

893 A Limitations and Impact Statement

894 **Limitations.** While AceRAG demonstrates strong empirical performance across a wide range of
 895 reasoning-intensive RAG benchmarks, several limitations remain. First, our framework is evaluated
 896 primarily on complex QA, fact verification, and document-level reasoning tasks; its applicability to
 897 other RAG tasks such as open-ended generation, dialogue, or use of real-time tools remains to be
 898 explored, though our scope is comparable (or even broader) compared to concurrent works [68, 33, 94].
 899 Second, AceRAG relies on a *fixed* retriever during training and inference. Joint optimization of
 900 retrieval and reasoning could offer further gains but is left for future work. Third, our decomposition-
 901 based pipeline introduces inference overhead, which may limit applicability in latency-sensitive
 902 settings. Nonetheless, as shown in Figure 3(c), AceRAG achieves favorable tradeoffs, and many
 903 strong baselines [76, 39, 77] also adopt multi-turn retrieval. Finally, due to resource constraints, we
 904 adopt iterative preference optimization (Online DPO) as a practical and efficient alternative to fully
 905 online reinforcement learning. While this approach achieves strong results in our setting, exploring
 906 more expressive or theoretically grounded RL formulations may offer further improvements.

907 **Impact Statement.** This work advances the development of retrieval-augmented systems capable of
 908 complex reasoning. By enabling smaller open-source LLMs to perform multi-hop reasoning more
 909 effectively, AceRAG reduces reliance on proprietary or extremely large models, which may have high
 910 computational or financial barriers. This can promote accessibility and democratization of advanced
 911 AI capabilities in low-resource or domain-specific applications, such as finance, scientific discovery,
 912 and healthcare.

913 B Derivation Step for Optimal Policy π^* and ρ^*

914 We aim to maximize the following objective:

$$\mathcal{J}_\theta = \mathbb{E}_q \left[\mathbb{E}_{z \sim \rho_\theta(\cdot|q), a \sim \pi_\theta(\cdot|q,z)} [r(q, a)] - \beta \mathcal{D}_{\text{KL}}(\rho_\theta \| \rho_{\text{ref}}) - \beta \mathbb{E}_{z \sim \rho_\theta(\cdot|q)} [\mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})] \right].$$

915 Since ρ and π appear in separate terms, we can optimize them independently.

916 **1. Optimal π for each z .** For fixed z , consider the Lagrangian

$$\mathcal{L}_z(\pi, \lambda_z) = \sum_{w, a'} \pi(w, a' | q, z) r(q, a', a) - \beta \sum_{w, a'} \pi(w, a' | q, z) \ln \frac{\pi(w, a' | q, z)}{\pi_{\text{ref}}(w, a' | q, z)} \quad (\text{B.1})$$

$$+ \lambda_z \left(\sum_{w, a'} \pi(w, a' | q, z) - 1 \right). \quad (\text{B.2})$$

917 Taking the functional derivative with respect to $\pi(w, a' | q, z)$ and setting to zero gives

$$r(q, a', a) - \beta \left(\ln \pi(w, a' | q, z) - \ln \pi_{\text{ref}}(w, a' | q, z) + 1 \right) + \lambda_z = 0.$$

918 Rearranging yields

$$\ln \pi(w, a' | q, z) = \ln \pi_{\text{ref}}(w, a' | q, z) + \frac{1}{\beta} r(q, a', a) + \underbrace{\left(\frac{\lambda_z}{\beta} - 1 \right)}_{\text{constant in } w, a'}.$$

919 Hence

$$\pi^*(w, a' | q, z) \propto \pi_{\text{ref}}(w, a' | q, z) \exp\left(\frac{1}{\beta} r(q, a', a)\right).$$

920 **2. Optimal ρ .** Substitute π^* back into \mathcal{J}_q . Denote

$$g(q, z) = \mathbb{E}_{(w, a') \sim \pi^*(\cdot|q, z)} [r(q, a', a)],$$

921 so that the remaining functional in ρ is

$$\sum_z \rho(z | q) g(q, z) - \beta \sum_z \rho(z | q) \ln \frac{\rho(z | q)}{\rho_{\text{ref}}(z | q)}$$

922 together with the constraint $\sum_z \rho(z|q) = 1$. Introduce multiplier μ and form

$$\mathcal{L}[\rho, \mu] = \sum_z \rho(z|q) g(q, z) - \beta \sum_z \rho(z|q) \ln \frac{\rho(z|q)}{\rho_{\text{ref}}(z|q)} + \mu \left(\sum_z \rho(z|q) - 1 \right).$$

923 Taking $\partial \mathcal{L} / \partial \rho(z|q) = 0$ gives

$$G(q, z) - \beta (\ln \rho(z|q) - \ln \rho_{\text{ref}}(z|q) + 1) + \mu = 0,$$

924 so

$$\ln \rho(z|q) = \ln \rho_{\text{ref}}(z|q) + \frac{1}{\beta} g(q, z) + \underbrace{\left(\frac{\mu}{\beta} - 1 \right)}_{\text{constant in } z}.$$

925 Thus

$$\rho^*(z|q) \propto \rho_{\text{ref}}(z|q) \exp \left(\frac{1}{\beta} \mathbb{E}_{(w, a') \sim \pi^*(\cdot|q, z)} [r(q, a', a)] \right).$$

926 Combining these two results yields exactly the stated closed-form solutions

$$p^*(z|q) \propto p_{\text{ref}}(z|q) \exp \left(\frac{1}{\beta} \mathbb{E}_{(w, a') \sim \pi_{\theta}(\cdot|q, z)} [r(q, a', a)] \right), \quad (\text{B.3})$$

$$p^*(w, a' | q, z) \propto p_{\text{ref}}(w, a' | q, z) \exp \left(\frac{1}{\beta} r(q, a', a) \right). \quad (\text{B.4})$$

927 C Omitted Theorems and Proofs

928 C.1 Notion

929 Let $B(r, x)$ represent the l_2 -ball of radius r centered at x . For two positive sequences $\{a_n\}$ and $\{b_n\}$,
 930 $a_n \gtrsim b_n$ if $a_n \geq C b_n$. The l_2 norm of a vector $x \in \mathbb{R}^d$ is defined as $\|x\|_2 := \left(\sum_{i=1}^d x_i^2 \right)^{1/2}$. A
 931 sequence of random variables X_n is said to be $o_P(1)$ if $X_n \xrightarrow{P} 0$, that is, X_n converges to 0 in prob-
 932 ability as $n \rightarrow \infty$. The Kullback–Leibler (KL) divergence from a discrete distribution p to a discrete
 933 distribution q (defined over a common support \mathcal{X}) is given by $\mathbb{D}_{\text{KL}}[p \| q] := \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right)$,
 934 assuming $p(x) > 0$ implies $q(x) > 0$ for all $x \in \mathcal{X}$. provided that p is absolutely continuous with
 935 respect to q .

936 C.2 Main theorem

937 Recall the losses (4.2) and (4.5) are defined as follows:

$$\mathbb{E}_q \left[\mathbb{E}_{z \sim \rho_{\theta}, (w, a') \sim \pi_{\theta}} [r(a', q, a)] - \beta \mathbb{D}_{\text{KL}}[u_{\theta}(a', z, w | q) \| u_{\text{ref}}(a', z, w | q)] \right]. \quad (\text{C.1})$$

938

$$\mathcal{L}_{\text{mDPO}} := -\mathbb{E}_{(x, g^+, g^-) \sim \mathcal{D}_{\text{pref}}^{(t)}} \log \sigma \left(\beta \left[\log \frac{p_{\theta}^{(t+1)}(g^+ | x)}{p_{\theta}^{(t)}(g^+ | x)} - \log \frac{p_{\theta}^{(t+1)}(g^- | x)}{p_{\theta}^{(t)}(g^- | x)} \right] \right). \quad (\text{C.2})$$

939 To enable decomposition into a decomposer and a solver, we require the following assumption:
assumption C.1 (Conditional Probability decomposition). We assume the following decomposition
 holds:

$$p_{\theta}(a | q) = \sum_z p_{\theta}(z | q) \left(\sum_w p_{\theta}(a | q, z, w) p_{\theta}(w | q, z) \right)$$

940 We present the informal version of our theorem below. Formal statements are given in Theorems C.2
 941 and C.3.

942 **Theorem C.1** (Informal). Under regularity conditions, with high probability, the minimizer of the
 943 loss (C.2) at step t is close to the minimizer of the loss (C.1). Furthermore, as t increases, the
 944 minimizer converges to the true parameter θ^* .

945 *Remark C.1.* The main theorem can be divided into two components. The first component establishes
 946 the equivalence between loss (C.1) and loss (C.2) are equivalent. The second component shows that,
 947 once the equivalence is established and the maximizer of loss (C.1) converges, the minimizer of
 948 loss (C.2) also converges.

949 The proof is organized as follows: In Appendices C.3 and C.4, we analyze the convergence properties
 950 of the maximizer of the population version loss (C.3) and sample version of loss (C.6) which
 951 corresponds exactly to loss (C.1). In Appendix C.5, we demonstrate the equivalence of loss (C.1)
 952 to loss (C.2). Finally, in Appendix C.6, building on these results, we prove that the minimizer of
 953 loss (C.2) converges as well.

954 C.3 Population Version

955 Based on the loss (C.1), define the population version loss as

$$L(\theta \mid \theta_{t-1}) = \mathbb{E}_{(q,a) \sim p_{\theta^*}(\cdot)} \left[\mathbb{E}_{z \sim \rho_{\theta}(\cdot \mid q)} \left[\mathbb{E}_{(w,a') \sim \pi_{\theta}(\cdot \mid z,q)} [r(a', q, a)] \right] \right] \\ - \beta \mathbb{D}_{\text{KL}}(u_{\theta}(a', z, w \mid q) \parallel u_{\theta_{t-1}}(a', z, w \mid q)). \quad (\text{C.3})$$

956 Define the operator $M : \Theta \rightarrow \Theta$,

$$M(\theta) = \arg \max_{\theta' \in \Theta} L(\theta' \mid \theta),$$

957 where Θ represents the parameter space. Notice that it is natural to assume that θ^* satisfy the
 958 self-consistency, i.e. $\theta^* = M(\theta^*)$. So the first assumption will be:

959 *assumption C.2* (Self-consistency). $\theta^* = M(\theta^*)$.

960 *assumption C.3* (λ -strong Concavity). There is some $\lambda > 0$ such that

$$L(\theta_1 \mid \theta^*) - L(\theta_2 \mid \theta^*) - \langle \nabla L(\theta_2 \mid \theta^*), \theta_1 - \theta_2 \rangle \leq -\frac{\lambda}{2} \|\theta_1 - \theta_2\|_2^2 \quad \text{for all } \theta_1, \theta_2 \in B(r, \theta^*). \quad (\text{C.4})$$

961 **Definition C.1** (First-order stability). The functions $\{L(\cdot \mid \theta), \theta \in \Theta\}$ satisfy the First-order stability
 962 condition over $B(r, \theta^*)$ if

$$\|\nabla L(M(\theta) \mid \theta^*) - \nabla L(M(\theta) \mid \theta)\|_2 \leq \mu \|\theta - \theta^*\|_2$$

963 for all $\theta \in B(r, \theta^*)$.

964 *assumption C.4.* Assume the functions $\{L(\cdot \mid \theta), \theta \in \Theta\}$ satisfy the First-order stability condition
 965 (C.1).

966 **Proposition C.1** (Population Version). For some radius $r > 0$ and pair (μ, λ) such that $0 \leq \mu < \lambda$,
 967 suppose that the Assumption C.1-C.4 hold, then the population operator M is contractive over
 968 $B(r, \theta^*)$, in particular with

$$\|M(\theta_{t-1}) - \theta^*\|_2 \leq \frac{\mu}{\lambda} \|\theta_{t-1} - \theta^*\|_2 \quad \text{for all } \theta \in B(r, \theta^*).$$

969 *Proof.* By the first order optimality condition, we have:

$$\begin{aligned} \langle \nabla L(\theta^* \mid \theta^*), \theta - \theta^* \rangle &\leq 0 \quad \forall \theta \\ \Rightarrow \langle \nabla L(\theta^* \mid \theta^*), M(\theta_{t-1}) - \theta^* \rangle &\leq 0 \\ \langle \nabla L(M(\theta_{t-1}) \mid \theta_{t-1}), \theta - M(\theta_{t-1}) \rangle &\leq 0 \quad \forall \theta \\ \Rightarrow \langle \nabla L(M(\theta_{t-1}) \mid \theta_{t-1}), \theta^* - M(\theta_{t-1}) \rangle &\leq 0. \end{aligned}$$

970 Combine the two terms,

$$\langle \nabla L(\theta^* \mid \theta^*) - \nabla L(M(\theta_{t-1}) \mid \theta_{t-1}), M(\theta_{t-1}) - \theta^* \rangle \leq 0.$$

971 Thus,

$$\begin{aligned} \langle \nabla L(\theta^* \mid \theta^*) - \nabla L(M(\theta_{t-1}) \mid \theta^*), M(\theta_{t-1}) - \theta^* \rangle &\leq \\ - \langle \nabla L(M(\theta_{t-1}) \mid \theta^*) - \nabla L(M(\theta_{t-1}) \mid \theta_{t-1}), M(\theta_{t-1}) - \theta^* \rangle. \end{aligned}$$

972 For the right-hand side, by Cauchy-Schwarz inequality,

$$-\langle \nabla L(M(\theta_{t-1}) \mid \theta^*) - \nabla L(M(\theta_{t-1}) \mid \theta_{t-1}), M(\theta_{t-1}) - \theta^* \rangle \leq \|\nabla L(M(\theta_{t-1}) \mid \theta^*) - \nabla L(M(\theta_{t-1}) \mid \theta_{t-1})\|_2 \|M(\theta_{t-1}) - \theta^*\|_2.$$

973 By Assumption C.4,

$$\|\nabla L(M(\theta_{t-1}) \mid \theta^*) - \nabla L(M(\theta_{t-1}) \mid \theta_{t-1})\|_2 \|M(\theta_{t-1}) - \theta^*\|_2 \leq \mu \|M(\theta_{t-1}) - \theta^*\|_2^2.$$

974 For the left-hand side, by Assumption C.3,

$$\begin{aligned} \frac{\lambda}{2} \|M(\theta_{t-1}) - \theta^*\|_2^2 &\leq L(\theta^* \mid \theta^*) - L(M(\theta_{t-1}) \mid \theta^*) + \langle \nabla L(\theta^* \mid \theta^*), M(\theta_{t-1}) - \theta^* \rangle, \\ \frac{\lambda}{2} \|M(\theta_{t-1}) - \theta^*\|_2^2 &\leq L(M(\theta_{t-1}) \mid \theta^*) - L(\theta^* \mid \theta^*) + \langle \nabla L(M(\theta_{t-1}) \mid \theta^*), \theta^* - M(\theta_{t-1}) \rangle. \end{aligned}$$

975 Hence,

$$\lambda \|M(\theta_{t-1}) - \theta^*\|_2^2 \leq \langle \nabla L(\theta^* \mid \theta^*) - \nabla L(M(\theta_{t-1}) \mid \theta^*), M(\theta_{t-1}) - \theta^* \rangle. \quad (\text{C.5})$$

976 Combining all,

$$\lambda \|M(\theta_{t-1}) - \theta^*\|_2^2 \leq \mu \|M(\theta_{t-1}) - \theta^*\|_2^2.$$

977

□

978 *Remark C.2.* This theorem follows the idea in [3]. It suggests that, under a self-play procedure, the
979 algorithm progressively approaches the true underlying distribution. This behavior is characterized
980 by a contraction parameter $\frac{\mu}{\lambda}$, which ensures convergence toward the ground-truth parameter θ^* . The
981 incorporation of an intermediate reasoning step smooths the local optimization landscape, rendering
982 the loss approximately convex and thereby facilitating convergence to the global optimum.

983 C.4 Sample Version

984 We define the below sample version: assume we have the data

$$\mathcal{D}_{q,a} = \{q_i, a_i\}_{i=1}^N.$$

985 The loss will be:

$$\begin{aligned} L_N(\theta \mid \theta_{t-1}) &= \mathbb{E}_{(q,a) \sim \tilde{p}_{\theta^*}(\cdot)} \left[\mathbb{E}_{z \sim \rho_{\theta}(\cdot \mid q)} \left[\mathbb{E}_{(w,a') \sim \pi_{\theta}(\cdot \mid z,q)} [r(a', q, a)] \right] \right. \\ &\quad \left. - \beta \mathbb{D}_{\text{KL}}(u_{\theta}(a', z, w \mid q) \parallel u_{\theta_{t-1}}(a', z, w \mid q)) \right], \quad (\text{C.6}) \end{aligned}$$

986 where \tilde{p} represents the empirical distribution defined as

$$\tilde{p}(q, a) = \frac{1}{N} \sum_{i=1}^N 1\{(q, a) = (q_i, a_i)\}.$$

987 We also have the similar convergence property. Similar to the population version, we define the
988 sample-based operator $M_N : \Theta \rightarrow \Theta$,

$$M_N(\theta) = \arg \max_{\theta' \in \Theta} L_N(\theta' \mid \theta).$$

989 For a given sample size N and tolerance parameter $\epsilon \in (0, 1)$, define $\zeta_M^{\text{unif}}(N, \epsilon)$ as the smallest scalar
990 such that

$$\sup_{\theta \in B_2(r; \theta^*)} \|M_N(\theta) - M(\theta)\|_2 \leq \zeta_M^{\text{unif}}(N, \epsilon) \quad (\text{C.7})$$

991 with probability at least $1 - \epsilon$.

992 **Proposition C.2** (Sample Version). Suppose that for all $\theta \in B(r, \theta^*)$, the mapping M satisfies

$$\|M(\theta_{t-1}) - \theta^*\|_2 \leq \frac{\mu}{\lambda} \|\theta_{t-1} - \theta^*\|_2$$

993 with probability at least $1 - \epsilon$. Then we have

$$\|M_N(\theta_{t-1}) - \theta^*\|_2 \leq \frac{\mu}{\lambda} \|\theta_{t-1} - \theta^*\|_2 + \zeta_M^{\text{unif}}(N, \epsilon), \quad \text{for all } \theta \in B(r, \theta^*)$$

994 with probability at least $1 - \epsilon$.

995 *Proof.* The result follows directly from the triangle inequality:

$$\begin{aligned} \|M_N(\theta_{t-1}) - \theta^*\|_2 &\leq \|M(\theta_{t-1}) - \theta^*\|_2 + \|M_N(\theta_{t-1}) - M(\theta_{t-1})\|_2 \\ &\leq \frac{\mu}{\lambda} \|\theta_{t-1} - \theta^*\|_2 + \zeta_M^{\text{unif}}(N, \epsilon). \end{aligned}$$

996 □

997 C.5 On the Equivalence with DPO

998 In the deterministic setting - where m or m' is fixed and the responses with
 999 the maximum and minimum rewards are selected - depending on the data tuple
 1000 $((a^{\max}, z^{\max}, w^{\max}), (a^{\min}, z^{\min}, w^{\min}), a, q)$, we note that in practice the construction of
 1001 positive and negative samples can vary, some containing z or (z, w) , and others including full triples
 1002 such as (a', z, w) . For simplicity, we unify the representation and consider the minimal component
 1003 shared across all formats, namely the tuple (a', z, w) . This process can thus be interpreted as
 1004 observing a finite dataset:

$$\mathcal{D} = \{((a_i^+, z_i^+, w_i^+), (a_i^-, z_i^-, w_i^-), a_i, q_i)\}_{i=1}^N,$$

1005 Then the DPO loss will be:

$$\begin{aligned} L_{\text{mDPO}}(\theta \mid \theta_{t-1}) &= -\mathbb{E}_{((a^+, z^+, w^+), (a^-, z^-, w^-), a, q) \sim \mathcal{D}} \\ &\log \sigma \left(\beta_{\text{mDPO}} \left[\log \frac{u_\theta(a^+, z^+, w^+ \mid q)}{u_{\theta_{t-1}}(a^+, z^+, w^+ \mid q)} - \log \frac{u_\theta(a^-, z^-, w^- \mid q)}{u_{\theta_{t-1}}(a^-, z^-, w^- \mid q)} \right] \right) \quad (\text{C.8}) \end{aligned}$$

1006 To demonstrate the closeness between the loss (C.6) and the loss (C.8), we first show that, with high
 1007 probability, optimizing the loss (C.8) over the dataset \mathcal{D} is equivalent to maximizing the original
 1008 reward up to a scaling factor.

1009 Specifically, we can derive a closed-form solution for Equation (C.6) at step t :

$$u_{\theta_t^*}(a', z, w \mid q) \propto u_{\theta_{t-1}}(a', z, w \mid q) \exp \left(\frac{1}{\beta} r(a', q, a) \right), \quad (\text{C.9})$$

1010 where θ_t^* denotes the ground truth parameter at step t . Accordingly, the reward function $r(a', q, a)$ can
 1011 be written as $r_{\gamma_t^*}(a', q, a)$ to emphasize its dependence on the true reward parameter γ_t^* . Specifically,
 1012 consider the dataset \mathcal{D} , which follows the following deterministic model:

$$\mathbb{P}((a^+, z^+, w^+) \succ (a^-, z^-, w^-) \mid q) = 1 \quad \text{if} \quad r_{\gamma_t^*}(a_i^+, q, a) > r_{\gamma_t^*}(a_i^-, q, a), \quad (\text{C.10})$$

1013 indicating that we always select a_i^+ as the positive sample. To approximate this deterministic behavior,
 1014 we introduce the α -BT model:

$$\mathbb{P}((a^+, z^+, w^+) \succ (a^-, z^-, w^-) \mid q) = \frac{e^{\alpha r_{\gamma_t^*}(a^+, q, a)}}{e^{\alpha r_{\gamma_t^*}(a^+, q, a)} + e^{\alpha r_{\gamma_t^*}(a^-, q, a)}}. \quad (\text{C.11})$$

1015 As $\alpha \rightarrow \infty$, the α -BT model becomes close to the deterministic model (C.10). Then given the above
 1016 dataset \mathcal{D} , we define the following data set

$$\mathcal{D}_\alpha = \{((a_{\alpha, i}^+, z_{\alpha, i}^+, w_{\alpha, i}^+), (a_{\alpha, i}^-, z_{\alpha, i}^-, w_{\alpha, i}^-), a_i, q_i)\}_{i=1}^n,$$

1017 where $((a_{\alpha, i}^+, z_{\alpha, i}^+, w_{\alpha, i}^+), (a_{\alpha, i}^-, z_{\alpha, i}^-, w_{\alpha, i}^-))$ is generated according to the α -BT model (C.11). To
 1018 ensure the closeness between the dataset \mathcal{D} and \mathcal{D}_α , we have the following lemma:

1019 *assumption C.5* (Reward Separation Condition). Assume that given (q, a) , for any
 1020 $((a^+, z^+, w^+), (a^-, z^-, w^-))$, there exists δ such that $|r_{\gamma_t^*}(a^+, q, a) - r_{\gamma_t^*}(a^-, q, a)| \geq \delta$.

1021 **Lemma C.1.** Suppose the Assumption C.5 holds, given ϵ , there exists $\alpha_0 \gtrsim \frac{\log \frac{N}{2\epsilon}}{\delta}$,

$$\mathbb{P}(\mathcal{D} = \mathcal{D}_{\alpha_0}) \geq 1 - \frac{\epsilon}{2}. \quad (\text{C.12})$$

1022 *Proof.* We start by bounding the probability of disagreement between two actions:

$$\mathbb{P}((a_{\alpha,i}^+, z_{\alpha,i}^+, w_{\alpha,i}^+) \neq (a_i^+, z_i^+, w_i^+)) = \frac{e^{\alpha r_{\gamma_t^*}(a^-, q, a)}}{e^{\alpha r_{\gamma_t^*}(a^+, q, a)} + e^{\alpha r_{\gamma_t^*}(a^-, q, a)}} \leq \frac{1}{1 + e^{\alpha_0 \delta}}.$$

1023 The total probability that the datasets \mathcal{D} and \mathcal{D}_α differ is bounded by

$$\mathbb{P}(\mathcal{D} \neq \mathcal{D}_\alpha) = \sum_{i=1}^N \mathbb{P}(a_i^+ \neq a_i^-) \leq \frac{N}{1 + e^{\alpha_0 \delta}}.$$

1024 Given $\alpha_0 \gtrsim \frac{\log \frac{N}{\delta \epsilon}}{\delta}$, we conclude that

$$\mathbb{P}(\mathcal{D} = \mathcal{D}_\alpha) = 1 - \mathbb{P}(\mathcal{D} \neq \mathcal{D}_\alpha) \geq 1 - \frac{\epsilon}{2}.$$

1025

□

1026 we can take our data generated according to the α_0 -BT model. In this case the new reward will be

$$\tilde{r}_{\alpha_0, \gamma_t^*}(a', q, a) = \alpha_0 r_{\gamma_t^*}(a', q, a). \quad (\text{C.13})$$

1027 Under this model, the minimizer of the loss (C.8) can be obtained via a two-step optimization
1028 procedure [61]:

1029 Step 1: minimize the negative log-likelihood to obtain the reward:

$$L_{N, \text{NLL}}(\gamma \mid \alpha_0) = -\mathbb{E}_{((a^+, z^+, w^+), (a^-, z^-, w^-), a, q) \sim \mathcal{D}} \log \sigma(\tilde{r}_{\alpha_0, \gamma}(a^+, q, a) - \tilde{r}_{\alpha_0, \gamma}(a^-, q, a)), \quad (\text{C.14})$$

1030 Denote the minimizer as $\tilde{r}_{\alpha_0, \hat{\gamma}_{N,t}}(a', q, a)$.

1031 Step 2: maximize the reward $\tilde{r}_{\alpha_0, \hat{\gamma}_{N,t}}(a', q, a)$:

$$L_{\text{REW}}(\theta \mid \theta_{t-1}) = \mathbb{E}_{(a, q) \sim \tilde{p}_{\theta^*}(\cdot)} \left[\mathbb{E}_{z, w \sim \tilde{p}_{\theta_{t-1}}(\cdot \mid a, q)} [\mathbb{E}_{a' \sim f_{\theta}(\cdot \mid z, w, q, a)} [\tilde{r}_{\alpha_0, \hat{\gamma}_{N,t}}(a', q, a)]] \right] \\ - \beta_{\text{mDPO}} \mathbb{D}_{\text{KL}}(u_{\theta}(a' \mid z, w, q, a) \parallel u_{\theta_{t-1}}(a' \mid z, w, q, a)). \quad (\text{C.15})$$

1032 The solution will be

$$u_{\hat{\theta}_t}(a' \mid z, w, q, a) \propto u_{\theta_{t-1}}(a' \mid z, w, q, a) \exp \left(\frac{1}{\beta_{\text{mDPO}}} \tilde{r}_{\alpha_0, \hat{\gamma}_{N,t}}(a', q, a) \right), \quad (\text{C.16})$$

1033 This expression is identical to Equation (C.9), except that it uses a different parameterization of the
1034 reward. Specifically, the reward function $\tilde{r}_{\alpha_0, \gamma_t^*}(a', q, a)$ is parameterized by the ground truth γ_t^*
1035 and a hyperparameter α_0 . To ensure uniform consistency of the maximum likelihood estimator, we
1036 invoke the following lemma, which is modified from Theorem 5.7 in [75]. This result guarantees that
1037 the minimizer in Step 1 converges to the true reward function $\tilde{r}_{\alpha_0, \gamma_t^*}(a', q, a)$. We need the following
1038 assumption:

1039 *assumption C.6.* Suppose that there exists a constant $c_\alpha > 0$, for every $\epsilon > 0$, such that:

$$\sup_{\alpha_0 \in [c_\alpha, \infty)} \sup_{\gamma \in \Gamma} |L_{N, \text{NLL}}(\gamma \mid \alpha_0) - L_{\text{NLL}}(\gamma \mid \alpha_0)| \xrightarrow{P} 0, \quad (\text{C.17})$$

1040 where Γ represents the parameter space and

$$\sup_{\alpha_0 \in [c_\alpha, \infty)} \sup_{\gamma: \|\gamma - \gamma_t^*\|_2 \geq \epsilon} -(L_{\text{NLL}}(\gamma \mid \alpha_0) - L_{\text{NLL}}(\gamma_t^* \mid \alpha_0)) < 0. \quad (\text{C.18})$$

1041 **Lemma C.2** (Uniform MLE Consistency). Let $L_{N, \text{NLL}}(\gamma \mid \alpha_0)$ be the negative log-likelihood
1042 function, and let $L_{\text{NLL}}(\gamma \mid \alpha_0)$ denote its expected version. Let Assumption C.6 holds, then for the
1043 sequence of estimators $\hat{\gamma}_{N,t}$ obtained from minimizing the loss (C.14), we have: given $\epsilon > 0$, there
1044 exists N_1 , when $N \geq N_1$, for any $\alpha_0 \in [c_\alpha, \infty)$,

$$\mathbb{P}(\|\hat{\gamma}_{N,t} - \gamma_t^*\|_2 \leq \epsilon) \geq 1 - \frac{\epsilon}{2}. \quad (\text{C.19})$$

1045 *Proof.* For given ϵ , according to the Equation (C.18), there exists $c_{\epsilon, NLL}$, such that:

$$\sup_{\alpha_0 \in [c_\alpha, \infty)} \sup_{\gamma: \|\gamma - \gamma_t^*\|_2 \geq \epsilon} -(L_{NLL}(\gamma | \alpha_0) - L_{NLL}(\gamma_t^* | \alpha_0)) < -c_{\epsilon, NLL}.$$

1046 For $c_{\epsilon, NLL}$, according to Equation (C.17), there exists N_1 , when $N \geq N_1$, for any $\alpha_0 \in [c_\alpha, \infty)$,

$$\begin{aligned} \mathbb{P}\left(|L_{N, NLL}(\hat{\gamma}_{N, t} | \alpha_0) - L_{NLL}(\hat{\gamma}_{N, t} | \alpha_0)| \leq \frac{c_{\epsilon, NLL}}{3}\right) &\geq 1 - \frac{\epsilon}{4}, \\ \mathbb{P}\left(|L_{N, NLL}(\gamma_t^* | \alpha_0) - L_{NLL}(\gamma_t^* | \alpha_0)| \leq \frac{c_{\epsilon, NLL}}{3}\right) &\geq 1 - \frac{\epsilon}{4}. \end{aligned}$$

1047 Since $\hat{\gamma}_{N, t}$ is the minimizer of loss (C.14), for any $\alpha_0 \in [c_\alpha, \infty)$, we have:

$$L_{N, NLL}(\hat{\gamma}_{N, t} | \alpha_0) \leq L_{N, NLL}(\gamma_t^* | \alpha_0)$$

1048 Consequently,

$$\mathbb{P}\left(-(L_{NLL}(\gamma | \alpha_0) - L_{NLL}(\gamma_0 | \alpha_0)) \geq -\frac{2c_{\epsilon, NLL}}{3}\right) \geq 1 - \frac{\epsilon}{2}.$$

1049 Thus for any $\alpha_0 \in [c_\alpha, \infty)$,

$$\mathbb{P}(\|\hat{\gamma}_{N, t} - \gamma_t^*\|_2 \leq \epsilon) \geq \mathbb{P}\left(-(L_{NLL}(\gamma | \alpha_0) - L_{NLL}(\gamma_0 | \alpha_0)) \geq -\frac{2c_{\epsilon, NLL}}{3}\right) \geq 1 - \frac{\epsilon}{2}.$$

1050 □

1051 Having established the necessary groundwork, we are now ready to present Theorem C.2, which
1052 establishes the equivalence between the minimizes of the two loss functions:

1053 **Theorem C.2.** Assume Assumptions C.5 and C.6 hold, given ϵ , there exists N_1 and $\beta_{\text{mDPO}} \gtrsim \frac{\log \frac{N_1}{2\epsilon}}{\delta} \beta$,
1054 the minimizer of loss (C.8) $\hat{\theta}_{t, \text{mDPO}}$ will satisfy:

$$\mathbb{P}\left(\|\hat{\theta}_{t, \text{mDPO}} - \theta_t^*\|_2 \geq \epsilon\right) < \epsilon, \quad (\text{C.20})$$

1055 where θ_t^* is defined in Equation (C.9).

1056 *Proof.* First, according to the Lemma C.2, there exists N_1 , if we define the event $\Omega_1 =$
1057 $\{\|\hat{\gamma}_{N_1, t} - \gamma_t^*\|_2 \leq \epsilon\}$, we have:

$$\mathbb{P}(\Omega_1) \geq 1 - \frac{\epsilon}{2}.$$

1058 Secondly, just choose the sample size of \mathcal{D} as NK , define the event $\Omega_2 = \{\mathcal{D} = \mathcal{D}_{\alpha_0}\}$, by Lemma

1059 C.1, when we take $\alpha_0 \gtrsim \left(\frac{\log \frac{N_1}{2\epsilon}}{\delta} \vee c_\alpha\right)$, we have:

$$\mathbb{P}(\Omega_2) \geq 1 - \frac{\epsilon}{2}.$$

1060 Since c_α is a constant, we may, without loss of generality, take $\alpha_0 \gtrsim \frac{\log \frac{N_1}{2\epsilon}}{\delta}$. Henceforth, we restrict
1061 our analysis to the event $\Omega_1 \cap \Omega_2$, which occurs with probability at least $1 - \epsilon$. Conditioned on this
1062 event, the data can be viewed as being generated from the α_0 -BT model (C.11). Consequently, the
1063 minimizer of the loss (C.8) coincides with that of Equation (C.16):

$$\begin{aligned} u_{\hat{\theta}_{t, \text{mDPO}}}(a', z, w | q) &\propto u_{\theta_{t-1}}(a', z, w | q) \exp\left(\frac{1}{\beta_{\text{mDPO}}} \tilde{r}_{\alpha_0, \hat{\gamma}_{N_1, t}}(a', q, a)\right) \\ &\propto u_{\theta_{t-1}}(a', z, w | q) \exp\left(\frac{\alpha_0}{\beta_{\text{mDPO}}} r_{\hat{\gamma}_{N_1, t}}(a', q, a)\right). \end{aligned}$$

1064 Compared to the solution in (C.9), when $\beta_{\text{mDPO}} = \alpha_0 \beta \gtrsim \frac{\log \frac{N_1}{2\epsilon}}{\delta} \beta$, controlling the distance be-
1065 tween $\hat{\theta}_{t, \text{mDPO}}$ and θ_t^* reduces to controlling the distance between $\hat{\gamma}_{N_1, t}$ and γ_t^* , as established by
1066 Lemma C.2. Consequently, we obtain:

$$\mathbb{P}\left(\|\hat{\theta}_{t, \text{mDPO}} - \theta_t^*\|_2 \geq \epsilon\right) < \epsilon.$$

1067 This concludes the proof. □

1068 C.6 Convergence Property of DPO

1069 Finally, combining Proposition C.2, we conclude that the sequence $\hat{\theta}_{t,\text{mDPO}}$ converges as t increases.
 1070 We formally state the following theorem:

1071 **Theorem C.3.** For a given iteration number T , for some radius $r > 0$ and pair (μ, λ) such that
 1072 $0 \leq \mu < \lambda$, suppose that the Assumption C.1-C.6 hold and assume $(\epsilon + \zeta_M^{\text{unif}}(N, \epsilon)) < (1 - \frac{\mu}{\lambda})r$,
 1073 then with probability at least $1 - (T + 1)\epsilon$, we have:

$$\left\| \hat{\theta}_{T,\text{mDPO}} - \theta^* \right\|_2 \leq \left(\frac{\mu}{\lambda} \right)^T \left\| \theta_{\text{ref}} - \theta^* \right\|_2 + \frac{1}{1 - \frac{\mu}{\lambda}} \zeta_M^{\text{unif}}(n, \epsilon)$$

1074 *Proof.* Notice that $\theta_t^* = M_N(\hat{\theta}_{t-1,\text{mDPO}})$, apply Proposition C.2, we get:

$$\left\| \theta_t^* - \theta^* \right\|_2 \leq \frac{\mu}{\lambda} \left\| \hat{\theta}_{t-1,\text{mDPO}} - \theta^* \right\|_2 + \zeta_M^{\text{unif}}(N, \epsilon)$$

1075 with probability at least $1 - \epsilon$. Combining Theorem C.2,

$$\left\| \hat{\theta}_{t,\text{mDPO}} - \theta^* \right\|_2 \leq \frac{\mu}{\lambda} \left\| \hat{\theta}_{t-1,\text{mDPO}} - \theta^* \right\|_2 + \epsilon + \zeta_M^{\text{unif}}(N, \epsilon),$$

1076 with probability at least $1 - 2\epsilon$. Notice that $(\epsilon + \zeta_M^{\text{unif}}(N, \epsilon)) \leq (1 - \frac{\mu}{\lambda})r$, then $\hat{\theta}_{t,\text{mDPO}} \in B(r, \theta^*)$.
 1077 Based on this, we can perform iteration:

$$\begin{aligned} \left\| \hat{\theta}_{T,\text{mDPO}} - \theta^* \right\|_2 &\leq \frac{\mu}{\lambda} \left\| \hat{\theta}_{T-1,\text{mDPO}} - \theta^* \right\|_2 + \epsilon + \zeta_M^{\text{unif}}(N, \epsilon) \\ &\leq \frac{\mu}{\lambda} \left(\frac{\mu}{\lambda} \left\| \hat{\theta}_{T-2,\text{mDPO}} - \theta^* \right\|_2 + \epsilon + \zeta_M^{\text{unif}}(N, \epsilon) \right) \\ &\leq \left(\frac{\mu}{\lambda} \right)^T \left\| \theta_{\text{ref}} - \theta^* \right\|_2 + \sum_{s=0}^{T-1} \left(\frac{\mu}{\lambda} \right)^s (\epsilon + \zeta_M^{\text{unif}}(N, \epsilon)) \\ &\leq \left(\frac{\mu}{\lambda} \right)^T \left\| \theta_{\text{ref}} - \theta^* \right\|_2 + \frac{1}{1 - \frac{\mu}{\lambda}} (\epsilon + \zeta_M^{\text{unif}}(N, \epsilon)) \end{aligned}$$

1078 with probability at least $1 - (T + 1)\epsilon$. □

1079 D Information for Test Datasets

1080 The information of the test datasets used in AceRAG is listed in the following table. Note that We
 1081 conduct evaluations on all questions from StrategyQA and Bamboogle, and the first 500 questions
 1082 from the development sets of the other datasets following existing studies [74, 64, 39]. For dataset in
 1083 DocMathEval, we use the testmini version as the evaluation set to compare the performance of
 1084 AceRAG and baselines.

Table 4: Descriptions of datasets used in AceRAG. For SimpLong and CompLong, we use `text-embedding-3` to retrieve top-10 relevant context before generate the answer.

Dataset	Description
2WikiMHQA [25]	2WikiMultiHopQA is a multi-hop question answering dataset built from Wikipedia, where each question requires reasoning over two distinct articles. It emphasizes information synthesis across multiple documents for accurate answer retrieval.
HotpotQA [85]	HotpotQA is a crowd-sourced multi-hop QA dataset where each question demands reasoning over multiple Wikipedia passages. It also includes supporting fact annotations to promote explainability in QA systems.
Bamboogle [58]	Bamboogle is a multi-hop QA dataset constructed using Bing search engine snippets. It presents naturally occurring, challenging questions requiring reasoning over diverse web snippets rather than structured sources like Wikipedia.
MusiQue [73]	MusiQue is a multi-hop QA dataset featuring real-world questions from community forums like Quora and Yahoo Answers. It targets complex questions requiring synthesis across multiple evidence passages, each carefully annotated.
HOVER [32]	HOVER is a multi-hop QA dataset with annotated supporting facts, built on entity-linked Wikipedia documents. It stresses explainable reasoning by providing intermediate evidence chains.
ExFEVER [50]	ExFEVER extends the FEVER dataset by introducing multi-hop claims requiring evidence from multiple documents. It is designed to support research on fact verification and evidence-based reasoning.
TAT-QA [96]	TAT-QA is a financial QA dataset combining text and tables, requiring multi-step numerical reasoning and answer-type classification.
FinQA [8]	FinQA is a numerical reasoning dataset in the financial domain, where each question requires multi-step operations over semi-structured financial reports. It evaluates models on their ability to perform accurate quantitative reasoning in high-stakes contexts.
MultiHiertt [92]	MultiHiertt is a hierarchical multi-document QA dataset with questions requiring reasoning over tree-structured, topically segmented documents. It tests models' capabilities in processing structured discourse and thematic hierarchies.
TAT-HQA [38]	TAT-HQA is a counterfactual QA dataset extending TAT-QA, requiring compositional numerical reasoning under hypothetical changes.
DM _{SS} (DocMath SimpShort)	A dataset reannotated from TAT-QA and FinQA, consisting of short financial documents with a single table for simple numerical reasoning.
DM _{CS} (DocMath CompShort)	A dataset reannotated from TAT-HQA, consisting of short single-table documents for complex numerical reasoning, including hypotheticals.
DM _{SL} (DocMath SimpLong)	A dataset reannotated from MultiHiertt, consisting of long multi-table financial documents for simple reasoning in realistic contexts.
DM _{CL} (DocMath CompLong)	A dataset of long, structured financial documents requiring multi-step compositional numerical reasoning.

1085 E Details of Training Data

1086 We provide the data composition for SFT and RFT, including their corresponding tasks, links to
 1087 access the data, and the number we use in each stage in Table 5.

Table 5: The data composition for SFT and RFT stages.

Dataset	Task	Link	Count
<i>Data composition for SFT</i>			
NarrativeQA [35]	Context-rich QA	https://huggingface.co/datasets/deepmind/narrativeqa	20000
SQuAD 1.1 [62]	Context-rich QA	https://rajpurkar.github.io/SQuAD-explorer/	10000
SQuAD 2.0 [62]	Context-rich QA	https://rajpurkar.github.io/SQuAD-explorer/	10000
TAT-QA [96]	Context-rich QA	https://github.com/NExTplusplus/TAT-QA/tree/master/dataset_raw	12000
FEVER [72]	Context-rich QA	https://fever.ai/dataset/fever.html	10000
DROP [17]	Context-rich QA	https://huggingface.co/datasets/ucinlp/drop	20000
Quoref [13]	Context-rich QA	https://huggingface.co/datasets/allenai/quoref	20000
ROPES [41]	Context-rich QA	https://huggingface.co/datasets/allenai/ropes	10000
NQ [36]	Context-rich QA	https://dl.fbaipublicfiles.com/dpr/data/retriever/biencoder-nq-train.json.gz	20000
GSM8K [12]	Question Decomposition	https://huggingface.co/datasets/openai/gsm8k/viewer/socratic	7000
ConvFinQA [10]	Question Decomposition	https://github.com/czyssrs/ConvFinQA	1000
StrategyQA [20]	Question Decomposition	https://huggingface.co/datasets/ChilleD/StrategyQA	1600
IfQA [87]	CoT	https://github.com/wyu97/IfQA/tree/main/dataset	2000
TabMWP [49]	CoT	https://promptpg.github.io/index.html#dataset	10000
GSM8K [12]	CoT	https://huggingface.co/datasets/openai/gsm8k/viewer/socratic	7000
MathInstruct-COT [89]	CoT	https://huggingface.co/datasets/TIGER-Lab/MathInstruct	10000
MathInstruct-POT [89]	CoT	https://huggingface.co/datasets/TIGER-Lab/MathInstruct	10000
TOTAL	—	—	180600
<i>Data composition for RFT</i>			
HotpotQA [85]	RAG	https://github.com/hotpotqa/hotpot	10000
2WikiMQA [25]	RAG	https://huggingface.co/datasets/xanhho/2WikiMultihopQA	10000
HOVER [32]	RAG	https://github.com/hover-nlp/hover	10000
GSM8K [12]	Context-rich Reasoning	https://huggingface.co/datasets/openai/gsm8k/viewer/socratic	7000
TabMWP [49]	Context-rich Reasoning	https://promptpg.github.io/index.html#dataset	10000
ConvFinQA [10]	Context-rich Reasoning	https://github.com/czyssrs/ConvFinQA	2000
Total	—	—	49000

1088 F Prompt Templates

1089 F.1 Prompts for Direct RAG

You have the following context passages: {context}

Given the question: “{question}” as well as the context above, please answer the above question with one or a list of entities with the given context as the reference. Your answer needs to be a span with one or a list of entities.

Figure 6: Prompt for direct RAG on complex question answering tasks.

Answer the following questions with SUPPORTED or NOT_SUPPORTED with the given context as the reference.

Question: {question}

Context: {context}

Your answer should only be SUPPORTED or NOT_SUPPORTED.

Figure 7: Prompt for direct RAG on fact verification tasks.

You have the following passages and table:

Passages:

{passage}

Tables:

{table}

For the question “{question}”, write a Python program to solve the question. Store the final result in the variable ans.

Figure 8: Prompt for direct RAG on document-level reasoning tasks with PoT.

You have the following passages and table:

Passages:

{passage}

For the question “{question}”, reason step by step to calculate the final answer. Please use \boxed{ } to wrap your final answer.

Figure 9: Prompt for direct RAG on document-level reasoning tasks with CoT.

Please break down the question “{question}” into multiple specific sub-questions that address individual components of the original question.

Mark each sub-question with ### at the beginning. If you need to refer to answers from earlier sub-questions, use #1, #2, etc., to indicate the corresponding answers.

Decomposed question:

Figure 10: Prompt for question decomposition on complex question answering tasks.

Please break down the claim “{claim}” into multiple smaller sub-claims that each focus on a specific component of the original statement, making it easier for a model to verify. Begin each sub-claim with ###. If needed, refer to answers from earlier sub-claims using #1, #2, etc.

Decomposed claim:

Figure 11: Prompt for question decomposition on fact verification tasks.

You have the following passages and table:

Passages:

{passages}

Tables:

{tables}

Please break down the question “{question}” into multiple specific sub-questions that address individual components of the original question, with the table and passages as the reference. Use ### to mark the start of each sub-question.

Decomposed question:

Figure 12: Prompt for question decomposition on document-level reasoning tasks.

You have the following context passages:

{passages}

Please answer the question “{subquestion}” with a short span using the context as reference. If no answer is found in the context, use your own knowledge. Your answer needs to be as short as possible.

Figure 13: Prompt for subquestion answering on complex question answering tasks.

You have the following context passages:

`{passages}`

Please verify whether the claim “`{subquestion}`” is correct using the context as reference. If no answer is found in the context, use your own knowledge. Please only output Yes or No and do not give any explanation.

Figure 14: Prompt for subquestion answering on fact verification tasks.

You have the following passages and tables:

Passage:

`{passages}`

Table:

`{tables}`

For the question “`{subquestion}`”, write a Python program to solve the question. Store the final result in the variable `ans`.

Figure 15: Prompt for subquestion answering on document-level reasoning tasks with PoT.

You have the following passages and tables:

Passage:

`{passages}`

Table:

`{tables}`

For the question “`{subquestion}`”, reason step by step to calculate the final answer. Please use `\boxed{ }` to wrap your final answer.

Figure 16: Prompt for subquestion answering on document-level reasoning tasks with CoT.

1092 E.4 Prompts for final answer generation

You have the following passages:

`{passages}`

You are also given some subquestions and their answers:

subquestion #1: `{subquestion_1}` **Answer:** `{answer_1}`

subquestion #2: `{subquestion_2}` **Answer:** `{answer_2}`

...

Please answer the question “`{the_original_question}`” with a short span using the documents and subquestions as reference.

Make sure your response is grounded in documents and provides clear reasoning followed by a concise conclusion. If no relevant information is found, use your own knowledge.

Wrap your answer with `<answer>` and `</answer>` tags.

Figure 17: Prompt for final answer generation on complex question answering tasks.

You are given some subquestions and their answers:

subquestion #1: {subquestion_1} Answer: {answer_1}

subquestion #2: {subquestion_2} Answer: {answer_2}

...

Please answer the question “{the_original_question}” with only Yes or No using the subquestions as reference. Provides clear reasoning followed by a concise conclusion. If no relevant information is found, use your own knowledge.

Wrap your answer with <answer> and </answer> tags.

Figure 18: Prompt for final answer generation on fact verification tasks.

You have the following passages and table:

Passages:

{passage}

For the question “{question}”, here is a referenced breakdown:

{decomposition}.

Write a Python program to solve the question. Store the final result in the variable ans.

Figure 19: Prompt for final answer generation on document-level reasoning tasks with PoT.

You have the following passages and table:

Passages:

{passage}

For the question “{question}”, here is a referenced breakdown:

{decomposition}.

Reason step by step to calculate the final answer. Please use \boxed{} to wrap your final answer.

Figure 20: Prompt for final answer generation on document-level reasoning tasks with CoT.

Read the following documents relevant to the given question: {question}

Documents:

{documents}

...

Please identify documents that are useful to answer the given question: “{question}”. If none of the documents is aligned with the answer, in that case, you have to explain the answer only based on your own knowledge, without referring to the provided information.

Note that the question may be compositional and require intermediate analysis to deduce the final answer. Make sure your response is grounded and provides clear reasoning details followed by a concise conclusion. Your answer should be in a short span with a few keywords. Use <answer> and </answer> tag to mark your final answer.

Figure 21: Prompt for InstructRAG on complex question answering tasks.

Read the following documents relevant to the given question: {question}

Documents:

{documents}

...

Please identify documents that are useful to answer the given question: “{question}”. If none of the documents is aligned with the answer, in that case, you have to explain the answer only based on your own knowledge, without referring to the provided information.

Note that the question may be compositional and require intermediate analysis to deduce the final answer. Make sure your response is grounded and provides clear reasoning details followed by a concise conclusion. Your answer should be yes or no only. Use <answer> and </answer> tag to mark your final answer.

Figure 22: Prompt for InstructRAG on fact verification tasks.

1094 **G Additional Implementation Details**

1095 **G.1 Implementation Details for SFT**

1096 For SFT, we set the batch size to 64 for every example, and set the learning rate as Table 7. With maximum number of tokens to 2560.

Table 6: Results for different model sizes for SFT.

Model Size	Learning Rate	Warmup Steps
AceRAG 1.5B	$5e - 6$	5%
AceRAG 8B	$1e - 6$	5%
AceRAG 14B	$1e - 6$	5%
AceRAG 32B (LoRA)	$1e - 5$	5%

1097

1098 **G.2 Implementation Details for RFT**

1099 We set the hyperparameters to $m = 3$, $m' = 4$, and $t = 1.0$ when generating multiple rollouts.
 1100 Examples with identical maximum and minimum rewards are discarded. For RFT, we use $\beta = 0.1$

Table 7: Results for different model sizes for RFT.

Model Size	Learning Rate	Warmup Steps
AceRAG 1.5B	$1e-6$	5%
AceRAG 8B	$5e-7$	5%
AceRAG 14B	$5e-7$	5%
AceRAG 32B (LoRA)	$1e-6$	5%

and run for the DPO for 2 iterations by default. All models are optimized using AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, and experiments are conducted on 8 NVIDIA A100 GPUs.

G.3 Implementation Details for Inference

During inference, we set the temperature $t = 0.0$, the number of retrieved passages in each step to $k = 10$ by default. We use E5 [78] as the default embedding model.

G.4 Implementation Details for Baselines

We implement and evaluate a variety of baselines using standardized decoding and prompting configurations to ensure fair comparison. For **Qwen-3**, we follow the official guidance⁵ to adopt distinct sampling strategies depending on the task setting. In *thinking mode* (`enable_thinking=True`), we use temperature = 0.6, top-p = 0.95, top-k = 20, and min-p = 0 to encourage diverse yet coherent generation. Greedy decoding is explicitly avoided to prevent performance degradation and repetitive outputs. In *non-thinking mode* (`enable_thinking=False`), we slightly increase the temperature to 0.7 and reduce top-p to 0.8 while keeping top-k and min-p unchanged. In practice, we find that using the thinking mode leads to slightly better performance despite being slower. For **R1-distill** models, we set the maximum generation length to 32,768 tokens and use temperature = 0.6, top-p = 0.95. In **Plan-RAG**, we incorporate 3-shot demonstrations in the prompt to guide the model toward producing outputs in the correct format. For **InstructRAG**, we use the same SFT training set as AceRAG and generate CoT-style demonstrations tailored to context-rich QA datasets. For **Llama-4**, **GPT-4.1**, and **GPT-4o**, we use greedy decoding (temperature = 0) for consistency with their default inference behavior. For **IRCOT** and **RAG-Star**, we reproduce results by following the original repositories and hyperparameter settings. For these methods, we tune the number of retrieved passages from {5, 10, 20} and report the best performance. We refer to other baselines’ reported numbers in the corresponding paper.

H Additional Experimental Results

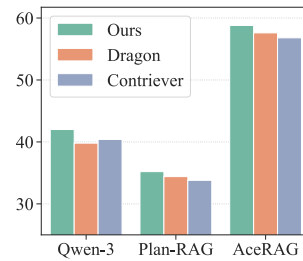
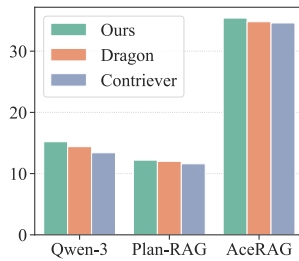
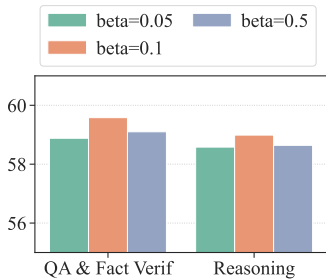


Figure 23: Parameter Study on β

Figure 24: Effect of different retrievers.

Effect of β . We study the effect of β in preference optimization with Llama-3.1-8B as the backbone, and find that AceRAG is generally robust to this parameter, with $\beta = 0.1$ leads to slightly better performance.

⁵<https://huggingface.co/Qwen/Qwen3-32B#best-practices>

Table 8: Performance comparison across models and prompting methods.

Model	Prompt Method	DM _{SimpShort}	DM _{CompShort}	DM _{SimpLong}	DM _{CompLong}	Avg.
AceRAG-32B	PoT	89.5	84.0	53.0	43.0	66.1
AceRAG-14B	PoT	84.0	82.0	49.0	39.3	62.4
AceRAG-8B	PoT	83.0	80.5	48.0	32.3	59.0
AceRAG-1.5B	PoT	66.5	77.5	39.0	18.0	47.6
AceRAG-32B	CoT	73.5	70.0	50.0	33.0	54.5
AceRAG-14B	CoT	78.5	75.5	44.0	34.7	57.0
AceRAG-8B	CoT	44.0	31.5	30.0	15.7	28.5
AceRAG-1.5B	CoT	37.5	32.0	18.0	9.7	23.2

Effect of Different Retrievers. We evaluate AceRAG and representative baselines (at the 8B scale) using two different retrievers: Dragon⁶ and Contriever⁷. Overall, the E5 retriever achieves the best performance, supporting our hypothesis that stronger retrieval models yield more relevant passages and thus enhance answer quality. Notably, AceRAG consistently outperforms baselines across different retrievers, demonstrating its robustness to retrieval choices.

Comparison of CoT and PoT for Document-level Reasoning. Table 8 presents a comparison between Program of Thought (POT) and Chain of Thought (COT) prompting methods across four evaluation settings. POT consistently outperforms COT across all tasks, with notable improvements on both simple and complex reasoning benchmarks. For example, across models, POT yields higher average scores than COT on DM_{CompLong} (e.g., 43.0 vs. 33.0 for AceRAG-32b) and DM_{SimpShort} (e.g., 89.5 vs. 73.5 for AceRAG-32b), demonstrating its advantage in guiding structured reasoning. These results highlight the effectiveness of POT in enhancing model performance on decision-making tasks requiring multi-step reasoning, regardless of model scale.

I Case Studies

Details of Human Study Our human study has received IRB approval from our institute. The below is the form used in human evaluation:

You are provided with:

- The original complex question
- A proposed decomposition into subquestions

Please assess the overall quality of the decomposition using the criteria below.

- Relevance: Do the subquestions help solve the original question?
- Completeness: Are all important aspects covered?
- Coherence: Is the breakdown logically structured and easy to follow?
- Usefulness: Does the decomposition make the reasoning process easier or more interpretable?

Please rate the overall quality of the subquestion decomposition.

- 1: Very Poor: Subquestions are irrelevant, incomplete, or incoherent.
- 2: Poor: Some relevance, but major gaps or unclear logic.
- 3: Fair: Moderately helpful with minor issues in coverage or clarity.
- 4: Good: Clear and mostly complete decomposition.
- 5: Excellent: Subquestions are highly relevant, complete, and well-structured, with proper reference to answers from previous subquestions.

⁶<https://huggingface.co/facebook/dragon-plus-context-encoder>

⁷<https://huggingface.co/facebook/contriever-msmarco>

1160 Details of Case Study

1161 We present two case studies from the MusiQue, 2WikiMHQA and DocMathEval datasets in Tables 9,
 1162 10 and 11. Despite its larger size, the Qwen3-32B-reasoning model struggles with handling complex
 1163 questions. In contrast, AceRAG effectively decomposes each question into manageable sub-questions,
 1164 generates accurate answers for each sub-questions, and performs multi-step reasoning to arrive at the
 1165 correct final answer.

Table 9: A case study from MusiQue dataset.

A sample from MusiQue
Question: In which state is Vera Barbosa’s place of birth located? Ground Truth Answer: Lisbon District
AceRAG-8B
Sub-question 1: Who is Vera Barbosa? Answer 1: a Portuguese track and field athlete Sub-question 2: Where was Vera Barbosa born? Answer 2: Vila Franca de Xira Sub-question 3: In which state is #2 located? Answer 3: Lisbon District Final Answer: The documents state that Vera Barbosa is a Portuguese track and field athlete and was born in Vila Franca de Xira. Additionally, Vila Franca de Xira is located in the Lisbon District. Therefore, Vera Barbosa’s place of birth, Vila Franca de Xira, is in the Lisbon District.
<answer>Lisbon District</answer>
Qwen3-32B Reasoning
<think> Okay, let’s see. The question is asking about the state where Vera Barbosa’s place of birth is located. First, I need to figure out which Vera Barbosa they’re referring to here because there are several people named Vera Barbosa in the provided contexts. Looking at the first entry: Vera Barbosa, born 13 January 1989 in Vila Franca de Xira, Portugal. That’s a Portuguese athlete. Then there’s another Vera Krasova, a Russian model, and Vera Begić, a Croatian athlete. There’s also a mention of a place called Barbosa in Colombia and Brazil. But the key here is the name Vera Barbosa specifically. The user is asking about Vera Barbosa’s birthplace state. The first context says she was born in Vila Franca de Xira, which is in Portugal. However, the question is phrased as "In which state is Vera Barbosa’s place of birth located?" So, states are typically administrative divisions in countries like the US, Brazil, etc. Portugal is divided into districts, not states. But maybe the user is using "state" in a general sense here. Wait, but looking at other entries, there’s a Carlos Barbosa in Brazil’s Rio Grande do Sul state, and Bárbara Barbosa is a Brazilian footballer. But the specific Vera Barbosa mentioned in the first paragraph is the Portuguese athlete. Her birthplace is Vila Franca de Xira, Portugal. Since Portugal doesn’t have states, but the question uses "state," perhaps the answer is Portugal. Alternatively, if the user is confused and actually refers to a different person, but based on the given data, the only Vera Barbosa with a birthplace listed is the Portuguese one. Therefore, the answer should be Portugal. </think>
<answer>Portugal</answer>

Table 10: A case study from 2WikiMHQA dataset.

A sample from 2WikiMHQA
Question: Which film has the director who was born later, The Silver Treasure or Taxi To Paradise? Ground Truth Answer: Taxi To Paradise
AceRAG-8B
Sub-question 1: Who directed The Silver Treasure? Answer 1: Rowland V. Lee Sub-question 2: Who directed Taxi To Paradise? Answer 2: Adrian Brunel Sub-question 3: When was the director of #1 born? Answer 3: September 6, 1891 Sub-question 4: When was the director of #2 born? Answer 4: 4 September 1892 Sub-question 5: Is the year of #3 later than #4? Answer 5: no Final Answer: From the documents, "The Silver Treasure" was directed by Rowland V. Lee, born on September 6, 1891 (Q3), and "Taxi To Paradise" was directed by Adrian Brunel, born on September 4, 1892 (Q4). Comparing these dates, Adrian Brunel was born later than Rowland V. Lee (Q5 confirms this). Thus, the film with the director born later is "Taxi To Paradise."
<answer>Taxi To Paradise</answer>
Qwen3-32B Reasoning
<think> Okay, let's see. The user is asking which film has the director born later between The Silver Treasure and Taxi to Paradise. First, I need to find the directors of both films. From the context: - The Silver Treasure (1926) is directed by Rowland V. Lee. - Taxi to Paradise (1933) is directed by Adrian Brunel. Now, I need to check their birth dates. The context doesn't provide birth dates for these directors. So I'll have to rely on my own knowledge. Rowland V. Lee was born on April 22, 1894. Adrian Brunel was born on October 2, 1887. Wait, Brunel was born earlier than Lee? Wait, no: if Brunel was born in 1887 and Lee in 1894, then Brunel was born earlier. Therefore, Rowland V. Lee was born later. So The Silver Treasure's director was born later than Taxi to Paradise's director. Therefore, the answer is The Silver Treasure. </think>
<answer>The Silver Treasure</answer>

Table 11: A case study from DM_{CS} subset of the DocMathEval dataset.**A sample from DM_{CS} subset of DocMathEval.****Table:**

Year	Segment	Americas	Europe	Asia	Total	% of Total
<i>(Amounts in thousands)</i>						
2019	TS	\$67,728	\$3,285	\$646	\$71,159	90%
2019	HPP	\$5,294	\$771	\$1,837	\$7,902	10%
2019	Total	\$72,522	\$4,056	\$2,483	\$79,061	100%
2019	% of Total	92%	5%	3%	100%	
2018	TS	\$52,034	\$9,059	\$1,344	\$62,437	86%
2018	HPP	\$8,424	\$1,266	\$789	\$10,479	14%
2018	Total	\$60,458	\$10,325	\$2,133	\$72,916	100%
2018	% of Total	83%	14%	3%	100%	

Table 12: The following table details the Company's sales by operating segment for fiscal years ended September 30, 2019 and 2018. The Company's sales by geographic area based on the location of where the products were shipped or services rendered are as follows: Substantially all Americas amounts are United States.

Question: What is the percentage change in Asia sales between 2018 and 2019 if the 2019 sales is doubled and increased by another 400 thousand? (in percent)

Ground Truth Answer: 151.5705

AceRAG-8B

Sub-question 1: What was the value of Asia sales in 2019?

Sub-question 2: What is the value of Asia sales in 2019 after doubling it?

Sub-question 3: What is the value of Asia sales in 2019 after doubling it and then adding 400,000?

Sub-question 4: What was the value of Asia sales in 2018?

Sub-question 5: What is the net change in the value of Asia sales from 2018 to 2019, after adjusting the 2019 value?

Sub-question 5: What is the percentage change in the value of Asia sales from 2018 to 2019, after adjusting the 2019 value?

Final Answer:

Given data

asia_sales_2019 = 2483 # in thousands (Q1)

asia_sales_2018 = 2133 # in thousands (Q4)

Adjust 2019 sales: double it and add 400 thousand (Q2, Q3)

adjusted_2019_sales = (asia_sales_2019 * 2) + 400

Compute value change (Q5)

change_in_sales = adjusted_2019_sales - asia_sales_2018

Compute percentage change (Q6)

ans = (change_in_sales / asia_sales_2018) * 100

151.57055789967183

Qwen3-32B Reasoning

Original 2019 Asia sales in thousands

original_2019_asia = 646

Adjusted 2019 sales after doubling and adding 400 thousand

adjusted_2019_asia = original_2019_asia * 2 + 400

Original 2018 Asia sales in thousands

original_2018_asia = 1344

Calculate percentage change

percentage_change = ((adjusted_2019_asia - original_2018_asia) / original_2018_asia) * 100

ans = percentage_change

25.892857142857146