.

5. RQ5: How does training of different models on the GPT3.5 generated data impact the performance? The original dataset contains a small amount of examples involving ambiguous and clarifying questions. Using GPT3.5, we generate more high quality clarifying question yeilding a dataset three times bigger than the original dataset. We evaluate and compare the performance of our models before and after being trained on the augmented dataset.

7 Experiments

7.1 Training and Hyper-parameters

All models, particularly FLAN-T5-xl, require large computation to be fine-tuned. A single *p3.8xlarge* AWS instance was spun up to fine-tune. Pre-trained models from HuggingFace Hub with the *transformers* library were used with PyTorch. All the models were trained for 3 epochs, with a learning rate of 5e-6 and used the Adam Optimizer. The batch size used for FLAN-T5-xl was 2 and for other models 4. Gradient accumulation was performed for 16 steps. A learning rate scheduler with a warmup of 10 steps was implemented.

7.2 Automatic Evaluation

Evaluation for text generation methods have traditionally followed automated metrics including BLEU, ROUGE and METEOR scores. Metrics like BLEU and ROUGE only take into account lexical overlaps and patterns without taking into account the semantics of the generated text. Given the nature of this task, we also need evaluation metrics that can adequately capture more subtle indicators like cohesiveness, continuity, fluency and diversity in a dialogue setting and better correlate with human judgement. To this end, we evaluate our models on UniEval and Vendi in addition to ROUGE, BLEU, and BERTScore.

Model	BLEU	ROUGE	BERT Score			Vendi Score
			Precision	Recall	F1	
bart	0.0484	0.1879	**0.6071**	**0.6005**	**0.6012**	1.458
t5_canard	0.0333	0.1558	0.5630	0.5881	0.5729	1.6236
t5_kldiv	0.0321	0.1548	0.5629	0.5859	0.5716	1.6323
ft5-small	0.0249	0.1302	0.5543	0.5607	0.5545	1.8671
ft5-base	0.0421	0.1684	0.5848	0.5790	0.5792	**1.9056**
ft5-large	0.0496	0.1708	0.5861	0.5792	0.5798	1.8268
ft5-xl	**0.0570**	**0.1969**	0.5994	0.5887	0.5914	1.8440

Table 4: Evaluation of all fine-tuned models. The BLEU and ROUGE scores measure the n-gram overlap between the generated clarifying questions and the ground truth. BERT score measures contextual similarity in the word embeddings. Vendi score is responsible for measuring diversity of generated responses.

UniEval

UniEval (Zhong et al., 2022) is an evaluation toolkit that can evaluate summarization, fact checking and dialogue response generation tasks. UniEval on dialogue, assesses generation quality on indicators including naturalness, coherence, engagement, understability and groundedness. From a diversity evaluation standpoint, we make use of the Vendi Score (Friedman and Dieng, 2022).

Vendi Score

The Vendi Score encodes each sentence using a sentence transformer model, following which it computes pairwise cosine similarity, and uses the Eigen values of the similarity matrix obtained to calculate a Shannon entropy score. The entropy obtained is a measure of the diversity in the sentences. The advantage with Vendi Score is that it uses sentence transformers and takes into account semantic diversity by construction.

GPTEval

To assess the performance of our trained models, we also employed GPT3.5 as an evaluator. This involved presenting it with a predetermined prompt and a set of grading criteria. GPT3.5 then assigned scores to each clarifying question based on a given story and ambiguous question, utilizing the provided criteria. In this evaluation method, GPT3.5 assigned a 4-tuple of scores to each received clarifying question. Each score in the 4-tuple ranged from 1 to 5, where a score of 1 represented the worst performance and a score of 5 indicated the best performance. The following criteria are presented for evaluation:

- ”How much sense does the clarifying question make according to the English language?” - This criteria checks for grammatical correctness of the generated questions.
- ”How different are the two questions from each other?” - This helps measure the diversity of generated clarifying questions.
- ”How effective is the clarifying question in removing the ambiguity of the ambiguous question?” - This criteria directly targets the effectiveness of clarifying question in resolving the ambiguity.
- ”How likely is it that the clarifying question is answerable based on the story and ambiguous question?” - This criteria helps us gauge whether the model is generating clarifying questions related to the underlying story and ambiguous question or not.

7.3 Human Evaluation

Humans are tasked with evaluating 560 clarifying questions from 40 passages. We task 4 humans with scoring each of the questions generated on a scale of 1-5 based on the following criteria:

- **Grammatical Correctness**: Judges whether a question generated is grammatical correct according to the syntactic structure of a language. Smaller models with fewer parameters and training data might make grammatical errors.
- **Diversity**: Judges the extent of diversity between two questions generated by a model. Models which generate the same question twice are assigned a lower score.
- **Clarity Index**: Metric to judge how capable the clarifying question is in resolving the ambiguity of the previously asked question.

- **Relevance**: Judges if the clarifying question asked is relevant to the story, history and the ambiguous question asked. A question which references substance outside of the given story is assigned a lower relevance score.

The *Clarity Index* metric judges if the clarifying question asked is helpful in resolving the ambiguity of the user asked question. We found out that the ambiguity of a question can be reduced by either providing multiple options for what the ambiguous question is referring to or by redirecting the user to the nearest match. For example, in Figure 1, the question ”Do you mean the trade name?” solves more ambiguity to ”What was it called?” than the question ”When was the active principle in epinephrine identified?”.

An annotation guideline was formed with a description of the task. Tricky edge-cases and possibilities for false-positives were highlighted. Scoring on 2 clarifying questions was performed for an unseen example passage. The annotation guideline can be accessed here. A markdown version of the annotation guideline can also be found within the source at $eval/human\_eval/annotation\_guideline.md$.

8 Results

A total of the 7 aforementioned fine-tuned models were evaluated on the test set. Each model generates two clarifying questions given the story, history turns and the ambiguous questions. The results can be seen in Figure 2 at the end of the report.

We observe that both *t5_kldiv* and *t5_canard* produce the same two clarifying questions which also explains the low diversity score they achieve during human evaluations. Smaller models like *bart* produce the second clarifying question by extending the first one. This can be attributed to limited vocabulary and fewer parameters of the model. *ft5-small* being the smallest of the Flan models produces grammatically incorrect and incoherent clarifying questions. The results start to improve with *ft5-base* and *ft5-large* with the latter producing coherent, diverse and relevant clarifying questions. Larger models like the *ft5-xl* although are capable of producing longer and diverse clarifying questions are not completely relevant.

9 Error analysis

9.1 Evaluation of outputs from different Models

In this section, the outputs of each model for a given test set example are analyzed.

Flan T5 XL

- AQ: what color was it?
- CQ 1: What color was the book found by Angie’s mother in the adult reading room?
- CQ 2: The book Angie’s mother found was what color?

Analysis : Flan T5 XL produces outputs that are much more specific and helpful as opposed to even the ground truth question. The model performs only moderately in terms of diversity, as both the clarifications are referring to the books found by Angie’s mother.

BART

- AQ: what color was it?
- CQ 1: What color was the book?
- CQ 2: What color was the book that Angie found?

Analysis: The diversity again is poor, and BART asks questions at a much lower granularity than Flan T5 xl or even the ground truth, which specifically asks if it is the first book. This is also captured by the lower helpfulness metric that is obtained in human evaluation as shown in Table **??**

FLAN T5 large

- AQ: what color was it?
- CQ 1: What color was the large book Angie’s mother found?
- CQ 2: What color was the book Angie made a drawing of her mother?

Analysis: FLAN T5 captures another aspect of the book, from the story given the fact the book could be the large book. The second question is very incoherent and in fact grammatically incorrect. This is a symptom of ancestral sampling which is used for diversification in text generation.

Again in terms of diversity the model is not very good, as both clarifications are about the same facet, pertaining to Angie's mother.

T5 CANARD with KL divergence regularization

- AQ: what color was it?
- CQ 1: what color was the book Angie and her mother found?
- CQ 2: what color was the book Angie and her mother found?

Analysis: KL Divergence was one of the regularization stratgies tried. KL divergence based regularization does not very much help in this case in that generated outputs are identical and indicate softmax probabilities are very skewed. This is a result that is well evidenced by the lowest score of 2.6 in terms of diversity on human evaluation. The primary reason why smaller models would fail as opposed to large and XL models is due to the lack of data. Scaling laws by (Kaplan et al., 2020) indicate that both LMs and the number of tokens seen need to increase at an equal rate in order for LM performance to really see drastic improvements. In this case, dataset size is a real bottleneck that hinders a model from achieving its full capabilities.

T5 canard with inverse entropy regularization

- AQ: what color was it?
- CQ 1: What color was the book Angie's mother found?
- CQ 2: what color was the book Angie and her mother found?

Analysis: Another regularization strategy which was tried is adding inverse of entropies to the loss function. This was done to penalize low entropy distributions. However this strategy also seems to be bottlenecked by a lack of data like the KL Divergence variant.

Flan T5 small

- AQ: what color was it ?
- CQ 1: How color was the book color?
- CQ 2: How color was the book the Red chair painted with?

Analysis: These outputs are very incoherent. Flan T5 small is our smallest model, and from the generated outputs, we can observe that there is a need for larger models, that can handle clarification question generation. This is evident from very low coherence scores obtained for flan t5 small in UniEval.

All in all, in this example, we observe that Flan T5 XL and Flan T5 large models are the ones that perform well in terms of asking very specific questions. BART model performs well in terms of coherence, but is unable to add more details to the clarification. Flan T5 small and T5 Base with both regularization methods is not effective at mitigating the issue of lack of diversity.

The diversity problem is mainly caused by lack of training data even after augmentation. After augmentation, we have 5 clarification questions per ambiguous question. In all, we still have only 800 ambiguous questions, which does not give sufficient error signals for a generative model to perform well on a task of clarification question generation. Similarly, model sizes play a very important role, as is evident from the performance of Flan T5 XL and Flan T5 large.

From the point of view of pretraining strategy that was used to train these models, FlanT5 models are the best performing tasks. Flan T5 models are trained on a large amount of data, much more so than BART and also with a different masked language modelling objective as compared to BART. The inference is that larger language models which are trained on a wide variety of tasks have better generalization capabilities, mainly due to the fact that they have seen diverse and large amounts of data during their pretraining. Scaling laws are visible here, wherein model effectiveness only increases with model size for the same amount of data.

9.2 Evaluation using GPT 3.5

We also use GPT3.5 to evaluate performance of our models by asking it to give a 4-tuple score (based on different criteria) to every clarifying question. The average score for all the clarifying questions generated by GPT3.5 can be seen in Table 1. We provided similar guidelines to humans to get scores for 560 clarifying questions generated by our model and the average score from human evaluation is shown in Table 2. From the tables 1 and 2 it is evident that GPT3.5 failed to give appropriate scores (in comparison to humans) to the clarifying questions generated by different models in our experiments. Consider the following example:

> Story: John was in the third grade, and nine years old. Every day he had to walk home from school. There were some kids in his class who were mean to him, and during the winter they would throw snowballs at him. John could have told the teacher, but one of the kids was a very pretty girl. She was mean, but John liked her because she was pretty and did not want her to get in trouble. One day, his teacher asked John to stay after class to wipe off the chalkboard and to empty the pencil sharpener. By the time he was done, the other kids had gone home. They could no longer throw snowballs at him. John did not mind helping out his teacher, and he soon stayed after class every day.John was not very good at math, and sometimes his teacher would help him when he stayed after school. She said if John could help her out for at least two weeks, he could pass his math class. John thought it was a good deal, and ended up being much better at math.
>
> Ambiguous Question: What happened when he helped?

The top two clarifying questions generated by the fine-tuned Flan-T5 Large model for this story and ambiguous question pair are:

- What did John do after his teacher asked him to stay after class to clean off the chalkboard and empty the pencil sharpener? - For this question, GPT3.5 generates a score tuple of (3,2,5,1) while a human annotator scored this question with (5,4,4,1).

- What happened the first time John helped the teacher? For this question, GPT3.5 generates a score tuple of (2,3,5,2) while a human annotator scored this question with (5,4,5,5)

The inability of GPT3.5 to perform well in comparison to humans can be attributed to the fact that there is still room for improvement in the prompt and the evaluation criteria that GPT3.5 was given.

9.3 Human Evaluation

The human evaluation scores can be seen in Table 2. While most of the models produce grammatically correct outputs, it is seen that smaller models produce less diverse results and tend to repeat the same question again. Both the T5 models rank higher in terms of resolving ambiguity. The high score might also be attributed to the fact that these models reuse words from the ambiguous question which makes their outputs seem more ambiguity resolving. FLAN-T5-XL owing to its huge size is able to remember more text from the story as result of which its outputs are majorly in context with the story provided and it consequently achieves a higher relevance score.

10 Contributions of group members

1. Rahul: Research on base paper and dataset, setting up evaluation pipelines for automatic metrics, implementing Entropy based regularization, and KL divergence regularization, finetuning BART and T5 models, literature survey

2. Aman: Data Augmentation with GPT3.5 using Zero Shot and Few Shot techniques, Evaluation of output of 7 different models using GPT3.5, Annotation for Human evaluation, literature survey, Setting up evaluation criteria for GPTEval for automatic evaluation.

3. Rohan: Fine-tuning all FLAN-T5 models (small, base, large, xl). All 4 evaluation metrics for BART, T5 models and the FLAN family. Annotation guidelines and metric aggregation for human evaluation.

4. Shreeya: Nearest Neighbor Few Shot prompting (with cosine similarity) implementation using GPT3.5, literature survey, analyzing GPT3.5 performance and evaluation of language models for clarifying questions generation, annotation for human evaluation.

11 Conclusion

In this project we explored multiple strategies of clarifying questions generation to ambiguous queries. We performed pre-processing and augmentation of the ABG-COQA dataset. We finetune multiple open source language models on this augmented dataset and evaluated their performance.

We used multiple evaluation strategies including GPTEval on 1722 instances and human evaluation on more than 560 instances. We also perform extensive experiments and calculate evaluation metrics for each of the models. A thorough error analysis reveals that large language models are capable of generating clarifying questions that are coherent, diverse and relevant. A major difficulty was parallelizing workloads of larger LMs like ft5-xl on both the GPU and the CPU. Using a clever prompt to augment the dataset while making sure questions were not template-based and that significant diversity was maintained was also tricky.

An extension of this project would be in-line with our previous idea of introducing multi-modality within the task. Introducing models like CLIP which can evaluate the similarity between the content of an image and a textual description would be an interesting approach. A textual description of an associated image would provide context in addition to the story and history turns to the model. Wikipedia articles can be a potential dataset in this regard, since images and correlated text are easily available.

As LLMs become more prominent, user behaviors towards them would continue to become more human like. In such a scenario, asking clarifying questions to an ambiguous user query would prove immensely useful to avoid giving incorrect or unexpected responses.

12 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - Yes. GPT3.5 and BARD were used for data augmentations and simplified defintions.

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - Data augmentation prompts are covered in the report above.
  - Single sentence definition prompts for BLEU, ROUGE and BERT score were used.
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - Astute prompting resulted in expected results. It was particularly tricky to find a suitable prompt for generating synthetic data.

References

Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., and Burtsev, M. (2020). Convai3: Generating clarifying questions for open-domain dialogue systems (clariq).

Aliannejadi, M., Zamani, H., Crestani, F., and Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Du, X., Shao, J., and Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Friedman, D. and Dieng, A. B. (2022). The vendi score: A diversity evaluation metric for machine learning.

Guo, M., Zhang, M., Reddy, S., and Alikhani, M. (2021). Abg-coQA: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Jiang, S., Ren, P., Monz, C., and de Rijke, M. (2019). Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference*, WWW '19, page 2879–2885, New York, NY, USA. Association for Computing Machinery.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103.

Rathod, M., Tu, T., and Stasaski, K. (2022). Educational multi-question generation for reading comprehension. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 216–223, Seattle, Washington. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2023). SBERT Semantic Search Example. https://www.sbert.net/examples/applications/semantic-search/README.html.

Shao, T., Cai, F., Chen, W., and Chen, H. (2022). Self-supervised clarification question generation for ambiguous multi-turn conversation. *Information Sciences*, 587:626–641.

Ueyama, A. and Kano, Y. (2020). Diverse dialogue generation with context dependent dynamic loss function. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4123–4127, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wang, J. and Li, W. (2021). Template-guided clarifying question generation for web search clarification. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management*, CIKM '21, page 3468–3472, New York, NY, USA. Association for Computing Machinery.

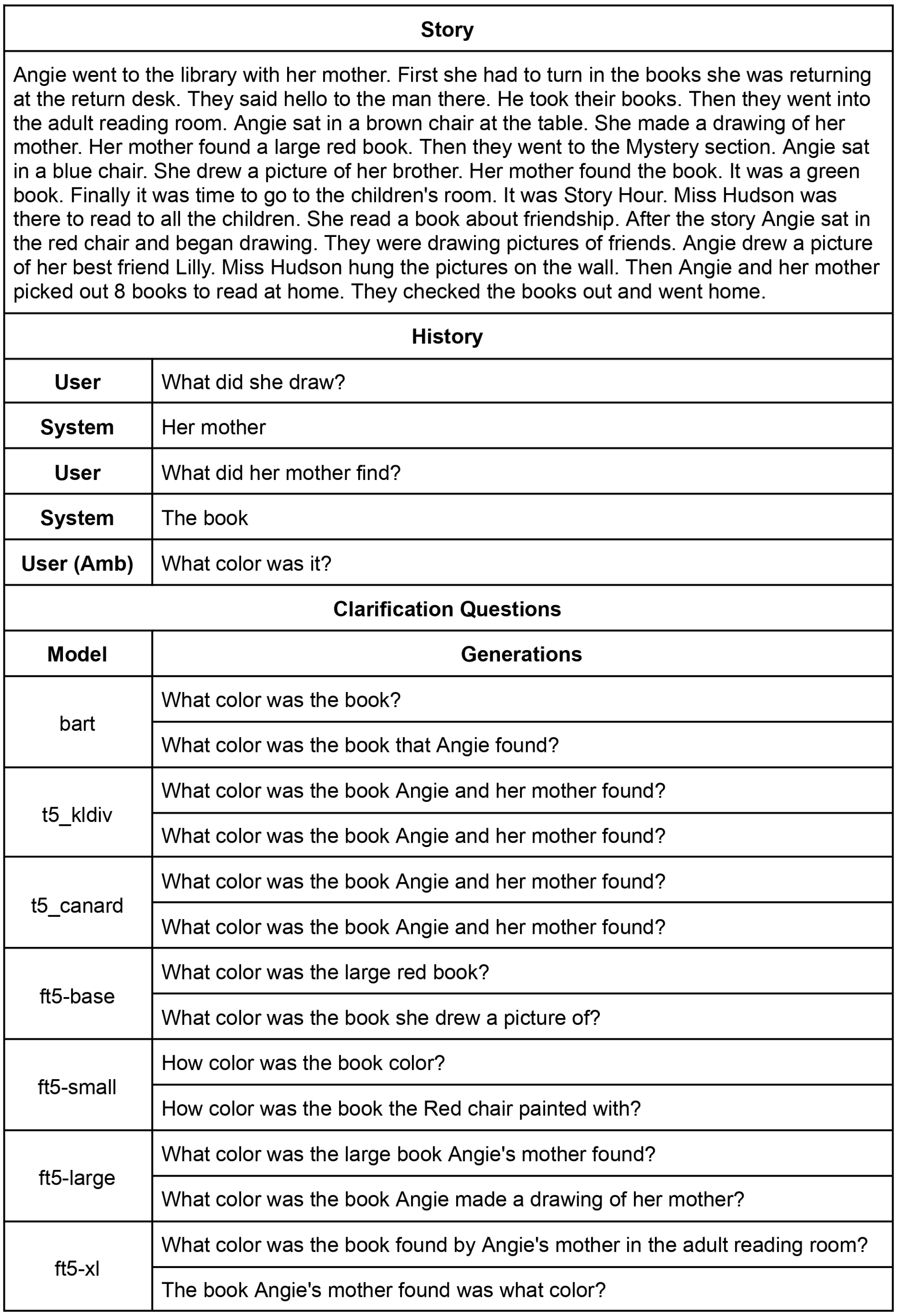
White, J., Poesia, G., Hawkins, R., Sadigh, D., and Goodman, N. (2021). Open-domain clarification question generation without question examples.

Xu, J., Wang, Y., Tang, D., Duan, N., Yang, P., Zeng, Q., Zhou, M., and Sun, X. (2019). Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629, Hong Kong, China. Association for Computational Linguistics.

Zamani, H., Dumais, S., Craswell, N., Bennett, P., and Lueck, G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, page 418–428, New York, NY, USA. Association for Computing Machinery.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., and Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2020). Fine-tuning language models from human preferences.



Story	
Angie went to the library with her mother. First she had to turn in the books she was returning at the return desk. They said hello to the man there. He took their books. Then they went into the adult reading room. Angie sat in a brown chair at the table. She made a drawing of her mother. Her mother found a large red book. Then they went to the Mystery section. Angie sat in a blue chair. She drew a picture of her brother. Her mother found the book. It was a green book. Finally it was time to go to the children's room. It was Story Hour. Miss Hudson was there to read to all the children. She read a book about friendship. After the story Angie sat in the red chair and began drawing. They were drawing pictures of friends. Angie drew a picture of her best friend Lilly. Miss Hudson hung the pictures on the wall. Then Angie and her mother picked out 8 books to read at home. They checked the books out and went home.	
History	
User	What did she draw?
System	Her mother
User	What did her mother find?
System	The book
User (Amb)	What color was it?
Clarification Questions	
Model	**Generations**
bart	What color was the book?
	What color was the book that Angie found?
t5_kldiv	What color was the book Angie and her mother found?
	What color was the book Angie and her mother found?
t5_canard	What color was the book Angie and her mother found?
	What color was the book Angie and her mother found?
ft5-base	What color was the large red book?
	What color was the book she drew a picture of?
ft5-small	How color was the book color?
	How color was the book the Red chair painted with?
ft5-large	What color was the large book Angie's mother found?
	What color was the book Angie made a drawing of her mother?
ft5-xl	What color was the book found by Angie's mother in the adult reading room?
	The book Angie's mother found was what color?

Figure 2: The story contains the context which is provided to each of the model during both training and inference time. History turns allow the model to gain a chronological context of the previous conversations. The outputs produced by each of the model can be seen above.