OMNI-TRUTH: A UNIFIED FRAMEWORK FOR LLM HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models must remain faithful to supplied evidence in closed-domain generation and answer reliably from parametric knowledge in open-domain generation without external evidence. Achieving both behaviors in a single model remains challenging, as improving one can compromise the other. We introduce OMNI-TRUTH, a unified post-training framework for closed-domain faithfulness and open-domain factuality. OMNI-TRUTH first develops a shared reflection-and-correction policy through supervised fine-tuning, then applies closed-domain followed by open-domain reinforcement learning to acquire complementary capabilities. Task-specific multi-teacher on-policy distillation consolidates these capabilities into a single model. Across six benchmarks, OMNI-TRUTH achieves the highest macro-average score among the evaluated Qwen3-8B-based methods (67.54), while balancing closed-domain faithfulness and open-domain factuality. On the closed-domain benchmarks, OMNI-TRUTH achieves an average score of 83.75, only 2.69 points below DeepSeek V4 Flash under our evaluation protocol.

1 INTRODUCTION

Hallucinations in large language models (LLMs) encompass multiple types of factual and faithfulness errors, distinguished in part by the evidence against which a response is evaluated (Ji et al., 2023; Huang et al., 2025). In a closed-domain task, a query is paired with a document, passage, or scenario, and correctness is defined relative to that supplied context. The model should extract and synthesize supported information, avoid contradictions, and not replace the evidence with a plausible parametric prior (Maynez et al., 2020). In our open-domain setting, the model receives only a question, without external evidence. It must answer using its parametric knowledge and determine whether that knowledge is sufficient, challenges studied by factuality and self-knowledge evaluations (Lin et al., 2022; Kadavath et al., 2022). The closed-domain setting therefore primarily tests *contextual faithfulness*, whereas the open-domain setting tests *parametric factuality and appropriate abstention*. Figure 1 illustrates the different evidence boundaries with a question about planets with rings. The supplied passage supports only Jupiter and Saturn, whereas an answer without supplied context must also include Uranus and Neptune to be complete.

Existing mitigation methods are also largely specialized by regime. For closed-domain faithfulness, recent methods provide step-level supervision, resample from faithful prefixes, or localize unsupported reasoning so that responses remain grounded in the supplied source (Gui et al., 2026; Nie et al., 2026; Li & Ng, 2025; Xie et al., 2026). For open-domain factuality, reinforcement-learning methods use knowledge verification, truthfulness-aware rewards, or retrieval-augmented reward computation to improve factual answering and calibrate abstention (Ren et al., 2026; Wei et al., 2026; Chen et al., 2025). These approaches produce strong domain specialists, but do not by themselves determine how to preserve both behaviors in a single policy. Recent joint studies and diagnostic evidence highlight the tension between factuality and contextual faithfulness (Wang et al., 2025b;a; Bi et al., 2025). Building on this work, we study how to consolidate complementary capabilities acquired through closed-domain and open-domain RL while preserving contextual faithfulness and factual reliability in a single policy.

Despite these advances, combining the two objectives remains non-trivial. A model's factually correct parametric knowledge may conflict with the supplied source (Li et al., 2026), making it difficult

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.

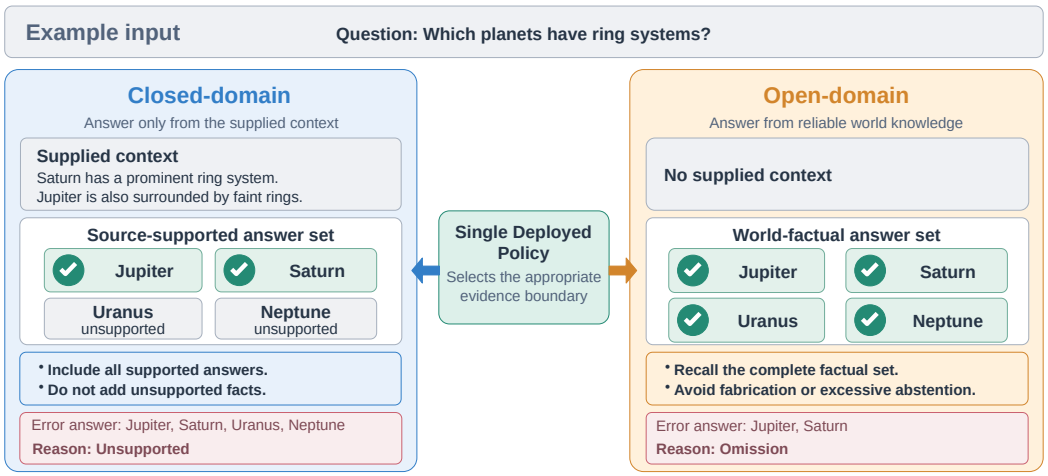


Figure 1: The same question requires different answer sets under different evidence boundaries. The supplied context supports Jupiter and Saturn, so adding Uranus or Neptune would be unsupported in the closed domain. Without supplied context, all four planets belong in the open-domain answer; returning only Jupiter and Saturn is factually correct but incomplete. A single policy must apply the appropriate boundary for each input.

to improve factuality and contextual faithfulness together. Sequential training can transfer capabilities across domains, but its later stage may weaken behaviors learned earlier (Ma et al., 2026). This motivates separating capability acquisition from integration: specialized policies can first learn complementary behaviors under different objectives and then be consolidated into one model. Multi-teacher on-policy distillation provides such a mechanism by distilling specialist teachers on the student's rollouts (Ma et al., 2026). The remaining challenge is to determine which teacher should guide each task so that consolidation preserves their complementary strengths.

To address this challenge, we introduce OMNI-TRUTH, a unified post-training framework for closed-domain faithfulness and open-domain factuality. The framework first trains a shared reflection-and-correction policy through supervised fine-tuning (SFT). We then apply closed-domain RL to obtain a Closed-only teacher that learns faithful use of supplied evidence, and continue training this policy with open-domain RL to improve factual answering, yielding a Closed-to-Open Cascade teacher. Task-specific multi-teacher on-policy distillation (MOPD) consolidates these policies: the Cascade guides question answering, while its Closed-only precursor guides source-grounded summarization. This design aims to preserve the Cascade's answering capability while recovering the Closed-only policy's source faithfulness in a single model.

Our contributions are as follows:

- We introduce OMNI-TRUTH, a three-stage post-training framework: shared reflection-and-correction SFT, closed-to-open Cascade RL, and task-specific MOPD. Together, these stages consolidate complementary capabilities into a single model that balances closed-domain faithfulness and open-domain factuality.
- We develop a bilingual, dual-domain post-training data construction pipeline for both SFT and RL. It converts closed- and open-domain examples into reflection-and-correction supervision and task-specific RL data through consistent generation and filtering.
- We evaluate OMNI-TRUTH across six benchmarks using a unified hallucination evaluation pipeline, comparing it with single-domain training, joint and sequential RL, and standard MOPD. It achieves the highest Macro score among the evaluated Qwen3-8B-based methods while balancing closed-domain faithfulness and open-domain factuality. Ablations examine student initialization and task-specific teacher guidance.

2 Related Work

Closed-domain faithfulness training. Closed-domain faithfulness training uses supplied evidence to supervise final answers and intermediate reasoning. Two concurrent methods use the name FaithRL. Gui et al. (2026) combine geometric step rewards with faithfulness-aware advantage modulation, while Nie et al. (2026) combine step-level faithfulness rewards with truncated resampling from faithful prefixes. Li & Ng (2025) propose FSPO, which verifies reasoning steps against task evidence and uses the resulting scores to adjust token-level advantages. Xie et al. (2026) propose FARCA, which models factual-error localization and verifier reliability separately for token-level credit assignment. These methods improve evidence-grounded reasoning in specialized training settings but do not integrate it with open-domain factuality in a single policy.

Open-domain factuality training. In our open-domain setting, models answer without supplied evidence and must recognize the limits of their parametric knowledge. Tian et al. (2024) construct preference pairs using automatically measured factual precision, while Zhang et al. (2024) train models to refuse questions beyond their knowledge. Among reinforcement learning methods, Ren et al. (2026) propose KnowRL with knowledge-verification rewards, and Wei et al. (2026) propose TruthRL with a ternary truthfulness signal. Chen et al. (2025) derive binary rewards from retrieved evidence. These methods target no-reference answering even when external knowledge is used to construct the training reward.

Joint faithfulness and factuality. Other studies consider both objectives. Bi et al. (2025) evaluate several factuality-enhancement methods and find inconsistent factuality gains alongside substantial declines in contextual faithfulness, which they associate with greater reliance on parametric knowledge. Wang et al. (2025a) propose SPACE, which identifies and edits activation subspaces shared by factuality and faithfulness. Wang et al. (2025b) jointly train a single policy with GRPO on short-form QA, reference-grounded long-form QA, and no-reference long-form QA. Our study complements these approaches by comparing joint RL, sequential RL, and policy distillation for closed-domain faithfulness and open-domain factuality, with a focus on consolidating complementary capabilities acquired through sequential training.

Capability integration in post-training. Ma et al. (2026) identify a limitation of sequential Cascade RL: earlier-stage capabilities may decay as later stages progress. They propose MOPD, which trains per-domain RL teachers independently and routes each prompt to a domain teacher for distillation on the student's own rollouts. Building on this paradigm, Omni-Truth uses teachers from two stages of sequential training: the Closed-only policy and the Closed-to-Open Cascade. It initializes the student from the Cascade and assigns these teachers by task within closed-domain integration data, with the Cascade guiding question answering and the Closed-only policy guiding summarization. This design aims to retain complementary capabilities for closed-domain faithfulness and open-domain factuality in a single policy.

3 Problem Formulation

Let q be a query, c a supplied context, y^* a reference answer, and y a model response. Reference answers support supervision and evaluation only. We consider a closed-domain distribution $(q, c, y^*) \sim \mathcal{D}_c$ and an open-domain distribution $(q, y^*) \sim \mathcal{D}_o$. A single policy π_θ must operate reliably in both regimes.

3.1 Closed-domain contextual faithfulness

Given supplied evidence c, the policy samples

$$y \sim \pi_\theta(\cdot \mid q, c). \tag{1}$$

A response should answer the query using claims supported by c, excluding unsupported parametric knowledge. Let $R_c(y; q, c, y^*)$ reward contextual faithfulness and task correctness. The objective is

$$J_c(\theta) = \mathbb{E}_{(q,c,y^*)\sim\mathcal{D}_c} \mathbb{E}_{y\sim\pi_\theta(\cdot|q,c)} \left[R_c(y; q, c, y^*) \right]. \tag{2}$$

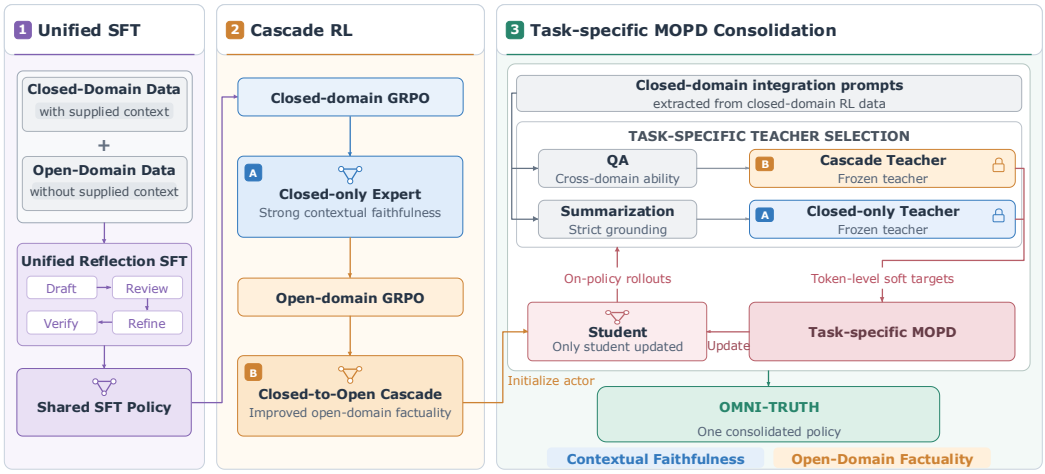


Figure 2: The three-stage OMNI-TRUTH training pipeline. Unified reflection SFT builds a shared policy; closed-domain followed by open-domain GRPO yields the Closed-only and Cascade experts. The Cascade initializes the student for task-specific MOPD on closed-domain prompts. Frozen Cascade and Closed-only teachers provide token-level guidance on student rollouts for QA and summarization, respectively; only the student is updated. Inference uses one consolidated policy.

3.2 OPEN-DOMAIN FACTUALITY

Without supplied context, the policy samples

$$y \sim \pi_\theta(\cdot \mid q). \tag{3}$$

A response should be factually correct or abstain when knowledge is insufficient. The implemented reward $R_o(y; q, y^*)$ favors correct answers and rewards incorrect answers and abstentions equally for valid-format responses (Equation 10). The training objective is

$$J_o(\theta) = \mathbb{E}_{(q,y^*)\sim\mathcal{D}_o} \mathbb{E}_{y\sim\pi_\theta(\cdot|q)} \left[R_o(y; q, y^*) \right]. \tag{4}$$

3.3 UNIFIED LEARNING GOAL

Our goal is to learn a single deployable policy that balances closed-domain faithfulness and open-domain factuality. We formulate this as a multi-objective problem:

$$\max_\theta \left(J_c(\theta),\, J_o(\theta) \right). \tag{5}$$

Our staged training seeks to balance these potentially competing objectives.

4 OMNI-TRUTH

4.1 METHOD OVERVIEW

Figure 2 summarizes the three stages of OMNI-TRUTH. First, reflection SFT on closed- and open-domain data produces a shared policy. Second, closed-domain GRPO produces a faithfulness specialist, and open-domain GRPO continues from it to produce a Closed-to-Open Cascade. Third, the Cascade checkpoint initializes a student that generates on-policy responses to closed-domain QA and summarization prompts. A frozen Cascade teacher provides token-level guidance on QA rollouts, while a frozen Closed-only teacher does the same for summarization; only the student is updated. The resulting policy handles both regimes at inference without any external model.

4.2 STAGE 1: SHARED REFLECTION SFT

Starting from a pretrained policy π_{θ_0}, we minimize the standard next-token objective over the closed- and open-domain SFT datasets $\mathcal{S}_c$ and $\mathcal{S}_o$ to obtain π_{θ_s}:

$$\mathcal{L}_{\mathrm{SFT}}(\theta_s) = -\sum_{(x,y)\in\mathcal{S}_c\cup\mathcal{S}_o}\sum_{t}\log \pi_{\theta_s}(y_t \mid x, y_{<t}), \tag{6}$$

where $x = (q, c)$ for closed-domain examples and $x = q$ otherwise, and y is the supervised reflection-and-correction response. This stage produces the shared SFT policy π_{θ_s}, which provides a common reasoning, review, and correction prior. Section 5 details data construction.

4.3 STAGE 2: SPECIALIST AND CASCADE CONSTRUCTION

Both RL stages use Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each prompt, GRPO samples a group of G responses and normalizes their rewards within the group to obtain advantages $\widehat{A}_i$. The standard clipped surrogate objective is

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\Big(\rho_{i,t}\widehat{A}_i, \mathrm{clip}(\rho_{i,t}, 1-\epsilon_l, 1+\epsilon_u)\widehat{A}_i\Big) - \beta\,\mathcal{D}_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})\Bigg], \tag{7}$$

where $\rho_{i,t} = \pi_\theta(y_{i,t} \mid x, y_{i,<t})/\pi_{\mathrm{old}}(y_{i,t} \mid x, y_{i,<t})$ is the per-token importance-sampling ratio. Here ϵ_l and ϵ_u are the lower and upper clipping offsets, and β is the KL-regularization coefficient. Implementation-specific aggregation and clipping settings are given in Appendix C.4. For RL reward computation, we train separate Qwen2.5-7B-Instruct generative reward models (GenRMs) for closed- and open-domain judgments using supervised fine-tuning.

4.3.1 CLOSED-DOMAIN GRPO

The closed-domain policy samples responses conditioned on context and query. A specialized generative reward model evaluates whether a candidate has the same meaning as the reference and contains no hallucinated material. Its judge prompt requests a binary Yes/No decision given the problem, ground truth, and candidate solution; the full template appears in Appendix A.3 (Figure 4). Let $f(y)$ indicate that the rollout contains the required completed reasoning tags, and let z_c denote the GenRM output. The implemented closed-domain reward is

$$R_c(y; q, c, y^*) = \begin{cases} 0, & f(y) = 0, \\ 2, & f(y) = 1 \text{ and } z_c = \textsc{Yes}, \\ 1, & f(y) = 1 \text{ and } \big(z_c = \textsc{No} \text{ or } z_c \notin \{\textsc{Yes}, \textsc{No}\}\big), \end{cases} \tag{8}$$

where an output outside $\{\textsc{Yes}, \textsc{No}\}$ is treated as unparsable. We initialize the actor and fixed reference policy from the shared SFT checkpoint π_{θ_s}. The resulting domain-specific objective can be written compactly as

$$\mathcal{J}_c(\theta_c; \theta_s) = \mathbb{E}_{(q,c,y^*)\sim\mathcal{D}_c,\, y\sim\pi_{\theta_c}}[R_c(y; q, c, y^*)] - \beta_c\,\mathcal{D}_{\mathrm{KL}}(\pi_{\theta_c} \,\|\, \pi_{\theta_s}), \tag{9}$$

which is optimized using Equation 7. This produces the Closed-only policy π_{θ_c} and is intended to establish assertive evidence use while penalizing contradictions and unsupported additions.

4.3.2 CASCADE OPEN-DOMAIN RL

The open-domain stage samples answers to question-only prompts. Its generative reward model receives the question, reference answer, and prediction and returns exactly one of Correct, Incorrect, or Not Attempted. Its full judge prompt appears in Appendix A.3 (Figure 4). With $f(y)$ defined above and the GenRM output denoted by z_o, the implemented reward is

$$R_o(y; q, y^*) = \begin{cases} 0, & f(y) = 0, \\ 2, & f(y) = 1 \text{ and } z_o = \textsc{Correct}, \\ 0.5, & f(y) = 1 \text{ and } z_o \in \{\textsc{Incorrect}, \textsc{Not Attempted}\}, \\ 1, & f(y) = 1 \text{ and } z_o \notin \{\textsc{Correct}, \textsc{Incorrect}, \textsc{Not Attempted}\}. \end{cases} \tag{10}$$

Appendix C.3 summarizes both domains' exact reward mapping and explains the shaping and fallback values. Continuing from the Closed-only policy, we initialize both the actor and fixed reference policy from π_{θ_c}. The corresponding objective is

$$\mathcal{J}_o(\theta_{co};\theta_c) = \mathbb{E}_{(q,y^*)\sim\mathcal{D}_o,\, y\sim\pi_{\theta_{co}}}[R_o(y;q,y^*)] - \beta_o\, \mathcal{D}_{\mathrm{KL}}(\pi_{\theta_{co}} \,\|\, \pi_{\theta_c}), \tag{11}$$

Optimizing this objective with Equation 7 yields the Closed-to-Open Cascade $\pi_{\theta_{co}}$.

The complete specialist-and-Cascade path is therefore

$$\theta_0 \xrightarrow{\text{SFT}} \theta_s \xrightarrow{\text{Closed RL}} \theta_c \xrightarrow{\text{Open RL}} \theta_{co}. \tag{12}$$

The resulting Cascade $\pi_{\theta_{co}}$ carries the open-domain adaptation on top of the closed-domain policy. This ordering first establishes evidence-conditioned answering and then improves open-domain factual answering, thereby separating the acquisition of the two capabilities. However, subsequent open-domain training may weaken the contextual faithfulness acquired during closed-domain training, motivating the consolidation stage that follows.

4.4 Stage 3: task-specific policy consolidation

The Cascade's stronger grounded QA and the Closed-only policy's stronger source faithfulness motivate task-specific consolidation (Table 1). We use $\pi_{\theta_{co}}$ as both the student initialization and the QA teacher, while retaining π_{θ_c} as the Closed-only teacher. To reduce the training time required for consolidation over the full training corpus, we focus MOPD on these complementary capabilities, selecting the corresponding QA and summarization examples from the closed-domain RL data: $\mathcal{D}_{\mathrm{int}} = \mathcal{D}_{\mathrm{QA}} \cup \mathcal{D}_{\mathrm{Sum}}$. Prompts are sampled with equal probability from the QA and summarization subsets. We select the teacher parameters as

$$\tau(x) = \begin{cases} \theta_{co}, & x \in \mathcal{D}_{\mathrm{QA}}, \\ \theta_c, & x \in \mathcal{D}_{\mathrm{Sum}}. \end{cases} \tag{13}$$

Integration uses only closed-domain prompts to reconcile grounded QA and summarization. Cascade initialization transfers open-domain adaptation to the student, while the frozen Cascade teacher guides QA to help retain these capabilities.

For each prompt x, the student generates an on-policy trajectory $y \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x)$. At prefix $s_t = (x, y_{<t})$, let $\mathcal{C}_t$ be the pre-update student's top-k candidate tokens and define the normalized student weight for each candidate as

$$w_t(v) = \frac{\pi_{\theta_{\mathrm{old}}}(v \mid s_t)}{\sum_{u\in\mathcal{C}_t} \pi_{\theta_{\mathrm{old}}}(u \mid s_t)}, \qquad v \in \mathcal{C}_t. \tag{14}$$

The routed teacher supplies the log probability of the same candidates. We then construct the candidate-level distillation signal

$$A_t(v) = w_t(v)\left[\log \pi_{\tau(x)}(v \mid s_t) - \log \pi_{\theta_{\mathrm{old}}}(v \mid s_t)\right]. \tag{15}$$

This signal is inspired by reverse KL: only w_t is normalized over the retained support; both log probabilities use full-vocabulary normalization. The support, weights, and $A_t(v)$ are fixed during each student update. With clipping offset $\epsilon > 0$, define the candidate ratio and clipped surrogate as

$$r_t(v) = \frac{\pi_\theta(v \mid s_t)}{\pi_{\theta_{\mathrm{old}}}(v \mid s_t)}, \tag{16}$$

$$h_t(v) = \min\{r_t(v)A_t(v),\, \mathrm{clip}(r_t(v), 1-\epsilon, 1+\epsilon)A_t(v)\}. \tag{17}$$

With dual-clip coefficient $d > 1$, let $\widetilde{h}_t(v) = \max\{h_t(v), dA_t(v)\}$ when $A_t(v) < 0$, and $\widetilde{h}_t(v) = h_t(v)$ otherwise. We minimize

$$\mathcal{L}_{\mathrm{MOPD}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}_{\mathrm{int}},\, y\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)}\left[\frac{1}{|y|}\sum_t \sum_{v\in\mathcal{C}_t} \widetilde{h}_t(v)\right], \tag{18}$$

where the expectation over x uses the balanced sampling scheme. Appendix C.5 specifies k, ϵ, and d. Candidate contributions are summed, then averaged over response tokens and examples. Only the current student receives gradients. After training, both teachers and the task selector are discarded, leaving the single Omni-Truth policy.

5 POST-TRAINING DATA CONSTRUCTION

Our post-training data support policy SFT, generative reward-model (GenRM) training, and online RL. The training sets comprise 50,455 policy-SFT examples, 411,342 GenRM judgment examples, and 8,231 RL prompts. Table 3 in Appendix A lists domain counts and validation splits.

5.1 UNIFIED REFLECTION SFT DATA

We use Qwen3-235B to synthesize reflection responses for policy SFT. Closed-domain data span question answering, summarization, context-conditioned generation, extraction and classification, rewriting and transformation, and analysis and reasoning; open-domain data draw from Natural Questions (Kwiatkowski et al., 2019), Chinese WebQA (Li et al., 2016), and CRAG (Yang et al., 2024). Our pipeline first filters queries by domain: closed-domain answers must depend on the supplied context, while open-domain questions must be unambiguous and concern temporally stable facts. We append an instruction to each query asking the model to draft, reflect on, and revise its answer within the reasoning process before producing the final response. We then apply response-quality, prompt-leakage, format, deduplication, and contamination filters. Appendices A.1 and A.2 detail the construction process and task composition.

5.2 GENERATIVE REWARD MODELS AND RL DATA

We synthesize GenRM training examples that pair a prompt, reference answer, and candidate response with either a closed-domain Yes/No label or an open-domain Correct/Incorrect/Not Attempted label. To validate these reward models, we evaluate their judgments on policy-generated responses. The closed-domain GenRM reaches 92.90% accuracy on the HHEM-labeled response set, while the open-domain GenRM reaches 95.08% accuracy across Chinese SimpleQA, CRAG, and WebQA. The supervision and validation protocols are detailed in Appendices A.3 and A.4. Section 6 further evaluates the practical effectiveness of these reward models through the downstream performance of policies trained with their rewards.

We extract the online RL prompts and references from source examples during SFT data construction. Closed-domain prompts cover summarization, question answering, and other context-conditioned tasks; open-domain prompts come from WebQA and CRAG. Each record contains a prompt, reference answer, and domain tag, while the policy generates responses during training. We use validation rewards to identify candidate checkpoints for evaluation. The task composition and leakage audit are described in Appendices A.5 and A.6.

6 EXPERIMENTS

We evaluate whether task-specific consolidation improves the balance between closed-domain faithfulness and open-domain factuality, and whether any aggregate gain is concentrated in a single benchmark or shared across both regimes.

6.1 COMPARED METHODS

The controlled comparison starts from Qwen3-8B and the shared Omni-SFT checkpoint at step 3945, which is also the SFT baseline in Table 1. Each in-house RL or distillation stage trains for 15 epochs. Single-domain controls apply only Closed RL or Open RL. Dual-domain baselines include Combined RL, Closed-to-Open and Open-to-Closed Cascade, and standard MOPD with independently trained Closed-only and Open-only teachers. OMNI-TRUTH instead initializes its student from the selected Closed-to-Open Cascade checkpoint and routes the Cascade and Closed-only teachers by task on closed-domain integration data, as described in Section 4. We also report compatible local adaptations or public checkpoints of the two FaithRL systems, TruthRL, and Binary RAR as specialized reference points (Gui et al., 2026; Nie et al., 2026; Wei et al., 2026; Chen et al., 2025). Appendix D details their selection, implementation status, and observed limitations. For cross-model context, Table 1 also includes the frontier large model DeepSeek V4 Flash evaluated on the same six test sets.

Table 1: Benchmark results (%). * marks a single reference model or checkpoint; other rows average three runs or teacher configurations. Checkpoint scores average three inference repetitions. Bold marks the best Qwen3-8B-based value per column. FG denotes FACTS Grounding and C-SQA denotes Chinese SimpleQA.

	Closed-domain			Open-domain			Overall
Method	**FG↑**	**HHEM↑**	**MultiHop↑**	**CRAG↑**	**WebQA↑**	**C-SQA↑**	**Macro↑**
Frontier large model							
DeepSeek V4 Flash*	79.26	98.44	81.61	68.58	85.60	80.43	82.32
Base and SFT references							
Base Qwen3-8B*	27.74	91.50	66.63	34.50	66.80	36.74	53.98
Omni-SFT (step 3945)*	43.17	93.97	70.10	38.10	69.30	35.60	58.37
Closed-domain methods							
Closed-only RL	75.15	98.30	75.09	36.56	69.52	39.23	65.64
FaithRL (Gui et al.)	42.69	93.66	74.62	36.45	69.46	37.94	59.13
FaithRL (Nie et al.)	44.59	94.93	76.88	38.11	70.39	38.48	60.56
Open-domain methods							
Open-only RL	41.59	93.90	75.10	38.71	71.68	40.23	60.20
TruthRL-style	44.38	93.96	75.97	39.49	71.97	**41.15**	61.15
Binary RAR*	29.96	92.58	61.45	36.43	68.12	36.35	54.15
Dual-domain integration							
Combined RL	69.68	97.91	74.54	38.93	73.00	39.91	65.66
Standard MOPD	**75.61**	98.11	75.34	38.66	72.89	40.18	66.80
Cascade (C→O)	70.30	97.82	79.43	39.78	73.01	40.44	66.80
Reverse Cascade (O→C)	73.11	**98.53**	76.46	38.90	72.50	39.54	66.51
OMNI-TRUTH	73.13	97.95	**80.16**	**40.37**	**73.34**	40.31	**67.54**

6.2 Benchmarks and metrics

We evaluate on FACTS Grounding (FG) (Jacovi et al., 2025), a source–response suite scored with HHEM-2.1-Open (Li et al., 2024), MultiHopRAG (Tang & Yang, 2024), CRAG (Yang et al., 2024), WebQA (Li et al., 2016), and Chinese SimpleQA (C-SQA) (He et al., 2024). In our protocol, FG reports the fraction of responses with no sentence judged unsupported or contradictory; HHEM reports the fraction with support scores ≥ 0.5. MultiHopRAG uses its public evaluator's answer-token-overlap accuracy. CRAG, WebQA, and C-SQA use the harmonic mean of overall correctness and attempted-answer accuracy. *Close avg.* averages FG, HHEM, and MultiHopRAG; *Open avg.* averages CRAG, WebQA, and C-SQA; *Macro* averages all six. Appendix B details scoring and abstention diagnostics.

6.3 Overall comparison

Single-domain training produces unbalanced policies. Closed-only RL reaches 75.15 on FG, whereas Open-only RL scores 41.59 but performs better on all three open-domain benchmarks (Table 1). These different performance profiles motivate integration across regimes. The results also support the practical utility of the domain-specific GenRM rewards for RL, although policy scores reflect the combined effects of rewards, data, and optimization.

Cascade and standard MOPD have complementary strengths. Sequential Closed-to-Open training raises MultiHopRAG from 75.09 for Closed-only RL to 79.43 for the Cascade, indicating beneficial cross-domain transfer, but reduces FG from 75.15 to 70.30. Standard MOPD recovers stronger FG (75.61) from its specialized teachers, yet reaches only 75.34 on MultiHopRAG; its independently trained teachers do not carry the Cascade's sequential transfer.

OMNI-TRUTH consolidates these advantages. Task-specific consolidation improves FG over the Cascade from 70.30 to 73.13 while improving MultiHopRAG from 79.43 to 80.16, exceeding stan-

dard MOPD by 4.82 points. Among Qwen3-8B-based methods, it achieves the best overall Macro score of 67.54. Its Close and Open averages are 83.75 and 51.34, compared with 82.52 and 51.08 for the Cascade, respectively. Its Close average is only 2.69 points below DeepSeek V4 Flash (86.44). The larger Open gap (51.34 versus 78.20) may partly reflect differences in model scale and parametric knowledge, which this comparison does not isolate. Training trajectories for SFT, closed-domain RL, open-domain RL, and MOPD are shown in Appendix E.6.

Response patterns reflect the same trade-off. Selected-checkpoint response distributions provide a complementary view of this trade-off (Appendix E.7, Table 18). Closed-only RL shortens grounding responses relative to SFT but increases explicit MultiHopRAG refusals. The Cascade reduces these refusals and lengthens short-form answers; OMNI-TRUTH largely retains this answer-coverage pattern while recovering FG. We use these shifts as descriptive evidence about response behavior.

6.4 ABLATION STUDY

Table 2: MOPD ablations (three-run means, %). FG denotes FACTS Grounding, C-SQA denotes Chinese SimpleQA, and Macro equally weights the six benchmarks. Higher is better.

Variant	FG	HHEM	MultiHop	CRAG	WebQA	C-SQA	Macro
Naive Cascade MOPD	**76.14**	97.82	76.05	39.79	73.14	40.21	67.19
OMNI-TRUTH (SFT init)	72.88	97.83	**80.51**	39.36	71.17	38.84	66.77
OMNI-TRUTH (summary-only)	75.13	**98.15**	77.01	38.42	72.48	39.94	66.86
OMNI-TRUTH	73.13	97.95	80.16	**40.37**	**73.34**	**40.31**	**67.54**

Table 2 compares three variants. Naive Cascade MOPD starts from SFT, routing closed-domain prompts to Closed-only and open-domain prompts to Cascade. Our method gains 4.11 points on MultiHopRAG and 0.35 in Macro, although the naive variant scores higher on FG. This suggests that domain-based routing misses part of the intended complementarity: it preserves strict grounding but provides no Cascade guidance on closed-domain QA to transfer the gains from two-stage training. OMNI-TRUTH (SFT init) changes only initialization; Cascade initialization improves all three open-domain scores. The summary-only variant retains Cascade initialization and the Closed-only teacher but removes QA prompts and the Cascade teacher. Relative to the full method, FG rises to 75.13, but MultiHopRAG falls to 77.01 and all three open-domain scores decline, indicating weaker retention of these capabilities. All variants train for 15 epochs on their respective data, with different update counts; the comparisons assess complete configurations rather than isolating teacher routing.

7 LIMITATIONS

Our study focuses on hallucination in source-grounded generation and knowledge-intensive question answering. Mathematical reasoning and agentic settings offer future opportunities to study multi-step computation, planning, tool use, and interactions with external environments. Our post-training experiments use Qwen3-8B; larger models offer a further test of how these gains scale. Extending the training objectives and evaluation framework to these settings and model scales is an important direction for future work.

8 CONCLUSION

We introduced OMNI-TRUTH for closed-domain faithfulness and open-domain factuality. It combines shared reflection SFT, closed-to-open sequential RL, and task-specific multi-teacher on-policy distillation. Across six benchmarks, it achieves 67.54 Macro, versus 66.80 for both standard MOPD and the Cascade. Among the evaluated Qwen3-8B-based methods, it achieves the highest Close and Open averages and leads on MultiHopRAG, CRAG, and WebQA. These results support task-specific consolidation as a way to combine strict closed-domain grounding with the gains of sequential closed-to-open training.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and language editing, literature discovery and retrieval, methodological review, interpretation of experimental results, and supplementary code review. They were also used to construct and filter parts of the training data, as described in Section 5. The authors reviewed the AI-assisted content and take full responsibility for the manuscript, code, analyses, and reported results.

ETHICS STATEMENT

This study uses automatically generated training data and model-based evaluation, both of which may contain factual errors and linguistic or cultural biases. We checked training-data quality and benchmark overlap using the filtering and audit procedures documented in Appendix A.

REPRODUCIBILITY STATEMENT

The appendices document data composition, construction and evaluator prompts, filtering and overlap checks, training configurations, checkpoint selection, inference settings, scoring rules, and per-run benchmark results. The controlled in-house methods use the Qwen3-8B backbone; the comparisons identify their different training prompts, teachers, and optimization settings. We will publicly release the data and code for this work upon acceptance.

REFERENCES

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. Is factuality enhancement a free lunch for LLMs? better factuality can lead to worse context-faithfulness. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=asGQQc7gNo.

Tong Chen, Akari Asai, Luke Zettlemoyer, Hannaneh Hajishirzi, and Faeze Brahman. Train for truth, keep the skills: Binary retrieval-augmented reward mitigates hallucinations. *arXiv preprint arXiv:2510.17733*, 2025. URL https://arxiv.org/abs/2510.17733.

Runquan Gui, Yafu Li, Xiaoye Qu, Ziyan Liu, Yeqiu Cheng, and Yu Cheng. FaithRL: Learning to reason faithfully through step-level faithfulness maximization. *arXiv preprint arXiv:2602.03507*, 2026. URL https://arxiv.org/abs/2602.03507.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Xuepeng Liu, Dekai Sun, Shirong Lin, Zhicheng Zheng, Xiaoyong Zhu, Wenbo Su, and Bo Zheng. Chinese SimpleQA: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*, 2024. URL https://arxiv.org/abs/2411.07140.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. doi: 10.1145/3703155. URL https://arxiv.org/abs/2311.05232.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The FACTS grounding leaderboard: Benchmarking LLMs' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025. URL https://arxiv.org/abs/2501.03200.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

Saurav Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

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/.

Junyi Li and Hwee Tou Ng. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.24630.

Miaoran Li, Rogger Luo, and Ofer Mendelevitch. HHEM-2.1-Open, 2024. URL https://huggingface.co/vectara/hallucination_evaluation_model.

Mingchen Li, Hanzhi Zhang, Heng Fan, Junhua Ding, and Yunhe Feng. Harmful factuality: LLMs correcting what they shouldn't. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 896–912, 2026. doi: 10.18653/v1/2026.findings-eacl.46. URL https://aclanthology.org/2026.findings-eacl.46/.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *arXiv preprint arXiv:1607.06275*, 2016. URL https://arxiv.org/abs/1607.06275.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026. URL https://arxiv.org/abs/2606.30406.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020. doi: 10.18653/v1/2020.acl-main.173.

Shuo Nie, Hexuan Deng, Chao Wang, Ruiyu Fang, Xuebo Liu, Shuangyong Song, Yu Li, Min Zhang, and Xuelong Li. Stop rewarding hallucinated steps: Faithfulness-aware step-level reinforcement learning for small reasoning models. *arXiv preprint arXiv:2602.05897*, 2026. URL https://arxiv.org/abs/2602.05897.

Baochang Ren, Shuofei Qiao, Ningyu Zhang, Da Zheng, and Huajun Chen. KnowRL: Exploring knowledgeable reinforcement learning for factuality. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39640–39658, 2026. doi: 10.18653/v1/2026.acl-long.1840. URL https://aclanthology.org/2026.acl-long.1840/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Yixuan Tang and Yi Yang. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t4eB3zYWBK.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.08401.

Pengbo Wang, Chaozhuo Li, Chenxu Wang, Liwen Zheng, Litian Zhang, and Xi Zhang. One SPACE to rule them all: Jointly mitigating factuality and faithfulness hallucinations in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://arxiv.org/abs/2506.11088.

Yudong Wang, Zhe Yang, Wenhan Ma, Zhifang Sui, and Liang Zhao. Enhancing reliability across short and long-form QA via reinforcement learning. *arXiv preprint arXiv:2512.08944*, 2025b. URL https://arxiv.org/abs/2512.08944.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. TruthRL: Incentivizing truthful LLMs via reinforcement learning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.25760.

Qiming Xie, Wenjie Zheng, Xiangqing Shen, and Rui Xia. FARCA: Fact-aligned reliability-aware credit assignment for reinforcement learning with factual supervision. *arXiv preprint arXiv:2608.24350*, 2026. URL https://arxiv.org/abs/2608.24350.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. CRAG—comprehensive RAG benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.04744.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-Tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, 2024. doi: 10.18653/v1/2024.naacl-long.394. URL https://arxiv.org/abs/2311.09677.

A Data Construction Details

This appendix details the synthesis, filtering, contamination-control, and evaluation procedures used in the reported experiments.

Table 3: Data used in the reported experiments. SFT counts are after quality, deduplication, and leakage filtering. GenRM denotes generative reward-model data. RL counts refer to the final training prompts.

Pipeline stage	Regime	Supervision	Train	Validation	Output or use
Policy SFT	Closed	Reflection response	25,528	–	Token sequence
Policy SFT	Open	Reflection response	24,927	–	Token sequence
GenRM	Closed	Judgment example	200,182	–	Yes/No
GenRM	Open	Judgment example	211,160	–	A/B/C
Policy RL	Closed	Prompt + reference	4,971	500	Online rollout
Policy RL	Open	Prompt + reference	3,260	992	Online rollout

A.1 SFT query filtering and response synthesis

Domain-specific query filtering. The two regimes require different candidate-selection criteria. For closed-domain SFT, we retain a query only when a correct response requires the supplied context and cannot be reliably obtained from parametric knowledge alone. This includes cases in which the model's prior knowledge is insufficient, ambiguous, or inconsistent with the supplied evidence. The filter therefore removes examples that remain answerable after the context is ignored. For open-domain SFT, we discard questions about rapidly changing or transient facts, such as current office holders or latest records. We retain questions that are fact-based, unambiguous, and anchored in stable knowledge, including historical events, scientific facts, and established definitions.

Initial response generation. We generate responses with Qwen3-235B; the synthesis script uses Qwen3-235B-A22B-Thinking-2507 with thinking enabled. We append a short instruction to the original query, retaining the supplied context for closed-domain examples. The suffix asks the model to ensure factual accuracy, first draft an answer within its reasoning, and then reflect on and revise it before returning the final response. English and Chinese versions of the suffix are selected according to the query language. The suffix also asks the model not to mention or quote these synthesis instructions in its reasoning or final answer. We retain the resulting reasoning trace and final response as a single SFT target; the reflection steps are not separate generation calls.

Post-generation filtering. We use Qwen3-8B as an automatic verifier with domain-specific criteria. A closed-domain response is rejected if its final answer contradicts the context or introduces claims that the context does not support. An open-domain response is rejected if comparison with the reference answer reveals a factual error or fabrication. A separate prompt-leakage check removes traces that explicitly refer to hidden synthesis instructions instead of presenting natural self-reflection. We then enforce the expected reasoning–answer format, retain at most one accepted response per prompt, and apply the contamination checks described in Appendix A.6.

A.2 Closed-domain SFT composition

The synthesized closed-domain SFT set contains 25,528 examples. We classify the retained records by their requested output form; records inherited from broad residual categories are reclassified at the sample level using a deterministic task-intent rubric. As shown in Figure 3, the resulting categories are question answering (8,587), context-conditioned generation (7,564), summarization (6,046), analysis and reasoning (1,405), rewriting and transformation (969), and extraction and classification (957).

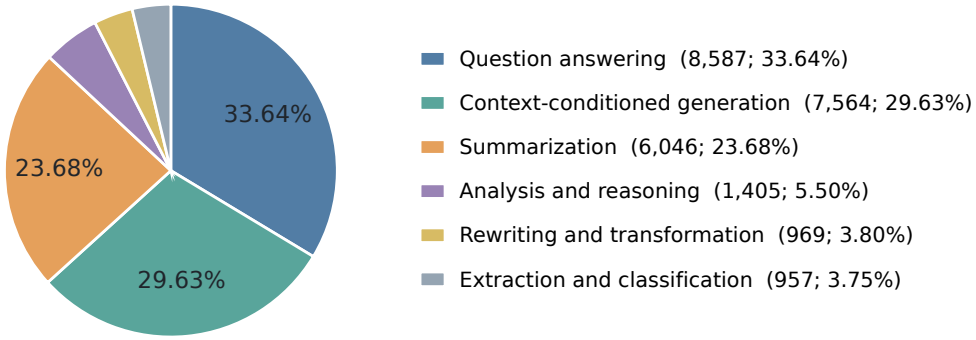


Figure 3: Task composition of the synthesized closed-domain policy-SFT data after quality, deduplication, and leakage filtering. The legend lists category counts and shares in descending order.

Table 4: Label composition of the GenRM training data. Percentages are computed within each domain.

Domain	Label	Meaning	Count	Share
Closed	Yes	Semantically matches the reference without hallucinated content	94,787	47.35%
Closed	No	Mismatches the reference or contains hallucinated content	105,395	52.65%
Open	A	Correct	99,990	47.35%
Open	B	Incorrect	109,585	51.90%
Open	C	Not Attempted	1,585	0.75%
		Closed-domain total	200,182	100.00%
		Open-domain total	211,160	100.00%

A.3 Generative reward-model supervision

Each GenRM training example consists of a problem or question, a reference answer, a candidate response, and a judgment label. The first three fields form the model input, with the reference answer presented as "Ground Truth" in the closed-domain prompt and "Correct Answer" in the open-domain prompt. The judgment label serves as the supervision target. Table 4 summarizes the label distribution: the closed-domain Yes/No labels are nearly balanced, while the open-domain data contain similar proportions of Correct and Incorrect examples, with Not Attempted accounting for 0.75%.

The two datasets are different supervision views over a largely shared tuple pool, rather than independently constructed collections. After whitespace normalization, 199,934 unique (query, reference, candidate) tuples occur in both sets, covering 99.88% of the closed-domain tuples and 94.68% of the open-domain tuples. On these shared tuples, Yes maps to A (Correct) and No maps to B (Incorrect). The open-domain set additionally contains all Not Attempted examples and a smaller number of additional Correct and Incorrect examples. This shared data design reflects the primary purpose of GenRM fine-tuning: adapting the models to produce judgments in the prescribed output format for each reward interface. We therefore do not partition the supervision by closed- versus open-domain data provenance. Instead, we distinguish the two GenRMs through their judgment prompts and output label spaces: Yes/No for the closed-domain reward interface and Correct/Incorrect/Not Attempted for the open-domain interface.

Table 5: GenRM validation results on benchmark responses. Values are percentages. Aggregate accuracy is computed over all evaluation examples for the corresponding judge.

Judge	Checkpoint	HHEM	C-SQA	CRAG	WebQA	Aggregate	Macro-F1
Closed	5,000	91.85	–	–	–	91.85	62.60
Closed	10,000	92.52	–	–	–	92.52	60.66
Closed	15,640	**92.90**	–	–	–	**92.90**	60.95
Open	5,000	–	93.30	93.33	95.77	94.13	83.39
Open	10,000	–	94.23	94.13	96.63	95.00	82.45
Open	16,500	–	**94.33**	**94.20**	**96.70**	**95.08**	**83.82**

Figure 4 shows the prompt templates used to convert each judgment tuple into generative classification supervision. The output target is a single Yes/No label for the closed-domain GenRM and a single A/B/C label for the open-domain GenRM.

Closed-domain GenRM prompt
You are an impartial referee expert. Given a problem and the ground truth, your task is to check whether the candidate answer is as correct as the ground truth. This means the candidate solution has the same meaning as the ground truth and does not contain any hallucinated content.
Problem: {`problem`}
Ground Truth: {`ground truth`}
Candidate Solution: {`candidate solution`}
Does the candidate solution match the ground truth (Yes/No)?

Open-domain GenRM prompt
Please evaluate whether the model's response is correct based on the given question, the correct answer, and the model's predicted answer. Rate the result as Correct, Incorrect, or Not Attempted. Reply with only one of A, B, or C, without apologizing or correcting any mistakes; only evaluate the response.
Question: {`question`}
Correct Answer: {`correct answer`}
Predicted Answer: {`predicted answer`}
Rate this predicted answer as one of the following:
A: [Correct]
B: [Incorrect]
C: [Not Attempted]
Return only the letter "A," "B," or "C" without adding any other text.

Figure 4: Prompt templates used to train the two generative reward models. Template input fields are shown in braces.

A.4 Generative reward-model validation

We evaluate the GenRMs on candidate responses produced by policies in our benchmark evaluation pipeline, rather than on synthesized GenRM training tuples. For the closed-domain judge, we convert 4,984 nonempty HHEM responses and their hallucination labels into the binary prompt format in Figure 4. For the open-domain judge, we use 9,000 responses distributed equally across Chinese SimpleQA, CRAG, and WebQA, with the existing Correct, Incorrect, and Not Attempted labels as references. All judge outputs are decoded as a single label; every output in these evaluations is valid and parseable.

Table 5 compares three checkpoints from each GenRM. Performance is stable across the evaluated checkpoints. The final closed-domain checkpoint obtains 92.90% accuracy on HHEM, and the final open-domain checkpoint obtains 95.08% aggregate accuracy and 83.82% macro-F1, with consistent accuracy across all three open-domain benchmarks. We exclude FACTS Grounding from this standalone GenRM validation because its strict all-claims-supported criterion differs from the answer-equivalence supervision used by the closed-domain GenRM; policy behavior on FACTS Grounding is instead included in the downstream six-benchmark evaluation. These results support using the two GenRMs as task-specific reward functions during online RL; the downstream effect of their rewards is evaluated through policy performance on all six benchmarks in Section 6.

A.5 RL POLICY-DATA COMPOSITION

The final RL training splits contain 4,971 closed-domain prompts and 3,260 open-domain prompts. We aggregate the English and Chinese closed-domain shards by requested task type, yielding summarization, question answering, and other context-conditioned tasks. The open-domain split is classified by its source benchmark. Figure 5 reports both compositions; the validation partitions are not included in these counts.

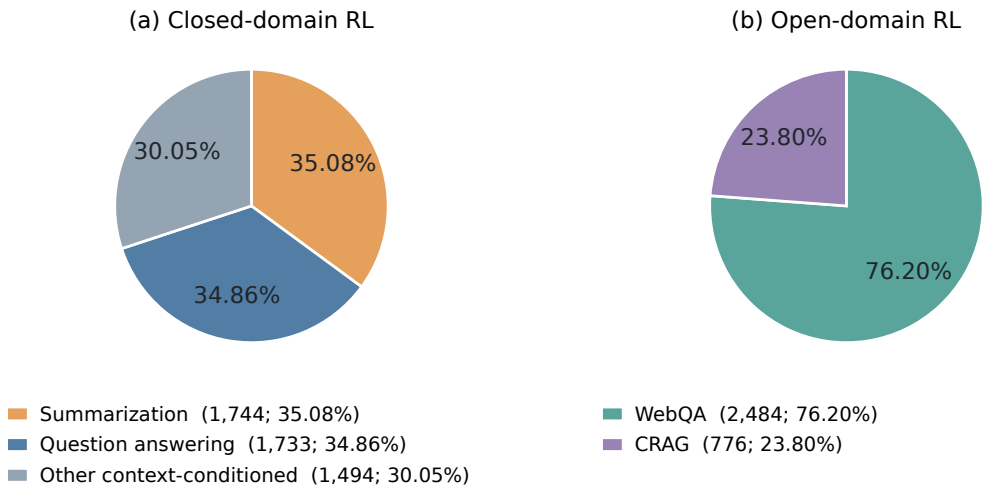


Figure 5: Composition of the closed- and open-domain RL training prompts. The closed-domain split is grouped by task type, while the open-domain split is grouped by benchmark.

A.6 THREE-STAGE LEAKAGE DETECTION PIPELINE

Before leakage detection, we normalize each example and assign it to question–answer (QA) matching, source matching, or both, according to the information it contains. QA matching identifies potentially repeated questions and answer targets, while source matching identifies overlapping source material. Grounded-QA datasets, including MultiHopRAG, undergo both checks. We normalize source URLs and divide long contexts into chunks to support lexical and embedding-based retrieval. The three stages below retrieve candidate overlaps and determine whether they constitute leakage.

Stage 1: lexical filtering. We build separate question and source indexes over the training records and retrieve exact question matches, fuzzy question matches, and source matches. The outputs preserve the originating train and evaluation record identifiers. This stage captures punctuation variants, minor edits, and other high-confidence lexical near-duplicates.

Stage 2: embedding candidate generation. We encode each indexable evaluation record with BAAI/bge-m3 and perform same-language cosine retrieval against the training indexes. QA retrieval uses the matching question as its primary representation and records question cosine, lexical similarity, answer equality, and answer cosine. A question cosine of at least 0.92 defines a strong candidate; scores from 0.80 up to, but excluding, 0.92 enter semantic review. Source retrieval combines document cosine, maximum and average top-chunk cosine, chunk overlap, and shared normalized URLs. A source pair is strong when its document cosine is at least 0.90, its maximum chunk cosine is at least 0.92, or a shared URL is supported by similarity above the review threshold. Additional review candidates satisfy document cosine of at least 0.84, maximum chunk cosine of at least 0.86, chunk overlap of at least 0.50, or a shared URL. The detector emits candidate reports and similarity evidence; final leakage labels are assigned in the next stage.

Stage 3: structured semantic adjudication. For each retrieved pair, we recover the original train and evaluation examples and obtain a structured LLM judgment. The SFT adjudication script uses GPT-5 (`gpt-5-2025-08-07`) with temperature zero. QA pairs are labeled as leakage when they

share both the question target and answer target, or as answer conflicts when the question target matches but the answers differ. Source pairs are labeled according to whether they share an underlying source and task target; a shared source with a different task target is recorded as source overlap. In the SFT audit, confirmed leakage, source overlaps, and answer conflicts are removed from training. Candidate pairs, supporting evidence, judgments, and removal decisions are retained for auditability.

A.7 Audit outcomes and evaluation policy

The SFT-to-benchmark audit produced 299 adjudicated QA candidate pairs: 74 were true leakage, four were answer conflicts, and 221 were false positives. These decisions removed 78 SFT examples. The source audit produced 26 candidate pairs: four true leakage, four source-overlap leakage, and 18 false positives, removing eight additional examples. In total, 86 examples were removed before forming the 50,455-example SFT set reported in Table 3.

For the RL policy splits, the same lexical, embedding, and semantic-review workflow identified within-split duplicates and cross-split contamination. The audit identifies 15 duplicate closed-domain training rows, 27 duplicate open-domain training rows, and 21 duplicate open-domain validation rows. The embedding review identified 17 closed-domain true-leakage pairs and 48 open-domain true-leakage pairs plus one answer conflict; deduplication removed 13 unique closed-domain training rows and 46 unique open-domain training rows. The open-domain RL review uses GPT-5 (`gpt-5-2025-08-07`); its single answer-conflict case is also removed after manual review. The closed-domain RL review uses GPT-5.4-mini to distinguish shared sources, task templates, and question targets in full-prompt candidates. It removes a training example only when both the source instance and question target match; source-only overlap is retained for review rather than automatically removed. A separate RL-train-to-external-evaluation audit found two open-domain overlaps, which were removed before the final RL splits were formed. This final audit uses GPT-5 (`gpt-5-2025-08-07`).

B Evaluation Protocol

For the controlled RL and MOPD comparisons, we evaluate the validation-selected checkpoints over six fixed test sets. This protocol is designed for controlled comparison within our study; it should not be interpreted as reproducing the official leaderboard protocol of every source benchmark. In particular, our CRAG subset is evaluated as closed-book question answering without retrieval, and the MultiHopRAG score uses the answer-overlap rule described below. Table 6 summarizes the inputs, sample counts, evaluators, and reported metrics. The controlled suite contains 5,747 raw examples and 5,739 evaluated examples per inference repetition. We use the same 2,548-example MultiHopRAG subset across the controlled runs, excluding eight prompts that exceeded the original inference prompt-length limit. Runs generated on all 2,556 examples are restricted to this fixed subset before aggregation. RL and MOPD checkpoints are selected from validation-reward evaluations during training; the six test sets are used to report the performance of the selected checkpoints.

B.1 Inference and repeated evaluation

For each evaluated checkpoint, we generate three stochastic inference repetitions over all six test sets. These repetitions measure sampling and evaluator variation for one checkpoint and are not independent training seeds. Generation uses temperature 0.6, top-p 0.95, top-k 20, presence penalty 1, a maximum of 16,384 generated tokens, and enabled hidden thinking. The evaluated model receives a single user message constructed from the benchmark prompt; reference answers and evaluator-only fields are not exposed to it. Hidden thinking and the final response are stored separately, and only the final response is scored. For the three short-answer QA datasets, residual text preceding `[unused17]` or `</think>` is removed before judging.

B.2 Grounding and MultiHopRAG scoring

For FACTS Grounding, the LLM judge labels each response sentence as supported, unsupported, contradictory, or not requiring attribution. A sample receives a grounding score of zero if any sen-

Table 6: The six-dataset local evaluation suite. Example counts for the controlled comparison are per inference repetition; MultiHopRAG contains 2,556 raw examples, of which eight exceed the inference prompt-length limit. "LLM judge" denotes the model-based evaluator. The correctness–attempt harmonic mean is defined in Equation 22 and is not standard token-level F1.

Dataset	**Examples**	**Evaluator**	**Information given to the evaluated model and reported metric**
FACTS Grounding	860	LLM judge	User request and context document; sample-level grounded rate
Source–response suite	831	HHEM-2.1-Open	Source text and generation request; thresholded faithful rate
MultiHopRAG	2,548	Lexical rule	Question and supplied evidence passages; answer-token-overlap accuracy
Chinese SimpleQA	500	LLM judge	Chinese question without retrieved evidence; Correct/Incorrect/Not Attempted and correctness–attempt HM
CRAG	500	LLM judge	English query without retrieved evidence; Correct/Incorrect/Not Attempted and correctness–attempt HM
WebQA	500	LLM judge	Chinese question without retrieved evidence; Correct/Incorrect/Not Attempted and correctness–attempt HM

tence is unsupported or contradictory and one otherwise. The reported runs use the same GPT-5.4-mini judge and temperature-zero decoding as the QA evaluation. We report

$$\mathrm{FG}_{\mathrm{grounded}} = 1 - \frac{N_{\mathrm{ungrounded}}}{N}. \tag{19}$$

This is a sample-level all-claims-supported rate, not a claim-level support percentage. Because an empty response or a response containing no attributable claim can pass this rule, the score is interpreted together with response length and abstention diagnostics. Temporary judge failures are repaired before aggregation; they are not treated as hallucinations.

For HHEM, a local detector assigns a continuous support score to the source and response pair. Scores below 0.5 are labeled hallucinated and scores at least 0.5 faithful. We report the fraction labeled faithful. This is consistency under a fixed detector and threshold rather than an estimate of absolute human factuality.

For MultiHopRAG, we follow the answer-matching rule in the benchmark's public QA evaluator (Tang & Yang, 2024)[1]: the reference answer and final response are lowercased and split on whitespace. A sample is correct when the two token sets have a nonempty intersection. The reported accuracy is the fraction of samples satisfying this condition. We fix the scoring subset to the 2,548 queries retained after the eight prompt-length exclusions and use these same query identifiers for every controlled run, including runs that generated all 2,556 responses.

B.3 Open-domain QA scoring

For Chinese SimpleQA, CRAG, and WebQA, an LLM judge compares the question, reference answer, and final response and assigns A = Correct, B = Incorrect, or C = Not Attempted. These are evaluator output labels rather than multiple-choice answers. All reported runs use GPT-5.4-mini as judge. The judge uses temperature zero and up to three explicit retries. Let N_{cor}, N_{inc}, and N_{na}

[1] https://github.com/yixuantt/MultiHop-RAG/blob/main/qa_evaluate.py

denote their corresponding counts, with $N = N_{\text{cor}} + N_{\text{inc}} + N_{\text{na}}$. We compute

$$\text{Correct} = N_{\text{cor}}/N, \qquad \text{Incorrect} = N_{\text{inc}}/N, \tag{20}$$

$$\text{Coverage} = (N_{\text{cor}} + N_{\text{inc}})/N, \qquad \text{AttemptedAcc} = \frac{N_{\text{cor}}}{N_{\text{cor}} + N_{\text{inc}}}, \tag{21}$$

$$\text{HM}_{\text{cor,att}} = \frac{2\,\text{Correct}\,\text{AttemptedAcc}}{\text{Correct} + \text{AttemptedAcc}}. \tag{22}$$

Zero-denominator cases are assigned zero. This correctness–attempt harmonic mean is not standard classification or answer-token F1. We therefore retain the full Correct, Incorrect, and Not Attempted distribution when analyzing refusal behavior. The parser extracts the first valid A/B/C label. An unparsable judge response defaults to Not Attempted. If all judge requests fail, the sample is recorded as an error and repaired rather than counted as model abstention.

B.4 Aggregation, completeness, and comparability

Each checkpoint-level benchmark value is first averaged over its three inference repetitions. Means and standard deviations across available training runs or teacher configurations are then computed separately. FACTS Grounding and HHEM are converted from hallucination rates to the positive-oriented grounded and faithful rates, so all six columns in Table 1 are higher-is-better. The reported six-benchmark macro-average gives every column equal weight and is used only as a compact within-study summary: its components are heterogeneous statistics and it is not itself a standardized benchmark metric.

Before aggregation, we verify sample completeness and re-evaluate failed judge calls so that missing evaluations are not counted as incorrect answers or abstentions.

C Training Configuration

C.1 Reflection SFT setup

We train Qwen3-8B with full-parameter SFT using LLaMA-Factory. The effective batch size is 64 examples. Training runs for five epochs at a learning rate of 2.5×10^{-6}, with cosine decay and a warmup ratio of 0.1. We use the Qwen3 chat template with thinking enabled and a maximum sequence length of 32,768 tokens. The shared initialization for subsequent controlled training is the final checkpoint at step 3945.

C.2 Generative reward-model setup

Both GenRMs start from Qwen2.5-7B-Instruct and undergo full-parameter supervised fine-tuning in LLaMA-Factory. We train them to generate the labels in Figure 4 using the Qwen chat template and a maximum sequence length of 32,768 tokens. Each model trains for five epochs with an effective batch size of 64. The learning rate is 5×10^{-6} with cosine decay and a warmup ratio of 0.1. The closed- and open-domain runs use 200,182 and 211,160 examples, respectively, and finish at steps 15,640 and 16,500. These final checkpoints provide the fixed GenRM rewards during policy RL.

C.3 RL reward implementation details

Table 7 gives the scalar mappings used by the single-domain and Cascade GRPO stages and by the reported Combined RL runs. These values are reward-shaping levels, not calibrated probabilities of correctness. GRPO standardizes rewards within the response group sampled for each prompt, so their primary role is to induce relative preferences among responses to the same prompt.

Table 7: Implemented scalar rewards for policy RL. The Combined RL column applies to both domains in the reported runs. A format failure means that the rollout lacks the required completed reasoning tags. An unparsable label is a reward-model output outside its requested label set.

Outcome	Closed	Open	Combined
Format failure	0	0	0
Correct (Yes / A)	2	2	2
Incorrect (No / B)	1	0.5	0.5
Not Attempted (C)	–	0.5	0.5
Unparsable GenRM label	1	1	1

For Combined RL, the reward for a closed-domain NO is reduced from 1 to 0.5, matching the incorrect-outcome reward on open-domain prompts in the reported runs. The positive value assigned to an incorrect but well-formed rollout preserves the distinction between task correctness and completion of the required response format; it should not be read as treating an incorrect answer as correct. In the open-domain stage, we assign the same reward to Incorrect and Not Attempted to reduce the incentive for reward hacking through excessive abstention. Assigning a higher reward to Not Attempted could encourage the policy to avoid answering even answerable questions. We therefore measure their frequencies separately in evaluation. An unparsable GenRM label receives an intermediate fallback value because it represents uncertainty in the evaluator output rather than a verified judgment about the policy response; it does not increment any correctness statistic. Reward-model requests are retried on connection or timeout failures, and an exhausted request raises an error instead of being converted into this fallback category.

C.4 SINGLE-DOMAIN AND CASCADE GRPO SETUP

The Closed-only and Open-only GRPO stages use a learning rate of 10^{-6}, 128 prompts per training batch, 64 prompts per policy mini-batch, and eight rollouts per prompt at temperature 1. The maximum prompt and response lengths are 4096 and 8192 tokens. The actor uses a low-variance KL penalty with coefficient 0.001 and no entropy bonus. The clipping offsets are $\epsilon_l = 0.2$ and $\epsilon_u = 0.28$ for closed-domain training or $\epsilon_u = 0.24$ for open-domain training, giving probability-ratio intervals $[0.8, 1.28]$ and $[0.8, 1.24]$, respectively. Cascade stages continue training from the selected single-domain checkpoints with the corresponding domain configuration. The implemented policy and KL losses use token-mean aggregation over all valid response tokens in the batch. For negative advantages, the policy loss additionally uses dual clipping with coefficient 3.0. These are implementation choices beyond the standard GRPO expression in Equation 7.

C.5 TASK-SPECIFIC CONSOLIDATION SETUP

For the reported task-specific consolidation runs, the integration training split contains 1,733 QA and 1,744 summarization prompts before length filtering. Applying the Qwen3 chat template, including the generation prefix, and removing prompts longer than 4,096 tokens leaves 1,471 QA and 1,692 summarization prompts (3,163 total). The validation split initially contains 186 prompts per task; the same filter retains 157 QA and 179 summarization prompts (336 total). The training sampler draws exactly 64 prompts per task in each 128-prompt batch, reshuffling and cycling a task's pool when exhausted. It emits $\lfloor 3163/128 \rfloor = 24$ full batches per epoch. The actor is initialized from the selected Cascade checkpoint and uses a learning rate of 10^{-6}, a training and policy mini-batch size of 128, one policy epoch per batch, rollout temperature 1, and maximum prompt and response lengths of 4096 and 8192 tokens. The distillation support contains the student's top 16 candidates at every response position. The actor and teachers use Qwen3-8B. The routed loss uses sequence-mean–token-mean aggregation, one rollout per prompt, and no rollout correction. Candidate probability ratios are clipped to $[0.8, 1.2]$, with a dual-clip coefficient of 3.0 for negative advantages. Optimization uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, gradient clipping at 1.0, and a constant learning-rate schedule without warmup. Student updates use no critic, external reward model, entropy bonus, or reference-policy KL regularization.

Training runs for 15 epochs ($24 \times 15 = 360$ updates) and saves every 25 updates. MOPD checkpoint selection follows the same validation-reward procedure as the RL stages: validation reward is

evaluated at steps 100, 200, 300, and 360, and the checkpoint is selected using that reward. Validation scoring is separate from the student update. The selected checkpoint is evaluated on the six test sets after training. The summary-only ablation uses the same Cascade initialization and Closed-only teacher, with 1,744 summarization prompts (1,692 after length filtering). Its $\lfloor 1692/128 \rfloor = 13$ batches per epoch give 195 updates over 15 epochs, and its reported checkpoints are steps 195, 100, and 100 for runs 1, 2, and 3, respectively. Standard and two-domain MOPD variants use token-mean loss aggregation; the routed consolidation uses sequence-mean–token-mean aggregation. All reported MOPD variants use the same top-16 token support, batch size, optimizer, rollout configuration, and maximum sequence length of 12,288 tokens.

D External Baseline Selection and Results

D.1 Selection and implementation status

Our primary comparison uses controlled single- and dual-domain variants built from the same Qwen3-8B backbone and shared SFT checkpoint. Table 8 lists the external methods considered, their implementation status, and whether comparable Qwen3-8B results are available. Combined RL, both Cascade orders, and standard MOPD are implemented in our framework as dual-domain controls.

Table 8: External baselines considered for comparison. Coverage describes independent training runs or available checkpoints, not repeated inference evaluations. A dash means that no comparable six-benchmark result is reported.

Method	Regime	Implementation status in this study	Coverage
FaithRL (Gui et al.)	Closed	Authors' implementation adapted to Qwen3-8B/MuSiQue.	3 seeds
FaithRL (Nie et al.)	Closed	Authors' implementation adapted locally to Qwen3-8B.	3 seeds
FSPO	Closed	Released code targets an older VERL/vLLM stack; no validated comparable Qwen3-8B run.	–
FARCA	Closed	No public implementation was available for comparable evaluation.	–
TruthRL	Open	TruthRL-style reward adapted in our VERL framework; not an official end-to-end reproduction.	3 seeds
Binary RAR	Open	Authors' public Qwen3-8B checkpoint; no local training.	1 checkpoint
KnowRL	Open	Local replication produced near-universal abstention; no comparable result.	–
SPACE	Dual	Public GitHub code is incomplete for reliable Qwen3-8B reproduction.	–
Wang et al. (2025b)	Dual	Single-policy GRPO across short-form and reference/no-reference long-form QA; no released implementation available.	–

SPACE (Wang et al., 2025a) has public activation and editing components, but its released repository does not provide a complete path to reproduce this comparison on Qwen3-8B. No released code repository was available for the joint GRPO method of Wang et al. (2025b). The TruthRL row instead uses a controlled reward adaptation: the authors' training pipeline uses CRAG-related data, so directly importing its results into our CRAG comparison would require a separate sample-level overlap audit. The external specialist references share our local evaluation protocol and Qwen3-8B backbone, but not necessarily our training data, initialization, or optimization budget.

D.2 External specialist results and diagnostic protocol

Table 9 reports all selected external-specialist results under our common six-benchmark evaluation. FACTS Grounding and HHEM are oriented so that higher scores indicate better grounding and

less hallucination. Both FaithRL adaptations and the TruthRL-style reward average selected checkpoints from three training seeds; Binary RAR is one public checkpoint evaluated repeatedly, not an independent-seed average. KnowRL and FSPO are excluded from the quantitative comparison because comparable six-benchmark results are unavailable. The controlled Combined RL, Cascade, and MOPD methods are our own implementations; their run-level results remain in Appendix E (Table 16).

Table 9: Selected external-specialist results in our local evaluation. FG denotes FACTS Grounding, C-SQA denotes Chinese SimpleQA, and Macro is the unweighted average of the six benchmarks. All values are percentages.

Method	Seeds	FG	HHEM	MultiHop	CRAG	WebQA	C-SQA	Macro
FaithRL (Gui et al.)	3	42.69	93.66	74.62	36.45	69.46	37.94	59.13
FaithRL (Nie et al.)	3	44.59	94.93	76.88	38.11	70.39	38.48	60.56
TruthRL-style	3	44.38	93.96	75.97	39.49	71.97	41.15	61.15
Binary RAR	–	29.96	92.58	61.45	36.43	68.12	36.35	54.15

For MultiHopRAG behavior diagnostics, we count a refusal only when a response equals `Insufficient Information` after trimming whitespace, lowercasing, and removing a final period. We separate unanswerable (*null*) from answerable (*non-null*) queries: correct refusal on null queries should be high, whereas refusal on non-null queries should be low. This refusal diagnostic uses 2,548 of the 2,556 MultiHopRAG examples per inference repetition, or 7,644 over three repetitions: 903 null and 6,741 non-null instances. Overall refusal alone can hide opposite changes in these groups.

D.3 FAITHRL (GUI ET AL.) ADAPTATION RESULTS

We adapted the FaithRL system of Gui et al. (2026) to Qwen3-8B and MuSiQue. Three independent training seeds are evaluated: seed 1 at step 270, seed 2 at step 240, and seed 3 at step 170. Each checkpoint is evaluated three times on all six benchmarks. Their method-level mean is .4269 FACTS Grounding, .9366 HHEM, .7462 MultiHopRAG, .3645 CRAG, .6946 WebQA, and .3794 Chinese SimpleQA, yielding a six-benchmark macro-average of $.5913 \pm .0036$.

Table 10: FaithRL (Gui et al.) per-inference-run results. FG and HHEM are reported as no-hallucination and faithful rates; MultiHop is accuracy; the remaining QA columns use the local correctness–attempt harmonic mean. All values are percentages and higher is better.

Seed/step	Run	FG	HHEM	MultiHop	CRAG	WebQA	C-SQA
1/270	1	46.51	93.98	74.37	35.98	69.01	37.79
	2	44.42	93.74	73.78	35.58	68.74	36.69
	3	41.74	93.98	74.14	34.91	69.62	38.15
2/240	1	41.98	93.50	75.08	36.98	69.68	38.28
	2	45.81	93.50	74.33	36.46	69.41	37.50
	3	46.05	94.71	74.84	35.48	70.95	36.98
3/170	1	37.56	93.62	75.35	36.42	69.95	39.52
	2	39.77	93.02	74.84	38.55	69.35	37.11
	3	40.35	92.90	74.84	37.66	68.40	39.40

In the selected checkpoints, the inspected MuSiQue responses place numbered reasoning outside the required reasoning span. Because the step verifier extracts steps from that span, this formatting limits the intended step-level supervision in our adaptation. The reported scores characterize the adapted Qwen3-8B policies; they do not imply the same behavior in the original system. The selected seed-level macro scores are 59.06%, 59.53%, and 58.81% for seeds 1, 2, and 3.

D.4 FAITHRL (NIE ET AL.) TRAINING TRAJECTORY

Our Qwen3-8B adaptation of the FaithRL system of Nie et al. (2026) was trained with three seeds. Here we examine the training trajectories of seeds 1 and 3. The main-table score averages selected checkpoints, not the last training step of each run. This matters because later checkpoints can show reduced abstention on unanswerable queries. Table 11 shows seed 1: FACTS Grounding rises from 46.43% to 61.86% between steps 200 and 500, while MultiHopRAG accuracy falls from 67.28% to 57.40%. Over the same interval, non-null refusal falls from 2.77% to 0.31%, while correct null refusal falls from 38.43% to 19.27%. Thus, the policy becomes less selective rather than simply more willing to answer correctly.

Table 11: FaithRL (Nie et al.) seed 1 trajectory. FG and HHEM are positive-oriented percentages. Refusal rates are split by MultiHopRAG answerability.

Step	FG	HHEM	MultiHop	Overall ref.	Non-null ref.	Null ref.
200	46.43	93.30	67.28	6.99	2.77	38.43
300	50.97	94.50	67.26	3.78	1.82	18.38
400	59.65	96.35	57.14	1.86	0.33	13.29
500	61.86	96.59	57.40	2.55	0.31	19.27

Seed 3 shows a similar reduction in refusal on unanswerable questions (Table 12), with no refusals observed among 903 null instances at step 300. We call this under-abstention or answer-all behavior observed in our adaptation, not an inevitable failure of the original FaithRL method.

Table 12: FaithRL (Nie et al.) seed 3 MultiHopRAG refusal diagnostics.

Seed/step	Scored	Overall ref.	Non-null ref.	Null ref.
3/100	7,644	14.72	8.01	64.78
3/200	7,644	11.94	4.04	70.99
3/300	7,644	0.18	0.21	0.00

D.5 OPEN-DOMAIN SPECIALIST OBSERVATIONS

Our KnowRL (Ren et al., 2026) replication showed near-universal abstention and did not yield a comparable six-benchmark result. We therefore exclude it from the quantitative comparison.

The public Binary RAR checkpoint refused 2,672 of 7,644 MultiHopRAG responses (34.96%). It refused 828 of 903 null instances (91.69%) but also 1,844 of 6,741 non-null instances (27.35%), indicating harmful over-abstention on answerable questions. Its FACTS Grounding score is 29.96%, so conservative answer selection did not translate into strong long-context grounding. The public checkpoint was not retrained locally, and its three inference repetitions are not three training seeds.

Our TruthRL-style reward adaptation uses three seeds, whose selected checkpoints are steps 300, 300, and 375, respectively. Their macro scores are 61.34%, 60.95%, and 61.17%, averaging 61.15%. This adaptation retains broader answer coverage than our KnowRL replication. For seed 1, non-null refusal decreased while correct null refusal increased over training (Table 13), but this trend was not uniformly monotonic across seeds: seeds 2 and 3 had non-null refusal ranges of 13.45–14.54% and 11.53–13.56%, respectively, and null refusal ranges of 82.06–87.60% and 80.62–84.72%. We report a controlled reward adaptation, not an official end-to-end TruthRL reproduction.

Table 13: MultiHopRAG refusal rates for seed 1 of our TruthRL-style reward adaptation. Null refusal is correct abstention; non-null refusal is an undesired refusal of an answerable question.

Step	Overall	Non-null	Null
100	20.55	12.42	81.28
200	19.65	11.20	82.72
300	19.52	10.92	83.72
375	19.14	10.32	84.94

D.6 Cross-baseline interpretation

The evaluated adaptations show different response patterns. In our FaithRL (Gui et al.) adaptation, reasoning outside the required span limits step-level verification. The FaithRL (Nie et al.) adaptation shows reduced refusal on unanswerable questions at later checkpoints. Our KnowRL replication exhibits near-universal refusal, while Binary RAR also refuses a substantial fraction of answerable questions. The TruthRL-style adaptation retains broader answer coverage, with refusal trends varying across seeds. These observations apply to the evaluated checkpoints and local adaptations. They motivate evaluating answer correctness, coverage, and source grounding together.

E In-House Training Results and Diagnostics

E.1 SFT checkpoint results

The final SFT checkpoint at step 3945 is both the SFT baseline in Table 1 and the shared initialization for controlled RL and SFT-initialized MOPD. It achieves Close and Open averages of 69.08% and 47.67%, respectively, and a Macro score of 58.37%.

E.2 Behavioral analysis of SFT

The analyses in this subsection use earlier SFT checkpoints (steps 3000 and 3600) to study response behavior; subsequent controlled training starts from step 3945. Figure 6 shows that SFT increases verification, backtracking, and subgoal actions across the evaluated datasets. The SFT baseline in Table 1 shows the largest benchmark improvement on FACTS Grounding, while open-domain score changes are smaller. Together, these results characterize the reflection behavior learned during SFT.

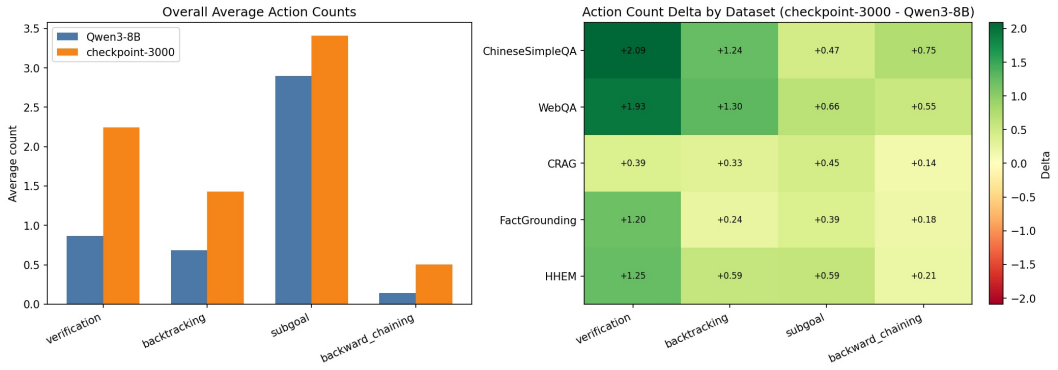


Figure 6: Changes in reflection-action frequencies from base Qwen3-8B to SFT step 3000. The left panel shows overall action counts, and the right panel shows changes by dataset.

Figure 7 compares hidden-thinking lengths on 500 matched examples: SFT step 3000 is longer than the base model on 463 of them. In the same paired subset, final responses become shorter on CRAG and WebQA but are approximately unchanged on FACTS Grounding. Longer hidden reasoning therefore can accompany different final-answer lengths across benchmarks.

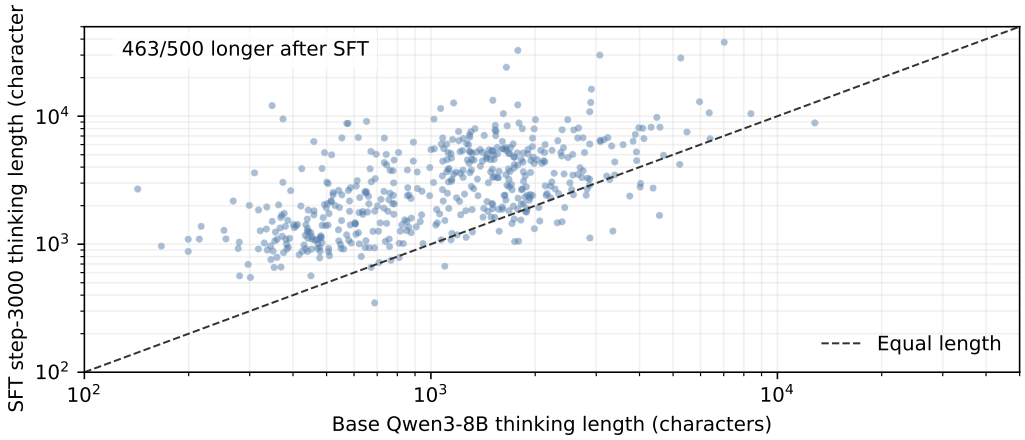


Figure 7: Paired hidden-thinking lengths for base Qwen3-8B and SFT step 3000 on 500 matched examples (three-inference-run mean per example). The dashed line indicates equal length. Both axes use logarithmic scales.

E.3 Single-domain checkpoint results

Table 14 gives the selected checkpoints for the three Closed-only and three Open-only training runs underlying Table 1. Each benchmark value averages three inference repetitions. Checkpoints are selected using validation reward measured during training. The six-benchmark Macro scores are reported for comparison, not used for checkpoint selection.

Table 14: Selected single-domain runs. FG denotes FACTS Grounding, C-SQA denotes Chinese SimpleQA, and Macro equally weights all six benchmarks. Higher is better.

Method	Run/step	FG	HHEM	MultiHop	CRAG	WebQA	C-SQA	Macro
Closed-only RL	1/500	.7442	.9848	.7558	.3477	.6897	.3954	.6529
	2/400	.7730	.9840	.7674	.3800	.7010	.3901	.6659
	3/300	.7373	.9803	.7295	.3690	.6950	.3913	.6504
Open-only RL	1/300	.4350	.9433	.7529	.3900	.7160	.3965	.6056
	2/375	.3867	.9350	.7395	.3900	.7210	.3964	.5948
	3/300	.4260	.9386	.7606	.3814	.7134	.4141	.6057

Closed-only RL consistently raises FG above the SFT checkpoint, but its selected runs retain weaker open-domain scores. Open-only RL instead improves short-answer QA while remaining far below Closed-only RL on FG. The selected Closed-only Macro scores range from 65.04 to 66.59, and the Open-only scores from 59.48 to 60.57; these ranges reflect training-run variation, not uncertainty from repeated inference on one checkpoint.

E.4 Single-domain RL behavior analysis

The following comparisons use representative evaluated checkpoints and their responses to characterize changes in answering behavior.

Closed-domain RL. For behavior analysis, we use Closed-RL-400, the step-400 checkpoint from a separate closed-domain RL training run initialized from SFT step 3945. This diagnostic run is separate from the three Closed-only runs in Table 14 and is excluded from the main comparison's method-level averages. Closed-RL-400 scores 75.17% on FACTS Grounding and 98.23% on HHEM, compared with 43.17% and 93.97%, respectively, for the earlier SFT step 3600 diagnostic checkpoint. MultiHopRAG, CRAG, and Chinese SimpleQA also improve, while WebQA

changes only slightly. The response-level diagnostics show shorter final answers on FACTS Grounding (1873.6 to 1325.7 characters) and longer hidden traces (5129.1 to 6448.6 characters). This shift is consistent with moving evidence checking into the hidden trace and suppressing unsupported surface text. Empty responses decrease on the diagnostic samples, but MultiHopRAG's "Insufficient Information" outputs increase from 1340 to 1913, indicating a stronger conservative boundary.

Table 15 compares the earlier SFT step 3600 with this diagnostic Closed-RL-400 checkpoint.

Table 15: Behavioral diagnostics for SFT step 3600 and Closed-RL-400, from a separate diagnostic training run initialized from SFT step 3945 and excluded from the main comparison's averages. Response and thinking lengths are in characters; empty counts are numbers of examples in the corresponding evaluation sample.

Dataset	Diagnostic	SFT-3600	Closed-RL-400
Chinese SimpleQA	response length	6.1	5.8
	empty responses	2	0
CRAG	response length	41.6	29.8
	empty responses	1	0
WebQA	response length	3.8	3.6
FACTS Grounding	response length	1873.6	1325.7
	thinking length	5129.1	6448.6
	empty responses	5	2
HHEM	response length	618.8	625.6
	thinking length	5486.8	4374.9
MultiHopRAG	response length	10.2	15.8
	empty responses	104	0

Comparing SFT step 3600 with Closed-RL-400, the marker analysis shows changes following closed-domain RL across both closed- and open-domain benchmarks. The fraction of CRAG traces mentioning a concise or direct answer rises from 60.2% to 97.2%, and the corresponding WebQA fraction rises from 93.0% to 99.4%. On FACTS Grounding, evidence or context verification rises from 84.8% to 92.2%, while concise/direct-answer markers rise from 64.7% to 77.0%. MultiHopRAG abstention phrases increase from 466 to 635 occurrences, so the elimination of empty responses can coincide with more explicit conservative answers.

Open-only RL. The selected Open-only run-1 checkpoint reaches 0.4350 FACTS Grounding, 0.9433 HHEM, 0.7529 MultiHopRAG, 0.390 CRAG, 0.716 WebQA, and 0.3965 Chinese SimpleQA. Relative to SFT, its open-domain gains are modest, while the large closed-domain improvement obtained by Closed-RL is absent. The response diagnostics associate this profile with a more answer-seeking distribution and less evidence-sensitive checking: direct answers help short factual QA, yet extra details can lower strict grounding scores. Across the examined Open-only checkpoints, later training also changes canonical answer selection and can reduce stability.

E.5 Dual-domain checkpoint results

Table 16 gives the selected run-level checkpoints underlying the controlled dual-domain rows of Table 1 and the ablations in Table 2. Each entry averages three inference repetitions. And we ran all methods using 3 different seeds or teacher configs. Dual-domain checkpoints are selected using validation reward measured during training. The full and SFT-initialized QA-plus-summarization variants use the shared checkpoint grid of steps 100, 200, 300, and 360.

Table 16: Selected dual-domain runs. FG denotes FACTS Grounding, C-SQA denotes Chinese SimpleQA, and Macro is the unweighted six-benchmark average. Higher is better.

Method	Run/checkpoint	FG	HHEM	MultiHop	CRAG	WebQA	C-SQA	Macro
Combined	run 1/750	.6587	.9820	.7606	.3950	.7260	.3957	.6530
	run 2/915	.7150	.9813	.7561	.3900	.7420	.4100	.6657
	run 3/250	.7167	.9740	.7194	.3830	.7220	.3916	.6511
Standard MOPD	run 1/100	.7667	.9844	.7763	.3993	.7285	.4038	.6765
	run 2/200	.7543	.9787	.7242	.3940	.7351	.4074	.6656
	run 3/200	.7473	.9803	.7595	.3665	.7232	.3941	.6618
Naive Cascade MOPD	run 1/600	.7655	.9803	.7742	.4160	.7360	.4052	.6795
	run 2/200	.7733	.9791	.7340	.3931	.7286	.3945	.6671
	run 3/100	.7453	.9751	.7732	.3846	.7296	.4064	.6690
Cascade C→O	run 1/300	.6883	.9827	.8021	.4100	.7230	.4030	.6682
	run 2/375	.7133	.9723	.7882	.3910	.7210	.4003	.6643
	run 3/375	.7074	.9795	.7928	.3923	.7462	.4098	.6713
Reverse O→C	run 1/300	.7487	.9807	.7597	.3830	.7140	.3981	.6640
	run 2/525	.7430	.9880	.7565	.4070	.7230	.3969	.6691
	run 3/300	.7017	.9873	.7776	.3770	.7380	.3911	.6621
OMNI-TRUTH (SFT init)	run 1/100	.7438	.9787	.8204	.3982	.7043	.3844	.6716
	run 2/200	.7260	.9795	.7875	.3915	.7145	.3917	.6651
	run 3/200	.7167	.9767	.8073	.3911	.7164	.3893	.6662
OMNI-TRUTH (summary-only)	run 1/195	.7717	.9884	.7830	.3966	.7214	.3903	.6752
	run 2/100	.7550	.9787	.7486	.3883	.7322	.3960	.6665
	run 3/100	.7271	.9775	.7789	.3677	.7207	.4118	.6640
OMNI-TRUTH	run 1/360	.7457	.9824	.8077	.4005	.7313	.3953	.6771
	run 2/100	.7109	.9779	.7960	.4011	.7333	.4070	.6710
	run 3/100	.7372	.9783	.8010	.4093	.7355	.4070	.6781

Across the selected runs, the Cascade's MultiHopRAG scores range from 78.82 to 80.21 while standard MOPD spans 72.42 to 77.63. MOPD instead has stronger FG in each reported teacher configuration. OMNI-TRUTH retains the Cascade's higher MultiHopRAG regime (79.60–80.77) while recovering FG above the Cascade mean across the selected runs.

E.6 TRAINING DYNAMICS

Figure 9 shows the training signals for the shared SFT, Closed-only RL, Closed-to-Open Cascade RL, and task-specific MOPD stages. SFT uses its training loss; both RL stages use their respective training rewards; MOPD uses the student policy loss from its on-policy distillation objective. Each panel has its own step and value scale. The three RL and MOPD curves use teacher-pair IDs 320, 605, and 42, corresponding to Cascade and OMNI-TRUTH runs 1, 2, and 3 in Table 16.

Figure 10 reports three additional logged training quantities for the same RL and MOPD configurations: actor entropy, mean response length in tokens, and actor gradient norm. Across the three MOPD runs, the student maintains nonzero entropy and comparable early- and late-stage response lengths while the distillation policy loss decreases.

E.7 DUAL-DOMAIN BEHAVIOR ANALYSIS

The following rollout diagnostics compare representative checkpoints across integration strategies. They describe answer coverage, response length, and grounding behavior rather than attributing each change to a single component.

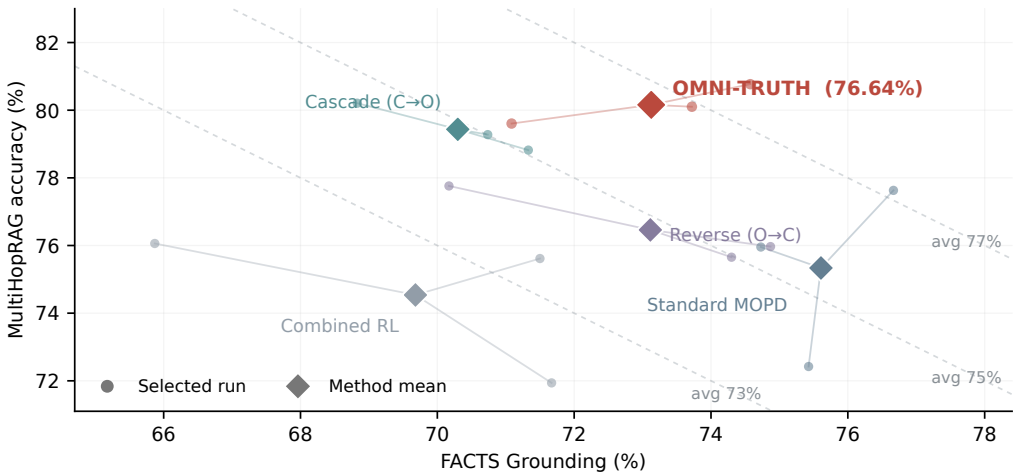


Figure 8: FACTS Grounding (FG) and MultiHopRAG results for the controlled dual-domain methods in Table 1. Small markers show the three selected run results from Table 16; larger markers show method means. Dashed diagonal lines indicate equal two-benchmark averages $(\mathrm{FG}+\mathrm{MultiHopRAG})/2$. OMNI-TRUTH has the highest two-benchmark mean among these methods (76.64%). Standard MOPD points are teacher configurations sharing one student-training seed; the other plotted methods have three runs as described above.

Sequential Closed-to-Open RL. The reported Cascade run-1 checkpoint is initialized from the reported Closed-only run-2 step-400 checkpoint. MultiHopRAG rises from 0.767 to 0.802, CRAG from 0.380 to 0.410, and WebQA from 0.701 to 0.723, whereas FACTS Grounding falls from 77.30% to 68.83%. Across the three stored evaluation repetitions, MultiHopRAG explicit-refusal counts fall from 1830 to 1429 and empty responses from 7 to 4. CRAG mean response length grows from 25.4 to 36.4 characters. Thus, the open stage improves answer coverage and cross-domain QA while weakening strict grounding in this aligned run.

Combined RL. Combined RL run 2 obtains 0.7150/0.9813 on FACTS Grounding/HHEM, 0.7561 on MultiHopRAG, 0.390 on CRAG, 0.742 on WebQA, and 0.410 on Chinese SimpleQA. Its diagnostic response lengths are 38.0 characters on CRAG, 5.7 on WebQA, and 7.8 on Chinese SimpleQA, with only one empty WebQA response. The open-domain Not Attempted rate is 0.67% on Chinese SimpleQA, 0.67% on CRAG, and 0.13% on WebQA for this checkpoint. These statistics show that joint training retains short-answer formatting with few Not Attempted responses on the open-domain tests.

Reverse sequential training. The reverse variant first trains on open-domain data and then applies closed-domain RL. The selected reverse checkpoints have a higher mean FG than the Cascade (73.11 versus 70.30), but a lower mean MultiHopRAG score (76.46 versus 79.43). The selected run-1 reverse checkpoint also has more explicit MultiHopRAG refusals than the selected run-1 Cascade checkpoint (1821 versus 1429). This pattern is consistent with a more conservative answer boundary after the final closed-domain stage.

MOPD. Across the three selected teacher configurations, MOPD obtains macro scores of 0.6765, 0.6656, and 0.6618. The student generally approaches the selected closed and open teachers, with some metric-level recovery rates above 100% (Table 19). The student's strong FACTS Grounding result is consistent with successful recovery of the closed-domain teacher's strict evidence-following capability. This is especially useful when closed-domain grounding and open-domain answer seeking favor different response boundaries. Conversely, MOPD trails sequential training on MultiHopRAG, CRAG, and Chinese SimpleQA, suggesting that direct capability restoration does not reproduce every gain that emerges from cross-domain sequential adaptation.

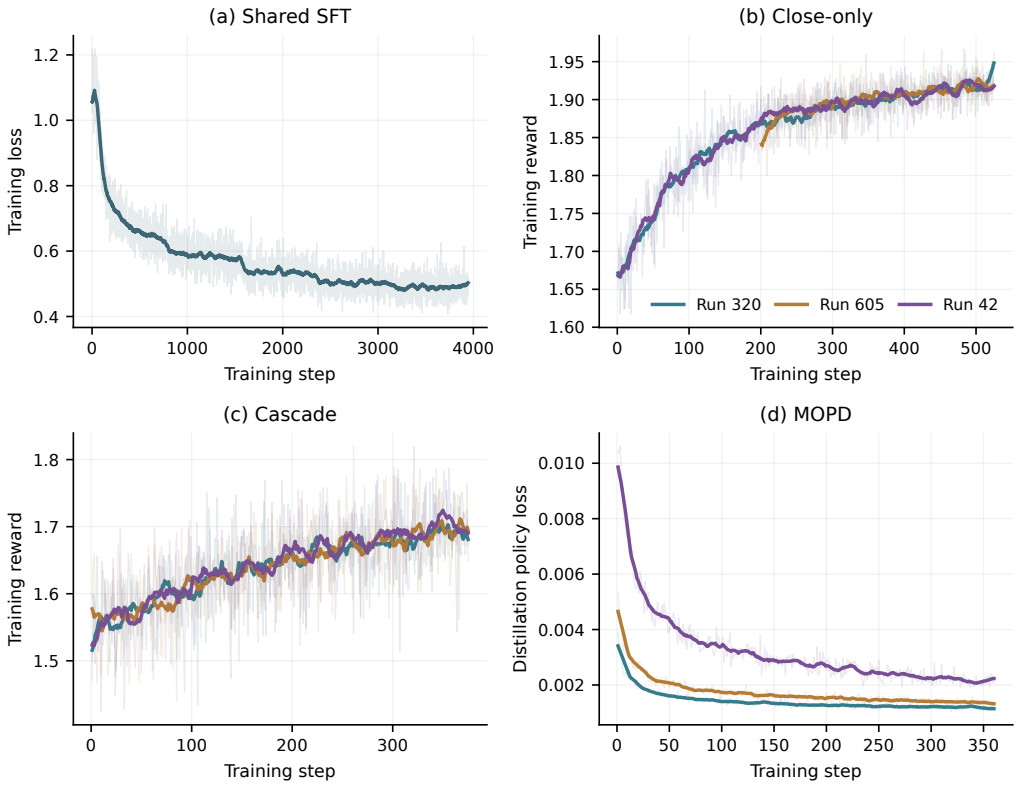


Figure 9: Training dynamics for SFT, closed-domain RL, open-domain RL, and task-specific MOPD. Faint curves are logged step values; solid curves are 15-step moving averages (51 steps for SFT). Closed-only and Cascade panels show training reward under their respective domain-specific reward functions. The MOPD panel shows distillation policy loss. The plotted Closed-only run-605 trace starts at step 201.

Task-specific MOPD. Across the three selected OMNI-TRUTH runs, MultiHopRAG remains in the Cascade's high-scoring regime (79.60–80.77), while FG improves from the Cascade mean of 70.30 to 73.13. The model thus retains the grounded-QA benefit of sequential training while improving strict grounding relative to the Cascade. OMNI-TRUTH (SFT init) uses the same task-specific teachers but averages 66.77 Macro and 49.79 Open avg., versus 67.54 and 51.34 with Cascade initialization. Cascade initialization therefore improves overall and open-domain performance in this comparison.

The selected run-level benchmark values are in Tables 14 and 16; the present table is intended to expose the response patterns associated with those scores. In particular, lower response length or lower Not Attempted rate should not be interpreted as universally better: short answers can affect local QA correctness judgments, whereas strict grounding metrics penalize unsupported details and multi-hop tasks can penalize excessive abstention.

Method-level rollout comparison. To make the distribution shift comparable across training strategies, we compute the same statistics from their evaluated responses. Table 18 uses a selected checkpoint for each listed method; standard MOPD reports the range over its selected runs. "Refusal" counts responses containing an explicit insufficient-information or cannot-answer phrase. These lexical counts are diagnostic indicators rather than evaluator labels. Each selected checkpoint has three evaluation repetitions for every benchmark.

For standard MOPD, configurations 1–3 pair Closed-only runs 2, 3, and 1 from Table 14 with Open-only teacher checkpoints 502/375, 605/375, and 107/300 (training identifier/step), respectively. These Open-only teacher identifiers are distinct from the table's run numbers;

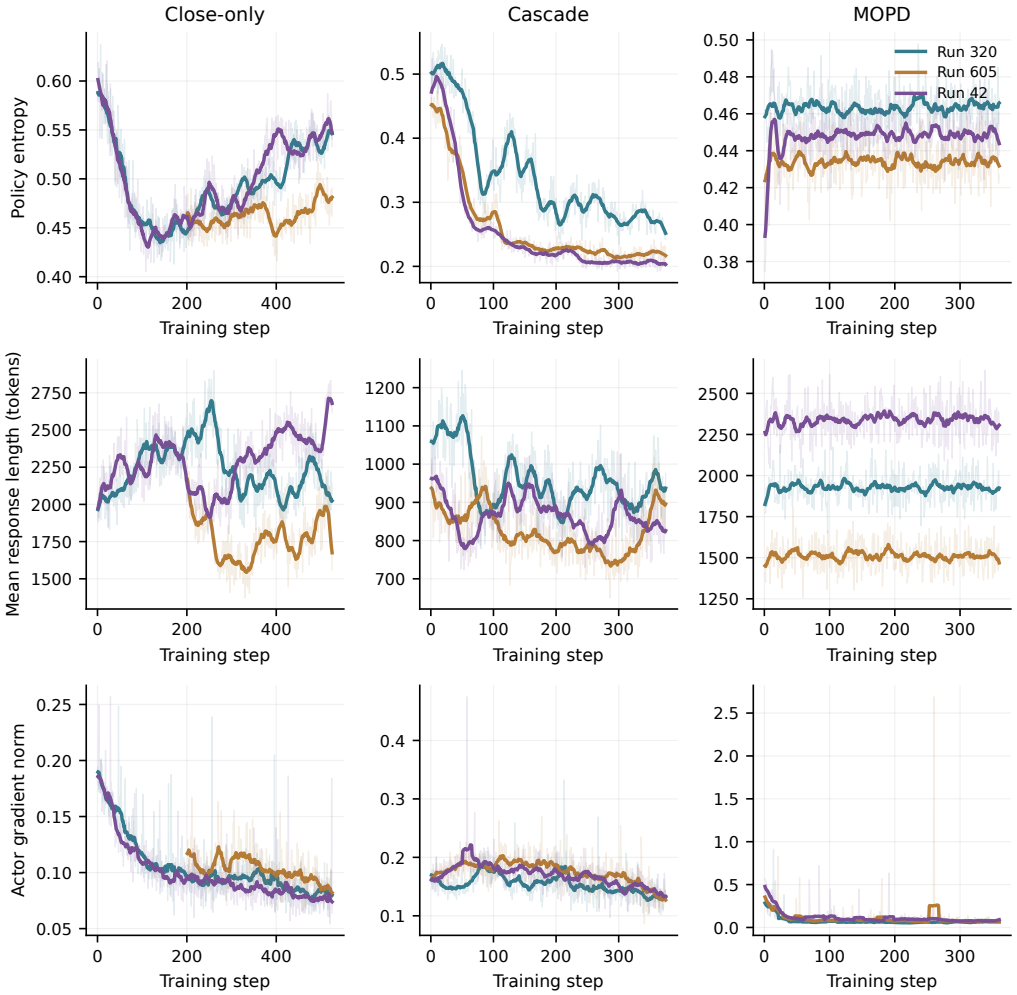


Figure 10: Training diagnostics across the Closed-only, Cascade, and MOPD stages. Each column uses its own training-step axis. Faint curves are logged step values; solid curves are 15-step moving averages. Response length is measured in output tokens. The plotted Closed-only run-605 trace starts at step 201.

Table 17: Response-length and coverage diagnostics for an earlier SFT diagnostic checkpoint, Closed-only run 2, and aligned Cascade and OMNI-TRUTH run-1 checkpoints, with Combined RL run 2 for context. Lengths are characters; counts pool three evaluation repetitions.

Dataset	Diagnostic	SFT	Closed-RL	Cascade	OMNI-TRUTH	Combined
C-SQA	response length	5.8	5.5	8.8	8.7	7.8
	empty responses	5	2	7	10	1
CRAG	response length (mean)	43.1	25.4	36.4	35.0	38.0
	response length (median)	15.0	14.0	15.0	14.0	16.5
WebQA	response length	3.8	3.5	6.5	6.3	5.7
	empty responses	2	0	5	5	1
FACTS Grounding	response length	1804.5	1077.2	1137.2	1131.2	1355.4
	thinking length	5752.8	8304.4	7584.3	7762.6	6630.4
	empty responses	36	6	0	0	19
HHEM	response length	613.6	532.5	534.7	526.2	652.7
MultiHopRAG	response length	10.6	12.2	12.2	10.6	16.6
	empty responses	489	7	4	5	66
	explicit refusals	1395	1830	1429	1404	1852

Table 18: Rollout distribution statistics for an earlier SFT diagnostic checkpoint and selected checkpoints from Tables 14 and 16. Lengths are response characters; counts pool the three evaluation repetitions. Standard MOPD entries are ranges over its three selected configurations.

Short-answer response format

Method	C-SQA length	C-SQA empty	WebQA length
SFT	5.8	5	3.8
Closed-RL	5.5	2	3.5
Open-only	8.8	6	6.1
Combined	7.8	1	5.7
Standard MOPD	6.6–8.7	3	5.6–6.2
Cascade C→O	8.8	7	6.5
Reverse	5.3	1	3.4
OMNI-TRUTH	8.7	10	6.3

CRAG and MultiHopRAG diagnostics

Method	CRAG mean	CRAG median	MH empty	MH refusal
SFT	43.1	15	489	1395
Closed-RL	25.4	14	7	1830
Open-only	36.9	14	37	1060
Combined	38.0	16.5	66	1852
Standard MOPD	29.5–39.0	14–16	4–21	1623–2176
Cascade C→O	36.4	15	4	1429
Reverse	25.6	14	8	1821
OMNI-TRUTH	35.0	14	5	1404

Table 19: Student-to-teacher score ratios at the selected MOPD checkpoints (%). Each entry divides the student score by a reference teacher's score and multiplies by 100. Standard and Naive Cascade MOPD use Closed-only as the reference for FG/HHEM/MultiHop, and Open-only and Cascade, respectively, for the three open-domain metrics. Both routed variants use Closed-only for FG/HHEM and Cascade for the four QA metrics. These are evaluation references; routed training uses only closed-domain prompts, so the open-domain ratios measure retention without open-domain distillation. Values above 100 indicate that the student exceeds its reference teacher. Run numbers identify aligned configurations within each variant, not additional training seeds.

Variant	Run	FG	HHEM	MultiHop	CRAG	WebQA	C-SQA
Standard MOPD	1	99.2	100.0	101.2	102.4	101.0	101.9
	2	102.3	99.8	99.3	101.0	101.4	104.9
	3	100.4	99.5	100.5	94.0	101.0	99.4
Naive Cascade MOPD	1	99.0	99.6	100.9	101.5	101.8	100.5
	2	104.9	99.9	100.6	100.5	101.1	98.6
	3	100.1	99.0	102.3	98.0	97.8	99.2
OMNI-TRUTH	1	96.5	99.8	100.7	97.7	101.1	98.1
	2	96.4	99.8	101.0	102.6	101.7	101.7
	3	99.1	99.3	101.0	104.3	98.6	99.3
OMNI-TRUTH (SFT init)	1	96.2	99.5	102.3	97.1	97.4	95.4
	2	98.5	99.9	99.9	100.1	99.1	97.9
	3	96.3	99.2	101.8	99.7	96.0	95.0

their CRAG/WebQA/C-SQA reference scores are 39.00/72.10/39.64, 39.00/72.50/38.82, and 39.00/71.60/39.65, respectively.

Cascade training lengthens CRAG answers relative to its Closed-only initialization but reduces MultiHopRAG refusals. Relative to that Cascade, OMNI-TRUTH slightly shortens CRAG answers (36.4 to 35.0 characters) and FACTS Grounding answers (1137.2 to 1131.2; Table 17) while preserving a similar MultiHopRAG refusal count. The response-length changes are small relative to the score changes. Combined has few empty short-answer responses, whereas Open-only has the fewest MultiHopRAG refusals among the selected methods. Reverse remains more conservative, and standard MOPD varies across its teacher configurations.

Benchmark-level comparison of integration strategies. Table 1 provides the overall comparison. Standard MOPD preserves more FACTS Grounding than the Cascade, whereas the Cascade transfers more strongly to MultiHopRAG. OMNI-TRUTH improves the Close, Open, and Macro averages through Cascade initialization and task-specific restoration. These patterns are descriptive comparisons of the selected checkpoints.

Behavioral patterns across training strategies. Table 20 relates the distribution statistics to response patterns observed in the selected model outputs.

Table 20: Representative response patterns and their interpretation.

Method / benchmark	Observed pattern	Interpretation
Closed-RL / MultiHopRAG	Run 2 has 1830 explicit refusals and only 7 empty responses	Evidence boundary is stronger, but answerability prior becomes conservative.
Cascade / CRAG	Run-1 mean response length rises from 25.4 to 36.4 characters relative to its Closed-only initialization	Open-domain updates encourage longer direct answers while improving CRAG score.
OMNI-TRUTH / FACTS Grounding	Run-1 response mean is 1131.2 characters, near the Cascade's 1137.2, while FG improves from .6883 to .7457	Grounding improves at a similar response length.
Combined / WebQA	Response mean 5.7 with one empty response across 1500 samples	Joint training preserves short-answer formatting.
Reverse / MultiHopRAG	Run-1 refusal count 1821 exceeds the selected Cascade run-1 count of 1429	Final closed-domain training can yield a more conservative answer boundary.
MOPD / cross-domain tests	Standard MOPD exceeds Cascade on FG but trails its MultiHopRAG mean by .0410	Teacher restoration preserves specialist capability but does not reproduce every cross-domain transfer gain.

The selected-run diagnostics distinguish answer coverage from final-answer length. Relative to its Closed-only initialization, the Cascade reduces explicit MultiHopRAG refusals from 1830 to 1429 while increasing mean CRAG response length from 25.4 to 36.4 characters. OMNI-TRUTH retains similar MultiHopRAG coverage (1404 refusals) and CRAG response length (35.0 characters), while improving aligned-run FG from 68.83% to 74.57%. Combined and reverse sequential training have higher MultiHopRAG refusal counts (1852 and 1821, respectively).