ONE ANSWER TO GRADE THEM ALL? REFERENCE-BASED EVALUATION AND ITS OPTIMIZATION EFFECTS IN SEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Search agents have become increasingly capable of tackling complex questions, with reliable judgments of answer correctness providing the basis for assessing their capabilities and guiding further learning. Search-agent evaluation and training commonly rely on reference answers, although recent work incorporates verified alternative answers into training rewards. However, the implications of *incomplete reference coverage* for evaluation and supervision remain insufficiently understood in more challenging web-search tasks. Questions combining multiple constraints may still admit valid solutions beyond the references, causing erroneous rejection and potentially misleading supervision. To examine this problem, this study uses evidence-based verification to quantify grading discrepancies and analyzes their sources through question constraints and supporting evidence. Policy updates under original and corrected rewards are then compared to examine how grading errors affect supervision of answers and search actions. Extensive experiments across five search benchmarks show that reference-based evaluation can underestimate performance by rejecting valid answers, with verified revisions increasing the acceptance rate by 2.12 pp. Source-based attribution identifies reference omissions as a distinct error category arising from one-to-many relations or unspecified selection rules. When used as rewards, these misjudgments can lower the likelihood of valid answers and useful search actions, depending on response content, group composition, and search context. These findings motivate checking for competing valid solutions when constructing questions and references for evaluation and learning.

1 INTRODUCTION

Search agents combine retrieval, browsing, and reasoning to address complex information needs (Yao et al., 2023; Wu et al., 2025a; Li et al., 2025). Evaluation judges whether final answers satisfy the questions' requirements. In outcome-supervised reinforcement learning, these judgments also reward the preceding search process (Jin et al., 2025; Chen et al., 2025b; Song et al., 2025). Reliable answer assessment therefore supports both capability evaluation and the reinforcement of search behaviors that lead to valid, evidence-supported answers.

Current methods commonly base these judgments on comparisons with supplied reference answers (Jin et al., 2025; Wei et al., 2025). These comparisons range from normalized matching to language-model judgments of semantic equivalence. To accommodate distinct valid answers beyond such equivalence, recent work on open-domain and multi-hop QA expands answer sets through evidence verification and incorporates multi-answer rewards into training (Zhang et al., 2026).

However, in more challenging web-search tasks, the extent of grading errors, their relation to *incomplete reference coverage*, and their effects on answer and search-action updates remain insufficiently understood. Questions combining multiple constraints may still admit distinct valid solutions under the same interpretation, while the references cover only some of them. The examples in Figure 1 illustrate this gap: different game editions meet the same requirements, and distinct films satisfy

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.

(a) Game editions WebWalkerQA · 210

Which version of Sid Meier's Civilization VII should a user purchase to start playing on **February 6, 2025**, and access a civilization with the '**Nepekifaki**' ability?

Constraint	Supplied reference: Founders Edition	Fixed answer: Deluxe Edition
Play on February 6, 2025	Early access included	Early access included
Civilization with Nepekifaki	Shawnee · Tecumseh and Shawnee Pack	Shawnee · Tecumseh and Shawnee Pack
Sources: pre-order announcement · Shawnee civilization guide		

Reference for fixed Deluxe answer	Founders	Deluxe	Founders + Deluxe	Deluxe + Founders
Reference-based verdict	**Reject**	Accept	Accept	Accept

(b) Film identities BrowseComp · Qwen3.7-Plus

Provide the name of the movie that fits the following criteria: • This movie was **released in the 1940's**. • The movie's director also took on **another key responsibility** in this project. • One of the actors had **a brother who was an author**. • Another actor was born in **London, England in the 1910s**. • The movie had **an additional English title** besides its original title.

Constraint	Supplied reference: *Murder at the Windmill*	Generated answer: *A Matter of Life and Death*
1940s release	1949	1946
Director's other role	Val Guest · writer / director	Powell & Pressburger · writers / directors
Actor → author brother	Garry Marsh → Digby George Gerahty	Richard Attenborough → David Attenborough
London-born in the 1910s	Jon Pertwee · 1919	David Niven · 1910
Other English title	*Mystery at the Burlesque*	*Stairway to Heaven*
	BFI · biographical sources	MoMA · BBFC · biographical sources

Figure 1: **Incomplete reference coverage in search tasks.** (a) Both game editions meet the requirements, yet a fixed concise Deluxe answer is rejected. (b) Two distinct films satisfy all five clues in the question. Supporting source documentation is provided in Appendix G.1.

all five clues. In both cases, supporting evidence establishes the reference's validity, not its uniqueness. Rejecting these alternatives can underestimate performance, and using the same judgments as rewards can penalize valid answers and useful search actions.

To address this problem, large-scale repeated sampling collects search trajectories across policies and benchmarks for reference-based grading. Rejected answers are then verified against external evidence without using the references, with strong-model checks and sampled human review supporting the resulting score revisions. Analysis of question constraints and source passages distinguishes reference omissions from other grading errors and identifies structures that permit multiple valid answers. Finally, comparisons of parameter updates under original and corrected rewards test how these errors affect supervision. Holding the initial model and trajectories for each question fixed isolates how reward correction changes answer and action likelihoods.

Experiments across five search benchmarks show that reference-based evaluation can underestimate performance by rejecting valid task answers, with the verified revisions increasing the acceptance rate by 2.12 pp. A separate source-attribution analysis identifies reference omissions alongside defective references and other grading errors. In these omission cases, one-to-many relations or unspecified selection rules permit several complete answers, while the references cover only some of them. These evaluation errors also affect learning when the judgments are used as rewards: parameter updates can reduce the likelihood of valid answers and useful searches. Answer-likelihood changes vary with response content and group composition. Search-likelihood changes depend on earlier reasoning and retrieved information and can extend to trajectories with valid answers. These findings motivate checks for competing solutions during question and reference construction.

This study makes three contributions:

- **Evaluation discrepancies.** Evidence-based verification, supported by strong-model checks and sampled human review, quantifies score revisions across five search benchmarks and multiple policies and documents rejected valid task answers.
- **Error sources and question structure.** The analysis distinguishes reference omissions from defective references and other grading errors, and identifies question structures that permit competing valid solutions to guide question and reference construction.
- **Optimization effects.** Paired comparisons measure how updates under original and corrected rewards affect valid answer and useful action likelihoods, revealing their dependence on response content, group composition, and search history.

2 Related Work

Search agents and outcome supervision. Search-agent research connects tool use with learning from answer-level rewards. ReAct interleaves reasoning with actions and observations (Yao et al., 2023). Building on such interactions, Search-R1, ReSearch, and R1-Searcher learn multi-turn search from outcome rewards (Jin et al., 2025; Chen et al., 2025b; Song et al., 2025). WebDancer, WebSailor, and WebExplorer extend training to longer web interactions, while AgentCPM-Explore studies compact agents (Wu et al., 2025a; Li et al., 2025; Liu et al., 2025; Chen et al., 2026). How these rewards guide learning also depends on the optimization procedure: group-relative policy optimization (GRPO) weights trajectories by comparing rewards for the same question (Shao et al., 2024). Noise-corrected GRPO analyzes reward corruption and corrects gradient bias using estimated flip probabilities (Mansouri et al., 2026), while experiments with spurious rewards show that training effects depend on the model family (Shao et al., 2026). Together, these studies show that answer-level supervision depends on both what the rewards measure and how the optimization procedure uses them to weight search trajectories.

Answer coverage and ambiguity. Determining what an answer-level reward should recognize requires accounting for both variation in expression and multiplicity of valid answers. Semantic-equivalence evaluation addresses the former by recognizing equivalent answers beyond token overlap (Bulian et al., 2022). Incomplete reference annotations pose a further problem: manual reviews show that they can underestimate generative QA performance (Kamalloo et al., 2023). For questions with multiple interpretations, AmbigQA annotates the corresponding answers, and ASQA evaluates long-form responses covering those interpretations (Min et al., 2020; Stelmakh et al., 2022). QAMPARI instead makes multiplicity explicit in the task by requesting sets of entities (Amouyal et al., 2023). A^{2}Search incorporates multiple-answer supervision into open-domain and multi-hop QA training by expanding answer sets through evidence verification and optimizing answer-level F1 rewards (Zhang et al., 2026). Together, these approaches broaden evaluation and supervision to recognize equivalent expressions, alternative interpretations, and multiple valid answers, while leaving unresolved how rejecting valid alternatives affects answer and search-action likelihoods through local parameter updates under original and corrected rewards.

Evidence-based verification. Verification methods differ in what they assess and how they establish correctness. Outcome and process verifiers assess complete solutions and intermediate steps, respectively (Cobbe et al., 2021; Lightman et al., 2024). For factual responses, FActScore assesses support for individual claims, while FacTool and SAFE retrieve external evidence for factual assessment (Min et al., 2023; Chern et al., 2023; Wei et al., 2024). The judgment procedure itself also affects reliability: evaluations of language-model judges document sensitivities to answer presentation (Zheng et al., 2023). Across these studies, verification reliability depends on the unit being assessed, the evidence available, and the behavior of the judge.

3 Errors in Reference-Based Grading

3.1 Problem definition

A question may admit several complete valid answers while its reference represents only one. Let q denote the question, t_{eval} the time at which its answer must be valid, and $\mathcal{A}^*(q, t_{\text{eval}})$ the set of answers satisfying all content, quantity, and temporal requirements of the question under a fixed interpretation. For the ith generated response, y_i denotes the full response and a_i its task answer, excluding any accompanying explanation.

Reference-based grading can nevertheless reject a valid task answer. Let a_{ref} denote the supplied reference and J_{ref} the grading function, which returns one for acceptance and zero for rejection. Task-answer validity r_i^* and the grading decision $\widetilde{r}_i$ are thus distinct:

$$r_i^* = \mathbf{1}[a_i \in \mathcal{A}^*(q, t_{\text{eval}})], \qquad \widetilde{r}_i = J_{\text{ref}}(q, a_{\text{ref}}, y_i). \tag{1}$$

Here $\mathbf{1}[\cdot]$ equals one when its condition holds and zero otherwise. A task-answer false rejection occurs when $r_i^* = 1$ but $\widetilde{r}_i = 0$. Full-response grading can also reject errors in the accompanying explanation, so attribution must establish what caused the rejection.

Incomplete reference coverage occurs when the references represent only some valid solutions. A false rejection is attributed to this gap when the candidate is absent from the references and both answers are complete valid solutions under the same interpretation and temporal scope. Source checks and the grading record must also exclude incorrect references, interpretation differences, equivalent expressions, and response-level errors as the cause.

3.2 Study design

To locate potential false rejections, repeated search trajectories are collected from multiple policies on BrowseComp (Wei et al., 2025), BrowseComp-ZH (Zhou et al., 2025), GAIA text-only (Mialon et al., 2024), WebWalkerQA (Wu et al., 2025b), and xBench-DeepSearch (2505) (Chen et al., 2025a). Each policy generates up to three trajectories per question, or eight for xBench-DeepSearch. Normalized exact match first checks answers against the reference. A reference-based LLM judge then evaluates nonmatching answers, allowing equivalent expressions to be accepted.

To check validity beyond reference agreement, the rejected answers are then assessed by three independent verifiers. Each searches for external evidence and checks the question's requirements without seeing the reference or the other verifiers' judgments. A rejected answer is newly accepted when all three verifiers support it. A stronger model then checks these revisions, followed by sampled human review. Majority voting provides a comparison requiring support from at least two verifiers, including cases in which the third assessment remains inconclusive.

To establish which disagreements reflect grading errors, a separately selected source-review sample includes rejections supported by the Qwen and Gemini verifiers, reference-based acceptances, and retained rejections. A stronger model checks and consolidates the evidence for human review. Reviewers check both the candidate and the reference against the question, assessing task-answer and full-response validity separately. These assessments distinguish reference omissions from other errors. When distinct solutions are both valid, analysis of the question constraints identifies why neither is excluded and which additional condition would distinguish them.

Table 1 reports evaluated questions (Q) and percentages of all outputs initially accepted (Ref.) or rejected with an answer available for verification (Eligible). Accepted gives unanimous acceptances as a percentage of eligible rejections. Increase gives the change in acceptance rate in percentage points (pp). Rates aggregate outputs across policies and benchmarks, weighting each output equally. Outputs without an assessable answer remain in the score calculation but do not enter verification.

3.3 Evaluation results and error attribution

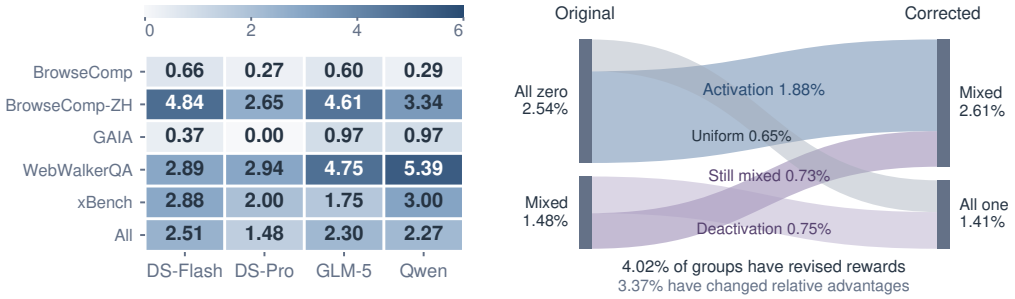
The verified revisions increase the acceptance rate by 2.12 pp, with changes across all five benchmarks. Table 1 shows larger changes on WebWalkerQA and BrowseComp-ZH than on BrowseComp and GAIA. The difference reflects more than the frequency of rejection: BrowseComp has the highest eligible rejection rate but the lowest acceptance rate among those rejections. Figure 2a shows that this benchmark contrast recurs across policies, while the size of the revision also varies by policy. Evaluation discrepancies therefore depend on both the task and the answers produced by the search agent, including the particular solutions it retrieves and presents.

Source attribution next examines why valid task answers are rejected. In the independently selected review sample, Figure 3a shows supported task answers in 71% of verifier-supported rejections. Full-response support is lower because explanations introduce additional claims. Attribution must therefore consider why the grader rejected the response, as well as whether the reference is valid and covers the task answer under a fixed interpretation of the question.

Figure 3b separates these causes: references are defective in 61% of reviewed rejections with supported task answers and supported in 31%. When both answers are supported, equivalent expressions indicate matching errors, while different interpretations can change the task. Distinct complete solutions establish incomplete reference coverage only under the same interpretation. This distinction matters for correction: repairing a reference or recognizing equivalent expressions cannot account for a distinct valid solution absent from a correct reference. Such omissions require examining why the question's constraints allow both solutions.

Table 1: **Score revisions after unanimous verification.** Results compare reference-based grading with verification of rejected answers across five benchmarks. Darker cells indicate larger values.

Benchmark	Q	Ref. (%)	Eligible (%)	Accepted (%)	Increase (pp)
BrowseComp	1,266	53.98	40.21	1.05	0.42
BrowseComp-ZH	289	63.26	27.25	14.18	3.86
GAIA text-only	103	79.82	14.51	4.02	0.58
WebWalkerQA	680	77.89	20.55	19.44	4.00
xBench-DS (2505)	100	83.31	12.97	18.55	2.41
Total / average	2,438	66.51	28.70	7.38	2.12



(a) Acceptance-rate increases (pp)

(b) Reward transitions

Figure 2: **Score and reward changes after unanimous verification.** (a) Acceptance-rate increases in pp, by benchmark and policy. DS denotes DeepSeek-V4, and Qwen denotes Qwen3.7-Plus. (b) Flows include only groups with revised rewards. Percentages use all recorded groups. Activation creates reward differences, deactivation removes them, and uniform zero-to-one changes leave every advantage at zero.

Examining these constraints reveals one-to-many relations as the most common structure among source-verified alternatives under a fixed interpretation. Clues identify an intermediate object but leave multiple qualifying answers connected to it. More detail about that object need not distinguish those answers. Other questions leave several objects satisfying all conditions, or request a subset without specifying how to select its members. In each case, additional clues establish uniqueness only if they distinguish the competing solutions. Question construction should therefore either distinguish competing solutions or ensure that the references cover them.

4 Reward Errors and Policy Updates

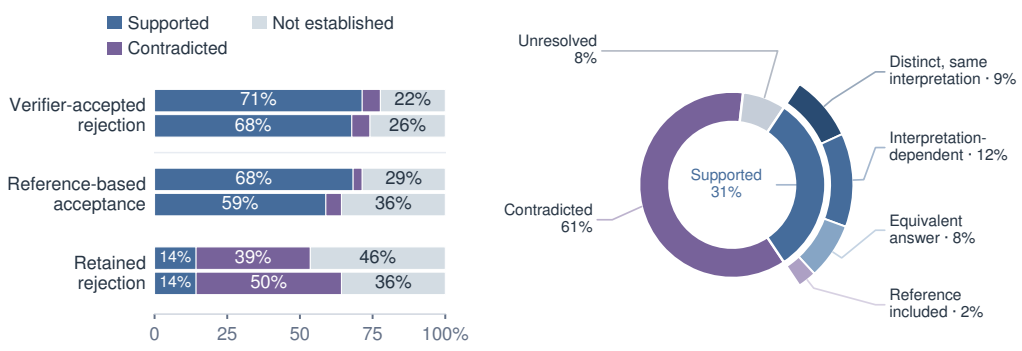
The grading errors identified in Section 3 also affect supervision when judgments reward search trajectories. Group-relative advantages connect these errors to parameter updates and, through those updates, to answer and search-action likelihoods.

4.1 Group-Relative Advantages

Group-relative learning weights each trajectory's contribution to the update by its advantage, computed from rewards for the same question (Shao et al., 2024). For trajectories grouped by question, policy, and decoding settings, rule $c \in \{\mathrm{ref}, \mathrm{corr}\}$ denotes reference-based or corrected rewards. Trajectory i in group j has advantage

$$A^c_{j,i} = \frac{r^c_{j,i} - \bar{r}^c_j}{\sigma^c_j + \epsilon_A}. \tag{2}$$

Here $r^c_{j,i} \in \{0, 1\}$ is the trajectory's reward, $\bar{r}^c_j$ and σ^c_j are the group's mean reward and reward standard deviation, and $\epsilon_A > 0$ prevents division by zero. When all rewards are equal, every advantage is zero. Otherwise, accepted trajectories have positive advantages and rejected trajectories have negative advantages relative to the other trajectories in the same group.



(a) Answer validity (b) Error attribution

Figure 3: **Answer validity and causes of rejection in the independently selected source-review sample.** (a) Task-answer validity (upper bars) and full-response validity (lower bars), with percentages within each review category. (b) Reference validity (inner ring) and answer relations for supported references (outer ring). Percentages in both rings refer to the full set of reviewed verifier-accepted rejections with supported task answers. Sample selection is described in Appendix D.2.

Verification changes advantages in about 3% of recorded groups. In Figure 2b, the most common such change is activation: verification creates reward differences in a previously uniform group. Other changes alter or remove existing differences. Changing all rewards from zero to one leaves advantages at zero. Among trajectories with changed advantages, 59% retain their rewards. This redistribution has an exact counterpart: when binary rewards change only from zero to one, corrected and unchanged trajectories account for equal total absolute advantage change. Their effects on answers and actions depend on the gradients weighted by those changes.

4.2 Reward-Weighted Policy Updates

Consider a source-verified false rejection. Correcting its reward from zero to one changes original rewards $r_i^{\text{ref}} = \widetilde{r}_i$ to r_i^{corr}, with all other rewards, initial parameters, and trajectories fixed.

Let π_θ be the search policy and θ_0 its initial parameters. In a group of G trajectories, $\mathcal{P}_i$ indexes model-generated tokens in trajectory i, and $L_i = |\mathcal{P}_i|$ counts them. Token x_{it} follows history h_{it}. The mean token log-likelihood ℓ_i and reward-dependent update direction g^c are

$$\ell_i(\theta) = \frac{1}{L_i} \sum_{t \in \mathcal{P}_i} \log \pi_\theta(x_{it} \mid h_{it}), \qquad g^c = \frac{1}{G} \sum_i A_i^c \nabla_\theta \ell_i(\theta_0). \tag{3}$$

Here A_i^c is the advantage under rule c, and ∇_θ differentiates with respect to model parameters θ. The fixed group index is omitted. Trajectories are weighted equally before applying advantages. Rewards, advantages, and recorded tokens remain fixed during differentiation.

Because the terminal advantage weights all model-generated tokens, a falsely rejected answer can give negative weight to reasoning and useful searches alike. Tool observations condition those predictions but are not generation targets. As the update changes shared parameters, its effects can also extend to answers and actions in other trajectories (Figure 4).

At the initial policy, g^c is the GRPO surrogate's reward-dependent gradient: importance ratios equal one and clipping is inactive (Shao et al., 2024; Schulman et al., 2017). Applying this direction without optimizer preconditioning or Kullback–Leibler (KL) regularization isolates the local effects of rewarding the same trajectories differently.

4.3 Answer and Action Likelihoods

For an answer or search with mean token log-likelihood $F(\theta)$ in a fixed context, a small shared step $\eta > 0$ gives the contrast with $\Delta g = g^{\text{corr}} - g^{\text{ref}}$:

$$F(\theta_0 + \eta g^{\text{corr}}) - F(\theta_0 + \eta g^{\text{ref}}) \approx \eta \nabla_\theta F(\theta_0)^\top \Delta g. \tag{4}$$

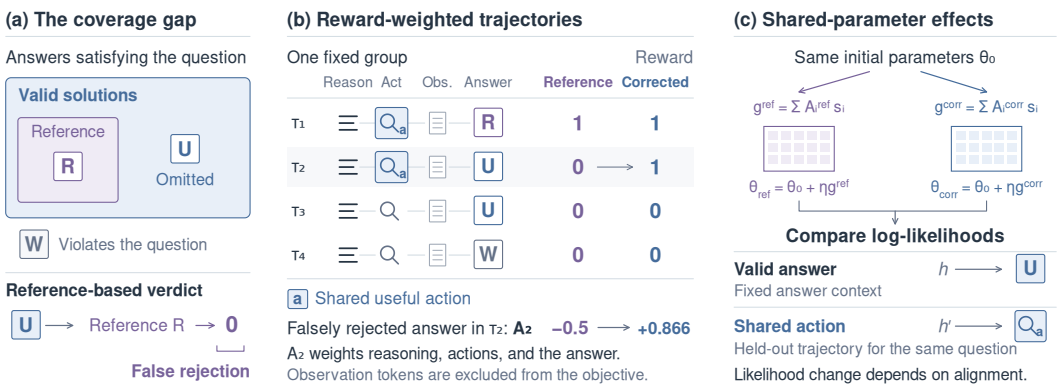


Figure 4: **Policy updates under original and corrected rewards.** (a) Reference R omits valid U. W is invalid. (b) Only the target reward is corrected. This changes all four advantages, including that of an accepted trajectory sharing action a. (c) Both updates share initial parameters and evaluation contexts. Here $s_i = G^{-1}\nabla_\theta \ell_i(\theta_0)$. Advantages use sample standard deviation with $\epsilon_A = 0$.

To resolve the contributions to this contrast, let $\mathcal{C}$ index corrected trajectories, $\Delta A_i = A_i^{\text{corr}} - A_i^{\text{ref}}$, and $s_i = G^{-1}\nabla \ell_i(\theta_0)$. When advantages change, define $M = \sum_{i\in\mathcal{C}} \Delta A_i > 0$. Then

$$\Delta g = M(\bar{s}_+ - \bar{s}_-), \qquad \bar{s}_+ = \sum_{i\in\mathcal{C}} \frac{\Delta A_i}{M} s_i, \qquad \bar{s}_- = \sum_{i\notin\mathcal{C}} \frac{-\Delta A_i}{M} s_i. \tag{5}$$

These averages use nonnegative weights summing to one. Correction favors the target to first order when corrected trajectories align more strongly with its likelihood gradient than unchanged trajectories. Labels determine the weights, while content and context determine the gradients. If advantages stay fixed, $\Delta g = 0$ and the two reward conditions induce the same parameter update.

A positive contrast measures improvement over original rewards. Comparing each update with $F(\theta_0)$ further distinguishes reduced suppression from an increase above the initial likelihood. Derivations appear in Appendix H.

5 Optimization Effects of Grading Errors

These distinctions guide comparisons under original and corrected rewards. The experiments test whether false rejections suppress valid answers and useful searches, and how these effects vary with response content, group composition and search history.

5.1 Experimental Design

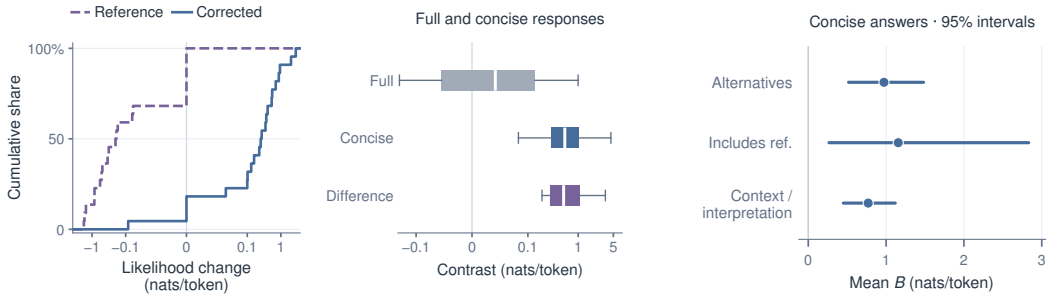
The experiments use groups of four AgentCPM (Chen et al., 2026) trajectories from WebWalkerQA and BrowseComp-ZH, each containing a false rejection verified through source checks before inspecting update effects. The analysis first considers distinct complete alternatives under the same interpretation, then extends to reference-containing and interpretation-dependent answers. Only the target's task-answer reward changes from 0 to 1, with the initial model and trajectories fixed.

Each condition applies one gradient step with shared loss weights and coefficient $\eta = \rho/\max(\|g^{\text{ref}}\|, \|g^{\text{corr}}\|)$, preserving relative gradient magnitudes. The displacement bound ρ equals 0.1 in the primary comparison. Additional settings appear in Appendix E.1.

For question q, let $F_q(\theta)$ be the target answer's mean token log-likelihood, averaged over two contexts for the same question from trajectories outside the update. With $\theta_c = \theta_0 + \eta g^c$ denoting the parameters after each update, define

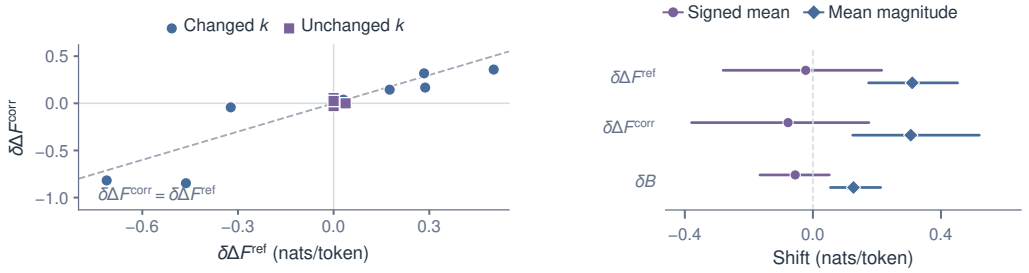
$$\Delta F_q^c = F_q(\theta_c) - F_q(\theta_0), \qquad B_q = \Delta F_q^{\text{corr}} - \Delta F_q^{\text{ref}}. \tag{6}$$

Thus $\Delta F_q^{\text{ref}} < 0$ indicates suppression under original rewards, while $B_q > 0$ indicates improvement from correction. Natural logarithms give changes in nats per token (nats/token). The 95% confidence intervals are obtained by resampling questions.



(a) Answer-likelihood changes (b) Full and concise responses (c) Answer attribution

Figure 5: **Answer suppression and its dependence on response content.** (a) Cumulative distributions of concise-answer likelihood changes ΔF. (b) Likelihood contrasts B for full and concise responses, and their within-question difference (Concise − Full). Boxes show quartiles and medians, whiskers the full range. Axes in (a–b) are logarithmic except near zero. (c) Mean concise-answer B by answer category, with 95% intervals from resampling questions. All panels use $\rho = 0.1$.



(a) Joint likelihood shifts (b) Signed changes and magnitudes

Figure 6: **Answer suppression varies across group compositions.** (a) Target-likelihood shifts after replacing another trajectory. The diagonal marks equal shifts under original and corrected rewards. The legend separates changes in the number of positive rewards k from replacements preserving k. (b) Mean signed and absolute shifts for replacements that change k. Bars show 95% intervals from resampling questions. All values use $\rho = 0.1$ and nats/token.

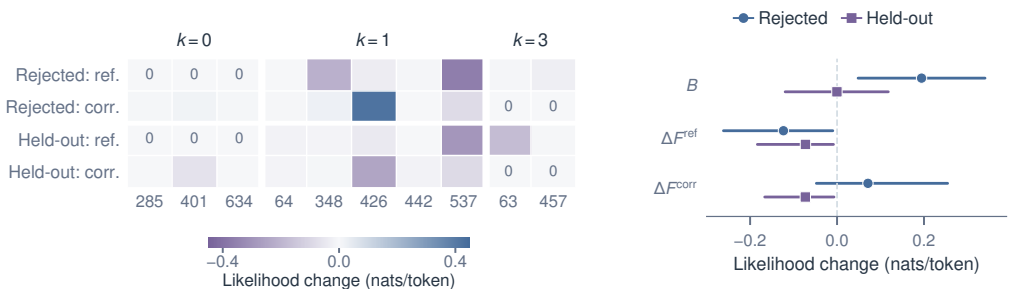
5.2 Answer Suppression and Response Content

The first comparison focuses on the valid task answer without its accompanying explanation. Earlier reasoning, tool observations, and other trajectories remain unchanged. The reference-based judge still rejects this concise answer. For distinct alternatives under the same interpretation, original rewards suppress the valid answer whenever the update is nonzero. Uniform-reward groups instead leave its likelihood unchanged. Correction yields higher likelihood in every case, with a mean contrast of 0.97 nats/token (Figure 5c). The broader grading-error comparison shows the same pattern, with a mean contrast of 0.92 nats/token. Figure 5a distinguishes these gains from recovery above the initial likelihood: some corrected updates still decrease likelihood, even though they improve on the original-reward update.

This consistent improvement concerns concise task answers. With full responses under the same task-answer label, correction can instead raise or lower target likelihood (Figure 5b). Most full responses contain incidental errors alongside the valid answer. Because shortening also changes token weighting and step normalization, the comparison measures their combined effect with response content. Task-answer validity alone therefore does not determine how rewarding the complete response affects the likelihood of the valid answer it contains.

5.3 Dependence on Group Composition

The answer-form comparison holds the group fixed. To examine group composition, one other trajectory is replaced while the target and its evaluation context remain fixed. Replacement changes



(a) Action likelihood across contexts

(b) Mean effects across contexts

Figure 7: **Effects of grading errors on matched search actions.** (a) Columns identify WebWalkerQA questions. Rows distinguish trajectories with falsely rejected answers (Rejected) from held-out valid-answer trajectories (Held-out). Ref./Corr. denote original/corrected rewards. k counts initially positive rewards. Violet indicates a likelihood decrease, blue an increase. (b) Means for $k = 1$, with 95% intervals from resampling questions. Both panels use $\rho = 0.1$.

trajectory content and may also change the number of positive rewards, jointly affecting advantages and gradients. All four updates share a step coefficient.

Replacing another trajectory changes nonzero suppression to 0.36–1.60 times its previous magnitude at $\rho = 0.1$, despite the target error staying fixed. Correction gives higher target likelihood in both group compositions on every evaluated question.

Figure 6a shows that original- and corrected-reward effects usually shift in the same direction after replacement. With δ denoting this shift, their common movement partly cancels in $\delta B = \delta\Delta F^{\mathrm{corr}} - \delta\Delta F^{\mathrm{ref}}$. The smaller mean absolute shift in B in Figure 6b shows why correction's relative benefit can remain similar even when suppression under original rewards changes substantially.

5.4 Effects on Useful Search Actions

Answer-level rewards also supervise earlier searches that provide useful evidence. To test effects within and beyond the updated trajectories, searches resolving the same question constraint are matched across a trajectory with a falsely rejected answer and a held-out counterpart with an accepted valid answer. Both final responses are reduced to concise task answers, so the comparison excludes errors in their accompanying explanations.

Query log-likelihood is evaluated given the preceding reasoning and retrieved information. Among five groups with nonzero updates under both reward conditions, original rewards suppress searches in all five rejected-answer trajectories and four held-out counterparts. Correction improves likelihood in four and three, respectively (Figure 7a). Its mean benefit is positive in rejected-answer trajectories but near zero in held-out contexts (Figure 7b). Thus correcting the same reward error need not consistently improve an equivalent search across contexts.

Sampling the next reasoning-and-action decision with matched random seeds shows no consistent change in search-intent selection. Together, these comparisons connect reward correction to the likelihood of valid continuations through trajectory weights, response content, and evaluation context.

6 Conclusion

Reference-based judgments connect search-agent evaluation with supervision. The benchmark analysis identifies incomplete reference coverage as one source of false rejection: multiple constraints can still permit distinct complete valid solutions. When these false rejections supply trajectory rewards, policy updates can reduce the likelihood of valid answers and useful searches, including in other trajectories. These effects vary with response content, group composition and search history, so answer validity alone does not determine the effect of rewarding a response. Reference coverage thus matters to both capability assessment and search supervision, motivating checks for competing solutions during question and reference construction.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, implementing methods, and assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results, and generating synthetic data sets is not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, creating or editing software code, and editing a research paper to improve readability. We have reviewed all AI-assisted work. We checked the mathematical arguments, verified and tested the implementation, and reviewed the edited figures and manuscript text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the reliability of reference-based evaluation and supervision in search agents. Its evaluations use public benchmarks and publicly accessible web evidence. Human review assesses answer validity and source support. The study may inherit biases from its source datasets, retrieved evidence, and model judgments. Public artifacts will follow the corresponding redistribution terms, with access credentials and reviewer-identifying information excluded.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3.2 and Appendices C–D specify benchmark evaluation, grading prompts, verification, and source attribution. The policy-update and action comparisons are described in Sections 5.1–5.4 and Appendices E–F. Appendices G and H document case evidence and mathematical derivations. The supplementary materials include evaluation and measurement code, frozen question-level measurements, documented case samples, and offline reproduction scripts. Original benchmark assets and model weights are obtained from their respective releases. The code and evaluation protocols will be publicly released.

REFERENCES

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. QAMPARI: A benchmark for open-domain questions with many answers. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz (eds.), *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 97–110, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.gem-1.9/.

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic Differentiation in Machine Learning: a Survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL http://jmlr.org/papers/v18/17-468.html.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 291–305, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.20. URL https://aclanthology.org/2022.emnlp-main.20/.

Haotian Chen, Xin Cong, Shengda Fan, Yuyang Fu, Ziqin Gong, Yaxi Lu, Yishan Li, Boye Niu, Chengjun Pan, Zijun Song, Huadong Wang, Yesai Wu, Yueying Wu, Zihao Xie, Yukun Yan, Zhong Zhang, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Agentcpm-explore: Realizing long-horizon deep exploration for edge-scale agents, 2026. URL https://arxiv.org/abs/2602.06485.

Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, Chen Sun, Han Hou, Hui Yang, James Pan, Jianan Lou, Jiayi Mao, Jizheng Liu, Jinpeng Li, Kangyi Liu, Kenkun Liu, Rui Wang, Run Li, Tong Niu, Wenlong Zhang, Wenqi Yan, Xuanzheng Wang, Yuchen Zhang, Yi-Hsin Hung, Yuan Jiang, Zexuan Liu, Zihan Yin, Zijian Ma, and Zhiwen Mo. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations, 2025a. URL https://arxiv.org/abs/2506.13651.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 85287–85307. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2858. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7b3a0d3a0864b66f229c0a9c84f9e950-Paper-Conference.pdf.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios, 2023. URL https://arxiv.org/abs/2307.13528.

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

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, 2025. URL https://arxiv.org/abs/2503.09516.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5591–5606, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.307. URL https://aclanthology.org/2023.acl-long.307/.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. Websailor: Navigating superhuman reasoning for web agent, 2025. URL https://arxiv.org/abs/2507.02592.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhu Chen, Pengyu Zhao, and Junxian He. Webexplorer: Explore and evolve for training long-horizon web agents, 2025. URL https://arxiv.org/abs/2509.06501.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. Noise-corrected grpo: From noisy rewards to unbiased gradients, 2026. URL https://arxiv.org/abs/2510.18924.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for General AI Assistants. In B. Kim,

Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25ae35b5b1738d80f1f03a8713e405ec-Paper-Conference.pdf.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

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

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in rlvr, 2026. URL https://arxiv.org/abs/2506.10947.

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, 2024. URL https://arxiv.org/abs/2402.03300.

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 reinforcement learning, 2025. URL https://arxiv.org/abs/2503.05592.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566/.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents, 2025. URL https://arxiv.org/abs/2504.12516.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 80756–80827. Curran Associates, Inc., 2024. doi: 10.52202/079017-2567. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/937ae0e83eb08d2cb8627fe1def8c751-Paper-Conference.pdf.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, Ding-Chu Zhang, Zekun Xi, Robert Tang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebDancer: Towards Autonomous Information Seeking Agency. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 120957–120985. Curran Associates, Inc., 2025a. doi: 10.52202/085713-4035. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/af043aee9cb137785d93195c6cf4cd96-Paper-Conference.pdf.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. WebWalker: Benchmarking LLMs in web traversal. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10290–10305, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.508. URL https://aclanthology.org/2025.acl-long.508/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023.

Fengji Zhang, Xinyao Niu, Chengyang Ying, Guancheng Lin, Zhongkai Hao, Zhou Fan, Chengen Huang, Jacky Keung, Bei Chen, and Junyang Lin. A^2Search: Ambiguity-Aware Question Answering with Reinforcement Learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 136394–136448, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/dcbdb995a239204228305ab34c5da096-Paper-Conference.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese, 2025. URL https://arxiv.org/abs/2504.19314.

A NOTATION

Table 2 collects the notation used in the main text and mathematical appendices. Group indices are omitted when a single group is fixed. Unless specified otherwise, gradients are taken with respect to model parameters at θ_0, vector norms are Euclidean, and log-likelihoods use nats/token.

Table 2: Notation for answer evaluation and policy-update analysis.

Symbol	Meaning
Questions, answers and grading	
q, Q	Question and number of evaluated questions.
t_{eval}	Time at which an answer must satisfy the question.
y_i, a_i, a_{ref}	Full generated response, its task answer, and supplied reference.
$\mathcal{A}^*(q, t_{\mathrm{eval}})$	Complete valid task answers under a fixed interpretation and temporal scope.
J_{ref}	Reference-based grader, returning one for acceptance and zero for rejection.
r_i^*, $\widetilde{r}_i$	Task-answer validity indicator and reference-based grading decision, respectively.
$\mathbf{1}[\cdot]$	Indicator equal to one when its condition holds and zero otherwise.
Trajectories, rewards and updates	
i, j, G, k	Trajectory index, group index, group size, and positive-reward count before correction.
$c \in \{\mathrm{ref}, \mathrm{corr}\}$	Reference-based or corrected reward condition. These labels do not denote reward values.
π_θ, θ_0, θ_c	Search policy, initial parameters, and parameters $\theta_c = \theta_0 + \eta g^c$ after an update.
t, x_{it}, h_{it}	Token position, recorded token, and its preceding history in trajectory i.
τ, $\mathcal{P}_i$, L_i	A trajectory, its policy-generated token positions, and their count $L_i = \lvert\mathcal{P}_i\rvert$. For a generic trajectory, the token set is $\mathcal{P}(\tau)$.
m_{it}	Token indicator: one for policy-generated tokens included in the objective and zero otherwise.
$r_{j,i}^c$, $\bar{r}_j^c$, σ_j^c	Terminal reward, group mean reward, and group reward standard deviation under condition c.
ϵ_A, ν	Positive stabilizer in advantage normalization and standard-deviation convention ($\nu = 0$: population, $\nu = 1$: sample).
σ_ν^c	Reward standard deviation under convention ν, with the group index omitted.
A_i^c, ΔA_i	Normalized advantage and its change $A_i^{\mathrm{corr}} - A_i^{\mathrm{ref}}$.
$\ell_i(\theta)$, b_i	Mean policy-token log-likelihood of trajectory i and its gradient $b_i = \nabla\ell_i(\theta_0)$.
ω_i, w_i	Fixed coefficient on each scored token of trajectory i, and total trajectory weight $w_i = \omega_i L_i$.
s_i	Weighted score gradient $w_i b_i$. Equal trajectory weighting gives $s_i = G^{-1} b_i$.
g^c, Δg	Reward-dependent update direction $\sum_i A_i^c s_i$ and difference $g^{\mathrm{corr}} - g^{\mathrm{ref}}$.
η, ρ	Shared gradient-step coefficient and upper bound on parameter displacement.
Answer and action likelihoods	
$F(\theta)$, $F_q(\theta)$	Mean token log-likelihood of an evaluated continuation. For answers, F_q averages the held-out contexts for question q.
$\mathcal{M}$, $F_{\mathcal{M}}(\theta)$	Token positions in a search argument and its mean token log-likelihood in a fixed context.
ΔF_q^c	Change from the initial model: $F_q(\theta_c) - F_q(\theta_0)$. The same definition applies to action likelihoods.
B_q	Correction contrast $\Delta F_q^{\mathrm{corr}} - \Delta F_q^{\mathrm{ref}}$. $B_q > 0$ does not imply $\Delta F_q^{\mathrm{corr}} > 0$.

Table 2: Notation for answer evaluation and policy-update analysis (continued).

Symbol	Meaning
U, R (subscripts)	Rejected target answer and valid supplied reference, used to distinguish their likelihood effects.
B_U^{concise}, B_U^{original}	Correction contrasts when the update uses the concise task answer or the original full response.
$\delta\Delta F^c$, δB	Changes after replacing another trajectory in the group. A prime denotes the replacement group, e.g., $\delta B = B' - B$.
Surrogate objective and differentiation	
$\mathcal{L}_c(\theta)$, $\varrho_{it}(\theta)$	Clipped group-relative surrogate and token probability ratio $\pi_\theta(x_{it} \mid h_{it})/\pi_{\theta_0}(x_{it} \mid h_{it})$.
$\varepsilon_{\text{clip}}$, β, $\mathcal{D}$	Clipping width, regularization coefficient, and reward-independent regularizer.
$\mathcal{V}$, $\mathbf{v}_{it}$, $\mathbf{p}_{it}$	Vocabulary, next-token logit vector, and its softmax probability vector. A subscript after the comma selects a vocabulary entry.
$\mathbf{e}_{x_{it}}$	One-hot vector for the recorded token x_{it}.
stopgrad	Holds its argument fixed during differentiation.
Expected policy gradients	
$P_\theta(\tau \mid q)$, $E(\tau \mid q)$	Trajectory probability and product of environment conditional probabilities, independent of θ.
$R(\tau)$, R_i, $\bar{R}$	Terminal reward function, sampled reward $R_i = R(\tau_i)$, and mean reward over the group.
$J_R(\theta)$, $b(q)$	Expected terminal reward and a baseline depending only on the question.
z_i	Trajectory score gradient $\nabla \log P_\theta(\tau_i \mid q)$.
Credit redistribution and local likelihood analysis	
p, γ, $d(p)$	Positive-reward fraction, factor $\sqrt{G/(G-\nu)}$, and normalization term $\gamma\sqrt{p(1-p)} + \epsilon_A$.
$A_+(p)$, $A_-(p)$	Advantage levels for rewards one and zero, respectively. Primes on these functions denote derivatives with respect to p.
$\mathcal{C}$, M	Indices of trajectories whose rewards change from zero to one and their total advantage change $M = \sum_{i\in\mathcal{C}} \Delta A_i = \|\Delta A\|_1/2$.
A^c, ΔA	Group advantage vector and difference $A^{\text{corr}} - A^{\text{ref}}$.
S, H	Matrix of weighted score gradients $[s_1, \ldots, s_G]$ and its Gram matrix $S^\top S$.
u, K_i	Evaluated continuation gradient $\nabla F(\theta_0)$ and alignment $u^\top s_i$.
α, $\mathbf{d}$	Scalar step and fixed parameter direction in the Taylor expansion.
N_g	Larger gradient norm $\max(\|g^{\text{ref}}\|, \|g^{\text{corr}}\|)$, giving $\eta = \rho/N_g$ when nonzero.
L_F, $\mathcal{R}_F$	Bound on the Hessian operator norm and remainder in the likelihood-contrast expansion.
$\bar{s}_+$, $\bar{s}_-$	Score-gradient averages over reward-corrected and reward-unchanged trajectories, weighted by positive and negative advantage changes.

B LIMITATIONS

Evaluation scope. Sampled outputs do not establish benchmark-wide omission prevalence. Verification can share model or source errors; the separate source review does not calibrate all score revisions. Question-structure analysis does not reconstruct benchmark authorship.

Local updates. Single steps on selected trajectories do not establish training gains. Full/concise comparisons jointly change content, token weighting and normalization; trajectory replacement can change content and reward.

Search behavior. Selected-query likelihoods do not cover all equivalent searches. Next-decision sampling shows no consistent selection advantage; generalization to other policies and sustained training remains untested.

C Evaluation Results and Protocol

C.1 Score revisions and affected trajectory groups

Table 3 separates policy-level score revisions from changes in trajectory-group rewards, showing where revised labels alter the learning signal. The main average uses unanimous three-verifier agreement for Qwen3.7-Plus, DeepSeek-V4-Pro, DeepSeek-V4-Flash-0731 and GLM-5. AgentCPM and WebExplorer-8B results remain supplementary and use their respective verification procedures.

Table 3: **Policy-level score revisions and reward changes.** Scores and shares use %, with acceptance-rate increases in pp as subscripts. † denotes matching to verifier-accepted answer classes. Accepted uses eligible rejected answers as its denominator; Affected and Activated use recorded trajectory groups. They count changed advantages and newly nonuniform rewards, respectively. Changes are computed before rounding.

	Score (%)		Rejections (%)	Groups (%)	
Policy	Ref.	Revised	Accepted	Affected	Activated
BrowseComp					
Qwen3.7-Plus	50.09	50.38$_{+0.29}$	0.62	0.63	0.47
DS-V4-Pro	62.01	62.28$_{+0.27}$	1.03	0.63	0.32
DS-V4-Flash-0731	72.01	72.67$_{+0.66}$	4.11	0.92	0.37
GLM-5	44.10	44.71$_{+0.60}$	1.11	1.34	1.03
AgentCPM	18.01	18.11$_{+0.11}$	0.13	0.24	0.24
WebExplorer-8B	12.50	12.67$_{+0.17}$	0.20	0.24	0.24
BrowseComp-ZH					
Qwen3.7-Plus	59.40	62.75$_{+3.34}$	13.24	5.19	2.77
DS-V4-Pro	68.51	71.16$_{+2.65}$	11.50	3.81	1.73
DS-V4-Flash-0731	72.43	77.28$_{+4.84}$	27.81	9.00	3.46
GLM-5	52.71	57.32$_{+4.61}$	10.67	8.30	5.88
AgentCPM	25.26	27.25$_{+1.99}$	2.81	4.15	2.08
WebExplorer-8B	27.85	30.71$_{+2.85}$	3.96	3.46	2.77
GAIA text-only					
Qwen3.7-Plus	75.08	76.05$_{+0.97}$	5.17	1.94	1.94
DS-V4-Pro	91.59	91.59$_{+0.00}$	0.00	0.00	0.00
DS-V4-Flash-0731	87.13	87.50$_{+0.37}$	4.17	0.97	0.97
GLM-5	66.34	67.31$_{+0.97}$	3.95	2.91	2.91
AgentCPM	57.52	58.25$_{+0.73}$	1.71	1.94	0.97
WebExplorer-8B	38.93	39.42$_{+0.49}$	0.80	1.94	0.97
WebWalkerQA					
Qwen3.7-Plus	73.58	78.97$_{+5.39}$	20.87	7.35	4.12
DS-V4-Pro	83.92	86.86$_{+2.94}$	20.91	4.41	2.65
DS-V4-Flash-0731	82.75	85.64$_{+2.89}$	18.79	3.68	2.06
GLM-5	71.32	76.08$_{+4.75}$	17.67	7.79	4.56
AgentCPM†	63.59	69.03$_{+5.44}$	14.98	7.35	2.65
WebExplorer-8B†	51.76	57.06$_{+5.29}$	10.98	—	—
xBench-DS (2505)					
Qwen3.7-Plus	84.00	87.00$_{+3.00}$	22.43	9.00	3.00
DS-V4-Pro	87.38	89.38$_{+2.00}$	22.22	4.00	0.00
DS-V4-Flash-0731	83.25	86.12$_{+2.88}$	25.27	8.00	3.00
GLM-5	78.62	80.38$_{+1.75}$	9.66	5.00	2.00
AgentCPM†	65.50	70.38$_{+4.88}$	14.13	12.00	1.00
WebExplorer-8B†	59.13	65.25$_{+6.13}$	14.98	22.00	5.00
Main evaluation average	**66.51**	**68.63$_{+2.12}$**	**7.38**	**3.37**	**1.88**

C.2 EVALUATION PROTOCOL

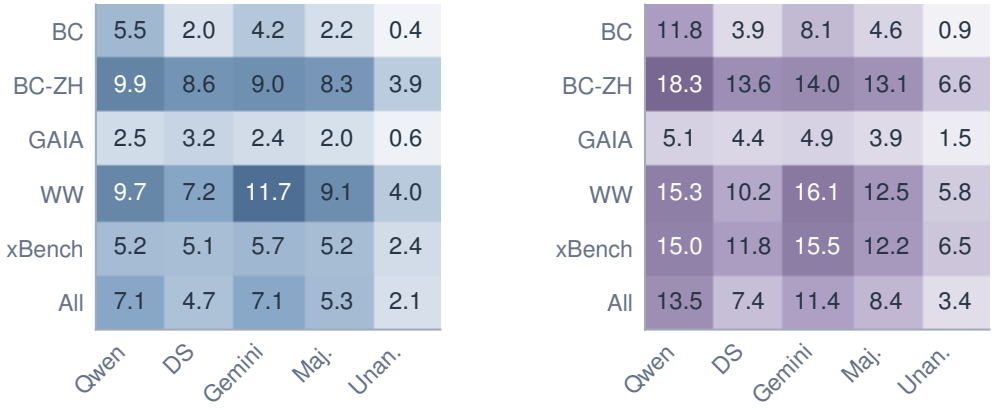
Table 4 specifies how responses are graded, what evidence verifiers receive, and which decisions enter the revised scores. The aggregate evaluation, source review and local-update comparisons use different selection rules; their percentages are not interchangeable.

Table 4: **Benchmark evaluation and verification protocol.** Source review and local-update eligibility are specified separately.

Component	Setting or decision rule
Benchmarks	BrowseComp, BrowseComp-ZH, GAIA text-only, WebWalkerQA and xBench-DeepSearch (2505).
Sampling	Up to three trajectories per question; eight for xBench-DeepSearch.
Exact match	Compare complete final responses after lowercasing and whitespace normalization; explanations remain included.
Reference judge	Qwen3.7-Plus (2026-05-26), thinking disabled; nonmatching responses use the benchmark-specific prompts below.
Verifiers	Independently run Qwen3.7-Flash, DeepSeek-V4-Flash-0731 and Gemini-3.6-Flash.
Verifier inputs	Question, complete response and temporal scope. Hide the reference, generating trajectory, policy identity and other verdicts.
Evidence rule	Retrieve sources with matched search budgets. Support requires the necessary conditions; contradiction establishes rejection; insufficient evidence remains inconclusive.
Primary revision	Accept a rejected response only if all three verifiers support it. Other labels remain unchanged.
Additional review	Stronger-model checking followed by sampled human review of accepted revisions.
Majority comparison	At least two support judgments, including when the third is inconclusive.
Denominators	Scores weight all recorded outputs equally. Acceptance among rejections uses only answers available for verification. Reward groups share question, policy and decoding settings.

C.3 Sensitivity to verifier agreement

Figure 8 examines how the agreement requirement changes both acceptance rates and the fraction of affected reward groups. The same outputs and matched per-verifier budgets yield acceptance-rate increases of 2.12 pp under unanimity and 5.30 pp under majority voting. The larger revision reflects a weaker agreement requirement, not independently established accuracy. Affected groups have changed advantages; uniform zero-to-one reward shifts leave advantages unchanged.



(a) Acceptance-rate increase (pp)

(b) Affected groups (%)

Figure 8: **Sensitivity to verifier agreement.** Columns compare Qwen3.7-Flash (Qwen), DeepSeek-V4-Flash-0731 (DS), Gemini-3.6-Flash (Gemini), majority voting (Maj.) and unanimity (Unan.). Acceptance-rate increases use pp. Affected groups are percentages of groups whose advantages change. BC/BC-ZH denote BrowseComp/BrowseComp-ZH, and WW denotes WebWalkerQA. The last row aggregates all benchmarks.

C.4 Reference-based evaluation prompts

The judge receives the question, reference and complete final response. GAIA/WebWalkerQA and BrowseComp templates are reproduced below with display-only formatting. BrowseComp-ZH uses the same criteria in Chinese, omitting the four additional examples on evasive answers, extra reference detail, inferable context and name transliteration; its duplicate rules are not repeated. The xBench box is an English translation, including its output labels.

GAIA and WebWalkerQA: answer equivalence

You are an evaluation assistant. Please determine if the predicted answer is equivalent to the labeled answer.

Question: {question}

Labeled Answer: {correct_answer}

Predicted Answer: {response}

Did the model give an answer equivalent to the labeled answer? Please respond with "Correct" if they are equivalent, or "Incorrect" if they are not equivalent. Do not include any other text.

BrowseComp: reference-based grading

Based on the given question, standard answer, and model-predicted answer, evaluate whether the model's response is correct. Your task is to classify the result as: [CORRECT] or [INCORRECT].

First, we'll list examples for each category, then you'll evaluate a new question's predicted answer.
Here are examples of [CORRECT] responses:
Question: What are the names of Barack Obama's children?
Standard Answer: Malia Obama and Sasha Obama
Model Prediction 1: Malia Obama and Sasha Obama

BrowseComp: reference-based grading (continued)

Model Prediction 2: Malia and Sasha
Model Prediction 3: Most would say Malia and Sasha, but I'm not sure, I should verify
Model Prediction 4: Barack Obama has two daughters, Malia Ann and Natasha Marian, commonly known as Malia Obama and Sasha Obama.
These responses are all [CORRECT] because they:
• Fully include the important information from the standard answer.
• Don't contain any information that contradicts the standard answer.
• Focus only on semantic content; language, capitalization, punctuation, grammar, and order aren't important.
• Vague statements or guesses are acceptable as long as they include the standard answer and don't contain incorrect information or contradictions.

Here are examples of [INCORRECT] responses:
Question: What are the names of Barack Obama's children?
Standard Answer: Malia Obama and Sasha Obama
Model Prediction 1: Malia
Model Prediction 2: Malia, Sasha and Susan or Sasha Obama or Malia Obama, or Natasha Marian, or Einstein
Model Prediction 3: While I don't know their exact names, I can tell you Barack Obama has two children.
Model Prediction 4: You might be thinking of Betsy and Olivia. But you should verify the details with the latest references. Is that the correct answer?
Model Prediction 5: Barack Obama's children
These responses are all [INCORRECT] because they:
• Contain factual statements that contradict the standard answer.
• Are empty or merely repeat the question.
• Enumerate multiple answers or repeat the answer.

Pay special attention to the following:
• The standard answer may contain responses to multiple aspects of the question, and within the same aspect, there might be different descriptions, all of which are correct and are given in the same bracket, connected by commas. For example, for the question "What is the name of ByteDance's AI model?", the standard answer is "[[Doubao, Skylark]]":
• Predicted answers "Doubao", "Doubao, Skylark", "Skylark", etc. are all [CORRECT].
• For standard answers containing responses to different aspects, the model needs to provide answers to all aspects to be considered correct; otherwise, it's directly judged as [INCORRECT]. There is no [PARTIALLY CORRECT] output option. These answers will be given in different brackets. For example, for the question "Who are the members of TFBOYS?", the standard answer is "[[Wang Junkai][Wang Yuan][Yi Yangqianxi]]":
• Predicted answers like "Wang Junkai, Wang Yuan, Yi Yangqianxi" that include all answers are [CORRECT].
• Predicted answers like "Wang Junkai, Yi Yangqianxi" that don't include all answers are [INCORRECT].

Also note the following points:
• For questions with numerical standard answers, the predicted answer should match the standard answer. For example, for the question "What is the total length in meters of the Huangpu River Bridge on the Jinshan Railway?", the standard answer is "3518.17":
• Predicted answers "3518", "3518.1", "3518.17" are all [CORRECT].
• Predicted answers "3520" and "3600" are [INCORRECT].
• If the model prediction doesn't directly answer the question, attempts to circumvent or fails to directly provide the standard answer, it's considered an [INCORRECT] answer.
• For example, for the question "Who is JJ Lin's wife?", with the standard answer "Ding Wenqi", model predictions like "JJ Lin's wife", "JJ Lin's wife should be excellent", "JJ Lin's wife might be a public figure" are all [INCORRECT].
• If the standard answer contains more information than the question asks for, the predicted answer only needs to include the information mentioned in the question.
• For example, for the question "What is the main chemical component of magnesite?", with the standard answer "Magnesium carbonate (MgCO3)", "Magnesium carbonate" or "MgCO3" are both considered [CORRECT] answers.
• If information omitted in the predicted answer can be clearly inferred from the question, it's considered correct.

BrowseComp: reference-based grading (continued)

- For example, for the question "The Nuragic ruins of Barumini were listed as a World Cultural Heritage by UNESCO in 1997, so where is this site located?", with the standard answer "Sardinia, Italy", the predicted answer "Sardinia" is considered [CORRECT].
- If it's clear that different translations of a name refer to the same person, it's considered correct.
- For example, if the standard answer is "Robinson", answers like "Lubinson" or "Lubinsun" are both correct.
- You should focus more on the match between the standard answer and the model prediction, rather than whether the standard answer itself is correct.

Below is a new question example. Please reply with only [CORRECT] or [INCORRECT], without apologies or corrections to your own errors, just evaluate the answer.
Question: {question}
Standard Answer: {correct_answer}
Predicted Answer: {response}

Evaluate this new question's predicted answer as one of the following:
A. [CORRECT]
B. [INCORRECT]

Return only the option representing [CORRECT] or [INCORRECT], i.e., just return A or B, without adding any other text.

xBench-DeepSearch: answer extraction and grading (English translation)

You are a general-purpose AI assistant. Given the [Correct Answer] below, judge whether the [Response] to the [Original Question] is correct.

[Original Question]: {question}
[Correct Answer]: {correct_answer}
[Response]: {response}

Your judgment must follow this format and these criteria:

Final Answer: Extract the precise final answer from [Response]. If there is no explicit final answer, write `None`.

Explanation: Explain why [Final Answer] is correct or incorrect according to [Correct Answer]. Focus only on substantive differences between the two. Do not discuss the question's background, solve it again, or defend an answer different from [Correct Answer]. Judge only whether they agree.

Conclusion: Write `Correct` if [Final Answer] agrees with [Correct Answer], or lies within an acceptable small tolerance for a numerical question. Otherwise, write `Incorrect`, including any inconsistency, ambiguity, nonequivalence or incorrect extracted answer.

D Error Attribution and Question Structure

D.1 Attribution criteria

Candidate validity is checked before comparison with the reference. Table 5 gives the evidence needed to distinguish strict reference omissions from other grading disagreements, assessing task-answer validity separately from claims in the full response.

Table 5: **Source-based attribution decisions.** All comparisons use the question's required completeness, temporal scope and fixed interpretation; inadequate evidence remains unresolved.

Decision	Required evidence
Answer validity	Check all required relations for each answer separately. Missing evidence is not contradiction; copied sources are not independent corroboration.
Reference omission	Both answers are valid, distinct and complete under one reading. Exclude rejection caused by explanation or format; a subset cannot answer an exhaustive-list request.
Other causes	Separate reference defects, equivalent expressions, reference-containing extensions and interpretation-dependent disagreements.

D.2 SOURCE-REVIEW SAMPLE

Questions are selected by benchmark and grading category (Qwen-and-Gemini-supported rejections, acceptances, retained rejections), retaining all responses and their frequencies. Qwen3.8-Max consolidates passages for human checks of both answers. Table 6 compares support for task answers and full responses across these categories, so support for a task answer is not treated as support for the entire response. This sample supports attribution, not calibration of three-verifier unanimity.

Table 6: **Source-review validity assessments (%).** Each row reports percentages within the indicated review category. Not established includes cases without sufficient evidence to assign validity. Values are descriptive response fractions, without sampling weights.

Review category	Assessed content	Supported	Contradicted	Not established
Verifier-accepted rejection	Task answer	71.4	6.3	22.3
	Full response	67.9	6.3	25.9
Reference-based acceptance	Task answer	68.3	3.1	28.6
	Full response	58.9	5.4	35.7
Retained rejection	Task answer	14.3	39.3	46.4
	Full response	14.3	50.0	35.7

D.3 QUESTION CONSTRAINTS AND ALTERNATIVE SOLUTIONS

Source and human checks compare generated answers, complete answers extracted from broader responses, and source-discovered alternatives (not policy outputs). Table 7 connects each source of nonuniqueness to a construction check, separating multiple solutions under fixed conditions from changes in context or interpretation. Each comparison receives one primary mechanism; this conditional collection does not estimate benchmark-wide prevalence.

Table 7: **Question structures and construction checks.** Shares describe the source-supported comparison collection and may not sum to 100% after rounding. Context and interpretation differences are separate from alternatives under fixed conditions.

Question structure	Share	Why the answer remains undetermined	Construction check
Jointly satisfied conditions	16%	Several target identities satisfy every stated condition.	Check the intersection of all conditions and distinguish or cover the remaining solutions.
One-to-many relations	33%	The clues identify an intermediate object with several qualifying relations or answers.	Distinguish the qualifying answers rather than only adding facts about the shared object.
Unspecified selection	7%	More objects qualify than the requested subset contains.	State a selection rule or accept qualifying subsets. Distinguish these from exhaustive-list answers.
Missing context	19%	The answer changes with an unstated session, track, round or year.	Supply the relevant context or index references by it.
Interpretive scope	26%	Different readings of a role, boundary or evaluative term change the requested relation.	Specify the intended role or convention before assessing coverage.

E Effects on Answer Likelihood

E.1 Comparison design

Table 8 defines the paired updates used to compare original and corrected rewards while holding the trajectory group fixed. Source checks precede inspection of update effects. The same concise task answer is measured after updates using either the full response or its concise version. A positive B compares corrected with original rewards; it does not imply recovery above the initial likelihood.

Table 8: **Design of the policy-update comparisons.** Within each answer form, only the target reward changes. Across forms, content, token weighting and step normalization change jointly.

Component	Setting
Policy and data	AgentCPM-Explore; selected WebWalkerQA and BrowseComp-ZH trajectories.
Eligibility	Source-supported target and reference; the concise target remains rejected by AgentCPM's original reference-based judge.
Update group	Four fixed trajectories; target reward changes from 0 to 1, companion rewards stay fixed.
Answer forms	Full response or concise task answer; preceding reasoning and observations remain unchanged. Task-answer support does not certify the full response.
Evaluation	Same concise target and reference; mean token log-likelihood over two held-out trajectories for the same question, ending immediately before the answer.
Loss	All parameters; policy-generated tokens only; equal trajectory weights. Tool observations supply context.
Advantage	Sample standard deviation with $\epsilon_A = 10^{-6}$; uniform rewards give zero update.
Step	$\eta = \rho / \max(\|g^{\mathrm{ref}}\|, \|g^{\mathrm{corr}}\|)$; $\rho \in \{0.01, 0.1, 1\}$, primary $\rho = 0.1$. Each form has its own coefficient.
Update rule	One gradient-ascent step from identical parameters; no adaptive preconditioning or KL penalty.
Aggregation	Average contexts within question; compute contrasts within question, then aggregate.

E.2 Control design

The controls examine dependence on companion trajectories, evaluation context and loss weighting. The companion comparison changes the group around a fixed target; Table 9 specifies the remaining controls, including a check of the local approximation across update scales. Replacement holds the target and held-out evaluation context fixed while changing one same-question companion, retaining its reward. Selection uses original labels and lengths before effects are inspected. All four directions share

$$\eta = \frac{\rho}{\max\{\|g^{\mathrm{ref}}\|, \|g^{\mathrm{corr}}\|, \|g^{\mathrm{ref}\prime}\|, \|g^{\mathrm{corr}\prime}\|\}}, \qquad \rho \in \{0.05, 0.1\}. \tag{7}$$

A prime denotes replacement; $\delta B = B' - B$. Changed-k comparisons jointly vary reward count and content. Unchanged-k comparisons cover uniform-reward boundaries, not groups remaining mixed in both conditions.

Table 9: **Auxiliary controls and interpretation.** All contrasts retain the same answers and paired reward conditions.

Control	Comparison	Scope
Context	Remove recorded reasoning and answer-stating messages; retain question, system instruction and tool observations, including answers within them.	Same updates and target answers; isolates evaluation-context dependence.
Weighting	Equal trajectory weights versus policy-token weights on the common question subset and both answer forms.	Same rewards and contexts; recompute the shared coefficient for each weighting rule.
Local linearity	Central finite-difference slopes at $\rho = 0.05$ and 0.1 under both reward rules and contexts; exclude zero directions.	Signs agree; relative discrepancies have median 6%, maximum 67%.

E.3 Attribution and scale

Attribution concerns the frozen target answer, not every alternative known for its question. Table 10 separates response forms and attribution categories. Table 11 then checks whether outcomes persist across update scales, retaining zero updates and distinguishing relative improvement from an increase above the initial likelihood. Leave-one-question-out mean B at the primary scale remains positive: 0.75–0.96 nats/token overall and 0.80–1.12 for same-interpretation alternatives.

Table 10: **Likelihood effects by response form and error attribution at $\rho = 0.1$.** Suppression ($\Delta F^{\mathrm{ref}} < 0$) and improvement ($B > 0$) are percentages of the questions in each row. Mean B uses nats/token. Full/concise describes the update response; both measure the same concise task answer. B is relative to the original-reward update. Attribution concerns the updated answer.

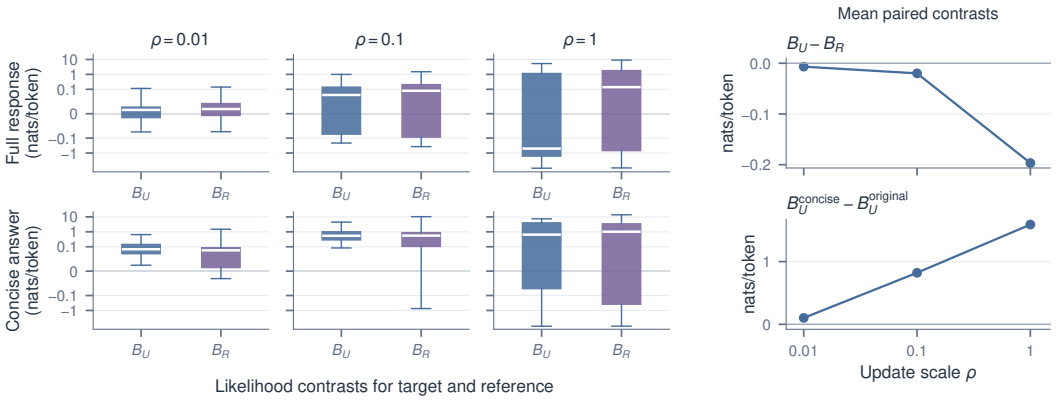
Condition	Suppression (%)	Improvement (%)	Mean B
Original full response	40.9	59.1	0.10
Concise task answer	68.2	100.0	0.92
Concise answers, by attribution			
Same-interpretation alternatives	42.9	100.0	0.97
Reference-containing extensions	60.0	100.0	1.16
Context / interpretation dependent	90.0	100.0	0.77

Table 11: **Concise-answer outcomes across update scales (%).** Each entry is the percentage of evaluated questions satisfying the condition. Improvement compares corrected with original rewards. A positive ΔF^{corr} indicates likelihood above its initial value.

Update scale ρ	Suppression $\Delta F^{\mathrm{ref}} < 0$	Improvement $B > 0$	Increase $\Delta F^{\mathrm{corr}} > 0$
0.01	68.2	100.0	81.8
0.1	68.2	100.0	81.8
1	68.2	72.7	27.3

E.4 TARGET AND REFERENCE CONTRASTS

A correction can change the likelihood of both the rejected target and the valid reference. Figure 9 compares these effects through $B_U - B_R$ and contrasts updates using full and concise responses. Figure 10 shows the individual scale profiles, making between-question and between-context variation visible alongside the aggregate comparisons.



(a) Likelihood contrasts across response forms

(b) Answer and form contrasts

Figure 9: **Effects of reward correction on the target and reference answers.** (a) B_U and B_R compare corrected with original rewards for the rejected target and valid reference. Boxes show medians and quartiles, with whiskers spanning the observed range. Symmetric logarithmic axes are linear within 0.03 of zero. (b) Mean within-question differences between answers (upper) and response forms (lower). Full and concise responses also differ in token weighting and step normalization.

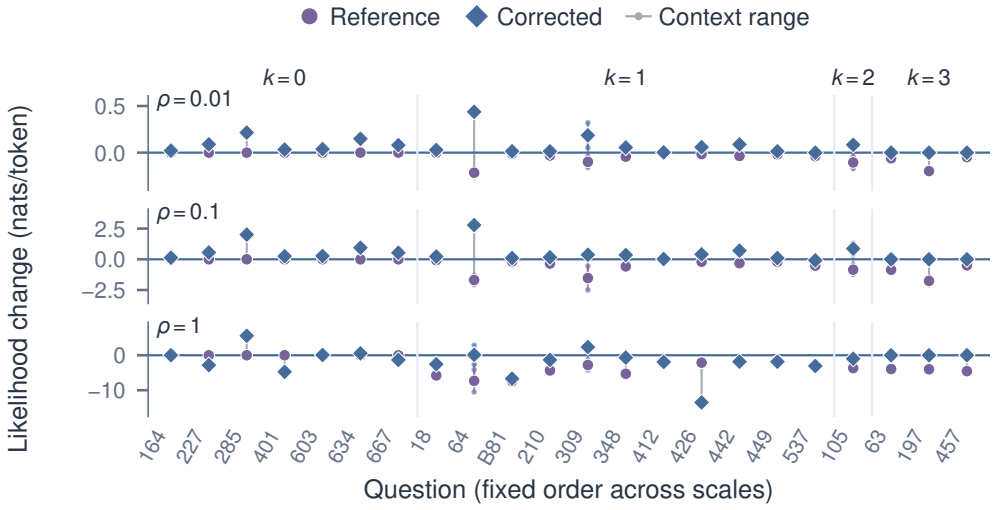


Figure 10: **Individual likelihood changes across update scales.** Each question keeps its horizontal position. Original- and corrected-reward changes are vertically aligned. Large markers average two contexts; small markers and segments show their range. Questions are ordered by initial positive-reward count k and ID. Vertical scales differ across rows. B81 denotes BrowseComp-ZH; other IDs denote WebWalkerQA.

E.5 CONTROL RESULTS

Figure 11 checks sensitivity to recorded reasoning and whether local slopes remain consistent across update scales. Table 12 compares the change in the correction contrast with the underlying likelihood shifts after companion replacement; changes shared by both reward conditions can cancel in the contrast. Table 13 compares loss-weighting rules on the same questions, so differences between rows do not arise from different question subsets.

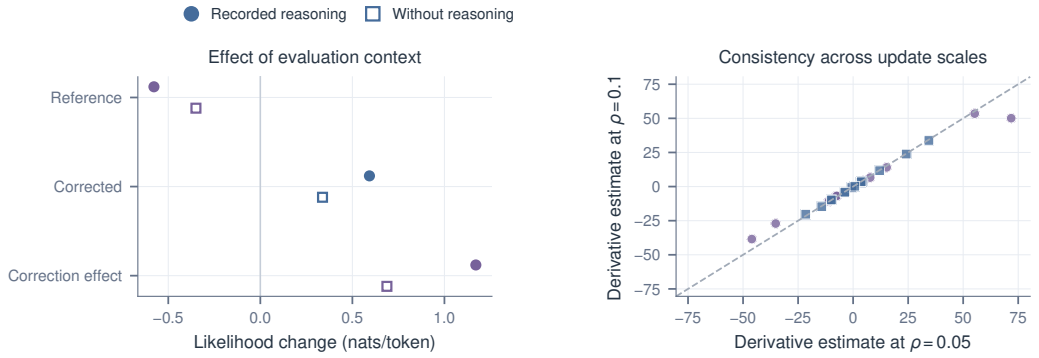


Figure 11: **Sensitivity to preceding context and update scale.** Left: mean answer-likelihood changes at $\rho = 0.1$, with recorded reasoning (filled circles) or without it (hollow squares). Right: central finite-difference slopes at two scales; the diagonal indicates equality. Slopes retain their signs but can differ in magnitude. Relative discrepancy is the absolute difference divided by the larger absolute slope.

Table 12: **Likelihood shifts after replacing another trajectory.** Entries are mean absolute within-question shifts in nats/token. All four updates share a step coefficient. Changed k and unchanged k indicate whether replacement alters the number of positive rewards. The unchanged-k comparisons include a uniform-reward condition.

Control	ρ	$\lvert\delta\Delta F^{\mathrm{ref}}\rvert$	$\lvert\delta\Delta F^{\mathrm{corr}}\rvert$	$\lvert\delta B\rvert$
Changed k	0.05	0.192	0.227	0.094
Changed k	0.1	0.310	0.306	0.127
Unchanged k	0.05	0.005	0.013	0.018
Unchanged k	0.1	0.009	0.017	0.026

Table 13: **Answer-likelihood changes under two loss-weighting rules.** Every row uses the same question subset at $\rho = 0.1$. The likelihood contrast $B = \Delta F^{\mathrm{corr}} - \Delta F^{\mathrm{ref}}$ uses nats/token. Ranges give observed minima and maxima. The final column is the percentage of questions with $B > 0$.

Answer form	Weighting	Mean	Median	Range	Improved (%)
Full	Trajectory	0.19	0.09	[-0.15, 1.00]	70
Full	Token	0.18	0.13	[-0.19, 0.79]	70
Concise	Trajectory	1.25	0.93	[0.08, 4.47]	100
Concise	Token	0.90	0.44	[0.09, 3.17]	100

F Effects on Search Actions

F.1 Protocol

Matched searches resolve the same information need in an updated trajectory and a held-out trajectory for the same question. Evidence checks precede inspection of update effects; acceptance or repeated wording alone does not establish usefulness. Table 14 defines useful searches, intent matching and the contexts used to assess whether a reward correction affects the matched queries.

Table 14: **Search-action comparisons.** The unit of inference is the question, not repeated searches.

Component	Definition
Useful search	Returned evidence resolves a required entity, relation or temporal constraint. Batched results must support the specific query.
Intent match	Equivalent information need under the same scope, not merely lexical similarity; refinements and different targets are separate.
Answer checks	Both target and held-out task answers require source support. Stronger-model review is followed by direct passage checks.
Update target	Reduce only the target's final response to its supported task answer; preceding decisions and observations stay fixed.
Held-out context	Excluded from the update. Its later final answer is checked for validity but is not part of the earlier query context.
Likelihood	Score query tokens given recorded reasoning and previous observations; exclude later observations and unscored prompt/tool tokens.
Aggregation	Average repeated occurrences within intent, intents within question, then questions with equal weight.

For query-token positions $\mathcal{M}$ and recorded histories h_t, the score is

$$F_{\mathcal{M}}(\theta) = |\mathcal{M}|^{-1} \sum_{t \in \mathcal{M}} \log \pi_\theta(x_t \mid h_t). \tag{8}$$

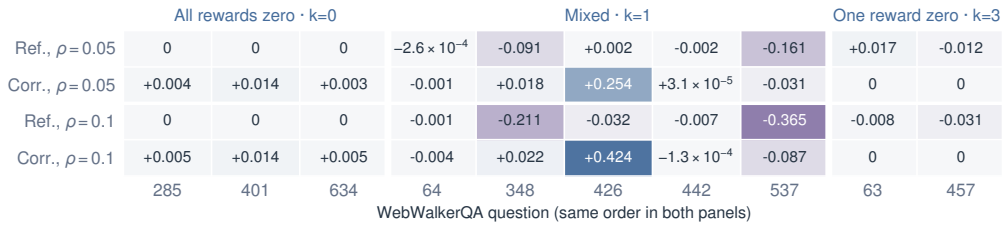
It measures the specified string, not the summed probability of equivalent searches.

F.2 Likelihood across contexts and scales

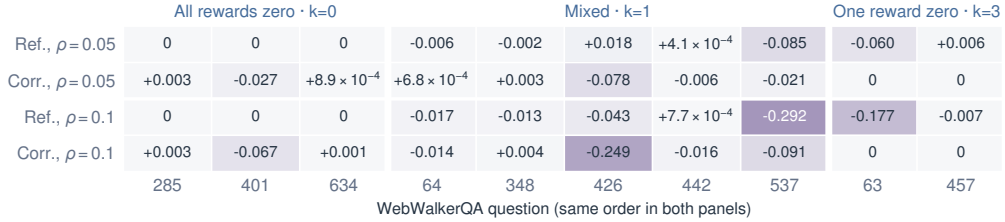
Figure 12 compares query-likelihood changes in updated and held-out trajectories across step sizes. Table 15 summarizes suppression, improvement and reversal at the primary scale, distinguishing relative gains from recovery above the initial likelihood. The groups include multiple answer-attribution categories; only one of the five $k = 1$ questions is a strict same-interpretation alternative. Both displays retain the $k = 0$ and $k = 3$ boundaries.

Table 15: **Matched search-action likelihood changes at $\rho = 0.1$.** Suppressed ($\Delta F^{\mathrm{ref}} < 0$), improved ($B > 0$) and reversal ($\Delta F^{\mathrm{ref}} < 0$, $\Delta F^{\mathrm{corr}} > 0$) are percentages of questions in each group. k counts initially positive rewards. Mean B uses nats/token and evaluates the search-query text.

Group	Context	Suppressed	Improved	Reversal	Mean B
$k = 0$	Rejected-answer trajectory	0.0	100.0	0.0	+0.008
$k = 0$	Held-out	0.0	66.7	0.0	-0.021
$k = 1$	Rejected-answer trajectory	100.0	80.0	40.0	+0.194
$k = 1$	Held-out	80.0	60.0	20.0	0.000
$k = 3$	Rejected-answer trajectory	100.0	100.0	0.0	+0.019
$k = 3$	Held-out	100.0	100.0	0.0	+0.092



(a) Search queries in trajectories with falsely rejected answers



(b) Matched search queries in held-out valid-answer trajectories

Figure 12: **Search likelihood changes within and beyond the updated trajectory.** Each column is one question, aligned across panels. Cells show the mean token log-likelihood change from initial parameters, in nats/token. Violet indicates a decrease and blue an increase, using the same scale in both panels. Ref./Corr. denote reference-based/corrected rewards, k counts initially positive rewards, and ρ bounds parameter displacement.

F.3 Next-action selection

A separate behavioral check asks whether corrected rewards change next-action selection. Unlike fixed-reasoning likelihood scoring, sampling regenerates reasoning before selecting an action. Temperature and top-p are one, seeds are matched, and generation stops before execution. Intent review hides condition and seed labels. Table 16 shows no consistent selection advantage; ambiguity places the corrected-minus-original difference between 0 and 3.1 pp.

Table 16: **Selection of a shared search intent (%).** Each cell gives the percentage of generated decisions selecting the matched search intent at $\rho = 0.1$. The range reflects an ambiguous intent match.

WebWalkerQA question	Initial	Reference	Corrected
64	87.5	100.0	100.0
348	100.0	100.0	100.0
442	100.0	100.0	100.0
537	25.0	12.5–25.0	25.0

G CASE STUDIES

Boxes contain questions, answers and source summaries; interpretation follows outside. Translations, abridgments and constructed comparisons are labeled.

G.1 EXAMPLES OF INCOMPLETE REFERENCE COVERAGE

Game editions (WebWalkerQA 210)

Original question. Which version of Sid Meier's Civilization VII should a user purchase to start playing on February 6, 2025, and access a civilization with the 'Nepekifaki' ability?

Reference. Founders Edition. **Constructed grading control.** Deluxe Edition.

Source evidence (summarized).

Requirement	Founders	Deluxe
February 6 early access	Yes	Yes
Tecumseh and Shawnee Pack	Included	Included

The publisher's "Sid Meier's Civilization VII Gameplay Revealed: Pre-Order Now", edition-contents paragraph, establishes both entitlements. The official "Shawnee" guide, "Unique Ability" entry, identifies Nepekifaki.

Grading observation. Fixed Deluxe is rejected against Founders but accepted against Deluxe or the combined reference in either order.

Both editions satisfy the conjunction. The fixed Deluxe response is a constructed grading control; a generated answer naming both editions instead contains the reference.

Film identities (BrowseComp 1212)

Original question. Provide the name of the movie that fits the following criteria: • This movie was released in the 1940's. • The movie's director also took on another key responsibility in this project. • One of the actors had a brother who was an author. • Another actor was born in London, England in the 1910s. • The movie had an additional English title besides its original title.

Reference. Murder at the Windmill.
Generated answer. A Matter of Life and Death (also known as Stairway to Heaven)
Reference-based judgment. Rejected.

Source evidence (summarized).

Constraint	Reference	Alternative
Film	Murder at the Windmill	A Matter of Life and Death
Release	1949	1946
Director's other role	Val Guest: writer	Powell/Pressburger: writers
Author brother	G. Marsh / D. G. Gerahty	R. / D. Attenborough
London-born, 1910s	Jon Pertwee, 1919	David Niven, 1910
Other title	Mystery at the Burlesque	Stairway to Heaven

Source locators (summarized). *Reference:* BFI's *10 great films set in Soho*, film entry (release, credits, cast, alternate title); *Digby George Gerahty* biography (authorship, brother); *The Independent*'s Jon Pertwee obituary (birth). *Alternative:* MoMA's 1946 screening record (release, title, credits); BBFC record (cast); National Portrait Gallery biographies (Attenborough brothers, Niven's birth); David Attenborough's *A Life on Our Planet* (2020), title and copyright pages (authorship).

Both films meet every condition. In each, the actor with an author brother differs from the actor satisfying the birthplace condition; neither comparison changes the reading of the question.

G.2 Reference omissions in the update comparisons

These are the concise answers actually used in the updates. Both the original full responses and concise versions received reference-based rejection. Task-answer support does not certify the removed explanations. The first four records are English translations; the remaining questions were already English.

Career honors (WebWalkerQA 164) — translated

Question. List three honorary titles received by Pan Gang, all related to his professional achievements.
Reference. China Youth May Fourth Medal; National May 1 Labour Medal; CCTV Economic Person of the Year.
Concise experimental answer. National Outstanding Entrepreneur; China Youth May Fourth Medal; Yuan Baohua Enterprise Management Gold Award.
Source evidence (summarized). Yili, "Chairman biography", opening biographical paragraph, lists the reference honors and the Yuan Baohua award. Xinhua, "Pan Gang awarded National Outstanding Entrepreneur" (23 September 2022, title translated), opening report paragraph, identifies the 2021–2022 award and its recipients. Both sources describe career achievements of the same person.

The person, achievement type and number of honors are fixed, but no condition selects a unique triple.

Conference roles (WebWalkerQA 227) — translated

Question. Who was both a program committee chair and a keynote speaker at the ninth International Conference on Information Science, Computer Technology and Transportation, hosted by the School of Computer Science and Technology of Southwest University of Science and Technology?
Reference. Xiaohua Xu.
Concise experimental answer. Zhenhua Li, Tsinghua University.
Source evidence (summarized). ISCTT 2024 Proceedings, "Preface", "Committee Member" section, lists Zhenhua Li and Xiaohua Xu as Technical Program Committee Chairs. The archived ISCTT 2024 "Keynote Speaker" page (March 2024) lists both speakers. The AIS Academy announcement for the same ninth conference identifies the host school and lists both under keynote guests.

Both people hold both roles at the same conference; the host and edition do not distinguish them.

Talks on the same date (WebWalkerQA 285) — translated

Question. Who gave the academic presentation at the School of Computer Science of Nanjing University of Posts and Telecommunications on May 8, 2024?
Reference. Jiankang Deng.
Concise experimental answer. Jun Cai.
Source evidence (summarized). The school issued separate notices titled "May 8, 2024 academic presentation: Jun Cai" and "May 8, 2024 academic presentation: Jiankang Deng" (titles translated). Both notices were published on 6 May 2024. Their titles establish the same requested event date, distinct from the publication date, and identify different speakers at the same school.

Both talks share the institution and date; no title or time selects one speaker.

Selecting qualifying papers (WebWalkerQA 634) — translated

Question. Who are the respective first authors of two transportation papers published in Transportation Research Part C under the supervision of Professor Zhigang Xu at the School of Information Engineering, Chang'an University?
Reference. Jinlong Li and Yuqin Zhang.
Concise experimental answer. Chenglin Liu and Yuqin Zhang.
Source evidence (summarized). Chang'an University reports identify Chenglin Liu as first author of "A multistep cooperative lane change strategy for connected and autonomous vehicle platoons departing from dedicated lanes" (university mirror dated 18 June 2024), Yuqin Zhang as first author of "A local traffic characteristic based dynamic gains tuning algorithm for cooperative adaptive cruise control considering wireless communication delay", and Jinlong Li as first author of "Domain adaptation from daytime to nighttime: A situation-sensitive vehicle detection and traffic flow parameter estimation framework" (team report dated 26 December 2022). The reports explicitly identify Zhigang Xu as supervisor and Transportation Research Part C as the journal.

Each pair supplies two qualifying papers. The question specifies neither titles nor a year and does not request an exhaustive list.

Guest countries of honor (WebWalkerQA 309)

Question. Which country is the guest of honor at the 2024 China International Import Expo?
Reference. Uzbekistan.
Concise experimental answer. France.
Source evidence (summarized). Shanghai Municipal Government, "7th CIIE to see record participation" (28 October 2024), the paragraph on the Comprehensive Country Exhibition, names both France and Uzbekistan as guest countries of honor. The event is the 2024 China International Import Expo specified in the question.

Both countries have the requested role at the same event; no ranking selects one.

Revenue-contributing games (WebWalkerQA 537)

Question. Which game contributed significantly to Rovio's revenue in both 2019 and 2022?
Reference. Angry Birds Dream Blast.
Concise experimental answer. Angry Birds 2.
Source evidence (summarized). Rovio's "Games revenue grew 5.7% in 2019; Angry Birds Dream Blast continued to grow in Q4", the CEO's review and game-performance discussion, identifies Angry Birds 2 as its largest game and Dream Blast as a substantial growth contributor. "Financial Statement Bulletin 2022", the CEO review, identifies both among the top live games achieving record annual revenue. The two reports establish contributions in each requested year.

Both games satisfy the two-year condition; the question does not require the largest contributor.

Games linked by event conditions (WebWalkerQA 63)

Question. Which two Angry Birds games released by Rovio incorporate innovative and environmental themes, one being a mobile AR game from 2019 and the other related to an Earth Day event in 2024?
Reference. Angry Birds AR: Isle of Pigs and Angry Birds Dream Blast.
Concise experimental answer. Angry Birds AR: Isle of Pigs and Angry Birds 2.
Source evidence (summarized). Rovio's announcement "Rovio and Resolution Games reveal first Angry Birds mobile AR game, Angry Birds AR: Isle of Pigs" (19 March 2019) identifies the mobile AR title and its 2019 release schedule. "Birds Helping Birds: Angry Birds Dream Blast Earth Day Campaign" (22 April 2024), opening event paragraph, explicitly lists both Angry Birds Dream Blast and Angry Birds 2 as participants. EARTHDAY.ORG's 17 April 2024 partnership release independently names Angry Birds 2.

The AR game is fixed, but either second game satisfies the Earth Day condition.

G.3 An unresolved relation in a multi-step question

A one-to-many relation can leave several complete answers even after an early entity in the reasoning chain has been identified.

Case material: film and television adaptations

Question (abridged). Give the number of story episodes, excluding recut episodes, in a television adaptation sharing the Chinese title and source novel of a film whose song was performed by a recurring *The Voice of China* mentor.

Source evidence (summarized).

Condition	Reference solution	Alternative solution
Recurring mentor	Na Ying	Na Ying
Song used in a film	*Mo*	*San Sheng San Shi Shi Li Tao Hua*
Film	*He Yi Sheng Xiao Mo*	*San Sheng San Shi Shi Li Tao Hua*
Same-title adaptation of the same novel	*He Yi Sheng Xiao Mo*, Gu Man novel	*San Sheng San Shi Shi Li Tao Hua*, Tang Qi novel
Story episodes	32 (four recuts excluded)	58

Na Ying's interview establishes her recurring mentor role. The adaptation report, film theme-song credit, and LeTV and Youku episode listings support the two song–adaptation chains.

Both chains meet the stated relations but yield different episode counts. Fixing the mentor leaves the song and adaptation unspecified; a condition on either would distinguish the solutions. This establishes an alternative without measuring whether a judge rejects it.

H Analysis of Reward-Dependent Policy Updates

The derivations specify the fixed-group surrogate, binary credit redistribution and local likelihood contrast. Histories and rewards remain fixed throughout.

H.1 The fixed-group update

For a fixed group, let r_i^c denote the terminal reward under rule c and $L_i > 0$ the number of scored policy tokens. Define

$$\bar{r}^c = \frac{1}{G}\sum_i r_i^c, \qquad (\sigma_\nu^c)^2 = \frac{1}{G-\nu}\sum_i (r_i^c - \bar{r}^c)^2, \qquad A_i^c = \frac{r_i^c - \bar{r}^c}{\sigma_\nu^c + \epsilon_A}, \tag{9}$$

where $\nu = 0$ or 1 selects population or sample standard deviation and $\epsilon_A > 0$. Constant groups receive zero advantages. Sample standard deviation requires $G \geq 2$. The benchmark reward analysis uses $(\nu, \epsilon_A) = (0, 10^{-8})$. Parameter measurements use $(1, 10^{-6})$. These conventions must not be interchanged when comparing magnitudes.

Proximal policy optimization (PPO)'s clipping objective and GRPO's group-relative terminal advantages motivate the following fixed-group surrogate (Schulman et al., 2017; Shao et al., 2024). Let θ_0 denote the fixed policy parameters in the ratio denominator and let ω_i be the fixed trajectory coefficient. Write

$$\varrho_{it}(\theta) = \frac{\pi_\theta(x_{it} \mid h_{it})}{\pi_{\theta_0}(x_{it} \mid h_{it})}, \tag{10}$$

$$\mathcal{L}_c(\theta) = \sum_i \omega_i \sum_{t \in \mathcal{P}_i} \min\{\varrho_{it}(\theta)A_i^c, \mathrm{clip}(\varrho_{it}(\theta), 1-\varepsilon_{\mathrm{clip}}, 1+\varepsilon_{\mathrm{clip}})A_i^c\} - \beta\mathcal{D}(\theta). \tag{11}$$

Here ϱ_{it} is the token probability ratio, $\varepsilon_{\mathrm{clip}} > 0$ is the clipping width, and $\beta \geq 0$ weights the reward-independent regularizer $\mathcal{D}$. The clip operator restricts the ratio to the indicated interval. A fixed policy used for KL regularization is distinct from the answer reference a_{ref}. All sampled histories, rewards, advantages, masks, coefficients, and ratio denominators are held fixed during differentiation.

At $\theta = \theta_0$, $\varrho_{it} = 1$ lies inside the clipping interval when $\varepsilon_{\mathrm{clip}} > 0$. Both branches consequently have the same local derivative, including zero when $A_i^c = 0$. Since $\nabla\varrho_{it}|_{\theta_0} = \nabla \log \pi_{\theta_0}(x_{it} \mid h_{it})$,

$$\nabla\mathcal{L}_c|_{\theta_0} = \sum_i A_i^c s_i - \beta\nabla\mathcal{D}(\theta_0), \qquad s_i = \omega_i \sum_{t \in \mathcal{P}_i} \nabla \log \pi_{\theta_0}(x_{it} \mid h_{it}). \tag{12}$$

The reported $g^c = \sum_i A_i^c s_i$ isolates the reward-dependent direction ($\beta = 0$). An identical additive regularizer cancels in a gradient difference on identical histories, but still changes each individual direction. That cancellation does not cover KL inserted into rewards before normalization. Away from θ_0, importance weights and active clipping can change the coefficients, so fixed s_i vectors need not describe gradients at later parameter values.

The local surrogate is exact for fixed histories from any sampling distribution. An on-policy interpretation additionally requires matching that distribution. Token probabilities use temperature one.

Token selection, weighting and differentiation. Let $\mathbf{v}_{it}(\theta)$ be the next-token logit vector, $\mathbf{p}_{it}$ its softmax, and $\mathbf{e}_{x_{it}}$ the one-hot vector for the observed token. Here $\mathcal{V}$ is the vocabulary, and subscripts after the comma identify its entries. Differentiating $\log p_{it,x_{it}} = v_{it,x_{it}} - \log\sum_{x\in\mathcal{V}} \exp v_{it,x}$ gives

$$\nabla_\theta \log \pi_\theta(x_{it} \mid h_{it}) = \left(\frac{\partial \mathbf{v}_{it}}{\partial\theta}\right)^\top (\mathbf{e}_{x_{it}} - \mathbf{p}_{it}). \tag{13}$$

The chain rule differentiates model representations, holding token identities, retrieved text and reward decisions fixed (Baydin et al., 2018).

Let $m_{it} = 1$ for policy-generated tokens and zero for unscored tokens. Put $L_i = \sum_t m_{it}$ and define the mean-token score gradient

$$b_i = \nabla_\theta \left[\frac{1}{L_i}\sum_t m_{it} \log \pi_\theta(x_{it} \mid h_{it})\right]\Bigg|_{\theta_0}. \tag{14}$$

The two token-weighting rules give

$$\begin{aligned} &\text{Policy-token mean:} && \omega_i = \frac{1}{\sum_{m=1}^{G} L_m}, && g^c = \sum_i \frac{L_i}{\sum_{m=1}^{G} L_m} A_i^c b_i, \\ &\text{Equal trajectory weight:} && \omega_i = \frac{1}{GL_i}, && g^c = \frac{1}{G}\sum_i A_i^c b_i. \end{aligned} \tag{15}$$

Thus $s_i = w_i b_i$, where $w_i = \omega_i L_i$ is the trajectory weight, equal to $L_i / \sum_{m=1}^{G} L_m$ or $1/G$. For unequal lengths they define different objectives: token normalization is not a general proof of an unbiased terminal-task policy gradient.

Equivalently, differentiate $\sum_i \omega_i \,\mathrm{stopgrad}(A_i^c) \sum_t m_{it} \log \pi_\theta(x_{it} \mid h_{it})$. The stop-gradient operator fixes advantages; minimizing the negative objective produces the ascent step $+\eta g^c$.

Relation to an expected policy gradient. Fix a question q. A trajectory τ interleaves policy-generated tokens x_t with tool observations. Let $\mathcal{P}(\tau)$ index policy tokens and h_t contain the complete preceding history. If the environment has no direct parameter dependence, its conditional factors can be collected into $E(\tau \mid q)$, giving

$$P_\theta(\tau \mid q) = E(\tau \mid q) \prod_{t \in \mathcal{P}(\tau)} \pi_\theta(x_t \mid h_t), \qquad \nabla_\theta \log P_\theta(\tau \mid q) = \sum_{t \in \mathcal{P}(\tau)} \nabla_\theta \log \pi_\theta(x_t \mid h_t). \tag{16}$$

Tool outputs condition later predictions but contribute no policy-token loss.

For a fixed terminal reward function $R(\tau)$, define $J_R(\theta) = \mathbb{E}_{P_\theta}[R(\tau)]$. Under differentiability, common support, and an interchange of differentiation and expectation, the likelihood-ratio identity yields the policy gradient (Sutton et al., 1999):

$$\nabla J_R(\theta) = \mathbb{E}_{P_\theta}\left[(R(\tau) - b(q)) \sum_{t \in \mathcal{P}(\tau)} \nabla \log \pi_\theta(x_t \mid h_t) \right]. \tag{17}$$

Here $b(q)$ is a reward baseline depending only on the question. Its expected contribution vanishes since $\mathbb{E}[\nabla \log P_\theta] = 0$. A within-group baseline instead depends on sampled answers. For G independent trajectories, let $R_i = R(\tau_i)$, $\bar{R} = G^{-1} \sum_{m=1}^{G} R_m$, and $z_i = \nabla \log P_\theta(\tau_i \mid q)$. Then

$$\mathbb{E}\left[\frac{1}{G} \sum_i (R_i - \bar{R}) z_i \right] = \left(1 - \frac{1}{G}\right) \nabla J_R(\theta). \tag{18}$$

To see this, expand $\bar{R}$: the $m = i$ term contributes $G^{-1} \nabla J_R$, while independence makes every $m \neq i$ cross-term zero. Division by a random group standard deviation introduces an additional reward-dependent factor. The group surrogate therefore requires separate analysis, rather than an assumption that its expectation equals Equation 17.

H.2 Redistribution of group-relative credit

For a binary reward vector, let $p = G^{-1} \sum_i r_i$ be its positive fraction. Define $d(p) = \gamma \sqrt{p(1-p)} + \epsilon_A$, where $\gamma = \sqrt{G/(G-\nu)}$. The positive and negative advantage levels are $A_+(p) = (1-p)/d(p)$ and $A_-(p) = -p/d(p)$. Differentiation on $0 < p < 1$ yields

$$A'_+(p) = -\frac{\epsilon_A + \gamma(1-p)/(2\sqrt{p(1-p)})}{d(p)^2} < 0, \qquad A'_-(p) = -\frac{\epsilon_A + \gamma p/(2\sqrt{p(1-p)})}{d(p)^2} < 0. \tag{19}$$

Write $\Delta A_i = A_i^{\mathrm{corr}} - A_i^{\mathrm{ref}}$ for the advantage change. When rewards change only from zero to one, the positive fraction cannot decrease. For a trajectory whose reward stays fixed, its advantage follows one of the decreasing functions above, so $\Delta A_i \leq 0$. A trajectory with a corrected reward moves from a nonpositive to a nonnegative advantage, so $\Delta A_i \geq 0$. Both advantage vectors sum to zero. Thus, for the corrected set $\mathcal{C}$,

$$\sum_{i \in \mathcal{C}} \Delta A_i = -\sum_{i \notin \mathcal{C}} \Delta A_i \quad \Longrightarrow \quad \sum_{i \in \mathcal{C}} |\Delta A_i| = \sum_{i \notin \mathcal{C}} |\Delta A_i| = \tfrac{1}{2} \sum_i |\Delta A_i|. \tag{20}$$

The zero-advantage convention covers activation, deactivation and uniform shifts. The identity concerns unweighted advantages: unequal weights or different score gradients need not preserve the equal split. It does not extend to arbitrary shaped rewards or simultaneous corrections in both directions.

H.3 Likelihood Changes Under Revised Advantages

Collect the advantages in $A^c = (A_1^c, \ldots, A_G^c)^\top$ and the weighted score gradients in $S = [s_1, \ldots, s_G]$. Let $H = S^\top S$ be their Gram matrix and $\Delta A = A^{\mathrm{corr}} - A^{\mathrm{ref}}$. Since $g^c = SA^c$,

$$\begin{aligned} \|g^c\|^2 &= (A^c)^\top H A^c, \\ (g^{\mathrm{ref}})^\top g^{\mathrm{corr}} &= (A^{\mathrm{ref}})^\top H A^{\mathrm{corr}}, \\ \|\Delta g\|^2 &= \Delta A^\top H \Delta A. \end{aligned} \tag{21}$$

The diagonal alone is insufficient because cross-trajectory inner products can reinforce or cancel.

For a fixed answer or action continuation in an independent context, let $F(\theta)$ be its mean token log-likelihood and $u = \nabla F(\theta_0)$. Its derivative along the update is $u^\top g^c = \sum_i K_i A_i^c$, where $K_i = u^\top s_i$ measures alignment with trajectory i's weighted score gradient. For a fixed parameter direction $\mathbf{d}$ and a small scalar step α,

$$F(\theta_0 + \alpha \mathbf{d}) - F(\theta_0) = \alpha u^\top \mathbf{d} + O(\alpha^2 \|\mathbf{d}\|^2). \tag{22}$$

This remainder requires suitable local smoothness and a bounded Hessian. For the norm-bounded comparisons, $\alpha = \rho$ and $\mathbf{d} = g^c / \max(\|g^{\mathrm{ref}}\|, \|g^{\mathrm{corr}}\|)$. The two directions share this normalization, preserving their relative magnitudes. A zero update leaves likelihood unchanged. A centered finite difference has $O(\alpha^2)$ truncation error when the third directional derivative is bounded. Floating-point error adds another term and need not decrease with α.

The contributions $\Delta g = \sum_{i \in \mathcal{C}} \Delta A_i s_i + \sum_{i \notin \mathcal{C}} \Delta A_i s_i$ need not yield equal likelihood effects: their projections onto u can differ in sign and magnitude.

The shared step coefficient. Let $N_g = \max(\|g^{\mathrm{ref}}\|, \|g^{\mathrm{corr}}\|) > 0$ and $\eta = \rho / N_g$. If the Hessian operator norm of F is bounded by L_F on both update segments, the full expansion underlying Equation 4 is

$$F(\theta_0 + \eta g^{\mathrm{corr}}) - F(\theta_0 + \eta g^{\mathrm{ref}}) = \eta \nabla F(\theta_0)^\top (g^{\mathrm{corr}} - g^{\mathrm{ref}}) + \mathcal{R}_F, \tag{23}$$

where the remainder $\mathcal{R}_F$ satisfies

$$|\mathcal{R}_F| \le \frac{L_F \rho^2}{2 N_g^2} \left(\|g^{\mathrm{ref}}\|^2 + \|g^{\mathrm{corr}}\|^2 \right) \le L_F \rho^2. \tag{24}$$

If every direction participating in the shared normalization is zero, all parameter conditions remain at θ_0. The bound specifies the form of the approximation error. The finite-difference measurements assess its behavior at the experimental scales.

Contributions of corrected and unchanged rewards. For binary rewards corrected only from zero to one, let $\mathcal{C}$ index the trajectories with corrected rewards and $M = \sum_{i \in \mathcal{C}} \Delta A_i = \frac{1}{2} \|\Delta A\|_1 > 0$. Define

$$\bar{s}_+ = \sum_{i \in \mathcal{C}} \frac{\Delta A_i}{M} s_i, \qquad \bar{s}_- = \sum_{i \notin \mathcal{C}} \frac{-\Delta A_i}{M} s_i. \tag{25}$$

Each set of weights is nonnegative and sums to one. For an evaluated log-likelihood F, let $u = \nabla F(\theta_0)$. Substitution into the fixed-group gradient gives

$$\Delta g = M(\bar{s}_+ - \bar{s}_-), \qquad u^\top \Delta g = M(u^\top \bar{s}_+ - u^\top \bar{s}_-). \tag{26}$$

The score vectors include the declared fixed loss coefficients. These coefficients need not be equal for the decomposition to hold. A differentiable local expansion gives $F(\theta_0 + \eta g^{\mathrm{corr}}) - F(\theta_0 + \eta g^{\mathrm{ref}}) = \eta M \, u^\top (\bar{s}_+ - \bar{s}_-) + o(\eta)$. Thus positive alignment difference is necessary and sufficient for positive *first-order* likelihood contrast, not for every finite perturbation. When $M = 0$, the gradient difference is zero without defining the two means.