Aligning Language Models to Explicitly Handle Ambiguity

Anonymous ACL submission

Abstract

In spoken languages, utterances are often shaped to be incomplete or vague for efficiency. This can lead to varying interpretations of the same input, based on different assumptions about the context. To ensure reliable user-model interactions in such scenarios, it is crucial for models to adeptly handle the inherent ambiguity in user queries. However, conversational agents built upon even the most recent large language models (LLMs) face challenges in processing ambiguous inputs, primarily due to the following two hurdles: (1) LLMs are not directly trained to handle inputs that are too ambiguous to be properly managed; (2) the degree of ambiguity in an input can vary according to the intrinsic knowledge of the LLMs, which is difficult to investigate. To address these issues, this paper proposes a method to align LLMs to explicitly handle ambiguous inputs. Specifically, we introduce a proxy task that guides LLMs to utilize their intrinsic knowledge to self-disambiguate a given input. We quantify the information gain from the disambiguation procedure as a measure of the extent to which the models perceive their inputs as ambiguous. This measure serves as a cue for selecting samples deemed ambiguous from the models' perspectives, which are then utilized for alignment. Experimental results from several question-answering datasets demonstrate that the LLMs fine-tuned with our approach are capable of handling ambiguous inputs while still performing competitively on clear questions within the task.

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 are frequently confronted with unexpected inputs from users, 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 a reliable user-friendly model, it is essential for the model to robustly handle such inputs. In this work, we seek to extend the scope of research to effectively handle invalid inputs. Specifically, we focus on managing "ambiguity" (Gleason, 1963; Mackay and Bever, 1967), which poses a significant challenge in Natural Language Processing (NLP) (Jurafsky, 1996).

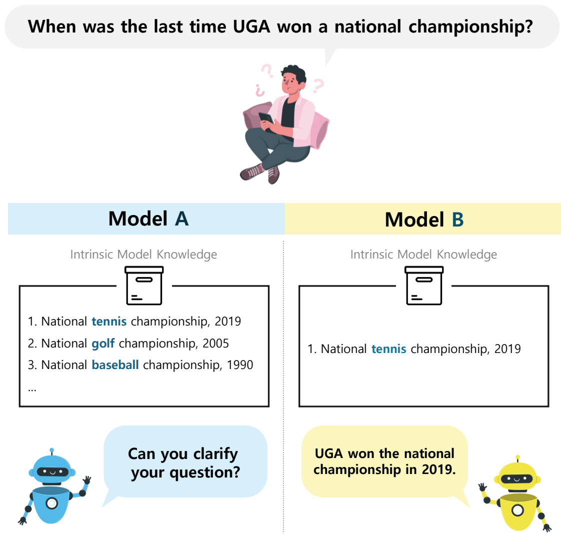

Figure 1: An example of an ambiguous query from AmbigQA (Min et al., 2020). The phrase "national championship" poses diverse denotations, causing ambiguity. Model possessing various related knowledge of the query could perceive it as ambiguous (left). On the other hand, if the model does not have sufficient related knowledge (right), the query can be perceived as unambiguous. Thus, the degree of ambiguity perceived by the model may vary even with identical inputs.

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, result in ambiguous requests. If the 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 especially pronounced in sensitive domains such as legal (Schane, 2002; Choi, 2024) or medical (Stevenson and Guo, 2010; Gyori et al., 2022) domains, where misinterpretations can lead to serious drawbacks. Despite the importance, approaches to robustly manage ambiguity are still significantly unexplored. In this paper, we endeavor to utilize the model's intrinsic knowledge to align the model in a manner that effectively handles ambiguity.

Properly processing ambiguous inputs is challenging primarily due to the following two hurdles. Firstly, models are **not directly trained to explicitly express ambiguity**. Even if a model perceives ambiguity, it is challenging to verify the recognition without explicit feedback. The second challenge is that the degree of ambiguity for the query can **vary depending on the intrinsic knowledge** of the model. 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". If a model possesses comprehensive knowledge across the possible denotations, it is plausible for the model to recognize the ambiguity (left). However, if the model's knowledge is limited to "national tennis championship", it would perceive the query as unambiguous (right). Therefore, it is essential to verify whether the input is deemed ambiguous from the model's point of view.

To overcome these issues, this paper proposes a method to align models to explicitly handle ambiguous queries. Specifically, we design a proxy task that guides the model to self-disambiguate a given query by utilizing its intrinsic knowledge. Then, we quantify the information gain from the disambiguation as an implicit measure of the extent to which the models perceive their inputs as ambiguous. This measure serves as a cue for selecting samples deemed ambiguous from the model's perspective, which are then utilized for alignment. Experimental results from several QA datasets demonstrate that the alignment process enables the model to properly clarify ambiguous inputs while maintaining its inherent capabilities. The findings underscore the value of assessing the perceived ambiguity, rather than relying solely on the ground-truth ambiguity. Furthermore, to provide a comprehensive framework for assessing ambiguity, we construct a new dataset dubbed AmbigTriviaQA. The dataset facilitates 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 defined as ambiguous if it has two or more distinct denotations (Wasow et al., 2005). Ambiguity challenges NLP applications by obscuring the intended meaning of expressions, leading to difficulties in accurately performing specific tasks. Efforts addressing 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 various answers, potentially not aligning 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) expands it to long-form generation. Furthermore, Cole et al. (2023) discovered that quantifying sampling repetition presents a reliable uncertainty measure for ambiguity, while Kim et al. (2023a) generates tree-of-clarification (ToC) that refines ambiguity within the inputs. As we share the goal of handling ambiguity, we adopt a novel approach of directly aligning the model to address ambiguity.

Alignment of LLMs LLMs are fundamentally 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). This has taken various forms, notably through Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a; Chakraborty et al., 2024), as well as 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) and safety (Bai et al., 2022b; Ji et al., 2023a; Liu et al., 2024b). Recent studies have concentrated on the factuality (Yang et al., 2023; Tian et al., 2024), avoiding hallucinations. Building on this foundation, our research extends the scope and aims to align models to effectively understand and handle ambiguities, which is a relatively unexplored area within the field of model alignment. This stands in contrast to previous methods which typically bypass the nuanced interpretation of ambiguous contexts inherent in language.

Data Quality Control for Alignment Data-centric AI (Chu et al., 2016; Majeed and Hwang, 2023; Kumar et al., 2024) emphasizes the importance of data quality in training models. In the context of the instruction following, LIMA (Zhou et al., 2024) demonstrates that models can be effectively aligned even with 1,000 human-curated samples. Similarly, AlpaGasus (Chen et al., 2024) shows effective alignment can be achieved by utilizing only a small subset of the Alpaca dataset (Taori et al., 2023) selected by ChatGPT. Various approaches for data selection have been explored, including those based on pre-defined quality factors such as length and complexity (Liu et al., 2024a), and utilizing gradient similarity from validation sets as a selection criterion (Xia et al., 2024). In this paper, we distinctly define data quality to effectively align models to handle ambiguity. To do so, we make use of the model's perceived ambiguity as an implicit cue for data quality.

3 Methodology

The objective of our approach is to align models to explicitly handle potentially ambiguous inputs leveraging intrinsic model knowledge. To this end, we propose a four-stage alignment pipeline, depicted in Figure 2. In this section, we first formulate the problem and describe each stage in detail.

Problem Formulation The goal of a QA task is to generate a factually correct answer y, given an unambiguous input x_{unambig}, a pre-defined inference template $t(\cdot)$, and a language model M. As we expand our input scope to ambiguous queries x_{ambig}, the model is expected to generate a clarification request y_{clarify}[1] for x_{ambig} to resolve the ambiguity, where the user is best positioned to clarify their intent.

[1]We have considered various approaches to handle ambiguity but were concluded to be impractical. Arbitrarily offering one of the valid answers may fail to reflect the user's intent, and presenting all possible answers is often impractical due to the potentially vast number of valid answers.

3.1 Explicit Prediction Stage

This initial stage involves assessing whether the model can appropriately handle each sample and identifying samples that the model currently fails to explicitly manage. By comparing the model's prediction with the ground-truth label, samples are categorized based on their response accuracy. We collect n correct samples, which the model can properly handle, as $D_{\text{correct}} = \{(x^i_{\text{correct}}, y^i_{\text{correct}})\}_{i=1}^n$ and incorrect samples are classified as $D_{\text{incorrect}}$.

3.2 Implicit Ambiguity Detection Stage

The objective of this stage is to identify samples that the model perceives as ambiguous from $D_{\text{incorrect}}$. Given that it is challenging for the model to explicitly express ambiguity, we construct a proxy task to estimate the ambiguity from the model's point of view.

The proxy task is designed to self-disambiguate x and implicitly measure the perceived ambiguity. Specifically, the model is first prompted to generate a disambiguation $\hat{x}_{\text{disambig}}$ for the input x. In this process, the model leverages its intrinsic knowledge related to x and generates further details. If x lacks specifications and the model possesses related knowledge necessary to compensate, then $\hat{x}_{\text{disambig}}$ would yield a higher certainty (lower entropy) for the model. 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 to 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:

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

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

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

where x is composed of I-tokens. We quantify the changes in input uncertainty by the difference in average entropy, which we define as **information gain** (INFOGAIN). The INFOGAIN from the disambiguation can be defined as the following:

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

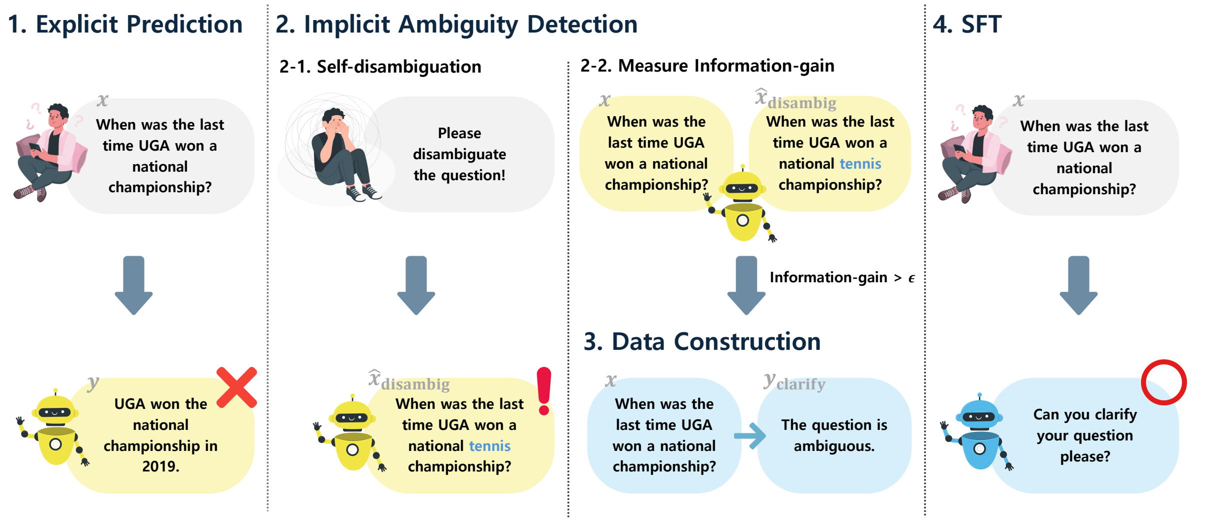

Figure 2: The overall process of our four-stage alignment pipeline. We initially filter samples that the model cannot explicitly handle (Stage 1), and self-disambiguate them to measure the information gain (Stage 2). Samples with high information gain are deemed ambiguous and utilized for supervised fine-tuning (Stage 3 & 4).

If the disambiguation results in a meaningful specification by utilizing intrinsic knowledge, a substantial INFOGAIN would be measured, suggesting that the model considers x as ambiguous. Conversely, a negligible INFOGAIN indicates that the model does not perceive x as ambiguous. Samples with INFOGAIN greater than the threshold ϵ are classified as ambiguous, denoted as x_{ambig}.

3.3 Data Construction Stage

In this stage, we construct datasets for the alignment process. This involves labeling m samples identified as ambiguous and constructing an ambiguous dataset $D_{\text{ambig}} = \{(x^j_{\text{ambig}}, y_{\text{clarify}})\}^m_{j=1}$. y_{clarify} serves as the ground-truth label for ambiguous samples, which are randomly selected from pre-defined clarification requests stipulated in Appendix D. To prevent the potential loss of the model's existing knowledge, we also incorporate D_{correct} for training. We balance the number of samples from both datasets so that $n = m$. The final training dataset is thus established as $D = D_{\text{correct}} + D_{\text{ambig}}$.

3.4 Supervised Fine-tuning (SFT) Stage

Utilizing the dataset $D = \{(x^k, y^k)\}^{n+m}_{k=1}$, the model is trained to generate ground-truth label y for input x, employing the identical inference template $t(\cdot)$. The model M with parameter θ is trained as follows:

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

4 Experimental Setting

4.1 Datasets

The capability of the model to perform within the trained domain is pivotal. However, its ability to generalize to out-of-distribution (OOD) is essential for real-world applicability, as queries that deviate from the training data are frequently confronted in the wild. To this end, we employ one training dataset and three OOD test sets to evaluate in diverse domains. All the datasets include both ambiguous and unambiguous queries.

AmbigQA Introduced by Min et al. (2020), AmbigQA is a derivative of the Natural Questions dataset (Kwiatkowski et al., 2019), designed to verify data points deemed ambiguous. The dataset covers diverse sources of ambiguity such as event and entity references. We set AmbigQA as the in-domain dataset and utilize it for training.

SituatedQA SituatedQA (Zhang and Choi, 2021) specifically focuses 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 split and the geographic split separately, denoted as Temp and Geo, respectively.

AmbigTriviaQA Since there are limited datasets for evaluating ambiguity in open-domain QA, we construct a new dataset, namely AmbigTriviaQA. By taking questions from the widely-used TriviaQA dataset (Joshi et al., 2017), we prompt gpt-3.5-turbo to ambiguate the initial query and verify the results. More details on dataset construction are described in Appendix B.

4.2 Baselines

To assess the effectiveness of our approach, we establish two sets of baselines: inference-only methods and trained methods. Further implementation details are described in Appendix A.

Inference-Only Methods Inference-only methods address ambiguity by directly prompting the model. We employ naïve prompting (**NAÏVE**) as a fundamental baseline, applying a simple QA prompt. Furthermore, we explore ambiguity-aware prompting (**AMBIGUITY-AWARE**), which additionally provides instructions on handling ambiguity. We also examine **SAMPLE REPETITION** (Cole et al., 2023) by measuring the consistency of the sampled generations. Finally, we evaluate **SELF-ASK** (Amayuelas et al., 2023), where the model initially generates an answer and subsequently determines the ambiguity based on the generation.

Trained Methods Given the lack of directly comparable prior work, we compare a fine-tuned baseline wherein the model is trained with the in-domain training set. **FULL-SET** applies the full in-domain training dataset and **SUBSET** is trained on a randomly selected subset of size $n+m$, which is equivalent in size to our approach. Additionally, we compare **HONESTY-TUNED** (Yang et al., 2023), which takes a similar approach utilizing an implicit measure named "expected accuracy" to estimate the model's factuality. The expected accuracy is measured as the average accuracy of sampled prediction for a single input query. We have revised the method so that incorrect samples based on "expected accuracy" are selected from the ground-truth ambiguous samples. Although HONESTY-TUNED shares similarities with our approach in the way of estimating the model's knowledge with an implicit measure, a notable distinction lies in our main focus on handling ambiguity beyond factuality.

4.3 Evaluation Metrics

As we expand the input scope to possibly ambiguous questions, the model should be capable of handling both unambiguous and ambiguous queries simultaneously. Therefore, we employ two widely used evaluation metrics to assess performance on both types of inputs. A successful alignment should preserve the model's capability to handle unambiguous inputs while successfully managing ambiguous queries. All evaluations are conducted by comparing the greedy generation to the ground truth.

Unambiguous Accuracy (Unambig. Acc.) While expanding the task scope, it remains crucial for the model to preserve the ability to handle unambiguous inputs. Thus, our analysis persists in exclusively evaluating the model's accuracy in processing unambiguous queries. We measure the quality of the generation by employing RougeL[2] (Lin and Och, 2004) with all the possible answers, where the prediction is regarded as correct if the score is above 0.3.

Ambiguity Detection F1-score (Ambig. F1) The model should be capable of detecting ambiguity and generating clarification requests for ambiguous inputs. However, especially for trained methods, models may exhibit biased predictions toward clarification requests. Taking these aspects into account, we evaluate the model's ambiguity detection capability with F1-score, which captures both the precision and recall of prediction, offering a balanced view of the model's ambiguity detection performance. Further details on the detection process are described in Appendix C.

4.4 Implementation Details

For our experiments, we utilize LLAMA2 7B & 13B (Touvron et al., 2023) and MISTRAL 7B (Jiang et al., 2023). We utilized QLoRA (Dettmers et al., 2023) to facilitate efficient training. Implementation details are stipulated in Appendix D.

5 Experimental Results

The main results of our experiments are presented in Table 1. **Inference-only methods exhibit a pronounced deficiency in handling ambiguous queries**. Specifically, NAÏVE establish poor performance in responding to ambiguous queries, resulting in a notably low F1-score. AMBIGUITY-AWARE demonstrates a strong bias towards clarification requests, as it achieves a relatively high F1-score at the expense of accuracy. Similarly, SAMPLE REPETITION exhibits a substantial trade-off between F1-score and accuracy. SELF-ASK displays a subpar F1-score, indicating that it is

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

Method	# Train Samples	AmbigQA		SituatedQA (Geo)		SituatedQA (Temp)		Ambig-TriviaQA	
		Unambig. Acc.	Ambig. F1	Unambig. Acc.	Ambig. F1	Unambig. Acc.	Ambig. F1	Unambig. Acc.	Ambig. F1
LLAMA2 7B									
NAÏVE	0	28.43	0.00	22.53	0.00	21.72	0.00	**62.46**	0.00
AMBIGUITY-AWARE	0	4.94	68.95	3.95	32.44	1.72	35.53	22.63	61.03
SAMPLE REPETITION	0	5.42	**74.96**	3.56	34.43	4.40	38.43	29.58	**65.88**
SELF-ASK	0	26.75	13.41	21.54	8.18	19.68	18.48	61.29	3.84
HONESTY-TUNED	3,088	18.19	68.00	9.49	36.56	7.98	37.57	52.42	54.60
SUBSET	3,088	29.16	64.23	23.72	36.85	13.95	36.74	51.71	55.70
FULL-SET	10,036	**37.11**	59.98	**26.09**	41.15	18.93	35.38	58.25	49.96
OURS	3,088	27.23	63.69	24.51	**42.05**	**21.90**	**40.77**	53.41	61.34
MISTRAL 7B									
NAÏVE	0	13.01	0.00	7.51	0.00	10.91	0.00	37.52	0.00
AMBIGUITY-AWARE	0	9.76	54.74	2.17	26.01	5.33	22.48	27.87	44.94
SAMPLE REPETITION	0	4.70	51.21	2.57	29.25	2.15	29.64	25.54	32.54
SELF-ASK	0	13.01	0.00	7.51	0.00	10.91	0.00	37.52	0.00
HONESTY-TUNED	1,382	16.51	**69.28**	11.66	33.84	8.09	39.16	41.70	**65.19**
SUBSET	1,382	33.49	60.91	29.05	35.64	19.11	37.38	58.87	54.61
FULL-SET	10,036	**43.73**	62.85	24.11	40.81	26.76	24.14	**65.04**	49.63
OURS	1,382	37.23	50.31	**32.21**	**42.18**	**35.74**	**40.17**	58.14	58.93
LLAMA2 13B									
NAÏVE	0	31.33	0.00	22.53	0.00	22.90	0.00	60.49	0.14
AMBIGUITY-AWARE	0	3.37	70.44	3.16	33.10	2.22	36.66	16.96	**64.17**
SAMPLE REPETITION	0	10.00	**71.10**	6.32	32.85	9.45	37.87	39.41	63.40
SELF-ASK	0	31.33	0.00	22.53	0.00	22.90	0.00	60.49	0.14
HONESTY-TUNED	3,216	17.83	68.57	3.16	33.29	3.58	38.03	48.13	60.39
SUBSET	3,216	35.18	62.89	23.12	36.19	20.50	39.00	59.72	57.26
FULL-SET	10,036	**43.61**	62.87	23.12	38.37	19.68	24.07	**66.80**	48.81
OURS	3,216	37.83	58.15	**24.51**	**41.59**	**24.36**	**41.09**	63.74	55.23

Table 1: Experimental results of in-domain dataset and three OOD datasets. We report unambiguous accuracy (Unambig. Acc.) and ambiguity detection F1-score (Ambig. F1). For each dataset, the **best method** is highlighted in bold and the second-best method is underlined. Our method demonstrates comparable performance in-domain and outperforms all the baselines in the OOD setting.

challenging to resolve ambiguity by explicitly "self-asking" the model.

Trained methods exhibit enhanced performance overall compared to inference-only approaches. HONESTY-TUNED struggles to handle ambiguity, as it also demonstrates biased detection performance. This is likely because the implicit measure from HONESTY-TUNED can be influenced by various factors but not specifically ambiguity. The results underscore the necessity of distinct methods for perceiving ambiguity. Compared to HONESTY-TUNED, SUBSET exhibits relatively balanced performance across both metrics. FULL-SET demonstrates the most superior performance among the baselines, particularly in the in-domain setting, as it has access to the ground-truth ambiguity.

Our approach yields comparable results in in-domain and demonstrates superior OOD performances. Despite employing identical inference templates as NAÏVE, our method demonstrates equal or improved unambiguous accuracy. This indicates the effectiveness of our alignment in managing ambiguity while preserving the inherent capabilities of the model. We can also observe an improvement in unambiguous accuracy, particularly in SituatedQA splits. It is especially surprising given that our method was trained on D_{correct}, which the model is already capable of handling. Compared to FULL-SET, we note a slight decline in the in-domain performance, an expected result given that FULL-SET is optimized with ground-truth ambiguity of the in-domain data. However, our method outperforms FULL-SET across OOD datasets in F1-score up to 17 points. This discrepancy underscores the effectiveness of utilizing perceived ambiguity for alignment, facilitating superior generalization and robustness. The efficacy

Method	AmbigQA		SituatedQA (Geo)	
	Unambig. Acc.	Ambig. F1	Unambig. Acc.	Ambig. F1
RANDOM	29.40	**64.74**	18.38	36.23
IMPLICIT	**29.52**	62.23	**24.90**	41.19
OURS	27.23	63.69	24.51	**42.05**
Method	**SituatedQA (Temp)**		**Ambig-TriviaQA**	
	Unambig. Acc.	Ambig. F1	Unambig. Acc.	Ambig. F1
RANDOM	16.99	38.76	**56.13**	50.63
IMPLICIT	21.36	40.11	55.65	59.16
OURS	**21.90**	**40.77**	53.41	**61.34**

Table 2: Ablation results of ambiguous data selection. In-domain dataset is highlighted in gray. **The best method** is in bold, and the second best method is underlined. Results show that perceived ambiguity measured by INFOGAIN is an effective cue for data selection.

Method	VAR (↑)	MCR (↓)	OAP (↑)
AmbigQA			
HONESTY-TUNED	**73.81**	44.49	20.48
SUBSET	62.12	33.05	20.79
FULL-SET	52.82	**18.64**	21.48
OURS	65.19	30.51	**22.65**
SituatedQA (Geo)			
HONESTY-TUNED	**93.80**	58.77	19.34
SUBSET	63.57	35.96	20.35
FULL-SET	72.09	35.96	23.08
OURS	89.15	**28.95**	**31.67**
SituatedQA (Temp)			
HONESTY-TUNED	**85.73**	56.34	18.71
SUBSET	68.61	54.86	15.48
FULL-SET	58.56	46.95	15.53
OURS	72.95	**36.08**	**23.31**
AmbigTriviaQA			
HONESTY-TUNED	42.39	11.64	18.73
SUBSET	48.10	17.57	19.83
FULL-SET	38.93	**9.81**	17.56
OURS	**57.45**	16.91	**23.87**

Table 3: VAR, MCR, and OAP of trained methods. A high OAP is preferred, with a high VAR and a low MCR. **The best performance** is in bold, and the second best performance is underlined. Our approach demonstrates the best OAP across all datasets, with the least trade-off between VAR and MCR.

of leveraging only the data perceived ambiguous (about 32% in the LLAMA2 family and 13% in MISTRAL) emphasizes the importance of data quality over quantity (Zhou et al., 2024; Chen et al., 2024).

6 Ablation Study

6.1 Impact of INFOGAIN for Data Selection

For a deeper analysis of the influence of INFOGAIN for data selection within our pipeline, we conduct an ablation study by varying the criteria for selecting ambiguous data. While maintaining the same D_{correct} for unambiguous samples, we alter the selection of m samples labeled as ambiguous. We compare the following data selection strategies:

- **Random Selection (RANDOM)** We randomly select m ground-truth ambiguous samples, without any consideration of INFOGAIN.

- **Implicit Measure-based Selection (IMPLICIT)** We select top-m samples with the largest INFOGAIN among those that are ground-truth ambiguous. It differs from our approach as our method utilizes samples perceived as ambiguous, allowing the potential inclusion of unambiguous samples.

Table 2 is the ablation results on LLAMA2 7B. For AmbigQA, baselines leveraging ground-truth ambiguity slightly outperform our method, which is a similar tendency from Section 5 where FULL-SET exhibit better in-domain performance. However, across OOD datasets, our approach demonstrates significantly superior performance. Specifically, RANDOM demonstrates a notable drop in F1-score by up to 10 points, illustrating the limitations of simply utilizing the ground-truth ambiguity, which might not align with the model's perceived ambiguity. Furthermore, IMPLICIT surpasses RANDOM by up to 5 points in F1-score, validating the effectiveness of INFOGAIN as a cue for data selection. Finally, our approach outperforms IMPLICIT across the majority of metrics, even with unambiguous samples selected for training, again highlighting the effectiveness of INFOGAIN as the data selection measure.

6.2 Analysis on Sample-level Prediction Change

Our method is designed to align the model to generate clarification requests for ambiguous queries. However, the process may lead to a potential

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

Table 4: Example of initial query x and its disambiguation $\hat{x}_{\text{disambig}}$ generated by LLAMA2 7B. **Additional specification** from the model is in bold. Unambig. and Ambig. refers to Unambiguous and Ambiguous, respectively.

trade-off, where the model erroneously generates clarification requests for unambiguous inputs that were previously well-handled. To assess this balance, we introduce three metrics: **Valid Alignment Rate (VAR)** measures the proportion of ambiguous samples incorrectly handled before alignment that are correctly addressed post-alignment and **Misaligned Clarification Rate (MCR)** measures the rate of correct unambiguous samples before training that erroneously generates clarification requests after alignment. A high VAR is desirable, whereas a low MCR is preferred simultaneously. Inspired by Yang et al. (2023), we additionally define **Overall Alignment Performance (OAP)** that measures the balance between VAR and MCR.

$$\mathsf{OAP} = \frac{\mathsf{VAR} + (1 - \mathsf{MCR})}{2} \qquad (5)$$

Table 3 compares the results on LLAMA2 7B. HONESTY-TUNED exhibits high VAR but poor MCR, implying a tendency to misinterpret known knowledge as ambiguous. This aligns with the previous results where HONESTY-TUNED displays biased generation towards clarification requests. On the other hand, FULL-SET and SUBSET demonstrate a good MCR and a relatively low VAR. Our method performs superior OAP, successfully addressing ambiguities (high VAR) while preserving existing capabilities (low MCR).

7 Self-disambiguation Case Study

Table 4 demonstrates examples of initial query x and its disambiguation $\hat{x}_{\text{disambig}}$ generated by LLAMA2 7B. 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, sharing the title "brave new world", leading to an erroneous integration of knowledge.

8 Conclusion

In this paper, we present a novel alignment pipeline designed to enhance the ability of LLMs to address ambiguities within queries, leveraging the model's intrinsic knowledge. Our method employs an implicit measure, dubbed INFOGAIN, to quantify ambiguity as perceived by the model. Through alignment based on the measure, the model learns to explicitly handle ambiguous as well as unambiguous queries. Experimental results demonstrate the effectiveness of our alignment, particularly pronounced in out-of-distribution scenarios. Results indicate the importance of alignment based on the model's perceived ambiguity. Future work may explore the extension of 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

For the experiments, we explore the most widely used models for evaluation, specifically LLAMA2 and MISTRAL. Despite this, a more comprehensive evaluation encompassing a broader consideration of LLMs could have enriched our findings, providing insights across different architectures and capabilities. Larger models in scale could demonstrate different tendencies and should be explored for future work. 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 towards our objective. Finally, the experiments are mainly focused on short-form QA tasks. The research scope could be expanded to long-form generation tasks such as detailed reasoning.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile 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. 2022b. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

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.

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.

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. $(\mathrm{QA})^2$: 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 Baseline Details

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

NAÏVE We make a direct inference using template from Table 5. We evaluate the greedy generation result with temperature 0.

AMBIGUITY-AWARE We utilize the template from Table 6, where we explicitly describe how to handle ambiguity. Identically, we use the greedy generations for evaluation.

SAMPLE REPETITION Template from Table 5 is used to generate a single greedy generation and 10 sampled generations with sampling temperature 1.0. We measure the rate of sampled generations that match the greedy generation, which is reported to be best-calibrated (Cole et al., 2023). Samples with the measure below a specific threshold is considered ambiguous. We empirically select a threshold that demonstrates the best accuracy and F1-score with the least trade-off.

SELF-ASK We initially prompt the model with the template from Table 5 and generate a greedy generation. Then, the initial query and the generated answer is utilized with the template from Table 7 and prompt the model to verify the ambiguity of the query. 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.

HONESTY-TUNED The approach involves measuring the "expected accuracy" of sampled generations. The expected accuracy is measured by generating 10 samples with temperature 1.0 and measuring the average accuracy among the generations. Ground-truth ambiguous samples with expected accuracy below the specific threshold, in this case 0.1 from Yang et al. (2023), are labeled ambiguous. The samples classified as ambiguous is re-labeled with y_{clarify}. The model M is trained to generate the ground-truth y given the input query x and the inference template $t(\cdot)$ from Table 5.

FULL-SET The full training set is utilized for training. The ground-truth ambiguous samples are

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

Table 5: Naïve prompting template.

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

Table 6: Ambiguity aware prompting. We explicitly describe how to handle ambiguity.

labeled with y_{clarify}. The model is trained to generate y given input question x and inference template $t(\cdot)$ from Table 5.

SUBSET We randomly select $|D|$ samples from the training data, with the same number ($|D|/2$) of ambiguous and unambiguous samples. SUBSET is trained in a same way as FULL-SET.

B AmbigTriviaQA Construction Details

AmbigTriviaQA is constructed by ambiguating the widely-used TriviaQA dataset (Joshi et al., 2017). We first prompt `gpt-3.5-turbo` to ambiguate the original question with the template from Table 8. To further validate the generation and control the quality of the dataset, we prompt `gpt-3.5-turbo` again for a secondary verification. We utilize the template in Table 9 and collect samples verified as ambiguous. This process yielded a total of 4,374 question pairs to examine the model's capability to interpret and generate responses to intentionally ambiguous queries. Examples from AmbigTriviaQA are demonstrated in Table 10.

C Ambiguity Detection Details

For ambiguous questions, we expect the model to generate clarification requests. Since there are various ways to express clarification requests, we use the following list of 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]`

```
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 7: 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.

```
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 from TriviaQA. We prompt `gpt-3.5-turbo` for the generation.

D Implementations Details

D.1 Pipeline Details

For explicit prediction (Stage 1), we utilize the same inference template as NAÏVE (Table 5) and the disambiguation is generated with the template from Table 11. We use the greedy generation for the disambiguated output. The threshold ϵ is set to 0.1 for filtering ambiguous inputs. For balancing training set size, if $n > m$, we randomly select m samples from D_{correct}, where $n = |D_{\text{correct}}|$ and $m = |D_{\text{ambig}}|$. If $n < m$, we select n samples from D_{ambig} with the largest INFOGAIN. Finally, for y_{clarify}, we randomly select 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`

```
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-3.5-turbo` for the validation. Samples with the output "Yes" are considered a valid ambiguation and are selected as AmbigTriviaQA.

`not clear. Can you clarify your question please?]`

D.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 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.

Original Question	Ambiguated Question
Which **volcano** in Tanzania is the highest mountain in Africa?	Which **geological formation** in Tanzania holds the title for the tallest landform in Africa?
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 10: Example of the original question and its ambiguation from AmbigTriviaQA. The **ambiguated phrase** is highlighted in bold.

```
The question seems ambiguous,
potentially requiring further details
for clarity. Please disambiguate
by providing specific context or
constraints related to the question.
This could include specifying any
relevant time periods, locations, or
other criteria necessary to narrow
down possible interpretations. No
additional specification is necessary
for an unambiguous question.

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 usa?
Disambiguation: What is the legal age
of marriage, without parental consent
or other authorization, in all but two
states in the usa?

Input Question: <question>
Disambiguation:
```

Table 11: Disambiguation template for implicit ambiguity measure. We provide 2-shot demonstration from AmbigQA train set.