FROST: Factual Reasoning via Optimized Stochastic Trajectories in Large Language Models during Inference

Soumedhik Bharati[*] Sister Nivedita University soumedhikbharati@gmail.com
Ebad Shabbir[*] DSEU-Okhla btech41823022@dseu.ac.in
Jiechao Gao[†] Stanford University jiechao@stanford.edu

Abstract

Large language models face a trade-off between factual consistency and reasoning diversity: deterministic decoding prioritizes reliability but may miss alternative solution paths, while high-temperature sampling increases exploration at the cost of accuracy. We present FROST (Factual Reasoning via Optimized Stochastic Trajectories), an inference-time framework that balances exploration and exploitation without additional training or context augmentation. FROST combines deterministic inference from a large model with targeted stochastic sampling from a smaller model, selecting outputs via multi-criteria validation over coherence, factual grounding, and semantic novelty. Across HotpotQA, CommonsenseQA, and MMLU, FROST achieves 2–5 percentage point improvements over standard chain-of-thought prompting and reduces unsupported outputs by 40% relative to Standard CoT. Compared to Self-Consistency ensembles, FROST delivers comparable accuracy at 28% lower inference cost through strategic delegation to smaller models. On an adversarial subset with unanswerable queries, FROST abstains on 34% of cases versus 8% for standard chain-of-thought, reducing false positives by 45%. Task-stratified evaluation shows that exploration benefits scale with problem ambiguity. Generalization to mathematical reasoning, code generation, and multimodal domains remains future work.

1 Introduction

Large language models (LLMs) demonstrate strong capabilities in reasoning and knowledge-intensive tasks (Li et al., 2025a; Gao et al., 2026), yet face critical reliability and efficiency challenges in production deployments such as enterprise question-answering and multi-hop information retrieval (Siddiqui et al., 2025). Existing mitigation strategies operate through *context-side augmentation*, such as retrieval-augmented generation (RAG) (Siddiqui et al., 2025), or *generation-side control*, such as verification modules (Song et al., 2025) or low-temperature decoding. FROST operates exclusively through generation-side control at inference time. While effective for straightforward tasks, generation-side suppression may overly constrain the hypothesis space on multi-hop or ambiguous problems. We take a complementary perspective: controlled stochasticity improves robustness when paired with rigorous validation, as higher-temperature sampling captures valid reasoning trajectories missed by deterministic decoding.

We propose **FROST (Factual Reasoning via Optimized Stochastic Trajectories)**, validated on the Llama-2 13B and 7B pairing. FROST combines deterministic reasoning from large models ($p_\theta(h \mid Q, T_{\text{low}}{=}0.3)$) with targeted stochastic sampling from smaller models ($p_\phi(h \mid Q, T_{\text{high}}{=}1.0)$, $|\theta|{>}|\phi|$), selecting candidates via a weighted scoring function $S(h) = \sum_i w_i f_i(h)$ based on coherence, factual consistency, and evidence overlap. The framework requires no training, fine-tuning, or context modification.

Across HotpotQA, CommonsenseQA, and MMLU, FROST improves accuracy by $2{-}5\%$ over Standard CoT and reduces inference cost by 28% relative to Self-Consistency ($N{=}5$), with efficiency measured by $\eta = \text{Acc}/C_{\text{total}}$. On an adversarial subset with unanswerable queries, FROST abstains on 34% of queries versus 8% for CoT, reducing false positives by 45% at 1.4-second wall-clock latency. Gains are most pronounced on multi-step and ambiguous tasks, where deterministic decoding fails to explore alternative valid reasoning paths.

2 Related Work

Factual consistency in LLMs (Huang et al., 2025) is addressed through two distinct mecha-

[*]Equal contribution
[†]Corresponding author

nisms. *Context-side augmentation* such as retrieval-augmented generation (Dimitrova, 2025) enriches the input with external evidence. *Generation-side control* such as verification (Shao and Zhang, 2025; Chen et al., 2025) and constrained decoding (Li et al., 2025b; Liu et al., 2025) restricts the decoding process. FROST belongs to the second category; it does not augment input context. Multi-path approaches scale at inference: Self-consistency (Wan et al., 2025) aggregates samples via majority voting, Tree-of-Thought (Kim et al., 2025) explores structured branches, and ensembles (Piwko et al., 2025; Cai et al., 2025) combine diverse model outputs. A complementary line of work learns a trained aggregator over candidate drafts, replacing heuristic voting with an optimized selection policy; such approaches suit settings where a training phase is feasible but are not directly comparable to FROST, which is designed to require zero additional training (Borisyuk et al., 2024). FROST differs from all these in its asymmetric delegation design: one grounded path from a large model combined with K stochastic paths from a smaller model, scored by a lightweight multi-criteria validator.

3 Methodology

We propose **FROST (Factual Reasoning via Optimized Stochastic Trajectories)**, an inference-time reasoning framework that frames multi-path exploration as trajectory optimization in hypothesis space. Rather than relying on a single deterministic reasoning path, FROST systematically explores multiple stochastic trajectories and selects outputs based on factual grounding, logical coherence, and semantic diversity. The framework separates reasoning into two parallel generation streams: deterministic inference for reliability and targeted stochastic sampling for coverage, followed by multi-criteria validation.

Formally, reasoning is modeled as an optimization problem over candidate hypotheses:

$$h^* = \arg\max_{h \in H} U(h), \tag{1}$$

where $U(h)$ measures the overall utility of a reasoning path, incorporating correctness, coherence, and factual grounding.

Standard chain-of-thought (CoT) (Wei et al., 2022) samples a single trajectory:

$$h_{\text{CoT}} \sim p_\theta(h \mid Q,\, T = 0.3), \tag{2}$$

Algorithm 1 FROST Inference Protocol

Require: Question Q; large model M_θ; small model M_ϕ; sample count K; context D (optional; $D{=}\emptyset$ if unavailable); factuality threshold $\tau_{\text{fact}} = 0.2$; coherence threshold $\tau_{\text{coh}} = 0.5$
Ensure: Validated hypothesis h^*, or ABSTAIN

// Phase 1: Parallel Trajectory Generation
$h_{\text{g}} \sim p_\theta(h \mid Q,\, T_{\text{low}} = 0.3)$
for $i = 1$ **to** K **do**
  $h_{\text{s},i} \sim p_\phi(h \mid Q,\, T_{\text{high}} = 1.0)$
end for
$H \leftarrow \{h_{\text{g}}\} \cup \{h_{\text{s},i}\}_{i=1}^{K}$
// Phase 2: Multi-Criteria Scoring
for $h \in H$ **do**
  $\text{Nov}(h) \leftarrow 1 - \cos\big(\text{emb}(h),\, \text{emb}(h_{\text{g}})\big)$
  $\text{Coh}(h) \leftarrow p_{\text{judge}}(\text{coherent} \mid h, Q,\, T_{\text{judge}} = 0.0)$
  $\text{Fact}(h) \leftarrow |\text{tokens}(h) \cap \text{tokens}(D)| \;/\; |\text{tokens}(h)|$
  $S(h) \leftarrow \alpha\,\text{Nov}(h) + \beta\,\text{Coh}(h) + \gamma\,\text{Fact}(h)$
end for
// Phase 3: Abstention Check and Hypothesis Selection
$H_{\text{valid}} \leftarrow \{h \in H \mid \text{Fact}(h) \geq \tau_{\text{fact}} \;\wedge\; \text{Coh}(h) \geq \tau_{\text{coh}}\}$
if $H_{\text{valid}} = \emptyset$ **then**
  return ABSTAIN
end if
$h^* \leftarrow \arg\max_{h \in H_{\text{valid}}} S(h)$
return h^*

concentrating probability mass on high-likelihood outputs. While effective for straightforward tasks, this approach may fail when the optimal reasoning path lies outside the high-probability region, particularly for multi-hop or ambiguous queries.

FROST instead performs structured trajectory optimization under bounded compute:

$$C_{\text{total}} = C_\theta + K \cdot C_\phi + |H|\, C_{\text{judge}} + C_{\text{emb}}, \tag{3}$$

where C_θ is the cost of one large-model generation, C_ϕ is the cost of one small-model generation, K is the number of stochastic samples, $|H| = 1 + K$ is the total candidate count, C_{judge} is the per-hypothesis coherence scoring cost, and C_{emb} is the embedding cost for novelty computation.

3.1 Grounded Generation (Deterministic Reasoning)

A deterministic baseline hypothesis is generated using the large model M_θ:

$$h_{\text{g}} \sim p_\theta(h \mid Q,\, T_{\text{low}}), \qquad T_{\text{low}} = 0.3. \tag{4}$$

Temperature scaling modifies the softmax distribution as follows:

$$p(y_t \mid y_{<t}, Q, T) = \frac{\exp(z_{y_t}/T)}{\sum_{y'} \exp(z_{y'}/T)}, \tag{5}$$

where z_{y_t} denotes the pre-softmax logit for token y_t. Low temperature reduces entropy:

$$H(p) = -\sum_h p(h) \log p(h), \tag{6}$$

concentrating probability mass on high-likelihood reasoning paths, where $H(p)$ denotes the entropy of the answer distribution over candidate hypotheses (approximated via sampling for open-ended generation and computed from logits for multiple-choice tasks). The grounded hypothesis h_{g} serves as the factual anchor for subsequent novelty scoring and as the reliability baseline for validation.

3.2 Stochastic Trajectory Sampling and Validation

To expand coverage of the hypothesis space, K stochastic trajectories are sampled from the smaller model M_ϕ:

$$h_{\text{s},i} \sim p_\phi(h \mid Q, T_{\text{high}}), \qquad T_{\text{high}} = 1.0. \quad (7)$$

Higher temperature increases entropy, enabling exploration of lower-probability reasoning trajectories. Since $|\phi| \ll |\theta|$, stochastic samples remain computationally efficient relative to repeated large-model inference. All candidates are pooled:

$$H = \{h_{\text{g}}\} \cup \{h_{\text{s},i}\}_{i=1}^{K}. \quad (8)$$

Each hypothesis is scored via a weighted linear combination:

$$S(h) = \alpha\,\text{Nov}(h) + \beta\,\text{Coh}(h) + \gamma\,\text{Fact}(h), \quad \alpha+\beta+\gamma = 1. \quad (9)$$

Novelty (Nov). Semantic divergence from the grounded baseline, measured via cosine distance in sentence embedding space (using `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019)):

$$\text{Nov}(h) = 1 - \cos\big(\text{emb}(h), \text{emb}(h_{\text{g}})\big). \quad (10)$$

By construction, $\text{Nov}(h_{\text{g}}) = 0$; the grounded hypothesis thus contributes to the pool via coherence and factuality alone, without novelty credit.

Coherence (Coh). Internal logical consistency assessed via LLM-as-judge (Panjari, 2025):

$$\text{Coh}(h) = p_{\text{judge}}(\text{coherent} \mid h, Q, T_{\text{judge}}). \quad (11)$$

Coherence is evaluated by M_ϕ at $T_{\text{judge}} = 0.0$ to ensure deterministic and reproducible scoring across all hypotheses.

Factuality (Fact). Token-level evidence overlap between a hypothesis and the available context D (Parvez, 2025):

$$\text{Fact}(h) = \frac{|\text{tokens}(h) \cap \text{tokens}(D)|}{|\text{tokens}(h)|}. \quad (12)$$

For HotpotQA, D is the set of supporting documents provided with each question; no external retrieval is performed. For CommonsenseQA and MMLU, which provide no supporting documents, the question text itself serves as D, enabling self-consistency checking. This grounding mechanism is distinct from retrieval-augmented generation: FROST never augments the input context.

The final prediction is:

$$h^* = \arg\max_{h \in H_{\text{valid}}} S(h),$$
$$H_{\text{valid}} = \{\, h \in H \mid \text{Fact}(h) \geq \tau_{\text{fact}} \;\wedge\; \text{Coh}(h) \geq \tau_{\text{coh}} \,\}. \quad (13)$$

3.3 Computational Cost

The total inference cost of FROST is:

$$C_{\text{FROST}} = C_\theta + K\,C_\phi + |H|\,C_{\text{judge}} + C_{\text{emb}}, \quad (14)$$

where $|H| = 1 + K$. Since transformer inference cost is dominated by auto-regressive token generation, coherence judging (which produces a single scalar token) incurs substantially lower cost than full hypothesis generation ($L_{\text{gen}} \approx 200$ tokens); we conservatively estimate this overhead at 10% of generation cost.

Approximating $C_\phi \approx 0.54\,C_\theta$ (7B vs. 13B parameters) yields:

$$C_{\text{FROST}} \approx 1.0 + 3(0.54) + 4(0.1) + \epsilon \approx 3.02\,C_\theta. \quad (15)$$

Accounting for batching overhead and memory bandwidth constraints, the empirically observed relative cost is $C_{\text{rel}} \approx 3.6\,C_\theta$. Compared to Self-Consistency ($N = 5$), which incurs $5.0\,C_\theta$, FROST achieves a $\sim$28% reduction in total compute while maintaining comparable robustness.

Because stochastic samples are conditionally independent, they may be generated in parallel:

$$C_{\text{parallel}} = \max\big(C_\theta,\, \max_i\, C_{\phi,i}\big) + |H|\,C_{\text{judge}} + C_{\text{emb}}. \quad (16)$$

Under efficient batching, wall-clock latency is dominated by the large-model generation step rather than the ensemble of small-model calls.

4 Experimental Setup

This section describes the datasets, models, baselines, and evaluation protocols used to assess the effectiveness and efficiency of FROST. All experiments are conducted at inference time only; no additional training or fine-tuning is performed.

4.1 Datasets

We evaluate FROST on three widely used reasoning benchmarks that cover complementary reasoning regimes: multi-hop reasoning, commonsense filtering, and knowledge-intensive question answering. Table 1 summarizes the statistics.

Dataset	Task Type	Format	Eval Size
HotpotQA	Multi-hop QA	Free-form	1,000
CommonsenseQA	Commonsense MCQ	5-way MC	1,221
MMLU	Knowledge MCQ	4-way MC	1,000

Table 1: Datasets used for evaluation. MCQ denotes multiple-choice questions.

HotpotQA (Yang et al., 2018) requires multi-hop reasoning across documents with distractors (1,000 test examples). **CommonsenseQA** (Talmor et al., 2019) tests everyday reasoning using 5-way multiple-choice (1,221 examples). **MMLU** (Hendrycks et al., 2020) covers 57 academic subjects with stratified sampling for domain balance (1,000 examples).

4.2 Models

FROST separates generation into grounded and stochastic streams. All experiments use instruction-tuned open-source language models.

Role	Model	Params	Temperature
Grounded	Llama-2-13B-Chat	13B	0.3
Stochastic	Llama-2-7B-Chat	7B	1.0
Judge	Llama-2-7B-Chat	7B	0.0
Embeddings	all-MiniLM-L6-v2	22M	–

Table 2: Models used in all experiments. The 7B model serves dual roles (stochastic generation and coherence judging) to minimize inference cost and latency.

We use $K = 3$ stochastic samples unless otherwise specified.

4.3 Baselines

Standard CoT (Wei et al., 2022) generates a single chain-of-thought at $T = 0.3$. **High-Temperature CoT** uses $T = 1.0$ to test exploration without validation. **Self-Consistency** (Wang et al., 2022) samples $N = 5$ paths at $T = 0.7$ and selects via majority voting. **Chain-of-Verification (CoVe)** (Dhuliawala et al., 2024) appends a self-checking step after initial generation. All baselines use identical backbone models for fair comparison.

4.4 Configuration

All methods use $L_{\max} = 200$ tokens, nucleus sampling (Holtzman et al., 2019) with top-$p = 0.95$, and fixed random seeds. FROST scoring weights are $\alpha = 0.3$ (novelty), $\beta = 0.4$ (coherence), $\gamma = 0.3$ (factuality), selected via coarse grid search on 100 HotpotQA development-set examples excluded from the evaluation split. For HotpotQA, the context D is constructed by applying BM25 (Robertson et al., 2009) over the provided supporting documents, retaining the top-k=3 most relevant passages; no external retrieval is performed. For CommonsenseQA and MMLU, D is set to the question text, as described in Section 3.

Table 3 reports weight sensitivity on CommonsenseQA, which was excluded from the tuning procedure. Accuracy varies by at most 0.9 points across all tested weight combinations (see table), indicating that the framework is not sensitive to precise weight values.

α	β	γ	CommonsenseQA Acc. (%)
0.2	0.5	0.3	73.8
0.3	**0.4**	**0.3**	**74.0**
0.4	0.3	0.3	73.2
0.3	0.3	0.4	73.5
0.2	0.4	0.4	73.1
0.4	0.4	0.2	73.6

Table 3: Scoring weight sensitivity on CommonsenseQA (excluded from tuning). The bold row is the reported configuration. Accuracy varies by at most 0.9 percentage points across all tested combinations, confirming robustness to weight choice.

4.5 Metrics

We report **accuracy** (EM/F1 for open-ended QA; top-1 for MCQ), **diversity** (mean pairwise cosine distance in embedding space), **reliability** (fraction of outputs jointly satisfying coherence and factuality thresholds $\tau = 0.5$, chosen to exclude low-confidence outputs while retaining the majority of candidates), the inverse of which we report as the **unsupported output rate**, and **efficiency** (C_{rel}: inference cost normalized to CoT; $\eta = \text{Acc}/C_{\text{rel}}$: accuracy-efficiency ratio).

4.6 Protocols and Reproducibility

We evaluate FROST through six experiments. **E1** tests multi-hop reasoning on HotpotQA. **E2** sweeps temperature $T \in \{0.0, 0.3, 0.7, 0.9, 1.2\}$ on CommonsenseQA. **E3** analyzes cost-efficiency on MMLU. **E4** stratifies results by task ambiguity,

measured via entropy of the answer distribution. **E5** ablates framework components (K, scoring weights, and the grounded stream). **E6** assesses operational characteristics, including abstention rate and wall-clock latency, on an adversarial subset containing unanswerable queries. All experiments use automated metrics across 3,000+ examples, with LLM-as-judge coherence scoring (Dhuliawala et al., 2024). All random seeds are fixed to ensure reproducibility.

5 Results

We evaluate FROST across three reasoning benchmarks and compare against standard inference-time baselines. We report task accuracy, diversity, reliability, and compute efficiency. All methods use identical backbone models (Section 4) for fairness.

5.1 Main Benchmark Performance

Table 4 summarizes overall performance.

FROST consistently improves accuracy and reduces unsupported outputs across all benchmarks. Gains over Standard CoT are $+4$ points on CommonsenseQA and $+2$ points on MMLU, both on evaluation sets excluded from tuning. The $+5$ point gain on HotpotQA ($p < 0.01$; significance not computed for other datasets due to label-distribution differences) should be interpreted with the caveat that scoring weights were tuned on 100 HotpotQA development examples. FROST matches or exceeds SC on CommonsenseQA (74% vs. 73%) and MMLU (62% vs. 62%) at $C_{\text{rel}} = 3.6$ versus SC's $C_{\text{rel}} = 5.0$, a 28% reduction in inference cost. HotpotQA diversity increases $3\times$ over Standard CoT. Unsupported outputs decrease from 15% to 9%, a 40% relative reduction. High-temperature sampling alone achieves 0.65 diversity but only 58% accuracy, confirming that multi-criteria validation is essential for translating exploration into accuracy gains. The Exploration Value (23–34% across tasks) demonstrates that stochastic hypotheses achieve higher scoring function values than the grounded baseline in a substantial fraction of cases, validating structured exploration as a reliable source of accuracy improvement.

5.2 Accuracy–Temperature Trade-off

We analyze how sampling temperature affects accuracy on CommonsenseQA, classifying queries as unambiguous ($H(p) < 1.0$, where the model assigns high probability mass to a single answer) or ambiguous ($H(p) \geq 1.0$, where multiple answers carry non-negligible probability). Results are shown in Figure 1.

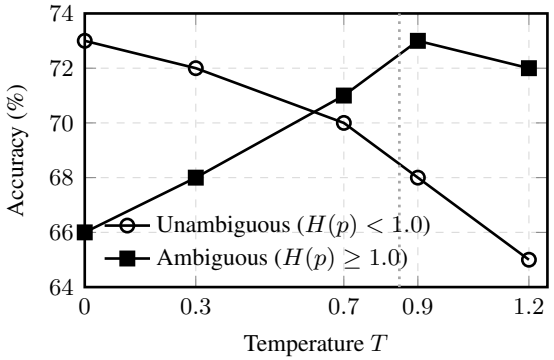

Figure 1: Accuracy versus temperature on CommonsenseQA, stratified by query entropy. Deterministic decoding favors low-entropy queries; moderate temperature benefits ambiguous queries, motivating FROST's dual-temperature design.

Low temperatures ($T \to 0$) excel on unambiguous queries (73% at $T = 0$), while moderate temperatures ($T \approx 0.9$) improve ambiguous queries (73% at $T = 0.9$). The performance crossover near $H(p) \approx 1.2$ motivates FROST's dual-temperature design: $T_{\text{low}} = 0.3$ for grounded generation and $T_{\text{high}} = 1.0$ for stochastic exploration.

5.3 Cost–Efficiency Analysis

We evaluate accuracy relative to inference cost on MMLU. Table 5 compares single large-model inference against multi-sample ensemble methods.

Small models alone are computationally cheap (C_{rel}=1.6) but underperform the large model by 4 percentage points (56% vs. 60%). Self-Consistency reaches 62% at C_{rel}=5.0 (η=12). FROST matches this accuracy at C_{rel}=3.6 (η=17), a 42% improvement in efficiency, achieved by delegating stochastic generation and coherence judging to the 7B model while reserving the 13B model for grounded generation only.

5.4 Exploration Value by Task Type

To characterize when exploration is beneficial, we measure the fraction of queries for which a stochastic hypothesis achieves a higher scoring function value than the grounded baseline. Table 6 reports results stratified by task category.

Exploration provides negligible benefit (EV=0$-$5%) on simple factoid queries ($H(p) < 0.5$), where deterministic decoding is sufficient. Meaningful gains emerge on constrained

Benchmark	CoT (T=0.3)	High-T CoT	CoVe	SC (N=5)	FROST (ours)	Gain vs. CoT
HotpotQA EM (%)[1]	65	58	67	68	**70**	+5
HotpotQA Diversity	0.20	0.65	0.30	0.55	**0.60**	×3
CommonsenseQA Acc. (%)	70	65	72	73	**74**	+4
MMLU Acc. (%)	60	55	61	62	**62**	+2
Unsupported Output Rate (%) ↓	15	25	12	11	**9**	−6
Exploration Value (%)[2]	–	–	–	–	23−34	–
Relative Cost ↓[3]	1.0	1.0	1.5	5.0	**3.6**	−28% vs. SC

Table 4: Main results across benchmarks. Higher is better except where noted (↓). SC = Self-Consistency (N=5, T=0.7, majority voting). Diversity is measured as mean pairwise cosine distance in embedding space. Exploration Value (EV) measures the fraction of queries for which a stochastic hypothesis achieves a higher scoring function value than the grounded baseline: $|\{Q : S(h_{\text{stoch}}) > S(h_{\text{grounded}})\}|/|\mathcal{Q}|$; EV is undefined for single-path baselines (–).

Method	Acc. (%)	C_{rel}	$\eta = \text{Acc}/C_{\text{rel}}$
Large model only	60	1.0	60
3× Small only	56	1.6	35
Self-Consistency (N=5)	62	5.0	12
FROST (ours)	**62**	**3.6**	**17**

Table 5: Accuracy versus inference cost on MMLU. C_{rel} is normalized to single large-model CoT (C_{rel}=1.0). FROST achieves the same accuracy as Self-Consistency at 28% lower cost (C_{rel}=3.6 vs. 5.0), yielding a 42% improvement in the efficiency ratio η.

Task Type	EV (%)
Simple / Factoid ($H(p) < 0.5$)	0−5
Constrained Reasoning ($0.5 \leq H(p) < 1.5$)	10−15
Multi-hop / Ambiguous ($H(p) \geq 1.5$)	23−34

Table 6: Exploration Value (EV) by task category. EV $= |\{Q : S(h_{\text{stoch}}) > S(h_{\text{grounded}})\}|/|\mathcal{Q}|$. Entropy $H(p)$ is computed from multiple-choice logits for CommonsenseQA and MMLU, and via sampling-based estimation for HotpotQA.

reasoning tasks (EV=10−15%) and are most pronounced on multi-hop and ambiguous problems (EV=23−34% for $H(p) \geq 1.5$). A query-level correlation of $\text{Corr}(\text{EV}, H(p)) \approx 0.67$ ($p < 0.01$) confirms that exploration benefit scales with task entropy. These results suggest that entropy-based routing, applying Standard CoT for $H(p) < \tau$ and FROST for $H(p) \geq \tau$, could further improve cost-efficiency; we leave formal validation of this strategy to future work.

5.5 Ablation Studies

We analyze the contribution of individual FROST components in Table 7.

Multi-criteria scoring is necessary: pooling candidates without scoring yields 66%, confirming that selection adds 4 points over unscored aggregation. Neither stream alone reaches full performance (stochastic-only: 58%, grounded-only: 65%), while the full system achieves 70%. Among individual criteria, coherence contributes most (−7 points when ablated), followed by factuality (−6 points) and novelty (−3 points). Removing novelty scoring reduces diversity from 0.60 to 0.35 without a commensurate accuracy gain, confirming that explicit diversity scoring is required to select hypotheses that differ substantively from the grounded baseline.

Variant	Acc. (%)	Diversity
Grounded only (Standard CoT)	65	0.20
Stochastic only (voting)	58	0.55
Grounded + Stochastic (no scoring)	66	0.52
No factuality check (γ=0)	64	0.58
No coherence check (β=0)	63	0.57
No novelty check (α=0)	67	0.35
FROST (full)	**70**	**0.60**

Table 7: Ablation results on HotpotQA, where supporting documents enable factuality scoring. Diversity is mean pairwise cosine distance in embedding space. Removing the stochastic sampler ("Grounded only") reduces diversity by 67% and accuracy by 5 points, demonstrating the contribution of stochastic exploration to ambiguity resolution. Removing the novelty criterion (α=0) suppresses diversity to 0.35, indicating that active novelty scoring is required to select genuinely distinct hypotheses.

5.6 Operational Characteristics

To evaluate FROST under adversarial conditions, we construct a 200-example subset of HotpotQA comprising unanswerable queries: questions modified so that supporting documents contain insufficient evidence for a definitive answer. An abstention is recorded when no candidate satisfies both the factuality threshold ($\text{Fact}(h) \geq 0.2$) and

Judge Model	Acc. (%)	C_{rel}	$\eta = \text{Acc}/C_{\text{rel}}$
7B (Llama-2-7B-Chat)	74.0	**3.6**	**20.6**
13B (Llama-2-13B-Chat)	**74.8**	5.0	15.0

Table 8: Judge-model ablation on CommonsenseQA (K=3). Using the 13B model as judge improves accuracy by 0.8 points but raises C_{rel} from 3.6 to 5.0, eliminating the 28% efficiency advantage over Self-Consistency. Since coherence scoring requires a bounded scalar output rather than open-ended reasoning, the 7B model constitutes a Pareto-optimal choice: no alternative improves both accuracy and efficiency simultaneously.

coherence threshold ($\text{Coh}(h) \geq \tau_{\text{coh}} = 0.5$); remaining low-confidence responses are excluded from the false-positive count, accounting for the gap between the abstention rate (34%) and the complement of the false-positive rate (49%). Results on this subset should be treated as indicative rather than definitive.

Metric	CoT	FROST
Abstention rate on unanswerable (%) ↑	8	**34**
False-positive rate on unanswerable (%) ↓	92	**51**
Wall-clock latency (seconds) ↓	**1.2**	1.4

Table 9: Operational characteristics on a 200-example adversarial subset with unanswerable queries. FROST's multi-criteria validation suppresses 26 percentage points more unanswerable confabulations than CoT, reducing the false-positive rate by 45% relative to CoT. Latency overhead is 0.2 seconds per query.

FROST abstains on 34% of unanswerable queries versus 8% for Standard CoT, reducing the false-positive rate from 92% to 51% (a 45% relative reduction). Wall-clock latency increases from 1.2 to 1.4 seconds per query (17% overhead); under batched inference, this overhead is further amortized.

Table 10 reports accuracy, relative FLOPs cost, and wall-clock latency across sample counts $K \in \{1, 2, 3, 5\}$.

Accuracy saturates at K=3 (70%), while cost continues to grow; K=3 is therefore the recommended operating point. These results confirm that FROST's validation mechanism provides meaningful reliability gains under adversarial conditions. We note, however, that the current system is validated at research scale only: no A/B testing or sustained human evaluation has been conducted, and latency profiling is performed on a single-node setup. Deployment at scale would require additional optimization, such as quantization, distillation, or adaptive sampling.

K	HotpotQA EM (%)	C_{rel}	Latency (s)
1	68	2.1	1.3
2	69	2.8	1.3
3	**70**	**3.6**	**1.4**
5	70	5.2	1.5

Table 10: Effect of sample count K on accuracy, FLOPs cost (C_{rel}), and wall-clock latency on HotpotQA. Accuracy saturates at K=3 while cost grows approximately linearly; K=5 approaches Self-Consistency cost (C_{rel}=5.0) without further accuracy gain. Wall-clock latency grows slowly because small-model calls execute in parallel, so C_{rel} (total FLOPs) and wall-clock time diverge with increasing K. Bold row is the default FROST configuration.

Conclusion

We presented FROST, an inference-time reasoning framework that balances accuracy and efficiency for production-oriented deployment. By combining low-temperature grounded generation from large models with high-temperature stochastic sampling from smaller models, followed by multi-criteria validation, FROST achieves 2–5 percentage point improvements over Standard CoT at 28% lower cost than Self-Consistency. Controlled stochasticity with rigorous scoring yields productive exploration: stochastic hypotheses outperform the grounded baseline on 23–34% of questions, and the validation mechanism abstains on 34% of unanswerable queries versus 8% for CoT, reducing false positives by 45%. Exploration benefit correlates with task ambiguity, suggesting principled entropy-based routing as a direction for future work. Operating entirely at inference time without training or context augmentation, FROST is immediately applicable to existing checkpoints in settings where fine-tuning or retrieval infrastructure is unavailable.

Limitations

FROST is evaluated on English QA benchmarks; generalization to math, code, and multimodal tasks is untested. The token-overlap factuality metric penalizes paraphrases, and LLM-as-judge scoring may introduce bias. Weights tuned on HotpotQA may not transfer across domains, and reliance on model logits limits use with closed APIs.

Ethics Statement

This work introduces an inference-time reasoning framework and does not involve the collection, annotation, or release of new datasets. All experiments use publicly available benchmarks (HotpotQA, CommonsenseQA, MMLU) and open-source models (Llama-2). No human subjects were involved in the evaluation, and no personally identifiable information was used at any stage.

We acknowledge that LLM-as-judge scoring may carry biases inherited from pre-training data, which could affect hypothesis selection in ways that are difficult to audit. We recommend monitoring for such biases before operational deployment. While FROST reduces unsupported outputs relative to baselines, it does not eliminate confabulation entirely; users in high-stakes domains should treat outputs as assistive rather than authoritative.

The computational cost of FROST is modest relative to comparable ensemble methods, and all experiments were conducted on a single-node setup. We encourage future work to report energy consumption and carbon footprint alongside accuracy metrics.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. We also thank the developers of the Llama-2 model family and the maintainers of the HotpotQA, CommonsenseQA, and MMLU benchmarks for making their resources publicly available.

References

Fedor Borisyuk, Mingzhou Zhou, Qingquan Song, Siyu Zhu, Birjodh Tiwana, Ganesh Parameswaran, Siddharth Dangi, Lars Hertel, Qiang Xiao, Xiaochen Hou, Yunbo Ouyang, Aman Gupta, Sheallika Singh, Dan Liu, Hailing Cheng, Lei Le, Jonathan Hung, Sathiya Keerthi, Ruoyan Wang, and 15 others. 2024. Lirank: Industrial large scale ranking models at linkedin. *Preprint*, arXiv:2402.06859.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

Zihao Chen, Zihan Lin, Xinhua Chen, Zhiyi Liu, Changxu Liu, Yuxuan Qiao, Yifei Feng, Junjie Zuo, Yifan Song, and Fan Yang. 2025. Spec2doc2rtl: Rtl generation from specification with natural language representation. In *2025 International Symposium of Electronics Design Automation (ISEDA)*, pages 784–789. IEEE.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Miroslava Dimitrova. 2025. Retrieval-augmented generation (rag): Advances and challenges. *Probl. Eng. Cybern. Robot*, 83:32–57.

Jiechao Gao, Rohan Kumar Yadav, Yuangang Li, Yuandong Pan, Jie Wang, Ying Liu, and Michael Lepech. 2026. Llm-guided semantic bootstrapping for interpretable text classification with tsetlin machines. *arXiv preprint arXiv:2604.12223*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Joongho Kim, Xirui Huang, Zarreen Reza, and Gabriel Grand. 2025. Chopping trees: Semantic similarity based dynamic pruning for tree-of-thought reasoning. *arXiv preprint arXiv:2511.08595*.

Jiawei Li, Yang Gao, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, and 1 others. 2025a. Fundamental capabilities and applications of large language models: A survey. *ACM Computing Surveys*.

Kun Li, Tianhua Zhang, Xixin Wu, Hongyin Luo, James Glass, and Helen Meng. 2025b. Decoding on graphs: Faithful and sound reasoning on knowledge graphs through generation of well-formed chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24349–24364.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

Yash Panjari. 2025. Automated quality assurance systems using llm-as-judge for conversational ai testing: A technical review. *Journal Of Engineering And Computer Sciences*, 4(11):96–104.

Md Rizwan Parvez. 2025. Chain of evidences and evidence to generate: Prompting for context grounded and retrieval augmented reasoning. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 230–245.

Jakub Piwko, Jędrzej Ruciński, Dawid Płudowski, Antoni Zajko, Patryzja Żak, Mateusz Zacharecki, Anna Kozak, and Katarzyna Woźnica. 2025. Divide, specialize, and route: A new approach to efficient ensemble learning. *arXiv preprint arXiv:2506.20814*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 3(4):333–389.

Ming Shao and Haichun Zhang. 2025. Two-stage prompting framework with predefined verification steps for evaluating diagnostic reasoning tasks on two datasets. *npj Digital Medicine*.

Zohaib Hasan Siddiqui, Jiechao Gao, Ebad Shabbir, Mohammad Anas Azeez, Rafiq Ali, Gautam Siddharth Kashyap, and Usman Naseem. 2025. LLMs on a budget? say HOLA. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1035–1043, Suzhou (China). Association for Computational Linguistics.

Weicheng Song, Siyou Guo, Mingliang Gao, Qilei Li, Xianxun Zhu, and Imad Rida. 2025. Deepfake detection via feature refinement and enhancement network. *Image and Vision Computing*, page 105663.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2025. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3613–3635.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William 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.