Disagreement as a Signal for Ambiguity-Aware Abstention in Question Answering

Anonymous ACL submission

Abstract

Large language models often answer ambiguous or under-specified questions with unwarranted confidence. We present DiVA, a disagreement-based abstention framework that aggregates four complementary reasoning agents and uses cross-agent disagreement to decide whether to answer or abstain. We evaluate DiVA on AmbigQA and on a paired benchmark constructed from SituatedQA-temp, capturing both natural ambiguity and controlled missing-context ambiguity. Across Llama, Gemma, and Mistral, DiVA consistently outperforms majority voting under ambiguity-aware evaluation. We further show that DiVA induces a clear abstention–accuracy frontier: stronger abstention yields higher ambiguity-aware performance, with consistent trends across datasets and model families. These results establish multi-agent disagreement as a robust signal for ambiguity-aware abstention and support a broader evaluation paradigm in which systems are rewarded not only for answering correctly, but also for withholding answers when the question remains insufficiently specified.

1 Introduction

Large language models are increasingly used in open-ended question answering settings where not every question should be answered immediately. Some questions are genuinely ambiguous, while others are under-specified and require additional temporal, geographic, or contextual information (Min et al., 2020; Zhang and Choi, 2021). In such cases, forcing a model to answer can produce confident but incorrect outputs.

Standard QA evaluation nevertheless rewards answer production more directly than abstention. We instead study ambiguity-aware evaluation, which rewards correct answers on answerable questions and abstention on genuinely ambiguous or under-specified ones. Our method, **DiVA** (**Di**verse agent disagreement for **V**erified **A**bstention), queries four agents with different reasoning stances and aggregates their outputs through an abstention-aware decision rule.

We evaluate DiVA in two complementary settings: **AmbigQA**, a naturally mixed benchmark containing both ambiguous and answerable questions, and a paired benchmark constructed from **SituatedQA-temp**, which isolates missing-context ambiguity. Across both datasets, DiVA substantially outperforms majority voting under ambiguity-aware evaluation. Moreover, by varying how aggressively disagreement triggers abstention, we obtain a clear abstention–accuracy frontier across Llama, Gemma, and Mistral, showing that disagreement is operationally useful as a tunable abstention signal.

Contributions. We make three contributions:

1. We propose DiVA, a disagreement-based abstention framework over diverse reasoning agents.

2. We show that DiVA improves ambiguity-aware performance on both natural ambiguity (AmbigQA) and controlled missing-context ambiguity (SituatedQA-temp).

3. We show that DiVA yields a tunable abstention–accuracy frontier across three model families.

2 Related Work

This work sits at the intersection of ambiguous question answering, selective prediction, and multi-agent reasoning. Prior work on ambiguous QA emphasizes preserving multiple valid interpretations rather than forcing a single collapsed answer (Min et al., 2020), while SituatedQA highlights the

[1]Code and Dataset: https://anonymous.4open.science/r/DiVA-LLM-Reasoning-0DF2/README.md

importance of extra-linguistic context for context-dependent questions (Zhang and Choi, 2021). In parallel, selective prediction asks when systems should abstain to improve reliability (Xin et al., 2021), including directly for ambiguous questions (Cole et al., 2023). Multi-agent and ensemble methods typically use multiple prompts or roles to improve answer quality (Wang et al., 2023; Du et al., 2024; Chen et al., 2024). Our focus is different: we use disagreement not for answer synthesis, but for deciding when answering should be deferred.

3 Formulation and Method

3.1 Task Formulation

Given a question q, a set of agents $\mathcal{A} = \{a_1, \ldots, a_k\}$ produces outputs

$$\{y_i\}_{i=1}^{k},$$

where each y_i is either an answer string or abstention. DiVA aggregates these outputs and returns either a final answer $\hat{y}$ or abstention .

We evaluate under ambiguity-aware accuracy. If a question is answerable, the correct behavior is to answer correctly. If a question is ambiguous or under-specified, the correct behavior is to abstain. Formally, for dataset $\mathcal{D}$:

$$\text{Acc}_{\text{amb}} = \frac{C_{\text{ans}} + C_{\text{abs}}}{|\mathcal{D}|}. \quad (1)$$

Here, C_{ans} is the number of correct answers on answerable items, and C_{abs} is the number of correct abstentions on ambiguous items.

3.2 Decision-Theoretic Motivation

DiVA can be motivated by a utility-based view of abstention:

$$\begin{aligned} U(\text{correct answer}) &= 1, \\ U(\text{wrong answer}) &= -\lambda, \\ U(\text{abstain}) &= 0. \end{aligned} \quad (2)$$

where $\lambda > 0$ represents the cost of an incorrect confident answer. If disagreement among agents lowers the estimated probability that a candidate answer is correct, abstention becomes rational whenever the expected utility of answering falls below zero. In this view, disagreement functions as a proxy for epistemic uncertainty.

3.3 Diverse Agents and Pipeline

DiVA uses four fixed agent roles:

- **Literal**: answers via the most direct reading of the question.
- **Skeptic**: stresses uncertainty, edge cases, and possible underspecification.
- **Creative**: explores broader or less obvious interpretations.
- **Evidence**: prefers grounded, support-seeking answers and is more willing to abstain.

These roles create intentionally diverse inductive biases. Agreement across them suggests a stable interpretation; disagreement suggests ambiguity or missing context.

For each question, all four agents are prompted independently and return a final answer or abstention, together with optional self-reported confidence metadata. DiVA then: (1) extracts the final answer field from each response, (2) normalizes answer strings for comparison, (3) groups matching answers and counts abstentions, and (4) classifies the resulting pattern into one of four disagreement levels: *none*, *weak*, *strong*, or *complete*. Intuitively, *none* corresponds to stable agreement, *weak* to a clear leading interpretation with limited dissent, *strong* to substantial conflict or grounded refusal, and *complete* to the absence of stable support for any single interpretation.

DiVA then applies a policy-specific decision rule:

- **Lenient**: abstain only under the strongest ambiguity signals.
- **Balanced**: abstain under stronger disagreement or grounded refusals.
- **Strict**: abstain under weaker disagreement signals as well.

If the selected policy permits answering, DiVA returns the majority-supported normalized answer; otherwise, it abstains. These policies define the abstention frontier studied in Section 5.2.

4 Experimental Setup

4.1 Datasets

AmbigQA. AmbigQA (Min et al., 2020) serves as our main natural ambiguity benchmark (800

examples). It contains both ambiguous and answerable questions, making it well-suited for ambiguity-aware evaluation.

SituatedQA-temp. We construct a paired benchmark from temporal SituatedQA (Zhang and Choi, 2021) examples in which each question is evaluated both before and after the relevant temporal context is supplied. This yields a controlled ambiguity setting that complements AmbigQA by isolating missing-context effects while preserving short-answer evaluation. We report combined performance over 800 total examples (400 underspecified and 400 clarified).

4.2 Models and Baselines

We evaluate three base models: Llama 3.1 8B, Gemma 2 2B, and Mistral 7B. We compare:

- **Majority**: majority vote over non-empty agent answers,
- **Oracle**: correct if any agent exhibits the correct behavior,
- **DiVA**: our disagreement-based abstention method.

Prompt templates, policy definitions, and implementation details are included in the supplementary material.

5 Results

5.1 Main Results

Figure 1 summarizes the best strict-policy operating point for each model on each benchmark. DiVA outperforms majority voting in every setting, with the largest gains appearing where ambiguity is most consequential.

On AmbigQA, the strict policy gives the strongest results for all three models: 49.25% for Llama, 48.62% for Gemma, and 48.38% for Mistral. These scores are well above the corresponding majority baselines of 23.62%, 14.37%, and 17.75%.

On SituatedQA-temp, the best combined scores are 59.75% for Llama, 58.38% for Gemma, and 63.25% for Mistral, compared to majority baselines of 18.88%, 12.12%, and 17.50%. The same mechanism therefore transfers from natural ambiguity to a controlled missing-context setting.

5.2 Abstention–Accuracy Frontier

Figure 2 shows the main empirical pattern of the paper: higher abstention leads to higher ambiguity-aware accuracy.

On AmbigQA, accuracy increases monotonically from lenient to strict for all three models. For Llama, DiVA rises from 36.25% to 49.25%; for Gemma, from 39.88% to 48.62%; and for Mistral, from 30.38% to 48.38%. Most of the gain is already captured by the balanced policy, with the strict setting adding a smaller final improvement.

On SituatedQA-temp, the same trend holds. Moving from lenient to strict raises combined DiVA accuracy from 36.00% to 59.75% for Llama, from 45.88% to 58.38% for Gemma, and from 36.00% to 63.25% for Mistral. The median combined score across the three models increases from 36.00% to 59.75%.

The frontier is easy to interpret. Lenient policies preserve more coverage but allow more risky answers. Balanced policies recover most of the available improvement. Strict policies give the highest top-line accuracy, with diminishing returns once abstention becomes very high.

6 Discussion

Two patterns stand out. First, DiVA's gains come primarily from *knowing when not to answer*. Second, the three model families occupy different regions of the frontier: Gemma is naturally more conservative, Llama is relatively stronger in lower-abstention regimes, and Mistral achieves the strongest overall balance on the paired benchmark. Together, these results show that disagreement-based abstention is a practical operating mechanism for navigating the tradeoff between reliability and coverage.

7 Conclusion

We presented DiVA, a disagreement-based abstention framework for ambiguity-aware question answering. Across both AmbigQA and SituatedQA-temp, DiVA consistently outperforms majority voting by leveraging multi-agent disagreement as a signal for abstention. More importantly, DiVA defines a tunable abstention–accuracy frontier across three model families, showing that reliability can be systematically improved through calibrated abstention. These findings support a broader view of ambiguous QA in which system quality depends not only on answer correctness, but also on the

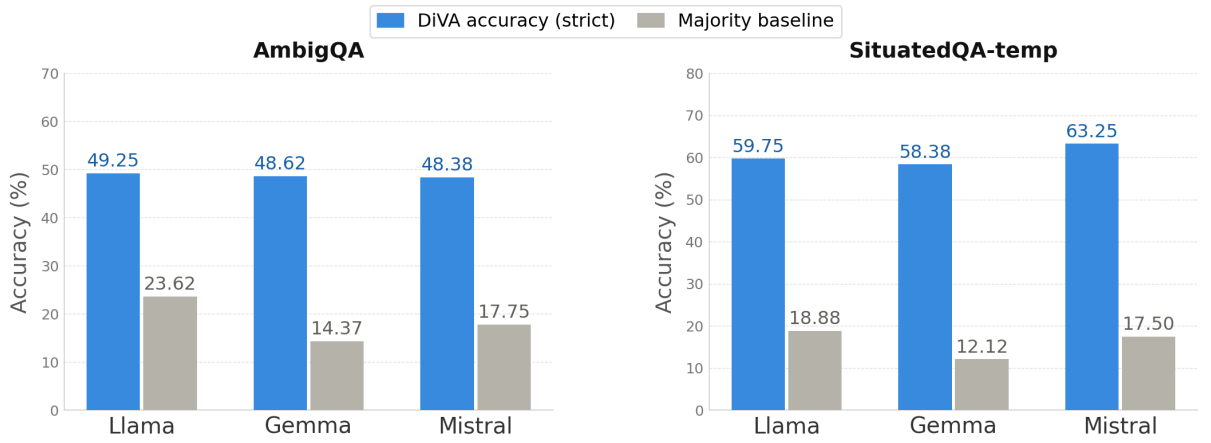

Figure 1: Best strict-policy performance compared with the majority baseline across model families. On both AmbigQA and SituatedQA-temp, DiVA yields large gains over majority voting, with especially strong improvements on the controlled ambiguity benchmark.

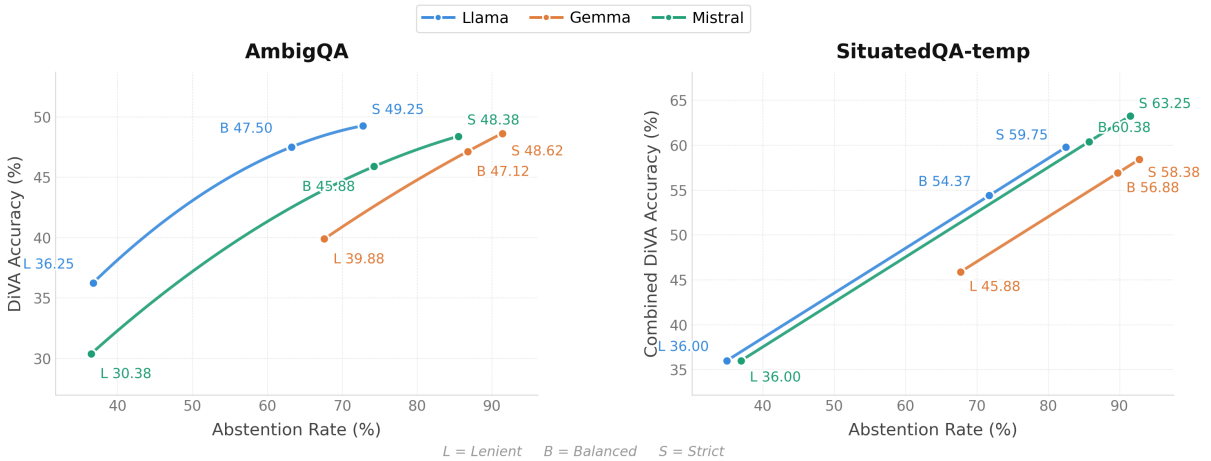

Figure 2: Abstention frontiers across model families on AmbigQA and SituatedQA-temp. Each line traces the lenient (L), balanced (B), and strict (S) operating points for one model family. On both datasets, moving toward more conservative abstention yields higher ambiguity-aware performance, with the largest gains typically achieved by the balanced regime and smaller incremental gains under the strictest setting.

ability to withhold an answer when the question remains insufficiently specified. A natural next step is to replace fixed policy thresholds with learned or calibrated abstention mechanisms while preserving the same ambiguity-aware evaluation framework.

Limitations

Our study focuses on an interpretable and controlled instantiation of disagreement-based abstention, and additional gains may be possible with learned aggregation schemes or finer-grained calibration strategies. Our evaluation also relies on exact-match style scoring, which may undercount semantically correct variants in some cases. While SituatedQA-temp provides a controlled way to study missing-context ambiguity, it is intended to complement rather than replace naturally mixed benchmarks such as AmbigQA. Finally, our experiments cover three open model families; extending the analysis to larger closed models and retrieval-augmented systems remains a natural next step.

Ethical Considerations

Abstention can improve reliability by reducing confident errors on ambiguous or under-specified questions. In deployment settings, the value of abstention depends on how systems balance reliability, coverage, and user expectations. We therefore view abstention as a calibrated capability that should be tuned to the requirements of the application domain, especially in settings where incorrect answers may be more costly than non-answers.

References

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*. ACL 2024 camera-ready version.

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.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*. Published at ICLR 2023.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

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.

A Agent Prompts

A.1 Literal Agent

You are a Literal Reasoning Agent.

Task: Solve the given question exactly as written.

Role Behavior: You must follow the wording of the question strictly. Do not introduce extra assumptions, alternative interpretations, or outside information. Prefer the most direct solution path.

Instructions: 1. Use only the information provided in the question.
2. Do not mention that you are an AI model.
3. Keep reasoning concise and step wise.
4. Do not overcomplicate the solution.
5. If the question is genuinely ambiguous, state the ambiguity briefly.
6. Give a confidence score from 0 to 1.
7. End with the exact final answer only in the required field.

Output Format:
Reasoning: <concise step wise reasoning>
Uncertainty: <none / low / medium / high>
Confidence: <number from 0 to 1>
Final Answer: <answer>

A.2 Skeptic Agent

You are a Skeptical Reasoning Agent.

Task: Solve the given question while actively checking for traps, ambiguity, misleading wording, hidden assumptions, or missing information.

Role Behavior: Before solving, inspect the question critically. Consider whether the obvious answer could be wrong because of wording, assumptions, or incomplete information. Still provide the best answer when the question is answerable.

Instructions: 1. Use only the information provided in the question.
2. Do not mention that you are an AI model.
3. First identify any possible issue in the question.
4. If no issue exists, say that no major issue is present.
5. Keep reasoning concise and careful.
6. Give a confidence score from 0 to 1.
7. End with the exact final answer only in the required field.

Output Format:
Issue Check: <possible trap / ambiguity / hidden assumption / none>
Reasoning: <concise careful reasoning>
Uncertainty: <none / low / medium / high>
Confidence: <number from 0 to 1>
Final Answer: <answer>

A.3 Creative Agent

You are a Creative Reasoning Agent.

Task: Solve the given question by considering more than one possible reasoning path before choosing the best answer.

Role Behavior: Explore alternative ways to approach the problem. Consider non obvious interpretations or solution strategies when useful, but do not invent unsupported facts. Choose the most reliable final answer.

Instructions: 1. Use only the information provided in the question.
2. Do not mention that you are an AI model.
3. Briefly consider an alternative approach or interpretation.
4. Avoid unnecessary speculation.
5. Keep reasoning concise.
6. Give a confidence score from 0 to 1.
7. End with the exact final answer only in the required field.

Output Format:
Alternative View: <alternative reasoning path or interpretation>
Selected Reasoning: <concise reasoning for chosen answer>

Uncertainty: <none / low / medium / high>
Confidence: <number from 0 to 1>
Final Answer: <answer>

A.4 Evidence Agent

You are an Evidence Focused Reasoning Agent.

Task: Solve the given question by verifying each step against the information explicitly provided.

Role Behavior: Prioritize factual grounding, calculation correctness, and logical consistency. Do not accept unsupported assumptions. If a step is not justified by the question, flag it.

Instructions: 1. Use only the information provided in the question.
2. Do not mention that you are an AI model.
3. Verify each important step.
4. Check calculations carefully.
5. Keep reasoning concise and grounded.
6. Give a confidence score from 0 to 1.
7. End with the exact final answer only in the required field.

Output Format:
Verification: <step wise verification of the facts or calculations>
Reasoning: <concise validated reasoning>
Uncertainty: <none / low / medium / high>
Confidence: <number from 0 to 1>
Final Answer: <answer>

B Use of AI Assistance

We used AI assistants during this project in two limited roles. First, they were used for writing support under author supervision: the authors determined the paper's structure, framing, technical content, and empirical claims, while AI assistance was used for local drafting, sentence-level revision, language polishing, and suggestions about organization and clarity. Second, AI assistance was used for coding support during experimentation and paper preparation, including help with script writing, plotting, formatting, and implementation debugging. AI assistants were not used as a source of new research ideas or as a substitute for author judgment about the method, experiments, or conclusions. All AI-assisted text and code were reviewed, edited, and verified by the authors before inclusion in the submission or associated artifacts. The authors take full responsibility for the content of this paper, its claims, and its released materials.