CREDO: EVIDENCE-CONDITIONED CREDIT ASSIGNMENT FOR TEMPORAL REASONING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal Knowledge Graph Question Answering (TKGQA) requires language model agents to retrieve evidence, select relevant facts, and perform multi-step temporal reasoning. However, the trajectory-level outcome signal in GRPO does not explicitly distinguish the quality of intermediate decisions: downstream failures can penalize useful evidence acquisition, while identical rewards in all-failed groups eliminate the relative outcome signal despite differences in intermediate progress. We propose Credit Redistribution with Evidence-Driven Opportunities (CREDO), which uses intermediate evidence to recover local supervision, including in all-failed groups, and determine where outcome supervision acts. CREDO assigns positive stage-level credit through comparisons among rollouts of the same question and uses retained answer overlap as an opportunity proxy to shift part of the outcome-policy weight toward reasoning and answering. The framework preserves terminal rewards and trajectory-level advantages, while redistribution conserves total outcome-policy weight per trajectory. Following the public evaluation protocol of prior work, CREDO raises Overall accuracy from 78.0% to 82.4% and Multiple accuracy from 55.0% to 65.6% on MultiTQ relative to the strongest published baseline in our comparison. On TimelineKGQA, it raises cross-KG accuracy from 30.9% to 36.4% relative to our reproduced baseline via direct transfer to ICEWS Actor under matched training and evaluation settings.

1 INTRODUCTION

Temporal Knowledge Graph Question Answering (TKGQA) requires reasoning over time-dependent facts under temporal constraints (Saxena et al., 2021; Chen et al., 2023). Recent advances in large language models have improved temporal reasoning through structured retrieval, evidence construction, and experience reuse (Hu et al., 2026; Liu et al., 2026; Tan et al., 2026). Agent-based approaches further integrate retrieval and reasoning into dynamic interaction trajectories, allowing models to adapt decisions to new evidence (Hu et al., 2025; Lv et al., 2026; Jiang et al., 2026; Huang et al., 2026). However, an agent may retrieve useful facts but fail to retain or reason over them correctly. The final answer therefore *only partially reflects intermediate decision quality*, motivating closer examination of how temporal agents learn from their interaction trajectories.

Reinforcement learning enables language models to learn multi-step reasoning and search strategies from outcome supervision (Shao et al., 2024; Guo et al., 2025; Jin et al., 2025; Song et al., 2025; Chen et al., 2025), with recent extensions to temporal agents (Gong et al., 2026; Huang et al., 2026). However, outcome-supervised GRPO assigns the *same terminal-reward-based advantage* to all trainable tokens within a trajectory, without explicitly distinguishing intermediate decision quality (Shao et al., 2024). Process supervision and fine-grained credit assignment provide more localized learning signals (Lightman et al., 2024; Wang et al., 2024; Xie et al., 2026; Zong et al., 2026; Zhang et al., 2026). For temporal agents, credit assignment must also account for decision dependencies: evidence selection relies on what retrieval exposes, while answer derivation relies on what selection retains. This raises a central question: how can intermediate evidence guide both what agents learn from unsuccessful trajectories and where terminal supervision should act?

Failed trajectories can still contain useful evidence. An agent may retrieve and retain relevant temporal facts yet answer incorrectly (Figure 1). In mixed-outcome groups, useful retrieval and selection decisions can receive negative outcome credit. When all rollouts fail with identical rewards,

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.

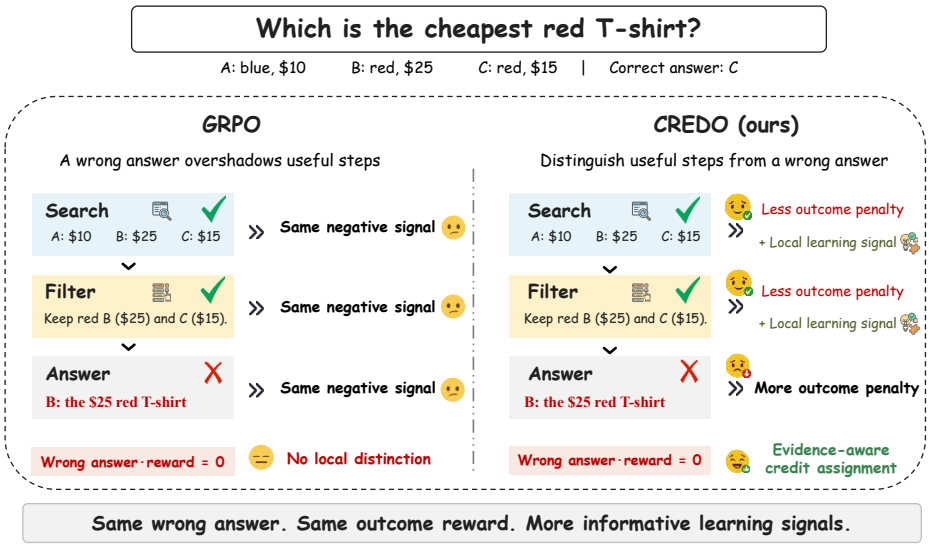

Figure 1: **Motivation of CREDO.** In a mixed-outcome group, useful intermediate decisions can receive negative outcome credit when the final answer is wrong. CREDO uses intermediate evidence for within-question comparisons and evidence-conditioned outcome redistribution.

outcome advantages vanish, leaving differences in intermediate evidence quality unexploited. These cases motivate distinguishing *local behavior quality* from *evidence opportunity*: how well a stage performs and how much answer-related evidence upstream decisions provide.

Credit redistribution with evidence-driven opportunities. Motivated by this analysis, we propose Credit Redistribution with Evidence-Driven Opportunities (CREDO), which uses intermediate evidence to guide both local supervision and outcome credit. CREDO compares retrieval coverage, evidence retention, and answer quality across rollouts of the same question, *rewarding locally superior behavior while conditioning downstream credit on upstream evidence*. These comparisons can recover local learning signals *even in all-failed groups*. For eligible trajectories with retained answer-related evidence, CREDO also shifts a controlled fraction of outcome-policy weight toward reasoning and answering tokens. This redistribution applies to both positive and negative advantages while *preserving total outcome-policy weight per trajectory*. Both components retain the original terminal reward and trajectory-level GRPO advantage without requiring an additional learned evaluator. Answer overlap with retrieved and retained facts serves as a proxy for evidence opportunity.

Contributions. Our main contributions are threefold. (1) We analyze credit assignment in temporal agents through dependencies among retrieval, evidence selection, and answering, distinguishing local behavior quality from upstream evidence opportunity when learning from failed trajectories. (2) We propose CREDO, which distinguishes local quality from upstream evidence opportunity to recover stage-level supervision through within-question comparisons, including in all-failed groups. (3) We develop evidence-conditioned outcome redistribution that shifts policy weight toward reasoning and answering for both positive and negative trajectory advantages while preserving total outcome-policy weight per trajectory, terminal rewards, and GRPO advantages.

2 Related Work

2.1 Temporal Knowledge Graph Question Answering

Early TKGQA approaches include embedding-based methods that score candidate answers (Saxena et al., 2021; Mavromatis et al., 2022; Chen et al., 2023; 2022), semantic parsing methods that generate executable queries (Jia et al., 2018; Neelam et al., 2021; Chen et al., 2024a; Ding et al., 2022), and graph-based methods that model structural dependencies (Jia et al., 2021; Sharma et al., 2023; Liu et al., 2023b;a). Recent LLM-based methods improve question interpretation through rewriting

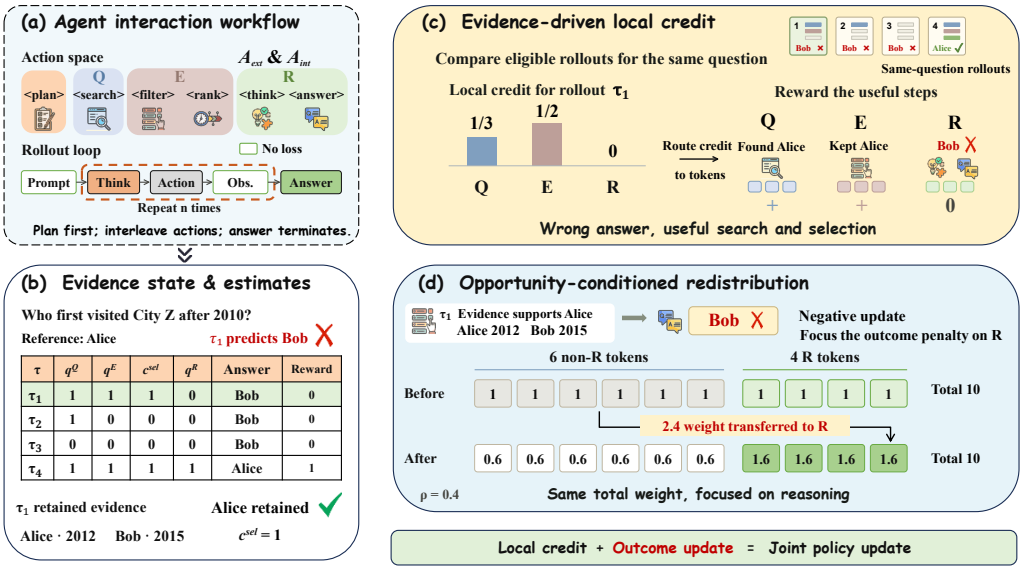

Figure 2: Framework of CREDO. CREDO evaluates retrieved and retained evidence to recover local credit through within-question comparisons and redistribute outcome-policy weight toward reasoning and answering. The illustration uses $\rho = 0.4$; the experimental default is $\rho = 0.04$.

in TimeR^4 (Qian et al., 2024) and decomposition in RTQA (Gong et al., 2025), organize reasoning through temporal tools in TempAgent (Hu et al., 2025) and planning in PoK (Qian et al., 2025), and reuse successful solution paths in MemoTime (Tan et al., 2026). Temp-R1 further explores end-to-end reinforcement learning for autonomous temporal reasoning (Gong et al., 2026). Building on this direction, CREDO investigates how intermediate evidence guides credit assignment during temporal-agent training, considering dependencies among retrieval, selection, and answering.

2.2 Search Agents and Reinforcement Learning

PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) optimize policies from outcome rewards without token-level process annotations. OpenAI-o1 and DeepSeek-R1 demonstrate effective reinforcement learning for reasoning (OpenAI et al., 2024; Guo et al., 2025), while Search-R1, R1-Searcher, and ReSearch extend it to agents interacting with external knowledge (Jin et al., 2025; Song et al., 2025; Chen et al., 2025). Recent methods supplement terminal feedback with finer-grained credit for intermediate decisions (Feng et al., 2025; Ma et al., 2026). TIPS derives turn-level rewards from changes in answer likelihood under a teacher model (Xie et al., 2026), while StepSearch uses intermediate search rewards and token-level supervision using information gain and redundancy penalties (Wang et al., 2025). CREDO instead uses answer-overlap proxies for within-question stage comparisons and evidence-conditioned outcome-weight redistribution, preserving terminal rewards and trajectory-level advantages without an additional learned evaluator.

3 Method

Credit Redistribution with Evidence-Driven Opportunities (CREDO) uses acquired and retained evidence to reward useful intermediate behavior through within-question comparisons and redistribute outcome supervision (Figure 2). Both components jointly optimize one agent while preserving terminal rewards and GRPO trajectory advantages. Algorithm 1 summarizes the update.

3.1 Problem Formulation

Given a temporal knowledge graph $\mathcal{K}$ and a training pair $(x, G) \in \mathcal{D}$, an agent π_θ interacts with $\mathcal{K}$ to answer x, where G contains acceptable answers. The behavior policy π_{old} samples K trajectories $\{\tau_i\}_{i=1}^{K}$ per question. Their original terminal rewards determine the GRPO advantages $\widehat{A}_i$.

Algorithm 1 CREDO training update

Input: Mini-batch $\mathcal{B}$, behavior policy π_{old}, rollout count K, and CREDO hyperparameters.

for each $(x, G) \in \mathcal{B}$ **do**
  Sample K trajectories with π_{old}; compute original rewards and GRPO advantages $\widehat{A}_i$.
  Extract evidence and stage masks; determine token availability t_i^s.
  Compute q_i^Q, c_i^{sel}, q_i^E, and q_i^R using Eqs. (1)–(2).
  Determine comparison validity v_i^s; set $\omega_i^E = q_i^Q$ and $\omega_i^R = c_i^{\text{sel}}$.
  for each stage $s \in \{Q, E, R\}$ **do**
    Compute peer comparisons F_i^s and local advantages a_i^s (Eqs. (4)–(5)).
    Compute the unnormalized stage budget; set it to zero if no positives exist (Appendix A.1).
  end for
  Normalize budgets across stages to obtain α_s; allocate u_{it}^s by Eq. (7).
  Identify eligible trajectories (Section 3.3); assign w_{it} by Eq. (6).
end for
Optimize $J_{\text{outcome}}^w + \lambda_{\text{local}} J_{\text{local}}$ using separately normalized clipped surrogates; omit zero-weight terms.

We distinguish retrieval Q (Search), selection E (Filter and Rank), and reasoning/answering R (post-retrieval Think and Answer), which may recur across rounds. Let $A_{it} \in \{0, 1\}$ indicate trainable model tokens and $M_{it}^s \leq A_{it}$ their membership in stage $s \in \{Q, E, R\}$. Stage masks are disjoint, exclude environment observations (Information), and locate credit; evidence states determine its assignment. Retrieval and selection masks can aggregate repeated actions; reasoning uses the final post-retrieval Think region and Answer contents. Appendix A.1 specifies localization.

3.2 Evidence-Driven Local Credit

Evaluating acquired and retained evidence. I_i contains all retrieved Information lines. Final selection S_i takes non-empty lines from the last complete Rank output, or the last complete Filter output if Rank is empty or absent. This answer-independent preference reflects Rank's organizational role. Line-level provenance checks yield ordered retained evidence $\widetilde{S}_i = (l_0, l_1, \ldots)$.

To estimate answer-related evidence, let $\text{cov}(l, g)$ denote the fraction of normalized answer tokens in g covered by fact line l, using multiset overlap. Retrieved and retained coverage are

$$q_i^Q = \max_{l \in I_i,\, g \in G} \text{cov}(l, g), \qquad c_i^{\text{sel}} = \max_{l_j \in \widetilde{S}_i,\, g \in G} \frac{\text{cov}(l_j, g)}{\log_2(j+2)}. \tag{1}$$

The positional discount gives greater credit to evidence placed earlier in the selection. Empty or unreliable evidence receives zero coverage. We then evaluate selection relative to its retrieved input and answering relative to the references:

$$q_i^E = \begin{cases} \min\{c_i^{\text{sel}}/q_i^Q, 1\}, & q_i^Q > \delta, \\ 0, & \text{otherwise}, \end{cases} \qquad q_i^R = \max_{g \in G} \text{F1}_{\text{token}}(\widehat{y}_i, g), \tag{2}$$

where $\delta > 0$ is a numerical threshold and $\widehat{y}_i$ is the locally extracted answer. Missing or unreliable answers receive zero quality. These coverage and overlap scores serve as local quality proxies; normalization, provenance, and extraction details are provided in Appendix A.1.

Accounting for evidence dependencies. We estimate evidence opportunity through upstream answer overlap:

$$\omega_i^E = q_i^Q, \qquad \omega_i^R = c_i^{\text{sel}}. \tag{3}$$

Here ω_i^E measures retrieved overlap and ω_i^R retained overlap. High retention quality q_i^E can coexist with low ω_i^R when little answer-related evidence was retrieved. Retrieval has no upstream dependency. These proxies do not verify relations, temporal constraints, or multi-fact completeness: unrelated facts may contain answer tokens, while evidence for computed answers may lack them. Provenance verifies the retrieved source rather than semantic sufficiency; controlled evidence perturbations illustrate this boundary (Appendix A.4).

Recovering useful local behavior. We compare stage quality among trajectories of the same question, allowing intermediate differences to remain informative when terminal rewards coincide. Let

$v_i^s \in \{0,1\}$ indicate comparison validity and $\mathcal{V}_s = \{i : v_i^s = 1\}$. For $i \in \mathcal{V}_s$ and $|\mathcal{V}_s| > 1$, define

$$F_i^s = \frac{1}{|\mathcal{V}_s| - 1} \sum_{j \in \mathcal{V}_s \setminus \{i\}} \operatorname{sgn}(q_i^s - q_j^s). \tag{4}$$

Ties contribute zero, so a positive score indicates more wins than losses against valid peers. We set $F_i^s = 0$ when comparison is unavailable. With $[z]_+ = \max(z, 0)$ and $t_i^s \in \{0,1\}$ indicating reliable stage-token localization, the auxiliary advantages are

$$a_i^Q = v_i^Q t_i^Q q_i^Q [F_i^Q]_+, \qquad a_i^s = v_i^s t_i^s \omega_i^s q_i^s [F_i^s]_+, \quad s \in \{E, R\}. \tag{5}$$

Locally superior behavior receives positive credit, scaled by upstream opportunity for dependent stages. Evaluable failures can be zero-quality peers without trainable tokens. Local quality differences can thus supply positive supervision even in all-failed groups.

Empirical illustration. Figure 3 shows recurring all-wrong groups during Stage 1. Of these, 38.29% retain active local credit despite zero terminal GRPO advantages. Activity requires positive local advantages, trainable tokens, and allocated weights. The trace demonstrates signal availability, without isolating its contribution to final accuracy. Counts and reward settings appear in Appendix A.3.

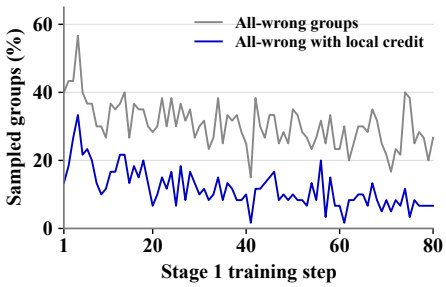

Figure 3: All-wrong groups and those with active local credit in CREDO Stage 1 (8B). Both denominators count all sampled groups.

3.3 Evidence-Conditioned Outcome Redistribution

Locating outcome supervision. The original negative GRPO advantage can still penalize useful evidence acquisition. When retained answer overlap indicates an answering opportunity, we therefore shift part of the outcome-policy weight toward reasoning and answering.

Let $n_i = \sum_t A_{it}$ and $m_i = \sum_t M_{it}^R$ be the numbers of trainable and reasoning/answering tokens. Redistribution requires $v_i^R = 1$, $\omega_i^R > 0$, $n_i, m_i > 0$, and $m_i/n_i \geq \eta$, where $\eta \in (0,1]$ prevents excessive concentration on a short token span. For eligible trajectories $\mathcal{E}$, we assign

$$w_{it} = \begin{cases} (1-\rho)A_{it} + \rho \dfrac{n_i}{m_i} M_{it}^R, & i \in \mathcal{E}, \\ A_{it}, & i \notin \mathcal{E}, \end{cases} \qquad \sum_t w_{it} = n_i. \tag{6}$$

Here $\rho \in [0,1)$ mixes the original token weights with a reasoning/answering-only allocation; the net weight moved into this region is $\rho(n_i - m_i)$. The factor n_i/m_i preserves total policy weight. Evidence opportunity determines eligibility rather than continuously scaling ρ. The conserved quantity is the sum of policy weights, not the gradient magnitude.

Redistribution applies to either sign of $\widehat{A}_i$. When $0 < \rho$ and $m_i < n_i$, it increases reasoning/answering weights and decreases others; if $m_i = n_i$, weights are unchanged. If $\widehat{A}_i = 0$, redistribution creates no outcome signal, but local credit can exploit intermediate quality differences.

3.4 Policy Optimization

Normalizing local supervision. To keep auxiliary learning from scaling with trajectory length or the number of positive examples, we assign each stage a normalized budget α_s within each question group (the group index is omitted here). This budget combines valid comparison support, upstream opportunity for dependent stages, within-group quality diversity, and remaining improvement; its complete definition appears in Appendix A.1. Let $\mathcal{H}_s = \{i : a_i^s > 0\}$, $h_s = |\mathcal{H}_s|$, and $n_i^s = \sum_t M_{it}^s$. For each positive example, define

$$u_{it}^s = \frac{\alpha_s}{h_s n_i^s} M_{it}^s, \qquad i \in \mathcal{H}_s, \qquad \sum_{i,t} u_{it}^s = \alpha_s. \tag{7}$$

Other token weights are zero; stages without valid positive examples receive zero budget. The allocation controls total stage weight, while advantage magnitudes still determine the weighted loss.

Table 1: Performance comparison on MultiTQ. CREDO follows the normalized substring-matching protocol in Temp-R1's public evaluation code. Baseline scores are adopted from Table 3 of Temp-R1 (Gong et al., 2026). The **best** and second-best scores under each metric are highlighted in colors, respectively; ties receive the same highlight.

Model	Overall	Question Type		Answer Type	
		Multiple	Single	Entity	Time
TKG Embedding-based methods					
EmbedKGQA	0.206	0.134	0.235	0.290	0.001
CronKGQA	0.279	0.134	0.337	0.328	0.156
MultiQA	0.293	0.159	0.347	0.349	0.157
Prompt-based LLMs					
ARI	0.380	0.210	0.680	0.394	0.344
TempAgent	0.702	0.316	0.857	0.624	0.870
MemoTime	0.730	0.459	0.829	0.677	0.846
RTQA	0.765	0.424	0.902	0.692	0.942
Fine-tuning-based LLMs					
Search-R1	0.352	0.094	0.474	0.230	0.705
TimeR4	0.728	0.335	0.887	0.639	0.945
PoK	0.779	0.409	**0.929**	0.696	0.962
Temp-R1	0.780	0.550	0.888	0.714	0.969
CREDO	**0.824**	**0.656**	0.903	**0.771**	**0.977**

Joint optimization. We combine the redistributed outcome signal with the recovered local credit:

$$J(\theta) = J^{w}_{\text{outcome}}(\theta) + \lambda_{\text{local}} J_{\text{local}}(\theta), \tag{8}$$

where J^{w}_{outcome} uses the original GRPO advantage $\widehat{A}_i$ with token weights w_{it}, and J_{local} uses stage advantages a^s_i with weights u^s_{it}. The outcome term uses token-mean aggregation within each static microbatch, with redistributed token weights in both the numerator and denominator. The local term uses PPO-minibatch weight normalization; Appendix A.1 specifies clipping, accumulation, and regularization. Zero-weight terms are omitted. The coefficient λ_{local} controls auxiliary supervision. Local credit determines which stage behavior to reinforce; redistribution determines where outcome supervision acts. Inference requires neither reference answers nor credit computations.

4 Experiments

4.1 Setup

Datasets. We train CREDO on two datasets: MultiTQ (Chen et al., 2023) and TimelineKGQA-CRON (Sun et al., 2025). We evaluate three settings: MultiTQ, CronQuestion KG, and ICEWS Actor. Following Temp-R1's evaluation layout (Gong et al., 2026), we report MultiTQ results by question and answer type, and evaluate both in-domain question answering and cross-KG transfer on TimelineKGQA. Tables 1 and 2 present these comparisons. Section 4.2 describes the tasks and evaluation protocol; dataset sizes, category distributions, and data-split audits appear in Appendix A.2.

Baselines. On MultiTQ, we compare CREDO with three baseline groups: (1) TKG embedding-based methods, including EmbedKGQA (Saxena et al., 2020), CronKGQA (Saxena et al., 2021), and MultiQA (Chen et al., 2023); (2) prompt-based LLM methods, including ARI (Chen et al., 2024b), TempAgent (Hu et al., 2025), MemoTime (Tan et al., 2026), and RTQA (Gong et al., 2025); and (3) fine-tuning-based LLM methods, including Search-R1 (Jin et al., 2025), TimeR4 (Qian et al., 2024), PoK (Qian et al., 2025), and Temp-R1 (Gong et al., 2026). TimelineKGQA baselines include RAG Baseline, LLaMA2-7B, GPT-4o, RTQA, and Temp-R1. For TimelineKGQA, we train Temp-R1 and CREDO under identical base configurations and evaluate them on identical inputs using Temp-R1's public scoring protocol (Appendix A.2), reducing confounding from protocol differences and making the comparison more directly reflect differences between methods. Other baseline scores come from Tables 3 and 2 of Temp-R1 for MultiTQ and TimelineKGQA, respectively.

Table 2: Performance comparison on TimelineKGQA across in-domain (CronQuestion KG) and out-of-domain (ICEWS Actor) settings. *Our reproduction, using the same base training configuration as CREDO and Temp-R1's public scoring protocol (Appendix A.2). Other baseline results are adopted from Temp-R1 (Gong et al., 2026). The **best** and second-best scores under each metric are highlighted in colors, respectively; ties receive the same highlight.

Model	CronQuestion KG (In Domain)				ICEWS Actor (Out of Domain)			
	Overall	Simple	Medium	Complex	Overall	Simple	Medium	Complex
RAG Baseline	0.235	0.704	0.092	0.009	0.265	**0.660**	0.128	0.011
LLaMA2-7B	0.169	0.049	0.143	0.282	0.111	0.035	0.066	0.322
GPT-4o	0.206	0.069	0.130	0.376	0.113	0.051	0.035	0.353
RTQA	0.298	0.608	0.218	0.135	–	–	–	–
Temp-R1*	0.508	**0.844**	0.241	0.446	0.309	0.447	0.123	0.324
CREDO	**0.523**	**0.844**	**0.278**	**0.456**	**0.364**	0.501	**0.170**	**0.393**

Implementation Details. We use Meta-Llama-3.1-8B-Instruct with E5-base-v2 retrieval and a FAISS Flat index, following Jin et al. (2025). Training follows Temp-R1's two-stage procedure (Gong et al., 2026). CREDO-family runs share initial assets and base training settings; selected models use a common evaluation configuration. Checkpoints are selected by monitoring accuracy (Appendix A.2). Defaults are $\lambda_{\text{local}} = 0.03$, $\rho = 0.04$, and $\eta = 0.05$. All 8B reinforcement learning experiments run on one machine with 8 Ascend 910B NPUs. Appendix A.2 gives the configurations and distinguishes these ablations from subsequent frozen-model diagnostics.

4.2 Benchmarks and Metrics

Temporal KGQA. Given a question and a temporal knowledge graph, the agent returns the requested entity, time, or other value under the question's temporal constraints. Evaluation concerns final answers without requiring a predefined tool-use trajectory, allowing different retrieval and reasoning paths across one or more interaction rounds to produce acceptable answers.

MultiTQ. MultiTQ (Chen et al., 2023) evaluates temporal reasoning at year, month, and day granularities. The official Single categories are Equal, Before/After, and First/Last; Multiple includes Equal Multi, Before Last, and After First. We use these official labels without treating them as measured constraint counts. Our fixed evaluation set contains 500 questions, comprising 340 Single and 160 Multiple questions. Table 1 reports results by question type and answer type.

TimelineKGQA. TimelineKGQA (Sun et al., 2025) categorizes questions as Simple, Medium, or Complex according to whether they require one, two, or three contextual facts. Tasks include temporally constrained retrieval, comparison, composition, and ranking; answers include entities, time points, intervals, durations, and temporal relations. Following Temp-R1, we evaluate on CronQuestion KG and directly transfer the same checkpoint to ICEWS Actor without further training, assessing cross-KG transfer. Table 2 reports Overall and complexity-level results.

Metrics. Following the evaluation protocol in the public Temp-R1 repository (Gong et al., 2026), we evaluate final answers using normalized substring matching. A prediction is correct if any acceptable answer occurs contiguously within the normalized final answer. Let $\hat{y}_i$ denote the predicted answer, G_i the set of acceptable answers, and $\nu(\cdot)$ the normalization function. Accuracy over N questions is defined as:

$$\text{Acc}_{\text{sub}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\exists\, g \in G_i : \text{substr}(\nu(g), \nu(\hat{y}_i))]\,, \tag{9}$$

where $\text{substr}(a, b)$ indicates contiguous occurrence. Answers are extracted from `<answer>` tags; normalization lowercases text, removes punctuation and English articles, and collapses whitespace. Missing answers and execution failures count as incorrect. This metric follows the baseline protocol and measures answer containment; it does not require the output to contain only a single unambiguous answer. Appendix A.1 provides extraction details.

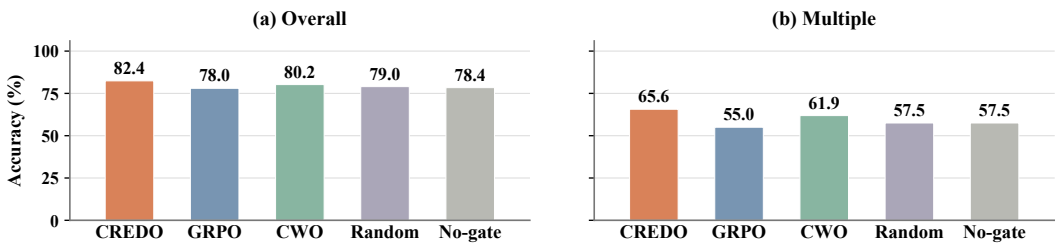

Figure 4: MultiTQ component comparisons: (a) Overall; (b) Multiple. GRPO uses the Temp-R1 reported results (Gong et al., 2026). CWO disables redistribution; Random changes its target tokens; No-gate adds ungated answer-region redistribution to CWO. See Appendix A.2 for selection details.

4.3 Experimental Results

Results on MultiTQ. CREDO achieves the best Overall and Multiple accuracy in Table 1, exceeding the strongest published baseline by 4.4 and 10.6 percentage points, respectively. The larger gain on Multiple questions highlights its effectiveness on compositional temporal tasks, where agents must coordinate evidence acquisition, selection, and answer derivation. This pattern aligns with the design of CREDO: local comparisons recover supervision from intermediate progress, while evidence-conditioned redistribution directs outcome-policy weight toward reasoning and answering. These mechanisms address the dependencies that make useful retrieval alone insufficient for solving complex questions. CREDO also ranks first on both entity and time answers, showing that its performance advantage extends across answer formats.

Results on TimelineKGQA. In Table 2, CREDO improves Overall accuracy over reproduced Temp-R1 on CronQuestion KG and direct ICEWS Actor transfer by 1.5 and 5.5 percentage points, respectively. On CronQuestion KG, the gains come from Medium and Complex questions while Simple accuracy is maintained, highlighting the benefit on tasks requiring multiple contextual facts to be integrated under the question's temporal constraints. The larger improvement on ICEWS Actor, with gains across all three complexity levels, further supports transfer beyond the training knowledge graph. The same Cron-trained checkpoint is applied directly without additional training, demonstrating that the learned policy remains effective in a new graph environment. Under matched training and evaluation settings, these results support evidence-conditioned credit assignment for both compositional temporal question answering and direct cross-KG transfer.

4.4 Analysis

Performance across training stages. Figure 5(a) summarizes the stage-wise evaluation records. CREDO achieves the highest final-stage accuracy, followed by CWO and our reproduced GRPO baseline. The matched final-stage comparison is consistent with a benefit from local credit and an additional improvement from outcome redistribution. Earlier intermediate-stage records are retained as reported; evaluation details appear in Appendix A.2.

Contribution of local credit and outcome redistribution. CWO (CREDO without outcome redistribution) retains local credit, allowing Figure 4 to examine the additional value of redistribution under the same two-stage schedule, selection criterion, and evaluation settings. CREDO improves Overall and Multiple accuracy by 2.2 and 3.75 percentage points, respectively, corresponding to net gains of 11 and six correct answers. Because both methods already recover local supervision, these observed differences suggest that controlling where outcome supervision acts can provide further value beyond local credit alone. The larger gain on Multiple questions is consistent with the motivation for emphasizing reasoning and answering in evidence-dependent tasks. Appendix A.3 reports the paired estimates and their uncertainty to contextualize the observed performance differences.

Redistribution variants. Random preserves eligibility, transferred mass, and target-token count; No-gate retains local credit and increases answer-region weights for trajectories satisfying structural and length conditions, without checking evidence overlap. CREDO exceeds both on Overall and Multiple under the common selection and evaluation protocol (Figure 4), favoring its evidence-conditioned configuration for determining where outcome supervision should act.

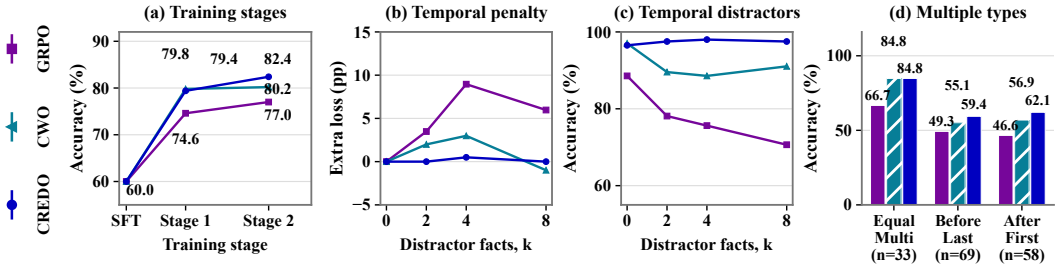

Figure 5: CREDO diagnostics. (a) Stage-wise accuracy on 500 MultiTQ inputs; the GRPO Stage 2 score is from our reproduction under the same training and evaluation settings as CREDO. (b) Additional substring-score loss from temporal rather than generic distraction. (c) Substring accuracy with temporal distractors, averaged over three decoding seeds on 67 questions. (d) Official Multiple types from the paired evaluation, with group sizes below the axis. Sources: Appendix A.2.

Official Multiple question types. In Figure 5(d), CREDO exceeds CWO on Before Last and After First, matches it on Equal Multi, and exceeds in-house GRPO throughout. These labels denote question types, not constraint counts (Appendix A.3).

Comparison under matched settings. Under identical base training and evaluation settings, CREDO exceeds our Temp-R1 reproduction by 5.4 percentage points in Overall accuracy and 13.75 points on Multiple questions. The improvement also holds against a baseline evaluated within our setup. The larger Multiple gain supports evidence-conditioned training's effectiveness on compositional temporal tasks. Appendix A.3 reports paired results and confidence intervals.

Answer containment under temporal distraction. Retaining supporting facts without retrieval, CREDO shows near-constant substring accuracy on 67 Single entity questions (Fig. 5(c)). On 60 Multiple questions, it leads in three evidence conditions despite CWO's smaller temporal-distractor drop. Both use three decoding seeds (construction and results: Appendix A.3).

Temporal-specific penalty. On the Single subset, Figure 5(b) compares generic and temporal distractors at matched counts and lengths, retaining supporting facts. CREDO combines a near-zero penalty with high substring accuracy (panel (c)), indicating stable target-answer containment with little additional loss from temporal distraction. GRPO's larger penalty highlights CREDO's advantage under fixed evidence. Negative penalties indicate higher accuracy with temporal distractors.

Hyperparameter sensitivity. Table 3 examines λ_{local} and ρ using Llama-3.2-3B-Instruct, varying one coefficient at a time while keeping all other training and evaluation settings fixed. For every setting, we evaluate the saved candidates on the same official 500-question MultiTQ set and report the highest accuracy on its 160 Multiple questions. The reported peaks are close across the tested values, indicating comparable attainable performance around the default configuration within this training budget. This exploratory comparison concerns peak performance rather than performance under the monitoring-based selection procedure used for the main 8B experiments.

Table 3: Peak Multiple accuracy on MultiTQ (3B), using the same evaluation settings as the 8B experiments. The default configuration is shown in bold.

λ_{local}	ρ	Multiple (%)
0.03	**0.04**	**45.00**
0.01	0.04	43.75
0.05	0.04	45.00
0.03	0.02	45.00
0.03	0.06	45.62

5 Conclusion

In this paper, we introduce CREDO for evidence-conditioned credit assignment in temporal reasoning agents. CREDO recovers positive local supervision through same-question comparisons and reallocates outcome-policy weight toward reasoning and answering based on retained answer overlap. It preserves terminal rewards, GRPO advantages, and total outcome-policy weight without an additional learned evaluator or inference-time credit computation. Results on MultiTQ and TimelineKGQA show gains in compositional question answering and cross-KG transfer, while component comparisons support the combined training procedure under a common evaluation protocol.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, code assistance, figure drafting, and research brainstorming. All AI-assisted text was reviewed and revised by the authors. AI-assisted code was manually inspected and tested against the intended experimental protocols, and all newly reported experimental results were obtained from our own implementations and runs. AI-assisted research suggestions were checked against the relevant literature, and generated figures were manually verified for consistency with the methods and results described in the paper. We take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies credit assignment for temporal question-answering agents using existing knowledge-graph benchmarks. Such agents may inherit biases, omissions, and outdated information from their base models and knowledge sources. Improved benchmark accuracy does not guarantee factual correctness or fairness outside the evaluated settings. Applications involving consequential decisions should therefore verify generated answers against reliable evidence and retain human oversight. Any reuse or redistribution of the underlying models and datasets should respect their licenses and access conditions.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 describe the credit-assignment procedure and policy objective. Sections 4.1 and 4.2 specify the model, retrieval setup, principal hyperparameters, benchmarks, and scoring protocol. Adopted baseline scores are distinguished from reproduced results. The paired analysis uses the main CREDO/CWO outputs and a separate in-house GRPO checkpoint. Monitoring/evaluation separation and model-selection details are disclosed in Appendix A.2. Reproducing these comparisons requires consistent checkpoints, data splits, retrieval settings, and answer normalization. The reported training trace and 3B sensitivity study describe individual runs and do not establish variability across random seeds.

REFERENCES

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025. doi: 10.48550/arXiv.2503.19470. URL https://arxiv.org/abs/2503.19470.

Zhuo Chen, Zhao Zhang, Zixuan Li, Fei Wang, Yutao Zeng, Xiaolong Jin, and Yongjun Xu. Self-improvement programming for temporal knowledge graph question answering. In *LREC/COLING*, pp. 14579–14594. ELRA and ICCL, 2024a.

Ziyang Chen, Xiang Zhao, Jinzhi Liao, Xinyi Li, and Evangelos Kanoulas. Temporal knowledge graph question answering via subgraph reasoning. *Knowl. Based Syst.*, 251:109134, 2022.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11378–11392. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.637. URL https://aclanthology.org/2023.acl-long.637/.

Ziyang Chen, Dongfang Li, and Xiang Zhao. Temporal knowledge question answering via abstract reasoning induction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 4872–4889, 2024b.

Wentao Ding, Hao Chen, Huayu Li, and Yuzhong Qu. Semantic framework based query generation for temporal question answering over knowledge graphs. In *EMNLP*, pp. 1867–1877. Association for Computational Linguistics, 2022.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Zhaoyan Gong, Juan Li, Zhiqiang Liu, Lei Liang, Huajun Chen, and Wen Zhang. RTQA : Recursive thinking for complex temporal knowledge graph question answering with large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9853–9870, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.499. URL https://aclanthology.org/2025.emnlp-main.499/.

Zhaoyan Gong, Zhiqiang Liu, Songze Li, Xiaoke Guo, Yuanxiang Liu, Xinle Deng, Zhizhen Liu, Lei Liang, Huajun Chen, and Wen Zhang. Temp-R1: A unified autonomous agent for complex temporal KGQA via reverse curriculum reinforcement learning, 2026. URL https://arxiv.org/abs/2601.18296.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Qianyi Hu, Xinhui Tu, Cong Guo, and Shunping Zhang. Time-aware ReAct agent for temporal knowledge graph question answering. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6013–6024, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.334. URL https://aclanthology.org/2025.findings-naacl.334/.

Qianyi Hu, Jiaxue Liu, Xinhui Tu, and Shoujin Wang. SAR: A structure-aligned reasoning framework for temporal knowledge graph question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31086–31094, 2026. doi: 10.1609/aaai.v40i37.40369. URL https://ojs.aaai.org/index.php/AAAI/article/view/40369.

Zicheng Huang, Yajuan Tong, Xinhui Tu, and Tingting He. TempTool-R1: Tool-augmented reinforcement learning for temporal knowledge graph question answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16224–16241, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.797. URL https://aclanthology.org/2026.findings-acl.797/.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. TEQUILA: temporal question answering over knowledge bases. In *CIKM*, pp. 1807–1810. ACM, 2018.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. Complex temporal question answering on knowledge graphs. In *CIKM*, pp. 792–802. ACM, 2021.

Zihao Jiang, Miao Peng, Zhenyan Shan, Wenjie Xu, Ben Liu, Gong Chen, Ziqi Gao, and Min Peng. TKG-Thinker: Towards dynamic reasoning over temporal knowledge graphs via agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2602.05818.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. doi: 10.48550/arXiv.2503.09516. URL https://arxiv.org/abs/2503.09516.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Shihao Liu, Xiaofei Zhou, Bo Wang, and Geyuan Zhang. Temporal evidence chain for temporal knowledge graph question answering with large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29515–29529, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1361. URL https://aclanthology.org/2026.acl-long.1361/.

Yonghao Liu, Di Liang, Fang Fang, Sirui Wang, Wei Wu, and Rui Jiang. Time-aware multiway adaptive fusion network for temporal knowledge graph question answering. In *ICASSP*, pp. 1–5. IEEE, 2023a.

Yonghao Liu, Di Liang, Mengyu Li, Fausto Giunchiglia, Ximing Li, Sirui Wang, Wei Wu, Lan Huang, Xiaoyue Feng, and Renchu Guan. Local and global: Temporal question answering via information fusion. In *IJCAI*, pp. 5141–5149. ijcai.org, 2023b.

Xufei Lv, Jiahui Yang, Yifu Gao, Linbo Qiao, and Houde Liu. Let the agent search: Autonomous exploration beats rigid workflows in temporal question answering. *arXiv preprint arXiv:2603.01853*, 2026. URL https://arxiv.org/abs/2603.01853.

Shichao Ma, Zhiyuan Ma, Ming Yang, Xiaofan Li, Xing Wu, Jintao Du, Yu Cheng, Weiqiang Wang, Qiliang Liu, Zhengyang Zhou, and Yang Wang. TSPO: Breaking the double homogenization dilemma in multi-turn search policy optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28607–28623, San Diego, California, United States, July 2026. Association for Computational Linguistics. URL https://aclanthology.org/2026.findings-acl.1428/.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N. Ioannidis, Adesoji Adeshina, Phillip Ryan Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. Tempoqr: Temporal question reasoning over knowledge graphs. In *AAAI*, pp. 5825–5833. AAAI Press, 2022.

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh K. Srivastava, Cezar Pendus, Saswati Dana, Dinesh Garg, Achille Fokoue, G. P. Shrivatsa Bhargav, Dinesh Khandelwal, Srinivas Ravishankar, Sairam Gurajada, Maria Chang, Rosario Uceda-Sosa, Salim Roukos, Alexander G. Gray, Guilherme Lima, Ryan Riegel, Francois P. S. Luus, and L. Venkata Subramaniam. SYGMA: system for generalizable modular question answering over knowledge bases. *CoRR*, abs/2109.13430, 2021.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng,

Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. Timer4 : Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *EMNLP*, pp. 6942–6952. Association for Computational Linguistics, 2024.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, and Xiaojie Yuan. Plan of knowledge: Retrieval-augmented large language models for temporal knowledge graph question answering, 2025. URL https://arxiv.org/abs/2511.04072.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6663–6676. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.520. URL https://aclanthology.org/2021.acl-long.520/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

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. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Aditya Sharma, Apoorv Saxena, Chitrank Gupta, Seyed Mehran Kazemi, Partha P. Talukdar, and Soumen Chakrabarti. Twirgcn: Temporally weighted graph convolution for question answering over temporal knowledge graphs. In *EACL*, pp. 2041–2052. Association for Computational Linguistics, 2023.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher: Incentivizing the search capability in LLMs via reinforce-

ment learning. *arXiv preprint arXiv:2503.05592*, 2025. doi: 10.48550/arXiv.2503.05592. URL https://arxiv.org/abs/2503.05592.

Qiang Sun, Sirui Li, Du Huynh, Mark Reynolds, and Wei Liu. TimelineKGQA: A comprehensive question-answer pair generator for temporal knowledge graphs. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 797–800. Association for Computing Machinery, 2025. doi: 10.1145/3701716.3715308.

Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, Liming Zhu, and Wenjie Zhang. MemoTime: Memory-augmented temporal knowledge graph enhanced large language model reasoning. In *Proceedings of the ACM Web Conference 2026*, pp. 4220–4231. Association for Computing Machinery, 2026. doi: 10.1145/3774904.3792581.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization, 2025. URL https://arxiv.org/abs/2505.15107.

Yutao Xie, Nathaniel Thomas, Nicklas Hansen, Yang Fu, Li Erran Li, and Xiaolong Wang. TIPS: Turn-level information-potential reward shaping for search-augmented LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fd881d3b625437354d4421818f81058f-Abstract-Conference.html.

Yaocheng Zhang, Haohuan Huang, Zijun Song, Zijie Zhao, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. CriticSearch: Fine-grained credit assignment for search agents via a retrospective critic. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12272–12290, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.596. URL https://aclanthology.org/2026.findings-acl.596/.

Zefang Zong, Dingwei Chen, Yang Li, Qi Yi, Bo Zhou, Chengming Li, Bo Qian, Peng Chen, and Jie Jiang. AT^{2}PO: Agentic turn-based policy optimization via tree search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24120–24143, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1106. URL https://aclanthology.org/2026.acl-long.1106/.

A Additional Details and Results

A.1 Implementation Details

Evidence and answer processing. Local scoring lowercases text, replaces ASCII punctuation with spaces, removes English articles, and collapses whitespace. Coverage is the fraction of reference-answer tokens found in a fact line, using multiset overlap; local answer F1 uses the same normalization. Retrieved evidence pools all complete Information blocks. The retained selection uses the last complete Rank block, falling back to the last complete Filter when that Rank is absent or empty. Document prefixes and title wrappers are removed before checking normalized selected lines against the retrieved-line set. Untraceable lines are discarded individually, and surviving lines are reindexed before positional discounting. Duplicate selected lines are retained. This response-wide provenance check verifies the retrieved source, not semantic or temporal sufficiency.

The local answer extractor accepts complete Answer tags case-insensitively and selects the last match shorter than 200 characters, or the shortest match if all are longer. The archived terminal extractor accepts lowercase tags and, when any short match exists, returns the last match regardless of length; otherwise it returns the shortest. Terminal normalization deletes punctuation rather than replacing it with spaces. The two branches can therefore evaluate different extracted or normalized text when outputs contain multiple Answer blocks. Missing answers receive zero quality or score.

Validity and token localization. Let T_i indicate truncation, C_i^Q reliable Search/Information structure with an Information response before the next Search, C_i^E reliable Filter/Rank structure, and N_i^E non-empty selection content. The stage-local rules are

$$\begin{aligned} v_i^Q &= C_i^Q \vee \neg T_i, \\ v_i^E &= v_i^Q \wedge (q_i^Q > \epsilon_0) \wedge [\neg T_i \vee (C_i^E \wedge N_i^E)], \\ v_i^R &= v_i^E \wedge (c_i^{\text{sel}} > \epsilon_0) \wedge \neg T_i, \qquad \epsilon_0 = 10^{-8}. \end{aligned} \tag{10}$$

Here ϵ_0 is the coverage threshold δ in the main text. These validity rules govern local comparisons and the evidence-conditioned CREDO route. No-gate retains them for local credit but uses structural and token-length eligibility for redistribution, without the overlap requirements in v_i^R (Appendix A.2). Separately, t_i^s requires a non-empty reliable token mask. Thus, an evaluable failure may contribute a zero-quality comparison without supplying teacher tokens.

The retrieval mask contains complete Search contents, and the selection mask aggregates complete Filter and Rank contents. The reasoning/answering mask contains complete Think contents starting after the last Information block (or the last Search when Information is absent), together with complete Answer contents, subject to structural validity. Tokens must lie entirely inside their content spans. Tokenizer offsets, with prefix-decoding checks when needed, align text spans to tokens. Global alignment failures, overlapping stage masks, or out-of-bounds masks disable local credit. Plan, tags, and unclassified model text receive no local mask but may retain outcome supervision. Information and padding are excluded from actor updates. These masks aggregate actions across rounds; the credit is stage-level rather than a separate estimate for each action.

Stage budgets. Budgets are computed per complete question group g before PPO minibatch partitioning. Write $V_{gs} = |\mathcal{V}_{gs}|$ for the valid comparison count and K_g for the sampled group size. Let $(o_i^Q, o_i^E, o_i^R) = (1, q_i^Q, c_i^{\text{sel}})$, and let $\mathcal{H}_{gs}$ contain positive local teachers with non-empty masks. For $V_{gs} \geq 2$, define

$$\begin{aligned} b_{gs} &= V_{gs}/K_g, & \bar{o}_{gs} &= \frac{1}{V_{gs}} \sum_{i \in \mathcal{V}_{gs}} o_i^s, \\ d_{gs} &= \frac{\sum_{i,j \in \mathcal{V}_{gs}} |q_i^s - q_j^s|}{V_{gs}(V_{gs}-1)}, & e_{gs} &= \left[1 - \frac{1}{V_{gs}} \sum_{i \in \mathcal{V}_{gs}} q_i^s \right]_+ . \end{aligned} \tag{11}$$

These factors represent comparison support, opportunity, quality dispersion, and remaining improvement. The budget is

$$\ell_{gs} = b_{gs} \bar{o}_{gs} d_{gs} e_{gs} \mathbf{1}[|\mathcal{H}_{gs}| > 0], \qquad \alpha_{gs} = \begin{cases} \ell_{gs} / \sum_{s'} \ell_{gs'}, & \sum_{s'} \ell_{gs'} > \epsilon_0, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Incomplete groups, unavailable comparisons, non-finite values, and stages without positive teachers receive zero corresponding credit. Frontier ties contribute zero. Equation 7 then divides each stage budget equally among its teachers and their stage tokens. Budgets are normalized across stages within a question, without additional question-difficulty weighting.

Policy loss and regularization. GRPO standardizes each trajectory reward within its question group using the sample standard deviation and denominator offset 10^{-6}. Local credit does not enter this reward. Outcome redistribution uses Equation 6, with a 10^{-6} tolerance in the minimum token-fraction check and implementation range $0 \le \rho < 1$. The net weight transferred from outside the reasoning/answering region is $\rho(n_i - m_i)$.

In minimization form, the PPO implementation uses the clipped ratio $r = \exp(\mathrm{clip}(\log \pi_\theta - \log \pi_{\mathrm{old}}, -20, 20))$ and dual clipping:

$$\begin{aligned} \phi(a, r) &= \max\{-ar, -a\,\mathrm{clip}(r, 0.8, 1.2)\}, \\ \widetilde{\phi}(a, r) &= \begin{cases} \min\{\phi(a, r), -3a\}, & a < 0, \\ \phi(a, r), & a \ge 0. \end{cases} \end{aligned} \tag{13}$$

The local advantages are non-negative and do not activate the negative-advantage branch. Saved old-policy log probabilities are used for both terms.

Disjoint stage masks define token-level local advantages $a^{\mathrm{loc}}_{it} = \sum_s M^s_{it} a^s_i$ and weights $u_{it} = \sum_s u^s_{it}$. For microbatch B_μ, the policy losses are

$$\begin{aligned} L_{\mathrm{out},\mu} &= \frac{\sum_{(i,t)\in B_\mu} w_{it} \widetilde{\phi}(\widehat{A}_i, r_{it})}{\sum_{(i,t)\in B_\mu} w_{it}}, \\ L_{\mathrm{loc},\mu} &= \frac{\sum_{(i,t)\in B_\mu} u_{it} \widetilde{\phi}(a^{\mathrm{loc}}_{it}, r_{it})}{\max(U_\mu, 10^{-6})}, \qquad U_\mu = \sum_{(i,t)\in B_\mu} u_{it}. \end{aligned} \tag{14}$$

Both training stages use the legacy actor implementation with static microbatches and token-mean loss aggregation (`use_legacy_worker_impl=enable`, `use_dynamic_bsz=False`, and `loss_agg_mode=token-mean`). The executed path receives no global token-count metadata, so the outcome denominator is computed within the current microbatch. Since redistribution preserves each trajectory's total weight, this denominator equals the number of valid actor tokens in that microbatch, excluding tool observations and padding. Redistribution therefore changes relative token weighting without changing the denominator for a fixed microbatch.

For a PPO minibatch, local weight is summed across the default distributed process group to obtain U_*, with world size W. After accumulation scaling, the loss contributions are

$$\sum_\mu \left[\gamma_\mu \big(L_{\mathrm{out},\mu} + \beta L_{\mathrm{KL},\mu} - \zeta H_\mu \big) + \lambda_{\mathrm{local}} \frac{W U_\mu}{U_*} L_{\mathrm{loc},\mu} \right]. \tag{15}$$

The local term is omitted if $U_* = 0$. Static microbatches use the reciprocal accumulation count for γ_μ; the local branch compensates for that scaling before backward. Stage budgets therefore retain their relative weights inside the local loss. Each PPO minibatch ends with an optimizer update. Outcome losses are averaged through static microbatch accumulation, rather than normalized by a single token count over the entire PPO minibatch.

KL is an independent actor loss and is not added to the checked training rewards. The low-variance estimator is $\mathrm{clip}(\exp(z) - z - 1, -10, 10)$ for $z = \mathrm{clip}(\log \pi_{\mathrm{ref}} - \log \pi_\theta, -20, 20)$. KL and entropy use the original actor response mask, not redistributed or local weights. Weight conservation thus does not imply invariant gradient magnitude.

A.2 Experimental Setup

Data and training. The fixed MultiTQ evaluation contains 340 Single and 160 Multiple questions, with 371 entity and 129 time answers. The year/month/day counts are 69/105/326. These are official categories, not inferred constraint counts. The processed data records contain 57,979 MultiTQ training-pool entries and 54,584 test entries, and 24,427 Cron training entries and 8,146

test entries; these describe the local processed files rather than the original benchmark release sizes. Normalized-question overlap between the processed MultiTQ training and test files is zero. The monitoring sets are seed-42 samples: 24 questions from `test_multiple` for Stage 1 and 30 questions from `test_all` for Stage 2. Both monitoring sets are disjoint from the fixed 500-question evaluation set. Checkpoints for the main comparisons are selected using monitoring accuracy; the 500-question evaluation scores are not used for this selection.

The policy is Meta-Llama-3.1-8B-Instruct, with E5-base-v2 retrieval and a FAISS Flat index. Table 4 gives the common MultiTQ settings for the in-house GRPO, CREDO, CWO, Random, and No-gate runs. Each starts from the same initial SFT weights and continues its own Stage 1 step-80 checkpoint into Stage 2 with a reinitialized optimizer. The main CREDO result uses Stage 2 step 400. Defaults are $\lambda_{\text{local}} = 0.03$, $\rho = 0.04$, and $\eta = 0.05$. CWO is the configuration named CF in archived files: local credit remains active, while outcome redistribution is disabled.

Table 4: Common base training configuration for the five in-house 8B methods on MultiTQ. Generation limits and response-container limits measure different quantities.

Setting	Stage 1	Stage 2
Scheduled updates	80	500
Questions per update × rollouts	60×5	2×8
PPO minibatch trajectories	30	2
Data / sequence parallelism	1 / 8	2 / 4
Learning rate	5×10^{-7}	5×10^{-7}
Gradient clipping	5	1
KL / entropy coefficients	0.01 / 0.001	0.001 / 0.001
Sampling temperature / top-p	1 / 0.95	1 / 0.95
Retrieval top-k	30	20
Maximum searches / model calls	3 / 4	3 / 4
Per-call generation limit	1,000	1,024
Configured generation / response limit	4,000 / 5,200	4,096 / 8,192
Per-call Information limit	1,400	1,500
Monitoring interval / questions	10 / 24	20 / 30

Controlled ablations and model selection. Both training stages use normalized substring rewards in GRPO, CREDO, CWO, Random, and No-gate. The archived configurations agree on batch sizes, rollout counts, optimizer settings, KL and entropy coefficients, token budgets, retrieval settings, and scheduled updates. The CREDO-family Stage 1 input manifests record identical hashes for 12 assets, including all four SFT weight shards, processed data files, the knowledge graph, and the retrieval index. GRPO uses outcome-only supervision; CWO removes outcome redistribution; Random changes its target positions while preserving eligibility, target count, and transferred mass. No-gate retains CWO's local-credit update. Its redistribution route requires reliable reasoning/answering-token localization, $n_i, m_i > 0$, and $m_i/n_i \geq \eta$, together with the structural validity checks, but does not require $v_i^R = 1$ or positive retrieved/retained answer overlap. Eligible trajectories use the same ρ and weight transformation in Eq. 6; local-credit validity is unchanged. This compares structurally eligible answer-region redistribution with CREDO's evidence-conditioned configuration.

For the main 8B comparisons, GRPO, CREDO, CWO, Random, and No-gate use monitoring-set accuracy to select among their available saved checkpoints, followed by evaluation under identical settings. All methods share the selection criterion, while candidate sets depend on checkpoint availability. In Stage 1, checkpoints for CREDO, CWO, Random, and No-gate were saved every 10 steps; in Stage 2, checkpoints were saved every 100 steps. All methods initialize Stage 2 from their own Stage 1 step-80 checkpoint. The final 500-question evaluation scores are not used for model selection in these main comparisons.

Evaluation protocol. We adopt normalized substring matching to retain compatibility with Temp-R1's public evaluation implementation (Gong et al., 2026). We verified the scoring behavior at commit `44ab58f`, with validation format and retrieval bonuses set to zero. Its evaluation entry point routes MultiTQ and TimelineKGQA to `qa_em_format.compute_score_em`; despite the function name, the underlying `em_check` tests substring containment rather than string equality. Extracted answers are lowercased, stripped of punctuation and English articles, and whitespace-normalized. A prediction receives one point if it contains any normalized acceptable answer; missing

answers and execution failures receive zero. The earlier Search-R1 code also provides a normalized substring scorer (Jin et al., 2025), alongside an exact-match scorer. Our choice follows the scoring branch used by the direct temporal-agent baseline, rather than assuming that all prior implementations use identical metrics. The same rule is applied to all in-house comparisons. This establishes protocol consistency while retaining the metric's answer-containment interpretation; additional candidates can still receive credit.

The main MultiTQ comparison uses one evaluation configuration for CREDO, CWO, Random, and No-gate, giving Overall scores of 82.4% (412/500), 80.2% (401/500), 79.0% (395/500), and 78.4% (392/500), respectively. Evaluation uses seed 42, temperature 1, top-p 0.95, generation/container limits of 5,000/6,200, Information limit 2,400, retrieval top-k 30, and up to three searches plus a final round.

MultiTQ baseline sources. Our in-house outcome-only GRPO baseline shares CREDO's SFT initialization, training data, terminal reward, answer-extraction rules, and base training and evaluation settings; local credit and outcome redistribution are disabled. It scores 385/500 (77.0%) and provides the GRPO Stage 2 endpoint in Figure 5(a), the predictions for Tables 5–6, and the frozen-model diagnostics. Table 1 and Figure 4 retain the published Temp-R1 reference of 78.0% (Gong et al., 2026). The literature comparison and our matched reproduction therefore use distinct result sources: CREDO's Overall improvements are 4.4 and 5.4 percentage points, respectively. The final-stage GRPO, CWO, and CREDO scores use the common evaluation configuration. Earlier intermediate-stage values in Figure 5(a) remain the reported records; the curve does not constitute a controlled re-evaluation of every intermediate checkpoint.

TimelineKGQA reproduction. The public materials available for our experiments did not provide the complete TimelineKGQA training pipeline needed to reproduce the published Temp-R1 models. We therefore train Temp-R1 and CREDO in our implementation with identical base configurations, including initialization, training data, optimizer settings, training budget, and retrieval environment. Evaluation follows the public normalized substring protocol, with identical inputs, decoding settings, token limits, interaction rules, and extraction routines. This provides a controlled comparison within our training setup. The reproduced Temp-R1 scores (50.8% Cron, 30.9% ICEWS) differ from the published 70.5%/64.2% (Gong et al., 2026); we do not claim numerical reproduction of those models or superiority over their reported scores.

Each model is evaluated on 800 Cron and 800 ICEWS rows, with correct counts of 406 versus 418 and 247 versus 291. Cron-trained checkpoints transfer directly to ICEWS without further training. Simple/Medium/Complex counts are 250/245/305 for Cron and 351/276/173 for ICEWS. The sets contain 797 and 793 unique question/answer keys; duplicate rows remain in the averages.

PoK is not included in our TimelineKGQA reproduction because a complete public implementation was unavailable to us. Its published Cron/ICEWS scores of 65.1%/60.2%, reported in Temp-R1, remain external reference results. Other adopted rows in Table 2 provide context; the controlled comparison is reproduced Temp-R1 versus CREDO, rather than an exhaustive ranking of published systems.

Table 3 uses Llama-3.2-3B-Instruct. All 3B runs use 80 training steps with a training batch size of 32; the corresponding 8B Stage 1 runs use 80 steps with a batch size of 60. Within the 3B study, all settings share the same SFT initialization, training data, base training configuration, and evaluation protocol. Only λ_{local} or ρ is varied relative to the default. Evaluation uses the same settings and official 500-question MultiTQ set as the main 8B experiments, comprising 340 Single and 160 Multiple questions. The table reports only Multiple accuracy.

A.3 Supplementary Results

Formal component comparisons. CREDO's observed Overall gain is 2.2 percentage points over CWO and 3.4 points over the selected Random model. The CREDO/CWO Multiple comparison is 105/160 versus 99/160, a 3.75-point difference. CREDO, CWO, Random, and No-gate use the same best-monitoring-checkpoint criterion and final evaluation settings. No-gate obtains 78.4% Overall and 57.5% Multiple (92/160); CREDO's point estimates are higher by 4.0 and 8.13 percentage points, respectively. The component figure also includes Random's reported Multiple accuracy of

57.5% and the literature GRPO reference (78.0% Overall, 55.0% Multiple) from Temp-R1. The latter is a reported baseline rather than a new independent run.

Question-type results. Table 5 gives the complete supplementary breakdown. Each model has 500 unique question identifiers with matching question/answer hashes. Missing extracted answers number 3, 3, and 6 for GRPO, CWO, and CREDO and remain in the denominator. CREDO's gains over GRPO cover all three Multiple categories, while Single-category gains are not uniform.

Table 5: Paired MultiTQ results: correct answers (accuracy, %). CREDO and CWO also provide the main results; GRPO is our matched in-house baseline. Aggregate rows overlap the individual categories.

Category	N	GRPO	CWO	CREDO
Equal	174	156 (89.66)	151 (86.78)	152 (87.36)
Before/After	94	81 (86.17)	89 (94.68)	87 (92.55)
First/Last	72	65 (90.28)	62 (86.11)	68 (94.44)
Equal Multi	33	22 (66.67)	28 (84.85)	28 (84.85)
Before Last	69	34 (49.28)	38 (55.07)	41 (59.42)
After First	58	27 (46.55)	33 (56.90)	36 (62.07)
Single	340	302 (88.82)	302 (88.82)	307 (90.29)
Multiple	160	83 (51.88)	99 (61.88)	105 (65.63)
Entity	371	263 (70.89)	282 (76.01)	286 (77.09)
Time	129	122 (94.57)	119 (92.25)	126 (97.67)
Overall	500	385 (77.00)	401 (80.20)	412 (82.40)

Paired comparisons. Table 6 reports corrections, regressions, and net differences on the same inputs. Intervals use 100,000 paired question bootstrap samples, seed 20260925, and percentile endpoints at 2.5% and 97.5%, without multiple-comparison correction. Against the matched in-house GRPO checkpoint, corrections exceed regressions by 27 questions Overall and 22 on Multiple, with both paired intervals above zero. The CWO intervals include zero and quantify uncertainty in the main CREDO–CWO difference. These intervals are conditional on the selected checkpoints and quantify question-sampling uncertainty, not training-seed variation or checkpoint-selection uncertainty.

Table 6: Paired supplementary comparisons. "Corrected" means only CREDO is correct; "Regressed" means only the comparator is correct. Differences and intervals are in percentage points.

Comparator / subset	N	Corrected	Regressed	Difference	95% CI
GRPO / Overall	500	50	23	$+5.40$	$[2.20, 8.80]$
CWO / Overall	500	34	23	$+2.20$	$[-0.80, 5.20]$
GRPO / Multiple	160	33	11	$+13.75$	$[6.25, 21.88]$
CWO / Multiple	160	19	13	$+3.75$	$[-3.13, 10.63]$

All-wrong groups. Across 80 Stage 1 updates, 4,800 question-group exposures yield 24,000 trajectories ($K = 5$). Of these groups, 1,473 are all-wrong (30.69%); 564 have active local credit (11.75% of all groups; 38.29% of all-wrong groups). Both curves use all groups as the denominator. Rewards are binary substring correctness, with format/retrieval bonuses disabled and KL excluded, so all-wrong groups have zero outcome advantages. Within active all-wrong groups, retrieval, selection, and reasoning teachers occur in 351, 350, and 59 groups, respectively (overlapping counts). These observations establish local-signal availability at all three stages, not gradient magnitude or an isolated performance contribution.

Single-question distractors. Among the fixed 500 inputs, 84 questions have parseable explicit dates, 72 pass rule-based KG answer checks and finite evidence construction, and 67 admit matched distractor conditions. All 67 are Single questions with entity answers: 60 Equal and 7 Before/After, with day/month/year counts 51/13/3. They are not an independently human-labeled evidence-sufficiency set.

The prompt instructs the model to answer from supplied evidence without tools and end with an Answer block. The user input contains the question and numbered Information lines. Retrieval is disabled; decoding uses temperature 1, top-p 0.95, generation limit 1,024, context limit 11,200, and seeds 42/43/44 with question-hash offsets. The first complete Answer block stops generation.

Over-budget inputs are rejected rather than truncating support. Generic and temporal distractors are matched in count, length, format, and insertion position. The former are unrelated facts; the latter are topically related but violate the question's temporal condition. Supporting facts remain fixed for $k \in \{0, 2, 4, 8\}$; an additional condition contains unrelated evidence only.

Table 7: Answer containment under fixed evidence. Entries are correct generations out of 201 (substring accuracy, %), from 67 questions and three decoding seeds.

Evidence condition	GRPO	CWO	CREDO
Supporting facts only	178 (88.56)	195 (97.01)	194 (96.52)
Generic, $k = 2$	164 (81.59)	184 (91.54)	196 (97.51)
Generic, $k = 4$	170 (84.58)	184 (91.54)	198 (98.51)
Generic, $k = 8$	154 (76.62)	181 (90.05)	196 (97.51)
Temporal, $k = 2$	157 (78.11)	180 (89.55)	196 (97.51)
Temporal, $k = 4$	152 (75.62)	178 (88.56)	197 (98.01)
Temporal, $k = 8$	142 (70.65)	183 (91.04)	196 (97.51)
Unrelated evidence only	4 (1.99)	3 (1.49)	15 (7.46)

The temporal penalty subtracts temporal-condition accuracy from generic-condition accuracy at matched k; negative values favor temporal distractors. Table 7 pools three decoding repetitions per question, with equal numbers of repetitions for every question. These are 67 questions, not 201 independent samples or training runs. This Single-question result concerns target containment with retained support; it does not directly test unique-answer selection, retrieval quality, or Multiple-question composition.

Multiple-question distractors. We additionally evaluate 60 Multiple questions, with 20 from each official type: Equal Multi, Before Last, and After First. The 8B models are GRPO at step 200 and CWO/CREDO at step 400. Each model answers every question under three decoding seeds in each condition, yielding 180 generations per model and condition. Complete supporting facts and sorting candidates are retained. Generic and temporal conditions each add four facts; inputs are identical across models, retrieval is disabled, and no input is truncated. Scoring follows the same official normalized substring protocol.

Table 8: Multiple-question answer containment with complete support retained. Entries are correct generations out of 180 (substring accuracy, %), from 60 questions and three decoding seeds. Bold indicates the highest accuracy in each condition.

Evidence condition	GRPO	CWO	CREDO
Complete evidence	60 (33.33)	114 (63.33)	**125 (69.44)**
Generic, $k = 4$	58 (32.22)	96 (53.33)	**119 (66.11)**
Temporal, $k = 4$	62 (34.44)	97 (53.89)	**100 (55.56)**

CREDO achieves the highest absolute accuracy in every condition (Table 8). Under temporal distraction, its advantage over CWO narrows to three correct generations (1.67 percentage points). Relative to complete evidence, CREDO loses 13.89 percentage points and CWO loses 9.44; relative to generic distraction, CREDO loses 10.56 points. Thus, the advantage in absolute accuracy coexists with greater sensitivity to temporal distraction in this Multiple setting. These results extend the evaluation to compositional question types, but do not establish uniformly stronger temporal robustness or isolate correct reasoning from answer containment. The 180 generations represent repeated decoding on 60 questions, not independent questions or training runs.

Hyperparameter sensitivity. Starting from the default $(\lambda_{\text{local}}, \rho) = (0.03, 0.04)$, we vary $\lambda_{\text{local}} \in \{0.01, 0.03, 0.05\}$ with ρ fixed, and vary $\rho \in \{0.02, 0.04, 0.06\}$ with λ_{local} fixed. All runs share the same 3B backbone, SFT initialization, training data, base training configuration, and evaluation protocol. Every row follows the same reporting rule: evaluate the available candidate checkpoints on the official 500-question MultiTQ set using normalized substring matching, and report the highest accuracy on the same 160 Multiple questions. This rule applies to the default configuration and all coefficient variants. The table therefore compares observed peak performance across parameter settings; unlike the main 8B comparisons, checkpoint selection in this exploratory analysis uses the reported evaluation subset. The results do not estimate performance under independent monitoring-based selection or variability across training seeds.

A.4 Discussion and Limitations

Evidence proxies. Answer overlap does not test semantic sufficiency. With answer tokens preserved, all 40 rule-checked Single examples retain positive coverage in each of the original, wrong-time, and wrong-relation conditions. This diagnoses coverage alone, not the full provenance/eligibility pipeline. Natural-trajectory gate precision and recall remain unmeasured; aggregated masks provide stage-level, not per-action, credit.

Mechanism and stability. Component comparisons do not isolate all-wrong updates, component interactions, or individual local-credit ingredients. Best-checkpoint comparisons describe the selection protocol rather than fixed-update effects. The training trace and 3B study use single runs; decoding repetitions and paired question intervals do not measure training-seed stability.