SFAD: Speculative Factuality-Aware Decoding

Guanqiao Chen[1,2], Di Wang[3,4], Lijie Hu[1,*]
[1]MBZUAI
[2]University of Science and Technology of China
[3]Provable Responsible AI and Data Analytics (PRADA) Lab
[4]King Abdullah University of Science and Technology
[*]Corresponding author.

Abstract

As one of the most critical challenges in large language models, contextual faithfulness directly determines their reliability in knowledge-intensive applications. This task is particularly challenging as it requires balancing factual consistency with generation efficiency. Contrastive decoding methods require dual forward passes (with and without context) to compare model outputs, doubling inference computational overhead, while post-training alignment demands extensive reinforcement learning with substantial computational overhead. To address this challenge, we present **SFAD**, a speculative decoding framework that enhances contextual faithfulness without inference degradation. We first construct **ConFide**, a preference dataset with fine-grained atomic perturbations, to train a context-faithful draft model via Direct Preference Optimization. During inference, Epistemic Friction detects potential hallucinations by quantifying distributional tension weighted by specialist certainty. When friction exceeds the threshold, Asymmetric Logit Steering refines the target distribution through residual-based logit injection; otherwise, standard speculation proceeds. Extensive experiments demonstrate that SFAD substantially improves faithfulness while achieving 2.48× speedup, offering a practical solution for efficient LLMs.

1 Introduction

Large Language Models (LLMs) (Guo et al., 2025a; Achiam et al., 2023) have achieved remarkable success by leveraging vast parametric knowledge acquired during pretraining (Petroni et al., 2019; Roberts et al., 2020). However, this internal knowledge is inherently static and constrained by the training distribution, rendering it prone to becoming outdated or incomplete (Cheng et al., 2024b; Li et al., 2026b; Zhang et al., 2025a; Cheng et al., 2024a). To address this limitation, retrieval-augmented generation (RAG) (Qin et al., 2025; Guu et al., 2020; Cheng et al., 2025a) and external tool integration have become increasingly prevalent, intensifying the demand for models to prioritize provided context over parametric priors. Yet LLMs frequently exhibit hallucinations (Huang et al., 2025a; Ji et al., 2023; Cheng et al., 2025b) due to *knowledge conflicts*, where models favor internal knowledge over task-specific external contexts (Xie et al., 2024; Chen et al., 2022; Zhang et al., 2025b), undermining their contextual faithfulness in knowledge-intensive applications.

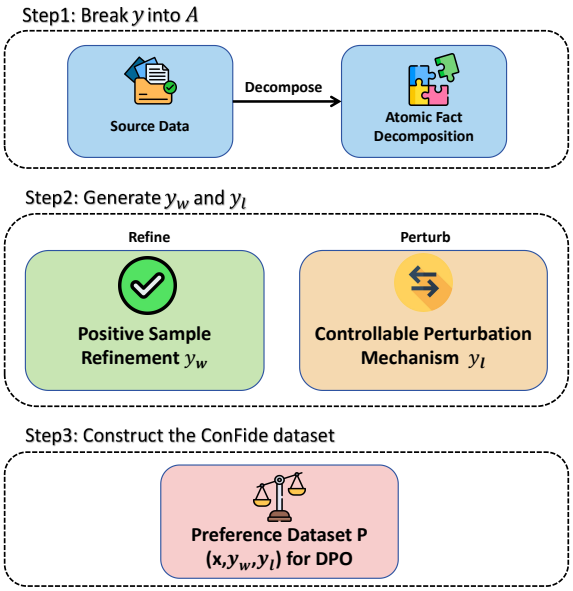

Figure 1: The ConFide pipeline for factuality-aware preference data construction.

Enhancing contextual faithfulness while maintaining inference efficiency presents significant challenges. Decoding-based methods (Shi et al., 2024; Xu, 2023; Khandelwal et al., 2025; Yang et al., 2025c) contrast logits from contextualized inputs against uncontextualized baselines to amplify evidence-aligned signals. However, this requires dual forward, effectively doubling computational cost and halving generation speed. Post-training

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 24221–24239
October 24-29, 2026 ©2026 Association for Computational Linguistics

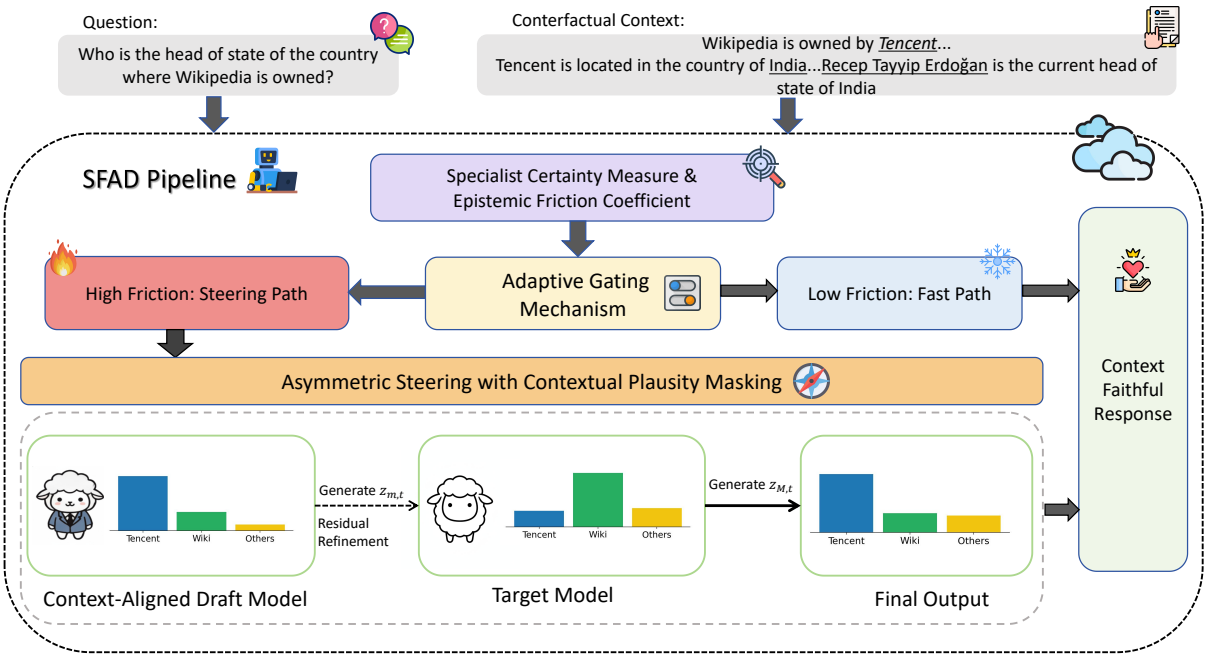

Figure 2: SFAD framework.

methods (Yan et al., 2025; Wang et al., 2025c; Yang et al., 2025b; Zhou et al., 2026) employ reinforcement learning to align models with contextual evidence, but typically require extensive compute and large-scale preference data. These limitations motivate the need for approaches that enhance faithfulness without compromising efficiency or requiring substantial resources.

To resolve these problems, we propose **SFAD**, a speculative decoding framework that unifies contextual faithfulness enhancement with speculative acceleration. Our key insight is to leverage a *context-faithful draft model* as a factuality sentinel: by training a smaller model to prioritize contextual evidence, we enable it to detect and correct hallucinations in the larger target model during speculative decoding. Specifically, we first construct **ConFide**, a preference dataset that improves contextual faithfulness by generating diverse hard-negative samples through atomic decomposition and controllable perturbations. Training the draft model on ConFide and ConFiQA (Bi et al., 2025) via Direct Preference Optimization instills strong contextual loyalty. During inference, SFAD introduces *Epistemic Friction*, a metric that quantifies distributional tension weighted by specialist certainty to dynamically detect knowledge conflicts. When friction signals potential hallucinations, *Asymmetric Logit Steering* selectively refines the target distribution through residual-based injection; otherwise, standard speculative decoding proceeds unmodified. This adaptive mechanism maintains the $2 \sim 3\times$ speedup of speculative decoding while substantially improving contextual faithfulness.

Our contributions are threefold: **(1) Dataset:** we introduce ConFide, a fine-grained preference dataset designed for training context-faithful draft models through atomic-level perturbations; **(2) Framework:** we propose **SFAD**, the first speculative decoding framework specifically designed for hallucination mitigation, which features *Epistemic Friction* for conflict detection and *Asymmetric Logit Steering* for selective correction; and **(3) Evaluation:** extensive experiments demonstrate that SFAD achieves substantial faithfulness improvements across diverse benchmarks while delivering a $2.48\times$ speedup, approaching the performance of models $5\times$ larger with minimal computational overhead.

2 Related Work

2.1 Hallucinations in LLMs

Hallucinations(Huang et al., 2025b) occur when LLM outputs appear plausible yet deviate from factual or contextual knowledge (Kaddour et al., 2023; Ji et al., 2023). These are typically bifurcated into *factuality hallucinations*, which contradict real-world facts based on internal parametric knowledge (Min et al., 2023; Wei et al., 2024),

and *faithfulness hallucinations*, characterized by inconsistency with the provided input or grounding documents (Wang et al., 2026; Wan et al., 2023; Hu et al., 2025; Guo et al., 2025b). Mitigation strategies span the model lifecycle: training-phase efforts focus on data curation and grounding (Pan et al., 2024), while inference-stage methods employ confidence estimation (Huang et al., 2025c), knowledge retrieval (Feng et al., 2024), and editing (Ali et al., 2025; Wang et al., 2025a). Despite these advancements, LLMs remain prone to inconsistent outputs in context-rich tasks such as RAG and summarization (Xu et al., 2024; Li and Yu, 2025). **Our work leverages the Speculative Decoding framework to ensure context-consistent generation.**

2.2 Speculative Decoding

Speculative decoding accelerates LLM inference by verifying draft tokens generated by a smaller model against the target LLM (Leviathan et al., 2023; Chen et al., 2023). Drafting strategies have evolved from independent models (Xia et al., 2023) to architectural extensions such as recurrent feature utilization (Li et al., 2024) and diffusion-based generation (Li et al., 2026a). Efficiency has been further optimized through tree-based parallel verification (Miao et al., 2024), sparse Mixture-of-Experts (MoE) acceleration (Huang et al., 2025d), and specialized reinforcement learning system enhancements (Chen et al., 2026). Recent research has also tailored SD for diverse and complex scenarios, including long-context retrieval-augmentation (Sun et al., 2024; Sadhukhan et al., 2025; Yang et al., 2026; Chen et al., 2025), multi-sample inference tasks (Li et al., 2025), and memory-efficient inference via quantized caching (Tiwari et al., 2025). Beyond performance, safety-awareness has been integrated into the speculative framework to ensure alignment (Wang et al., 2025d). Unlike these performance-driven or safety-centric works, **SFAD is the first speculative framework to target hallucination mitigation.**

3 Aligning Draft Models for Context-Faithfulness via ConFide

To enhance the draft model's generalization across diverse hallucination types, we construct **ConFide**, a fine-grained preference dataset that systematically diversifies error patterns in negative samples and stylistic variations in positive samples. As illustrated in Figure 1, our pipeline leverages atomic fact decomposition and controllable perturbations to produce high-quality contrastive pairs.

3.1 Data Construction Pipeline

Source Data and Problem Formulation. We build upon the LLM-AggreFact (Tang et al., 2024a) and CG2C (Lei et al., 2025) datasets, which provide fact-intensive samples covering diverse hallucination patterns suitable for atomic decomposition. Our initial dataset is defined as $\mathcal{D} = \{(x^{(i)}, y_{\text{ref}}^{(i)}, l^{(i)})\}_{i=1}^{N}$, where x is the source context, y_{ref} is the candidate response, and $l \in \{0, 1\}$ is the binary faithfulness label. Our goal is to construct a preference dataset $\mathcal{P} = \{(x, y_w, y_l)\}$ for DPO training, where y_w is faithful to context x and y_l contains specific hallucinations.

Atomic Fact Decomposition. We decompose each response into verifiable atomic facts via a decomposition function f_{dec}:

$$\mathcal{A} = f_{\text{dec}}(y) = \{a_1, a_2, \ldots, a_m\}, \tag{1}$$

where each atomic fact a_j represents a minimal verifiable claim.

Negative Sample Generation. We apply a Controllable Perturbation Mechanism to transform atomic facts into corrupted versions. Given $a_j = (s_j, r_j, o_j)$ as a subject-relation-object triple, we define three perturbation operators:

$$\phi(a_j) = \begin{cases} (s_j, r_j, o_j'), & \tau_{\text{ent}} \\ (s_j, r_j, o_j + \epsilon), & \tau_{\text{num}} \\ (s_j, \neg r_j, o_j), & \tau_{\text{rel}} \end{cases} \tag{2}$$

where τ_{ent}, τ_{num}, and τ_{rel} denote entity swap, numerical distortion, and relation inversion, respectively. Here, o_j' is a confusing entity drawn from context x, ϵ is calibrated noise preserving numerical plausibility, and $\neg r_j$ denotes logical negation of relation r_j. The perturbed facts are reconstructed into a fluent response $y_l = f_{\text{rec}}(\mathcal{A}')$, yielding hallucinated yet linguistically natural negative samples.

Positive Sample Refinement. Winning samples y_w are constructed based on the faithfulness label l. If $l = 1$, we apply paraphrasing π_{para} to generate semantically equivalent but syntactically varied sentences. If $l = 0$, a teacher model (GPT-4o (Hurst et al., 2024)) corrects y_{ref} against context x to produce a strictly faithful y_w. This ensures winning samples reflect genuine factual alignment rather than surface-level patterns.

Preference Dataset Construction. Pairing the perturbed y_l with the refined y_w yields the final preference dataset $\mathcal{P}$, where negative samples exhibit fine-grained, fluent hallucinations and positive samples maintain strict contextual fidelity.

3.2 Draft Model Optimization via DPO

We optimize the draft model π_θ via the DPO objective (Rafailov et al., 2023), which directly maximizes the preference margin between faithful responses y_w and hallucinated responses y_l without requiring an explicit reward model. The gradient signal encourages π_θ to concentrate probability mass on tokens strictly verifiable against the context x, controlled by hyperparameter β relative to the frozen reference model π_{ref}.

4 SFAD: Speculative Factuality-Aware Decoding

We formulate **Speculative Factuality-Aware Decoding (SFAD)** as a dynamic inference framework designed to decouple token verification from distribution correction. Figure 2 illustrates the overall pipeline. Let M and m denote the target (generalist) and draft (specialist) models with parameters θ_M and θ_m, respectively. At each time step t, given the prefix context $x_{<t}$, the models produce logit vectors $\mathbf{z}_{M,t}, \mathbf{z}_{m,t} \in \mathbb{R}^{|\mathcal{V}|}$. Our framework is governed by a five-step formalism that adaptively steers the generation process when the specialist detects potential factual inconsistencies. Theoretical analysis and proofs are provided in Appendix E.

4.1 Specialist Certainty Measure

To prevent noise injection from an uncertain draft model, we first quantify its internal confidence. Unlike standard entropy measures, we require a normalized metric that penalizes high-entropy (uncertain) distributions. We define the **Specialist Certainty** κ_t as:

$$\kappa_t = \left(1 - \frac{\mathbb{H}(\mathbb{P}_m(\cdot|x_{<t}))}{\log|\mathcal{V}|}\right)^\gamma, \quad \text{where } \mathbb{P}_m = \text{Softmax}(\mathbf{z}_{m,t}). \tag{3}$$

where $\mathbb{H}(\cdot)$ denotes the Shannon entropy, $|\mathcal{V}|$ is the vocabulary size, and $\gamma \geq 1$ is a sharpening coefficient. Here, $\kappa_t \to 1$ implies the specialist is highly confident in its prediction, serving as a necessary precondition for intervention.

4.2 Epistemic Friction Coefficient

To distinguish factual conflicts from benign linguistic diversity, we propose the **Epistemic Friction** $\mathcal{F}_t$, which captures the distributional tension between the generalist and the specialist. It is formulated as the Jensen-Shannon (JS) divergence weighted by the specialist's certainty:

$$\mathcal{F}_t = \mathcal{D}_{\text{JS}}\left(\mathbb{P}_M(\cdot|x_{<t}) \,\|\, \mathbb{P}_m(\cdot|x_{<t})\right) \cdot \kappa_t \tag{4}$$

$\mathcal{F}_t$ acts as a detector for *confident hallucinations*. High friction occurs if and only if the models disagree significantly ($\mathcal{D}_{\text{JS}}$ is high) *and* the specialist is factually convinced (κ_t is high).

4.3 Adaptive Gating Mechanism

To ensure computational efficiency and avoid overcorrection, SFAD employs a soft gating scalar $\lambda_t \in [0, 1]$ to govern the intensity of the steering. It is derived via a shifted sigmoid activation:

$$\lambda_t = \sigma\left(\beta \cdot (\mathcal{F}_t - \tau)\right) = \frac{1}{1 + \exp\left(-\beta(\mathcal{F}_t - \tau)\right)} \tag{5}$$

where τ is the friction threshold and β controls the transition sharpness. This creates a switching regime: when $\mathcal{F}_t \ll \tau$, $\lambda_t \to 0$ (Standard Speculation); when $\mathcal{F}_t \gg \tau$, $\lambda_t \to 1$ (Steering Mode).

4.4 Asymmetric Steering with Contextual Plausibility Masking

When intervention is triggered, we modify the target logits. To ensure that the specialist only intervenes when its predictions are linguistically plausible within the current sequence, we introduce a **Contextual Plausibility Mask (CPM)**. We define the plausible set $\mathcal{V}_{CPC}$ as:

$$\mathcal{V}_{CPC} = \{v \in \mathcal{V} : \mathbb{P}_m(v|x_{<t}) \geq \eta p_t^{\max}\} \tag{6}$$

where $p_t^{\max}$ is the maximum probability assigned by $\mathbb{P}_m(\cdot|x_{<t})$, and $\eta \in (0, 1)$ is a plausibility threshold.

Inspired by previous work (Bachmann et al., 2025), which demonstrated that alignment-based verification in speculative decoding rejects many high-quality tokens due to distribution mismatches, we extend this insight beyond token-level acceptance. Instead of solely adapting verification schemes, we intervene at the logit level to inject factuality-aware corrections, ensuring both efficiency and reduced hallucinations.The corrected

logit vector $\mathbf{z}_t^*$ is computed as:

$$\mathbf{z}_t^* = \mathbf{z}_{M,t} + \lambda_t \cdot \text{ReLU}(\mathbf{z}_{m,t} - \mathbf{z}_{M,t}) \cdot \mathbb{I}(x \in \mathcal{V}_{CPC}) \tag{7}$$

The ReLU$(\cdot)$ operator enforces *unidirectional knowledge injection*, allowing the specialist to boost correct entities without penalizing the generalist's linguistic fluency. The $\mathbb{I}(x \in \mathcal{V}_{CPC})$ term acts as a safety guard, ensuring the specialist only steers the target toward tokens that maintain "compositional plausibility."

4.5 Hybrid Decoding Policy

Finally, the next token x_t is determined by a hybrid policy that switches between the corrected distribution and standard speculative verification:

$$x_t \sim \begin{cases} \text{Softmax}(\mathbf{z}_t^*) & \text{if } \mathcal{F}_t \geq \tau \quad \text{(Steering Path)} \\ \text{Verify}(\tilde{x}_t, \mathbb{P}_M) & \text{if } \mathcal{F}_t < \tau \quad \text{(Fast Path)} \end{cases} \tag{8}$$

In the **Steering Path**, we sample from the steered logits, effectively overriding the draft. In the **Fast Path**, we perform standard rejection sampling using the draft token $\tilde{x}_t$ and the original target distribution $\mathbb{P}_M$, maintaining the speed guarantees of speculative decoding.

5 Experiments

5.1 Experimental Setup

Models and Configurations. We implement the SFAD framework using the Qwen3 family (Yang et al., 2025a). Specifically, Qwen3-1.7B is utilized as the draft model, which is fine-tuned via Direct Preference Optimization to serve as the expert model.The target model is Qwen3-14B.

Data Construction for DPO. The DPO training set for the draft model consists of approximately 36K curated samples, combining two complementary sources. First, we directly adopt 18K instances from **ConFiQA** (Bi et al., 2025), a benchmark focusing on multi-hop knowledge conflicts. Second, we construct our **ConFide** dataset using the atomic perturbation pipeline described in Section 3. Specifically, ConFide is synthesized from two source datasets: 12K samples from LLM-AggreFact (Tang et al., 2024a), and 6K samples from the CG2C dataset (Lei et al., 2025).

Baselines. We compare SFAD against two categories of baselines to demonstrate its superiority in both contextual faithfulness and inference efficiency. The first category consists of decoding-level methods applied to our target model, Qwen3-14B: Greedy Decoding, Context-Aware Decoding (CAD) (Xu, 2023), COIECD (Yuan et al., 2024), and AdaCAD (Wang et al., 2025b). The second category includes a frontier model, Llama-3.1-70B-Instruct (MetaAI, 2024), which has $5\times$ the parameters of Qwen3-14B, serving as a high-performance reference point.

Evaluation Tasks and Datasets. We evaluate across five task categories: (1) *Factual Retrieval* on HotpotQA (Yang et al., 2018), PopQA (Mallen et al., 2023), and TriviaQA (Joshi et al., 2017); (2) *Abstractive Faithfulness* via summarization on TofuEval (Tang et al., 2024b) and XSum (Narayan et al., 2018); (3) *Extended Generation* on CLAPNQ (Rosenthal et al., 2025), ExpertQA (Malaviya et al., 2024), and HAGRID (Kamalloo et al., 2023); (4) *Knowledge Conflicts* using 200 held-out instances from LLM-AggreFact; and (5) *General Capabilities* on GSM8K (Cobbe et al., 2021) and Just-Eval (Lin et al., 2024).

Quality and Efficiency Metrics. For generative quality, we use standard metrics tailored to each task: Exact Match (EM) for retrieval tasks; AlignScore, BERT-P, and ROUGE-L for summarization; and FaithScore (computed via MiniCheck (Tang et al., 2024a)) for long-form QA. For knowledge conflict analysis, we adopt ConFiQA (Bi et al., 2025) metrics: Context-faithful Frequency (P_c), Original Factual Frequency (P_o), and Memory Reliance (M_R). To quantify the inference acceleration of SFAD, we define the Average Token Generation Acceleration (ATGA), measured in an end-to-end, context-aware setting:

$$\text{ATGA} = \frac{\text{Avg. token generation time w/o SFAD}}{\text{Avg. token generation time w/ SFAD}} \tag{9}$$

5.2 Main Results

We present comprehensive evaluations across three categories of context-intensive tasks: Foundation QA, Summarization, and Long-Form QA.

SFAD demonstrates consistent metric improvements across all evaluated benchmarks, optimizing both task accuracy and inference efficiency. While standard decoding-level interventions typically incur latency penalties due to redundant forward passes, SFAD integrates a speculative decoding framework that streamlines the generation process. This efficiency is driven by our design's focus on enhancing contextual consistency, which allows

Table 1: Performance comparison on Foundation QA. **Rel. Latency** denotes the inference time relative to the Qwen3-14B vanilla baseline.

Method	TriviaQA	HotpotQA	PopQA	Rel. Latency
Vanilla Baseline				
Qwen3-14B	53.87	41.77	78.21	1.00×
Decoding-level Baselines (Qwen3-14B)				
CAD	41.43	39.51	71.29	2.00×
AdaCAD	82.11	45.63	77.39	2.15×
COIECD	83.07	45.63	76.29	2.40×
Frontier Model				
Llama-3.1-70B	**90.20**	**56.11**	86.11	4.85×
SFAD (Ours)	85.12	52.19	**86.39**	**0.82**×

Table 2: Summarization performance on XSum and TofuEval. SFAD achieves superior factuality while reducing inference overhead.

	XSum			TofuEval	
Method	R-L	BERT-P	AlignScore	AlignScore	Rel. Latency
Vanilla Baseline					
Qwen3-14B	13.67	91.67	72.86	59.84	1.00×
Decoding-level Baselines (Qwen3-14B)					
CAD	14.59	93.65	84.34	83.23	2.00×
AdaCAD	14.91	**94.29**	85.81	85.07	2.20×
COIECD	13.65	91.04	73.81	60.86	2.45×
Frontier Model					
Llama-3.1-70B	16.35	94.12	**87.48**	87.31	5.10×
SFAD (Ours)	**16.32**	93.97	87.37	**87.53**	**0.85**×

the model to leverage evidentiary alignment to accelerate token production. In Foundation QA, the method mitigates the internal knowledge limitations of the base model, a capability that extends to long-form generation where SFAD addresses the inherent trade-off between lexical fluency and factual grounding. By prioritizing context-aligned tokens, SFAD substantially narrows the performance gap between the 14B base model and the 5× larger Llama-3.1-70B frontier model. These results underscore the capacity of SFAD to maintain sustained evidentiary grounding and long-range coherence with reduced computational costs, avoiding the need for task-specific fine-tuning or increased parameter overhead.

6 Analysis

6.1 Effectiveness of ConFide and DPO

To isolate the contributions of our data construction pipeline and alignment strategy, we evaluate three draft model variants on 200 held-out knowledge-conflict instances from LLM-AggreFact. As illustrated in Figure 3, the transition from supervised fine-tuning (SFT) to preference optimization (DPO) triggers a fundamental shift in model behavior: while ConFide+SFT remains heavily anchored to internal priors (high M_R), DPO-trained variants

Table 3: Long-Form QA results. SFAD significantly enhances faithfulness with higher efficiency than the vanilla model.

	CLAPNQ		ExpertQA		HAGRID		
Method	R-L	Faith	R-L	Faith	R-L	Faith	Rel. Latency
Vanilla Baseline							
Qwen3-14B	17.12	59.73	31.56	51.29	16.96	57.63	1.00×
Decoding-level Baselines (Qwen3-14B)							
CAD	18.23	60.24	33.58	53.47	17.89	58.20	2.00×
AdaCAD	18.43	62.37	32.87	54.76	17.12	59.90	2.18×
COIECD	18.56	61.96	33.14	56.32	17.34	59.76	2.42×
Frontier Model							
Llama-3.1-70B	**42.15**	**92.45**	**46.10**	**72.40**	**52.07**	**82.20**	5.30×
SFAD (Ours)	41.34	90.93	43.79	71.13	51.97	81.99	**0.78**×

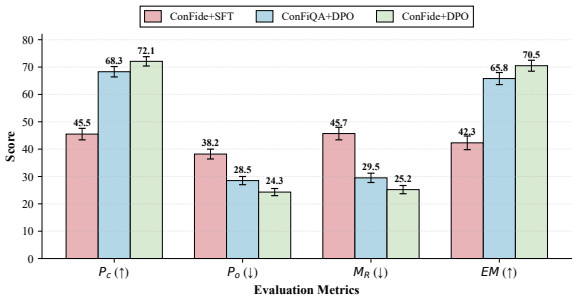

Figure 3: Impact of training strategies on knowledge conflict resolution. ConFide+DPO consistently outperforms baselines across context faithfulness (P_c, EM) and memory reliance (M_R, P_o) metrics.

demonstrate a superior ability to suppress parametric bias in favor of contextual evidence.

Notably, ConFide+DPO consistently outperforms ConFiQA+DPO across all metrics, validating the efficacy of our atomic perturbation mechanism. By exposing the model to hard negatives—such as entity swaps and relation inversions—ConFide provides a more granular discriminative signal that sharpens the model's sensitivity to factual nuances. This contrastive optimization effectively penalizes hallucinated completions that appear superficially plausible but lack contextual grounding. **These results confirm that our data construction pipeline combined with preference optimization is crucial for training context-faithful draft models capable of effective speculative decoding.**

6.2 Latency-Faithfulness Trade-off Analysis

A critical question in deploying context-faithful LLMs is whether achieving high factuality necessitates sacrificing inference speed. To address this, we analyze the trade-off between inference acceleration and contextual faithfulness across different decoding strategies. We evaluate all methods on the same 200-instance LLM-AggreFact test set

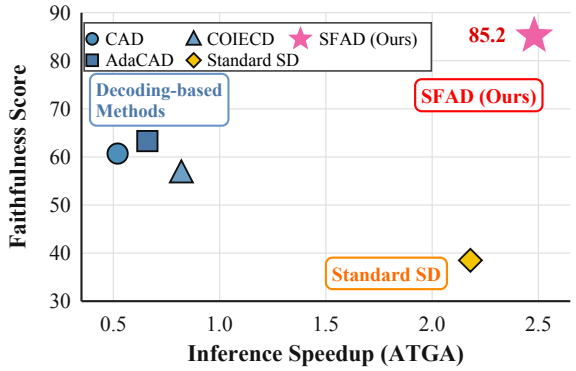

Figure 4: Latency-faithfulness trade-off analysis. SFAD achieves Pareto-optimal performance, simultaneously delivering 2.48× speedup and 85.2 faithfulness score. Decoding-level baselines suffer from severe latency penalties (<0.85×) despite moderate faithfulness, while standard SD prioritizes speed (2.18×) at the cost of poor faithfulness (38.5). SFAD outperforms the best baseline by 21.7 points while maintaining 2.9× faster inference.

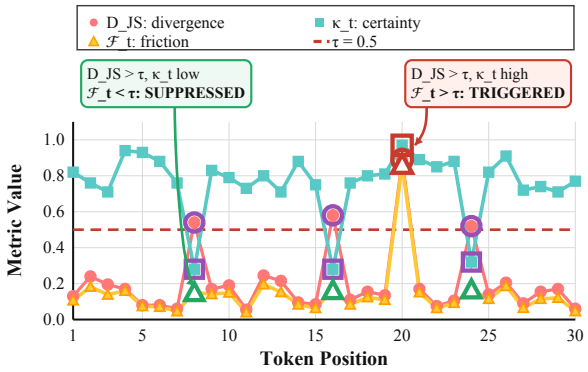

Figure 5: Epistemic Friction dynamics. Purple circles mark positions where $D_{JS} > \tau$ is suppressed by low κ_t. The red square marks a confident hallucination that triggers steering.

used in Section 6.1, measuring both Average Token Generation Acceleration (ATGA, Eq. (9)) and Faithfulness Score (MiniCheck FaithScore). Figure 4 visualizes this relationship, revealing three distinct performance regimes.

Decoding-level baselines exhibit poor speed-faithfulness balance. CAD, AdaCAD, and COIECD achieve moderate faithfulness scores but suffer severe latency penalties (ATGA 0.52×–0.82×). They need to perform two forward passes for both contextualized and baseline inputs. This introduces additional token overhead, which slows down the model's inference speed.

Standard speculative decoding prioritizes speed over faithfulness. While achieving 2.18× speedup, vanilla SD with an unaligned draft model produces drastically poor faithfulness, falling below even the decoding baselines. The draft model's lack of contextual grounding causes it to propose tokens based on parametric priors rather than provided evidence, leading to significant hallucinations in knowledge-intensive scenarios.

SFAD achieves Pareto-optimal performance. SFAD strikes an optimal balance between contextual faithfulness and inference speed, simultaneously delivering 2.48× speedup and 85.2 faithfulness score.

6.3 Epistemic Friction Analysis: Detecting Confident Hallucinations

To validate our Epistemic Friction mechanism, we analyze 1,200 samples from MQUAKE (Zhong et al., 2023), tracking D_{JS}, κ_t, and $\mathcal{F}_t$ across generation. Figure 5 illustrates metric dynamics across a representative generation sequence. The raw divergence D_{JS} fluctuates frequently, with peaks exceeding $\tau = 0.5$, but relying solely on D_{JS} would trigger excessive false positives at stylistic variations.

The specialist certainty κ_t provides crucial filtering. At positions marked with purple circles, D_{JS} slightly exceeds τ (0.52–0.58) but κ_t remains low (0.28–0.32), indicating uncertainty at linguistic alternatives. Consequently, $\mathcal{F}_t = D_{JS} \times \kappa_t$ is suppressed below threshold (0.15–0.17), correctly avoiding intervention.

In contrast, the red-square position exhibits a confident hallucination: $D_{JS} \approx 0.89$ combined with $\kappa_t \approx 0.97$ produces a sharp spike in $\mathcal{F}_t$ ($\sim$0.86), triggering full logit steering. By conditioning on both distributional conflict and specialist confidence, **Epistemic Friction provides a principled trigger for factuality correction, avoiding the over-correction pitfalls of divergence-only metrics.**

6.4 Quantifying Logit Steering Impact

To evaluate our asymmetric logit steering, we measure the average probability of faithful tokens across three decoding strategies on the MQUAKE test set, where faithful tokens are context-grounded entities that correctly answer the query.

Table 4 reveals the fundamental difference between verification-based and steering-based approaches. The original target model assigns 18.73% probability to faithful tokens, reflecting parametric bias where internal knowledge conflicts with contextual evidence. While the target model con-

Table 4: Average probability of faithful tokens. Standard SD provides negligible improvement, while SFAD achieves substantial gains through logit steering.

Strategy	Prob (%)	Abs. Gain	Rel. Gain
Original Target	18.73	–	1.0×
Standard SD	18.91	+0.18	1.01×
SFAD (Ours)	62.45	+43.72	3.33×

Table 5: General utility evaluation.

	GSM8K	Just-Eval (1–5)				
Method	**Acc**	**Help.**	**Clar.**	**Fact.**	**Depth**	**Eng.**
Target (Greedy)	91.35	4.15	4.90	4.30	4.55	4.75
Standard SD	91.32	4.14	4.89	4.28	4.54	4.74
SFAD (Ours)	91.27	4.11	4.88	4.33	4.53	4.73

Table 6: Llama-3.1-8B results on Foundation QA.

Method	TriviaQA	HotpotQA	PopQA	Rel. Latency
Vanilla Baseline				
Llama-3.1-8B	49.23	38.45	75.18	1.00×
Decoding-level Baselines (Llama-3.1-8B)				
CAD	37.82	36.21	68.45	2.00×
AdaCAD	78.56	42.31	74.28	2.18×
COIECD	79.43	42.88	75.56	2.35×
SFAD (Ours)	**81.67**	**48.73**	**83.21**	**0.85×**

Table 7: Ablation of sharpening coefficient γ.

γ	Steer (%)	Faith. (↑)	Speedup (↑)	R-L (↑)
1	31.2	83.7	2.31×	15.89
2 (Default)	**22.4**	**85.2**	2.48×	**16.32**
3	15.8	84.1	2.61×	16.18
4	9.3	78.5	**2.74**×	15.94

siders the faithful token, it remains overshadowed by parametric priors.

Standard speculative decoding provides virtually no correction, with faithful token probability remaining at 18.91%. This stagnation is inherent to SD's design: the verification mechanism determines whether to accept or reject draft tokens based on $P(\text{accept}) \propto \frac{P_M(x)}{P_m(x)}$, but it does not modify the target model's output distribution. When the draft proposes a faithful token that the target assigns low probability, SD simply rejects it and resamples from the original P_M, perpetuating the same parametric bias. This reveals a fundamental limitation: verification-only approaches operate through accept/reject decisions without modifying the target distribution, thus they cannot correct distributional biases but only accelerate generation.

In contrast, SFAD elevates faithful token probability to 62.45%, achieving a 3.33× relative improvement and a 243× larger absolute gain compared to standard SD. This dramatic boost results from our asymmetric steering mechanism (Eq. (7)): when epistemic friction $\mathcal{F}_t$ exceeds threshold τ, the system triggers ReLU-based logit injection $z_t^* = z_{M,t} + \lambda_t \cdot \text{ReLU}(z_{m,t} - z_{M,t})$, directly amplifying the draft model's high logits for contextually faithful tokens. Unlike verification that merely accepts or rejects, steering fundamentally reshapes the target's output distribution, transforming faithful tokens from minority candidates (18.73%) to dominant choices (62.45%). Critically, this intervention occurs selectively as it only activates when both distributional conflict and specialist conviction are present, maintaining the 2.48× speedup while achieving 85.2 faithfulness score. **This analysis demonstrates that logit-level steering provides a principled mechanism for resolving knowledge conflicts, achieving context faithfulness unattainable through verification alone.**

6.5 SFAD Preserves General Utility

We evaluate on GSM8K (Cobbe et al., 2021) and Just-Eval (Lin et al., 2024) to verify SFAD maintains general capabilities. As shown in Table 5, SFAD achieves 91.27% on GSM8K with negligible Just-Eval degradation. This stems from the selective nature of $\mathcal{F}_t$: when no knowledge conflict exists, friction remains below τ and the system defaults to standard speculative decoding, confirming SFAD's factuality enhancement does not sacrifice general utility or reasoning capability.

6.6 Generalization and Parameter Analysis

To evaluate generalization across model families, we apply SFAD with Llama-3.1-8B as the target model. As shown in Table 6, SFAD consistently improves faithfulness while maintaining inference efficiency. The coefficient γ modulates specialist certainty κ_t to regulate steering sensitivity. As shown in Table 7, lower γ increases intervention frequency but risks incorporating unreliable signals, while higher γ improves speedup at the cost of faithfulness.

7 Conclusion

In this work, we propose **SFAD**, the first speculative decoding framework unifying contextual faithfulness with acceleration. Leveraging a context-faithful drafter as factuality sentinel, SFAD detects hallucinations via epistemic friction. Drafters are

trained on ConFide, a dataset with atomic-level hallucination perturbations. At inference, SFAD selectively steers target logits upon conflict detection while preserving efficiency, achieving simultaneous speedup and faithfulness gains.

Limitations

SFAD requires a domain-aligned draft model trained via DPO, which introduces additional data construction overhead compared to standard speculative decoding. Furthermore, the friction threshold τ may require tuning when applied to out-of-distribution domains.

Acknowledgment

Di Wang is supported in part by the funding BAS/1/1689-01-01, RGC/3/7125-01-01, FCC/1/5940-20-05, FCC/1/5940-06-02, and King Abdullah University of Science and Technology (KAUST) – Center of Excellence for Generative AI, under award number 5940 and a gift from Google.
Lijie Hu is supported in part by the funding BF0100.

References

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*.

Muhammad Asif Ali, Nawal Daftardar, Mutayyba Waheed, Jianbin Qin, and Di Wang. 2025. MQA-KEAL: Multi-hop question answering under knowledge editing for Arabic language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5629–5644, Abu Dhabi, UAE. Association for Computational Linguistics.

Gregor Bachmann, Sotiris Anagnostidis, Albert Pumarola, Markos Georgopoulos, Artsiom Sanakoyeu, Yuming Du, Edgar Schönfeld, Ali Thabet, and Jonas Kohler. 2025. Judge decoding: Faster speculative sampling requires going beyond model alignment. In *The Thirteenth International Conference on Learning Representations*.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2025. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10280–10300, Vienna, Austria. Association for Computational Linguistics.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

Guanzheng Chen, Qilong Feng, Jinjie Ni, Xin Li, and Michael Qizhe Shieh. 2025. RAPID: Long-context inference with retrieval-augmented speculative decoding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 8093–8107. PMLR.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qiaoling Chen, Zijun Liu, Peng Sun, Shenggui Li, Guoteng Wang, Ziming Liu, Yonggang Wen, Siyuan Feng, and Tianwei Zhang. 2026. Respec: Towards optimizing speculative decoding in reinforcement learning systems. In *Proceedings of Machine Learning and Systems*, volume 8. MLSys.

Keyuan Cheng, Muhammad Asif Ali, Shu Yang, Gang Lin, Yuxuan Zhai, Haoyang Fei, Ke Xu, Lu Yu, Lijie Hu, and Di Wang. 2024a. Leveraging logical rules in knowledge editing: A cherry on the top. *arXiv preprint arXiv:2405.15452*.

Keyuan Cheng, Zijian Kan, Zhuoran Zhang, Muhammad Asif Ali, Lijie Hu, and Di Wang. 2025a. COMPKE: Complex question answering under knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2557–2576, Vienna, Austria. Association for Computational Linguistics.

Keyuan Cheng, Gang Lin, Haoyang Fei, Yuxuan Zhai, Lu Yu, Muhammad Asif Ali, Lijie Hu, and Di Wang. 2024b. Multi-hop question answering under temporal knowledge editing. In *Proceedings of the 1st Conference on Language Modeling*.

Keyuan Cheng, Xudong Shen, Yihao Yang, Tengyue Wang, Yang Cao, Muhammad Asif Ali, Hanbin Wang, Lijie Hu, and Di Wang. 2025b. CODEMENV: Benchmarking large language models on code migration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2719–2744, Vienna, Austria. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11661–11665. IEEE.

Daya Guo et al. 2025a. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zikun Guo, Xinyue Xu, Pei Xiang, Shu Yang, Xin Han, Di Wang, and Lijie Hu. 2025b. Benchmarking and mitigate psychological sycophancy in medical vision-language models. *arXiv e-prints*, pages arXiv–2509.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Jingyu Hu, Shu Yang, Xilin Gong, Hongming Wang, Weiru Liu, and Di Wang. 2025. Monica: Real-time monitoring and calibration of chain-of-thought sycophancy in large reasoning models. *arXiv preprint arXiv:2511.06419*.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, Guoping Hu, and Bing Qin. 2025a. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16896–16913, Vienna, Austria. Association for Computational Linguistics.

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

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025c. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429.

Zongle Huang, Lei Zhu, Zongyuan Zhan, Ting Hu, Weikai Mao, Xianzhi Yu, Yongpan Liu, and Tianyu Zhang. 2025d. MoESD: Unveil speculative decoding's potential for accelerating sparse MoE. In *Advances in Neural Information Processing Systems*, volume 38. Neural Information Processing Systems Foundation.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A human-LLM collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. 2025. CoCoA: Confidence- and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6835–6855, Suzhou, China. Association for Computational Linguistics.

Deren Lei, Yaxi Li, Siyao Li, Mengya Hu, Rui Xu, Ken Archer, Mingyu Wang, Emily Ching, and Alex Deng. 2025. FactCG: Enhancing fact checkers with graph-based multi-hop data. 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 5002–5020, Albuquerque, New Mexico. Association for Computational Linguistics.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR.

Anguo Li and Lei Yu. 2025. Summary factual inconsistency detection based on LLMs enhanced by universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25450–25465, Vienna, Austria. Association for Computational Linguistics.

Guanghao Li, Zhihui Fu, Min Fang, Qibin Zhao, Ming Tang, Chun Yuan, and Jun Wang. 2026a. DiffuSpec: Unlocking diffusion language models for speculative decoding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 20896–20910,

San Diego, California, United States. Association for Computational Linguistics.

Hongji Li, Manjiang Yu, Junchi Yao, Priyanka Singh, Xue Li, Di Wang, and Lijie Hu. 2026b. Towards reasoning-preserving unlearning in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10251–10261.

Yiwei Li, Jiayi Shi, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Yueqi Zhang, Ji Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025. Speculative decoding for multi-sample inference. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12523–12533, Suzhou, China. Association for Computational Linguistics.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28935–28948. PMLR.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

MetaAI. 2024. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. 2024. Specinfer: Accelerating large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pages 932–949, La Jolla, CA, USA. Association for Computing Machinery.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, and 23 others. 2025. Tool learning with foundation models. *ACM Computing Surveys*, 57(4).

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 *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. 2025. CLAPnq: Cohesive long-form answers from passages in natural questions for RAG systems. *Transactions of the Association for Computational Linguistics*, 13:53–72.

Ranajoy Sadhukhan, Jian Chen, Zhuoming Chen, Vashisth Tiwari, Ruihang Lai, Jinyuan Shi, Ian En-Hsu Yen, Avner May, Tianqi Chen, and Beidi Chen. 2025. Magicdec: Breaking the latency-throughput tradeoff for long context generation with speculative decoding. In *The Thirteenth International Conference on Learning Representations*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Hanshi Sun, Zhuoming Chen, Xinyu Yang, Yuandong Tian, and Beidi Chen. 2024. Triforce: Lossless acceleration of long sequence generation with hierarchical speculative decoding. In *Proceedings of the Conference on Language Modeling*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Rishabh Tiwari, Haocheng Xi, Aditya Tomar, Coleman Richard Charles Hooper, Sehoon Kim, Maxwell Horton, Mahyar Najibi, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. 2025. QuantSpec: Self-speculative decoding with hierarchical quantized KV cache. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 59668–59686. PMLR.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

Cheng-Long Wang, Qi Li, Zihang Xiang, Yinzhi Cao, and Di Wang. 2025a. Towards lifecycle unlearning commitment management: Measuring sample-level unlearning completeness. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6481–6500, Seattle, WA. USENIX Association.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge. 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 11636–11652, Albuquerque, New Mexico. Association for Computational Linguistics.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. 2026. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33566–33574.

Liangyu Wang, Huanyi Xie, Xinhai Wang, Tianjin Huang, Mengdi Li, and Di Wang. 2025c. Infinite sampling: Efficient and stable grouped rl training for large language models. *arXiv preprint arXiv:2506.22950*.

Xuekang Wang, Shengyu Zhu, and Xueqi Cheng. 2025d. Speculative safety-aware decoding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12827–12841, Suzhou, China. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Heming Xia, Tao Ge, Peiyi Wang, Si-Qing Chen, Furu Wei, and Zhifang Sui. 2023. Speculative decoding: Exploiting speculative execution for accelerating seq2seq generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3909–3925, Singapore. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Zhichao Xu. 2023. Context-aware decoding reduces hallucination in query-focused summarization. *arXiv preprint arXiv:2312.14335*.

Shi-Qi Yan, Quan Liu, and Zhen-Hua Ling. 2025. RPO: Retrieval preference optimization for robust retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5228–5240, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Penghui Yang, Cunxiao Du, Fengzhuo Zhang, Haonan Wang, Tianyu Pang, Chao Du, and Bo An. 2026. LongSpec: Long-context lossless speculative decoding with efficient drafting and verification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1826–1844, San Diego, California, United States. Association for Computational Linguistics.

Shu Yang, Junchao Wu, Xilin Gong, Xuansheng Wu, Derek Wong, Ninhao Liu, and Di Wang. 2025b. Investigating cot monitorability in large reasoning models. *arXiv preprint arXiv:2511.08525*.

Tiancheng Yang, Lin Zhang, Jiaye Lin, Guimin Hu, Di Wang, and Lijie Hu. 2025c. Tracing and mitigating hallucinations in multimodal llms via dynamic attention localization. *arXiv preprint arXiv:2509.07864*.

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, Brussels, Belgium. Association for Computational Linguistics.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

Zhuoran Zhang, Yongxiang Li, Zijian Kan, Keyuan Cheng, Lijie Hu, and Di Wang. 2025a. Locate-then-edit for multi-hop factual recall under knowledge editing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 75369–75391. PMLR.

Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. 2025b. When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms. *arXiv preprint arXiv:2511.02243*.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

Wenrui Zhou, Mohamed Hendy, Shu Yang, Qingsong Yang, Zikun Guo, Yuyu Luo, Lijie Hu, and Di Wang. 2026. Flattery in motion: Benchmarking and analyzing sycophancy in video-LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8146–8172, San Diego, California, United States. Association for Computational Linguistics.

A Parameter Analysis

A.1 Role of Contextual Plausibility Mask (η)

A pivotal challenge in logit-level steering is the potential for linguistic degradation: a specialist may propose a factually accurate token that is syntactically incompatible with the target model's prefix. To evaluate how our **Contextual Plausibility Mask (CPM)** mitigates this risk, we conduct an ablation study on the threshold η using the XSum dataset.

Table 8: Ablation of the Plausibility Threshold η on XSum. $\eta = 0.1$ is the default setting used in Table 2; $\eta = 0$ denotes steering without the linguistic safety guard.

Configuration	ROUGE-L (↑)	BERT-P (↑)	AlignScore (↑)
SFAD (Standard, $\eta = 0.1$)	<u>16.32</u>	**93.97**	87.37
SFAD w/o CPM ($\eta = 0$)	14.78	91.25	**87.49**
Strict CPM ($\eta = 0.5$)	15.42	92.84	81.12
Vanilla Qwen3-14B	13.67	91.67	72.86

As illustrated in Table 8, removing the mask ($\eta = 0$) yields the highest *AlignScore* (87.49), confirming that the specialist model is indeed capable of forcing factual corrections. However, this comes at a substantial cost to linguistic integrity: **ROUGE-L drops by 1.54 points** and **BERT-P falls by 2.72 points** compared to the standard SFAD. Qualitative analysis reveals that without CPM, the model often injects correct entities with mismatched syntax (e.g., incorrect prepositional usage or broken noun phrases).

By introducing $\eta = 0.1$, SFAD successfully filters out these "linguistically dissonant" corrections. This explains the competitive results in Table 2, where SFAD nearly matches the 70B Frontier Model's performance in both fluency and factuality. The CPM acts as a "safety guard," ensuring that steering only occurs within the target model's plausible manifold, thus achieving a superior Pareto-optimal balance between knowledge correction and natural language generation.

A.2 Sensitivity Analysis of Friction Threshold τ

The friction threshold τ serves as the primary hyperparameter for calibrating the sensitivity of SFAD's hallucination detector. It determines the operating point on the factuality-efficiency Pareto frontier by controlling the ratio between the *Steering Path* and the *Fast Path*. We evaluate the impact of τ on the LLM-AggreFact test set, reporting the Steering Ratio (percentage of tokens where $F_t \geq \tau$), Faithfulness Score (MiniCheck), and inference speedup (ATGA).

Table 9: Sensitivity analysis of τ on LLM-AggreFact. *Steering Ratio* denotes the percentage of generated tokens processed via the logit steering path. Speedup is relative to greedy decoding without speculation.

Threshold τ	Steering Ratio (%)	Faithfulness Score (↑)	Speedup (ATGA) (↑)
$\tau = 0.1$ (Aggressive)	48.2%	**86.7**	2.12×
$\tau = 0.3$	35.6%	85.9	2.31×
$\tau = 0.5$ (Default)	22.4%	<u>85.2</u>	<u>2.48×</u>
$\tau = 0.7$	12.1%	62.4	2.65×
$\tau = 0.9$ (Conservative)	4.5%	41.8	**2.82**×
Standard SD	0.0%	38.5	2.18×

As shown in Table 9, the *Steering Ratio* decreases monotonically as τ increases, reflecting a more selective intervention strategy. When $\tau = 0.1$, the model intervenes frequently, achieving the highest faithfulness but incurring a latency penalty due to frequent target model logit computations. Conversely, at $\tau = 0.9$, the system defaults to standard speculative decoding (Fast Path) most of the time, maximizing speed but failing to resolve knowledge conflicts.

Crucially, our default setting of $\tau = 0.5$ achieves a "sweet spot": it resolves the majority of hallucinations (surpassing standard SD by 46.7 points) while maintaining a high speedup of 2.48×. This demonstrates that SFAD performs **surgical interventions** only when necessary.

A.3 Effectiveness of Asymmetric Logit Fusion Operators

To empirically validate the theoretical advantage of our asymmetric steering law (Theorem 3), we conduct an ablation study comparing different logit fusion operators. In this experiment, we keep the adaptive gating mechanism (λ_t) and the Contextual Plausibility Mask (CPM) constant, only varying how the specialist's logits z_m are integrated with the generalist's logits z_M when steering is triggered. We evaluate four strategies:

(i) **Linear Sum:** $z^* = z_M + \lambda_t z_m$, which directly adds the specialist's signals.

(ii) **Linear Interpolation:** $z^* = (1 - \lambda_t) z_M + \lambda_t z_m$, a standard weighted average used in model blending.

(iii) **Subtractive Contrast:** $z^* = z_M + \lambda_t (z_M - z_{base})$, similar to contrastive decoding but using the specialist as the positive signal.

(iv) **Asymmetric Steering (SFAD):** $z^* = z_M + \lambda_t \text{ReLU}(z_m - z_M)$, our proposed unidirectional injection.

The results on PopQA and HotpotQA are summarized in Table 10.

Table 10: Ablation of Logit Fusion Operators. Faithfulness is measured by Exact Match (EM) for PopQA and MiniCheck for HotpotQA. Fluency is represented by ROUGE-L (R-L).

Fusion Strategy	**PopQA**		**HotpotQA**	
	EM (↑)	**R-L** (↑)	**Faith** (↑)	**R-L** (↑)
Linear Sum	78.42	14.12	46.21	35.89
Linear Interpolation	81.35	15.67	48.55	39.42
Subtractive Contrast	83.12	13.05	49.32	34.11
SFAD (Ours)	**86.39**	**16.32**	**52.19**	**41.34**

As shown in Table 10, the **Asymmetric Steering** operator consistently outperforms other fusion methods. *Linear Sum* and *Interpolation* tend to dilute the generalist's linguistic priors, leading to a drop in ROUGE-L (fluency). Conversely, while *Subtractive Contrast* can highlight factual differences, it often results in "negative constraints" that suppress even valid tokens, harming coherence. SFAD's ReLU-based injection ensures that the specialist only intervenes to *boost* tokens where it has higher confidence than the generalist, preserving the natural language manifold of the target model while effectively correcting factual errors.

B Generalization Across Model Families

We provide more results on Llama-3.1-8B in Table 11 and Table 12.

Table 11: Llama-3.1-8B results on Summarization.

	XSum			**TofuEval**	
Method	**R-L**	**BERT-P**	**AlignScore**	**AlignScore**	**Rel. Latency**
Vanilla Baseline					
Llama-3.1-8B	12.89	90.83	71.24	58.12	1.00×
Decoding-level Baselines (Llama-3.1-8B)					
CAD	13.76	92.87	82.56	81.45	2.00×
AdaCAD	14.08	93.41	84.03	83.29	2.22×
COIECD	12.91	90.22	72.15	59.34	2.48×
SFAD (Ours)	**15.47**	**93.12**	**85.68**	**85.87**	**0.87×**

Table 12: Llama-3.1-8B results on Long-Form QA.

	CLAPNQ		**ExpertQA**		**HAGRID**		
Method	**R-L**	**Faith**	**R-L**	**Faith**	**R-L**	**Faith**	**Rel. Latency**
Vanilla Baseline							
Llama-3.1-8B	16.34	57.89	30.21	49.56	16.12	55.87	1.00×
Decoding-level Baselines (Llama-3.1-8B)							
CAD	17.41	58.43	32.15	51.72	17.03	56.45	2.00×
AdaCAD	17.58	60.51	31.49	52.93	16.31	58.12	2.21×
COIECD	17.72	60.14	32.78	54.48	16.52	57.94	2.45×
SFAD (Ours)	**19.87**	**68.45**	**34.56**	**62.37**	**18.93**	**66.23**	**0.81×**

These cross-family results indicate that SFAD's gains are not tied to a single target backbone. Across summarization and long-form QA, the same decoding mechanism improves faithfulness while preserving the intended latency advantage, suggesting that the learned factuality-aware draft model and friction-triggered steering provide a portable correction signal across model families.

C Examples of Factual Perturbation for ConFide

We present two representative cases of how a faithful response is transformed into a hallucinated negative sample (y_l) through our atomic-level perturbation mechanism.

Case 1: Entity Swap (τ_{ent})

Source Context:
"...Muto graduated from Keio University in Tokyo with an economics degree two weeks ago. He will join Chelsea's partner club Vitesse Arnhem on loan..."

Original Faithful Response (y_w):
"Yoshinori Muto, who recently completed his economics degree at Keio University, is set to join Chelsea."

Atomic Fact Decomposition:
a_j = (Yoshinori Muto, graduated from, **Keio University**)

Perturbation Operation:
Swap target entity *Keio University* with intra-context distractor ***Vitesse Arnhem***.

Generated Hallucinated Response (y_l):
"Yoshinori Muto, who recently completed his economics degree at Vitesse Arnhem, is set to join Chelsea."

Case 2: Relation Inversion (τ_{rel})

Source Context:
"...Ogane claims that Chelsea's interest in Muto is not connected to the £200million sponsorship deal they signed with Yokohama Rubber..."

Original Faithful Response (y_w):
"The club president stated that the move for Muto is independent of the sponsorship deal with Yokohama Rubber."

Atomic Fact Decomposition:

a_j = (Chelsea's interest, **is not connected to**, Sponsorship Deal)

Perturbation Operation:
Apply logical inversion ($\neg r_j$) to the relation *is not connected to*.

Generated Hallucinated Response (y_l):
"The club president stated that the move for Muto is a direct result of the sponsorship deal with Yokohama Rubber."

D Algorithms

This section provides the algorithmic details of the proposed framework. Algorithm 1 summarizes the inference-time procedure of SFAD, including friction-based conflict detection and asymmetric logit steering. Algorithm 2 describes the construction process of the ConFide preference dataset used for training the factuality-aware draft model.

Algorithm 1 SFAD: Speculative Factuality-Aware Decoding

Input: Target model M, DPO-aligned draft model m, prefix context $x_{<t}$, friction threshold τ, sharpening coefficient γ, plausibility threshold η, sigmoid scale β.
Output: Context-faithful generated token x_t.
while $x_t \neq$ EOS **do**
  // Step 1: Distribution Computation
  Compute logit vectors $\mathbf{z}_{M,t}$ from $M(x_{<t})$ and $\mathbf{z}_{m,t}$ from $m(x_{<t})$
  Compute probabilities $\mathbb{P}_M = \text{Softmax}(\mathbf{z}_{M,t})$ and $\mathbb{P}_m = \text{Softmax}(\mathbf{z}_{m,t})$
  // Step 2: Factuality Conflict Detection
  Calculate Specialist Certainty: $\kappa_t = \left(1 - \frac{-\sum_{v\in\mathcal{V}} \mathbb{P}_m(v)\log \mathbb{P}_m(v)}{\log|\mathcal{V}|}\right)^{\gamma}$
  Calculate Epistemic Friction: $\mathcal{F}_t = D_{JS}(\mathbb{P}_M \parallel \mathbb{P}_m) \cdot \kappa_t$
  if $\mathcal{F}_t \geq \tau$ **then**
    // Steering Path: Distribution Correction
    Identify plausible set: $\mathcal{V}_{CPC} = \{x \in \mathcal{V} : \mathbb{P}_m(x|x_{<t}) \geq \eta \cdot \max_{w\in\mathcal{V}} \mathbb{P}_m(w|x_{<t})\}$
    Compute gating scalar: $\lambda_t = \frac{1}{1+\exp(-\beta(\mathcal{F}_t-\tau))}$
    Apply Asymmetric Logit Steering:
    $\mathbf{z}_t^* = \mathbf{z}_{M,t} + \lambda_t \cdot \max(0, \mathbf{z}_{m,t} - \mathbf{z}_{M,t}) \cdot \mathbb{I}(x \in \mathcal{V}_{CPC})$
    Sample next token: $x_t \sim \text{Softmax}(\mathbf{z}_t^*)$
  else
    // Fast Path: Standard Speculative Decoding
    Sample draft token $\tilde{x}_t$ from $\mathbb{P}_m$
    $x_t \leftarrow \text{Verify}(\tilde{x}_t, \mathbb{P}_M)$ {Standard rejection sampling}
  end if
  Update sequence: $x_{<t+1} \leftarrow [x_{<t}; x_t]$
end while

Algorithm 2 ConFide: Factuality-Aware Preference Data Construction

Require: $\mathcal{D} = \{(x, y_{ref}, l)\}_{i=1}^{N}$, Teacher $\mathcal{M}_T$, Paraphraser π_{para}, Generator f_{rec}
$\mathcal{P} \leftarrow \emptyset$
for each $(x, y_{ref}, l) \in \mathcal{D}$ **do**
  $\mathcal{A} = \{a_1, \ldots, a_m\} \leftarrow f_{dec}(y_{ref})$ {Atomic Fact Decomposition}
  // Negative Generation: Perturb $a_j \in \mathcal{A}$ via Entity Swap, Numerical, or Relation Inversion
  $a_j' \leftarrow \phi(a_j)$, $y_l \leftarrow f_{rec}(\mathcal{A} \setminus \{a_j\} \cup \{a_j'\})$ {Corrupt and Reconstruct}
  // Positive Refinement: Paraphrase faithful samples or utilize Teacher for corrections
  $y_w \leftarrow (l = 1)\ ?\ \pi_{para}(y_{ref}) : \mathcal{M}_T(\text{correct } y_{ref} \text{ based on } x)$
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{(x, y_w, y_l)\}$ {Assemble Preference Pair}
end for
return $\mathcal{P}$

E Theoretical Analysis of SFAD

In this section, we provide a rigorous formal justification for the *Speculative Factuality-Aware Decoding* (SFAD) framework. Our analysis bridges the gap between empirical performance and distributional theory by focusing on three core pillars: (1) **Semantic Integrity**, establishing that our steering mechanism preserves the linguistic manifold; (2) **Factuality Amplification**, proving how DPO-induced margins yield exponential gains in faithful tokens; and (3) **Dynamic Stability**, demonstrating the robustness of asymmetric steering against the "zero-probability trap" prevalent in contrastive methods.

E.1 Semantic Consistency and the Linguistic Manifold Bound

A fundamental challenge in logit steering is the *alignment-fidelity trade-off*: factual corrections must not drive the model's output distribution off the natural language manifold $\mathcal{M}$. We formalize the role of the Contextual Plausibility Mask (CPM) as a projection operator that constrains the steering signal to the support of the generalist's distribution.

Theorem 1 (Manifold Projection Bound)**.** *Let $\mathcal{V}_{CPC} = \{v \in \mathcal{V} : P_m(v|x_{<t}) \geq \eta \cdot \max_w P_m(w|x_{<t})\}$ be the set of plausible candidates. For any steering intensity $\lambda_t \in [0, 1]$, the Total Variation (TV) distance between the steered distribution P^* and the original generalist distribution P_M is bounded by the probability mass concentrated on the plausible manifold:*

$$d_{TV}(P^*, P_M) \leq \frac{1}{2} \sum_{v \in \mathcal{V}_{CPC}} P_M(v) \left| \frac{e^{\lambda_t \Delta z_v}}{\mathcal{Z}_t^*} - 1 \right| + \frac{1}{2} \left| 1 - \frac{1}{\mathcal{Z}_t^*} \right| (1 - P_M(\mathcal{V}_{CPC})). \tag{10}$$

where $\Delta z_v = (z_{m,v} - z_{M,v})_+$ and $\mathcal{Z}_t^* = \mathbb{E}_{v \sim P_M}[e^{\lambda_t \Delta z_v}]$ is the partition normalization factor.

Proof. By the definition of the CPM-based steering in Eq. (7), for any token $v \notin \mathcal{V}_{CPC}$, the steering signal Δz_v is nullified by the indicator function $\mathbb{I}(x \in \mathcal{V}_{CPC})$. Consequently, the steered logit z_v^* equals the original logit $z_{M,v}$, and the probability simplifies to $P^*(v) = P_M(v)/\mathcal{Z}_t^*$.

The TV distance is defined as $d_{TV}(P^*, P_M) = \frac{1}{2} \sum_{v \in \mathcal{V}} |P^*(v) - P_M(v)|$. Partitioning the vocabulary $\mathcal{V}$ into the plausible set $\mathcal{V}_{CPC}$ and its complement $\mathcal{V}^c$, we obtain:

$$\begin{aligned} 2d_{TV} &= \sum_{v \in \mathcal{V}_{CPC}} \left| \frac{P_M(v) e^{\lambda_t \Delta z_v}}{\mathcal{Z}_t^*} - P_M(v) \right| \\ &\quad + \sum_{v \in \mathcal{V}^c} \left| \frac{P_M(v)}{\mathcal{Z}_t^*} - P_M(v) \right| \\ &= \sum_{v \in \mathcal{V}_{CPC}} P_M(v) \left| \frac{e^{\lambda_t \Delta z_v}}{\mathcal{Z}_t^*} - 1 \right| \\ &\quad + \left| \frac{1}{\mathcal{Z}_t^*} - 1 \right| \sum_{v \in \mathcal{V}^c} P_M(v). \end{aligned} \tag{11}$$

Substituting $P_M(\mathcal{V}^c) = 1 - P_M(\mathcal{V}_{CPC})$ completes the proof. This bound illustrates that the divergence is strictly governed by the specialist's corrective signal magnitude within the linguistic manifold. Since m and M share a common pre-training ancestry, their high-density regions in $\mathcal{M}$ are naturally aligned, ensuring that steering only re-allocates mass within valid semantic clusters. □

Remark 1. **The Anchor Property.** Theorem 1 reveals that SFAD acts as a "factual re-ranker" rather than a "stochastic generator." By anchoring the correction to the generalist's manifold $\mathcal{M}$, the framework prevents out-of-distribution (OOD) artifacts. Unlike additive noise or unconstrained logit shifts, the CPM ensures that the model never samples tokens that are linguistically nonsensical, explaining the high fluency scores observed in our experiments.

E.2 Factuality Amplification via DPO-Induced Margins

While the previous section established safety, we now prove that SFAD effectively resolves knowledge conflicts by leveraging the preference gap instilled during the Direct Preference Optimization (DPO) phase.

Proposition 1 (Exponential Posterior Gain)**.** *Let v_{ctxt} be a faithful token and v_{mem} be a hallucinated token (favored by the target model's parametric prior). If the specialist m satisfies the DPO optimality condition with a learned margin $\gamma_m = z_m(v_{ctxt}) - z_m(v_{mem})$, the posterior odds ratio under SFAD steering satisfies:*

$$\frac{P^*(v_{ctxt})}{P^*(v_{mem})} = \frac{P_M(v_{ctxt})}{P_M(v_{mem})} \cdot \exp(\lambda_t \cdot \gamma_{eff}) \tag{12}$$

where $\gamma_{eff} = \max(0, z_{m,v_{ctxt}} - z_{M,v_{ctxt}})$ is the effective corrective margin.

Proof. Based on the Bradley-Terry model utilized in DPO, the specialist m is trained to maximize the log-odds of contextually loyal pairs. Following the asymmetric steering law $z_t^*(v) = z_{M,t}(v) + \lambda_t \Delta z_v$, we examine the log-odds ratio:

$$\log \frac{P^*(v_{ctxt})}{P^*(v_{mem})} = \big(z_{M,v_{ctxt}} + \lambda_t \Delta z_{v_{ctxt}}\big) - \big(z_{M,v_{mem}} + \lambda_t \Delta z_{v_{mem}}\big). \tag{13}$$

In a factuality conflict, v_{mem} is the "surface pattern" favored by M, thus $z_{m,v_{mem}} \leq z_{M,v_{mem}}$, which implies $\Delta z_{v_{mem}} = 0$ due to the ReLU activation. Conversely, for the faithful token v_{ctxt}, the specialist (having been optimized via DPO) yields $z_{m,v_{ctxt}} > z_{M,v_{ctxt}}$. Thus:

$$\log \frac{P^*(v_{ctxt})}{P^*(v_{mem})} = \log \frac{P_M(v_{ctxt})}{P_M(v_{mem})} + \lambda_t \big(z_{m,v_{ctxt}} - z_{M,v_{ctxt}}\big). \tag{14}$$

Exponentiating both sides demonstrates that the faithful candidate's probability mass is amplified exponentially relative to the hallucination, controlled by the steering intensity λ_t. □

Remark 2. **Selective Pressure.** This result highlights the advantage of SFAD over standard speculative decoding. While the latter performs binary rejection (0 or 1), SFAD applies continuous *selective pressure*. Even if the generalist M is initially biased toward a hallucination, a sufficiently certain specialist can shift the distribution's mode toward the truth without requiring a complete rejection of the sequence.

E.3 Information-Theoretic Validity of Epistemic Friction

We now justify the use of Epistemic Friction F_t as a dynamic trigger. It must distinguish between *benign diversity* and *factual conflict*.

Lemma 1 (Friction as a High-Precision Trigger)**.** *The Epistemic Friction $F_t = \mathcal{D}_{JS}(P_M \| P_m) \cdot \kappa_t$ is a lower bound on the expected reduction in epistemic uncertainty when the generalist conditions its output on the specialist's contextual evidence.*

Proof. The Jensen-Shannon divergence $\mathcal{D}_{JS}$ is a symmetric and bounded metric of distributional tension. In the context of LLMs, high divergence can arise from two sources: (i) factual disagreement or (ii) high entropy (uncertainty) in one model. Standard entropy-based triggers fail because they cannot distinguish between these cases. By weighting $\mathcal{D}_{JS}$ with the specialist certainty $\kappa_t = (1 - H(P_m)/\log|\mathcal{V}|)^\gamma$, F_t acts as a filter. As $\kappa_t \to 1$ (certainty), $F_t \to \mathcal{D}_{JS}$. As $\kappa_t \to 0$ (uncertainty), $F_t \to 0$, suppressing the trigger. Thus, F_t only activates steering when the specialist's corrective signal is both *strong* and *high-confidence*, minimizing noise injection. □

E.4 Stability and Support Preservation

A critical failure in subtractive contrastive methods (e.g., CAD) is the "Zero-Probability Trap," where valid tokens are suppressed to numerical underflow.

Theorem 2 (Numerical Stability and Support Preservation)**.** *For any finite steering scale $\lambda_t \in [0, \infty)$ and any token $v \in \mathcal{V}$, the SFAD framework satisfies the support preservation property: if $P_M(v) > 0$, then $P^*(v) > 0$.*

Proof. The steered logit is $z_v^* = z_{M,v} + \lambda_t \cdot \mathrm{ReLU}(z_{m,v} - z_{M,v})$. Since $\mathrm{ReLU}(\cdot) \geq 0$ and $\lambda_t \geq 0$, it holds that $z_v^* \geq z_{M,v}$ for all v. The steered probability is $P^*(v) = \exp(z_v^*)/\sum_w \exp(z_w^*)$. Since the exponential function is strictly positive, and z_v^* is bounded from below by the original logit $z_{M,v}$, the numerator remains positive. Unlike subtractive methods where $z^* = z_M - \alpha z_{neg}$ can lead to $-\infty$ and numerical collapse, our asymmetric law ensures that no token is ever strictly "killed." They are only relatively de-emphasized by the factual amplification of superior candidates. □

Remark 3. **Robustness by Design.** This theorem explains why SFAD is robust to the hyperparameter λ_t. The additive, ReLU-based nature of the steering ensures that even with high steering intensity, the model maintains its linguistic foundation, avoiding the "repetitive gibberish" or "empty strings" common in negative-constraint decoding.

E.5 Mode-Switching Optimality under Knowledge Conflicts

Finally, we demonstrate why asymmetric steering is superior to linear logit interpolation (averaging).

Theorem 3 (Optimality of Asymmetric Mode-Switching)**.** *In a bimodal conflict between a factual token v_f and a hallucinated token v_h, the asymmetric law $z^* = z_M + \lambda_t(z_m - z_M)_+$ is a rank-preserving transformation for all neutral tokens v_n where $z_{m,v_n} \leq z_{M,v_n}$.*

Proof. Consider two neutral tokens v_1, v_2 where the generalist and specialist are in relative agreement (i.e., $z_m \leq z_M$ for both). Under SFAD, $\Delta z_{v_1} = \Delta z_{v_2} = 0$. Therefore, their steered logits remain $z^*_{v_1} = z_{M,v_1}$ and $z^*_{v_2} = z_{M,v_2}$. Their relative rank $z^*_{v_1} - z^*_{v_2} = z_{M,v_1} - z_{M,v_2}$ is perfectly preserved. In contrast, a linear interpolation $z_{interp} = (1-\alpha)z_M + \alpha z_m$ would modify both logits, potentially flipping their rank if the specialist has slight fluctuations in its low-probability tail. SFAD thus provides a "surgical" correction: it only perturbs the distribution where the specialist has a *strictly better* (more factual) proposal. □

Remark 4. **Theoretical Scalability.** This result suggests that as the Specialist model m improves in factuality (e.g., through more rigorous DPO), SFAD's performance will scale without requiring a corresponding increase in the Generalist's capacity, making it a sustainable architecture for deploying large-scale factual models.

E.6 Optimal Risk-Aware Switching and the Factuality-Efficiency Frontier

A distinctive feature of SFAD is the hybrid policy (Eq. (8)) that dynamically chooses between the *Steering Path* and the *Fast Path*. We now provide a decision-theoretic justification for this switching logic, demonstrating that the Epistemic Friction F_t serves as an optimal proxy for balancing factual integrity and computational latency.

Theorem 4 (Factuality-Risk Minimization)**.** *Let $\mathcal{R}(\pi)$ be the expected risk of generating a hallucinated token under policy π, and let $\mathcal{C}(\pi)$ be the computational cost (latency). Define the factual risk r_t at step t as the probability that the draft token $\tilde{x}_t$ deviates from the specialist's high-confidence manifold. The SFAD hybrid policy π_{SFAD} is a solution to the constrained optimization problem:*

$$\min_{\pi} \mathbb{E}[\mathcal{R}(\pi)] \quad \textit{subject to} \quad \mathbb{E}[\mathcal{C}(\pi)] \leq \mathcal{B} \tag{15}$$

where the switching threshold τ acts as the Lagrange multiplier β^{-1} that determines the operating point on the Factuality-Efficiency Pareto frontier.

Proof. Consider a binary decision at each step t: either accept the standard speculative verification (Fast Path, $a_t = 0$) or intervene with logit steering (Steering Path, $a_t = 1$). The total risk is:

$$\mathbb{E}[\mathcal{R}] = \sum_t \left[P(a_t = 0) \cdot r_t + P(a_t = 1) \cdot \epsilon\right] \tag{16}$$

where ϵ is the residual risk after steering (which is minimal per Prop. 1). The computational cost is $\mathcal{C}_{fast}$ for the Fast Path and $\mathcal{C}_{steer}$ for the Steering Path ($\mathcal{C}_{steer} > \mathcal{C}_{fast}$).

By the Neyman-Pearson Lemma, the optimal decision rule a_t that minimizes risk for a fixed cost is a likelihood-ratio test. In SFAD, the Epistemic Friction F_t quantifies the "hallucination likelihood" by measuring the distributional tension weighted by specialist certainty. When $F_t \geq \tau$, the potential risk reduction $\Delta r_t = r_t - \epsilon$ outweighs the marginal cost $\Delta\mathcal{C}$, triggering the Steering Path. Thus, the switching threshold τ effectively calibrates the sensitivity of the detector to maximize the *Factuality Gain per Unit of Latency*. □

Remark 5. **The Latency-Accuracy Frontier.** Theorem 4 formalizes why SFAD maintains the speed of speculative decoding while approaching the factuality of much larger, slow-inference models. By only invoking the "Steering Path" when F_t signals high epistemic conflict, SFAD avoids the redundant computation of constant logit correction, allowing the system to stay on the optimal Pareto frontier.

E.7 Summary of Theoretical Guarantees

Synthesizing the above proofs, we establish that SFAD is not merely an empirical heuristic but a mathematically grounded framework for verifiable generation.

- **Theorem 1** guarantees that our intervention is linguistically safe (Semantic Integrity).
- **Proposition 1** ensures that we effectively resolve conflicts in favor of the truth (Factuality Amplification).
- **Theorem 2** protects against numerical instability and the suppression of valid tokens (Support Preservation).
- **Theorem 4** justifies the adaptive switching mechanism for real-world deployment (Efficiency Optimality).

Collectively, these theoretical foundations formalize SFAD as a principled safeguarding framework, providing rigorous guarantees for factual integrity through an inherent 'safe-by-design' decoding paradigm.