# Aligning Language Models to Explicitly Handle Ambiguity

**Anonymous ACL submission**

## Abstract

In interactions between users and language model agents, user utterances frequently exhibit ellipsis (omission of words or phrases) or imprecision (lack of exactness) to prioritize efficiency. This can lead to varying interpretations of the same input based on different assumptions or background knowledge. It is thus crucial for agents to adeptly handle the inherent ambiguity in queries to ensure reliability. However, even state-of-the-art large language models (LLMs) still face challenges in such scenarios, primarily due to the following hurdles: (1) LLMs are not explicitly trained to deal with ambiguous utterances; (2) the degree of ambiguity perceived by the LLMs may vary depending on the possessed knowledge. To address these issues, we propose **Alignment with Perceived Ambiguity (APA)**, a novel pipeline that aligns LLMs to manage ambiguous queries by leveraging their own assessment of ambiguity (i.e., **perceived ambiguity**). Experimental results on question-answering datasets demonstrate that APA empowers LLMs to explicitly detect and manage ambiguous queries while retaining the ability to answer clear questions. Furthermore, our finding proves that APA excels beyond training with gold-standard labels, especially in out-of-distribution scenarios.

## 1 Introduction

Large Language Models (LLMs) (Ouyang et al., 2022; Team et al., 2023; Achiam et al., 2023) have demonstrated remarkable capabilities in text generation, proving particularly effective for question-answering (QA) tasks (Zhang et al., 2023; Etezadi and Shamsfard, 2023). QA systems in the wild frequently encounter unexpected user input, such as unanswerable (Kim et al., 2023b; Yin et al., 2023) or ambiguous questions (Cole et al., 2023; Lee et al., 2023; Kim et al., 2023a). To build an agent that is both reliable and user-friendly, it is essential for the model to robustly handle such inputs. In this work, we seek to extend the scope of research to manage invalid inputs effectively. Specifically, we focus on managing "ambiguity" (Gleason, 1963; Mackay and Bever, 1967), which poses a significant challenge in Natural Language Processing (NLP) (Jurafsky, 1996).

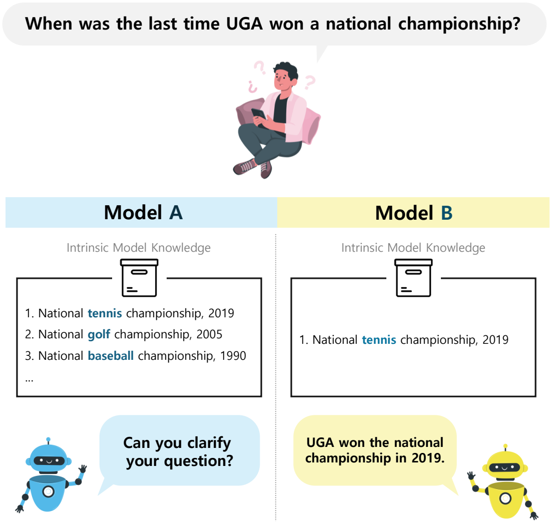

When was the last time UGA won a national championship?
Model A
Model B
Intrinsic Model Knowledge
1. National tennis championship, 2019
2. National golf championship, 2005
3. National baseball championship, 1990
...
Intrinsic Model Knowledge
1. National tennis championship, 2019
Can you clarify your question?
UGA won the national championship in 2019.

Figure 1: An example of an ambiguous query from AmbigQA. The term "national championship" poses diverse denotations, causing ambiguity. (Left) A model with diverse relevant knowledge might perceive the case as ambiguous. (Right) In contrast, the query can be deemed unambiguous when the model lacks substantial related knowledge. Thus, the perceived ambiguity may differ depending on the model's intrinsic knowledge.

**Ambiguity** refers to cases where an expression conveys multiple denotations (Wasow et al., 2005). Users may pose queries with clear intentions that, possibly due to insufficient domain knowledge or omission during the utterance, result in ambiguous requests. If a model arbitrarily responds to such ambiguity, there is a risk of misinterpreting the user's original intent, potentially harming the model's reliability. This is particularly evident in domains requiring high reliability, such as legal (Schane, 2002; Choi, 2024) or medical (Stevenson and Guo, 2010; Gyori et al., 2022), where misinterpretations may lead to severe consequences. Despite such importance, approaches to manage ambiguity robustly are still significantly unexplored.

Properly processing ambiguous inputs is challenging primarily due to the following two hurdles. Firstly, models are **not trained to express ambiguity explicitly**. Even if a model is capable of recognizing ambiguity, confirming this recognition requires explicit cues from the model itself, such as expressing uncertainty or offering multiple interpretations. The second challenge is that the **degree of ambiguity perceived by the model can vary** based on its intrinsic knowledge. Consider the scenario depicted in Figure 1. The initial query is ambiguous as the phrase “national championship” poses various denotations, such as “national *tennis* championship” or “national *golf* championship”. With comprehensive knowledge across possible denotations, a model can likely recognize the query’s ambiguity (Figure 1, left). However, limited knowledge would lead the model to perceive the query as unambiguous (Figure 1, right). Therefore, how a model interprets ambiguity hinges on its knowledge scope, which we define as **perceived ambiguity**.

To overcome these issues, this paper proposes **Alignment with Perceived Ambiguity (APA)**— a novel alignment pipeline for models to **explicitly handle** ambiguous queries by leveraging their **perceived ambiguity**. Specifically, we design a proxy task that guides the model in utilizing its intrinsic knowledge for self-disambiguation of a given query. We then quantify the information gained from this disambiguation as an implicit measure of the extent to which the model perceives the input as ambiguous. This measure serves as a cue for ambiguous sample selection. For the selected ambiguous query and its disambiguation, the model generates a clarification request regarding the ambiguity. Finally, the model is trained to request explicit clarification in response to ambiguous queries.

Experimental results from a range of QA datasets demonstrate that APA enables a language model to properly handle ambiguous inputs while maintaining its inherent capabilities of answering unambiguous queries. Furthermore, we present three new datasets to provide a comprehensive framework for assessing ambiguity: AmbigTriviaQA, AmbigWebQuestions, and AmbigFreebaseQA. These datasets facilitate a more extensive evaluation of models’ robustness in addressing ambiguity, thus contributing to the further expansion of related research.

## 2 Related Work

**Ambiguity in NLP** An expression is ambiguous if it has two or more distinct denotations (Wasow et al., 2005). Ambiguity poses a significant challenge to NLP applications by obscuring the intended meaning of expressions, preventing models from accurately performing specific tasks. Efforts to address this issue span across various domains, including machine translation (Pilault et al., 2023), coreference resolution (Poesio and Artstein, 2005; Yuan et al., 2023), and natural language inference (Liu et al., 2023). The challenge intensifies in the scope of QA, as ambiguous questions may yield multiple answers that may not align with the user’s initial intent. Min et al. (2020) introduce the AmbigQA dataset to tackle ambiguity in open-domain QA and Stelmakh et al. (2022) expand it to long-form generation. Furthermore, Cole et al. (2023) demonstrate that quantifying sampling repetition presents a reliable uncertainty measure for ambiguity, while Kim et al. (2023a) generate tree-of-clarification (ToC) that refines input ambiguity. While we share the goal of handling ambiguity, we propose a method of directly aligning the model.

**Alignment of LLMs** LLMs are typically trained through causal language modeling, a process essential for understanding and generating text of high fluency and consistency. To better harness these models, approaches have been developed to align them with human preferences (Leike et al., 2018; Ji et al., 2023b) through various forms, such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Chakraborty et al., 2024), and Supervised Fine-tuning (SFT) (Dong et al., 2023; Yang et al., 2023; Zhou et al., 2024). Previous works focused on preferences such as helpfulness (Ding et al., 2023; Köpf et al., 2023; Xu et al., 2024), safety (Bai et al., 2022; Ji et al., 2023a; Liu et al., 2024b), and factuality (Yang et al., 2023; Tian et al., 2024). Building on this foundation, our research expands the scope of research by focusing on aligning models to understand and manage ambiguity effectively.

**Data Quality Control for Alignment** Data-centric AI (Chu et al., 2016; Majeed and Hwang, 2023; Kumar et al., 2024) highlights the importance of data quality in model training. In the context

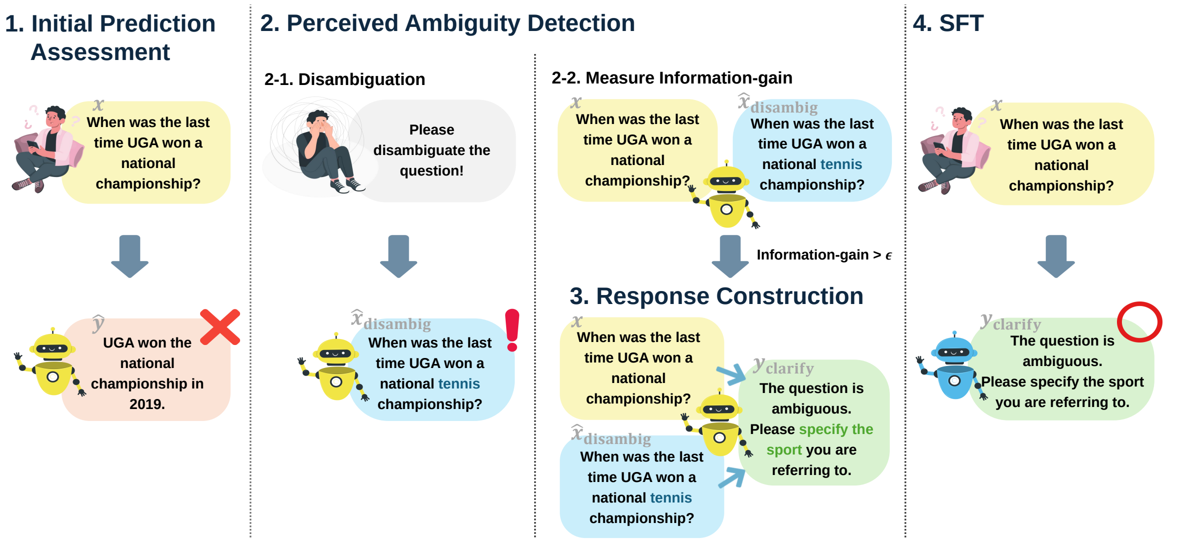

1. Initial Prediction Assessment
$x$
When was the last time UGA won a national championship?
$\hat{y}$
UGA won the national championship in 2019.
2. Perceived Ambiguity Detection
2-1. Disambiguation
Please disambiguate the question!
$\hat{x}_{\text{disambig}}$
When was the last time UGA won a national tennis championship?
2-2. Measure Information-gain
$x$
When was the last time UGA won a national championship?
$\hat{x}_{\text{disambig}}$
When was the last time UGA won a national tennis championship?
Information-gain > $\epsilon$
3. Response Construction
$x$
When was the last time UGA won a national championship?
$\hat{x}_{\text{disambig}}$
When was the last time UGA won a national tennis championship?
$y_{\text{clarify}}$
The question is ambiguous. Please specify the sport you are referring to.
4. SFT
$x$
When was the last time UGA won a national championship?
$y_{\text{clarify}}$
The question is ambiguous. Please specify the sport you are referring to.

Figure 2: The overall process of APA. We select incorrect samples from the model (Stage 1) and let the model self-disambiguate them with the intrinsic knowledge. We measure the information gain (INFOGAIN) between the input and the disambiguation and select samples with high INFOGAIN as ambiguous (Stage 2). Then, the model generates a clarification request regarding the ambiguity (Stage 3), which is used as the label for training (Stage 4).

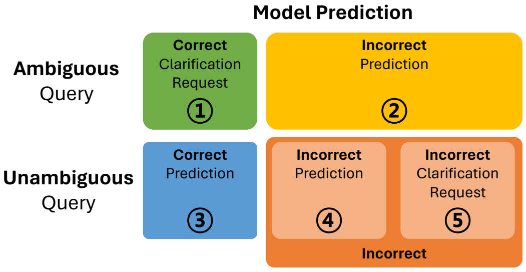


Figure 3: Illustration of five possible results from our scenario. For ambiguous queries, the prediction is correct (①) if the model generates a clarification request; otherwise, all the other responses are classified as incorrect (②). When evaluating unambiguous queries, we compare the predictions to the ground-truth labels and categorize them as the correct prediction (③), incorrect prediction (④), or incorrect clarification request (⑤).

of instruction-following techniques, LIMA (Zhou et al., 2024) demonstrates that effective model alignment can be achieved with just 1,000 high-quality, human-curated samples. Similarly, AlpaGasus (Chen et al., 2024) leverages only a small subset of the Alpaca dataset (Taori et al., 2023), filtered by ChatGPT, for an effective alignment. Various approaches for data selection have been explored, including those based on factors such as length and complexity (Liu et al., 2024a), and gradient similarity from validation sets (Xia et al., 2024). This work proposes a new viewpoint on data quality estimation: assessing how well data aligns models for ambiguity management. For this purpose, we utilize the model's perceived ambiguity as an implicit cue for measuring data quality.

## 3 Methodology

The primary goal of our research is to align models in a way that they can explicitly handle potentially ambiguous inputs, leveraging the model's perceived ambiguity. To this end, we propose **Alignment with Perceived Ambiguity (APA)**, a four-stage alignment pipeline, illustrated in Figure 2. In this section, we first formulate the problem and describe each stage in detail regarding the five possible results depicted in Figure 3. Further implementation details are stipulated in Appendix A.

**Problem Formulation** In this study, we focus on open-domain QA. The model $M$ is expected to generate a prediction $\hat{y}_{\text{unambig}}$ for an unambiguous query $x_{\text{unambig}}$ given a pre-defined inference template $t(\cdot)$. $\hat{y}_{\text{unambig}}$ is compared to the ground-truth label $y$ and categorized as correct prediction (③), incorrect prediction (④), or incorrect clarification request (⑤). As we expand our input scope to ambiguous queries[1], the model prediction for the ambiguous query $\hat{y}_{\text{ambig}}$ is anticipated to serve as a clarification request $y_{\text{clarify}}$ to resolve the ambiguity. This approach is grounded on the assumption that

[1] Separating ambiguous from unambiguous queries is inherently challenging due to subjective factors such as various perspectives and underlying assumptions. Despite the complexity, we simplify the problem and follow the pre-defined ambiguity from the training dataset for the alignment.

the user is best positioned to clarify their intent.[2] $\hat{y}_{\text{ambig}}$ is considered correct (①) if it is a proper clarification request. Otherwise, responses that fail to address the ambiguity are classified as incorrect (②). The final objective of the alignment is to increase the number of samples corresponding to ① while simultaneously maintaining or improving the proportion of responses classified as ③.

### 3.1 Initial Prediction Assessment

The initial stage focuses on identifying samples that the model currently fails to handle. To do so, we compare the model's prediction with the ground-truth label, where samples are categorized based on accuracy. Specifically, we assess the correctness by matching $\hat{y}_{\text{unambig}}$ with $y$ and $\hat{y}_{\text{ambig}}$ with $y_{\text{clarify}}$. A total of $n$ correct samples, included in ① and ③, are collected as $D_{\text{correct}} = \{(x^i_{\text{correct}}, y^i_{\text{correct}})\}^n_{i=1}$. Incorrect samples falling under categories ②, ④, and ⑤ are unified as a separate dataset, $D_{\text{incorrect}}$.

### 3.2 Perceived Ambiguity Detection

This stage aims to identify samples from $D_{\text{incorrect}}$ that the model perceives as ambiguous. Given that it is challenging for the model to express ambiguity explicitly, we construct a proxy task to estimate the ambiguity from the model's perspective. Specifically, the model is prompted to self-disambiguate the given query $x$ and generate a disambiguation $\hat{x}_{\text{disambig}}$. The model leverages its intrinsic knowledge related to $x$ to generate further details in this process. If $x$ is underspecified and the model possesses related knowledge necessary to compensate, then $\hat{x}_{\text{disambig}}$ would yield a higher certainty (lower entropy) from the model's perspective. On the other hand, if $x$ requires no specification or the model lacks the necessary knowledge, $\hat{x}_{\text{disambig}}$ would exhibit a similar level of uncertainty as $x$. To quantify the uncertainty associated with $x$ and $\hat{x}_{\text{disambig}}$, we employ the model's average entropy (Malinin and Gales, 2021; Abdar et al., 2021). Formally, the entropy of an output distribution is defined as follows:

$$\mathcal{H}_{x,i} = -\sum_{v\in\mathcal{V}} p_{x,i}(v)\log p_{x,i}(v) \qquad (1)$$

where $p_{x,i}(v)$ is the probability of the $i^{\text{th}}$ token $v$ of a sentence $x$ from the full vocabulary set $\mathcal{V}$. The average entropy for $x$ can be defined as:

$$\mathcal{H}_x = \frac{1}{N}\sum_i \mathcal{H}_{x,i} \qquad (2)$$

with $x$ composed of $N$-tokens. We quantify the additional information gained from $\hat{x}_{\text{disambig}}$ by the difference in average entropy, which we define as **information gain (INFOGAIN)**.

$$\text{INFOGAIN}_{x,\hat{x}_{\text{disambig}}} = \mathcal{H}_x - \mathcal{H}_{\hat{x}_{\text{disambig}}} \qquad (3)$$

A meaningful specification from $\hat{x}_{\text{disambig}}$ would result in a substantial INFOGAIN, suggesting that the model perceives $x$ as ambiguous. Regardless of the ground-truth ambiguity, samples with INFOGAIN greater than the threshold $\epsilon$ are classified as ambiguous, denoted as $x_{\text{ambig}}$.

### 3.3 Response Construction

In this stage, we define $y_{\text{clarify}}$, which represents the clarification request the model should generate in response to an ambiguous query. We explore two approaches for response generation: Fixed response and Generated response.

**Fixed Response** We utilize a pre-defined clarification request as $y_{\text{clarify}}$ for $x_{\text{ambig}}$.

**Generated Response** The model is prompted to generate a clarification request specifying the source of the ambiguity. To do so, we provide the model with $x_{\text{ambig}}$ and $\hat{x}_{\text{disambig}}$ to identify the aspect that causes the ambiguity, thereby generating $y_{\text{clarify}}$ specific to the identified factor.

### 3.4 Supervised Fine-Tuning (SFT)

The objective of this stage is to construct datasets for the alignment. Specifically, We label $m$ samples identified as ambiguous and construct an ambiguous dataset $D_{\text{ambig}} = \{(x^j_{\text{ambig}}, y^j_{\text{clarify}})\}^m_{j=1}$, where $y_{\text{clarify}}$ serves as the ground-truth label. To prevent the potential loss of the model's existing knowledge, we also incorporate $D_{\text{correct}}$ for training. The number of samples from both datasets are balanced so that $n = m$. The final training dataset is thus established as $D = D_{\text{correct}} + D_{\text{ambig}}$. Utilizing the dataset $D = \{(x^k, y^k)\}^{n+m}_{k=1}$, the model is trained to generate $y$ for $x_{\text{unambig}}$ and $y_{\text{clarify}}$ for $x_{\text{ambig}}$, employing the identical inference template $t(\cdot)$. The model $M$ with parameter $\theta$ is trained as follows:

$$\min_{\theta} \sum_{(x,y)\in D} \sum_{i=1}^{|y|} -\log M_\theta(y_i|y_{<i}, t(x)) \qquad (4)$$

[2] We explored alternatives for ambiguity management but found them to be impractical. For instance, arbitrarily selecting one of the valid answers may not accurately capture the user's intent. Presenting all possible answers is often unfeasible due to the potentially vast number of valid responses.

Two versions of APA are trained based on the type of $y_{\text{clarify}}$: APA$_{\text{FIXED}}$ and APA$_{\text{GEN}}$, which utilizes fixed and generated responses, respectively.

## 4 Experimental Setting

### 4.1 Datasets

The capability of the model to perform within the trained domain is pivotal. However, for real-world applicability, the model must generalize to out-of-distribution (OOD) queries, as queries that diverge from the training data are frequently confronted in practice. Therefore, we utilize AmbigQA (Min et al., 2020) as the in-domain dataset for training and validation. The dataset includes both ambiguous and unambiguous queries, with unambiguous queries labeled with ground-truth answers. SituatedQA (Zhang and Choi, 2021) is used as a held-out OOD test dataset with two different splits, denoted as SituatedQA-Geo and SituatedQA-Temp, each focusing on geographical and temporal ambiguities. To further evaluate ambiguity across diverse QA domains, we have constructed three additional datasets: **AmbigTriviaQA**, **AmbigWebQuestions**, and **AmbigFreebaseQA**, each derived from TriviaQA (Joshi et al., 2017), WebQuestions (Berant et al., 2013), and FreebaseQA (Jiang et al., 2019) respectively. We prompt `gpt-4o`[3] to ambiguate the initial query from the original dataset and verify the generation. To mitigate the potential biases in the validation process, we further evaluate the verified samples with human annotators and select samples for the final dataset. More details on the datasets and the construction process are described in Appendix B.

### 4.2 Baselines

To evaluate the effectiveness of our approach, we introduce two sets of baselines: inference-only methods and trained methods. Specific implementation details are described in Appendix C.

**Inference-Only Methods** Inference-only methods address ambiguity by utilizing different prompting strategies. We employ direct prompting (**DIRECT**) as a fundamental baseline, applying a simple QA prompt. Furthermore, we explore ambiguity-aware prompting (**AMBIG-AWARE**), which incorporates additional instructions on handling ambiguous inputs. We also examine Sample Repetition (**SAMPLE REP**) (Cole et al., 2023) by measuring the consistency of the sampled generations. Finally, we compare **SELF-ASK** (Amayuelas et al., 2023), where the model generates an answer and subsequently determines the ambiguity based on the generation.

**Trained Methods** Given the lack of directly comparable prior work, we compare APA with fine-tuned baselines wherein the model is trained with the in-domain training set. We follow the ambiguity as defined within the in-domain dataset, and train the model accordingly. We compare **FULL-SET**, which applies the entire training dataset. Furthermore, we compare two variations that leverages the equal number of training samples with APA. **SUBSET$_{\text{RAND}}$** is trained on a randomly selected subset with an equal number of ambiguous and unambiguous samples. **SUBSET$_{\text{ENT}}$** applies the entropy of the model's prediction of the ambiguous query as the uncertainty measure. Ambiguous samples with the most significant entropy are selected, and unambiguous samples are selected at random.

### 4.3 Evaluation Metrics

A successful alignment should preserve the model's capability to handle unambiguous inputs while effectively managing ambiguous queries. Based on the five possible results illustrated in Figure 3, we define two distinct metrics to quantify such capabilities. Further details of the evaluation process are described in Appendix D.

**Unambiguous Prediction F1 (F1$_u$)** The model must generate accurate answers to unambiguous queries while minimizing arbitrary responses to ambiguous queries. To measure this, we utilize the unambiguous prediction F1 score, which is the harmonic mean of precision ($\frac{\textcircled{3}}{\textcircled{2}+\textcircled{3}+\textcircled{4}}$) and recall ($\frac{\textcircled{3}}{\textcircled{3}+\textcircled{4}+\textcircled{5}}$) for ambiguous queries.

**Ambiguity Detection F1 (F1$_a$)** Given an ambiguous input, the model should be able to detect them and generate clarification requests accordingly. However, models may exhibit biased predictions toward clarification requests. Taking these aspects into account, we evaluate the model's ambiguity detection capability with the F1-score, which captures both the precision ($\frac{\textcircled{1}}{\textcircled{1}+\textcircled{5}}$) and recall($\frac{\textcircled{1}}{\textcircled{1}+\textcircled{2}}$).

### 4.4 Implementation Details

For our experiments, we utilize LLAMA2 7B & 13B (Touvron et al., 2023), and MISTRAL 7B

[3] https://openai.com/index/hello-gpt-4o/

| Method | # Training Samples | SituatedQA-Geo | | SituatedQA-Temp | | Ambig-TriviaQA | | Ambig-WebQuestions | | Ambig-FreebaseQA | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1$_u$ | F1$_a$ | F1$_u$ | F1$_a$ | F1$_u$ | F1$_a$ | F1$_u$ | F1$_a$ | F1$_u$ | F1$_a$ |
| **Llama2 7B** | | | | | | | | | | | |
| Direct | 0 | 30.44 | 0.00 | 28.38 | 0.00 | 47.68 | 0.00 | 24.87 | 0.00 | 50.07 | 0.00 |
| Ambig-aware | 0 | 7.33 | 32.44 | 3.23 | 35.53 | 27.23 | 68.14 | 14.53 | 62.40 | 51.27 | 76.62 |
| Sample Rep | 0 | 6.83 | 34.43 | 8.28 | 38.43 | 53.11 | 72.63 | 13.31 | 69.21 | 63.11 | 78.70 |
| Self-Ask | 0 | 29.66 | 8.18 | 26.97 | 18.48 | 48.04 | 4.99 | 20.81 | 3.02 | 48.54 | 5.03 |
| Subset$_{\text{Rand}}$ | 3,088 | 31.90 | 37.17 | 29.48 | 33.68 | 54.71 | 70.97 | 38.69 | 73.84 | 63.59 | 77.70 |
| Subset$_{\text{Ent}}$ | 3,088 | 39.33 | 40.84 | 34.28 | 34.62 | 58.83 | 74.98 | 42.39 | 75.86 | 72.18 | 83.89 |
| Full-set | 10,036 | 37.67 | 41.45 | 29.59 | 36.92 | 58.10 | 71.25 | 40.46 | 73.84 | 69.97 | 80.34 |
| Apa$_{\text{Fixed}}$ | 3,088 | 39.99 | 41.86 | 31.74 | 39.63 | **62.97** | 75.50 | **49.15** | **77.07** | **73.37** | 84.19 |
| Apa$_{\text{Gen}}$ | 3,088 | **41.01** | **43.10** | **34.38** | **41.89** | 59.27 | **75.74** | 47.26 | 76.64 | 73.18 | **84.90** |
| **Mistral 7B** | | | | | | | | | | | |
| Direct | 0 | 11.29 | 0.00 | 15.34 | 0.00 | 33.19 | 0.00 | 17.85 | 0.00 | 31.37 | 0.00 |
| Ambig-aware | 0 | 3.66 | 26.01 | 8.43 | 22.48 | 26.26 | 48.43 | 8.39 | 30.52 | 32.96 | 54.91 |
| Sample Rep | 0 | 7.64 | 25.31 | 7.83 | 21.13 | 29.52 | 17.04 | 8.99 | 12.10 | 27.25 | 16.31 |
| Self-Ask | 0 | 11.29 | 0.00 | 15.34 | 0.00 | 33.19 | 0.00 | 17.85 | 0.00 | 31.37 | 0.00 |
| Subset$_{\text{Rand}}$ | 1,382 | 41.42 | 33.95 | 34.14 | 37.01 | 60.57 | 67.82 | 45.16 | 71.74 | 70.60 | 75.93 |
| Subset$_{\text{Ent}}$ | 1,382 | **47.34** | 29.49 | 42.00 | 32.04 | 62.17 | 67.16 | 50.93 | 71.11 | 72.94 | 77.17 |
| Full-set | 10,036 | 35.99 | 41.28 | 31.16 | 33.72 | 66.67 | 76.38 | 41.83 | 74.72 | 76.98 | 84.67 |
| Apa$_{\text{Fixed}}$ | 1,382 | 38.43 | 41.84 | **45.01** | **43.95** | **70.70** | **83.48** | **54.02** | **81.07** | **80.84** | **90.12** |
| Apa$_{\text{Gen}}$ | 1,382 | 39.55 | **42.07** | 43.29 | 40.70 | 67.73 | 82.14 | 51.41 | 79.54 | 80.27 | 89.22 |
| **Llama2 13B** | | | | | | | | | | | |
| Direct | 0 | 30.44 | 0.00 | 29.69 | 0.00 | 46.43 | 0.00 | 27.59 | 0.00 | 49.17 | 0.00 |
| Ambig-aware | 0 | 5.99 | 33.10 | 4.22 | 36.66 | 24.80 | 68.19 | 4.81 | 65.28 | 43.81 | 73.40 |
| Sample Rep | 0 | 11.57 | 32.85 | 16.56 | 37.87 | 49.93 | 72.44 | 7.89 | 67.26 | 61.05 | 79.33 |
| Self-Ask | 0 | 30.44 | 0.00 | 29.69 | 0.00 | 46.43 | 0.00 | 27.59 | 0.00 | 49.17 | 0.00 |
| Subset$_{\text{Rand}}$ | 3,216 | 33.11 | 36.87 | 28.57 | 37.84 | 63.19 | 73.52 | 44.31 | 72.99 | 70.40 | 78.29 |
| Subset$_{\text{Ent}}$ | 3,216 | **40.19** | 38.39 | 31.03 | 38.00 | 64.95 | 76.03 | 48.70 | 77.43 | 73.38 | 81.93 |
| Full-set | 10,036 | 37.58 | 38.39 | 29.41 | 34.37 | 68.33 | 76.82 | 47.20 | 75.27 | 76.56 | 83.00 |
| Apa$_{\text{Fixed}}$ | 3,216 | 31.31 | **40.23** | **36.45** | **42.18** | **70.83** | **80.99** | **53.69** | **79.22** | **79.92** | **88.03** |
| Apa$_{\text{Gen}}$ | 3,216 | 34.04 | 39.89 | 31.72 | 39.36 | 69.25 | 79.57 | 52.96 | 78.46 | 79.80 | 87.61 |

Table 1: Experimental results for five different datasets. We report the unambiguous and ambiguous F1-scores as F1$_u$ and F1$_a$, respectively. For each dataset, the **best method** is highlighted in bold and the second-best method is underlined. Apa outperforms all the baselines by utilizing the perceived ambiguity.

(Jiang et al., 2023). We employ QLoRA (Dettmers et al., 2023) to facilitate efficient training. Results are averaged over three different random seeds.

## 5 Experimental Results

The main results are presented in Table 1.

**Inference-only methods exhibit significant limitations in handling ambiguous queries**. Direct fails to manage ambiguous queries, as evidenced by its consistent zero F1$_a$ scores. Ambig-aware and Sample Rep demonstrate a strong bias towards clarification requests, exhibiting deficient F1$_u$. Self-Ask displays a subpar F1$_a$, indicating it is challenging to resolve ambiguity by just "asking" the model without task-specific training.

**Trained methods present enhanced performance compared to inference-only approaches.** Specifically, Subset$_{\text{Rand}}$ exhibits improved performance across both metrics compared to inference-only methods. Full-set demonstrates superior performance among the baselines, leveraging the entire training set. Notably, Subset$_{\text{Ent}}$ surpasses Subset$_{\text{Rand}}$ by a large margin and even outperforms Full-set in some datasets. The results of Subset$_{\text{Ent}}$ verify that entropy is capable of capturing ambiguity to some extent and is beneficial when incorporated into the alignment process.

**Apa achieves superior performance across all datasets.** Despite employing an identical inference template, Apa achieves a notable enhancement

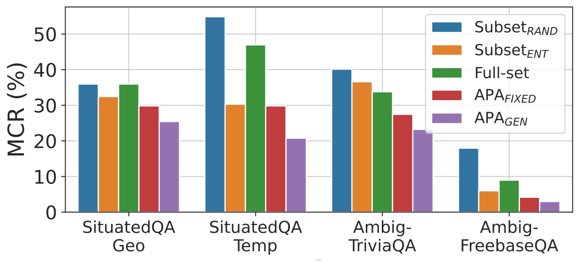
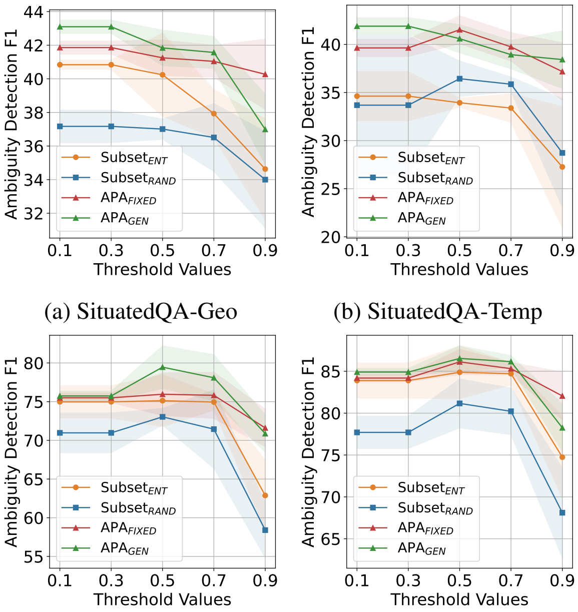


Figure 4: Misaligned Clarification Request Rate (MCR) of trained methods. Low MCR indicates that the model retains its intrinsic knowledge even after the alignment process. In all instances, APA exhibits the lowest MCR.



(a) SituatedQA-Geo

(b) SituatedQA-Temp

(c) AmbigTriviaQA

(d) AmbigFreebaseQA

Figure 5: Changes in the F1$_a$ score according to the threshold value. Regardless of the threshold value, APA consistently outperforms all the baselines.

in F1$_u$ compared to DIRECT. This improvement is especially surprising considering that APA was trained on $D_{\text{correct}}$, which consists of samples that the model is already capable of handling. Moreover, APA consistently outperforms across all the datasets in terms of F1$_a$, achieving gains up to 6 points. The results highlight the effectiveness of leveraging perceived ambiguity for alignment, enhancing generalization and robustness. When compared to SUBSET$_{\text{ENT}}$, the improvement of APA suggests that INFOGAIN provides better quantification of ambiguity than entropy. The efficacy of leveraging only the data perceived ambiguous, comprising approximately 32% in the LLAMA2 family and 13% in MISTRAL, again emphasizes the importance of data quality over quantity (Zhou et al., 2024; Chen et al., 2024). Furthermore, APA$_{\text{FIXED}}$ generally exhibits enhanced performance compared to APA$_{\text{GEN}}$. This is because APA$_{\text{GEN}}$ engages in a more challenging task of generating specific clarification requests.

| Method | SituatedQA-Geo | SituatedQA-Temp | Ambig-TriviaQA | Ambig-FreebaseQA |
|---|---|---|---|---|
| RAND | 39.31<br>(1.28) | 38.34<br>(0.44) | 72.05<br>(0.58) | 81.28<br>(1.88) |
| MIN | 34.95<br>(1.71) | 36.03<br>(0.90) | 70.30<br>(1.50) | 79.19<br>(2.02) |
| MAX | 40.96<br>(0.71) | 39.33<br>(0.88) | 73.95<br>(1.03) | 82.23<br>(1.31) |
| APA | 43.10<br>(0.39) | 41.89<br>(2.02) | 75.74<br>(1.52) | 84.90<br>(0.40) |

Table 2: Average and standard deviation (in parentheses) of F1$_a$ scores of different data selection methods. The first, second, and third best results are highlighted. Results show that utilizing INFOGAIN regardless of the ground-truth ambiguity is effective for data selection.

## 6 Ablation Study

In this section, we perform a series of ablation studies to further evaluate APA. Unless otherwise specified, all experiments are conducted on LLAMA2 7B across four datasets: SituatedQA-Geo, SituatedQA-Temp, AmbigTriviaQA, and AmbigFreebaseQA. Additional details are stipulated in Appendix E.

### 6.1 Analysis on Sample-level Misalignment

The alignment process of generating clarification requests for ambiguous queries may lead to a potential trade-off, where the model incorrectly generates clarification requests for unambiguous inputs that were previously well-handled. To assess such a case, we define **Misaligned Clarification Request Rate (MCR)**, which measures the proportion of unambiguous samples that were correctly answered (③ in Figure 3) before training but incorrectly shifted to erroneously generating clarification requests (⑤ in Figure 3) after alignment. A low MCR is desirable, representing that the model preserves its existing capabilities even after the alignment. We can observe from Figure 4 that, overall, APA consistently demonstrates the lowest MCR, indicating that the model successfully learns to handle ambiguity while effectively preserving the existing capabilities.

### 6.2 The Effect of Threshold Values

The number of training samples used for alignment depends on the threshold value $\epsilon$. To understand the

| Type | Generations |
|---|---|
| $x$<br>$\hat{x}_{\text{disambig}}$<br>$y_{\text{clarify}}$ | How many pages in a brave new world?<br>How many pages in **the 1932 edition of the book** brave new world **by Aldous Huxley**?<br>Your question is ambiguous. Which edition of the book are you interested in? |
| $x$<br>$\hat{x}_{\text{disambig}}$<br>$y_{\text{clarify}}$ | Who was the commander of the british forces in boston?<br>Who was the commander of the british forces in boston **during the american revolution?**<br>Your question seems ambiguous. Can you be more specific about the event or time? |

Table 3: Examples of generated $y_{\text{clarify}}$ and $\hat{x}_{\text{disambig}}$ from the initial query $x$. **Additional specification** from the disambiguation is highlighted in bold and the specification of the clarification requests are underlined.

impact of $\epsilon$ on performance, we conduct an analysis by applying different $\epsilon$ for ambiguous data selection. We compare SUBSET$_{\text{ENT}}$ and SUBSET$_{\text{RAND}}$, each with an equal number of training samples. Figure 5 presents the F1$_a$ scores measured under different $\epsilon$. In general, larger $\epsilon$ reduces the data available for training, resulting in declined performance. SUBSET$_{\text{RAND}}$ consistently demonstrates subpar performance, whereas SUBSET$_{\text{ENT}}$ is a strong baseline across all scenarios. Nevertheless, APA outperforms all the baselines across different $\epsilon$ values.

### 6.3 Impact of INFOGAIN for Data Selection

For a deeper analysis of INFOGAIN on data selection within APA, we conducted an ablation study by varying the criteria for selecting ambiguous data. With the correct dataset $D_{\text{correct}}$ held constant, we alter the strategies of selecting $m$ ambiguous samples as follows:

- **Random Selection (RAND)** We randomly select $m$ ground-truth ambiguous samples.
- **INFOGAIN-based Selection** We explore two different selection methods leveraging INFOGAIN: **MAX** selects top-$m$ samples with the largest INFOGAIN from the ground-truth ambiguous samples. **MIN** selects the bottom-$m$ samples with the minimum INFOGAIN among those that are ground-truth ambiguous.

APA differs from the baselines by utilizing samples perceived as ambiguous, allowing the potential inclusion of ground-truth unambiguous samples.

Table 2 demonstrates the overall results. RAND consistently lags behind MAX by a margin of 1 to 4 points. The disparity underscores the effectiveness of data selection based on INFOGAIN, even with ground-truth ambiguous samples. Moreover, APA outperforms all the baselines across all the datasets. Notably, even though the perceived ambiguity does not always coincide with ground-truth ambiguity, results show that exploiting model-perceived ambiguity significantly enhances alignment. MIN demonstrates the worst performance among the methods evaluated. We speculate that this decline is because the training samples with low INFOGAIN are perceived as unambiguous, yet are trained as ambiguous. This misalignment likely accounts for the degradation in performance.

### 6.4 Case Study

Table 3 demonstrates examples of generated disambiguation $\hat{x}_{\text{disambig}}$ and the clarification request $y_{\text{clarify}}$ from the query $x$. We can observe that the model generates factual specifications about the query leveraging its intrinsic knowledge (e.g., *1932 edition of the book*). Furthermore, given $x$ and $\hat{x}_{\text{disambig}}$, the model successfully generates a clarification request, specifically mentioning the factor that causes the ambiguity (e.g., *Which edition*). Further examples of disambiguations and failure cases are in Appendix F.

## 7 Conclusion

In this work, we present a novel alignment pipeline, dubbed **Alignment with Perceived Ambiguity (APA)**, designed to enhance the ability of LLMs to address ambiguities within queries, leveraging the model's intrinsic knowledge. Our method employs an implicit measure INFOGAIN to quantify the ambiguity perceived by the model itself. The model learns to effectively manage (un)ambiguous queries through alignment based on this metric. Experimental results demonstrate the effectiveness of APA, which outperform all the baselines across various QA datasets. As a future avenue, we plan to explore extending this methodology to broader domains and more complex types of ambiguities, further solidifying the role of LLMs in managing the inherent uncertainty present in NLP tasks.

## Limitations

The scope of our research is mainly focused on short-form QA tasks. The research scope could be expanded to long-form generation tasks such as detailed reasoning. Furthermore, there are cases when a query becomes ambiguous by considering additional contexts, e.g., cases in conversational QA (Guo et al., 2021). As our research focuses solely on situations where a single query is given, future work may consider scenarios where additional context is provided to the model. For experiments, we explore the most widely used models for evaluation, specifically LLAMA2 and MISTRAL. Despite this, a more comprehensive evaluation encompassing a broader range of LLMs could have enriched our findings, providing insights across different architectures and capabilities. Larger-scale models may exhibit different tendencies and, therefore, should be explored in future research. Furthermore, our work mainly focuses on supervised fine-tuning (SFT) as the alignment method. However, alternative methods, such as Reinforcement Learning from Human Preference (RLHF) (Ouyang et al., 2022) or Direct Preference Optimization (DPO) (Rafailov et al., 2023), could offer distinct advantages toward our objective.

## References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *Preprint*, arXiv:2305.13712.

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 Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi 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. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *Preprint*, arXiv:2402.08925.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. Alpagasus: Training a better alpaca model with fewer data. In *The Twelfth International Conference on Learning Representations*.

Jonathan H Choi. 2024. Measuring clarity in legal text. *U. Chi. L. Rev.*, 91:1.

Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, page 2201–2206, New York, NY, USA. Association for Computing Machinery.

Jeremy Cole, Michael 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*, pages 530–543, Singapore. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023.

RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*.

Romina Etezadi and Mehrnoush Shamsfard. 2023. The state of the art in open domain complex question answering: a survey. *Applied Intelligence*, 53(4):4124–4144.

H.A. Gleason. 1963. *Linguistics and English Grammar*. H.A. Gleason jr.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coQA: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Benjamin M Gyori, Charles Tapley Hoyt, and Albert Steppi. 2022. Gilda: biomedical entity text normalization with machine-learned disambiguation as a service. *Bioinformatics Advances*, 2(1):vbac034.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023a. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2023b. Ai alignment: A comprehensive survey. *Preprint*, arXiv:2310.19852.

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Daniel Jurafsky. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, 20(2):137–194.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023a. 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*, pages 996–1009, Singapore. Association for Computational Linguistics.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023b. $(\text{QA})^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 47669–47681. Curran Associates, Inc.

Sushant Kumar, Sumit Datta, Vishakha Singh, Sanjay Kumar Singh, and Ritesh Sharma. 2024. Opportunities and challenges in data-centric ai. *IEEE Access*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. 2023. Asking clarification questions to handle ambiguity in open-domain QA. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11526–11544, Singapore. Association for Computational Linguistics.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *CoRR*, abs/1811.07871.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024a. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. 2024b. Enhancing llm safety via constrained direct preference optimization. *Preprint*, arXiv:2403.02475.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Donald G Mackay and Thomas G Bever. 1967. In search of ambiguity. *Perception & Psychophysics*, 2:193–200.

A. Majeed and S. Hwang. 2023. Data-centric artificial intelligence, preprocessing, and the quest for transformative artificial intelligence systems development. *Computer*, 56(05):109–115.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–483, Nusa Dua, Bali. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. 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 *Thirty-seventh Conference on Neural Information Processing Systems*.

Sanford Schane. 2002. Ambiguity and misunderstanding in the law. *T. Jefferson L. Rev.*, 25:167.

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.

Mark Stevenson and Yikun Guo. 2010. Disambiguation in the biomedical domain: the role of ambiguity type. *Journal of biomedical informatics*, 43(6):972–981.

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`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Thomas Wasow, Amy Perfors, and David Beaver. 2005. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pages 265–282.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *Preprint*, arXiv:2402.04333.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Yuewei Yuan, Chaitanya Malaviya, and Mark Yatskar. 2023. AmbiCoref: Evaluating human and model sensitivity to ambiguous coreference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1023–1030, Dubrovnik, Croatia. Association for Computational Linguistics.

Lingxi Zhang, Jing Zhang, Xirui Ke, Haoyang Li, Xinmei Huang, Zhonghui Shao, Shulin Cao, and Xin Lv. 2023. A survey on complex factual question answering. *AI Open*, 4:1–12.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

# A Implementations Details

## A.1 Pipeline Details

For initial prediction assessment (Stage 1), we utilize the same inference template as DIRECT (Table 4) and disambiguate the given query with the template from Table 5. We use the greedy generation for the disambiguation. The threshold $\epsilon$ is empirically set to 0.1 for selecting ambiguous inputs. When balancing training set size, if $n > m$, we randomly select $m$ samples from $D_{\text{correct}}$, where $n = |D_{\text{correct}}|$ and $m = |D_{\text{ambig}}|$. If $n < m$, we select $n$ samples from $D_{\text{ambig}}$ with the largest INFOGAIN. For APA$_{\text{GEN}}$, we use the template from Table 6 to generate specific clarification requests for each ambiguous queries. Furthermore, for APA$_{\text{FIXED}}$, we randomly set $y_{\text{clarify}}$ from the following pre-defined phrases : [`The questions is ambiguous. Please clarify your question. Your question is ambiguous. Can you clarify your question? Your question is not clear. Can you clarify your question please?`]

## A.2 Training Details

For training, we applied AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 32. We selected the model with the best performance in the validation set from learning rates {`1e-3, 5e-4, 1e-4`} and training epochs {`1, 2, 3`}. All the experiments were implemented with Pytorch (Paszke et al., 2019) and Huggingface Transformers library (Wolf et al., 2020). For efficient training, we applied QLoRA from Huggingface PEFT library (Mangrulkar et al., 2022) with *r=4* and *alpha=16*. The training takes about half an hour on a single Tesla V100 GPU. All experiments are averaged over three different random seeds.

```
Answer the following question.
Question: <question>
Answer:
```

Table 4: Direct prompting template.

```
Evaluate the clarity of the input question. If the question is ambiguous, enhance it by adding specific details such as relevant locations, time periods, or additional context needed to resolve the ambiguity. For clear questions, simply repeat the query as is.

Example:
Input Question: When did the Frozen ride open at Epcot?
Disambiguation: When did the Frozen ride open at Epcot?

Input Question: What is the legal age of marriage in the USA?
Disambiguation: What is the legal age of marriage in each state of the USA, excluding exceptions for parental consent?

Input Question: <question>
Disambiguation:
```

Table 5: Disambiguation template used in Perceived Ambiguity Detection Stage of APA. We provide 2-shot demonstrations from AmbigQA train set.

```
Engage with the provided ambiguous question by extracting the key point of ambiguity, and interactively ask for clarification based on the disambiguated question.

Example 1:
Ambiguous Question: Who won?
Disambiguation: Who won the 2020 U.S. presidential election?
Clarification Request: Your question seems ambiguous. Could you specify which competition or event you are asking about?

Example 2:
Ambiguous Question: What’s the weather like?
Disambiguation: What’s the weather like in Miami today?
Clarification Request: Your question is ambiguous. Where are you interested in the weather report for?

Ambiguous Question: <ambiguous question>
Disambiguation: <disambiguation>
Clarification Request:
```

Table 6: Template for generating clarification request for the given ambiguous query. The model is prompted to extract the factor that causes the ambiguity and generate a clarification request based on the extracted factor.

The full results of APA and trained baseline methods with the standard deviation are demonstrated in Table 18.

## B Dataset Overview

### B.1 Dataset Details

This section stipulates the details of the datasets we used in the experiments. The statistics of ambiguous and unambiguous samples for each dataset is specified in Table 7.

**AmbigQA** (Min et al., 2020) is a derivative of the Natural Questions dataset (Kwiatkowski et al., 2019), designed to verify ambiguous data points. The dataset covers diverse sources of ambiguity, such as event and entity references. The dataset consists of pre-defined ambiguous and unambiguous queries, where unambiguous queries are labeled with ground-truth answers. We set AmbigQA as the in-domain dataset and utilize it for training and validation. Specifically, we follow the ambiguity defined by the dataset and train the model to generate ground-truth answers for unambiguous queries and pre-defined clarification requests for ambiguous queries. Further training details are stipulated in Appendix C.

**SituatedQA** (Zhang and Choi, 2021) focuses explicitly on temporal and geographic ambiguity from the input query. As the cause of ambiguity and its construction process are distinct, we assess performance on the temporal and geographic split separately, denoted as Temp and Geo, respectively.

| Dataset | Train | | Validation / Test | |
|---|---|---|---|---|
| | Unambig. | Ambig. | Unambig. | Ambig. |
| AmbigQA | 5,287 | 4,749 | 830 | 1,172 |
| SituatedQA-Geo | - | - | 506 | 129 |
| SituatedQA-Temp | - | - | 2,795 | 876 |
| AmbigTriviaQA | - | - | 500 | 500 |
| AmbigWebQuestions | - | - | 500 | 500 |
| AmbigFreebaseQA | - | - | 500 | 500 |

Table 7: Number of ambiguous and unambiguous samples for each datasets. We utilize AmbigQA for in-domain training and validation. The rest of the datasets are evaluated as OOD test sets.

```
Please make the following question
ambiguous. Your task is to introduce
ambiguity by altering the specificity
of the noun phrase or omitting crucial
details from the statement. Keep the
rest of the sentence unchanged except
for the modified sections. Generate
only the revised statement.

Question: <question>
Ambiguation:
```

Table 8: Template to ambiguate the input query for dataset construction. We prompt gpt-4o for the generation.

**TriviaQA** (Joshi et al., 2017) consists of question-answer-evidence triplets collected from Wikipedia and the web. For our experiments, we only utilize the question-answer pairs. We ambiguiate the subset of TriviaQA to build AmbigTriviaQA.

**WebQuestions** (Berant et al., 2013) is a question-answering dataset that uses Freebase as the knowledge base. The dataset consists of questions from the Google Suggest API and then answers obtained from Amazon Mechanical Turk. In creating AmbigWebQUestions, we applied ambiguity to the subset of WebQuestions.

**FreebaseQA** (Jiang et al., 2019) is an open-domain QA over the Freebase knowledge graph. The question-answer pairs are collected from various sources such as TriviaQA, QuizBalls, and QuizZone. AmbigFreebaseQA is derived from the subset of FreebaseQA.

```
An ambiguous question has multiple
valid answers. Is the following
question ambiguous with multiple
possible answers? Answer only in Yes
or No.

Question: <ambiguous generation>

Yes or No:
```

Table 9: Template for validating the generated ambiguated queries. We prompt gpt-4o for the validation. Samples with the output "Yes" are considered a valid ambiguation.

```
You are given an ambiguous question
and its possible ambiguation. Please
verify whether the ambiguous question
poses proper ambiguity. An ambiguous
question must have multiple valid
answers.

Original Question: <original question>
Ambiguous Question: <ambiguated
question>

Yes or No:
```

Table 10: Instructions for human validation for dataset construction. Samples selected as "Yes" are considered a valid ambiguation.

### B.2 Dataset Construction Details

To further examine the model's capability to interpret and generate responses to intentionally ambiguous queries, we constructed AmbigTriviaQA, AmbigWebQuestions, and AmbigFreebaseQA by ambiguating the TriviaQA, WebQuestions, and FreebaseQA, respectively. We first prompt gpt-4o to ambiguate the original question with the template from Table 8. To further validate the generation and control the dataset's quality, we again prompt gpt-4o for secondary verification. We utilize the template in Table 9 and collect samples verified as ambiguous. Validating the generations from the same model may pose unnecessary biases. To mitigate the potential biases in the validation process, we evaluate the verified samples with human annotators and select samples for the final dataset. (Table 10) This human-in-the-loop data construction ensures the quality and fairness of the dataset. The process yielded 1,000 question-answer pairs, with 500 ambiguous and 500 unambiguous pairs. Examples from AmbigTriviaQA are demon-

```
Answer the following question. If
the question is ambiguous, it is
proper to answer with “The question is
ambiguous”.
Question: <question>
Answer:
```

Table 11: Ambiguity-aware prompting. We explicitly describe how to handle ambiguity.

```
Answer the following question. Given
the question and answer, is the
question ambiguous or unambiguous?
Answer only ambiguous or unambiguous.
Question: <question>
Answer: <generated answer>

Is the question ambiguous or
unambiguous? Answer only ambiguous or
unambiguous.
Ambiguous or Unambiguous:
```

Table 12: Verification template for SELF-ASK. With the generated answer and the original question, the model is prompted to verify the ambiguity of the initial query.

strated in Table 14.

## C Baseline Details

In this section, we describe implementation details of the baselines.

**DIRECT** We make a direct inference using the template from Table 4. The greedy generation result with temperature 0 is used for evaluation.

**AMBIG-AWARE** We utilize the template from Table 11, where we explicitly describe how to handle ambiguity. Identically, we use the greedy generations for evaluation.

**SAMPLE REP** The template from Table 4 is used to generate a single greedy generation and ten sampled generations with sampling temperature of 1.0. We quantify the rate of sampled generations that match the greedy generation as the uncertainty measure, where 1.0 is the most certain and 0.0 being the least certain. Samples with the measure below a specific threshold are considered ambiguous. For instance, if three out of ten samples exactly match the greedy generation, then the uncertainty for the given query is 0.3. We empirically select a threshold that demonstrates the best F1$_u$ and F1$_a$ with the least trade-off.

**SELF-ASK** We initially prompt the model with the template from Table 4 and generate a greedy generation. Then, the initial query and the generated answer are utilized with the template from Table 12 and prompt the model to verify the query's ambiguity. We modified the prompt from Amayuelas et al. (2023) so that the model can specifically focus on ambiguity. The ambiguity detection is determined based on the model's final verification of "Yes" or "No".

**FULL-SET** The entire training set is utilized for training. Following APA$_{\text{FIXED}}$, we label the ground-truth ambiguous samples with pre-defined clarification requests as $y_{\text{clarify}}$. (Pre-defined clarification requests are listed in Appendix A.1.) The model is trained to generate $y$ for $x_{\text{unambig}}$ and $y_{\text{clarify}}$ for $x_{\text{ambig}}$ with the inference template from Table 4.

**SUBSET$_{\text{RAND}}$** The training method is identical to FULL-SET, but SUBSET$_{\text{RAND}}$ utilizes a subset of the training set. We randomly select $|D|$ samples from the training data, with the equal number ($|D|/2$) of ambiguous and unambiguous samples.

**SUBSET$_{\text{ENT}}$** The training of SUBSET$_{\text{RAND}}$ is identical to SUBSET$_{\text{RAND}}$ except the ambiguous sample selection method. When $x_{\text{ambig}}$ is given, we measure the entropy of the generated result from the model. A high entropy value indicates that the model is uncertain about the prediction of the ambiguous query. Therefore, among the $x_{\text{ambig}}$ in the train set, we select $|D|/2$ samples with the highest output entropy and use them as ambiguous samples.

## D Evaluation Details

In this section, we describe the evaluation details of our experiments. We utilize the greedy generation from the model for the evaluation.

### D.1 Unambiguous Query Evaluation

For unambiguous queries, we measure the quality of the generation by employing RougeL[4] (Lin and Och, 2004) with all the possible valid answers. The prediction from the model is regarded as correct if the score is above 0.3.

[4] https://huggingface.co/spaces/evaluate-metric/rouge

| Threshold | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| # Samples | 3,088 | 3,088 | 1,860 | 886 | 396 |

Table 13: Number of training samples for different threshold values. We vary the threshold value from 0.1 to 0.9.

### D.2 Ambiguous Query Evaluation

For ambiguous questions, we expect the model to generate clarification requests. Since there are various ways to express clarification requests, we use the following phrases to detect the requests. The presence of pre-defined ambiguity-related phrases in the model's output is treated as a successful detection. The pre-defined phrases are the follows: `[ambiguous, ambig, unclear, not clear, not sure, confused, confusing, vague, uncertain, doubtful, doubt, questionable, clarify, not clear]`

## E Details of Ablation Experiments

### E.1 Details of Sample-level Misalignment Analysis

To measure Misaligned Clarification Request rate (MCR), we start with a base model (e.g., LLAMA2 7B or MISTRAL 7B) which has not undergone any alignment training. We prompt the model using the template in Table 4 and select the correct, unambiguous samples. Subsequently, we evaluate the aligned models, such as FULL-SET, SUBSET$_{\text{ENT}}$, or APA$_{\text{GEN}}$, leveraging these pre-selected samples. We then count the cases where the aligned model's predictions shifted from providing correct answers to generating wrong clarification requests post-alignment. MCR is measured as the proportion of these shifted samples relative to the total number of initially correct, unambiguous samples. The metric quantified the extent to which the model's alignment process leads to unnecessary clarification requests for previous well-handled unambiguous queries.

### E.2 Details of Threshold Ablation

To measure the performance with different threshold values, we apply $\epsilon \in$ `{0.1, 0.3, 0.5, 0.7, 0.9}`. The number of selected samples for training is illustrated in Table 13.

### E.3 Details of Data Selection Ablation

This section details the data selection methods from Section 6.3, with the corresponding visualization in Figure 6. Consider the case where the ground-truth ambiguous and unambiguous queries are sorted based on their INFOGAIN. APA selects $m$-samples with the largest INFOGAIN regardless of the ground-truth ambiguity, focusing on perceived ambiguity. In contrast, RAND randomly selects $m$-samples as ambiguous from the ground-truth ambiguous queries (highlighted in blue in Figure 6). MAX and MIN select top-$m$ and bottom-$m$ samples regarding the INFOGAIN from the ground-truth ambiguous queries, respectively. Unlike the baseline methods, which only consider the ground-truth ambiguity, APA leverages the perceived ambiguity, which may not always align with the ground-truth ambiguity.

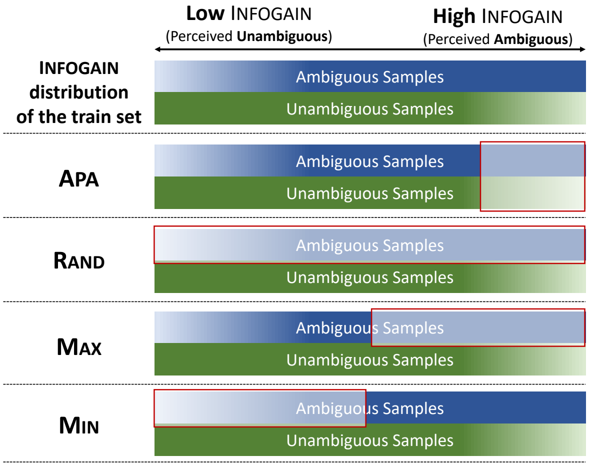


Figure 6: Illustration of ground-truth ambiguous and unambiguous samples sorted by the INFOGAIN. We highlight the chosen samples for each data selection method. APA selects samples with the largest INFOGAIN regardless of the ground-truth ambiguity. On the other hand, baseline methods select training data from ground-truth ambiguous samples with different selection strategies.

## F Additional Case Studies

### F.1 Failure Cases Before Alignment

Table 15 demonstrates generations by models before alignment for ambiguous queries from SituatedQA-Geo. Given the diverse denotations of the query, each model interprets the query differently based on their intrinsic knowledge. For instance, the first question is ambiguous due to the numerous possible "revolution" it could reference. Each model interprets "revolution" differently: LLAMA2 7B as the "*Russian* revolution", MISTRAL 7B as the "*French* revolution", and LLAMA2 13B as the "*American* Revolutionary War". Consequently, each model generates factual responses corresponding to its interpretation. We regard this phenomenon as problematic since the user likely has a specific “revolution” in mind while querying the model. However, the model may misinterpret the input and generate responses not aligned with the user’s intended reference. Consequently, this misalignment can lead to providing incorrect or irrelevant answers.

### F.2 Case Study of Disambiguations

Table 16 demonstrates examples of initial query $x$ and its disambiguation $\hat{x}_{\text{disambig}}$. The first example is when $x$ is inherently ambiguous, yet the model perceives it as unambiguous. Specifically, the model generates hallucination ("in the 1960s") where the song "don’t mess around with jim" was originally released in 1972. This non-factual generation would not provide any information gain to the model, classifying $x$ as ambiguous. In such a case, $x$ should be considered "unknown" with no related knowledge within the model. The second and third examples are correctly classified, as the model properly applies its intrinsic knowledge to perceive ambiguity. Regardless of the quantity of additional context generated, the model is capable of verifying its ambiguity. The last example is a misclassification as ambiguous. Despite disambiguation provides factually correct information ("1932 novel" and "by Aldous Huxley") for "brave new world", we speculate that the misclassification may arise from the existence of various media, such as movies and songs or even different versions of the book, sharing the title "brave new world".

### F.3 Failure Cases of Clarification Request Generation

Table 17 presents failure cases of clarification request generation. Even when the model successfully provides valid disambiguation (e.g., *in the usa* or *in 2015*), in some cases the model fails to consider the aspect that causes the ambiguity while generating clarification requests. For example, the first case generates "What is the book the title refers to?", which does not address the relevant ambiguity. Furthermore, the second example only requests for clarification and fails to provide further specifications regarding the ambiguity.

| Original Question | Ambiguated Question |
|---|---|
| Who wrote the 19th century **novel 'Anna Karenina'**? | Who wrote a 19th century **novel**? |
| What was **President Gerald Ford's** middle name? | What was the middle name of a **former U.S. president?** |
| Where in England was **actor Nigel Hawthorne** born? | Where in the UK was **the actor** born? |

Table 14: Examples of the original question and its ambiguation from AmbigTriviaQA. The **ambiguated phrase** is highlighted in bold.

| Question | Llama2 7B | Mistral 7B | Llama2 13B |
|---|---|---|---|
| When did the revolution begin? | The revolution began in 1917. *(Russian revolution)* | The revolution began in 1789. *(French revolution)* | The revolution began in 1775. *(American Revolutionary War)* |
| When did we host the world cup? | 1992 *(Not factual)* | 1994 *($15^{th}$ World Cup)* | 1996. *(Not factual)* |
| Who was the first elected president? | George Washington | George Washington | The first elected president of India was Dr. Rajendra Prasad. |

Table 15: Model predictions of ambiguous queries from SituatedQA-Geo before alignment. Ambiguous query, due to its variety of denotations, leads the model to interpret the query differently based on its intrinsic knowledge, consequently generating diverse answers.

| Model Prediction | Ground Truth | Type | Generated Text |
|---|---|---|---|
| Unambig. | Ambig. | $x$<br>$\hat{x}_{\text{disambig}}$ | Who sings don't mess around with jim?<br>Who sings don't mess around with jim, **in the 1960s?** |
| Unambig. | Unambig. | $x$<br>$\hat{x}_{\text{disambig}}$ | Who is winner in bigg boss season 5 kannada?<br>Who is the winner of the fifth season of the kannada **version of the indian reality television series bigg boss?** |
| Ambig. | Ambig. | $x$<br>$\hat{x}_{\text{disambig}}$ | How many jury members in a criminal trial?<br>How many jury members are required in a criminal trial **in the united states?** |
| Ambig. | Unambig. | $x$<br>$\hat{x}_{\text{disambig}}$ | How many pages in a brave new world?<br>How many pages are in the **1932 novel** brave new world **by aldous huxley?** |

Table 16: Example of initial query $x$ and its disambiguation $\hat{x}_{\text{disambig}}$. **Additional specification** from the model is in bold. Unambig. and Ambig. refers to Unambiguous and Ambiguous, respectively.

| Type | Generations |
|---|---|
| $x$ | Where was the old man and the sea published? |
| $\hat{x}_{\text{disambig}}$ | Where was the old man and the sea published **in the usa**? |
| $y_{\text{clarify}}$ | Your question is ambiguous. What is the book the title refers to? |
| $x$ | Who won last triple crown of horse racing? |
| $\hat{x}_{\text{disambig}}$ | Who won the triple crown of horse racing **in 2015**? |
| $y_{\text{clarify}}$ | I'm not sure about your question, could you provide some more information. |

Table 17: Failure cases of generated clarification request $y_{\text{clarify}}$ from the initial query $x$ and its disambiguation $\hat{x}_{\text{disambig}}$. **Additional specification** from the disambiguation is highlighted in bold. Despite the correct disambiguations, the model fails to generate clarification requests regarding the ambiguity.

| Method | LLAMA2 7B | | MISTRAL 7B | | LLAMA2 13B | |
|---|---|---|---|---|---|---|
| | F1$_u$ | F1$_a$ | F1$_u$ | F1$_a$ | F1$_u$ | F1$_a$ |
| **SituatedQA-Geo** | | | | | | |
| SUBSET$_{\text{RAND}}$ | 31.90 (3.29) | 37.17 (0.97) | 41.42 (3.08) | 33.95 (1.62) | 33.11 (3.21) | 36.87 (0.85) |
| SUBSET$_{\text{ENT}}$ | 39.33 (3.77) | 40.84 (0.28) | **47.34** (1.41) | 29.49 (4.36) | **40.19** (0.95) | 38.39 (1.80) |
| FULL-SET | 37.67 (1.87) | 41.45 (1.19) | 35.99 (1.18) | 41.28 (0.40) | 37.58 (1.71) | 38.39 (1.01) |
| APA$_{\text{FIXED}}$ | 39.99 (0.96) | 41.86 (0.39) | 38.43 (1.17) | 41.84 (0.39) | 31.31 (3.32) | **40.23** (0.40) |
| APA$_{\text{GEN}}$ | **41.01** (0.89) | **43.10** (0.39) | 39.55 (5.14) | **42.07** (1.13) | 34.04 (4.59) | 39.89 (2.10) |
| **SituatedQA-Temp** | | | | | | |
| SUBSET$_{\text{RAND}}$ | 29.48 (7.72) | 33.68 (7.24) | 34.14 (5.02) | 37.01 (0.82) | 28.57 (3.09) | 37.84 (1.39) |
| SUBSET$_{\text{ENT}}$ | 34.28 (1.52) | 34.62 (2.56) | 42.00 (1.71) | 32.04 (2.73) | 31.03 (2.02) | 38.00 (1.33) |
| FULL-SET | 29.59 (0.85) | 36.92 (1.43) | 31.16 (4.97) | 33.72 (8.36) | 29.41 (8.25) | 34.37 (8.93) |
| APA$_{\text{FIXED}}$ | 31.74 (1.16) | 39.63 (0.89) | **45.01** (2.06) | **43.95** (2.07) | **36.45** (0.38) | **42.18** (3.37) |
| APA$_{\text{GEN}}$ | **34.38** (0.40) | **41.89** (2.02) | 43.29 (3.69) | 40.70 (2.98) | 31.72 (3.24) | 39.36 (1.45) |
| **AmbigTriviaQA** | | | | | | |
| SUBSET$_{\text{RAND}}$ | 54.71 (2.26) | 70.97 (2.57) | 60.57 (0.81) | 67.82 (4.14) | 63.19 (3.06) | 73.52 (3.94) |
| SUBSET$_{\text{ENT}}$ | 58.83 (1.42) | 74.98 (2.09) | 62.17 (0.81) | 67.16 (4.14) | 64.95 (1.17) | 76.03 (0.86) |
| FULL-SET | 58.10 (0.66) | 71.25 (1.53) | 66.67 (0.66) | 76.38 (0.53) | 68.33 (0.82) | 76.82 (0.91) |
| APA$_{\text{FIXED}}$ | **62.97** (0.63) | 75.50 (0.62) | **70.70** (1.16) | **83.48** (0.59) | **70.83** (1.43) | **80.99** (1.67) |
| APA$_{\text{GEN}}$ | 59.27 (1.07) | **75.74** (1.52) | 67.73 (1.11) | 82.14 (1.76) | 69.25 (1.59) | 79.57 (1.74) |
| **AmbigWebQuestions** | | | | | | |
| SUBSET$_{\text{RAND}}$ | 38.69 (1.83) | 73.84 (1.67) | 45.16 (2.03) | 71.74 (1.75) | 44.31 (3.51) | 72.99 (2.36) |
| SUBSET$_{\text{ENT}}$ | 42.39 (1.36) | 75.86 (0.94) | 50.93 (5.43) | 71.11 (4.74) | 48.70 (1.19) | 77.43 (1.34) |
| FULL-SET | 40.46 (4.04) | 73.84 (1.67) | 41.83 (1.95) | 74.72 (0.40) | 47.20 (1.59) | 75.27 (0.75) |
| APA$_{\text{FIXED}}$ | **49.15** (2.57) | **77.07** (1.67) | **54.02** (2.17) | **81.07** (1.26) | **53.69** (0.97) | **79.22** (0.35) |
| APA$_{\text{GEN}}$ | 47.26 (1.01) | 76.64 (0.50) | 51.41 (0.92) | 79.54 (0.24) | 52.96 (3.46) | 78.46 (2.00) |
| **AmbigFreebaseQA** | | | | | | |
| SUBSET$_{\text{RAND}}$ | 63.59 (2.53) | 77.70 (1.93) | 70.60 (1.27) | 75.93 (4.66) | 70.40 (7.06) | 78.29 (5.35) |
| SUBSET$_{\text{ENT}}$ | 72.18 (0.87) | 83.89 (1.10) | 72.94 (2.97) | 77.17 (4.66) | 73.38 (0.89) | 81.93 (0.25) |
| FULL-SET | 69.97 (1.33) | 80.34 (1.19) | 76.98 (2.62) | 84.67 (3.08) | 76.56 (1.13) | 83.00 (0.69) |
| APA$_{\text{FIXED}}$ | **73.37** (0.40) | 84.19 (0.45) | **80.84** (0.69) | **90.12** (0.27) | **79.92** (2.82) | **88.03** (1.51) |
| APA$_{\text{GEN}}$ | 73.18 (0.74) | **84.90** (0.40) | 80.27 (1.32) | 89.22 (0.96) | 79.80 (2.14) | 87.61 (2.82) |

Table 18: Average and standard deviation (in parentheses) of the trained methods over three different random seeds. The **best method** is highlighted in bold and the second-best method is underlined.