GROUNDED WITHOUT DRIFTING FROM PRETRAINING: ROLLOUT-FREE TARGET OPTIMIZATION WITH IMPLICIT DISTILLATION FOR RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training for retrieval-augmented generation (RAG) faces two coupled challenges: human-written supervision can be out-of-distribution with respect to the pretrained model's response distribution, yielding a supervision mismatch that induces undesirable distributional drift, while the learned model must remain faithful to the retrieved context. We propose TIDE (**T**arget Optimization with **I**mplicit **D**istillation for Cont**E**xt-Grounded RAG), a principled and efficient rollout-free approach that retains human supervision while simultaneously encouraging pretrained-distribution compatibility and faithfulness. We first characterize the analytical form of a desired target LLM for RAG by optimizing a novel value-based objective that jointly captures length-normalized likelihood under the pretrained model and faithfulness to retrieved evidence. While the target LLM is intractable due to the unknown human-supervision distribution, we introduce implicit target-LLM distillation, which distills the analytically defined target without explicitly constructing or sampling from it and instead enables training directly on the observed human-written data. This distinguishes our approach from standard teacher–student distillation. Our theoretical analysis characterizes how the optimized target shifts probability mass away from low-compatibility and unfaithful responses and provides a rejection-sampling interpretation of TIDE. Unlike approaches that mitigate supervision mismatch through model-generated demonstrations or constructed preference pairs, TIDE requires no response rollouts or generate-and-filter supervision, reducing the cost of constructing training targets. Across a broad range of RAG benchmarks, TIDE achieves the highest average answer-quality F1, outperforming the strongest trained baseline by 23.2% with Qwen3-4B and 21.7% with Qwen3-14B, while also achieving strong faithfulness to the retrieved context.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) has become an important approach in improving the reliability of language models on knowledge-intensive tasks by grounding generation in externally retrieved evidence (Lewis et al., 2020), allowing the model to incorporate information that is more current or specific to the query. Although RAG substantially improves knowledge-intensive generation, simply providing retrieved evidence does not ensure that a pretrained language model can use it effectively (Shi et al., 2024a; Wu et al., 2024; Zhang et al., 2025). This has motivated retrieval-aware post-training methods that explicitly teach language models how to incorporate external evidence. For example, RA-DIT jointly adapts the language model and retriever for retrieval-conditioned generation (Lin et al., 2024) and RAFT trains models to identify and use relevant evidence in the presence of distractors (Zhang et al., 2024). These studies demonstrate the importance of post-training for improving how language models use retrieved evidence.

Despite this progress, RAG post-training faces two important challenges. The first is distribution shift from the pretrained model. Human-written responses provide valuable supervision but may become out-of-distribution under the pretrained response distribution. Directly fitting such responses can therefore induce unnecessary distributional shift and degrade downstream performance (Yang et al., 2024; Kang et al., 2024). Addressing this challenge is difficult because post-training must learn

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from high-quality human responses while preserving behavior already acquired during pretraining. Self-Generated Demonstration training mitigates this mismatch by replacing externally written responses with filtered responses sampled from the pretrained model (Finlayson et al., 2025), but requires repeated generation and filtering and may discard useful human supervision.

The second challenge is faithfulness to retrieved evidence. Even after retrieval-aware post-training, a language model may solely rely on its parametric knowledge, or produce claims unsupported by the retrieved evidence. SELF-RAG solves this issue by training the model to retrieve, generate, and critique its own responses (Asai et al., 2024). SSFO constructs preference pairs from responses generated with and without retrieved context and applies Direct Preference Optimization to favor context-grounded generation (Rafailov et al., 2023; Tang et al., 2025). Nevertheless, this approach also relies on generating additional responses for constructing its training supervision. These approaches improve grounding through additional model-generated signals or constructed preference supervision, whereas we focus on a complementary setting with fixed human-written RAG responses and no additional response generation, where jointly promoting pretrained-distribution compatibility and faithfulness has not been directly addressed by existing approaches.

To address these challenges, we propose TIDE (**T**arget Optimization with **I**mplicit **D**istillation for Cont**E**xt-Grounded RAG), a simple yet effective approach for faithful and in-distribution RAG post-training in the accurate-retrieval setting, which must learn from valuable human responses even when they are not well aligned with the pretrained model's response distribution, while ensuring the faithfulness of the resulting model to the retrieved evidence. TIDE addresses these challenges by constructing a novel analytically characterized target LLM that retains human supervision while favoring responses both compatible with the pretrained model and supported by the retrieved context. Because this desired LLM depends on the unknown human-response distribution, we introduce an implicit distillation approach that enables the model to learn from it directly using the observed training data. Consequently, TIDE requires neither additional response generation nor generate-and-filter supervision. Our theoretical analysis characterizes the resulting target LLM and its connection to rejection sampling, while experiments with Qwen3-4B and Qwen3-14B across five knowledge-intensive RAG benchmarks demonstrate consistent improvements in answer quality and faithfulness. We summarize our main contributions as follows:

- We propose TIDE, a rollout-free RAG post-training framework that retains fixed human-written supervision while promoting both pretrained-distribution compatibility and faithfulness to retrieved context.
- We derive a tractable value–based likelihood objective via implicit distillation of the optimized target model, enabling direct training on observed human responses without modeling the unknown response distribution.
- Theoretically, we characterize the optimized target distribution and show how it shifts probability mass away from low-compatibility and unfaithful responses, with a rejection-sampling interpretation and conditional guarantees for the learned model.
- We conduct extensive experiments across five RAG benchmarks and two model scales, including ablation and distribution-shift analyses. TIDE achieves the highest overall average answer-quality F1 among the evaluated methods, improving over the strongest trained baseline by 23.2% and 21.7% with Qwen3-4B and Qwen3-14B, respectively, while maintaining strong faithfulness.

2 PRELIMINARIES

In this section, we first present some background for post-training of RAG, and then define the two major issues in the post-training under our setup: preserving compatibility with the pretrained response distribution while ensuring faithfulness to retrieved evidence. These challenges motivate the LLM-optimization framework developed in the next section.

2.1 PROBLEM SETUP

Let $\mathcal{V}$ denote a vocabulary, and $\pi(\cdot \mid X)$ denote a candidate response LLM. A large language model parameterized by θ induces the LLM $\pi_\theta(\cdot \mid X)$. In RAG, a query q is paired with a retrieved context

$c = (d_1, \ldots, d_K)$ containing K retrieved documents, and we denote the complete RAG input by $X = (q, c)$. Given X, the model defines a conditional distribution over a response $y = (y_1, \ldots, y_{T_y})$ as $\pi_\theta(y \mid X) = \prod_{t=1}^{T_y} \pi_\theta(y_t \mid X, y_{<t})$, where T_y is the number of response tokens.

RAG post-training adapts a pretrained LLM using human-supervised responses conditioned on queries and the retrieved contexts. To formalize this, let $\mathcal{D} = \{(X_i, g_i)\}_{i=1}^N$ denote the human-supervised RAG training set with $g_i \sim \pi_0(\cdot \mid X_i)$, where $\pi_0(\cdot \mid X)$ is the unknown ideal human-response distribution and $\mathcal{D}_X$ is the empirical distribution over the training inputs. In this work, RAG post-training optimizes a pretrained LLM model $\pi_{\theta_0}(\cdot \mid X)$ initialized at θ_0 using the observed prompt–response pairs $(X_i, g_i) \in \mathcal{D}$, producing the final post-trained LLM $\pi_{\hat{\theta}}^*(\cdot \mid X)$.

2.2 Distribution Mismatch in RAG Post-Training

In RAG post-training, distribution mismatch denotes the discrepancy between the human-response distribution π_0 and the pretrained LLM π_{θ_0}. To formally characterize this issue and motivate our approach, we characterize the compatibility of a response y with the pretrained LLM π_{θ_0} using the length-normalized log-likelihood $\ell_{\theta_0}(X, y) = T_y^{-1} \sum_{t=1}^{T_y} \log \pi_{\theta_0}(y_t \mid X, y_{<t})$ and the corresponding score $s_{\theta_0}(X, y) = \exp\{\ell_{\theta_0}(X, y)\} = [\pi_{\theta_0}(y \mid X)]^{1/T_y}$. Length normalization prevents longer responses from receiving mechanically smaller compatibility scores. The out-of-distribution issue arises because the pretrained LLM may assign a small compatibility score $s_{\theta_0}(X_i, g_i)$ to a human-written response $g_i \sim \pi_0(\cdot \mid X_i)$, even when that response provides useful supervision. Directly fitting such responses can move the adapted LLM unnecessarily far from its pretrained behavior.

Finlayson et al. (2025) addresses this issue by replacing each g_i with a response sampled from π_{θ_0} and judged correct with respect to g_i , thereby producing training targets naturally compatible with the pretrained model. However, this requires repeatedly generating, evaluating, and filtering candidate responses; depends on the pretrained model producing suitable candidates; and may discard useful human supervision. We therefore develop a method that retains the human responses by constraining the optimized LLM toward π_0 through $D_{\mathrm{KL}}\left(\pi(\cdot \mid X) | \pi_0(\cdot \mid X)\right)$, and assigns greater importance to responses with high compatibility scores $s_{\theta_0}(X_i, g_i)$.

2.3 Faithfulness to Retrieved Evidence

The faithfulness issue is a longstanding challenge in language modeling, where a model's output may contain claims that are not supported by the provided source information (Wu et al., 2024; Zhang et al., 2025). In RAG post-training, this issue refers to the learned LLM generating responses that ignore or contradict the retrieved evidence, or rely on unsupported parametric knowledge, even when relevant evidence is provided in the input. To formally characterize this issue in our setting, we consider accurate retrieval, where the retrieved context provides reliable evidence for answering the query, and define $F(X, y) \in [0, 1]$ as the faithfulness score measuring the extent to which response y is supported by the retrieved evidence contained in X, with larger values indicating stronger support.

Pretrained-model compatibility does not guarantee faithfulness: a response may have large $s_{\theta_0}(X, y)$ while ignoring the retrieved context or relying on unsupported parametric knowledge (Shi et al., 2024a). RAG post-training must therefore retain human supervision while jointly accounting for pretrained-model compatibility and context-grounded faithfulness. Section 3 combines these criteria in the TIDE value function and constructs the corresponding desired LLM.

3 Methodology

In this section, we introduce a novel distribution-aware framework. The developed framework involves two procedures, target optimization for characterizing the desired LLM and implicit distillation for solving the intractable issues for learning such LLM using empirical data.

3.1 In-Distribution and Faithful Target Optimization LLM Formulation

Our goal is to obtain a post-trained LLM that is both compatible with the pretrained response distribution and faithful to the retrieved evidence. Instead of first specifying a training algorithm and then

analyzing the LLM it produces, we begin by characterizing the desired target LLM itself. Specifically, the target should retain the useful supervision encoded by the human-response distribution, remain compatible with the pretrained LLM, and assign greater probability to responses that are well supported by the retrieved context. We formalize these requirements through the following value-aware, KL-regularized optimization problem:

$$\max_{\pi} \ \mathbb{E}_{X\sim\mathcal{D}_X}\left[\alpha\mathbb{E}_{y\sim\pi(\cdot|X)}[\log s_{\theta_0}(X,y)+F(X,y)]-D_{\mathrm{KL}}(\pi(\cdot\mid X)\,\|\,\pi_0(\cdot\mid X))\right], \tag{1}$$

The KL term in the objective anchors the desired LLM to the human-response distribution π_0, while $\log s_{\theta_0}(X,y)+F(X,y)$ favors responses compatible with the pretrained model and faithful to the retrieved evidence. The parameter $\alpha>0$ controls this trade-off, with larger values emphasizing pretrained-model compatibility and faithfulness more than KL regularization. Interestingly, we find that the optimization problem induced by Equation (1) has a closed-form solution which can be characterized via validating the KKT conditions. We present this solution in the following theorem.

Theorem 3.1 (Optimal TIDE LLM). *Let $\alpha>0$ and $0<\mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[e^{\alpha(\log s_{\theta_0}(X,y)+F(X,y))}\right]<\infty$ for all X. Then the* TIDE *objective in Equation (1) admits the unique optimizer*

$$\pi^*_\alpha(y\mid X)=\frac{\pi_0(y\mid X)e^{\alpha(\log s_{\theta_0}(X,y)+F(X,y))}}{\mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[e^{\alpha(\log s_{\theta_0}(X,y)+F(X,y))}\right]}. \tag{2}$$

Thus, by anchoring the optimized LLM to the human-response distribution while exponentially favoring responses that are both compatible with the pretrained model and faithful to the retrieved evidence, Theorem 3.1 provides a closed-form analytical characterization of the desired TIDE target LLM. Consequently, π^*_α shifts probability mass away from low-compatibility and unfaithful responses. An equivalent auxiliary joint-distribution formulation induces the same conditional target LLM and will be used in the next subsection to formulate implicit distillation.

3.2 Implicit Distillation of the Optimized LLM

Although the desired target LLM π^*_α provides an analytical characterization of the desired LLM, it depends on the unknown human-response distribution π_0 and therefore cannot be explicitly constructed. This raises the central computational challenge of transferring the optimized target into a parameterized language model without explicitly estimating π_0 or sampling replacement responses. To address this challenge, we propose a novel implicit-distillation approach that transfers the optimized target LLM π^*_α into π_θ, initialized at θ_0, through forward-KL LLM distillation.

To simplify the derivation, we define the response-level value function as $V(X,y)=\log s_{\theta_0}(X,y)+F(X,y)$, which combines compatibility with the pretrained model and faithfulness to the retrieved evidence. The conditional optimizer in Theorem 3.1 admits an equivalent auxiliary joint-distribution formulation that exponentially reweights the original human-response joint distribution $\mathcal{D}_X(X)\pi_0(y\mid X)$ according to $V(X,y)$. Conditioning this auxiliary joint distribution on X yields the optimized LLM $\pi^*_\alpha(y\mid X)$ in Equation (2). The same auxiliary joint distribution also induces an input marginal; marginalizing it over y gives

$$\widetilde{q}_{\alpha,X}(X)=\frac{\mathcal{D}_X(X)\mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[e^{\alpha V(X,y)}\right]}{\mathbb{E}_{X'\sim\mathcal{D}_X}\mathbb{E}_{y'\sim\pi_0(\cdot|X')}\left[e^{\alpha V(X',y')}\right]}$$

which assigns greater probability to inputs for which the human-response distribution places more mass on high-value responses.Thus, $\pi^*_\alpha(\cdot\mid X)$ and $\widetilde{q}_{\alpha,X}$ are, respectively, the conditional response distribution and the input marginal induced by the same optimized auxiliary joint distribution. We use this induced marginal to average the conditional LLM-distillation objective. It is needed only for the population-level derivation and does not need to be estimated or sampled from during training. Accordingly, we distill the optimized target LLM into the parameterized LLM π_θ by minimizing:

$$\mathcal{L}_{\mathrm{distill}}(\theta)=\mathbb{E}_{X\sim\widetilde{q}_{\alpha,X}}\left[D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot\mid X)\,\|\,\pi_\theta(\cdot\mid X)\right)\right]. \tag{3}$$

At first sight, Equation (3) appears intractable because the target LLM depends on the unknown human-response distribution π_0. Our following description shows that, under an appropriate construction, this LLM-distillation principle admits a tractable learning objective over the observed

human responses. Expanding the sequence-level forward-KL objective and collecting all terms independent of θ into C_α gives

$$\begin{aligned}\mathcal{L}_{\text{distill}}(\theta) &= C_\alpha - \mathbb{E}_{X\sim\widetilde{q}_{\alpha,X}}\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[\log \pi_\theta(y \mid X)\right] \\ &= C_\alpha - \mathbb{E}_{X\sim\mathcal{D}_X}\left[\frac{\mathbb{E}_{y'\sim\pi_0(\cdot|X)}\left[e^{\alpha V(X,y')}\right]}{\mathbb{E}_{X'\sim\mathcal{D}_X}\mathbb{E}_{y'\sim\pi_0(\cdot|X')}\left[e^{\alpha V(X',y')}\right]}\right. \\ &\qquad\left.\times\,\mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[\frac{e^{\alpha V(X,y)}}{\mathbb{E}_{y'\sim\pi_0(\cdot|X)}\left[e^{\alpha V(X,y')}\right]}\log\pi_\theta(y\mid X)\right]\right] \\ &= C_\alpha - \frac{\mathbb{E}_{\substack{X\sim\mathcal{D}_X\\ y\sim\pi_0(\cdot|X)}}\left[e^{\alpha V(X,y)}\log\pi_\theta(y\mid X)\right]}{\mathbb{E}_{X'\sim\mathcal{D}_X}\mathbb{E}_{y'\sim\pi_0(\cdot|X')}\left[e^{\alpha V(X',y')}\right]}.\end{aligned}$$

In the second equality, substituting $\widetilde{q}_{\alpha,X}$ and π^*_α reveals the prompt-specific factor $\mathbb{E}_{y'\sim\pi_0(\cdot|X)}[e^{\alpha V(X,y')}]$ in the numerator and denominator respectively, which therefore cancels. This removes the unknown prompt-specific normalizer and leaves an expectation over $X \sim \mathcal{D}_X$ and $y \sim \pi_0(\cdot \mid X)$. The remaining denominator and C_α are independent of θ, so neither affects the optimizer. Replacing the remaining population expectation with its empirical average over the observed training pairs $\mathcal{D} = \{(X_i, g_i)\}_{i=1}^N$ yields the tractable training objective

$$\mathcal{L}_{\text{TIDE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} e^{\alpha V(X_i,g_i)}\log\pi_\theta\left(g_i \mid X_i\right). \tag{4}$$

Thus, we show how the desired target LLM becomes tractable despite its dependence on the unknown human-response distribution. At the population level, the forward-KL distillation objective can be rewritten as a value-weighted expectation over human responses. Replacing this expectation with the observed training examples yields the data-dependent objective in Equation (4), whose optimization produces the final trained LLM $\pi^*_{\widehat{\theta}}$. The key novelty is that $\pi^*_{\widehat{\theta}}$ can approximate the analytically optimized target π^*_α without explicitly estimating the unknown human-response distribution or constructing samples from π^*_α while being in the support of π_{θ_0}. It preserves the human-written supervision and avoids generating, evaluating, and filtering replacement responses.

This procedure differs from standard teacher–student distillation. Standard distillation assumes access to a teacher model whose output probabilities, logits, or generated responses can be directly queried. In contrast, the TIDE target π^*_α depends on the unknown human-response distribution π_0 and can neither be queried nor sampled from directly. TIDE therefore performs distillation implicitly: the preceding reformulation converts the forward-KL objective into the tractable objective in Equation (4), which is optimized directly using the observed human-supervised data. Optimizing this objective produces the final post-trained LLM $\pi^*_{\widehat{\theta}}$ without explicitly constructing the target LLM or performing a generate-and-filter procedure.

4 Theoretical Analysis

In this section, we characterize theoretical properties of TIDE. In particular, we first establish suppression of out-of-distribution and unfaithful responses and support preservation. In second part, we establish an interesting connection of TIDE with rejection-sampling and compare with existing baseline methodologies under this framework. Complete proofs are provided in Appendix B.

4.1 Properties of the Optimized Target and Learned LLM

Theorem 4.1 (Bad-response suppression for the final learned LLM)**.** *Let $\alpha > 0$, and let ν be any distribution over RAG inputs. For every prompt X, define the bad- and good-response sets by*

$$\mathcal{B}_X(\tau,\eta) = \{y : s_{\theta_0}(X,y) \le \tau,\ F(X,y) \le \eta\}, \mathcal{G}_X(\kappa,\gamma) \quad = \{y : s_{\theta_0}(X,y) \ge \kappa,\ F(X,y) \ge \gamma\}$$

respectively. Suppose that $0 < \tau < \kappa \le 1$, $0 \le \eta < \gamma \le 1$, and, for ν-almost every X, $\pi_0(\mathcal{G}_X(\kappa,\gamma) \mid X) \ge \rho$ for some $\rho > 0$. Further, suppose that the final learned LLM satisfies

the forward-KL approximation condition $\mathbb{E}_{X\sim\nu}\left[D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot \mid X)\,\|\,\pi^*_{\hat{\theta}}(\cdot \mid X)\right)\right] \le \varepsilon_{\mathrm{seq}}$. *Then*

$$\mathbb{E}_{X\sim\nu}\left[\pi^*_{\hat{\theta}}\left(\mathcal{B}_X(\tau,\eta) \mid X\right)\right] \le \frac{1}{\rho}\left(\frac{\tau}{\kappa}\right)^{\alpha} e^{-\alpha(\gamma-\eta)} + \sqrt{\frac{\varepsilon_{\mathrm{seq}}}{2}}. \tag{5}$$

Theorem 4.1 shows that the final learned LLM inherits the target LLM's suppression of responses that are both incompatible with the pretrained model and insufficiently faithful, up to an additive approximation term. The result characterizes how closeness to the optimized target transfers its suppression guarantee to the final learned LLM.

Proposition 4.2. *Suppose that* $\pi_0(\cdot \mid X)$ *and* $\pi_{\theta_0}(\cdot \mid X)$ *are untruncated autoregressive softmax LLMs over the same vocabulary and valid response space, with finite logits. Then*

$$\operatorname{supp}\pi^*_{\hat{\theta}}(\cdot \mid X) = \operatorname{supp}\pi_{\theta_0}(\cdot \mid X). \tag{6}$$

Proposition 4.2 establishes that, under the stated autoregressive softmax assumptions, the final learned LLM preserves the pretrained model's valid response support. Thus, TIDE reweights probability mass within the pretrained response space rather than introducing responses outside it.

4.2 Connection to Rejection Sampling

Although TIDE directly optimizes the observed human responses through weighted likelihood, its target LLM can equivalently be obtained by proposing responses from the human-response distribution and accepting them with probabilities determined by their values, admitting a natural rejection-sampling interpretation. This is an analytical interpretation rather than algorithm used to train TIDE; it makes explicit how exponential reweighting moves probability mass toward high-value responses.

Theorem 4.3 (Rejection-sampling representation of TIDE)**.** *Fix* X *and let* $\alpha > 0$. *Define* $V_{\max}(X) = \sup_{y\in\mathcal{Y}_X} V(X,y)$. *If a response* $Y \sim \pi_0(\cdot \mid X)$ *is accepted with probability* $A_{\mathrm{TIDE}}(X,Y) = \exp\{\alpha(V(X,Y) - V_{\max}(X))\}$, *then the accepted response follows exactly* $\pi^*_\alpha(\cdot \mid X)$. *The probability of accepting one proposal is*

$$a^*_{\mathrm{TIDE}}(X) = \mathbb{E}_{y\sim\pi_0(\cdot|X)}[e^{\alpha V(X,y)}]e^{-\alpha V_{\max}(X)}. \tag{7}$$

Before normalization, an accepted response y has probability mass $\pi_0(y \mid X)A_{\mathrm{TIDE}}(X,y) = \pi_0(y \mid X)\exp\{\alpha(V(X,y) - V_{\max}(X))\}$. Summing this expression over responses gives $a^*_{\mathrm{TIDE}}(X)$ in Equation (7). Conditioning on acceptance then cancels the common factor $e^{-\alpha V_{\max}(X)}$ and yields

$$\Pr(Y = y \mid X, \mathrm{accept}) = \frac{\pi_0(y \mid X)e^{\alpha V(X,y)}}{\mathbb{E}_{y\sim\pi_0(\cdot|X)}[e^{\alpha V(X,y)}]} = \pi^*_\alpha(y \mid X). \tag{8}$$

The construction has three key steps. It propose human responses from π_0, assigns larger acceptance probabilities to responses with larger values, and conditions on acceptance. The last step normalizes the accepted responses and produces exactly the exponentially reweighted TIDE target. Importantly, this is only a probabilistic interpretation: TIDE does not sample from π_0 or reject responses during training; it realizes the same reweighting directly through its weighted likelihood objective.

We next place Self-Generated Demonstration (Self-Demo) training (Finlayson et al., 2025) in the same framework. Self-Demo repeatedly samples responses from the pretrained model and rejects those that fail its correctness criterion, whereas TIDE uses rejection sampling only as an analytical interpretation and performs no sampling in practice. Their one-proposal acceptance probabilities therefore provide a common basis for comparing training-target construction costs.

Let g denote the observed human-written response and let $q_{\mathrm{SelfDemo}}(\cdot \mid X)$ denote the distribution from which Self-Demo generates candidate responses. We define the correct-response set as $\mathcal{S}_{X,g} = \{y \in \mathcal{Y}_X : y \text{ is judged correct relative to } g\}$. Under an idealized rejection-sampling view, Self-Demo repeatedly draws responses from $q_{\mathrm{SelfDemo}}(\cdot \mid X)$ until a candidate belonging to $\mathcal{S}_{X,g}$ is obtained. Its one-proposal success probability is therefore $a^*_{\mathrm{SelfDemo}}(X,g) = q_{\mathrm{SelfDemo}}(\mathcal{S}_{X,g} \mid X)$. Thus, the computational efficiency of Self-Demo depends directly on how much probability its candidate generator assigns to responses that pass the correctness criterion.

To compare the two acceptance probabilities, we propose the following theorem:

Theorem 4.4 (Acceptance-probability comparison). *Suppose that, for some $C \geq 0$ and $R > 0$, every response $y \in \mathcal{S}_{X,g}$ satisfies $\alpha(V_{\max}(X) - V(X,y)) \leq C$, and suppose that $\pi_0(\mathcal{S}_{X,g} \mid X) \geq R\, q_{\text{SelfDemo}}(\mathcal{S}_{X,g} \mid X)$. If $Re^{-C} > 1$ and $a^*_{\text{SelfDemo}}(X,g) > 0$, then $a^*_{\text{TIDE}}(X) > a^*_{\text{SelfDemo}}(X,g)$.*

Theorem 4.4 shows that TIDE has a larger one-proposal acceptance probability when correct responses have sufficiently high value and receive adequate probability under π_0 relative to the Self-Demo generator. This directly reduces the number of proposals required to obtain a usable response.

To obtain at least one accepted response with probability at least $1 - \delta$, Self-Demo requires

$$N_{\text{SelfDemo}}(X, g, \delta) = \left\lceil \frac{\log \delta}{\log\left(1 - a^*_{\text{SelfDemo}}(X,g)\right)} \right\rceil \tag{9}$$

candidate responses. The corresponding general rejection-sampling result and additional candidate-complexity details are provided in Appendix B.12.

This comparison is interpretive: our training algorithm does not sample from π_0 or perform rejection sampling. Rather it transfers the same exponential preference into the model through value-weighted likelihood training using the observed human responses. Self-Demo, by contrast, realizes its selection mechanism explicitly by generating and evaluating multiple candidates for each training input. The rejection-sampling view therefore explains why Self-Demo incurs additional generation and evaluation cost, whereas our method performs a single weighted update on the available response.

5 Experiments

Benchmarks and Models. Under accurate retrieval, we train TIDE on MS MARCO (Bajaj et al., 2016) and NewsQA (Trischler et al., 2017), and evaluate its in-domain and cross-domain performance on MS MARCO, HotpotQA (Yang et al., 2018), Natural Questions (NQ) (Kwiatkowski et al., 2019), Wizard of Wikipedia (WoW) (Dinan et al., 2019), and T-REx (Elsahar et al., 2018). We compare TIDE against Self-Demo (Finlayson et al., 2025), supervised fine-tuning (SFT), and the pretrained base model. Experiments use Qwen3-4B and Qwen3-14B as the base and trainable LLMs. Complete dataset construction, training, and evaluation details are provided in Appendix C.

Choice of baselines. We compare TIDE with SFT, which directly fits the human-written responses, and Self-Demo, which addresses supervision mismatch using filtered model-generated targets. Other faithfulness-oriented methods rely on additional generated supervision, preference construction, or different training mechanisms, and are therefore not direct comparisons under our fixed human-supervision setting.

Evaluation. For answer quality, the judge compares the generated response with the reference answers while using the retrieved context. A response is attempted if it is nonempty and not a refusal, and an example is answerable if its context contains sufficient evidence. If A, C, and N_{ans} denote the numbers of attempted, correct, and answerable responses, respectively, we denote precision and recall as $P_{\text{ans}} = \frac{C}{A}$ and $R_{\text{ans}} = \frac{C}{N_{\text{ans}}}$ respectively, and report $F1 = \frac{2P_{\text{ans}}R_{\text{ans}}}{P_{\text{ans}}+R_{\text{ans}}}$. Following prior RAG evaluation work (Es et al., 2023; Adlakha et al., 2024), we evaluate faithfulness separately from answer correctness. The judge ignores reference-answer agreement and marks a generation as supported only when every factual claim is entailed by the retrieved context. With $S_i \in \{0,1\}$ denoting this judgment for generation i, we report $\text{Faith}_{\text{supported}} = \frac{1}{N}\sum_{i=1}^{N} S_i = \frac{N_{\text{supported}}}{N}$. The complete prompts and scoring instructions are provided in Appendix E.

Main Results: Table 1 shows that TIDE achieves the highest overall F1 at both model scales, with averages of 0.452 for Qwen3-4B and 0.448 for Qwen3-14B, and ranks first in seven of the ten train–evaluation settings at each scale. More importantly, its gains are concentrated on cross-domain evaluations such as HotpotQA, WoW, and T-REx. This pattern suggests that directly fitting human responses through SFT can improve performance on the training distribution but may transfer less reliably when the response style, required reasoning, or answer format changes.

Baseline results clarify this pattern. SFT performs best on in-domain MS MARCO, where the training and evaluation distributions match, while Self-Demo leads on NQ, plausibly because its short,

Table 1: F1 score across RAG evaluation benchmarks. The best result is highlighted in light blue and the two highest results are shown in bold.

		Qwen3-4B Model				Qwen3-14B Model			
Train	**Eval**	**TIDE**	**Self-Demo**	**SFT**	**Base**	**TIDE**	**Self-Demo**	**SFT**	**Base**
MSMARCO	MSMARCO	**0.588**	0.486	**0.634**	0.422	**0.594**	0.522	**0.664**	0.484
	HotpotQA	**0.722**	**0.623**	0.569	0.609	**0.728**	0.681	0.640	**0.692**
	NQ	0.350	**0.400**	**0.371**	0.330	0.311	**0.332**	**0.317**	0.256
	WoW	**0.302**	0.194	**0.289**	0.012	**0.290**	0.143	**0.186**	0.010
	T-REx	**0.314**	0.150	0.140	**0.186**	**0.327**	0.205	**0.251**	0.191
NewsQA	MSMARCO	**0.584**	**0.517**	0.405	0.422	**0.559**	**0.513**	0.395	0.484
	HotpotQA	**0.713**	**0.650**	0.258	0.609	**0.717**	**0.695**	0.273	0.692
	NQ	**0.353**	**0.422**	0.186	0.330	**0.308**	**0.334**	0.164	0.256
	WoW	**0.307**	0.060	**0.266**	0.012	**0.337**	0.067	**0.259**	0.010
	T-REx	**0.292**	0.170	0.145	**0.186**	**0.310**	**0.192**	0.149	0.191
Overall Average		**0.452**	**0.367**	0.326	0.312	**0.448**	**0.368**	0.330	0.326

model-generated targets suit NQ's extractive format. TIDE, relying on neither train–test alignment nor a particular answer style, performs more consistently across datasets. Similar rankings across model scales suggest its gains are not capacity-specific. These results support TIDE's motivation: combining pretrained-behavior compatibility with evidence-grounded supervision balances in-domain adaptation and cross-domain generalization.

Table 2 reports $\mathrm{Faith}_{\mathrm{supported}}$ for Qwen3-14B models evaluated by the fixed Qwen3-32B context-faithfulness judge. TIDE achieves the highest average faithfulness under both training distributions, reaching 0.949 when trained on MS MARCO and 0.943 when trained on NewsQA. It ranks first in seven of the ten train–evaluation settings and remains among the two strongest methods in every setting. This consistency is important as the evaluation benchmarks differ substantially in their domains, response styles, and evidence structures. The results therefore indicate that the grounding behavior learned by TIDE is not restricted to the pretraining distribution.

The largest gains over the pretrained model occur on NQ and WoW, where plausible answers from parametric knowledge may be insufficient: responses must remain supported by the retrieved passage or dialogue context. TIDE also maintains high faithfulness on MS MARCO, HotpotQA, and T-REx, indicating its gain not relying on a single evidence format. This aligns with the faithfulness term in the TIDE value, which downweights unsupported responses and emphasizes grounded ones.

These results complement the answer-quality findings in Table 1. The faithfulness judge evaluates only whether the generated claims are supported by the retrieved context and does not reward agreement with reference answer. Consequently, the strong scores cannot be explained solely by reproducing reference-answer wording. Together, the answer-quality and faithfulness results suggest that TIDE improves task performance while consistently being faithful to the supplied evidence.

Effectiveness of the Faithfulness Component. TIDE achieves an overall $\mathrm{Faith}_{\mathrm{supported}}$ score of 0.946, compared with 0.799 achieved by TIDE$-$Faithfulness, which removes $F(X, y)$ while retaining the pretrained-model compatibility term, and performs better in eight of the ten train–evaluation settings, as shown in Table 3. The improvement is especially pronounced on T-REx, where removing faithfulness reduces the score from 0.944 to 0.588 under MS MARCO training and from 0.934 to 0.132 under NewsQA training. Large improvements also appear on HotpotQA and on WoW under NewsQA training. Although TIDE$-$Faithfulness performs better on MS MARCO and WoW under MS MARCO training, its substantially lower cross-domain average shows that pretrained-model compatibility alone does not consistently encourage grounding in retrieved evidence. These results demonstrate that the faithfulness term is essential for the context-grounded behavior of TIDE.

Distributional Shift and Task Performance. We examine the trade-off between F1 and deviation from the pretrained model, measured by KL divergence, for NewsQA-trained models. Lower KL indicates greater similarity to the pretrained model along generated responses. Figure 1(a) shows that KL-divergence alone does not ensure strong performance. Self-Demo achieves low KL by training on pretrained-model generations, but lower F1 than TIDE. Conversely, SFT fits human responses without considering pretrained-model compatibility, yielding the highest KL and lowest F1 among the post-trained methods. Removing the compatibility term from TIDE also increases

Table 2: $\text{Faith}_{\text{supported}}$ across RAG benchmarks. The best result is highlighted in light blue, and the two highest results are shown in bold.

	MSMARCO-Trained				NewsQA-Trained			
Evaluation	**TIDE**	**Self-Demo**	**SFT**	**Base**	**TIDE**	**Self-Demo**	**SFT**	**Base**
MSMARCO	**0.972**	**0.972**	0.970	0.958	**0.972**	**0.963**	0.926	0.958
HotpotQA	**0.985**	0.916	0.917	**0.984**	**0.980**	0.931	0.902	**0.984**
NQ	**0.986**	0.945	**0.952**	0.762	**0.982**	**0.967**	0.932	0.762
WoW	**0.858**	**0.624**	0.431	0.586	**0.846**	0.570	**0.739**	0.585
T-REx	**0.944**	0.688	0.330	**0.981**	**0.934**	0.408	0.157	**0.981**
Average	**0.949**	0.829	0.720	**0.854**	**0.943**	0.768	0.731	**0.854**

Table 3: Supported faithfulness for Qwen3-14B models with and without the faithfulness component. TIDE−Faithfulness removes $F(X, y)$ while retaining pretrained-model compatibility. The best result within each training setting is highlighted.

Train	**Method**	**MSMARCO**	**HotpotQA**	**NQ**	**WoW**	**T-REx**	**Average**
MSMARCO	TIDE	0.972	**0.985**	**0.986**	0.858	**0.944**	**0.949**
	TIDE−Faithfulness	**0.987**	0.938	0.974	**0.911**	0.588	0.880
NewsQA	TIDE	**0.972**	**0.980**	**0.982**	**0.846**	**0.934**	**0.943**
	TIDE−Faithfulness	0.955	0.931	0.959	0.617	0.132	0.719
Overall Average			**TIDE: 0.946**			TIDE−Faithfulness: 0.799	

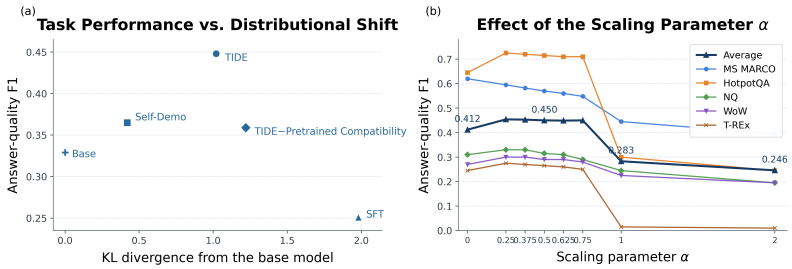

Figure 1: (a) F1 versus KL divergence from the base model; TIDE−Pretrained Compatibility removes the pretrained-model compatibility term. (b) F1 across α, averaged over the five benchmarks.

KL and reduces F1. Full TIDE achieves the highest F1 with substantially lower KL than SFT, demonstrating the value of balancing human supervision with pretrained behavior preservation.

Effect of the Scaling Parameter. Figure 1(b) shows how α controls the strength of value-based reweighting relative to the reference distribution. At $\alpha = 0$, all responses receive equal weight, reducing the objective to SFT, yielding an average F1 of 0.412. Increasing α initially improves performance, with the average F1 reaching 0.45 at $\alpha = 0.25$ and remaining stable between 0.449 and 0.454 for $\alpha \in [0.25, 0.75]$, indicating limited sensitivity within that range. Performance declines sharply to 0.283 at $\alpha = 1$ and 0.246 at $\alpha = 2$, with particularly large degradation on HotpotQA and T-REx. Large α concentrates training on high-value responses underemphasizing human supervision; moderate values balance pretrained-model compatibility, faithfulness and human supervision.

6 Discussion

TIDE is a rollout-free RAG post-training framework that retains human-written supervision while promoting pretrained-distribution compatibility and faithfulness. Its implicit-distillation objective trains directly on observed responses without modeling the unknown response distribution. Our analysis shows suppression of low-compatibility and unfaithful responses and gives a rejection-sampling interpretation. Across five RAG benchmarks and two model scales, TIDE achieves the highest average answer-quality F1 while maintaining strong faithfulness. This study assumes reliable retrieval and evaluates Qwen3 models on selected knowledge-intensive benchmarks; future work should extend TIDE to noisy retrieval, broader architectures, and additional domains.

AI USE STATEMENT

In this work, we used generative AI tools, including OpenAI ChatGPT and Codex, to assist the implementation and revision of research code. We also used these tools to draft and edit portions of the manuscript, improve its organization and readability and revise scientific figures. We did not use generative AI to generate synthetic training or evaluation data, translate research materials, clean or reformat datasets, or conduct qualitative or thematic data analysis; tasks involving surveys, interviews, or transcription were not applicable.

All AI-assisted material was reviewed by the authors. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work uses publicly available RAG benchmarks and pretrained language models and does not involve new human-subject data collection. TIDE is designed to improve the use of retrieved evidence during RAG post-training and may support more reliable, context-grounded generation. As with other RAG methods, its performance depends on the quality and coverage of the retrieved evidence, and inherited biases in the underlying models or datasets may remain. We therefore encourage appropriate source verification and human oversight when applying the method in consequential settings. We report the models, datasets, evaluation procedures, and limitations necessary to support transparent and responsible use of the proposed approach.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete proofs of our theoretical results in Appendix B. The datasets, preprocessing, training examples, decoding settings, and evaluation protocol are described in Appendix C, while the implementation and optimization details, together with the anonymous source code, are reported in Appendix D. The complete prompts and scoring instructions used by the LLM-based evaluators are provided in Appendix E.

REFERENCES

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024. doi: 10.1162/tacl_a_00667.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. SELF-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016. URL https://arxiv.org/abs/1611.09268.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1811.01241.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association, 2018. URL https://aclanthology.org/L18-1544/.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Roman Vashurin, Shehzaad Dhuliawala, Artem Shelmanov, Timothy Baldwin, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation. *arXiv preprint arXiv:2505.21072*, 2025.

Matthew Finlayson, Ilia Kulikov, Daniel M. Bikel, Barlas Oguz, Xilun Chen, and Aasish Pappu. Post-training an LLM for RAG? train on self-generated demonstrations. *arXiv preprint arXiv:2502.10596*, 2025. doi: 10.48550/arXiv.2502.10596. URL https://arxiv.org/abs/2502.10596.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yuqing Huang, Rongyang Zhang, Qimeng Wang, Chengqiang Lu, Yan Gao, Yi Wu, Yao Hu, Xuyang Zhi, Guiquan Liu, Xin Li, Hao Wang, and Enhong Chen. SelfAug: Mitigating catastrophic forgetting in retrieval-augmented generation via distribution self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14175–14190, 2025.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 874–880, 2021.

Yihan Jiao, Zhehao Tan, Dan Yang, Duolin Sun, Jie Feng, Yue Shen, Jian Wang, and Peng Wei. HIRAG: Hierarchical-thought instruction-tuning retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5111–5130, 2025.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*, 2024.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Wanlong Liu, Junying Chen, Ke Ji, Li Zhou, Wenyu Chen, and Benyou Wang. RAG-instruct: Boosting LLMs with diverse retrieval-augmented instructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3865–3888, 2025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 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*, pp. 9802–9822, 2023.

David Mueller, Mark Dredze, and Nicholas Andrews. Multi-task transfer matters during instruction-tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14880–14891, 2024.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 10862–10878, 2024.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 338–354, 2024.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 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*, pp. 783–791, 2024a.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8371–8384, 2024b.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*, 2023.

Xiaqiang Tang, Yi Wang, Keyu Hu, Rui Xu, Chuang Li, Weigao Sun, Jian Li, and Sihong Xie. SSFO: Self-supervised faithfulness optimization for retrieval-augmented generation. *arXiv preprint arXiv:2508.17225*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-2623. URL https://aclanthology.org/W17-2623/.

Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, and Kai-Wei Chang. Synchronous faithfulness monitoring for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9390–9406, 2024.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1028–1043, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://arxiv.org/abs/1809.09600.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*, 2024.

Miaosen Zhang, Yishan Liu, Shuxia Lin, Xu Yang, Qi Dai, Chong Luo, Weihao Jiang, Peng Hou, Anxiang Zeng, Xin Geng, and Baining Guo. Towards on-policy SFT: Distribution discriminant theory and its applications in LLM training. *arXiv preprint arXiv:2602.12222*, 2026.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *arXiv preprint arXiv:2506.08938*, 2025.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*, 2024.

A Additional Related Work

Retrieval-Augmented Language Models. Large language models trained to predict the next token over large text corpora have achieved strong performance across a wide range of natural language processing tasks (Rae et al., 2021; Hoffmann et al., 2022; Touvron et al., 2023). However, the knowledge stored in their parameters is limited by the information available during pre-training and may become incomplete or outdated. Retrieval-augmented language models address this limitation by combining parametric knowledge with information retrieved from an external corpus (Guu et al., 2020; Lewis et al., 2020; Izacard & Grave, 2021). Subsequent work has explored different ways of connecting retrieval and generation. For example, REPLUG treats the language model as a frozen black box and trains its retriever using feedback from the model (Shi et al., 2024b), while other work shows that retrieval is particularly useful for long-tail knowledge that is poorly represented in model parameters (Mallen et al., 2023). Although retrieval provides access to relevant external information, the model may still ignore the retrieved context, rely instead on its parametric knowledge, or generate claims that are unsupported by the supplied evidence.

Post-Training for RAG. Beyond improving retrieval itself, a growing line of work adapts the language model through retrieval-augmented post-training. SELF-RAG trains a model to retrieve information, generate responses, and critique the relevance and factual support of its own generations (Asai et al., 2024). SSFO constructs preference pairs from responses generated with and without retrieved context and applies Direct Preference Optimization to favor context-grounded responses (Tang et al., 2025; Rafailov et al., 2023). More broadly, RA-DIT adapts both the language model and retriever to improve retrieval-conditioned generation (Lin et al., 2024), whereas RAFT trains the model to identify useful evidence in the presence of distracting documents (Zhang et al., 2024). Recent work also constructs diverse retrieval-augmented instructions through RAG-INSTRUCT (Liu et al., 2025) or teaches models to progressively filter, combine, and reason over retrieved information through HIRAG (Jiao et al., 2025). These methods demonstrate that post-training can substantially improve how language models use retrieved evidence. Nevertheless, post-training does not by itself guarantee that the resulting responses remain faithful to the retrieved context or that the supervision is compatible with the pretrained model's response distribution.

Faithfulness Evaluation and Mitigation. Despite advances in retrieval-augmented post-training, generated responses may still ignore, contradict, or introduce claims unsupported by the retrieved evidence. This has motivated the development of specialized benchmarks, evaluation frameworks, and mitigation methods. RAGTRUTH provides fine-grained human annotations of hallucinations in retrieval-augmented outputs (Niu et al., 2024), while ARES evaluates RAG systems along the dimensions of context relevance, answer faithfulness, and answer relevance (Saad-Falcon et al., 2024). Work on attributable generation further evaluates whether individual claims in a generated response are adequately supported by cited evidence (Gao et al., 2023). Fadeeva et al. (2025) study uncertainty-based detection of unreliable RAG outputs and emphasize the distinction between faithfulness to the retrieved context and factual correctness with respect to the external world.

Several methods seek to mitigate unsupported generation at inference time. Context-Aware Decoding contrasts the model's predictions with and without the retrieved context, thereby shifting generation toward context-supported responses (Shi et al., 2024a). SYNCHECK monitors fine-grained signals during generation, including sequence likelihood, uncertainty, context influence, and semantic alignment, to identify and correct unsupported sentences (Wu et al., 2024). FAITHFULRAG explicitly identifies conflicts between retrieved evidence and the model's parametric knowledge and introduces a fact-level reasoning process to resolve these conflicts before generation (Zhang et al., 2025). Yoran et al. (2024) fine-tune language models on a mixture of relevant and irrelevant passages to make generation robust to retrieval noise. RECOMP compresses retrieved documents and can selectively omit retrieval augmentation when the evidence is not useful (Xu et al., 2024), while CRAG evaluates retrieval quality and triggers corrective retrieval actions when the original documents are unreliable (Yan et al., 2024). Although effective, these methods primarily address faithfulness through evaluation, filtering, monitoring, corrective reasoning, or decoding-time interventions. They do not explicitly optimize faithfulness together with compatibility between the post-training supervision and the pretrained response distribution.

Distribution Compatibility in Post-Training. Alongside faithfulness, recent work has increasingly examined the compatibility between post-training supervision and the pretrained model's response distribution. Fine-tuning on examples containing knowledge unfamiliar to the pretrained model can influence how the resulting model hallucinates, suggesting that compatibility between the training examples and the model distribution matters (Kang et al., 2024). Instruction tuning may also reduce transfer to unseen tasks, highlighting the possibility that post-training overwrites previously acquired capabilities (Mueller et al., 2024).

Self-generated supervision offers one approach to reducing this mismatch. Self-Distillation Fine-Tuning rewrites task supervision using responses produced by the model itself, thereby preserving more of its original instruction-following behavior (Yang et al., 2024). Similarly, SELF-ALIGN trains language models on their own principle-guided responses and demonstrates that self-generated outputs can provide effective supervision with limited human annotation (Sun et al., 2023). More recently, Distribution Discriminant Theory characterizes the relationship between the training-data distribution and the model-induced distribution and motivates in-distribution fine-tuning and hinted decoding for reducing this mismatch (Zhang et al., 2026). SELFAUG connects distribution shift with catastrophic forgetting in the RAG setting and aligns the fine-tuned model's logits with its original semantic distribution (Huang et al., 2025).

Self-Generated Demonstration training instead replaces externally written targets with responses sampled from the reference model (Finlayson et al., 2025). This produces targets that are naturally compatible with the model's pretrained behavior. However, the procedure requires repeated candidate generation, evaluation, and filtering to obtain suitable demonstrations. Its success also depends on the reference model generating an acceptable response among the sampled candidates, while replacing the original human-written response may discard useful supervision encoded in that response. Consequently, faithfulness and distributional compatibility remain distinct but coupled objectives that should be addressed jointly.

Positioning of TIDE. In this work, we construct an optimized response distribution that retains the supervision encoded by the human-response distribution while reweighting responses according to both their length-normalized likelihood under the pretrained model and their faithfulness to the retrieved evidence. We then distill this analytically characterized target into the language model through a tractable value-weighted likelihood objective. Unlike methods that address faithfulness only through decoding-time correction, TIDE incorporates faithfulness directly into post-training. Unlike Self-Generated Demonstration training, its practical objective retains the available human responses and does not require repeatedly generating, evaluating, and filtering multiple candidate responses for each training example. Thus, TIDE jointly addresses evidence-grounded generation and compatibility with pretrained behavior without additional response rollouts.

B TECHNICAL PROOFS AND DERIVATIONS

For compactness in the appendix, define the prompt-specific normalizer

$$Z_\alpha(X) = \mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[e^{\alpha V(X,y)}\right]$$

and the corresponding global normalizer

$$\mathcal{Z}_\alpha = \mathbb{E}_{X\sim\mathcal{D}_X}\left[Z_\alpha(X)\right] = \mathbb{E}_{\substack{X\sim\mathcal{D}_X \\ y\sim\pi_0(\cdot|X)}}\left[e^{\alpha V(X,y)}\right].$$

These abbreviations are used only in the appendix; the main text writes the corresponding expectations explicitly.

B.1 PROOF OF THEOREM 3.1

Proof. For any conditional LLM $\pi(\cdot \mid X)$ that is absolutely continuous with respect to $\pi_0(\cdot \mid X)$, Equation (2) gives

$$\log \frac{\pi^*_\alpha(y \mid X)}{\pi_0(y \mid X)} = \alpha V(X,y) - \log Z_\alpha(X).$$

Therefore,

$$D_{\mathrm{KL}}(\pi(\cdot \mid X) \,\|\, \pi^*_\alpha(\cdot \mid X)) = D_{\mathrm{KL}}(\pi(\cdot \mid X) \,\|\, \pi_0(\cdot \mid X)) - \alpha\, \mathbb{E}_{y\sim\pi(\cdot|X)}[V(X,y)] + \log Z_\alpha(X). \tag{10}$$

Rearranging this identity and averaging over $X \sim \mathcal{D}_X$ yields

$$\begin{aligned} &\mathbb{E}_{X\sim\mathcal{D}_X}\left[\alpha \mathbb{E}_{y\sim\pi(\cdot|X)}[V(X,y)] - D_{\mathrm{KL}}(\pi(\cdot \mid X) \,\|\, \pi_0(\cdot \mid X))\right] \\ &\quad = \mathbb{E}_{X\sim\mathcal{D}_X}\left[\log Z_\alpha(X) - D_{\mathrm{KL}}(\pi(\cdot \mid X) \,\|\, \pi^*_\alpha(\cdot \mid X))\right]. \end{aligned} \tag{11}$$

The first term is independent of π. Nonnegativity of KL divergence therefore shows that the unique maximizer is $\pi(\cdot \mid X) = \pi^*_\alpha(\cdot \mid X)$ for $\mathcal{D}_X$-almost every X, up to equality almost everywhere in y. □

B.2 Equivalent Joint-Distribution Formulation and Normalizer Cancellation

To connect the conditional LLM optimization in the main paper to the value-weighted empirical training objective, let $p_{\mathrm{data},X} = \mathcal{D}_X$ and define

$$p_{\mathrm{data}}(X,y) = p_{\mathrm{data},X}(X)\pi_0(y \mid X).$$

Introduce an auxiliary joint distribution $q(X,y)$ and the objective

$$\mathcal{J}^{\mathrm{joint}}_\alpha(q) = \alpha \mathbb{E}_{(X,y)\sim q}[V(X,y)] - D_{\mathrm{KL}}(q \,\|\, p_{\mathrm{data}}). \tag{12}$$

This is the joint-distribution analogue of the conditional objective in Equation (1). This joint formulation is used only to characterize how training examples are reweighted; the deployed language model remains a conditional LLM over $y \mid X$.

Proposition B.1 (Equivalence of conditional and joint formulations)**.** *Suppose* $0 < \mathcal{Z}_\alpha < \infty$. *Then Equation* (12) *has the unique optimizer*

$$\widetilde{q}_\alpha(X,y) = \frac{p_{\mathrm{data}}(X,y)e^{\alpha V(X,y)}}{\mathcal{Z}_\alpha},$$

and its conditional response LLM is exactly the optimizer in Theorem 3.1:

$$\widetilde{q}_\alpha(y \mid X) = \pi^*_\alpha(y \mid X).$$

Proof. We present the proof for discrete prompt and response spaces; the general case follows by replacing sums with integrals and probability mass functions with Radon–Nikodym derivatives.

Recall

$$\mathcal{J}^{\mathrm{joint}}_\alpha(q) = \alpha \sum_{X,y} q(X,y)V(X,y) - \sum_{X,y} q(X,y)\log\frac{q(X,y)}{p_{\mathrm{data}}(X,y)},$$

where $q(X,y) \geq 0$, $\sum_{X,y} q(X,y) = 1$, and $q(X,y) = 0$ whenever $p_{\mathrm{data}}(X,y) = 0$.

Introducing a multiplier λ for normalization and $\mu_{X,y} \geq 0$ for nonnegativity gives

$$\begin{aligned} \mathscr{L}(q,\lambda,\{\mu_{X,y}\}) = \alpha \sum_{X,y} q(X,y)V(X,y) - \sum_{X,y} q(X,y)\log\frac{q(X,y)}{p_{\mathrm{data}}(X,y)} \\ - \lambda\left(\sum_{X,y} q(X,y) - 1\right) + \sum_{X,y}\mu_{X,y}q(X,y). \end{aligned} \tag{13}$$

The stationarity condition is

$$\alpha V(X,y) - \log\frac{q(X,y)}{p_{\mathrm{data}}(X,y)} - 1 - \lambda + \mu_{X,y} = 0.$$

For (X,y) in the support of p_{data}, the optimal solution is interior, so complementary slackness implies $\mu_{X,y} = 0$. Hence,

$$\log\frac{q(X,y)}{p_{\mathrm{data}}(X,y)} = \alpha V(X,y) - \lambda - 1,$$

and therefore

$$q(X, y) = p_{\text{data}}(X, y) \exp\{\alpha V(X, y) - \lambda - 1\}. \tag{14}$$

Imposing $\sum_{X,y} q(X, y) = 1$ yields

$$1 = e^{-\lambda-1} \sum_{X,y} p_{\text{data}}(X, y) e^{\alpha V(X,y)} = e^{-\lambda-1} \mathcal{Z}_\alpha,$$

Thus $e^{-\lambda-1} = 1/\mathcal{Z}_\alpha$, and Equation (14) gives

$$\widetilde{q}_\alpha(X, y) = \frac{p_{\text{data}}(X, y) e^{\alpha V(X,y)}}{\mathcal{Z}_\alpha}. \tag{15}$$

The assumption $0 < \mathcal{Z}_\alpha < \infty$ guarantees that $\widetilde{q}_\alpha$ is well defined.

It remains to verify global optimality. From Equation (15),

$$\log \frac{\widetilde{q}_\alpha(X, y)}{p_{\text{data}}(X, y)} = \alpha V(X, y) - \log \mathcal{Z}_\alpha.$$

Therefore, for any $q \in \mathcal{Q}$,

$$D_{\text{KL}}(q \,\|\, \widetilde{q}_\alpha) = D_{\text{KL}}(q \,\|\, p_{\text{data}}) - \alpha \mathbb{E}_q[V(X, y)] + \log \mathcal{Z}_\alpha, \tag{16}$$

or equivalently,

$$\mathcal{J}_\alpha^{\text{joint}}(q) = \log \mathcal{Z}_\alpha - D_{\text{KL}}(q \,\|\, \widetilde{q}_\alpha). \tag{17}$$

Since KL divergence is nonnegative, $\mathcal{J}_\alpha^{\text{joint}}(q)$ is uniquely maximized by $q = \widetilde{q}_\alpha$, up to equality almost everywhere.

Finally, using

$$p_{\text{data}}(X, y) = p_{\text{data},X}(X) \pi_0(y \mid X)$$

and using the prompt-specific normalizer $Z_\alpha(X)$ defined above, marginalization of Equation (15) gives

$$\widetilde{q}_{\alpha,X}(X) = \frac{p_{\text{data},X}(X) Z_\alpha(X)}{\mathcal{Z}_\alpha}.$$

Consequently,

$$\widetilde{q}_\alpha(y \mid X) = \frac{\widetilde{q}_\alpha(X, y)}{\widetilde{q}_{\alpha,X}(X)} = \frac{\pi_0(y \mid X) e^{\alpha V(X,y)}}{Z_\alpha(X)} = \pi_\alpha^*(y \mid X). \tag{18}$$

Thus the global normalizer $\mathcal{Z}_\alpha$ appears in both the joint numerator and the optimized prompt marginal and cancels after conditioning. The only remaining normalizer is the prompt-specific $Z_\alpha(X)$ from the main conditional formulation. Hence the conditional and joint optimization problems lead to exactly the same target LLM π_α^*. This completes the proof. □

B.3 Sequence-Level Form of Implicit Distillation

We now establish the precise sequence-level equivalence underlying implicit distillation. The length-normalized pretrained-model likelihood remains part of the response value $V(X, y)$, whereas the distillation loss compares the complete response distributions $\pi_\alpha^*(\cdot \mid X)$ and $\pi_\theta(\cdot \mid X)$.

Corollary B.2 (Implicit-distillation equivalence)**.** *Under the assumptions of Proposition B.1, the sequence-level forward-KL objective*

$$\mathcal{L}_{\text{distill}}(\theta) = \mathbb{E}_{X \sim \widetilde{q}_{\alpha,X}} \left[D_{\text{KL}}(\pi_\alpha^*(\cdot \mid X) \,\|\, \pi_\theta(\cdot \mid X)) \right] \tag{19}$$

satisfies

$$\mathcal{L}_{\text{distill}}(\theta) = C_\alpha + \frac{1}{\mathcal{Z}_\alpha} \mathcal{L}_{\text{TIDE}}^{\text{pop}}(\theta),$$

where C_α is independent of θ and $\mathcal{L}_{\text{TIDE}}^{\text{pop}}$ is the population counterpart of Equation (4). Consequently,

$$\arg\min_\theta \mathcal{L}_{\text{distill}}(\theta) = \arg\min_\theta \mathcal{L}_{\text{TIDE}}^{\text{pop}}(\theta).$$

Proof. Expanding the sequence-level forward-KL objective gives

$$\begin{aligned}\mathcal{L}_{\text{distill}}(\theta) &= \mathbb{E}_{X\sim\widetilde{q}_{\alpha,X}}\left[D_{\text{KL}}\left(\pi^*_\alpha(\cdot \mid X)\,\|\,\pi_\theta(\cdot \mid X)\right)\right] \\ &= C_\alpha - \mathbb{E}_{X\sim\widetilde{q}_{\alpha,X}}\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[\log \pi_\theta(y \mid X)\right], \end{aligned} \tag{20}$$

where C_α is independent of θ.

Since $\widetilde{q}_{\alpha,X}$ and $\pi^*_\alpha(\cdot \mid X)$ are the marginal and conditional distributions induced by the same optimized joint distribution $\widetilde{q}_\alpha$, Equation (20) can be written as

$$\mathcal{L}_{\text{distill}}(\theta) = C_\alpha - \mathbb{E}_{(X,y)\sim\widetilde{q}_\alpha}\left[\log \pi_\theta(y \mid X)\right]. \tag{21}$$

By Proposition B.1,

$$\widetilde{q}_\alpha(X,y) = \frac{p_{\text{data}}(X,y)e^{\alpha V(X,y)}}{\mathcal{Z}_\alpha}. \tag{22}$$

Substituting this expression into Equation (21) yields

$$\begin{aligned}\mathcal{L}_{\text{distill}}(\theta) &= C_\alpha - \frac{1}{\mathcal{Z}_\alpha}\mathbb{E}_{(X,y)\sim p_{\text{data}}}\left[e^{\alpha V(X,y)}\log \pi_\theta(y \mid X)\right] \\ &= C_\alpha + \frac{1}{\mathcal{Z}_\alpha}\mathcal{L}^{\text{pop}}_{\text{TIDE}}(\theta), \end{aligned} \tag{23}$$

where $\mathcal{L}^{\text{pop}}_{\text{TIDE}}$ denotes the population counterpart of Equation (4). The finite-sample objective in that equation is its empirical average over $\mathcal{D} = \{(X_i, g_i)\}_{i=1}^N$.

Since C_α and $\mathcal{Z}_\alpha > 0$ are independent of θ,

$$\arg\min_\theta \mathcal{L}_{\text{distill}}(\theta) = \arg\min_\theta \mathcal{L}^{\text{pop}}_{\text{TIDE}}(\theta). \tag{24}$$

Equivalently,

$$\nabla_\theta \mathcal{L}_{\text{distill}}(\theta) = \frac{1}{\mathcal{Z}_\alpha}\nabla_\theta \mathcal{L}^{\text{pop}}_{\text{TIDE}}(\theta). \tag{25}$$

Thus, the sequence-level value-weighted objective is exactly equivalent, up to an additive θ-independent term and a positive multiplicative constant, to sequence-level forward-KL distillation of the optimized TIDE target LLM. The length normalization used to measure pretrained-model compatibility remains inside $V(X,y)$ and is not applied to the distilled response log-likelihood. □

B.4 Proof of Proposition 4.2

Proof. We first show that

$$\operatorname{supp}\pi^*_{\widehat{\theta}}(\cdot \mid X) \subseteq \operatorname{supp}\pi_{\theta_0}(\cdot \mid X). \tag{26}$$

Since $\pi_0(\cdot \mid X)$ and $\pi_{\theta_0}(\cdot \mid X)$ are untruncated autoregressive softmax LLM policies over the same vocabulary and valid response space, with finite logits, both policies assign strictly positive probability to every valid finite response. Therefore,

$$\operatorname{supp}\pi_0(\cdot \mid X) = \operatorname{supp}\pi_{\theta_0}(\cdot \mid X). \tag{27}$$

Suppose, for contradiction, that

$$\operatorname{supp}\pi_{\theta_0}(\cdot \mid X) \subsetneq \operatorname{supp}\pi^*_{\widehat{\theta}}(\cdot \mid X). \tag{28}$$

Then there exists a response

$$y^0 \in (\operatorname{supp}\pi_{\theta_0}(\cdot \mid X))^c \quad \text{such that} \quad \pi^*_{\widehat{\theta}}(y^0 \mid X) > 0. \tag{29}$$

By Equation (27), $y^0 \notin \operatorname{supp}\pi_0(\cdot \mid X)$, and hence $\pi_0(y^0 \mid X) = 0$.

Now choose $y' \in \operatorname{supp}\pi_{\theta_0}(\cdot \mid X)$ and define a new LLM $\widehat{\pi}(\cdot \mid X)$ by transferring the probability assigned to y^0 to y'. More precisely,

$$\widehat{\pi}(y \mid X) = \pi^*_{\widehat{\theta}}(y \mid X), \qquad y \neq y^0, y', \tag{30}$$

and

$$\begin{aligned}\widehat{\pi}(y' \mid X) &= \pi^*_{\widehat{\theta}}(y' \mid X) + \pi^*_{\widehat{\theta}}(y^0 \mid X),\\ \widehat{\pi}(y^0 \mid X) &= 0.\end{aligned} \tag{31}$$

Thus, $\widehat{\pi}$ is a valid probability distribution. Moreover,

$$\widehat{\pi}(y' \mid X) > \pi^*_{\widehat{\theta}}(y' \mid X),$$

and therefore

$$\log \widehat{\pi}(y' \mid X) > \log \pi^*_{\widehat{\theta}}(y' \mid X).$$

Because $y' \in \operatorname{supp} \pi_0(\cdot \mid X)$, we also have $\pi_0(y' \mid X) > 0$.

The objective is unchanged at every response other than y^0 and y'. Removing probability from y^0 does not decrease the objective because $\pi_0(y^0 \mid X) = 0$, whereas increasing the probability assigned to y' strictly increases the objective. Consequently,

$$\begin{aligned}&\mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[\log \widehat{\pi}(y \mid X)\exp\{\alpha V(X,y)\}\right]\\ &\quad> \mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[\log \pi^*_{\widehat{\theta}}(y \mid X)\exp\{\alpha V(X,y)\}\right].\end{aligned} \tag{32}$$

This contradicts the optimality of $\pi^*_{\widehat{\theta}}$. Therefore,

$$\operatorname{supp} \pi^*_{\widehat{\theta}}(\cdot \mid X) \subseteq \operatorname{supp} \pi_{\theta_0}(\cdot \mid X). \tag{33}$$

We next prove the reverse inclusion,

$$\operatorname{supp} \pi_{\theta_0}(\cdot \mid X) \subseteq \operatorname{supp} \pi^*_{\widehat{\theta}}(\cdot \mid X). \tag{34}$$

Suppose, for contradiction, that there exists a response y such that

$$y \in \operatorname{supp} \pi_{\theta_0}(\cdot \mid X), \qquad y \notin \operatorname{supp} \pi^*_{\widehat{\theta}}(\cdot \mid X). \tag{35}$$

Then

$$\pi^*_{\widehat{\theta}}(y \mid X) = 0 \quad \text{and hence} \quad \log \pi^*_{\widehat{\theta}}(y \mid X) = -\infty.$$

By Equation (27), $y \in \operatorname{supp} \pi_0(\cdot \mid X)$ and thus $\pi_0(y \mid X) > 0$. Consequently,

$$\begin{aligned}&\max_{\pi} \mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[\log \pi(y \mid X)\exp\{\alpha V(X,y)\}\right]\\ &\quad= \mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[\log \pi^*_{\widehat{\theta}}(y \mid X)\exp\{\alpha V(X,y)\}\right] = -\infty.\end{aligned} \tag{36}$$

However, the pretrained LLM π_{θ_0} is a feasible candidate and satisfies $\pi_{\theta_0}(y \mid X) > 0$ for every valid finite response. Therefore,

$$\mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[\log \pi_{\theta_0}(y \mid X)\exp\{\alpha V(X,y)\}\right] > -\infty. \tag{37}$$

Thus, π_{θ_0} achieves a strictly larger objective value than $\pi^*_{\widehat{\theta}}$, contradicting the optimality of $\pi^*_{\widehat{\theta}}$. Hence,

$$\operatorname{supp} \pi_{\theta_0}(\cdot \mid X) \subseteq \operatorname{supp} \pi^*_{\widehat{\theta}}(\cdot \mid X). \tag{38}$$

Combining Equations (33) and (38) gives

$$\operatorname{supp} \pi^*_{\widehat{\theta}}(\cdot \mid X) = \operatorname{supp} \pi_{\theta_0}(\cdot \mid X), \tag{39}$$

which completes the proof. □

B.5 Existence and Finiteness of the Normalizing Constant

Proposition B.3. *Under the definition $V(X,y) = \log s_{\theta_0}(X,y) + F(X,y)$ and Equation (2), suppose $\alpha > 0$ and*

$$\operatorname{supp} \pi_0(\cdot \mid X) \cap \operatorname{supp} \pi_{\theta_0}(\cdot \mid X) \neq \varnothing.$$

Then

$$0 < Z_\alpha(X) \leq e^\alpha < \infty.$$

Proof. By the definition of the value function,

$$\begin{aligned}\exp\{\alpha V(X,y)\} &= \exp\{\alpha\left[\log s_{\theta_0}(X,y) + F(X,y)\right]\} \\ &= \left[s_{\theta_0}(X,y)\right]^{\alpha}\exp\{\alpha F(X,y)\}\,.\end{aligned} \tag{40}$$

Since $0 < s_{\theta_0}(X,y) \le 1$ and $\alpha > 0$, we have

$$0 \le \left[s_{\theta_0}(X,y)\right]^{\alpha} \le 1. \tag{41}$$

Moreover, because $0 \le F(X,y) \le 1$, it follows that

$$1 \le \exp\{\alpha F(X,y)\} \le e^{\alpha}. \tag{42}$$

Therefore,

$$0 \le \left[s_{\theta_0}(X,y)\right]^{\alpha}\exp\{\alpha F(X,y)\} \le e^{\alpha}. \tag{43}$$

Using the definition of $Z_\alpha(X)$ and Equation (40), we obtain

$$\begin{aligned}Z_\alpha(X) &= \sum_{y\in\mathcal{Y}_X}\pi_0(y\mid X)\left[s_{\theta_0}(X,y)\right]^{\alpha}\exp\{\alpha F(X,y)\} \\ &= \mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[\left(s_{\theta_0}(X,y)\right)^{\alpha}\exp\{\alpha F(X,y)\}\right].\end{aligned} \tag{44}$$

Applying the upper bound in Equation (43) gives

$$Z_\alpha(X) \le \sum_{y\in\mathcal{Y}_X}\pi_0(y\mid X)e^{\alpha} = e^{\alpha}\sum_{y\in\mathcal{Y}_X}\pi_0(y\mid X) = e^{\alpha}. \tag{45}$$

Hence,

$$Z_\alpha(X) \le e^{\alpha} < \infty. \tag{46}$$

Finally, by assumption, there exists at least one response $y \in \mathcal{Y}_X$ such that $\pi_0(y \mid X) > 0$ and $s_{\theta_0}(X,y) > 0$. For this response,

$$\pi_0(y\mid X)\left[s_{\theta_0}(X,y)\right]^{\alpha}\exp\{\alpha F(X,y)\} > 0. \tag{47}$$

Since every term in the sum defining $Z_\alpha(X)$ is nonnegative, the existence of this strictly positive term implies

$$Z_\alpha(X) > 0. \tag{48}$$

Combining Equations (46) and (48), we obtain

$$0 < Z_\alpha(X) \le e^{\alpha} < \infty, \tag{49}$$

which completes the proof. □

B.6 Proof of Proposition B.4

Proposition B.4 (Sequence-level approximation stability)**.** *Let ν be any distribution over RAG inputs, and let $\pi^*_\alpha(\cdot \mid X)$ and $\pi^*_{\widehat{\theta}}(\cdot \mid X)$ denote the optimized target LLM and the final learned LLM, respectively. Suppose that*

$$\mathbb{E}_{X\sim\nu}\left[D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot\mid X)\,\|\,\pi^*_{\widehat{\theta}}(\cdot\mid X)\right)\right] \le \varepsilon_{\mathrm{seq}}. \tag{50}$$

For any bounded response-quality functional $g : \mathcal{X}\times\mathcal{Y}\to[0,1]$,

$$\left|\mathbb{E}_{X\sim\nu}\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[g(X,y)\right] - \mathbb{E}_{X\sim\nu}\mathbb{E}_{y\sim\pi^*_{\widehat{\theta}}(\cdot|X)}\left[g(X,y)\right]\right| \le \sqrt{\frac{\varepsilon_{\mathrm{seq}}}{2}}. \tag{51}$$

Proof. The proposition uses the same sequence-level KL appearing in the implicit distillation objective in Equation (3).

For a fixed prompt X, define

$$P_X = \pi^*_\alpha(\cdot\mid X), \qquad Q_X = \pi^*_{\widehat{\theta}}(\cdot\mid X). \tag{52}$$

Since $g(X, y) \in [0, 1]$, the difference between its expectations under P_X and Q_X is bounded by their total variation distance:

$$|\mathbb{E}_{y\sim P_X}[g(X, y)] - \mathbb{E}_{y\sim Q_X}[g(X, y)]| \le D_{\mathrm{TV}}(P_X, Q_X). \tag{53}$$

By Pinsker's inequality,

$$D_{\mathrm{TV}}(P_X, Q_X) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_X \| Q_X)}. \tag{54}$$

Combining Equations (53) and (54) gives

$$\begin{aligned} &\left|\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}[g(X, y)] - \mathbb{E}_{y\sim\pi^*_{\widehat{\theta}}(\cdot|X)}[g(X, y)]\right| \\ &\le \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot \mid X) \,\|\, \pi^*_{\widehat{\theta}}(\cdot \mid X)\right)}. \end{aligned} \tag{55}$$

Taking expectation over $X \sim \nu$ and using the triangle inequality yields

$$\begin{aligned} &\Big|\mathbb{E}_{X\sim\nu}\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}[g(X, y)] \\ &\quad - \mathbb{E}_{X\sim\nu}\mathbb{E}_{y\sim\pi^*_{\widehat{\theta}}(\cdot|X)}[g(X, y)]\Big| \\ &\le \mathbb{E}_{X\sim\nu}\left[\sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot \mid X) \,\|\, \pi^*_{\widehat{\theta}}(\cdot \mid X)\right)}\right]. \end{aligned} \tag{56}$$

Since the square-root function is concave, Jensen's inequality implies

$$\begin{aligned} &\mathbb{E}_{X\sim\nu}\left[\sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot \mid X) \,\|\, \pi^*_{\widehat{\theta}}(\cdot \mid X)\right)}\right] \\ &\le \sqrt{\frac{1}{2}\mathbb{E}_{X\sim\nu}\left[D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot \mid X) \,\|\, \pi^*_{\widehat{\theta}}(\cdot \mid X)\right)\right]} \le \sqrt{\frac{\varepsilon_{\mathrm{seq}}}{2}}, \end{aligned} \tag{57}$$

where the final inequality follows from Equation (50). Therefore,

$$\begin{aligned} &\Big|\mathbb{E}_{X\sim\nu}\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}[g(X, y)] \\ &\quad - \mathbb{E}_{X\sim\nu}\mathbb{E}_{y\sim\pi^*_{\widehat{\theta}}(\cdot|X)}[g(X, y)]\Big| \le \sqrt{\frac{\varepsilon_{\mathrm{seq}}}{2}}, \end{aligned} \tag{58}$$

which completes the proof. □

B.7 Proof of Proposition B.5

Proposition B.5 (Suppression under the optimized target LLM). *Fix a prompt X and define*

$$\mathcal{B}_X(\tau, \eta) = \{y : s_{\theta_0}(X, y) \le \tau,\ F(X, y) \le \eta\}, \tag{59}$$

$$\mathcal{G}_X(\kappa, \gamma) = \{y : s_{\theta_0}(X, y) \ge \kappa,\ F(X, y) \ge \gamma\}. \tag{60}$$

Suppose that $0 < \tau < \kappa \le 1$, $0 \le \eta < \gamma \le 1$, and $\pi_0(\mathcal{G}_X(\kappa, \gamma) \mid X) \ge \rho$ for some $\rho > 0$. Then the target LLM in Equation (2) satisfies

$$\pi^*_\alpha\left(\mathcal{B}_X(\tau, \eta) \mid X\right) \le \frac{1}{\rho}\left(\frac{\tau}{\kappa}\right)^\alpha e^{-\alpha(\gamma-\eta)}. \tag{61}$$

Proof. By Equation (2),

$$\pi^*_\alpha\left(\mathcal{B}_X(\tau, \eta) \mid X\right) = \frac{\sum_{y\in\mathcal{B}_X(\tau,\eta)} \pi_0(y \mid X)\,[s_{\theta_0}(X, y)]^\alpha\, e^{\alpha F(X,y)}}{Z_\alpha(X)}. \tag{62}$$

For every $y \in \mathcal{B}_X(\tau, \eta)$, $s_{\theta_0}(X, y) \le \tau$ and $F(X, y) \le \eta$. Therefore,

$$[s_{\theta_0}(X, y)]^\alpha\, e^{\alpha F(X,y)} \le \tau^\alpha e^{\alpha\eta}. \tag{63}$$

It follows that the numerator of Equation (62) satisfies

$$\begin{aligned}&\sum_{y\in\mathcal{B}_X(\tau,\eta)} \pi_0(y \mid X)\,[s_{\theta_0}(X,y)]^\alpha\, e^{\alpha F(X,y)}\\ &\le \pi_0(\mathcal{B}_X(\tau,\eta)\mid X)\,\tau^\alpha e^{\alpha\eta}.\end{aligned} \tag{64}$$

We next lower-bound the normalizer using only the good-response set. Since every term defining $Z_\alpha(X)$ is nonnegative,

$$\begin{aligned}Z_\alpha(X) &\ge \sum_{y\in\mathcal{G}_X(\kappa,\gamma)} \pi_0(y \mid X)\,[s_{\theta_0}(X,y)]^\alpha\, e^{\alpha F(X,y)}\\ &\ge \kappa^\alpha e^{\alpha\gamma} \sum_{y\in\mathcal{G}_X(\kappa,\gamma)} \pi_0(y \mid X)\\ &= \kappa^\alpha e^{\alpha\gamma}\pi_0(\mathcal{G}_X(\kappa,\gamma)\mid X) \ge \rho\kappa^\alpha e^{\alpha\gamma}.\end{aligned} \tag{65}$$

Combining Equations (64) and (65) gives

$$\begin{aligned}\pi^*_\alpha(\mathcal{B}_X(\tau,\eta)\mid X) &\le \frac{\pi_0(\mathcal{B}_X(\tau,\eta)\mid X)\,\tau^\alpha e^{\alpha\eta}}{\rho\kappa^\alpha e^{\alpha\gamma}}\\ &= \frac{\pi_0(\mathcal{B}_X(\tau,\eta)\mid X)}{\rho}\left(\frac{\tau}{\kappa}\right)^\alpha e^{-\alpha(\gamma-\eta)}.\end{aligned} \tag{66}$$

Since $\pi_0(\mathcal{B}_X(\tau,\eta)\mid X)\le 1$, we conclude that

$$\pi^*_\alpha(\mathcal{B}_X(\tau,\eta)\mid X) \le \frac{1}{\rho}\left(\frac{\tau}{\kappa}\right)^\alpha e^{-\alpha(\gamma-\eta)}, \tag{67}$$

which completes the proof. □

B.8 Proof of Theorem 4.1

Proof. For a fixed prompt X, let $P_X = \pi^*_\alpha(\cdot \mid X)$ and $Q_X = \pi^*_{\hat\theta}(\cdot \mid X)$. Applying the total-variation bound to the indicator of $\mathcal{B}_X(\tau,\eta)$ and then Pinsker's inequality gives

$$\begin{aligned}Q_X(\mathcal{B}_X(\tau,\eta)) &\le P_X(\mathcal{B}_X(\tau,\eta)) + D_{\mathrm{TV}}(P_X,Q_X)\\ &\le P_X(\mathcal{B}_X(\tau,\eta)) + \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_X\|Q_X)}.\end{aligned} \tag{68}$$

Under the good-response mass and threshold assumptions stated in the theorem, Proposition B.5 yields

$$P_X(\mathcal{B}_X(\tau,\eta)) \le \frac{1}{\rho}\left(\frac{\tau}{\kappa}\right)^\alpha e^{-\alpha(\gamma-\eta)}. \tag{69}$$

Taking expectation over $X\sim\nu$ in Equation (68), applying Equation (69), and using Jensen's inequality gives

$$\begin{aligned}&\mathbb{E}_{X\sim\nu}\left[Q_X(\mathcal{B}_X(\tau,\eta))\right]\\ &\le \frac{1}{\rho}\left(\frac{\tau}{\kappa}\right)^\alpha e^{-\alpha(\gamma-\eta)} + \sqrt{\frac{1}{2}\mathbb{E}_{X\sim\nu}\left[D_{\mathrm{KL}}(P_X\|Q_X)\right]}\\ &\le \frac{1}{\rho}\left(\frac{\tau}{\kappa}\right)^\alpha e^{-\alpha(\gamma-\eta)} + \sqrt{\frac{\varepsilon_{\mathrm{seq}}}{2}},\end{aligned} \tag{70}$$

where the final inequality follows from Equation (50). This is exactly Equation (5). □

B.9 Proof of Proposition B.6

Proposition B.6 (α-controlled value–deviation trade-off)**.** *Fix X and $\alpha>0$, and suppose that*

$$0 < \mathbb{E}_{y\sim\pi_0(\cdot|X)}\left[e^{\alpha V(X,y)}\right] < \infty, \qquad \mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[V(X,y)^2\right] < \infty. \tag{71}$$

Then

$$\frac{\partial}{\partial\alpha}\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}[V(X,y)] = \mathrm{Var}_{y\sim\pi^*_\alpha(\cdot|X)}(V(X,y)) \geq 0, \tag{72}$$

and

$$\frac{\partial}{\partial\alpha}D_{\mathrm{KL}}(\pi^*_\alpha(\cdot \mid X)\|\pi_0(\cdot \mid X)) = \alpha\ \mathrm{Var}_{y\sim\pi^*_\alpha(\cdot|X)}(V(X,y)) \geq 0. \tag{73}$$

Thus, increasing α improves expected response value at the cost of greater deviation from the reference LLM.

If $\mathrm{Var}_{y\sim\pi^*_\alpha(\cdot|X)}(V(X,y)) > 0$, then both inequalities are strict for $\alpha > 0$. Hence, α explicitly controls the trade-off between increasing expected value and remaining close to the reference distribution. Setting $\alpha = 1$ applies the value function at its natural scale, whereas $\alpha = 0$ removes value-based reweighting and reduces the objective to standard supervised fine-tuning.

Proof. For a fixed prompt X, define the log-partition function

$$\psi_X(\alpha) = \log Z_\alpha(X). \tag{74}$$

Differentiating the partition function gives

$$Z'_\alpha(X) = \sum_{y\in\mathcal{Y}_X} \pi_0(y \mid X)V(X,y)\exp\{\alpha V(X,y)\}. \tag{75}$$

Therefore,

$$\begin{aligned}\psi'_X(\alpha) &= \frac{Z'_\alpha(X)}{Z_\alpha(X)} \\ &= \sum_{y\in\mathcal{Y}_X} \frac{\pi_0(y \mid X)\exp\{\alpha V(X,y)\}}{Z_\alpha(X)}V(X,y) \\ &= \mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[V(X,y)\right].\end{aligned} \tag{76}$$

Differentiating the partition function a second time gives

$$Z''_\alpha(X) = \sum_{y\in\mathcal{Y}_X} \pi_0(y \mid X)V(X,y)^2\exp\{\alpha V(X,y)\}. \tag{77}$$

Hence,

$$\begin{aligned}\psi''_X(\alpha) &= \frac{Z''_\alpha(X)Z_\alpha(X) - [Z'_\alpha(X)]^2}{Z_\alpha(X)^2} \\ &= \frac{Z''_\alpha(X)}{Z_\alpha(X)} - \left[\frac{Z'_\alpha(X)}{Z_\alpha(X)}\right]^2 \\ &= \mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[V(X,y)^2\right] - \left(\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[V(X,y)\right]\right)^2 \\ &= \mathrm{Var}_{y\sim\pi^*_\alpha(\cdot|X)}\left(V(X,y)\right) \geq 0.\end{aligned} \tag{78}$$

Combining Equations (76) and (78) yields

$$\frac{\partial}{\partial\alpha}\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[V(X,y)\right] = \mathrm{Var}_{y\sim\pi^*_\alpha(\cdot|X)}\left(V(X,y)\right) \geq 0. \tag{79}$$

We next study the divergence from the reference LLM. By Equation (2),

$$\log\frac{\pi^*_\alpha(y \mid X)}{\pi_0(y \mid X)} = \alpha V(X,y) - \log Z_\alpha(X). \tag{80}$$

It follows that

$$\begin{aligned}D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot \mid X)\,\|\,\pi_0(\cdot \mid X)\right) &= \alpha\mathbb{E}_{y\sim\pi^*_\alpha(\cdot|X)}\left[V(X,y)\right] - \log Z_\alpha(X) \\ &= \alpha\psi'_X(\alpha) - \psi_X(\alpha).\end{aligned} \tag{81}$$

Differentiating Equation (81) with respect to α gives

$$\begin{aligned}\frac{\partial}{\partial\alpha}D_{\mathrm{KL}}\left(\pi^*_\alpha(\cdot \mid X)\,\|\,\pi_0(\cdot \mid X)\right) &= \psi'_X(\alpha) + \alpha\psi''_X(\alpha) - \psi'_X(\alpha) \\ &= \alpha\psi''_X(\alpha) \\ &= \alpha\,\mathrm{Var}_{y\sim\pi^*_\alpha(\cdot|X)}\left(V(X,y)\right) \geq 0,\end{aligned} \tag{82}$$

which completes the proof. □

B.10 Rejection-Sampling Interpretation of the Optimized LLM

We show that the conditional LLM induced by the optimized joint distribution can be obtained exactly by applying rejection sampling to responses drawn from the reference LLM. For a fixed prompt X, recall that

$$\pi^*_\alpha(y \mid X) = \frac{\pi_0(y \mid X)\exp\{\alpha V(X,y)\}}{Z_\alpha(X)}, \qquad Z_\alpha(X) = \sum_{y' \in \mathcal{Y}_X} \pi_0(y' \mid X)\exp\{\alpha V(X,y')\}, \tag{83}$$

where

$$V(X,y) = \log s_{\theta_0}(X,y) + F(X,y). \tag{84}$$

Here, $\pi_0(\cdot \mid X)$ is the proposal distribution and $\pi^*_\alpha(\cdot \mid X)$ is the target distribution.

For a fixed prompt X, independently generate

$$Y_1, Y_2, \ldots \overset{\text{i.i.d.}}{\sim} \pi_0(\cdot \mid X), \qquad U_1, U_2, \ldots \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1),$$

where the two sequences are mutually independent. Define the stopping index

$$\sigma = \inf\left\{ n \geq 1 : U_n \leq a_{\text{RS}}(X)\frac{\pi^*_\alpha(Y_n \mid X)}{\pi_0(Y_n \mid X)} \right\}, \tag{85}$$

where $a_{\text{RS}}(X) > 0$ satisfies

$$a_{\text{RS}}(X)\frac{\pi^*_\alpha(y \mid X)}{\pi_0(y \mid X)} \leq 1 \tag{86}$$

for every y such that $\pi^*_\alpha(y \mid X) > 0$. The returned response is Y_σ, the first proposal that passes the acceptance test.

Proposition B.7 (Rejection-sampling equivalence)**.** *Suppose that*

$$0 < Z_\alpha(X) < \infty, \qquad M_\alpha(X) := \sup_{y \in \mathcal{Y}_X} \exp\{\alpha V(X,y)\} < \infty. \tag{87}$$

Choose

$$\begin{aligned} a^*_{\text{RS}}(X) &= \frac{Z_\alpha(X)}{M_\alpha(X)} \\ &= Z_\alpha(X)\exp\left\{ -\alpha \sup_{y \in \mathcal{Y}_X} V(X,y) \right\}. \end{aligned} \tag{88}$$

Then the acceptance probability for a proposed response y is

$$\begin{aligned} A_\alpha(X,y) &= \frac{\exp\{\alpha V(X,y)\}}{M_\alpha(X)} \\ &= \exp\left\{ \alpha \left(V(X,y) - \sup_{y'} V(X,y') \right) \right\}. \end{aligned} \tag{89}$$

Moreover,

$$Y_\sigma \sim \pi^*_\alpha(\cdot \mid X). \tag{90}$$

Proof. For notational simplicity, define

$$f_X(y) = \pi^*_\alpha(y \mid X), \qquad g_X(y) = \pi_0(y \mid X).$$

The rejection-sampling constant must satisfy

$$a_{\text{RS}}(X)\frac{f_X(y)}{g_X(y)} \leq 1$$

for every y such that $f_X(y) > 0$, or equivalently, $a_{\mathrm{RS}}(X) \le g_X(y)/f_X(y)$. Therefore, the largest admissible constant is

$$\begin{aligned} a^*_{\mathrm{RS}}(X) &= \inf_{y: f_X(y)>0} \frac{g_X(y)}{f_X(y)} \\ &= \inf_{y: f_X(y)>0} \frac{\pi_0(y \mid X)}{\pi_0(y \mid X) e^{\alpha V(X,y)}/Z_\alpha(X)} \\ &= \inf_{y: f_X(y)>0} Z_\alpha(X) e^{-\alpha V(X,y)} \\ &= Z_\alpha(X) \exp\left\{-\alpha \sup_y V(X,y)\right\} = \frac{Z_\alpha(X)}{M_\alpha(X)}. \end{aligned} \tag{91}$$

Substituting this constant into the acceptance probability gives

$$\begin{aligned} a^*_{\mathrm{RS}}(X) \frac{f_X(y)}{g_X(y)} &= \frac{Z_\alpha(X)}{M_\alpha(X)} \frac{\pi_0(y \mid X) e^{\alpha V(X,y)}/Z_\alpha(X)}{\pi_0(y \mid X)} \\ &= \frac{e^{\alpha V(X,y)}}{M_\alpha(X)}, \end{aligned} \tag{92}$$

which belongs to $[0,1]$ by the definition of $M_\alpha(X)$.

Let $\mathcal{A}_n$ denote the event that the nth proposal is accepted. For every response y,

$$\Pr(Y_n = y, \mathcal{A}_n \mid X) = \pi_0(y \mid X) \frac{e^{\alpha V(X,y)}}{M_\alpha(X)}. \tag{93}$$

The probability of accepting an arbitrary proposal is

$$\begin{aligned} p_{\mathrm{acc}}(X) &= \sum_{y \in \mathcal{Y}_X} \pi_0(y \mid X) \frac{e^{\alpha V(X,y)}}{M_\alpha(X)} \\ &= \frac{Z_\alpha(X)}{M_\alpha(X)} = a^*_{\mathrm{RS}}(X). \end{aligned} \tag{94}$$

Because the trials are independent, $\{\sigma = n\}$ means that the first $n-1$ proposals were rejected and the nth proposal was accepted. Hence,

$$\begin{aligned} \Pr(Y_\sigma = y \mid X) &= \sum_{n=1}^{\infty} (1 - p_{\mathrm{acc}}(X))^{n-1} \pi_0(y \mid X) \frac{e^{\alpha V(X,y)}}{M_\alpha(X)} \\ &= \frac{\pi_0(y \mid X) e^{\alpha V(X,y)}/M_\alpha(X)}{p_{\mathrm{acc}}(X)} \\ &= \frac{\pi_0(y \mid X) e^{\alpha V(X,y)}}{Z_\alpha(X)} = \pi^*_\alpha(y \mid X). \end{aligned} \tag{95}$$

Therefore, $Y_\sigma \sim \pi^*_\alpha(\cdot \mid X)$. □

A convenient acceptance rule under bounded faithfulness. Under TIDE value function,

$$V(X,y) = \log s_{\theta_0}(X,y) + F(X,y).$$

Since $0 < s_{\theta_0}(X,y) \le 1$ and $0 \le F(X,y) \le 1$, we have

$$V(X,y) \le 1 \qquad \text{and hence} \qquad e^{\alpha V(X,y)} \le e^\alpha. \tag{96}$$

Thus, even when $\sup_y V(X,y)$ is unavailable, we may use the valid upper bound

$$M_{\mathrm{safe}}(\alpha) = e^\alpha.$$

This gives the conservative rejection-sampling constant

$$a^{\mathrm{safe}}_{\mathrm{RS}}(X) = \frac{Z_\alpha(X)}{e^\alpha}. \tag{97}$$

The corresponding acceptance probability is

$$\begin{aligned} A_{\mathrm{safe},\alpha}(X,y) &= a_{\mathrm{RS}}^{\mathrm{safe}}(X)\frac{\pi_\alpha^*(y \mid X)}{\pi_0(y \mid X)} \\ &= \frac{Z_\alpha(X)}{e^\alpha}\frac{\pi_0(y \mid X)e^{\alpha V(X,y)}/Z_\alpha(X)}{\pi_0(y \mid X)} \\ &= \exp\{\alpha\,(V(X,y)-1)\}\,. \end{aligned} \tag{98}$$

Because $V(X,y) \le 1$, this acceptance probability belongs to $[0,1]$. Consequently, sampling $Y \sim \pi_0(\cdot \mid X)$ and accepting it with probability

$$\exp\{\alpha\,(V(X,Y)-1)\}$$

also produces samples from $\pi_\alpha^*(\cdot \mid X)$.

The exact and conservative constants satisfy

$$a_{\mathrm{RS}}^{\mathrm{safe}}(X) \le a_{\mathrm{RS}}^*(X), \tag{99}$$

with equality when $\sup_{y\in\mathcal{Y}_X} V(X,y) = 1$. Thus, $Z_\alpha(X)e^{-\alpha}$ is always a valid rejection-sampling constant under the bound $V(X,y) \le 1$, although it need not be the largest admissible constant.

Remark B.8 (Interpretation of the acceptance probability)**.** *Using the definition of V, the conservative acceptance probability can be written as*

$$\begin{aligned} A_{\mathrm{safe},\alpha}(X,y) &= \exp\{\alpha\,[\log s_{\theta_0}(X,y) + F(X,y) - 1]\} \\ &= [s_{\theta_0}(X,y)]^\alpha \exp\{\alpha(F(X,y)-1)\}\,. \end{aligned} \tag{100}$$

The first factor favors responses with high length-normalized likelihood under the pretrained model, while the second favors responses that are faithful to the retrieved evidence. This provides an accept–reject interpretation of the two components of the value function. The rejection sampler is used only as a theoretical interpretation; TIDE *training algorithm applies the corresponding exponential weight directly and does not generate candidates repeatedly.*

B.11 Candidate Complexity for Obtaining a Good Response

For a fixed prompt X and thresholds $\kappa \in (0,1]$ and $r \in [0,1]$, define the set of good responses

$$\mathcal{G}_X(\kappa,r) = \{y \in \mathcal{Y}_X : s_{\theta_0}(X,y) \ge \kappa,\ F(X,y) \ge r\}\,, \tag{101}$$

and let

$$p_X(\kappa,r) = \pi_0(\mathcal{G}_X(\kappa,r) \mid X)\,. \tag{102}$$

Proposition B.9. *Consider the conservative rejection-sampling rule*

$$A_{\mathrm{safe},\alpha}(X,y) = \exp\{\alpha\,(V(X,y)-1)\}\,, \tag{103}$$

where $V(X,y) = \log s_{\theta_0}(X,y) + F(X,y)$. Every $y \in \mathcal{G}_X(\kappa,r)$ is accepted with probability at least

$$\kappa^\alpha \exp\{\alpha(r-1)\}\,. \tag{104}$$

Consequently, one proposal produces a good accepted response with probability at least

$$q_X := p_X(\kappa,r)\kappa^\alpha \exp\{\alpha(r-1)\}\,. \tag{105}$$

If $q_X \in (0,1)$, then a sufficient condition for obtaining at least one good accepted response among N independent proposals with probability at least $1-\delta$, where $\delta \in (0,1)$, is

$$N \ge \left\lceil \frac{\log \delta}{\log(1-q_X)} \right\rceil. \tag{106}$$

Moreover, it is sufficient to choose

$$N \ge \left\lceil \frac{\log(1/\delta)}{p_X(\kappa,r)\kappa^\alpha e^{\alpha(r-1)}} \right\rceil. \tag{107}$$

Proof. For every $y \in \mathcal{G}_X(\kappa, r)$, $s_{\theta_0}(X, y) \geq \kappa$ and $F(X, y) \geq r$. Therefore,

$$\begin{aligned} A_{\text{safe},\alpha}(X, y) &= \exp\{\alpha \left[\log s_{\theta_0}(X, y) + F(X, y) - 1\right]\} \\ &= \left[s_{\theta_0}(X, y)\right]^{\alpha} e^{\alpha(F(X,y)-1)} \geq \kappa^{\alpha} e^{\alpha(r-1)}. \end{aligned} \tag{108}$$

Since a proposal lies in $\mathcal{G}_X(\kappa, r)$ with probability $p_X(\kappa, r)$, the probability of obtaining a good accepted response from one proposal is at least

$$q_X = p_X(\kappa, r)\kappa^{\alpha} e^{\alpha(r-1)}. \tag{109}$$

For N independent proposals, the probability of obtaining no good accepted response is at most $(1 - q_X)^N$. Hence,

$$\Pr(\text{at least one good accepted response} \mid X) \geq 1 - (1 - q_X)^N. \tag{110}$$

Requiring the right-hand side to be at least $1 - \delta$ gives $(1 - q_X)^N \leq \delta$, which is equivalent to

$$N \geq \frac{\log \delta}{\log(1 - q_X)}. \tag{111}$$

Taking the ceiling proves Equation (106). Finally, using $1 - q_X \leq e^{-q_X}$ gives $(1 - q_X)^N \leq e^{-N q_X}$. Thus, $N \geq \log(1/\delta)/q_X$ is sufficient, yielding Equation (107). □

Remark B.10. *The candidate requirement scales inversely with the reference mass $p_X(\kappa, r)$ and the minimum acceptance probability $\kappa^{\alpha} e^{\alpha(r-1)}$. Thus, more candidates are required when good responses are rare under π_0, low in length-normalized likelihood under the pretrained LLM, or only weakly faithful. This candidate complexity describes the theoretical rejection-sampling interpretation only.* TIDE *training objective applies $e^{\alpha V(X,y)}$ directly to observed prompt–response pairs and does not generate or filter multiple candidates for each training prompt.*

B.12 Comparison with Self-Generated Demonstrations

We compare the rejection-sampling interpretation of TIDE optimized conditional LLM with an idealized rejection-sampling interpretation of Self-Generated Demonstrations (Self-Demo). We fix a prompt X throughout.

Rejection-sampling constant for TIDE method. Define

$$V_{\max}(X) := \sup_{y \in \mathcal{Y}_X} V(X, y), \qquad M_{\alpha}(X) = e^{\alpha V_{\max}(X)}. \tag{112}$$

From Proposition B.7, the maximal rejection-sampling constant for TIDE method is

$$\begin{aligned} a^*_{\text{TIDE}}(X) = \frac{Z_{\alpha}(X)}{M_{\alpha}(X)} &= Z_{\alpha}(X) e^{-\alpha V_{\max}(X)} \\ &= \mathbb{E}_{Y \sim \pi_0(\cdot|X)} \left[\exp\{\alpha \left(V(X, Y) - V_{\max}(X)\right)\}\right]. \end{aligned} \tag{113}$$

Idealized interpretation of Self-Demo. Let G denote the observed human-written response. At the population level, $G \sim \pi_0(\cdot \mid X)$, whereas for an observed training example we condition on $G = g$. Self-Demo generates candidates from

$$q_{\text{SelfDemo}}(\cdot \mid X) = \pi_{\theta_0}(\cdot \mid X). \tag{114}$$

Define the correctness indicator

$$C(X, y, g) = \mathbf{1}\left\{y \text{ is judged correct relative to } g\right\}. \tag{115}$$

The corresponding idealized target is

$$\pi_{\text{SelfDemo}}(y \mid X, g) = \frac{q_{\text{SelfDemo}}(y \mid X) C(X, y, g)}{Z_{\text{SelfDemo}}(X, g)}, \tag{116}$$

where

$$Z_{\text{SelfDemo}}(X,g) = \sum_y q_{\text{SelfDemo}}(y \mid X) C(X,y,g) = \Pr_{Y \sim q_{\text{SelfDemo}}(\cdot|X)} (C(X,Y,g)=1)\,. \tag{117}$$

Since

$$\frac{\pi_{\text{SelfDemo}}(y \mid X,g)}{q_{\text{SelfDemo}}(y \mid X)} = \frac{C(X,y,g)}{Z_{\text{SelfDemo}}(X,g)}, \tag{118}$$

the maximal rejection-sampling constant for Self-Demo is

$$a^*_{\text{SelfDemo}}(X,g) = Z_{\text{SelfDemo}}(X,g) = \Pr_{Y \sim q_{\text{SelfDemo}}(\cdot|X)} (C(X,Y,g)=1)\,. \tag{119}$$

Thus, $a^*_{\text{SelfDemo}}(X,g)$ is precisely the probability that one pretrained-model candidate is judged correct relative to the observed response.

Exact comparison. Equations (113) and (119) imply

$$a^*_{\text{TIDE}}(X) > a^*_{\text{SelfDemo}}(X,g) \iff \mathbb{E}_{Y \sim \pi_0(\cdot|X)} \left[e^{\alpha (V(X,Y) - V_{\max}(X))} \right] > \Pr_{Y \sim q_{\text{SelfDemo}}(\cdot|X)} (C(X,Y,g)=1)\,. \tag{120}$$

A universal constant-factor comparison. Recall the correct-response set $\mathcal{S}_{X,g}$ as

$$\mathcal{S}_{X,g} = \{ y \in \mathcal{Y}_X : y \text{ is judged correct relative to } g \}\,. \tag{121}$$

Under the idealized judge notation used here, membership is equivalently written as $C(X,y,g)=1$.

Assumption 1 (Uniform value gap)**.** *There exists a constant $C \geq 0$ such that, for every X, g, and $y \in \mathcal{S}_{X,g}$,*

$$\alpha \left(V_{\max}(X) - V(X,y) \right) \leq C. \tag{122}$$

Assumption 2 (Uniform correct-mass ratio)**.** *There exists a constant $R > 0$ such that, for every X and g,*

$$\pi_0(\mathcal{S}_{X,g} \mid X) \geq R\, q_{\text{SelfDemo}}(\mathcal{S}_{X,g} \mid X). \tag{123}$$

Proposition B.11 (Universal constant-factor comparison)**.** *Under Assumptions 1 and 2,*

$$a^*_{\text{TIDE}}(X) \geq R e^{-C} a^*_{\text{SelfDemo}}(X,g) \tag{124}$$

for every X and g. Consequently, if $Re^{-C} > 1$ and $a^*_{\text{SelfDemo}}(X,g) > 0$, then $a^*_{\text{TIDE}}(X) > a^*_{\text{SelfDemo}}(X,g)$.

Proof. Restricting the sum defining $a^*_{\text{TIDE}}(X)$ to $\mathcal{S}_{X,g}$ and applying Assumption 1 gives

$$a^*_{\text{TIDE}}(X) \geq \sum_{y \in \mathcal{S}_{X,g}} \pi_0(y \mid X) e^{\alpha (V(X,y) - V_{\max}(X))} \geq e^{-C} \pi_0(\mathcal{S}_{X,g} \mid X). \tag{125}$$

By Assumption 2,

$$a^*_{\text{TIDE}}(X) \geq R e^{-C} q_{\text{SelfDemo}}(\mathcal{S}_{X,g} \mid X) = R e^{-C} a^*_{\text{SelfDemo}}(X,g), \tag{126}$$

which proves the result. □

Candidate-complexity comparison. For a rejection sampler with one-proposal acceptance probability $a \in (0, 1)$, define

$$N_{\min}(a, \delta) := \left\lceil \frac{\log \delta}{\log(1 - a)} \right\rceil, \qquad \delta \in (0, 1). \tag{127}$$

This is the minimum number of independent proposals required to obtain at least one accepted response with probability at least $1 - \delta$. Hence,

$$\begin{aligned} N_{\mathrm{TIDE}}(X, \delta) &:= N_{\min}(a^*_{\mathrm{TIDE}}(X), \delta)\,, \\ N_{\mathrm{SelfDemo}}(X, g, \delta) &:= N_{\min}(a^*_{\mathrm{SelfDemo}}(X, g), \delta)\,. \end{aligned} \tag{128}$$

Since $N_{\min}(a, \delta)$ is decreasing in a,

$$a^*_{\mathrm{TIDE}}(X) > a^*_{\mathrm{SelfDemo}}(X, g) \quad \Longrightarrow \quad N_{\mathrm{TIDE}}(X, \delta) \le N_{\mathrm{SelfDemo}}(X, g, \delta). \tag{129}$$

Remark B.12. *This comparison concerns the idealized rejection-sampling interpretations of the two methods. Self-Demo explicitly generates and filters candidates, whereas* TIDE *implemented objective directly weights observed prompt–response pairs by $e^{\alpha V(X,y)}$. Thus,* TIDE *training algorithm does not incur the candidate-generation cost described by $N_{\mathrm{TIDE}}(X, \delta)$.*

C Experimental Setup

C.1 Datasets and Evaluation Splits

We train TIDE on two source datasets. The MSMARCO-trained models use 75,000 examples from the MS MARCO training split. The NewsQA-trained models use 71,719 examples from the NewsQA training split. MSMARCO contains short-answer annotations grounded in retrieved passages, whereas NewsQA contains question–answer pairs over news articles with extractive answer spans.

We evaluate on MSMARCO, HotpotQA, Natural Questions (NQ), Wizard of Wikipedia (WoW), and T-REx. Table 4 reports the evaluation splits and sizes.

Table 4: Evaluation datasets, splits, and numbers of evaluated examples.

Benchmark	Split and construction	Size
MSMARCO	`microsoft/ms_marco`, version 1.1, validation; rows with a selected passage and reference answer after filtering	6,789
HotpotQA	`hotpotqa/hotpot_qa`, distractor configuration, validation	7,000
NQ	DPR Natural Questions test split with top-1 retrieved context	3,610
WoW	Wizard of Wikipedia validation	7,866
T-REx	T-REx validation	7,000

MSMARCO measures in-domain passage QA performance. HotpotQA and NQ evaluate transfer to multi-hop and open-domain short-answer QA. WoW evaluates grounded conversational generation, and T-REx evaluates factual entity–value extraction from retrieved evidence.

C.2 Accurate-Retrieval Construction

The accurate-retrieval and answerability checks are constructed from the retrieved evidence and reference answers, not from predictions made by an evaluated model. For MSMARCO, we use the passage marked with `is_selected == 1`. For HotpotQA, we use the resolved supporting-fact sentences provided by the dataset. For NQ, we use the top-1 DPR-retrieved passage and gate correctness on whether a normalized reference answer appears in or is supported by that passage. For WoW, we use the provided dialogue and knowledge evidence; scoring emphasizes grounded relevance rather than exact span matching. For T-REx, we use the provided relation and evidence context and check predictions against the reference object answers.

C.3 Generation and Judge Decoding

Final evaluation generation uses deterministic greedy decoding with `do_sample=False`; temperature and top-p are therefore inactive. Table 5 gives the dataset-specific output limits.

Dataset	Maximum new tokens	Strategy
MSMARCO	32	Greedy (`do_sample=False`)
HotpotQA	24	Greedy (`do_sample=False`)
NQ	32	Greedy (`do_sample=False`)
WoW	64	Greedy (`do_sample=False`)
T-REx	16	Greedy (`do_sample=False`)

Table 5: Decoding settings used for final evaluation generation.

Judge decoding is also greedy. The Qwen3-32B judge uses `max_new_tokens=192` for MSMARCO, HotpotQA, NQ, and WoW, and `max_new_tokens=160` for T-REx.

D Implementation Details

Table 6 summarizes the implementation details used across all training and evaluation settings. The table includes the base LLM model, judge model, optimization configuration, distributed training setup, batch size, number of epochs, and the value-function components used by TIDE. We use the same optimizer, learning rate, batch size, and number of training epochs for all trainable methods to ensure a fair comparison. Unlike methods that train a separate reward network, TIDE uses a fixed value function composed of the base-model compatibility term and a faithfulness score, while evaluation is performed with a fixed Qwen3-32B judge. Our code is provided at tide-research-2027.

Component	Value
Base LLM model	Qwen3-14B
Judge model	Qwen3-32B
Training method	Full-parameter fine-tuning
Distributed training	DeepSpeed ZeRO-3
Optimizer	AdamW 8-bit
Learning rate	5×10^{-6}
Warmup ratio	0.03
Epochs	1
Per-device batch size	1
Gradient accumulation steps	4
Number of GPUs	8
Effective batch size	32
MSMARCO training examples	75,000
NewsQA training examples	71,719
TIDE α	0.5
Faithfulness scorer in Training	MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli
Evaluation datasets	MSMARCO, HotpotQA, NQ, WoW, T-REx

Table 6: Implementation details for training and evaluation.

D.1 Self-Demo

Self-Demonstration follows an SD-RA-IT/RAM-style generate, score, select, and fine-tune pipeline. Starting from `Qwen/Qwen3-14B`, it generates self-demonstration candidate answers using prompt optimization and scores and selects only model-generated candidates. No reference-answer fallback is used as a training target. Strict target filtering requires a minimum reference F1 of 0.70, minimum reference precision of 0.50, and minimum support recall of 0.75. We use four retrieved documents (`ndocs=4`), with maximum answer lengths of 14 words for MSMARCO and 16 words for NewsQA. The base model is then trained on the selected self-demonstrations.

D.2 QWEN3-4B SCALING EXPERIMENTS

The smaller-scale experiments use `Qwen/Qwen3-4B` as the base and trainable LLM and `Qwen/Qwen3-8B` as the fixed validation judge. All trainable methods use full-parameter fine-tuning without LoRA, QLoRA, or other adapters. We use AdamW 8-bit with a learning rate of 5×10^{-6}, a warmup ratio of 0.03, and one training epoch. The per-device batch size is 1, gradient accumulation is 8, and the executable launcher requests four GPUs, giving an effective batch size of 32.

Training Efficiency. We additionally compare the end-to-end runtime of the training methods on MSMARCO. As shown in Table 7, TIDE requires 6:39:16, compared with 6:01:38 for standard SFT and 11:04:05 for Self-Demo. The reported Self-Demo runtime includes both candidate-generation and model-training stages, since candidate generation is an essential part of its training procedure. TIDE is only moderately more expensive than standard SFT because it requires computing the fixed response-level value weights, while avoiding the repeated candidate generation and selection required by Self-Demo. Consequently, TIDE reduces the total runtime of Self-Demo by approximately 40% in this experimental setting.

E EVALUATION PROMPTS

E.1 ANSWER-QUALITY JUDGE PROMPT

The answer-quality evaluator used the following prompt template, where `QUESTION`, `PASSAGE`, `REFERENCE_ANSWERS`, and `MODEL_ANSWER` are replaced by the corresponding example fields.

```
You are a strict fixed judge for retrieval-augmented question answering.
You receive QUESTION, CONTEXT, REFERENCE_ANSWERS, and MODEL_ANSWER.
Use only the supplied information; do not use outside knowledge.

Set correct=true only when MODEL_ANSWER gives the intended answer expressed by
REFERENCE_ANSWERS. Allow aliases, paraphrases, abbreviations, and harmless
formatting differences. Set correct=false for a refusal, wrong entity, partial
answer that misses essential information, contradiction, prompt copying, or
unsupported extra factual detail that materially changes the answer.

Set supported=true only when every factual claim in MODEL_ANSWER is directly
stated by or clearly entailed by CONTEXT. A context-copied sentence is not
automatically correct. Prefer false when uncertain.

Return ONLY valid JSON with exactly these keys:
{"correct": true, "supported": true, "reason": "short reason"}
```

E.2 FAITHFULNESS JUDGE PROMPT

For supported-faithfulness evaluation, we used Qwen3-32B with the following chat-formatted system and user messages. The reference response is supplied only for orientation and is explicitly excluded from the faithfulness decision.

System message.

```
You are a strict context-faithfulness judge for retrieval-augmented generation. An answer
↪  is faithful only if every factual claim in it is directly stated by, or clearly
↪  entailed by, the provided context. Ignore reference-answer correctness. Return JSON
↪  only.
```

User message.

```

Judge only faithfulness to CONTEXT. Do not mark unsupported just because the answer differs
↪  from the reference; mark unsupported only when the answer adds factual content not
↪  entailed by the context. For conversation, allow harmless social wording, but factual
↪  claims still need context support.
```

Table 7: End-to-end runtime on the MSMARCO training set. Self-Demo runtime includes candidate generation, evaluation, selection, and model training. All runtimes are reported in hours:minutes:seconds.

Method	Total Runtime
SFT	**6:01:38**
TIDE	6:39:16
Self-Demo	11:04:05

```
Return exactly one JSON object: {"faithful": true/false, "reason": "short reason",
↪  "unsupported_claims": ["..."]}.
```