Answering Has Consequences: Value-Guided Selective Reasoning for Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used as reasoning systems for question answering, mathematical problem solving, and expert-facing decision support. In these settings, reliability depends not only on generating plausible answers, but also on deciding when answering is worse than abstaining. Existing selective reasoning methods often reduce this decision to confidence or uncertainty thresholding. This is insufficient for open-ended generation, where token likelihood is sensitive to surface form, verbalized confidence can be miscalibrated, and sampled reasoning paths may converge on the same wrong conclusion. Moreover, abstention is not a valueless rejection action, because a refused query may be routed to retrieval, a stronger model, human review, or delayed verification. We therefore formulate selective reasoning as an expected gain comparison between answering and abstaining. Based on this formulation, we introduce the Value-Guided Selective Reasoning Gate (VALUEGATE), a model-agnostic inference-time framework that combines sampled candidate responses, semantic agreement, calibrated self-evaluation, and task-dependent fallback value. VALUEGATE answers only when the estimated gain of answering is positive. Experiments on mathematical, commonsense, multi-task, and open-ended reasoning benchmarks evaluate VALUEGATE under selective accuracy, coverage-accuracy trade-offs, calibration, utility, and inference cost. The results show that value-guided selection provides a more deployment-aware decision rule than confidence-only abstention, while also revealing the importance of calibration, fallback specification, and sampling cost.

1 Introduction

Large language models (LLMs) have become a common interface for natural language reasoning, ranging from few-shot problem solving to instruction-following question answering (Brown et al., 2020). Their capabilities have been further strengthened by chain-of-thought prompting and large-scale alignment (Wei et al., 2022). However, the same models that generate fluent explanations can also answer confidently when their reasoning is wrong, when the question is underspecified, or when the required knowledge lies outside their effective boundary (Kadavath et al., 2022). In real deployments, such errors are not merely failed benchmark items. They can mislead users, trigger unsafe downstream actions, or waste scarce expert attention.

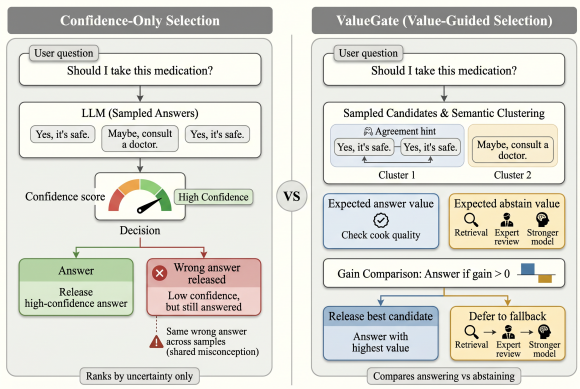

Figure 1: **Motivation of value-guided selective reasoning.** Confidence-only selection ranks questions by uncertainty and may release a high-confidence wrong answer. Value-guided selection compares the expected value of answering with the expected value of abstaining under a concrete fallback.

Selective prediction provides a natural mechanism for controlling such failures by allowing a model to reject inputs with high expected risk (Chow, 1970). In classification, this usually amounts to answering only when the confidence score exceeds a threshold. Recent work has adapted this idea to LLMs through token probability, verbalized confidence, sample entropy, or verifier confidence (Lin et al., 2022). Although simple, this formulation hides a strong assumption. It treats

the decision variable as uncertainty alone. For language reasoning, the deployment question is broader. A system must decide whether releasing an answer improves the final outcome relative to the available fallback.

This broader decision view raises three problems. **Problem I)** asks how to estimate answer quality when confidence is not aligned with semantic correctness. Token likelihood is local and sensitive to wording, whereas answer correctness is usually defined over semantic content (Kuhn et al., 2023). Verbalized confidence is easier to present to users, but it can be distorted by prompt wording, answer format, and alignment training (Zhang et al., 2024). **Problem II)** asks how to aggregate reasoning evidence from multiple generations. Sampling can reveal competing solution paths, hidden ambiguity, or unstable intermediate steps, but agreement can still concentrate around a shared misconception. **Problem III)** asks how to make abstention depend on the target workflow. A tentative answer may be useful in tutoring but unacceptable in medical triage or legal assistance if expert review is available.

To address these problems, we introduce the Value-Guided Selective Reasoning Gate (VALUEGATE), an inference-time framework for selective reasoning with LLMs. The central idea is to replace confidence ranking with expected gain comparison. Given an input, VALUEGATE samples candidate responses, estimates their task-level quality, aggregates semantic evidence, and compares the expected answer value with the abstention value. The system releases an answer only when the estimated gain is positive.

For **Problem I)**, VALUEGATE uses calibrated self-evaluation to estimate answer-level quality. Instead of treating raw token probability or verbalized confidence as the decision score, the evaluator judges whether a candidate response answers the question and whether its reasoning supports the final answer. Because raw self-evaluation can be overconfident, VALUEGATE calibrates the score on held-out validation data when available. This turns self-evaluation into a quality estimator rather than an unverified verbal judgment.

For **Problem II)**, VALUEGATE groups sampled responses by semantic equivalence and uses agreement as evidence rather than as a hard voting rule. This design avoids treating harmless wording differences as disagreement. It also avoids assuming that the majority answer is always correct. A large cluster can still be wrong when samples share the same misconception, while a minority candidate can be valuable when it receives strong calibrated quality evidence.

For **Problem III)**, VALUEGATE introduces an operational fallback value. Abstention is not modeled as the absence of an answer, but as routing the query to a fallback source such as retrieval, a stronger model, human review, or delayed verification. The fallback has its own expected quality and cost. This allows the same model output to be accepted in a low-cost setting and rejected in a high-stakes setting.

The resulting framework is model-agnostic. It requires only generated outputs and evaluation prompts, while optional validation data can improve calibration. This distinguishes VALUEGATE from standard calibration methods, which estimate correctness probability but do not decide whether answering is better than deferring. It also extends classical reject-option theory to open-ended generation, where the answer space is semantic and the decision depends on user workflow.

Our contributions are summarized as follows.

- We formulate selective reasoning for LLMs as expected gain comparison rather than pure uncertainty ranking, making fallback quality and error cost explicit.
- We introduce VALUEGATE, a model-agnostic inference-time framework that integrates sampled reasoning, semantic agreement, calibrated self-evaluation, and fallback value.
- We define operational fallback settings for fixed abstention reward, retrieval or stronger-model fallback, simulated human review, and oracle diagnostic analysis.
- We design a comprehensive evaluation protocol covering selective accuracy, coverage-risk tradeoffs, calibration, utility, inference cost, ablation, and error analysis.

2 Related Work

2.1 Selective Abstention

Selective prediction studies predictors that may abstain when the expected risk of answering is too high. Classical formulations connect rejection with error risk and decision cost (Chow, 1970). Later work developed learning-theoretic foundations and selective classifiers for modern neural networks

(El-Yaniv and Wiener, 2010). Deep selective classification usually evaluates the coverage-risk tradeoff, where a useful model should keep high accuracy on answered examples while avoiding unnecessary abstention (Geifman and El-Yaniv, 2017). Recent work adapts abstention to generative and vision-language models, often using uncertainty, answer probability, or verifier confidence as the decision score (Srinivasan et al., 2024). Our work follows the reject-option tradition but changes the estimated quantity. VALUEGATE estimates the expected gain of answering relative to a concrete fallback.

2.2 LLM Calibration

Calibration asks whether predicted confidence corresponds to empirical correctness (Guo et al., 2017). In NLP, this relation is complicated by length, tokenization, prompt format, and the many possible verbalizations of the same answer (Desai and Durrett, 2020). LLMs further complicate calibration because instruction tuning and human feedback can improve helpfulness while weakening the link between output probability and factual correctness (Ouyang et al., 2022). Empirical studies show that LLM confidence remains unstable across tasks and prompting conditions (Geng et al., 2024). Semantic uncertainty methods aggregate responses by meaning rather than by surface string and are useful for detecting unstable answers (Farquhar et al., 2024). These methods improve confidence estimation, but they do not by themselves define when abstention is preferable under a concrete fallback.

2.3 Reasoning Evidence

LLMs can often critique, compare, or refine their own responses (Manakul et al., 2023). Chain-of-thought prompting and self-consistency improve reasoning by generating multiple solution paths and aggregating answers (Kojima et al., 2022). Self-consistency shows that sampled generations contain useful evidence about answer quality, but majority agreement can still reflect a shared misconception (Wang et al., 2023). Self-evaluation is also not automatically calibrated, especially when the evaluator shares biases with the generator. VALUEGATE separates these roles. Sampling exposes candidate evidence, semantic agreement estimates answer stability, calibrated self-evaluation estimates task-level quality, and fallback value determines whether answering is better than abstaining.

3 Methodology

3.1 Problem Setting

We study selective reasoning with an input x and a language model p_θ. The system must either release an answer y or abstain and route the query to a fallback source, such as retrieval, a stronger model, human review, or delayed verification. Let $Q(y, x) \in [0, 1]$ denote task-level answer quality, defined by the task protocol as exact match, normalized final-answer correctness, semantic equivalence, model-judge score, or human evaluation.

We define answer and abstention values as

$$U_{\text{ans}}(x) = \mathbb{E}_{y \sim p_\theta(\cdot|x)}[Q(y, x)] - C_{\text{ans}}(x), \quad (1)$$

$$U_{\text{abs}}(x) = Q_{\text{fb}}(x) - C_{\text{fb}}(x), \quad (2)$$

where C_{ans} is the cost of releasing an answer, and Q_{fb} and C_{fb} are the fallback quality and cost. The gain of answering is

$$G(x) = U_{\text{ans}}(x) - U_{\text{abs}}(x). \quad (3)$$

The ideal policy answers when $G(x) > 0$. Since $G(x)$ is unobserved, VALUEGATE estimates it from sampled responses, semantic agreement, calibrated self-evaluation, and fallback value.

3.2 Semantic Evidence

Given x, VALUEGATE samples K candidate responses

$$\mathcal{Y}(x) = \{y_1, \ldots, y_K\}. \quad (4)$$

Each response is mapped to a semantic answer representation a_i. This can be a selected option, a normalized numeric answer, an extracted entity, or an embedding-based cluster label. Let $m(a)$ be the number of candidates assigned to semantic answer a. The agreement score of candidate y_i is

$$A_i = \frac{m(a_i)}{K}. \quad (5)$$

Agreement measures stability but not correctness. A majority cluster can still be wrong, while a minority answer can be correct. Therefore, VALUEGATE uses agreement as evidence rather than as a hard voting rule.

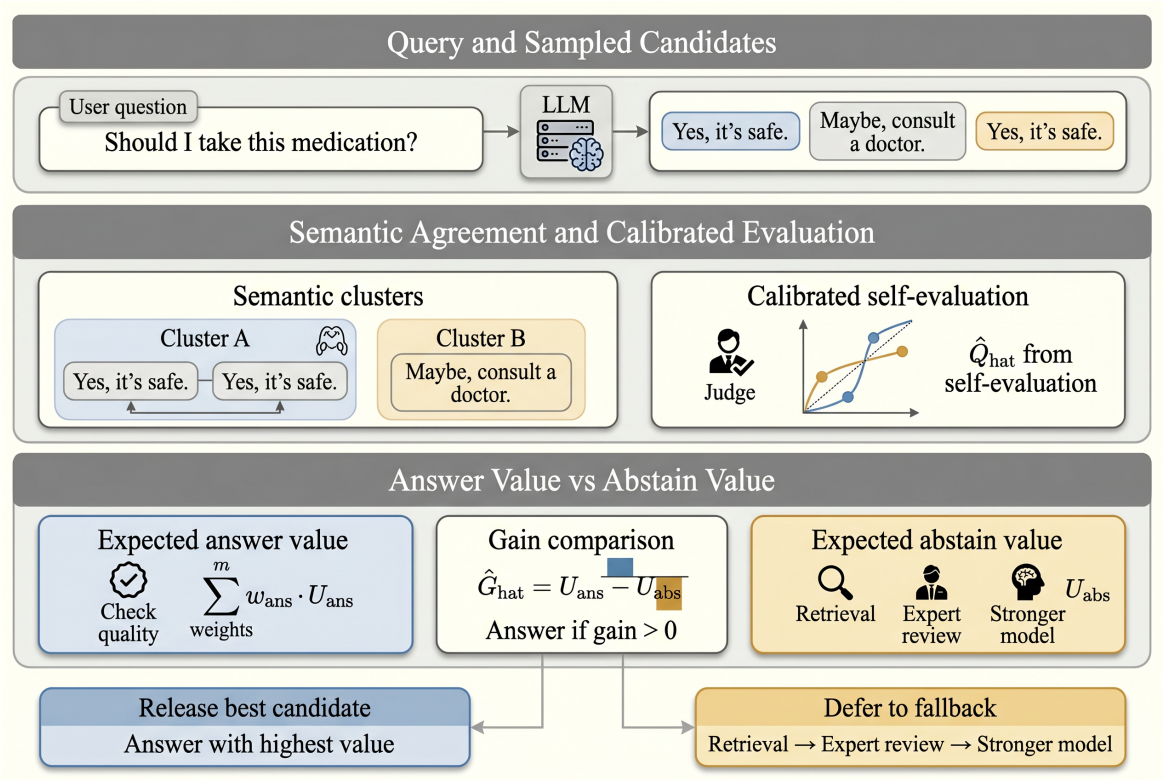

Figure 2: **Overview of VALUEGATE.** VALUEGATE samples candidate responses, groups them by semantic answer, estimates calibrated answer quality, compares answer value with fallback value, and answers only when the expected gain is positive.

3.3 Calibrated Evaluation

For each candidate y_i, an evaluator produces a raw quality score $s_i \in [0, 1]$ from the input, response, and task-specific judging instructions. The evaluator can be the same model under a judging prompt or a separate judge model. Because raw self-evaluation can be overconfident, VALUEGATE calibrates it on validation data when available.

We use a logistic calibration layer that combines self-evaluation and semantic agreement,

$$\widehat{Q}(y_i, x) = \sigma(\alpha \operatorname{logit}(\widetilde{s}_i) + \beta \operatorname{logit}(A_i) + b)\,, \tag{6}$$

where $\widetilde{s}_i$ is the normalized evaluator score, and α, β, and b are fitted on held-out data. The estimated answer value is

$$\widehat{U}_{\text{ans}}(x) = \sum_{i=1}^{K} w_i \widehat{Q}(y_i, x) - C_{\text{ans}}(x), \tag{7}$$

where w_i is a normalized candidate weight. By default, $w_i \propto A_i \widehat{Q}(y_i, x)$, which gives more weight to candidates that are both semantically supported and highly rated.

3.4 Fallback Value

Abstention is treated as a routing action rather than a valueless rejection. For a fallback source f, VALUEGATE estimates

$$\widehat{U}_{\text{abs}}(x; f) = \widehat{Q}_{\text{fb}}(x; f) - C_{\text{fb}}(x; f). \tag{8}$$

The fallback quality can be a fixed reward, validation-estimated accuracy, verifier score, stronger-model estimate, or simulated human-review quality. The cost term can represent latency, monetary cost, or expert effort after normalization to the same utility scale. This makes the answer-abstention decision depend on the target workflow.

3.5 Decision Protocol

The estimated gain is

$$\widehat{G}(x) = \widehat{U}_{\text{ans}}(x) - \widehat{U}_{\text{abs}}(x). \tag{9}$$

VALUEGATE answers when $\widehat{G}(x) > \tau$ and abstains otherwise. The default threshold is $\tau = 0$, while a positive τ gives a more conservative deployment policy. Algorithm 1 summarizes the inference procedure.

Algorithm 1: Value-Guided Selective Reasoning

Data: Input x, generator p_θ, evaluator J, fallback source f, sample size K, safety margin τ

Result: Released answer $y^\star$ or abstention

$\mathcal{Y} \leftarrow \mathrm{Sample}(p_\theta, x, K)$
$\mathcal{A} \leftarrow \mathrm{ExtractSemanticAnswers}(\mathcal{Y})$
$\mathcal{C} \leftarrow \mathrm{Cluster}(\mathcal{A})$
foreach *candidate* $y_i \in \mathcal{Y}$ **do**
  $s_i \leftarrow J(x, y_i)$
  $A_i \leftarrow |\mathrm{cluster}(y_i)|/K$
  $\widehat{Q}_i \leftarrow \mathrm{Calibrate}(s_i, A_i)$
$\widehat{U}_{\mathrm{ans}} \leftarrow \sum_i w_i \widehat{Q}_i - C_{\mathrm{ans}}$
$\widehat{U}_{\mathrm{abs}} \leftarrow \widehat{Q}_{\mathrm{fb}}(x; f) - C_{\mathrm{fb}}(x; f)$
$\widehat{G} \leftarrow \widehat{U}_{\mathrm{ans}} - \widehat{U}_{\mathrm{abs}}$
if $\widehat{G} > \tau$ **then**
  $y^\star \leftarrow \arg\max_{y_i \in \mathcal{Y}} \widehat{Q}_i$
  return $y^\star$
else
  return $\mathrm{Abstain}(x, f)$

4 Experiments

In this section, we evaluate VALUEGATE from four aspects. We first examine whether value-guided selection improves selective accuracy at matched coverage. We then study calibration and utility, focusing on whether gain estimates correspond to observed benefit over fallback. Next, we analyze component contributions through ablation. Finally, we study fallback sensitivity, inference cost, and qualitative failure modes.

4.1 Experimental Setup

Datasets. We evaluate on reasoning benchmarks with different answer formats. GSM8K measures mathematical word problems (Cobbe et al., 2021). ARC Challenge measures science reasoning (Clark et al., 2018). MMLU covers broad-domain knowledge (Hendrycks et al., 2021). HellaSwag evaluates commonsense completion (Zellers et al., 2019). StrategyQA measures implicit yes-or-no reasoning (Geva et al., 2021), and HotpotQA measures multi-hop question answering (Yang et al., 2018). Together, these datasets cover short-answer, multiple-choice, and open-ended reasoning.

Baselines. We compare against answer-all prediction, token-probability confidence, verbalized confidence, entropy over sampled answers, self-consistency majority confidence, semantic entropy, and validation-calibrated confidence. Probability-based baselines are used when answer probabilities are available. In black-box settings, baselines rely on verbalized confidence and sample agreement. This separation tests whether VALUEGATE remains useful without direct access to model internals.

Table 1: Method suite for value-guided selective reasoning.

Method	Sample	Eval.	Fallback	Calls
Answer all	–	–	–	1.0
Token conf.	–	–	–	1.0
Verbal conf.	–	✓	–	2.0
Self-cons.	✓	–	–	8.0
Semantic entropy	✓	–	–	8.0
Calibrated conf.	–	✓	–	2.0
VALUEGATE	✓	✓	✓	10.0

Metrics. The primary selective metric is accuracy at fixed coverage. We also report area under the coverage-accuracy curve, selective risk, expected calibration error, utility calibration error, cost-sensitive utility, and inference cost measured by model calls per query. Utility metrics are evaluated under fixed, validation-estimated, stronger-model, and simulated-human fallback settings. For fixed-coverage evaluation, selection thresholds are tuned on the validation split and then applied to the test split.

4.2 Main Results

Value-guided selection improves answer-abstention decisions. Table 2 reports selective accuracy at 70% coverage. Answer-all is evaluated by random subsampling at the same coverage, so its score is close to full-coverage accuracy. Confidence-based methods improve over answer-all because they filter some low-quality responses. Semantic entropy further improves selection by grouping meaning-level uncertainty rather than surface-form variation. VALUEGATE obtains the best average selective accuracy, showing that fallback-aware gain estimation ranks answerable examples more effectively than uncertainty alone.

Coverage curves reveal operating-point robustness. A single coverage value can hide whether a method is robust under different deployment tolerances. Figure 3 compares coverage-accuracy behavior across operating points. A tutoring sys-

Table 2: Selective accuracy at 70% coverage across reasoning benchmarks.

Method	GSM8K	ARC-C	MMLU	HellaSwag	StrategyQA	HotpotQA	Avg.
Answer all	74.2	71.8	73.6	82.4	69.5	62.8	72.4
Token confidence	78.6	75.0	77.1	85.4	72.6	67.1	76.0
Verbalized confidence	80.1	76.2	78.6	86.0	74.4	68.4	77.3
Calibrated confidence	81.5	77.5	79.4	86.4	75.5	69.2	78.3
Self-consistency	82.8	78.1	79.5	86.7	76.3	70.5	79.0
Semantic entropy	84.3	79.6	80.7	87.5	77.8	72.0	80.3
VALUEGATE	**86.5**	**81.3**	**82.6**	**88.8**	**80.1**	**74.1**	**82.2**

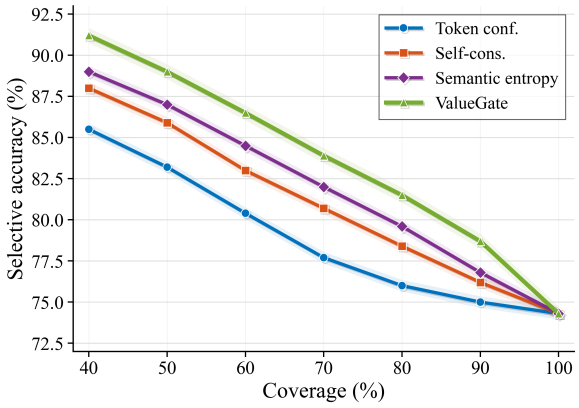

Figure 3: Coverage-accuracy curves for confidence-based selection, uncertainty-based selection, and VALUEGATE.

tem may operate at high coverage, whereas expert support may require lower coverage. VALUEGATE maintains stronger accuracy across the middle-coverage region, where abstention is selective enough to remove risky answers but not so conservative that the system refuses most queries.

4.3 Calibration and Utility Analysis

Gain calibration measures the decision quantity directly. Correctness calibration bins examples by predicted quality and compares each bin with empirical correctness. Utility calibration bins examples by predicted gain and compares each bin with observed improvement over fallback. As shown in Table 3, VALUEGATE reduces both expected calibration error and utility gap. This indicates that the gain score is not only better at ranking examples, but also better aligned with the quantity used by the answer-abstention policy.

Self-evaluation requires calibration. Figure 4 analyzes the relationship between self-evaluation scores and empirical correctness. Same-model and separate-judge variants are reported separately, because evaluator bias can affect the decision boundary. Calibration reduces overconfident self-judgment and makes the estimated gain more interpretable. This supports treating self-evaluation as a noisy quality signal rather than as a direct probability of correctness.

Table 3: Calibration and inference-cost comparison. Lower ECE and utility gap are better.

Method	ECE	Utility gap	Calls
Verbalized confidence	0.146	0.083	2.0
Self-consistency	0.118	0.069	8.0
Semantic entropy	0.103	0.057	8.0
VALUEGATE	**0.061**	**0.028**	10.0

4.4 Ablation and Cost

Table 4 studies both component necessity and practical cost. Removing semantic agreement weakens performance because surface-form diversity is no longer separated from meaning-level disagreement. Removing calibration causes a larger utility drop than accuracy drop, indicating that raw self-evaluation can still rank some examples but gives poorly scaled decision values. Removing fallback value damages utility most directly, confirming that abstention quality must be modeled rather than treated as a fixed rejection action. The cluster-representative variant recovers most of the utility of full VALUEGATE while reducing calls.

4.5 Fallback Sensitivity and Error Analysis

Fallback value changes the decision boundary. Figure 5 reports sensitivity to fallback value, sample size K, and safety margin τ. Increasing fallback value makes the system more conservative because abstention becomes more attractive. Increasing K improves evidence quality until marginal gains saturate. Increasing τ reduces harmful over-answering while increasing abstention. Table 5 summarizes representative fallback settings. The stronger-model and simulated-human settings pro-

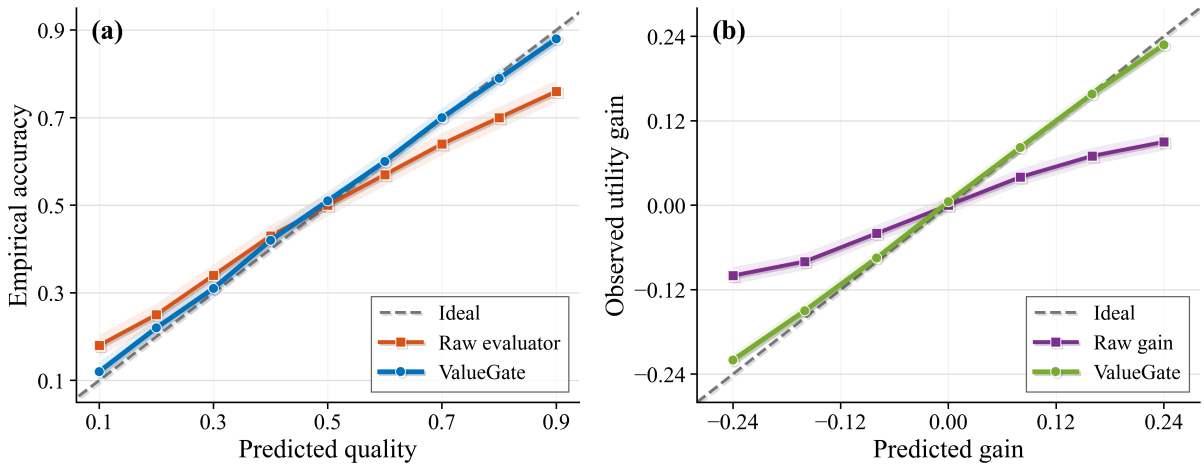

Figure 4: Reliability analysis for calibrated self-evaluation and gain estimation. The left panel compares predicted answer quality with empirical accuracy, while the right panel compares predicted gain with observed utility gain.

Table 4: Ablation study and evaluation cost. Results are averaged across benchmarks at 70% coverage.

Setting	Sem.	Calib.	Fallback	Multi-sample	Sel. Acc.	Utility	Calls
Full VALUEGATE	✓	✓	✓	✓	**82.2**	**0.142**	10.0
w/o semantic agreement	–	✓	✓	✓	80.7	0.119	10.0
w/o calibration	✓	–	✓	✓	80.1	0.103	10.0
w/o fallback value	✓	✓	–	✓	80.9	0.081	10.0
Single sample	–	✓	✓	–	79.0	0.093	3.0
Cluster representative	✓	✓	✓	partial	81.3	0.132	6.1

Table 5: Sensitivity to fallback settings.

Fallback	Coverage	Utility	Error rate
Fixed reward	72.0	0.118	12.4
Retrieval	68.5	0.132	10.7
Stronger model	61.8	0.147	8.1
Human review	54.2	0.154	5.9
Oracle	49.5	0.188	2.4

Table 6: Error categories for value-guided selection.

Category	Count	Share
False answer	31	31.0%
Unnecessary abstention	28	28.0%
Wrong fallback preference	25	25.0%
Other	16	16.0%

duce higher utility but lower coverage, reflecting a more conservative policy.

Error categories connect failures to the utility formulation. We categorize failures into false answers, unnecessary abstentions, and wrong fallback preferences. A false answer occurs when the system answers incorrectly. An unnecessary abstention occurs when the system abstains although the selected answer would have been correct. A wrong fallback preference occurs when the sign of estimated gain disagrees with the observed utility comparison between answering and abstaining. Table 6 shows that false answers remain the largest category, but wrong fallback preferences are also substantial, suggesting that fallback estimation is a central bottleneck.

5 Discussion

The central implication of VALUEGATE is that selective reasoning should be evaluated as a decision problem. A confidence threshold asks whether the model appears to believe its answer, while a gain score asks whether releasing the answer improves the expected outcome compared with abstention. This distinction is small in balanced classification benchmarks but large in real deployments, where fallback options and error costs vary across tasks and users. The framework also explains why common signals can fail in isolation. Token probability may be weakly related to open-ended correctness, verbalized confidence may be distorted by helpful-sounding explanations, self-consistency

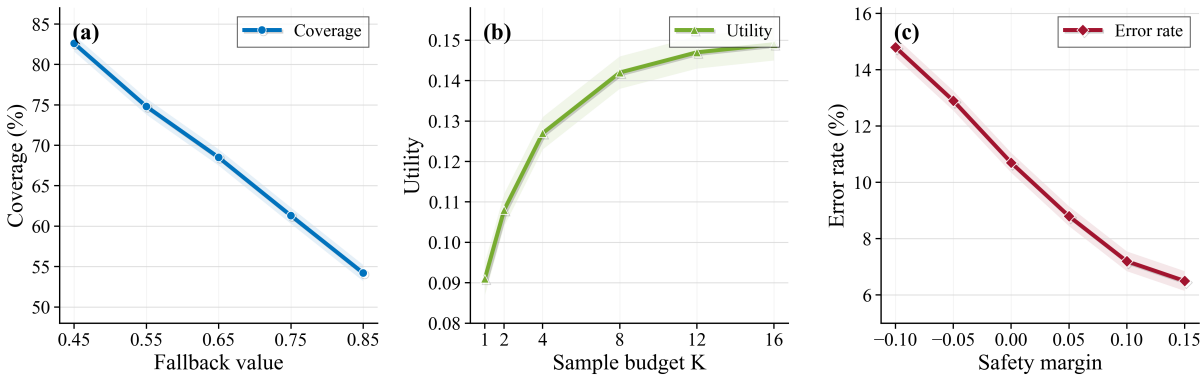

Figure 5: Sensitivity of selective utility to fallback value, sample budget, and safety margin.

can over-trust popular wrong answers, and semantic entropy can detect disagreement without determining whether abstention is preferable. VALUEGATE does not discard these signals. It places them inside a utility-aware decision model and evaluates the resulting decision by observed gain. The practical boundary of the approach is cost and calibration. Sampling and self-evaluation increase inference cost, same-model evaluation can share the generator's misconceptions, and fallback value must be specified carefully because a poor fallback estimate can move the decision boundary in the wrong direction. These limitations motivate explicit reporting of model calls, self-evaluation reliability, and fallback sensitivity whenever selective reasoning is used in consequential settings.

6 Conclusion

This paper introduces VALUEGATE, a value-guided framework for selective reasoning with large language models. Instead of deciding by confidence alone, VALUEGATE estimates whether answering yields positive expected gain over abstaining. It combines sampled reasoning, semantic agreement, calibrated self-evaluation, and operational fallback value into a single decision rule. This formulation connects selective prediction, language-model calibration, and deployment-level utility, offering a practical path toward safer and more adaptive use of LLMs.

Ethics Statement

This work aims to reduce harmful over-answering by allowing language models to abstain when answering is not expected to improve the outcome. It should not be used as a substitute for domain expertise in high-stakes settings. Abstention can also create access issues if a system refuses too often for certain topics or user groups. Evaluations for consequential deployments should therefore include subgroup analysis, fallback quality analysis, and human oversight.

Limitations

VALUEGATE requires multiple samples and self-evaluation calls, making it more expensive than single-pass generation. The cluster-representative variant reduces cost but may miss minority candidates that contain the correct answer. The framework also depends on the quality of self-evaluation and the specification of fallback value. If the evaluator shares the generator's misconception, or if fallback value is poorly estimated, the decision boundary may be misaligned with the target workflow. Future work should study adaptive sampling, stronger black-box calibration, and human-in-the-loop fallback estimation.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Chi-Keung Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias

Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 295–302. Association for Computational Linguistics.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6577–6595.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 346–361.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models mostly know what they know. *arXiv preprint arXiv:2207.05221*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. 2024. Selective selective prediction: Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*. 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. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800. Association for Computational Linguistics.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the confidence of large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.