EVESKILLDEBATE: EVIDENCE-VERIFIABLE MULTI-AGENT DEBATE WITH SELF-EVOLVING AGENT REASONING SKILL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-Agent Debate (MAD) can improve language-model reasoning through iterative proposal, critique, and revision. However, agreement under a fixed protocol does not ensure that decisions are evidence-backed, nor does it convert recurring failures into persistent procedural improvements. To address these limitations, we propose a novel paradigm, EVESKILLDEBATE, which consists of *Evidence-Verifiable Multi-Agent Debate* and *Self-Evolving Agent Reasoning Skill*. The debate module aims to construct candidate-specific reasoning paths that bind candidate-level claims to document spans and knowledge-graph edges. A frozen verifier assesses citation validity, evidential support, contradiction, and graph consistency, after which an evidence-aware judge aggregates the debated proposals and verification signals. The evolution module contrasts successful and failed (*i.e.*, higher- and lower-utility) execution trajectories to identify recurring skill-level failures and propose targeted updates to the agents' external inference-time policies while keeping the underlying LLM frozen. Candidate skills are evaluated on subsequent batches and deployed only when they satisfy a predeclared feedback gate; post-deployment regressions trigger rollback. Across five public datasets spanning four domains and three LLM families, EVESKILLDEBATE delivers its strongest and most consistent gains on evidence-grounded ranking tasks while achieving high evidence traceability. Evolved skills win 72.4% of non-tied comparisons and achieve 93.7% non-regression, showing that EVESKILLDEBATE converts trajectory feedback into beneficial skill updates in the MAD systems.

1 INTRODUCTION

"Groups are remarkably intelligent, and are often smarter than the smartest people in them."
— James Surowiecki, *The Wisdom of Crowds* Chafetz (2005).

Large language models (LLMs) have achieved strong performance across a broad range of tasks, yet reliable complex reasoning remains challenging Zhu et al. (2023); Du et al. (2024c); OpenAI (2026); Anthropic (2026). Inference-time methods based on chain-of-thought, graph-structured, and meta-reasoning can strengthen an individual model's reasoning process Wei et al. (2022); Besta et al. (2024); Gao et al. (2024), but unsupported intermediate claims and compounding errors can persist. Multi-agent debate (MAD) offers a complementary direction: multiple LLM agents propose, critique, and revise candidate solutions, allowing the system to combine diverse perspectives and potentially correct individual errors Yin et al. (2023); Du et al. (2024b); Becker et al. (2025); Sun et al. (2025); Xiong et al. (2023); Guo et al. (2024); Tran et al. (2025); Chen et al. (2024a).

Research on MAD has progressively enriched this paradigm Wang et al. (2024a). Early studies established basic debate protocols through role specialization Liang et al. (2024a); Chan et al. (2024); Wang et al. (2024b) and examined how the number of debate rounds, interaction order, and team size affect reasoning performance Du et al. (2024b;a); Xu et al. (2023); Liang et al. (2024b); Xiong et al. (2023). Subsequent work extended these protocols to cross-model communication and consensus-oriented interaction Chen et al. (2024a); Yin et al. (2023), introducing structured communication, inter-model information exchange, and consensus rules to improve coordination and reasoning consistency. More recent methods have targeted efficiency, robustness, and deliberation quality

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.

through sparse communication Li et al. (2024), structured interaction Liu et al. (2024), budget-aware coordination Yang et al. (2026c), and diversity-oriented deliberation Li et al. (2026).

Despite this progress, existing MAD frameworks still face two tightly coupled limitations. **1) Static Protocols.** Many MAD systems employ roles, prompts, and coordination policies that are fixed before deployment Liang et al. (2024a); Wang et al. (2024b); Chan et al. (2024); Du et al. (2024a); Zhang et al. (2026b); Xiong et al. (2023); Liang et al. (2024b); Li et al. (2024); Cui et al. (2025). Although often effective on their target tasks, these protocols lack an explicit mechanism for converting observed successes and failures into persistent procedural improvements. Recurring errors may therefore remain unlocalized and reappear as tasks or data distributions change, limiting adaptation even when the underlying LLM is strong. **2) Unverifiable Decisions.** Many MAD systems assess arguments through generated natural-language rationales or inter-agent agreement, without requiring each factual claim to be bound to independently inspectable evidence Liang et al. (2024a); Wang et al. (2024b); Chan et al. (2024); Du et al. (2024a); Zhang et al. (2026b); Xiong et al. (2023); Liang et al. (2024b); Li et al. (2024); Cui et al. (2025); Zeng et al. (2025); Yang et al. (2026c); Li et al. (2026). Unsupported claims may therefore propagate across agents, while agreement may be mistaken for correctness when agents share the same error. Without explicit claim–evidence links and provenance, users cannot determine why an answer was accepted, which reasoning step failed, or whether a revision is better grounded. The two limitations reinforce each other: skill evolution requires audited trajectories, while evidence-grounded debate benefits from reusable corrections to recurring failures.

To address these limitations, we propose EVESKILLDEBATE, a closed-loop framework that integrates evidence-grounded multi-agent deliberation with feedback-driven skill evolution through two tightly coupled modules: *Evidence-Verifiable Multi-agent Debate* and *Self-Evolving Agent Reasoning Skill*. The first module comprises evidence-path grounding, evidence-guided debate, and evidence-aware closing. Agents retrieve document, graph, and tool evidence to construct candidate-specific reasoning paths. A frozen verifier assesses citation validity, support, contradiction, and graph consistency, guiding ranking revisions. After bounded debate, an evidence-aware judge aggregates the verified proposals and returns an answer with supporting evidence. The second module evolves versioned inference-time policies while keeping the LLM frozen. Under a label-delayed protocol, deployed and shadow policies are evaluated on the same batches, with outputs and traces sealed before label release. Successful–failed trajectory pairs are aligned to localize faulty clauses and propose targeted edits. Candidate policies are evaluated on later batches, deployed through a predeclared feedback gate, and rolled back upon subsequent regression. Across five cross-domain benchmarks and three LLM families, EVESKILLDEBATE achieves the best MRR and NDCG@3 estimates in every ranking-task–backbone pairing. Evolved skills win 72.4% of non-tied comparisons with a 93.7% non-regression rate, demonstrating targeted, low-regression adaptation. Our main contributions are summarized as:

- We formulate skill evolution in MAD as a controlled lifecycle for external inference-time policies: audited trajectories identify localized failures, candidate edits are evaluated prospectively, and deployment is governed by explicit acceptance and rollback rules.
- We propose EVESKILLDEBATE, comprising *Evidence-Verifiable Multi-Agent Debate* and *Self-Evolving Agent Reasoning Skill*. The former binds candidate claims and critiques to auditable evidence; the latter contrasts successful and failed trajectories to localize clause-level failures and evaluate targeted policy updates on subsequent batches.
- We evaluate EVESKILLDEBATE on five benchmarks spanning four domains and multiple LLM families. The experiments measure task performance, evidence traceability, skill-evolution behavior, and computational cost, while identifying weaker gains on single-answer mathematical reasoning.

2 RELATED WORK

2.1 MULTI-AGENT DEBATE

MAD emulates human deliberation by coordinating diverse agents toward a shared solution. Early work established role-based debate protocols and examined the effects of debate rounds, turn order, and agent count Liang et al. (2024a); Chan et al. (2024); Wang et al. (2024b); Du et al. (2024b;a); Xu et al. (2023); Liang et al. (2024b); Xiong et al. (2023). Role placement and viewpoint assignment were also shown to influence reasoning performance Zhang et al. (2026b). Subsequent studies extended

MAD to cross-model communication and consensus-oriented interaction Chen et al. (2024a); Yin et al. (2023), with Exchange-of-Thought enabling information exchange across models Yin et al. (2023). Recent methods improve efficiency, robustness, and reasoning diversity through sparse communication Li et al. (2024), structured interaction Liu et al. (2024), budget-aware design Yang et al. (2026c), and diversity-oriented deliberation Cui et al. (2025); Liu et al. (2025b); Li et al. (2026). Nevertheless, most MAD systems use static protocols and lack verifiable evidence.

2.2 Self-Evolving Agent Skills

Self-evolving skill systems accumulate execution experience to produce persistent improvements in agent capabilities Jiang et al. (2026); Zhang et al. (2026a); Tian et al. (2026). Existing methods broadly use three feedback sources: successful trajectories Liu et al. (2026a), failed trajectories Liu et al. (2026c;b), or both Yang et al. (2026a;d); Ni et al. (2026); Ma et al. (2026). Success-aware methods retain reusable discoveries from successful executions; for example, Liu et al. (2026a) manages skills as lifecycle artifacts and admits only successful, attributable updates. Failure-driven methods diagnose execution errors and revise the affected skills using verifier feedback Liu et al. (2026c). Mixed-trajectory methods combine successful procedures with failure-derived insights; for instance, Yang et al. (2026a) distills reusable skills from diverse solution trajectories grouped by problem archetype. However, existing evaluations rarely isolate these feedback sources under matched conditions, leaving their contributions to persistent skill improvement unclear.

3 EVESkillDebate

To address static coordination and unverifiable decisions, we propose EVESkillDebate, coupling *Evidence-Verifiable Multi-Agent Debate* with *Self-Evolving Agent Reasoning Skill*. Figure 1 presents the framework, and Appendix B provides the complete procedure.

3.1 Setup and Notation

An *instance* is $x = (q, \mathcal{C}, \mathcal{D}, \mathcal{G})$, where q is a query, $\mathcal{C} = \{c_1, \dots, c_n\}$ is a candidate set with $n \geq 2$, $\mathcal{D}$ is a document collection, and $\mathcal{G} = (\mathcal{V}, \mathcal{R}, E_{\mathcal{G}})$ is a knowledge graph. For each $c \in \mathcal{C}$, y_c denotes the candidate-level proposition that c satisfies the factual requirements of q; nonfactual preference fit is handled through F_i and J_ϕ. A ranking $\sigma : \mathcal{C} \to [n]$ is a bijection, with rank one being best; all ties are broken by a fixed order $\prec_{\mathcal{C}}$. We use $K \geq 2$ agents, define $[m] = \{1, \dots, m\}$, and write $\mathrm{top}(\sigma) = \sigma^{-1}(1)$.

An *evidence unit* e is a document span, an edge in $E_{\mathcal{G}}$, or a replayable output produced by a registered deterministic tool. For each $c \in \mathcal{C}$, an *evidence assertion* $a = (e, y_c)$ binds e to the candidate-level claim y_c. A frozen verifier returns citation validity $\nu(e) \in \{0, 1\}$, support $\gamma(e, y_c) \in [0, 1]$, and contradiction $\gamma_\neg(e, y_c) \in [0, 1]$, with graph consistency serving as an auxiliary verification signal.

A *reasoning skill* is an external, versioned inference-time policy represented by natural-language procedural clauses supplied to the frozen LLM. Agent i uses $s_i = (\pi_i, R_i, T_i)$, where π_i contains the inference-visible clauses with stable identifiers, R_i is an evolution-only record, and T_i specifies the tool interface. Only $\bar{s}_i = (\pi_i, \mathrm{spec}(T_i))$ enters the inference context; skill evolution modifies π_i rather than the underlying model parameters. The retriever obtains task evidence, not skills from a memory store. Appendix A provides the notation summary.

3.2 Evidence-Verifiable Multi-Agent Debate

This module first grounds candidate claims in explicit evidence paths, then permits revisions only through audited critiques, and finally applies a frozen evidence-aware judge after bounded debate.

First, evidence-path grounding: every opening claim is linked to a verifiable evidence path. Before observing peers, agent i retrieves:

$$(\mathcal{D}_i, \mathcal{G}_i, \mathcal{O}_i) = \mathrm{Retrieve}_i(q, \mathcal{C}, \bar{s}_i; \mathcal{D}, \mathcal{G}), \quad \mathcal{D}_i \subseteq \mathcal{D}, \quad \mathcal{G}_i \preceq \mathcal{G}, \tag{1}$$

where $\preceq$ denotes subgraph inclusion and $\mathcal{O}_i$ contains replayable tool outputs. Let $\mathcal{B}_i$ be the resulting evidence units and $\mathcal{B} = \bigcup_{i=1}^{K} \mathcal{B}_i$. Retrieval uses fixed budgets and deterministic tie-breaking and has

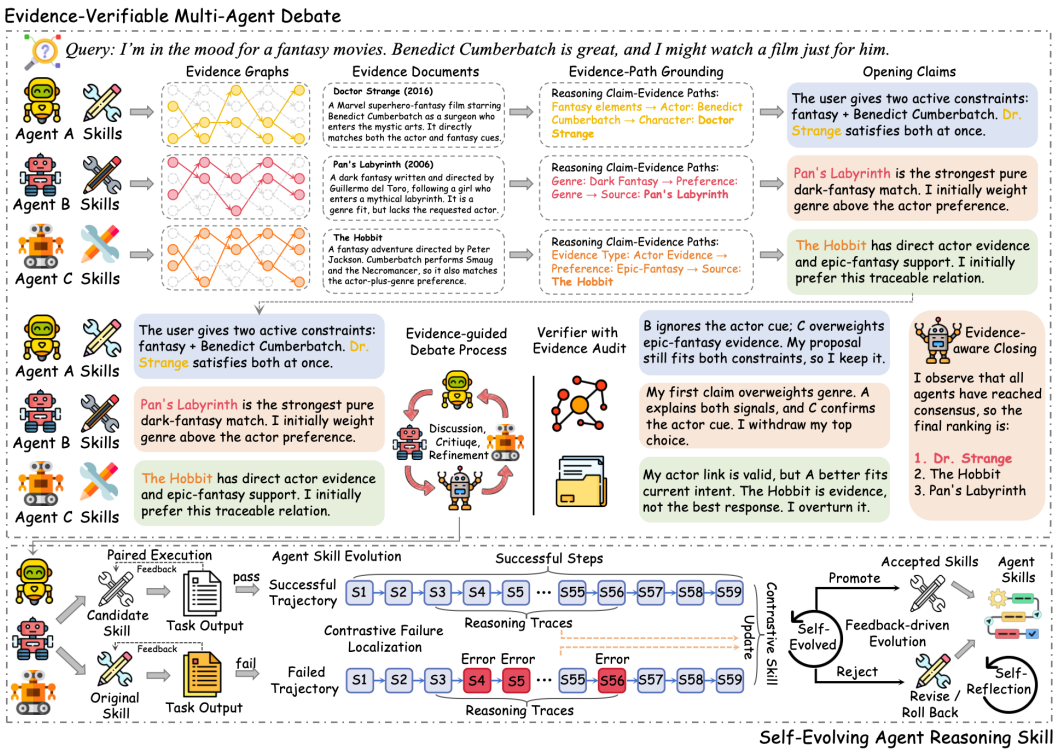


Figure 1: **Overview of EVESKILLDEBATE.** *Evidence-Verifiable Multi-Agent Debate* grounds deliberation in auditable evidence and produces evidence-aware decisions with traceable reasoning trajectories. *Self-Evolving Agent Reasoning Skill* contrasts paired successful and failed trajectories to derive localized policy updates, which undergo feedback-guided deployment, rollback, and freezing.

no access to gold labels or peer outputs. For $t \in [T_{\max}]$, let $\mathcal{B}_i^1 = \mathcal{B}_i$ and $\mathcal{B}_i^t = \mathcal{B}$ for $t \geq 2$. For candidate c, agent i submits:

$$p_i^t(c) = \big((e_{i,l}^t, y_c)_{l=1}^{L_{i,c}^t}; \zeta_i^t(c)\big), \tag{2}$$

where $e_{i,l}^t \in \mathcal{B}_i^t$ and $\zeta_i^t(c)$ records how the cited evidence supports the candidate-level claim y_c. An empty path denotes that the agent makes no factual claim about the candidate; a factual claim may be asserted only through a nonempty evidence path.

For a nonempty path p, the frozen verifier computes:

$$V_t(p, y) = \mathcal{V}_\phi(\mathrm{Src}(p), \mathrm{Sup}(p, y), 1 - \mathrm{Con}_t(p, y), \mathrm{KG}(p, y)), \qquad V_t(\varnothing, y) = 0, \tag{3}$$

where $\mathcal{V}_\phi : [0,1]^4 \to [0,1]$ is a frozen verifier map and each of Src, Sup, Con_t, and KG lies in $[0,1]$. Let $V_i^t(c) = V_t(p_i^t(c), y_c)$, and let $\mathcal{H}^{t-1}$ collect earlier proposals and verifier feedback, with $\mathcal{H}^0 = \varnothing$. The agent score and ranking can be described as:

$$\begin{aligned} B_i^t(c) &= F_i\big(q, c, \mathcal{B}_i^t, \mathcal{H}^{t-1}; \bar{s}_i\big) \in [0,1], \\ \sigma_i^t &= \mathrm{Sort}_{\prec_c}^{\downarrow}\{c \mapsto B_i^t(c) : c \in \mathcal{C}\}. \end{aligned} \tag{4}$$

The verifier score $V_i^t(c)$ and agent score $B_i^t(c)$ are passed jointly to subsequent debate and closing.

Then, evidence-guided debate: critiques are verified, and rankings are revised using evidence and verifier feedback. Agents synchronously submit revised paths, rationales, and claim-bound counterevidence; messages are revealed only after all submissions are sealed. The verifier assesses whether cited evidence supports a proposed claim or contradicts an alternative. We represent agent i's belief by (B_i^t, σ_i^t). Revisions may use evidence paths, peer rationales, verifier feedback, and renewed task reasoning; agreement alone has no update rule. Appendix C.3 specifies the protocol.

Finally, evidence-aware closing: a frozen judge aggregates proposals and verification signals. At the final round $T_{\max}$, the judge sets:

$$A_i = \mathrm{Agg}_\phi\Big(\{(B_i^{T_{\max}}(c), V_i^{T_{\max}}(c), p_i^{T_{\max}}(c)) : c \in \mathcal{C}\}\Big),$$
$$J(c) = J_\phi\Big(q, c, \mathcal{B}, \{(B_i^{T_{\max}}, \sigma_i^{T_{\max}}, A_i, p_i^{T_{\max}})\}_{i=1}^K\Big) \in [0,1]. \tag{5}$$

Here A_i summarizes agent i's proposal and verifier signals, while the frozen J_ϕ assesses query fit, evidential support, contradiction, and the reported reasoning. Agreement is an input to the judge rather than an activation condition, so a better-supported minority proposal may be selected. Let $c^\dagger$ maximize $J(c)$ under $\prec_{\mathcal{C}}$. Define the ordered list as:

$$L_J = \mathrm{Sort}^{\downarrow}_{\prec_{\mathcal{C}}}\{c \mapsto J(c) : c \in \mathcal{C}\}. \tag{6}$$

For a ranking task, let $k \in [n]$ be fixed in advance. The user-facing output is:

$$\widehat{y}(x) = \begin{cases} L_J[1\!:\!k], & \text{ranking task,} \\ c^\dagger, & \text{single-answer task.} \end{cases} \tag{7}$$

For each selected candidate, the system assembles its cited documents and associated graph paths into an inspectable evidence record. Verification is relative to the retrieved sources, registered tools, and frozen verifier rather than a guarantee of real-world truth.

3.3 Self-Evolving Agent Reasoning Skill

Only π_i changes inference behavior; R_i is updated externally, while all other components remain frozen. Version v indexes the procedural profile $\boldsymbol{\pi}^{(v)} = (\pi_1^{(v)}, \ldots, \pi_K^{(v)})$.

First, paired execution: matched trajectories provide evidence for skill evolution. The stream is partitioned into ordered batches $\mathcal{X}_{1:M}$. Within batch b, a deployed champion v_b^{ch} and an optional distinct shadow v_b^{sh} are frozen and run on the same instances under identical predeclared decoding conditions, using common random seeds when the backend supports them; only the champion produces user-facing answers. Version v records the observable trace:

$$\tau_v(x) = \big((i_m, t_m, \alpha_m, c_m, a_m, \ell_m, \mathbf{g}_m)_{m=1}^{L_v(x)}; \widehat{y}_v(x)\big), \tag{8}$$

containing actions, evidence assertions, stable clause tags, and pre-label audit features—not private chain of thought. The tuple records agent, round, action, candidate, assertion, clause tag, and audit features; $L_v(x)$ is its length. Outputs and traces are sealed before labels are revealed. After release, let $c^\star(x)$ denote the gold target. With a predeclared scoring function m_x whose range is $[0,1]$, define $s_v(x) = m_x(\widehat{y}_v(x), c^\star(x))$. A contrastive pair comprises two active versions with different utility:

$$\mathcal{P}_b^{\mathrm{tr}} = \big\{(x, \tau_{v^+}(x), \tau_{v^-}(x)) : x \in \mathcal{X}_b,\ v_b^{\mathrm{sh}} \neq \perp,$$
$$\{v^+, v^-\} = \{v_b^{\mathrm{ch}}, v_b^{\mathrm{sh}}\},\quad s_{v^+}(x) > s_{v^-}(x)\big\}. \tag{9}$$

Here v^+ is the higher-utility version and v^- is the other member of the matched comparison. Equal-utility outputs provide no directional preference. Thus, v^+ and v^- are uniquely identified as the successful and failed members of the active champion–shadow pair, where these labels denote higher and lower utility within that matched pair rather than absolute success or failure. Without an active shadow, the set is empty. Only pairs with $v^- = v_b^{\mathrm{ch}}$ may generate edits to the champion, as detailed in Appendix C.6. Appendix C.5 defines sealing and the trace schema.

Then, contrastive failure localization: each policy edit is tied to a localized skill-level failure. A fixed contrastive alignment combines stable clause identity, observable action semantics, and audit features for each successful–failed pair. An unmatched, well-supported action in the successful trajectory may identify an omitted reasoning operation. Missing audits are never treated as failures. Let $\overline{\Lambda}_b$ contain the implicated clause identifiers; their diagnosed failure types are recorded separately. Every proposed diff obeys:

$$\mathrm{Loc}(\Delta\boldsymbol{\pi}_b) \subseteq \overline{\Lambda}_b, \qquad |\Delta\boldsymbol{\pi}_b| \leq H_{\max}. \tag{10}$$

Here $H_{\max} \geq 1$ is fixed and $|\Delta\boldsymbol{\pi}_b|$ counts edited clauses; the allowed operation at each skill-level location must also match its diagnosed failure type. Successful trajectories serve as references rather than edit targets. The aligned diagnostics form a batch-level contrast representation, and its scorer is updated only between batches. Alignment, local auditing, representation construction, scorer learning, and edit validation are defined in Appendix C.6.

Finally, feedback-guided acceptance, rollback, and freezing: only empirically supported edits are promoted. A frozen candidate is evaluated against the frozen champion only on later batches. For the first $N \geq 1$ disagreements sealed before label release, define after release $d_j = s_{\mathrm{cand},j} - s_{\mathrm{ch},j} \in [-1, 1]$; fix $h \geq 1$ and let

$$\bar{d}_N = \frac{1}{N}\sum_{j=1}^{N} d_j, \qquad \bar{d}_N^{\mathrm{rec}} = \frac{1}{h_N}\sum_{j=N-h_N+1}^{N} d_j, \quad h_N = \min(h, N). \tag{11}$$

Let $G_N = (N, \bar{d}_N, \bar{d}_N^{\mathrm{rec}})$ summarize the number of disagreements, cumulative advantage, and recent advantage. For a fixed $N_{\min} \geq 1$, promotion requires $\mathrm{Gate}_{\mathrm{dep}}(G_N) = 1$ after at least $N_{\min}$ disagreements. After promotion, a fresh comparison analogously constructs G_N^{rb} using $d_j^{\mathrm{rb}} = s_{\mathrm{new},j} - s_{\mathrm{prev},j}$ and triggers rollback when $\mathrm{Gate}_{\mathrm{rb}}(G_N^{\mathrm{rb}}) = 1$. These gates are predeclared maps $\mathbb{N}_{\geq 1} \times [-1, 1]^2 \to \{0, 1\}$ for positive deployment and negative rollback evidence; they do not imply a time-uniform confidence guarantee or distribution-free improvement. After the final batch, the deployed version is frozen. All reported online metrics use predictions sealed before the corresponding labels were revealed; past instances are never re-evaluated under later skill versions.

4 Experiments and Analyses

In this section, we conduct experiments to fully evaluate our proposed method and answer the following questions: **RQ1:** How does EVESkillDebate compare with representative baselines across benchmarks and LLM backbones? **RQ2:** How verifiable and reliable are the decisions (i.e., response) produced by EVESkillDebate? **RQ3:** How effective is skill evolution, and how valid, behaviorally diverse, and transferable are the evolved skills? **RQ4:** How does each component contribute to task performance? **RQ5:** How does EVESkillDebate balance task performance and computational efficiency? **RQ6:** How sensitive is EVESkillDebate to its key hyperparameters?

Table 1: Main comparison with multi-agent baselines across representative backbones.

Model	Method	ReDial					TG-ReDial					RareBench					HLE	SGPQA
		H@1	H@2	H@3	MRR	N@3	H@1	H@2	H@3	MRR	N@3	H@1	H@2	H@3	MRR	N@3	Acc.	Acc.
GPT-5.4 Nano	Classic MAD	0.239	0.458	0.639	0.409	0.468	0.084	0.161	0.213	0.140	0.159	0.636	0.692	0.762	0.687	0.706	0.212	0.428
	Debate-or-Vote (Dec.)	0.249	0.250	0.250	0.250	0.250	0.079	0.079	0.079	0.079	0.079	0.641	0.641	0.641	0.641	0.641	0.141	0.457
	Debate-or-Vote (Sparse)	0.260	0.265	0.266	0.263	0.263	0.078	0.080	0.082	0.079	0.080	0.631	0.647	0.647	0.639	0.641	0.176	0.449
	Debate-or-Vote (Cent.)	0.257	0.257	0.257	0.257	0.257	0.082	0.082	0.082	0.082	0.082	0.624	0.624	0.624	0.624	0.624	0.165	0.468
	MATTRL	0.248	0.480	0.654	0.422	0.481	0.074	0.138	0.201	0.127	0.146	0.609	0.678	0.765	0.672	0.696	0.259	0.396
	MADKE	0.230	0.444	0.587	0.385	0.436	0.036	0.064	0.094	0.060	0.069	0.561	0.648	0.749	0.638	0.666	0.129	0.411
	Adaptive Judge	0.240	0.457	0.629	0.406	0.463	0.070	0.138	0.198	0.124	0.143	0.631	0.685	0.765	0.685	0.705	0.200	0.407
	iMAD	0.236	0.448	0.628	0.402	0.460	0.079	0.135	0.179	0.122	0.137	0.611	0.683	0.764	0.674	0.697	0.176	0.392
	Identity (Anon.)	0.228	0.454	0.638	0.402	0.462	0.075	0.135	0.195	0.125	0.143	0.628	0.698	0.785	0.692	0.716	0.212	0.417
	Identity (Exposed)	0.235	0.447	0.629	0.402	0.460	0.072	0.141	0.207	0.129	0.149	0.615	0.686	0.778	0.681	0.706	0.212	0.400
	ED2D	0.230	0.433	0.580	0.380	0.431	0.033	0.076	0.129	0.072	0.087	0.567	0.655	0.744	0.641	0.667	0.165	0.342
	MAD-Logic	0.199	0.395	0.560	0.352	0.405	0.023	0.059	0.094	0.052	0.063	0.590	0.631	0.714	0.638	0.657	0.176	0.390
	DoM	0.179	0.329	0.435	0.289	0.327	0.036	0.060	0.100	0.062	0.071	0.542	0.627	0.706	0.611	0.635	0.118	0.306
	EVESkillDebate (Ours)	0.272	0.528	0.652	0.441	0.495	0.363	0.446	0.506	0.424	0.445	0.670	0.832	0.898	0.773	0.805	0.200	0.444
Gemini 3.5 Flash-Lite	Classic MAD	0.277	0.478	0.619	0.425	0.475	0.099	0.155	0.201	0.142	0.157	0.690	0.714	0.756	0.716	0.726	0.176	0.597
	Debate-or-Vote (Dec.)	0.285	0.285	0.285	0.285	0.285	0.107	0.107	0.107	0.107	0.107	0.676	0.676	0.676	0.676	0.676	0.212	0.612
	Debate-or-Vote (Sparse)	0.289	0.297	0.297	0.293	0.294	0.110	0.112	0.112	0.111	0.111	0.682	0.686	0.686	0.684	0.684	0.247	0.623
	Debate-or-Vote (Cent.)	0.295	0.296	0.296	0.295	0.296	0.112	0.112	0.112	0.112	0.112	0.681	0.682	0.682	0.681	0.681	0.224	0.637
	MATTRL	0.261	0.495	0.650	0.429	0.486	0.096	0.151	0.195	0.139	0.153	0.672	0.696	0.732	0.696	0.705	0.129	0.526
	MADKE	0.241	0.427	0.519	0.365	0.405	0.032	0.059	0.078	0.052	0.058	0.634	0.661	0.703	0.662	0.672	0.153	0.558
	Adaptive Judge	0.261	0.485	0.648	0.428	0.484	0.096	0.161	0.209	0.145	0.161	0.690	0.718	0.770	0.721	0.733	0.153	0.574
	iMAD	0.255	0.447	0.585	0.397	0.445	0.092	0.150	0.197	0.137	0.152	0.686	0.715	0.752	0.713	0.723	0.118	0.560
	Identity (Anon.)	0.241	0.480	0.652	0.418	0.478	0.100	0.169	0.218	0.151	0.168	0.684	0.722	0.773	0.720	0.734	0.153	0.602
	Identity (Exposed)	0.253	0.485	0.661	0.427	0.487	0.098	0.161	0.218	0.148	0.166	0.679	0.713	0.773	0.716	0.730	0.153	0.625
	ED2D	0.246	0.471	0.619	0.408	0.462	0.063	0.115	0.151	0.101	0.114	0.662	0.699	0.755	0.699	0.714	0.165	0.556
	MAD-Logic	0.222	0.446	0.611	0.389	0.446	0.044	0.091	0.145	0.085	0.101	0.652	0.681	0.738	0.685	0.699	0.141	0.581
	DoM	0.236	0.419	0.562	0.375	0.423	0.062	0.099	0.143	0.095	0.107	0.675	0.701	0.735	0.699	0.709	0.106	0.560
	EVESkillDebate (Ours)	0.333	0.551	0.711	0.496	0.551	0.312	0.388	0.456	0.373	0.394	0.640	0.803	0.857	0.739	0.770	0.224	0.646
Claude Haiku 4.5	Classic MAD	0.250	0.487	0.681	0.433	0.496	0.067	0.120	0.155	0.105	0.118	0.588	0.649	0.741	0.649	0.672	0.094	0.472
	Debate-or-Vote (Dec.)	0.278	0.303	0.303	0.290	0.294	0.052	0.059	0.059	0.056	0.056	0.604	0.613	0.614	0.609	0.610	0.141	0.474
	Debate-or-Vote (Sparse)	0.275	0.311	0.311	0.293	0.298	0.055	0.080	0.084	0.069	0.073	0.602	0.628	0.630	0.616	0.619	0.141	0.476
	Debate-or-Vote (Cent.)	0.290	0.290	0.290	0.290	0.290	0.060	0.060	0.060	0.060	0.060	0.595	0.595	0.595	0.595	0.595	0.106	0.486
	MATTRL	0.244	0.470	0.657	0.419	0.480	0.056	0.112	0.173	0.104	0.122	0.561	0.627	0.702	0.619	0.641	0.118	0.407
	MADKE	0.244	0.459	0.653	0.416	0.476	0.013	0.025	0.056	0.030	0.036	0.509	0.580	0.686	0.580	0.607	0.082	0.417
	Adaptive Judge	0.232	0.460	0.641	0.407	0.467	0.062	0.106	0.149	0.098	0.111	0.584	0.641	0.735	0.644	0.667	0.082	0.444
	iMAD	0.239	0.461	0.654	0.415	0.476	0.056	0.103	0.155	0.097	0.112	0.579	0.651	0.738	0.644	0.668	0.106	0.421
	Identity (Anon.)	0.240	0.459	0.655	0.415	0.476	0.055	0.106	0.146	0.094	0.107	0.596	0.643	0.739	0.652	0.674	0.094	0.442
	Identity (Exposed)	0.235	0.464	0.654	0.413	0.474	0.058	0.110	0.151	0.098	0.111	0.591	0.644	0.737	0.648	0.671	0.106	0.442
	ED2D	0.246	0.458	0.635	0.411	0.468	0.044	0.083	0.123	0.077	0.089	0.542	0.613	0.713	0.611	0.637	0.129	0.417
	MAD-Logic	0.211	0.445	0.635	0.391	0.453	0.033	0.060	0.108	0.063	0.074	0.561	0.640	0.749	0.636	0.665	0.071	0.417
	DoM	0.229	0.443	0.660	0.408	0.472	0.039	0.075	0.110	0.068	0.079	0.586	0.640	0.735	0.645	0.668	0.106	0.436
	EVESkillDebate (Ours)	0.285	0.525	0.702	0.464	0.525	0.319	0.387	0.438	0.370	0.387	0.645	0.797	0.865	0.744	0.775	0.106	0.532

4.1 EXPERIMENTAL PROTOCOL

1) Datasets. We evaluate EVESKILLDEBATE on five public benchmarks spanning recommendation, medical diagnosis, mathematics, and education: ReDial (Li et al., 2018), TG-ReDial (Zhou et al., 2020), RareBench (Chen et al., 2024b), HLE Math (Phan et al., 2025), and SuperGPQA (shortly, S-GPQA) (M-A-P Team et al., 2025). Appendix F details the dataset sizes, construction procedures, filtering criteria, and evidence sources. **2) Evaluation metrics.** For the ranking tasks (ReDial, TG-ReDial, and RareBench), we report Hit@k ($k \in \{1, 2, 3\}$), MRR, and NDCG@3; the two single-answer tasks (HLE Math, and SuperGPQA) use accuracy. Skill evolution is assessed under a label-delayed online protocol. Appendix G provides detailed metric definitions and control settings. **3) Compared methods.** We compare EVESKILLDEBATE with 13 configurations from ten recent representative MAD: Classic MAD (Du et al., 2024b), Debate-or-Vote (Choi et al., 2025), MAT-TRL (Hu et al., 2026), MADKE (Wang et al., 2025), Adaptive Judge (Hu et al., 2025), iMAD (Fan et al., 2026), Identity Anonymization (Choi et al., 2026), ED2D (Han et al., 2026), MAD-Logic (Yang et al., 2026b), and DoM (Liu et al., 2025a). Further details are provided in Appendix H.

4.2 OVERALL PERFORMANCE (RQ1)

We evaluate one LLM backbone from each of OpenAI, Google, and Anthropic under identical settings. Table 1 compares EVESKILLDEBATE with 13 baselines across five benchmarks. On the three ranking tasks, EVESKILLDEBATE achieves the highest H2, MRR, and NDCG3 point estimates for every backbone–task combination, indicating more reliable placement of the gold candidate near the top of the ranking rather than improvement at a single cutoff. The largest gains occur on TG-ReDial: relative to the strongest backbone-specific baseline, H1 improves by 20.0–27.9 points, MRR by 22.2–28.4 points, and NDCG3 by 22.6–28.6 points. On ReDial, the MRR and NDCG3 gains are 1.9–6.7 and 1.4–6.4 points; on RareBench, they are 1.8–9.2 and 3.6–10.1 points. Two top-1 exceptions remain—RareBench with Gemini and ReDial with Claude—although both settings improve at H2, MRR, and NDCG3. Results on single-answer tasks are less uniform: EVESKILLDEBATE performs best on SuperGPQA with Gemini and Claude, but not with GPT, and does not surpass the strongest HLE-Math baseline. The observed advantage is therefore clearest on evidence-grounded ranking tasks. Appendix D provides additional backbone results.

4.3 EVIDENCE VERIFIABILITY AND RESPONSE RELIABILITY (RQ2)

Table 2 measures whether the evidence can be independently traced and whether the final decision is supported by the verification and debate processes. **1) Evidence verifiability.** Provenance validity, candidate coverage, and KG-path validity are 100.0% across all three backbone families. Because these properties are explicitly enforced by the framework, these values primarily measure compliance with its structural constraints. Overall citation resolution is 99.5%, indicating that 0.5% of citations cannot be mapped to retained evidence records. Dual-view coverage is 96.2% overall and ranges from 94.0% to 99.9% across providers. Thus, most evaluated decisions contain both documentary and relational evidence, although their relative use differs across backbone families. **2) Response reliability.** Audit coverage is 100.0%, indicating that every returned candidate receives explicit support and contradiction scores. The positive-margin rate is 94.9% overall, ranging from 89.3% for Anthropic to 97.3% for OpenAI. Ranking agreement between the aggregate and audited agent rankings is 90.6%, while 86.1% of decisions receive support from at least two agents using distinct evidence signatures. Independent corroboration is therefore the lowest of the reported decision-level measures. Rationale traceability reaches 99.6%, showing that nearly all returned decisions retain both a rationale and its associated evidence chain. Collectively, these results describe high observed rates of traceability and procedural consistency. **3) Claim–evidence human evaluation.** We evaluate a blinded, provider-balanced sample of 450 final-decision claims, with 150 claims from each backbone family. Human annotators label each claim– evidence pair as supported, contradicted, or insufficient. Overall, 76.0% of claims are supported, 2.9% are contradicted, and 21.1% have insufficient evidence. Treating supported claims as the positive class, the verifier score achieves an AUROC of 0.792 and an AUPRC of 0.817. The support rate increases from 8.9% in the lowest-score decile to 95.6% in the highest-score decile, indicating that the verifier score is informative about evidential support.

Table 2: Evidence verifiability, answer reliability, and claim–evidence evaluation.

	Evidence verifiability					Answer reliability					Claim–evidence evaluation				
Backbone family	**Source provenance**	**Candidate retrieval**	**Citation traceability**	**Graph-path validity**	**Dual-view trace**	**Candidate verification**	**Support consistency**	**Aggregation consistency**	**Agent corroboration**	**Explanation coverage**	**Citation faithfulness ↑**	**Contradiction error ↓**	**AUROC ↑**	**AUPRC ↑**	**Support rate high / low conf.**
OpenAI	100.0%	100.0%	99.6%	100.0%	94.0%	100.0%	97.3%	91.3%	86.6%	100.0%	75.3%	1.3%	0.836	0.866	100.0% / 6.7%
Anthropic	100.0%	100.0%	99.4%	100.0%	99.9%	100.0%	89.3%	89.5%	84.8%	100.0%	74.0%	3.3%	0.780	0.784	86.7% / 6.7%
Google	100.0%	100.0%	99.5%	100.0%	95.8%	100.0%	96.2%	90.5%	86.4%	99.0%	78.7%	4.0%	0.816	0.838	100.0% / 13.3%
Overall	**100.0%**	**100.0%**	**99.5%**	**100.0%**	**96.2%**	**100.0%**	**94.9%**	**90.6%**	**86.1%**	**99.6%**	**76.0%**	**2.9%**	**0.792**	**0.817**	**95.6% / 8.9%**

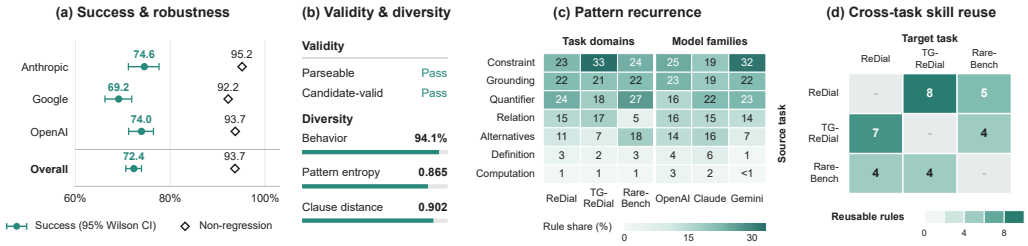


Figure 2: Evolution quality in terms of (a) success and non-regression, (b) validity and diversity, (c) error-pattern recurrence, and (d) cross-task skill reuse.

4.4 Evolution Success and Evolved-Skill Quality (RQ3)

We analyze the sealed online evaluation trajectories to assess evolution effectiveness, structural validity, behavioral diversity, and transferability, as summarized in Figure 2.

1) Evolution success. On matched examples, we compare evolved and unevolved skills using NDCG@3 for ranking and accuracy for single-answer tasks. Evolved skills win 72.4% of non-tied pairs (descriptive 95% Wilson interval: [70.7%,74.0%]) and achieve 93.7% non-regression when including ties. Success/non-regression rates are 74.6%/95.2% for Claude, 69.2%/92.2% for Gemini, and 74.0%/93.7% for OpenAI. Because some pairs share instances or settings, the interval is descriptive rather than a cluster-robust significance test. **2) Validity and diversity.** All agents equipped with evolved skills produce parseable, schema-compliant outputs and select labels exclusively from the prescribed candidate sets, establishing structural validity. On 94.1% of the examples, at least one evolved challenger differs from its unevolved counterpart. The accepted edits span seven error codes, with a normalized entropy of 0.865, and their edited clauses have a mean pairwise distance of 0.902. These measurements establish observed output and clause variation, but do not by themselves establish semantic diversity. **3) Transferability.** Across the evaluated ranking domains and backbone families, all seven evolved reasoning patterns recur, yielding cross-setting coverage of 100.0% and a mean Jensen–Shannon similarity of 83.0%. Recurrence and distributional similarity do not alone establish source-to-target transfer; Figure 2(d) therefore separately reports direct cross-task reuse among the three ranking tasks for GPT-4o-mini.

Table 3: Ablation studies on two backbones.

Backbone	Configuration	ReDial					TG-ReDial					RareBench					HLE Math	S-GPQA
		H@1	H@2	H@3	MRR	N@3	H@1	H@2	H@3	MRR	N@3	H@1	H@2	H@3	MRR	N@3	Acc.	Acc.
GPT-4o-mini	w/o documents	30.2	51.2	62.9	44.6	49.3	32.3	39.0	44.0	37.3	39.0	36.3	53.3	67.0	49.4	53.9	15.3	30.4
	w/o graph	29.3	50.8	62.4	43.9	48.6	20.3	28.6	36.1	27.0	29.3	63.9	79.9	88.2	74.6	78.1	11.8	31.4
	w/o evolution	27.0	49.9	61.8	42.4	47.4	16.7	23.2	30.1	22.3	24.3	55.2	78.2	86.7	69.5	74.0	14.1	30.8
	w/o paired comparison	25.7	47.8	61.0	41.1	46.2	15.3	24.4	33.6	22.9	25.6	62.8	79.4	87.8	73.9	77.5	14.1	31.7
	w/o localization	29.1	50.5	62.2	43.7	48.5	32.4	39.5	45.6	38.0	40.0	63.4	79.6	87.8	74.2	77.7	12.9	31.2
	w/o rollback	29.9	51.0	63.0	44.4	49.2	33.1	39.8	45.9	38.5	40.4	63.2	79.3	87.6	74.1	77.5	16.5	29.8
	w/o verifier	29.8	50.9	62.6	44.3	49.0	19.5	29.0	37.6	27.2	29.8	63.5	79.7	88.0	74.4	77.9	16.5	31.0
	w/o debate	29.3	50.4	62.5	43.9	48.6	32.5	39.4	45.6	38.0	40.0	63.1	79.4	87.5	73.9	77.4	15.3	31.2
	Full model	**30.1**	**51.4**	**63.2**	**44.7**	**49.4**	**32.8**	**39.9**	**46.1**	**38.4**	**40.4**	**63.6**	**80.0**	**88.3**	**74.6**	**78.1**	**16.5**	**31.7**
GPT-5.4-Nano	w/o documents	27.3	51.5	64.0	43.6	48.8	35.6	41.1	45.0	39.6	41.0	61.9	74.1	81.7	70.5	73.4	15.3	44.7
	w/o graph	26.9	52.1	64.5	43.6	49.0	18.7	28.1	36.1	26.1	28.7	65.7	82.4	89.0	76.3	79.5	14.1	43.6
	w/o evolution	26.7	50.8	63.8	43.1	48.4	15.3	21.3	29.0	20.9	22.9	65.9	82.1	88.8	76.2	79.5	16.5	42.8
	w/o paired comparison	20.8	45.4	61.7	38.5	44.5	11.2	18.3	25.0	17.0	19.1	64.7	80.5	87.8	75.0	78.3	18.8	44.2
	w/o localization	27.0	51.9	64.4	43.6	49.0	36.5	44.4	50.3	42.5	44.5	66.5	82.5	89.4	76.8	80.1	18.8	43.8
	w/o rollback	19.4	52.2	64.2	39.8	46.1	34.1	41.2	47.3	39.7	41.6	65.1	81.2	88.2	75.4	78.7	21.2	43.2
	w/o verifier	27.1	52.3	64.8	43.9	49.3	19.1	27.0	33.7	25.3	27.5	67.2	83.0	89.6	77.3	80.5	18.8	43.4
	w/o debate	26.8	51.6	64.3	43.5	48.8	35.7	43.8	49.8	41.8	43.8	66.8	82.5	89.5	77.0	80.2	18.8	43.6
	Full model	**27.2**	**52.8**	**65.2**	**44.1**	**49.5**	**36.3**	**44.6**	**50.6**	**42.4**	**44.5**	**67.0**	**83.2**	**89.8**	**77.3**	**80.5**	**20.0**	**44.4**

4.5 Ablation Studies (RQ4)

Table 3 confirms the complementary roles of documents and graphs, whose removal most affects RareBench and TG-ReDial, respectively. Evolution and paired comparison are critical: removing the latter reduces TG-ReDial MRR from 38.4 to 22.9 and from 42.4 to 17.0 across the two backbones. The verifier provides substantial gains, while localization, rollback, and debate offer smaller improvements. Overall, the full model delivers the strongest and most consistent ranking performance, while its advantage is clearest on evidence-grounded recommendation and diagnosis tasks.

Table 4: Task performance and LLM-call efficiency averaged across backbones.

Method	Task performance (%) ↑					Computational cost ↓		
	ReDial N@3	TG-ReDial N@3	RareBench N@3	HLE Acc.	S-GPQA Acc.	Inference calls / ex.	Total calls / ex.	Full-budget usage (%)
Debate-or-Vote (Dec.)	25.75	9.12	60.41	22.35	51.21	30.00	30.00	100.00
Debate-or-Vote (Sparse)	26.01	9.43	61.04	24.41	51.10	30.00	30.00	100.00
Debate-or-Vote (Cent.)	25.56	9.21	59.80	22.79	51.49	30.00	30.00	100.00
MATTRL	44.18	14.82	65.66	21.18	48.93	8.28	8.28	27.61
MADKE	43.35	6.90	62.04	19.12	49.69	8.29	8.29	27.64
Adaptive Judge	48.15	15.20	71.29	20.15	51.89	12.32	12.32	41.06
Identity (Anon.)	47.71	15.31	71.65	20.29	52.78	9.00	9.00	30.00
Identity (Exposed)	47.66	15.23	71.51	20.59	52.88	9.00	9.00	30.00
ED2D	46.05	10.33	67.85	21.18	48.61	16.00	16.00	53.33
DoM	43.16	10.37	69.71	18.97	47.64	9.00	9.00	30.00
EVESkillDebate (Ours)	52.47	39.25	79.18	22.94	55.63	13.77	13.88	46.28

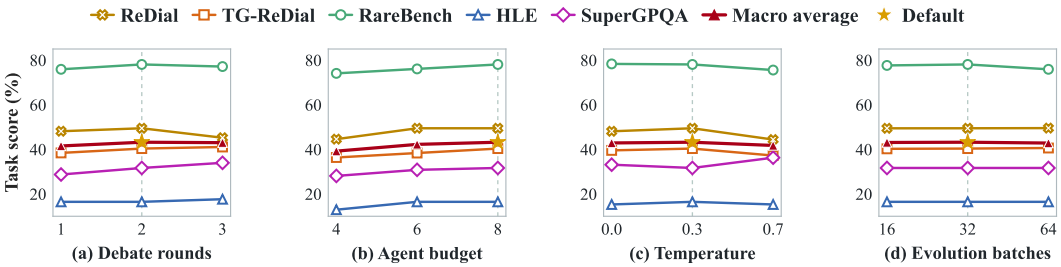


Figure 3: Hyperparameter sensitivity of EVESkillDebate.

4.6 Performance–efficiency trade-off (RQ5)

As shown in Table 4, EVESkillDebate achieves the highest performance on ReDial, TG-ReDial, RareBench, and S-GPQA, exceeding the strongest baselines by 4.32, 23.94, 7.53, and 2.75 points, respectively, while ranking second on HLE Math by 1.47 points. It requires 13.77 inference calls and 13.88 total calls, corresponding to 46.28% of the full Debate-or-Vote budget and fewer calls than ED2D and Debate-or-Vote variants. Thus, EVESkillDebate offers a favorable performance–efficiency trade-off without minimizing inference cost. Further analysis is provided in Appendix I.

4.7 Hyperparameter Sensitivity (RQ6)

Figure 3 shows stable performance near the defaults. Macro scores vary by at most 1.7% across rounds and 0.4% across batch sizes, but drop to 39.2% with four agents and 41.8% at temperature 0.7. We use two rounds, eight agents, 32 batches, and temperature 0.3. Details appear in Appendix J.

5 Conclusion

We introduced a novel paradigm EVESkillDebate, which is comprised of *Evidence-Verifiable Multi-Agent Debate* with *Self-Evolving Agent Reasoning Skill*. The former grounds candidate claims in document and graph evidence, while the latter uses successful–failed trajectories to update inference-time policies while keeping the LLM frozen. Across five datasets, our method achieves its strongest gains on evidence-grounded ranking tasks with high evidence traceability.

AI USE STATEMENT

Generative AI tools were used solely to improve writing clarity, grammar, and presentation. All research ideas, methods, experiments, analyses, and conclusions were developed and verified by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The appendix provides complete algorithmic details, dataset construction procedures, evaluation metrics, prompt templates, hyperparameters, and baseline configurations. All online experiments follow a fixed label-delayed protocol, with predictions and trajectories sealed before labels are revealed. Code, processed data, and experiment configurations are available in our anonymous repository: `https://anonymous.4open.science/r/EVESkillDebate/`.

REFERENCES

Anthropic. Models overview. In *https://platform.claude.com/docs/en/about-claude/models/overview*, 2026.

Jonas Becker, Lars Benedikt Kaesberg, Niklas Bauer, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Mallm: Multi-agent large language models framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 418–439, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-demos.29. URL https://aclanthology.org/2025.emnlp-demos.29/.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Josh Chafetz. It's the aggregation, stupid (reviewing james surowiecki, the wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations (2004), 2005.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International conference on learning representations*, volume 2024, pp. 9079–9093, 2024.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. RareBench: Can LLMs serve as rare diseases specialists? *arXiv preprint arXiv:2402.06341*, 2024b. doi: 10.48550/arXiv.2402.06341. URL https://arxiv.org/abs/2402.06341.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2508.17536.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. When identity skews debate: Anonymization for bias-reduced multi-agent reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://arxiv.org/abs/2510.07517.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate, September 2025. URL https://arxiv.org/abs/2509.11035. ICLR 2026 withdrawn submission.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, July 2024b. URL https://proceedings.mlr.press/v235/du24e.html.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024c.

Wei Fan, JinYi Yoon, and Bo Ji. iMAD: Intelligent multi-agent debate for efficient and accurate LLM inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.11306.

Peizhong Gao, Ao Xie, Shaoguang Mao, Wenshan Wu, Yan Xia, Haipeng Mi, and Furu Wei. Meta reasoning for large language models. *arXiv preprint arXiv:2406.11698*, 2024.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large Language Model based multi-agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8048–8057, 2024. doi: 10.24963/ijcai.2024/890. Survey Track.

Chen Han, Yijia Ma, Jin Tan, Wenzhen Zheng, and Xijin Tang. Beyond detection: Exploring evidence-based multi-agent debate for misinformation intervention and persuasion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.07267.

Tianyu Hu, Zhen Tan, Song Wang, Huaizhi Qu, and Tianlong Chen. Multi-agent debate for LLM judges with adaptive stability detection. *arXiv preprint arXiv:2510.12697*, 2025. doi: 10.48550/arXiv.2510.12697. URL https://arxiv.org/abs/2510.12697.

Zhiyuan Hu, Yunhai Hu, Juncheng Liu, Shuyue Stella Li, Yucheng Wang, Zhen Xu, See-Kiong Ng, Anh Tuan Luu, Xinxing Xu, Bryan Hooi, Cynthia Breazeal, and Hae Won Park. Collaborative multi-agent test-time reinforcement learning for reasoning. *arXiv preprint arXiv:2601.09667*, 2026. doi: 10.48550/arXiv.2601.09667. URL https://arxiv.org/abs/2601.09667.

Guanyu Jiang, Zhaochen Su, Xiaoye Qu, and Yi R Fung. Xskill: Continual learning from experience and skills in multimodal agents. *arXiv preprint arXiv:2603.12056*, 2026.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. *Advances in Neural Information Processing Systems*, 31, 2018.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.427. URL https://aclanthology.org/2024.findings-emnlp.427/.

Zhenghao Li, Zhi Zheng, Wei Chen, Jielun Zhao, Yong Chen, Tong Xu, and Enhong Chen. Dynadebate: Breaking homogeneity in multi-agent debate with dynamic path generation, January 2026. URL https://arxiv.org/abs/2601.05746.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024a.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Hongyi Liu, Haoyan Yang, Tao Jiang, Bo Tang, Feiyu Xiong, Yuyu Luo, and Zhiyu Li. Skillsvote: Lifecycle governance of agent skills from collection, recommendation to evolution. *arXiv preprint arXiv:2605.18401*, 2026a.

Jilong Liu, Pengyang Shao, Wei Qin, Fei Liu, Yonghui Yang, and Richang Hong. Debate over mixed-knowledge: A robust multi-agent reasoning framework for incomplete knowledge graph question answering. *arXiv preprint arXiv:2511.12208*, 2025a. doi: 10.48550/arXiv.2511.12208. URL https://arxiv.org/abs/2511.12208.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion, September 2024. URL https://arxiv.org/abs/2409.14051.

Xingyan Liu, Xiyue Luo, Linyu Li, Ganghong Huang, Jianfeng Liu, and Honglin Qiao. Skillforge: Forging domain-specific, self-evolving agent skills in cloud technical support. *arXiv preprint arXiv:2604.08618*, 2026b.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*, Singapore, April 2025b. URL https://openreview.net/forum?id=t6QHYUOQL7. ICLR 2025 Poster.

Yuxuan Liu, Zhaochen Su, Lingyun Xie, Yuhao Zhang, Qing Zong, Jiahe Guo, Zhongwei Xie, Yiyan Ji, Yauwai Yim, Hongyu Luo, et al. Skillrevise: Improving llm-authored agent skills via trace-conditioned skill revision. *arXiv preprint arXiv:2606.01139*, 2026c.

M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025. doi: 10.48550/arXiv.2502.14739. URL https://arxiv.org/abs/2502.14739.

Yuchen Ma, Yue Huang, Han Bao, Haomin Zhuang, Swadheen Shukla, Michel Galley, Xiangliang Zhang, and Stefan Feuerriegel. Skillgen: Verified inference-time agent skill synthesis. *arXiv preprint arXiv:2605.10999*, 2026.

Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158*, 2026.

OpenAI. Introducing gpt-5.5. In *https://openai.com/index/introducing-gpt-5-5*, 2026.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. doi: 10.48550/arXiv.2501.14249. URL https://arxiv.org/abs/2501.14249.

Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. Cortexdebate: Debating sparsely and equally for multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9503–9523, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.495. URL https://aclanthology.org/2025.findings-acl.495/.

Yu Tian, Jiawei Chen, Lifan Zheng, Mingxiang Tao, Xinyi Zeng, Zhaoxia Yin, Hang Su, and Xian Sun. Skills-coach: A self-evolving skill optimizer via training-free grpo. *arXiv preprint arXiv:2604.27488*, 2026.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. Multi-Agent Collaboration Mechanisms: A Survey of LLMs, jan 2025.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing*, 618:129063, 2025. doi: 10.1016/j.neucom.2024.129063. URL https://doi.org/10.1016/j.neucom.2024.129063.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of computer science*, 18(6):186345, 2024a.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 257–279, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7572–7590, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.508. URL https://aclanthology.org/2023.findings-emnlp.508/.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. Towards reasoning in large language models via multi-agent peer review collaboration, November 2023. URL https://arxiv.org/abs/2311.08152.

Haochen Yang, Ke Zhao, Mengyuan Ma, Xingyu Lu, Xiangfeng Wang, and Hong Qian. Optskills: Learning generalizable optimization skills from problem archetypes via cluster-based distillation. *arXiv preprint arXiv:2605.29829*, 2026a.

Haocheng Yang, Fengxiang Cheng, Tianjun Yao, Mengyue Yang, Jiajun Chai, Xiaohan Wang, Guojun Yin, Wei Lin, Soummya Kar, Fenrong Liu, Haoxuan Li, and Yisen Wang. MAD-Logic: Multi-agent debate enhances symbolic translation and reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=rdE9qxGfIv.

Liming Yang, Junyu Luo, Xuanzhe Liu, Yiling Lou, and Zhenpeng Chen. BAMAS: Structuring budget-aware multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026c. Oral paper.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, et al. Skillopt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026d.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15135–15153, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.936. URL https://aclanthology.org/2023.emnlp-main.936/.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. S^2-mad: Breaking the token barrier to enhance multi-agent debate efficiency. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9393–9408, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.475. URL https://aclanthology.org/2025.naacl-long.475/.

Hanrong Zhang, Shicheng Fan, Henry Peng Zou, Yankai Chen, Zhenting Wang, Jiayu Zhou, Chengze Li, Wei-Chieh Huang, Yifei Yao, Kening Zheng, et al. Coevoskills: Self-evolving agent skills via co-evolutionary verification. *arXiv preprint arXiv:2604.01687*, 2026a.

Qian Zhang, Jinyi Liu, Yan Zheng, Hebin Liang, and Lanjun Wang. Key decision-makers in multi-agent debates: who holds the power? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29883–29891, 2026b.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125*, 2020.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. Solving math word problems via cooperative reasoning induced language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4471–4485, 2023.

A NOTATION

Table 5: Notation used in EVESKILLDEBATE.

Symbol	Definition
Problem setup and reasoning skills	
$x = (q, \mathcal{C}, \mathcal{D}, \mathcal{G})$	Instance consisting of a query, candidate set, document collection, and knowledge graph.
$\mathcal{C} = \{c_1, \ldots, c_n\}$	Candidate set with $n \geq 2$.
$\mathcal{G} = (\mathcal{V}, \mathcal{R}, E_{\mathcal{G}})$	Knowledge graph with entities, relation types, and edges.
y_c	Candidate-level proposition that c satisfies the factual requirements of q.
z	Generic claim variable used only in the appendix's type-specific verifier semantics; it is instantiated as y_c in candidate-level assertions.
$[m]$	Index set $\{1, \ldots, m\}$.
K	Number of debate agents.
$\sigma : \mathcal{C} \to [n]$	Complete candidate ranking, with rank one being best.
$\mathrm{top}(\sigma)$	Top-ranked candidate $\sigma^{-1}(1)$.
$\prec_{\mathcal{C}}$	Fixed total order used to break candidate-score ties.
$s_i = (\pi_i, R_i, T_i)$	Reasoning skill of agent i, comprising inference-visible procedural clauses π_i, an evolution-only record R_i, and a tool interface T_i.
$\bar{s}_i$	Inference-visible skill $(\pi_i, \mathrm{spec}(T_i))$; R_i is excluded from the inference context.
$v, \boldsymbol{\pi}^{(v)}$	Skill-version identifier and its joint procedural profile $(\pi_1^{(v)}, \ldots, \pi_K^{(v)})$.
Evidence-path grounding	
e	Evidence unit: a document span, a knowledge-graph edge, or a replayable output produced by a registered deterministic tool.
$a = (e, y_c)$	Evidence assertion binding evidence unit e to candidate-level claim y_c.
$\nu(e)$	Binary citation-validity indicator in $\{0, 1\}$ for evidence unit e.
$\gamma(e, y_c)$	Frozen support score in $[0, 1]$ for evidence unit e and claim y_c.
$\gamma_{\neg}(e, y_c)$	Frozen contradiction score in $[0, 1]$ for evidence unit e and claim y_c.
$\mathcal{D}_i, \mathcal{G}_i, \mathcal{O}_i$	Documents, KG subgraph, and replayable tool outputs retrieved by agent i.
$\mathcal{B}_i$	Evidence units retrieved by agent i.
$\mathcal{B}$	Public evidence pool $\bigcup_{i=1}^{K} \mathcal{B}_i$.
$\mathcal{B}_i^t$	Evidence available to agent i in round t: $\mathcal{B}_i^1 = \mathcal{B}_i$ and $\mathcal{B}_i^t = \mathcal{B}$ for $t \geq 2$.
$p_i^t(c)$	Evidence path submitted by agent i for candidate c in round t.
y	Generic claim argument of the verifier, instantiated as y_c for candidate c.
$y(p)$	Unique candidate-level claim bound to the assertions in path p.
$L_{i,c}^t$	Number of evidence assertions in $p_i^t(c)$.
$\zeta_i^t(c)$	Observable explanation of how $p_i^t(c)$ supports y_c.
$\mathrm{WF}(p, y)$	Indicator that path p is well formed for claim y.
$\mathrm{Src}(p)$	Citation-validity signal in $[0, 1]$ for evidence path p.
$\mathrm{Sup}(p, y)$	Evidential-support signal in $[0, 1]$ for claim y from path p.
$\mathrm{Con}_t(p, y)$	Contradiction signal in $[0, 1]$ for path p and claim y in round t.
$\mathrm{KG}(p, y)$	Graph-consistency signal in $[0, 1]$ for path p and claim y.
$\mathcal{V}_\phi$	Frozen verifier map from four audit signals to $[0, 1]$.
$V_t(p, y)$	Verification score assigned to path p for claim y in round t.
$V_i^t(c)$	Shorthand for $V_t(p_i^t(c), y_c)$.
Evidence-guided debate and evidence-aware closing	
$\mathcal{H}^{t-1}$	Public proposals, critiques, evidence, and verifier feedback available before round t, with $\mathcal{H}^0 = \varnothing$.
F_i	Skill-conditioned candidate-scoring map used by agent i.
$B_i^t(c)$	Score in $[0, 1]$ assigned by agent i to candidate c in round t.
σ_i^t	Deterministic ranking induced by B_i^t.
$T_{\max}$	Maximum number of debate rounds.
A_i	Final audited summary of agent i's scores, paths, and verification signals.
Agg_ϕ	Frozen map that constructs A_i from agent i's final-round records.
J_ϕ	Frozen evidence-aware judge.

Symbol	Definition
$J(c)$	Final judge score in $[0, 1]$ assigned to candidate c.
L_J	Candidate list obtained by sorting $J(c)$ in descending order.
$c^\dagger$	Highest-scoring candidate under J, with ties broken by $\prec_{\mathcal{C}}$; equivalently, $c^\dagger = L_J[1]$.
k	Predeclared ranking-output cutoff in $[n]$.
$\widehat{y}(x), \widehat{y}_v(x)$	User-facing output for instance x, and the output under skill version v, respectively.
Paired execution and trajectory construction	
$\mathcal{X}_{1:M}$	Ordered sequence of label-delayed evaluation batches.
$v_b^{\mathrm{ch}}, v_b^{\mathrm{sh}}$	Champion and optional shadow skill versions in batch b.
$\perp$	Absence of an active shadow version.
$\mathfrak{V}_b$	Active versions in batch b: the champion and, when present, the shadow.
$\tau_v(x)$	Observable execution trajectory produced by version v on instance x.
$L_v(x)$	Number of observable steps in $\tau_v(x)$.
$(i_m, t_m, \alpha_m, c_m, a_m, \ell_m, \mathbf{g}_m)$	Agent, round, observable action, candidate, evidence assertion, stable clause tag, and pre-label audit features at trajectory step m.
$c^\star(x)$	Gold target revealed only after predictions and trajectories are sealed.
m_x	Predeclared task-specific evaluation metric for instance x.
$s_v(x)$	Utility $m_x(\widehat{y}_v(x), c^\star(x)) \in [0, 1]$ of version v.
$\mathcal{P}_b^{\mathrm{tr}}$	Strict-utility trajectory pairs from batch b; equal-utility comparisons are excluded.
v^+, v^-	Versions producing successful/failed (higher-/lower-utility) trajectories in a matched comparison; the terms do not imply perfect success or zero utility.
Contrastive failure localization	
$A^*(x)$	Deterministic monotone alignment between the successful and failed trajectories for instance x.
$\mathrm{Aud}_{v,m}$	Local audit score for trajectory step m, or `NA` when no audit rule applies.
η_A, δ_A	Local-audit threshold and successful–failed audit margin.
Λ_b	Stable skill-clause identifiers implicated by localized failures in batch b; diagnosed failure types are stored separately.
$r_b(u)$	Diagnosed failure type stored separately for implicated clause identifier $u \in \Lambda_b$.
$\mathcal{W}_b$	Predeclared batch indices in the active comparison window.
$\overline{\Lambda}_b$	Accumulated implicated clause identifiers $\bigcup_{s \in \mathcal{W}_b} \Lambda_s$.
$\Delta\boldsymbol{\pi}_b$	Proposed localized edit to the champion's procedural profile.
$\mathrm{Loc}(\Delta\boldsymbol{\pi}_b)$	Stable identifiers of the clauses modified by $\Delta\boldsymbol{\pi}_b$.
$H_{\max}$	Maximum number of clause-level operations allowed in one edit.
$\Delta\Phi(x)$	Fixed-length contrast representation using the sign convention successful minus failed.
ω_b	Parameters of the contrast scorer in batch b.
Feedback-guided acceptance, rollback, and freezing	
d_j	Paired candidate advantage on disagreement j: $s_{\mathrm{cand},j} - s_{\mathrm{ch},j} \in [-1, 1]$.
N	Number of sealed champion–candidate disagreements observed so far.
h, h_N	Predeclared recent-window length and available length $h_N = \min\{h, N\}$.
$\bar{d}_N, \bar{d}_N^{\mathrm{rec}}$	Cumulative and recent mean paired advantages.
G_N	Deployment summary $(N, \bar{d}_N, \bar{d}_N^{\mathrm{rec}})$.
G_N^{rb}	Corresponding summary for the rollback comparison.
$N_{\min}, N_{\min}^{\mathrm{rb}}$	Minimum disagreement counts for deployment and rollback.
H_{sh}	Predeclared maximum evaluation horizon for a shadow candidate.
$v_r^{\mathrm{cand}}, v_r^{\mathrm{prev}}$	Candidate in a deployment window and retained predecessor in a rollback window.
$\delta_{\mathrm{dep}}, \delta_{\mathrm{dep}}^{\mathrm{rec}}, \delta_{\mathrm{rb}}, \delta_{\mathrm{rb}}^{\mathrm{rec}}$	Predeclared cumulative and recent advantage margins for deployment and rollback.
$\mathrm{Gate}_{\mathrm{dep}}, \mathrm{Gate}_{\mathrm{rb}}$	Predeclared deployment and rollback gates.
v_{final}	Champion skill version frozen after the final evolution batch.

Algorithm 1 End-to-end procedure of EVESKILLDEBATE.

Require: Ordered batches $\mathcal{X}_{1:M}$; initial champion v_1^{ch}; optional shadow v_1^{sh}; round limit $T_{\max}$; frozen retriever, verifier, and judge
Ensure: Sealed champion predictions and final version v_{final}

function DEBATE(x, v)
Evidence-path grounding
  for all $i \in [K]$ **do**
    Retrieve $(\mathcal{D}_i, \mathcal{G}_i, \mathcal{O}_i)$ under $\bar{s}_i^{(v)}$ and form $\mathcal{B}_i$
  end for
  $\mathcal{B} \leftarrow \bigcup_{i=1}^{K} \mathcal{B}_i$ and $\mathcal{H}^0 \leftarrow \varnothing$
  for all $i \in [K]$ **do**
    for all $c \in \mathcal{C}$ **do**
      Construct $p_i^1(c)$ and compute $V_i^1(c)$ and $B_i^1(c)$
    end for
    Compute σ_i^1 from $\{B_i^1(c) : c \in \mathcal{C}\}$
  end for
  Seal and reveal the first-round messages synchronously
Evidence-guided debate
  for $t = 2, \ldots, T_{\max}$ **do**
    for all $i \in [K]$ **do**
      for all $c \in \mathcal{C}$ **do**
        Construct $p_i^t(c)$ using $\mathcal{B}$ and $\mathcal{H}^{t-1}$; compute $V_i^t(c)$ and $B_i^t(c)$
      end for
      Compute σ_i^t from $\{B_i^t(c) : c \in \mathcal{C}\}$
    end for
    Seal and reveal the round-t messages synchronously
  end for
Evidence-aware closing
  Compute $\{A_i\}_{i=1}^{K}$, $\{J(c)\}_{c \in \mathcal{C}}$, L_J, and $\widehat{y}_v(x)$ using Eqs. equation 5–equation 7
  Finalize $\tau_v(x)$ and **return** $(\widehat{y}_v(x), \tau_v(x))$
end function
for $b = 1, \ldots, M$ **do**
  Freeze the batch state and set $(v_{b+1}^{\mathrm{ch}}, \omega_{b+1}) \leftarrow (v_b^{\mathrm{ch}}, \omega_b)$
  $\mathfrak{V}_b \leftarrow \{v_b^{\mathrm{ch}}\}$ if $v_b^{\mathrm{sh}} = \perp$; otherwise $\mathfrak{V}_b \leftarrow \{v_b^{\mathrm{ch}}, v_b^{\mathrm{sh}}\}$
Paired execution
  for all $x \in \mathcal{X}_b$ and $v \in \mathfrak{V}_b$ **do**
    $(\widehat{y}_v(x), \tau_v(x)) \leftarrow$ DEBATE(x, v) under matched conditions
  end for
  Seal all records; release champion outputs; then reveal labels
  if $v_b^{\mathrm{sh}} \neq \perp$ **then**
    Form strict-utility successful/failed pairs $\mathcal{P}_b^{\mathrm{tr}}$
Contrastive failure localization
    if $\mathcal{P}_b^{\mathrm{tr}} \neq \varnothing$ **then**
      Align the paired trajectories and update ω_{b+1}
      if some pair has $v^- = v_b^{\mathrm{ch}}$ **then**
        Localize champion failures and enqueue edits satisfying Eq. equation 10
      end if
    end if
Feedback-guided acceptance, rollback, and freezing
    Update the controller and v_{b+1}^{ch} using Eq. equation 41
  end if
  Schedule at most one comparator for batch $b + 1$
end for
$v_{\mathrm{final}} \leftarrow v_{M+1}^{\mathrm{ch}}$ and freeze it
return sealed champion predictions and v_{final}

B ALGORITHM

Algorithm 1 provides the end-to-end procedure of our proposed EVESKILLDEBATE.

C Technical Definitions and Protocol Details

This appendix specifies the objects introduced in Section 3. It follows the effective main-paper formulation: candidate-level evidence verification, evidence-aware closing, strict-utility trajectory pairing, and empirical feedback gates. No gold label or evolution-only record enters inference for the batch in which it is produced.

C.1 Evidence Semantics

Write $e = (\mathrm{type}(e), \mathrm{payload}(e))$, where $\mathrm{type}(e) \in \{\mathrm{doc}, \mathrm{kg}, \mathrm{tool}\}$. A document payload is (d, ξ) with exact span ξ of d; a KG payload is an edge $g = (u, r, v)$; and a tool payload is $(T, \mathrm{prog}, \mathrm{out})$, where prog contains canonicalized inputs. Citation validity is

$$\nu(e) = \begin{cases} \mathbf{1}[d \in \mathcal{D} \ \wedge \ \xi \sqsubseteq d], & \mathrm{type}(e) = \mathrm{doc}, \\ \mathbf{1}[g \in E_{\mathcal{G}}], & \mathrm{type}(e) = \mathrm{kg}, \\ \mathbf{1}[T \in \mathfrak{T} \ \wedge \mathrm{Exec}_T(\mathrm{prog}) = \mathrm{out}], & \mathrm{type}(e) = \mathrm{tool}, \end{cases} \tag{12}$$

where $\mathfrak{T}$ is the registry of deterministic tools with fixed semantics. Let $M_{\mathcal{G}}(g, z)$ and $M_T(\mathrm{prog}, \mathrm{out}, z)$ take values in $\{\mathrm{ent}, \mathrm{con}, \mathrm{unk}\}$. Define

$$\gamma(e, z) = \begin{cases} P_{\mathrm{NLI}}(\mathrm{entail} \mid \xi, z), & \mathrm{type}(e) = \mathrm{doc}, \\ \mathrm{conf}(g)\mathbf{1}[M_{\mathcal{G}}(g, z) = \mathrm{ent}], & \mathrm{type}(e) = \mathrm{kg}, \\ \mathbf{1}[M_T(\mathrm{prog}, \mathrm{out}, z) = \mathrm{ent}], & \mathrm{type}(e) = \mathrm{tool}, \end{cases} \tag{13}$$

and

$$\gamma_{\neg}(e, z) = \begin{cases} P_{\mathrm{NLI}}(\mathrm{contradict} \mid \xi, z), & \mathrm{type}(e) = \mathrm{doc}, \\ \mathrm{conf}(g)\mathbf{1}[M_{\mathcal{G}}(g, z) = \mathrm{con}], & \mathrm{type}(e) = \mathrm{kg}, \\ \mathbf{1}[M_T(\mathrm{prog}, \mathrm{out}, z) = \mathrm{con}], & \mathrm{type}(e) = \mathrm{tool}. \end{cases} \tag{14}$$

Here $\mathrm{conf}(g) \in [0, 1]$. An absent KG edge is treated as unknown rather than contradictory. Tool scores are exact only relative to registered semantics and successful replay. All probabilistic and symbolic components are frozen before evaluation; the resulting values are operational verification signals rather than probabilities of real-world truth.

C.2 Evidence-Path Grounding

A nonempty path for candidate c has the form $p = ((e_l, y_c)_{l=1}^{L}; \zeta)$, with $1 \le L \le L_{\max}$. Thus $y(p) = y_c$ is uniquely determined by the assertions stored in p. Let $\mathrm{WF}(p, y)$ indicate that the path has valid syntax, contains no duplicate citation identifiers, and that ζ references only evidence present in p. Define

$$\begin{aligned} \mathrm{Src}(p) &= \mathbf{1}[\mathrm{WF}(p, y(p))] \min_{l \in [L]} \nu(e_l), \\ \mathrm{Sup}(p, y) &= \mathbf{1}[\mathrm{WF}(p, y)] \left(\prod_{l=1}^{L} \gamma(e_l, y) \right)^{1/L}. \end{aligned} \tag{15}$$

At round t, let $\mathcal{K}_{\le t}(y)$ be the set of submitted counterevidence assertions for the negation of y. Their strongest valid challenge is

$$\mathrm{Chal}_t(y) = \max_{\widetilde{e} \in \mathcal{K}_{\le t}(y)} \nu(\widetilde{e}) \min\{\gamma(\widetilde{e}, \neg y), \gamma_{\neg}(\widetilde{e}, y)\}, \qquad \max \varnothing := 0. \tag{16}$$

The path-level contradiction factor is

$$\mathrm{Con}_t(p, y) = \max\left\{ \mathrm{Chal}_t(y), \max_{l \in [L]} \gamma_{\neg}(e_l, y) \right\}. \tag{17}$$

Let $\mathrm{Req}_{\mathcal{G}}(p, y)$ indicate that p or ζ invokes a graph relation. Graph consistency is

$$\mathrm{KG}(p, y) = \begin{cases} 1, & \mathrm{Req}_{\mathcal{G}}(p, y) = 0, \\ 0, & \mathrm{Req}_{\mathcal{G}}(p, y) = 1 \ \wedge \ p \text{ contains no KG edge}, \\ \mathrm{Path}_{\mathcal{G}}(p, y), & \text{otherwise}, \end{cases} \tag{18}$$

where $\mathrm{Path}_{\mathcal{G}} \in \{0,1\}$ checks edge membership, endpoint connectivity, relation direction, and consistency with ζ. The verifier in Eq. (3) is therefore

$$V_t(p,y) = \mathcal{V}_\phi(\mathrm{Src}(p), \mathrm{Sup}(p,y), 1-\mathrm{Con}_t(p,y), \mathrm{KG}(p,y)), \quad V_t(\varnothing, y) = 0, \tag{19}$$

with $\mathcal{V}_\phi : [0,1]^4 \to [0,1]$ frozen before evaluation. We impose the structural boundary condition $\mathcal{V}_\phi(s_{\mathrm{src}}, s_{\mathrm{sup}}, s_{\mathrm{con}}, s_{\mathrm{kg}}) = 0$ whenever $s_{\mathrm{src}} = 0$ or $s_{\mathrm{sup}} = 0$, so a malformed, unresolvable, or unsupported path cannot receive positive verification solely from the remaining inputs. The four inputs are verifier signals; no independence or probabilistic-factorization assumption is made.

C.3 Evidence-Guided Debate

At round t, agent i submits

$$m_i^t = \big((p_i^t(c), B_i^t(c))_{c\in\mathcal{C}}, r_i^t, \kappa_i^t\big), \tag{20}$$

where r_i^t is the observable rationale and κ_i^t contains claim-bound counterevidence. Every factual citation that influences a proposal must occur in a path or critique. Messages are sealed before simultaneous reveal. The public history is updated by

$$\mathcal{H}^t = \mathcal{H}^{t-1} \cup \{m_i^t, V_i^t, \sigma_i^t : i \in [K]\}. \tag{21}$$

For the next round, F_i may use the query, candidate, agent-specific evidence, public evidence and rationales in $\mathcal{H}^t$, and verifier feedback. It receives no gold label. Agreement has no privileged numerical update: it can affect later reasoning only through the same recorded history available to the other debate inputs. Document citations for candidate c are induced directly by its submitted path:

$$Z_i^t(c) = \{(d,\xi) : (\mathrm{doc},(d,\xi)) \text{ occurs in } p_i^t(c)\}. \tag{22}$$

C.4 Evidence-Aware Closing

At $T_{\max}$, the frozen aggregation maps used in Eq. (5) are evaluated once:

$$\begin{aligned} A_i &= \mathrm{Agg}_\phi\big(\{(B_i^{T_{\max}}(c), V_i^{T_{\max}}(c), p_i^{T_{\max}}(c)) : c \in \mathcal{C}\}\big), \\ J(c) &= J_\phi\Big(q, c, \mathcal{B}, \{(B_i^{T_{\max}}, \sigma_i^{T_{\max}}, A_i, p_i^{T_{\max}})\}_{i=1}^K\Big). \end{aligned} \tag{23}$$

Neither map has access to the gold label. $J(c)$ may favor a better-supported minority proposal because inter-agent agreement is an input rather than a hard activation condition. The complete deterministic list is $L_J = \mathrm{Sort}^{\downarrow}_{\prec_{\mathcal{C}}}\{c \mapsto J(c) : c \in \mathcal{C}\}$, and $c^\dagger = L_J[1]$. Consistent with Eq. (7), a ranking task returns the predeclared prefix $L_J[1\!:\!k]$, whereas a single-answer task returns $c^\dagger$. When skill version v is active, this output is denoted by $\widehat{y}_v(x)$.

For every returned candidate c, define its inspectable evidence record as

$$\mathcal{E}_J(c) = \bigcup_{i=1}^K \{(e, y_c) : e \text{ occurs in } p_i^{T_{\max}}(c)\}, \tag{24}$$

together with the associated ν, γ, $\gamma_\neg$, and graph-path checks. This record makes the decision traceable to the retrieved substrate; it does not guarantee that the substrate itself is complete or correct.

C.5 Paired Execution, Sealing, and Trace Schema

Write $\pi_i = ((u_{i,1}, \varphi_{i,1}), \ldots, (u_{i,M_i}, \varphi_{i,M_i}))$, where $u_{i,m}$ is a stable clause identifier and $\varphi_{i,m}$ is its text. Rewriting preserves an identifier, deletion retires it, and insertion creates a fresh identifier. An observable action tag satisfies

$$\ell(\alpha_{i,k}) \in \mathcal{I}_i(v) \cup \{\varnothing\}, \tag{25}$$

where $\mathcal{I}_i(v)$ is agent i's set of active clause identifiers in version v; controller actions use $\varnothing$. In the trajectory notation of Eq. (8), $\ell_m = \ell(\alpha_m)$.

Let $\mathcal{F}_{b-1}$ contain all sealed executions and labels revealed through batch $b-1$. Before batch b, the controller state

$$\Xi_b = (v_b^{\mathrm{ch}}, v_b^{\mathrm{sh}}, \omega_b, \mathbf{R}_b, \mathfrak{S}_b) \quad \text{is } \mathcal{F}_{b-1}\text{-measurable} \tag{26}$$

and remains frozen within $\mathcal{X}_b$. Here $\mathbf{R}_b$ collects the evolution-only records and is excluded from all inference contexts, while $\mathfrak{S}_b$ records the serialized controller phase, retained predecessor, and queued candidate edits. The active versions are

$$\mathfrak{V}_b = \begin{cases} \{v_b^{\mathrm{ch}}, v_b^{\mathrm{sh}}\}, & v_b^{\mathrm{sh}} \neq \perp, \\ \{v_b^{\mathrm{ch}}\}, & v_b^{\mathrm{sh}} = \perp. \end{cases} \tag{27}$$

Champion and shadow agents communicate only within their own version.

For each step, the frozen feature constructor records only information available before label release:

$$\mathbf{g}_m = G_\phi(i_m, t_m, \alpha_m, c_m, a_m, \ell_m, V_m, B_m, \nu_m, \gamma_m, \gamma_{\neg,m}, \mathrm{KG}_m) \in \mathbb{R}^{d_g}, \tag{28}$$

with a mask for inapplicable fields. G_ϕ and d_g are fixed before evaluation. The vector contains no label, task score, or regenerated private chain of thought.

The controller atomically seals

$$\mathcal{Z}_b = \mathrm{Seal}(\mathrm{id}(\Xi_b), \{(u_b(x), \{(\mathrm{id}(v), \tau_v(x)) : v \in \mathfrak{V}_b\}) : x \in \mathcal{X}_b\}), \tag{29}$$

where $u_b(x)$ records the matched execution conditions and a shared random seed when supported. Labels Y_b are revealed only afterward, and

$$\mathcal{F}_b = \sigma(\mathcal{F}_{b-1}, \mathcal{X}_b, \mathcal{Z}_b, Y_b). \tag{30}$$

Here $\sigma(\cdot)$ denotes the generated sigma-field, not a ranking.

After label release, $s_v(x) = m_x(\widehat{y}_v(x), c^\star(x))$. If a distinct shadow is active, the strict-utility pair set is

$$\begin{aligned} \mathcal{P}_b^{\mathrm{tr}} = \big\{ & (x, \tau_{v^+}(x), \tau_{v^-}(x)) : x \in \mathcal{X}_b, \\ & \{v^+, v^-\} = \{v_b^{\mathrm{ch}}, v_b^{\mathrm{sh}}\}, \quad s_{v^+}(x) > s_{v^-}(x) \big\}. \end{aligned} \tag{31}$$

The set is empty without a shadow, and equal-score cases have no directional pair. We call the two members successful/failed (i.e., higher-/lower-utility) trajectories; these labels are relative to a matched instance and do not require perfect correctness or zero utility.

C.6 Contrastive Failure Localization and Editing

Let $\mathbf{h}_{v,m} = \mathrm{Enc}_\phi(\alpha_m, a_m)$ be a unit-normalized observable-action embedding from an encoder frozen before evaluation. For successful step j and failed step k, define

$$\begin{aligned} \mathrm{sim}_{jk} = \mathbf{1}[i_j^+ = i_k^-] \Big(& \lambda_\ell \mathbf{1}[\ell_j^+ = \ell_k^- \ \wedge \ \ell_j^+ \neq \varnothing] + \lambda_c \mathbf{1}[c_j^+ = c_k^-] \\ & + \lambda_h \frac{1 + (\mathbf{h}_j^+)^\top \mathbf{h}_k^-}{2} + \lambda_g \exp(-\|\mathbf{g}_j^+ - \mathbf{g}_k^-\|_1 / d_g) \Big), \end{aligned} \tag{32}$$

where nonnegative coefficients sum to one. Let $\mathfrak{A}(\tau^+, \tau^-)$ contain monotone one-to-one partial alignments whose matched pairs satisfy a fixed similarity threshold ρ. The selected alignment is

$$A^*(x) = \mathrm{TB}_{\prec_A} \operatorname*{arg\,max}_{A \in \mathfrak{A}(\tau^+, \tau^-)} \left\{ \sum_{(j,k) \in A} \mathrm{sim}_{jk} - \lambda_{\mathrm{gap}}(L^+ + L^- - 2|A|) \right\}. \tag{33}$$

All thresholds and tie rules are fixed before evaluation.

Local audits use verifier outputs rather than task labels:

$$\mathrm{Aud}_{v,m} = \begin{cases} \nu(e_m)\gamma(e_m, y_{c_m})(1 - \gamma_\neg(e_m, y_{c_m})), & a_m = (e_m, y_{c_m}), \\ V_{t_m}(p_m, y_{c_m}), & m \text{ records a path-verification step}, \\ B_{i_m}^{t_m}(c_m), & m \text{ records a candidate-scoring step}, \\ \texttt{NA}, & \text{otherwise}. \end{cases} \tag{34}$$

For an aligned pair, a failed step is implicated when both audits exist, the successful audit is at least η_A, and its margin over the failed audit is at least δ_A. An unmatched failed step is implicated only

when it has an observed audit below η_A. An unmatched successful step may justify an insertion only when its observed audit is at least η_A and a deterministic neighboring failed-trace anchor with a nonempty clause tag exists. Missing audits are never converted into failures.

Let $\mathcal{E}_{\mathrm{al}}(x)$, $\mathcal{E}_{\mathrm{gap}}(x)$, and $\mathcal{E}_{\mathrm{miss}}(x)$ denote these three sets. With a frozen error coder Err_ϕ, define the typed localization records

$$\Lambda_b = \bigcup_{(x,\tau^+,\tau^-)\in\mathcal{P}_b^{\mathrm{tr}}:\, v^-=v_b^{\mathrm{ch}}} \{(u, \mathrm{Err}_\phi(x,u)) : u \text{ is tagged by } \mathcal{E}_{\mathrm{al}}(x)\cup\mathcal{E}_{\mathrm{gap}}(x)\cup\mathcal{E}_{\mathrm{miss}}(x)\}\,. \quad (35)$$

Thus each element of Λ_b is a pair (u, r) consisting of a stable skill-clause location and its diagnosed failure type. Accordingly, $\mathrm{Loc}(\Delta\boldsymbol{\pi}_b)$ in Eq. (10) denotes the same type of pair, making the subset relation well typed. Pairs in which the shadow is the failed member can train the contrast scorer but cannot authorize edits to champion clauses.

For a fixed number of position bins R, the aligned diagnostics are pooled into a fixed-length vector $\Delta\Phi(x) \in \mathbb{R}^{d_\Phi}$ using the sign convention successful minus failed; matched differences encode degraded behavior, while unmatched positive and negative terms encode omitted and extraneous behavior. The scorer is updated only after batch b:

$$\omega_{b+1} = \arg\min_{\omega\in\mathbb{R}^{d_\Phi}} \left\{ \frac{1}{|\mathcal{P}_b^{\mathrm{tr}}|} \sum_{(x,\tau^+,\tau^-)\in\mathcal{P}_b^{\mathrm{tr}}} \log\Big(1 + e^{-\omega^\top \Delta\Phi(x)}\Big) + \lambda_\omega \|\omega - \omega_b\|_2^2 \right\}, \quad (36)$$

when $\mathcal{P}_b^{\mathrm{tr}} \neq \varnothing$; otherwise $\omega_{b+1} = \omega_b$. The fixed $\lambda_\omega > 0$ makes the objective strongly convex. Labels determine pair direction and do not enter $\Delta\Phi(x)$.

Let $\mathcal{W}_b \subseteq [b]$ be the predeclared set of batch indices in the active comparison window. The accumulated localization records are $\overline{\Lambda}_b = \bigcup_{s\in\mathcal{W}_b} \Lambda_s$. The edit generator may rank valid diagnostics using ω_{b+1} but cannot create new edit locations. Every candidate edit must satisfy Eq. (10), preserve the registered tool interfaces, and match the diagnosed operation type. Stale, overlapping, syntactically invalid, or unlocalized edits are rejected. A valid edit can first run as a shadow on a later batch.

C.7 Feedback-Guided Acceptance, Rollback, and Freezing

For a frozen champion–candidate comparison, the candidate is the active shadow version in an acceptance window. Define the pre-label disagreement set in batch r as

$$\mathcal{U}_r = \{x \in \mathcal{X}_r : \widehat{y}_{\mathrm{cand}}(x) \neq \widehat{y}_{\mathrm{ch}}(x)\}. \quad (37)$$

Membership is sealed before labels are observed. Fix a predeclared recent window length $h \geq 1$, index accumulated disagreements by j, and set

$$d_j = s_{\mathrm{cand},j} - s_{\mathrm{ch},j}, \quad \bar{d}_N = \frac{1}{N}\sum_{j=1}^{N} d_j, \quad h_N = \min\{h, N\}, \quad \bar{d}_N^{\mathrm{rec}} = \frac{1}{h_N}\sum_{j=N-h_N+1}^{N} d_j. \quad (38)$$

The empirical summary is $G_N = (N, \bar{d}_N, \bar{d}_N^{\mathrm{rec}})$. For predeclared nonnegative margins $\delta_{\mathrm{dep}}, \delta_{\mathrm{dep}}^{\mathrm{rec}}$, a concrete deployment rule is

$$\mathrm{Gate}_{\mathrm{dep}}(G_N) = \mathbf{1}[N \geq N_{\min} \ \wedge\ \bar{d}_N > \delta_{\mathrm{dep}} \ \wedge\ \bar{d}_N^{\mathrm{rec}} > \delta_{\mathrm{dep}}^{\mathrm{rec}}]. \quad (39)$$

If the gate is not met within the predeclared shadow horizon H_{sh}, the candidate is rejected and the champion is unchanged.

After acceptance and deployment, the former champion is retained as the predecessor and used as the shadow comparator in a fresh rollback window. Define $d_j^{\mathrm{rb}} = s_{\mathrm{dep},j} - s_{\mathrm{prev},j}$ and let $\bar{d}_N^{\mathrm{rb}}$ and $\bar{d}_N^{\mathrm{rb,rec}}$ denote its cumulative and recent means. Write $G_N^{\mathrm{rb}} = (N, \bar{d}_N^{\mathrm{rb}}, \bar{d}_N^{\mathrm{rb,rec}})$. For nonnegative rollback margins $\delta_{\mathrm{rb}}, \delta_{\mathrm{rb}}^{\mathrm{rec}}$, rollback is triggered by

$$\mathrm{Gate}_{\mathrm{rb}}(G_N^{\mathrm{rb}}) = \mathbf{1}[N \geq N_{\min}^{\mathrm{rb}} \ \wedge\ \bar{d}_N^{\mathrm{rb}} < -\delta_{\mathrm{rb}} \ \wedge\ \bar{d}_N^{\mathrm{rb,rec}} < -\delta_{\mathrm{rb}}^{\mathrm{rec}}]. \quad (40)$$

Deployment and rollback windows are serialized, so exactly one of the two gates is applicable at any controller update. The deployed version changes only through

$$v_{r+1}^{\mathrm{ch}} = \begin{cases} v_r^{\mathrm{cand}}, & \mathrm{Gate}_{\mathrm{dep}}(G_N) = 1, \\ v_r^{\mathrm{prev}}, & \mathrm{Gate}_{\mathrm{rb}}(G_N^{\mathrm{rb}}) = 1, \\ v_r^{\mathrm{ch}}, & \text{otherwise.} \end{cases} \quad (41)$$

These are predeclared empirical rules, not confidence sequences; they provide no time-uniform or distribution-free improvement guarantee. In this paper, "monotonic" means that an edit cannot become the champion without satisfying the deployment gate and that rollback restores only the retained predecessor. After batch M, $v_{\text{final}} = v_{M+1}^{\text{ch}}$ is frozen. Online metrics retain the predictions sealed before their labels were revealed; past instances are never rescored under a later version.

D Results Across Additional Backbone Families

Tables 6–7 provide the complete results for OpenAI, Gemini, and Claude backbones. The clearest cross-family pattern appears on the candidate-ranking tasks: among methods returning ranked prefixes through depth three, EVESkillDebate has the highest reported MRR, NDCG@3, and H@2 point estimates for every displayed backbone–task combination. This pattern is not confined to one cutoff or one provider, although the tables alone do not establish statistical significance.

OpenAI backbones. Table 6 shows consistent improvements across GPT-4o-mini, GPT-5.4-Nano, and GPT-5.4-Mini. On ReDial, the gains over the strongest backbone-specific baselines range from 0.019 to 0.046 in MRR and from 0.013 to 0.033 in NDCG@3. Although the proposed method is slightly below the strongest baseline on H@3, it achieves higher H@1, H@2, MRR, and NDCG@3 under all three backbones, indicating better ordering of the highest ranked candidates. The advantage is substantially larger on TG-ReDial, with absolute gains of 0.239–0.284 in MRR and 0.245–0.286 in NDCG@3. RareBench exhibits a similarly consistent pattern: MRR improves by 0.041–0.118, while NDCG@3 improves by 0.050–0.112. The largest OpenAI-family margins therefore occur when candidate selection combines multiple query constraints with heterogeneous evidence.

Table 6: Results across OpenAI backbones.

Model	Method	ReDial					TG-ReDial					RareBench					HLE	SGPQA
		H@1	H@2	H@3	MRR	N@3	H@1	H@2	H@3	MRR	N@3	H@1	H@2	H@3	MRR	N@3	Acc.	Acc.
GPT-4o-mini	Classic MAD	0.224	0.459	0.639	0.401	0.462	0.054	0.119	0.175	0.105	0.123	0.527	0.625	0.754	0.619	0.653	0.129	0.314
	Debate-or-Vote (Dec.)	0.274	0.275	0.275	0.274	0.274	0.063	0.063	0.063	0.063	0.063	0.571	0.583	0.583	0.577	0.579	0.071	0.346
	Debate-or-Vote (Sparse)	0.264	0.278	0.278	0.271	0.273	0.067	0.075	0.075	0.071	0.072	0.576	0.601	0.603	0.589	0.593	0.118	0.348
	Debate-or-Vote (Cent.)	0.279	0.279	0.279	0.279	0.279	0.063	0.063	0.063	0.063	0.063	0.555	0.555	0.555	0.555	0.555	0.094	0.333
	MATTRL	0.214	0.428	0.618	0.384	0.444	0.079	0.134	0.189	0.125	0.141	0.489	0.600	0.714	0.582	0.616	0.094	0.302
	MADKE	0.207	0.471	0.642	0.396	0.459	0.020	0.033	0.075	0.041	0.049	0.422	0.530	0.668	0.522	0.559	0.153	0.283
	Adaptive Judge	0.217	0.432	0.618	0.387	0.446	0.062	0.114	0.170	0.106	0.123	0.492	0.611	0.722	0.589	0.623	0.082	0.291
	iMAD	0.225	0.431	0.596	0.383	0.437	0.063	0.123	0.185	0.114	0.132	0.474	0.599	0.719	0.576	0.613	0.118	0.310
	Identity (Anon.)	0.226	0.461	0.634	0.401	0.461	0.058	0.126	0.182	0.110	0.129	0.515	0.631	0.742	0.610	0.644	0.094	0.289
	Identity (Exposed)	0.225	0.465	0.632	0.400	0.460	0.062	0.116	0.171	0.107	0.124	0.511	0.633	0.751	0.611	0.647	0.106	0.302
	ED2D	0.217	0.432	0.586	0.376	0.430	0.033	0.083	0.154	0.082	0.100	0.515	0.629	0.742	0.610	0.644	0.106	0.296
	MAD-Logic	0.195	0.429	0.586	0.364	0.421	0.039	0.088	0.146	0.083	0.099	0.418	0.568	0.712	0.541	0.585	0.118	0.266
	DoM	0.211	0.414	0.574	0.365	0.419	0.052	0.104	0.162	0.098	0.114	0.507	0.674	0.787	0.628	0.669	0.118	0.281
	EVESkillDebate (Ours)	0.301	0.514	0.632	0.447	0.495	0.328	0.399	0.461	0.384	0.404	0.636	0.800	0.883	0.746	0.781	0.165	0.317
GPT-5.4 Nano	Classic MAD	0.239	0.458	0.639	0.409	0.468	0.084	0.161	0.213	0.140	0.159	0.636	0.692	0.762	0.687	0.706	0.212	0.428
	Debate-or-Vote (Dec.)	0.249	0.250	0.250	0.250	0.250	0.079	0.079	0.079	0.079	0.079	0.641	0.641	0.641	0.641	0.641	0.141	0.457
	Debate-or-Vote (Sparse)	0.260	0.265	0.266	0.263	0.263	0.078	0.080	0.082	0.079	0.080	0.631	0.647	0.647	0.639	0.641	0.176	0.449
	Debate-or-Vote (Cent.)	0.257	0.257	0.257	0.257	0.257	0.082	0.082	0.082	0.082	0.082	0.624	0.624	0.624	0.624	0.624	0.165	0.468
	MATTRL	0.248	0.480	0.654	0.422	0.481	0.074	0.138	0.201	0.127	0.146	0.609	0.678	0.765	0.672	0.696	0.259	0.396
	MADKE	0.230	0.444	0.587	0.385	0.436	0.036	0.064	0.094	0.060	0.069	0.561	0.648	0.749	0.638	0.666	0.129	0.411
	Adaptive Judge	0.240	0.457	0.629	0.406	0.463	0.070	0.138	0.198	0.124	0.143	0.631	0.685	0.765	0.685	0.705	0.200	0.407
	iMAD	0.236	0.448	0.628	0.402	0.460	0.079	0.135	0.179	0.122	0.137	0.611	0.683	0.764	0.674	0.697	0.176	0.392
	Identity (Anon.)	0.228	0.454	0.638	0.402	0.462	0.075	0.135	0.195	0.125	0.143	0.628	0.698	0.785	0.692	0.716	0.212	0.417
	Identity (Exposed)	0.235	0.447	0.629	0.402	0.460	0.072	0.141	0.207	0.129	0.149	0.615	0.686	0.778	0.681	0.706	0.212	0.400
	ED2D	0.230	0.433	0.580	0.380	0.431	0.033	0.076	0.129	0.072	0.087	0.567	0.655	0.744	0.641	0.667	0.165	0.342
	MAD-Logic	0.199	0.395	0.560	0.352	0.405	0.023	0.059	0.094	0.052	0.063	0.590	0.631	0.714	0.638	0.657	0.176	0.390
	DoM	0.179	0.329	0.435	0.289	0.327	0.036	0.060	0.100	0.062	0.071	0.542	0.627	0.706	0.611	0.635	0.118	0.306
	EVESkillDebate (Ours)	0.272	0.528	0.652	0.441	0.495	0.363	0.446	0.506	0.424	0.445	0.670	0.832	0.898	0.773	0.805	0.200	0.444
GPT-5.4 Mini	Classic MAD	0.258	0.521	0.694	0.447	0.510	0.099	0.155	0.187	0.138	0.151	0.660	0.717	0.787	0.712	0.731	0.200	0.593
	Debate-or-Vote (Dec.)	0.280	0.282	0.282	0.281	0.281	0.094	0.094	0.094	0.094	0.094	0.680	0.681	0.681	0.680	0.680	0.318	0.597
	Debate-or-Vote (Sparse)	0.269	0.284	0.284	0.277	0.279	0.092	0.096	0.096	0.094	0.095	0.682	0.687	0.687	0.685	0.685	0.329	0.570
	Debate-or-Vote (Cent.)	0.275	0.275	0.275	0.275	0.275	0.091	0.091	0.091	0.091	0.091	0.675	0.675	0.675	0.675	0.675	0.282	0.579
	MATTRL	0.259	0.506	0.684	0.442	0.504	0.082	0.137	0.189	0.127	0.142	0.652	0.695	0.772	0.699	0.717	0.247	0.566
	MADKE	0.256	0.478	0.621	0.415	0.468	0.032	0.056	0.076	0.051	0.057	0.634	0.669	0.720	0.668	0.682	0.188	0.526
	Adaptive Judge	0.257	0.504	0.687	0.442	0.504	0.087	0.150	0.189	0.131	0.146	0.670	0.710	0.769	0.709	0.725	0.247	0.539
	iMAD	0.271	0.499	0.686	0.447	0.508	0.074	0.137	0.190	0.123	0.140	0.675	0.715	0.782	0.717	0.734	0.259	0.556
	Identity (Anon.)	0.265	0.508	0.684	0.445	0.506	0.092	0.149	0.193	0.135	0.150	0.668	0.711	0.774	0.710	0.727	0.235	0.581
	Identity (Exposed)	0.252	0.504	0.687	0.439	0.502	0.079	0.129	0.187	0.123	0.140	0.674	0.732	0.784	0.721	0.737	0.176	0.577
	ED2D	0.251	0.482	0.659	0.425	0.485	0.055	0.087	0.130	0.085	0.097	0.655	0.720	0.793	0.712	0.733	0.235	0.532
	MAD-Logic	0.259	0.482	0.654	0.428	0.486	0.054	0.090	0.133	0.086	0.098	0.633	0.684	0.756	0.683	0.701	0.224	0.526
	DoM	0.235	0.439	0.605	0.392	0.447	0.041	0.075	0.107	0.069	0.079	0.660	0.707	0.765	0.703	0.719	0.235	0.505
	EVESkillDebate (Ours)	0.312	0.527	0.677	0.470	0.523	0.323	0.394	0.450	0.377	0.396	0.684	0.809	0.856	0.762	0.787	0.247	0.606

Gemini backbones. As reported in Table 7, EVESkillDebate has the strongest reported point estimates across all ranking metrics on ReDial and TG-ReDial for Gemini 2.5 Flash, Gemini 3.5 Flash-Lite, and Gemini 3.6 Flash. On ReDial, the improvements over the strongest baselines range from 0.033 to 0.067 in MRR and from 0.025 to 0.064 in NDCG@3. On TG-ReDial, the corresponding gains increase to 0.160–0.222 and 0.163–0.226, respectively. On RareBench, the proposed method

Table 7: Results across Gemini and Claude backbones.

Model	Method	ReDial					TG-ReDial					RareBench					HLE	SGPQA
		H@1	H@2	H@3	MRR	N@3	H@1	H@2	H@3	MRR	N@3	H@1	H@2	H@3	MRR	N@3	Acc.	Acc.
Gemini 2.5 Flash	Classic MAD	0.111	0.205	0.277	0.182	0.206	0.071	0.124	0.186	0.118	0.136	0.237	0.403	0.523	0.360	0.402	0.082	0.354
	Debate-or-Vote (Dec.)	0.068	0.092	0.100	0.082	0.087	0.067	0.092	0.100	0.082	0.087	0.195	0.224	0.224	0.209	0.213	0.071	0.174
	Debate-or-Vote (Sparse)	0.062	0.087	0.093	0.077	0.081	0.070	0.079	0.079	0.074	0.076	0.202	0.231	0.234	0.217	0.222	0.082	0.193
	Debate-or-Vote (Cent.)	0.068	0.079	0.096	0.079	0.084	0.067	0.087	0.106	0.083	0.089	0.199	0.219	0.227	0.212	0.216	0.106	0.187
	MATTRL	0.080	0.154	0.210	0.135	0.154	0.064	0.118	0.185	0.113	0.132	0.206	0.374	0.500	0.332	0.375	0.082	0.344
	MADKE	0.136	0.273	0.382	0.241	0.277	0.079	0.126	0.190	0.124	0.141	0.218	0.389	0.509	0.343	0.386	0.082	0.407
	Adaptive Judge	0.236	0.499	0.688	0.430	0.496	0.091	0.147	0.216	0.142	0.161	0.643	0.686	0.765	0.691	0.710	0.106	0.516
	iMAD	0.239	0.497	0.677	0.428	0.492	0.078	0.153	0.209	0.134	0.153	0.659	0.693	0.772	0.702	0.720	0.094	0.495
	Identity (Anon.)	0.216	0.427	0.594	0.377	0.433	0.094	0.147	0.201	0.138	0.154	0.622	0.716	0.779	0.690	0.713	0.082	0.476
	Identity (Exposed)	0.208	0.425	0.582	0.369	0.424	0.084	0.165	0.221	0.143	0.163	0.600	0.709	0.776	0.677	0.702	0.071	0.463
	ED2D	0.229	0.443	0.619	0.395	0.452	0.067	0.111	0.149	0.102	0.114	0.528	0.577	0.644	0.575	0.592	0.047	0.379
	MAD-Logic	0.225	0.421	0.578	0.376	0.427	0.059	0.108	0.149	0.097	0.110	0.522	0.619	0.683	0.592	0.615	0.129	0.354
	DoM	0.209	0.413	0.575	0.365	0.419	0.059	0.099	0.158	0.099	0.114	0.618	0.652	0.716	0.656	0.671	0.059	0.352
	EVESkillDebate (Ours)	0.305	0.541	0.691	0.473	0.529	0.264	0.356	0.424	0.333	0.356	0.735	0.851	0.890	0.806	0.828	0.106	0.495
Gemini 3.5 Flash-Lite	Classic MAD	0.277	0.478	0.619	0.425	0.475	0.099	0.155	0.201	0.142	0.157	0.690	0.714	0.756	0.716	0.726	0.176	0.597
	Debate-or-Vote (Dec.)	0.285	0.285	0.285	0.285	0.285	0.107	0.107	0.107	0.107	0.107	0.676	0.676	0.676	0.676	0.676	0.212	0.612
	Debate-or-Vote (Sparse)	0.289	0.297	0.297	0.293	0.294	0.110	0.112	0.112	0.111	0.111	0.682	0.686	0.686	0.684	0.684	0.247	0.623
	Debate-or-Vote (Cent.)	0.295	0.296	0.296	0.295	0.296	0.112	0.112	0.112	0.112	0.112	0.681	0.682	0.682	0.681	0.681	0.224	0.637
	MATTRL	0.261	0.495	0.650	0.429	0.486	0.096	0.151	0.195	0.139	0.153	0.672	0.696	0.732	0.696	0.705	0.129	0.526
	MADKE	0.241	0.427	0.519	0.365	0.405	0.032	0.059	0.078	0.052	0.058	0.634	0.661	0.703	0.662	0.672	0.153	0.558
	Adaptive Judge	0.261	0.485	0.648	0.428	0.484	0.096	0.161	0.209	0.145	0.161	0.690	0.718	0.770	0.721	0.733	0.153	0.574
	iMAD	0.255	0.447	0.585	0.397	0.445	0.092	0.150	0.197	0.137	0.152	0.686	0.715	0.752	0.713	0.723	0.118	0.560
	Identity (Anon.)	0.241	0.480	0.652	0.418	0.478	0.100	0.169	0.218	0.151	0.168	0.684	0.722	0.773	0.720	0.734	0.153	0.602
	Identity (Exposed)	0.253	0.485	0.661	0.427	0.487	0.098	0.161	0.218	0.148	0.166	0.679	0.713	0.773	0.716	0.730	0.153	0.625
	ED2D	0.246	0.471	0.619	0.408	0.462	0.063	0.115	0.151	0.101	0.114	0.662	0.699	0.755	0.699	0.714	0.165	0.556
	MAD-Logic	0.222	0.446	0.611	0.389	0.446	0.044	0.091	0.145	0.085	0.101	0.652	0.681	0.738	0.685	0.699	0.141	0.581
	DoM	0.236	0.419	0.562	0.375	0.423	0.062	0.099	0.143	0.095	0.107	0.675	0.701	0.735	0.699	0.709	0.106	0.560
	EVESkillDebate (Ours)	0.333	0.551	0.711	0.496	0.551	0.312	0.388	0.456	0.373	0.394	0.640	0.803	0.857	0.739	0.770	0.224	0.646
Gemini 3.6 Flash	Classic MAD	0.268	0.505	0.715	0.457	0.523	0.137	0.206	0.264	0.191	0.209	0.729	0.765	0.824	0.767	0.781	0.459	0.721
	Debate-or-Vote (Dec.)	0.326	0.326	0.326	0.326	0.326	0.145	0.145	0.145	0.145	0.145	0.732	0.732	0.732	0.732	0.732	0.494	0.719
	Debate-or-Vote (Sparse)	0.320	0.325	0.325	0.322	0.323	0.139	0.146	0.146	0.143	0.143	0.722	0.729	0.729	0.726	0.726	0.576	0.725
	Debate-or-Vote (Cent.)	0.308	0.308	0.308	0.308	0.308	0.147	0.147	0.147	0.147	0.147	0.732	0.732	0.732	0.732	0.732	0.482	0.723
	MATTRL	0.262	0.496	0.678	0.440	0.501	0.122	0.193	0.246	0.175	0.193	0.717	0.746	0.796	0.748	0.760	0.435	0.704
	MADKE	0.258	0.487	0.662	0.431	0.490	0.050	0.084	0.114	0.077	0.086	0.676	0.705	0.751	0.706	0.717	0.447	0.704
	Adaptive Judge	0.275	0.512	0.712	0.460	0.525	0.131	0.207	0.261	0.187	0.206	0.731	0.763	0.823	0.767	0.781	0.447	0.700
	iMAD	0.282	0.511	0.688	0.455	0.515	0.134	0.198	0.254	0.185	0.203	0.731	0.759	0.805	0.760	0.772	0.400	0.694
	Identity (Anon.)	0.273	0.511	0.710	0.458	0.522	0.138	0.214	0.274	0.196	0.216	0.721	0.764	0.823	0.762	0.777	0.471	0.725
	Identity (Exposed)	0.287	0.511	0.715	0.467	0.530	0.131	0.206	0.270	0.190	0.211	0.723	0.763	0.825	0.764	0.780	0.482	0.725
	ED2D	0.257	0.487	0.672	0.434	0.495	0.075	0.131	0.157	0.112	0.123	0.681	0.725	0.785	0.723	0.739	0.471	0.688
	MAD-Logic	0.255	0.483	0.708	0.444	0.511	0.060	0.108	0.151	0.099	0.112	0.711	0.747	0.813	0.751	0.766	0.412	0.704
	DoM	0.230	0.470	0.691	0.424	0.492	0.099	0.171	0.224	0.153	0.171	0.725	0.758	0.817	0.761	0.775	0.482	0.709
	EVESkillDebate (Ours)	0.337	0.554	0.717	0.500	0.555	0.285	0.386	0.447	0.356	0.379	0.722	0.813	0.851	0.780	0.798	0.424	0.721
Claude Haiku 4.5	Classic MAD	0.250	0.487	0.681	0.433	0.496	0.067	0.120	0.155	0.105	0.118	0.588	0.649	0.741	0.649	0.672	0.094	0.472
	Debate-or-Vote (Dec.)	0.278	0.303	0.303	0.290	0.294	0.052	0.059	0.059	0.056	0.056	0.604	0.613	0.614	0.609	0.610	0.141	0.474
	Debate-or-Vote (Sparse)	0.275	0.311	0.311	0.293	0.298	0.055	0.080	0.084	0.069	0.073	0.602	0.628	0.630	0.616	0.619	0.141	0.476
	Debate-or-Vote (Cent.)	0.290	0.290	0.290	0.290	0.290	0.060	0.060	0.060	0.060	0.060	0.595	0.595	0.595	0.595	0.595	0.106	0.486
	MATTRL	0.244	0.470	0.657	0.419	0.480	0.056	0.112	0.173	0.104	0.122	0.561	0.627	0.702	0.619	0.641	0.118	0.407
	MADKE	0.244	0.459	0.653	0.416	0.476	0.013	0.025	0.056	0.030	0.036	0.509	0.580	0.686	0.580	0.607	0.082	0.417
	Adaptive Judge	0.232	0.460	0.641	0.407	0.467	0.062	0.106	0.149	0.098	0.111	0.584	0.641	0.735	0.644	0.667	0.082	0.444
	iMAD	0.239	0.461	0.654	0.415	0.476	0.056	0.103	0.155	0.097	0.112	0.579	0.651	0.738	0.644	0.668	0.106	0.421
	Identity (Anon.)	0.240	0.459	0.655	0.415	0.476	0.055	0.106	0.146	0.094	0.107	0.596	0.643	0.739	0.652	0.674	0.094	0.442
	Identity (Exposed)	0.235	0.464	0.654	0.413	0.474	0.058	0.110	0.151	0.098	0.111	0.591	0.644	0.737	0.648	0.671	0.106	0.442
	ED2D	0.246	0.458	0.635	0.411	0.468	0.044	0.083	0.123	0.077	0.089	0.542	0.613	0.713	0.611	0.637	0.129	0.417
	MAD-Logic	0.211	0.445	0.635	0.391	0.453	0.033	0.060	0.108	0.063	0.074	0.561	0.640	0.749	0.636	0.665	0.071	0.417
	DoM	0.229	0.443	0.660	0.408	0.472	0.039	0.075	0.110	0.068	0.079	0.586	0.640	0.735	0.645	0.668	0.106	0.436
	EVESkillDebate (Ours)	0.285	0.525	0.702	0.464	0.525	0.319	0.387	0.438	0.370	0.387	0.645	0.797	0.865	0.744	0.775	0.106	0.532
Claude Sonnet 5	Classic MAD	0.223	0.448	0.661	0.407	0.472	0.095	0.150	0.207	0.142	0.158	0.714	0.745	0.797	0.747	0.760	0.353	0.692
	Debate-or-Vote (Dec.)	0.262	0.263	0.263	0.263	0.263	0.099	0.099	0.099	0.099	0.099	0.701	0.702	0.702	0.701	0.701	0.341	0.717
	Debate-or-Vote (Sparse)	0.266	0.273	0.273	0.269	0.270	0.103	0.104	0.104	0.104	0.104	0.707	0.714	0.714	0.711	0.712	0.282	0.704
	Debate-or-Vote (Cent.)	0.257	0.257	0.257	0.257	0.257	0.092	0.092	0.092	0.092	0.092	0.705	0.705	0.705	0.705	0.705	0.365	0.706
	MATTRL	0.235	0.467	0.673	0.420	0.484	0.094	0.146	0.206	0.140	0.157	0.696	0.725	0.782	0.729	0.743	0.329	0.669
	MADKE	0.213	0.445	0.639	0.394	0.457	0.029	0.052	0.075	0.048	0.055	0.622	0.657	0.717	0.660	0.674	0.294	0.669
	Adaptive Judge	0.225	0.455	0.650	0.405	0.467	0.100	0.158	0.216	0.148	0.166	0.711	0.745	0.800	0.746	0.760	0.294	0.679
	iMAD	0.233	0.462	0.663	0.415	0.478	0.095	0.151	0.209	0.142	0.159	0.701	0.723	0.783	0.732	0.745	0.318	0.683
	Identity (Anon.)	0.228	0.464	0.667	0.414	0.478	0.099	0.153	0.202	0.142	0.158	0.694	0.726	0.794	0.733	0.748	0.282	0.690
	Identity (Exposed)	0.228	0.460	0.663	0.412	0.476	0.091	0.146	0.205	0.138	0.155	0.694	0.730	0.793	0.733	0.749	0.341	0.696
	ED2D	0.227	0.442	0.638	0.400	0.461	0.050	0.094	0.146	0.089	0.104	0.652	0.683	0.745	0.689	0.703	0.376	0.679
	MAD-Logic	0.197	0.426	0.643	0.384	0.450	0.039	0.080	0.120	0.073	0.085	0.667	0.710	0.773	0.710	0.726	0.271	0.677
	DoM	0.204	0.433	0.646	0.389	0.455	0.047	0.090	0.131	0.082	0.095	0.681	0.716	0.773	0.717	0.731	0.294	0.662
	EVESkillDebate (Ours)	0.303	0.512	0.692	0.467	0.525	0.320	0.386	0.420	0.364	0.379	0.678	0.817	0.868	0.764	0.791	0.365	0.690

achieves the best complete metric profile with Gemini 2.5 Flash. With Gemini 3.5 Flash-Lite and Gemini 3.6 Flash, its H@1 is lower than the strongest baseline by 0.050 and 0.010, respectively; nevertheless, it achieves higher H@2, H@3, MRR, and NDCG@3 in both cases. Thus, the correct diagnosis appears more often among the leading candidates even when top-1 selection is weaker.

Claude backbones. Table 7 reports an analogous ranking pattern for Claude Haiku 4.5 and Claude Sonnet 5. On ReDial, MRR improves by 0.031–0.047 and NDCG@3 by 0.029–0.041. The only exception among the ReDial top-1 results occurs with Claude Haiku 4.5, where H@1 is 0.285 compared with 0.290 for the strongest baseline; all remaining ReDial metrics favor our method. On TG-ReDial, EVESKILLDEBATE has higher MRR by 0.216–0.265 and higher NDCG@3 by 0.213–0.265. On RareBench, it achieves the strongest complete metric profile with Claude Haiku 4.5. With Claude Sonnet 5, H@1 is lower than the strongest baseline, whereas H@2, H@3, MRR, and NDCG@3 are higher. The latter pattern reflects stronger reported ranking quality beyond the first position.

Single-answer tasks and interpretation. The results on HLE Math and SuperGPQA-Education are less uniform. On HLE Math, EVESKILLDEBATE outperforms the compared methods with GPT-4o-mini but does not consistently exceed the strongest baseline under the remaining backbones. On SuperGPQA-Education, it achieves the highest accuracy with GPT-5.4-Mini, Gemini 3.5 Flash-Lite, and Claude Haiku 4.5, while remaining competitive but not uniformly superior elsewhere. The empirical advantage is therefore strongest on evidence-grounded candidate-ranking tasks rather than single-answer mathematical reasoning. This pattern is consistent with the design of EVESKILLDEBATE: evidence-verifiable debate supports explicit comparison among competing candidates, while skill evolution converts recurring evidence and ranking failures into reusable procedural corrections. The largest margins on TG-ReDial are consistent with this task requiring the joint resolution of conversational constraints, documentary evidence, and graph relations; they do not by themselves identify which module causes the improvement.

E TRAJECTORY CHANGES AFTER SKILL EVOLUTION

We qualitatively examine five recorded trajectories spanning four evaluated benchmarks to characterize how the proposed framework changes observable reasoning behavior. The first four cases compare trajectories before and after skill evolution, whereas the final HLE case isolates the effect of structured reasoning. Across these examples, the revisions do not replace an entire trajectory; instead, they preserve valid intermediate steps and introduce a targeted check at the localized point of failure.

In HLE Math (Figure 4), the original trajectory correctly identifies that b decreases but extrapolates this observation to an overly broad blow-up region without conditioning on the sign of $a(0)$. The audit localizes the missing boundary distinction, and the evolved skill adds the check that $a(0) \leq 0$ cannot yield $a(t) \to +\infty$, while $a(0) > 0$ defines the relevant positive-data strip. This correction changes the estimate to $10 \leq m(\Omega) < 20$ while retaining the valid preceding analysis. In SuperGPQA-Education (Figure 5), the original trajectory associates repeated exposure with a broad learning category and incorrectly selects associative learning. The evolved skill instead tests the defining mechanism—a reduced response to repeated harmless stimulation without reward, punishment, or paired events—and correctly identifies habituation. In RareBench (Figure 6), evolution reduces reliance on generic phenotype overlap and introduces a specificity-based comparison using discriminative phenotype combinations, restoring the gold diagnosis to the leading candidates. In ReDial (Figure 7), the evolved trajectory prioritizes the user's current superhero-action preference over incidental entity and co-mention links, thereby recovering a candidate that better satisfies the expressed request. Finally, Figure 8 contrasts reasoning-disabled and reasoning-enabled HLE trajectories. The former invokes Parseval's identity without checking orthonormality and prematurely selects a simplex, whereas the latter constructs the Gram-matrix representation, verifies the induced quadratic constraint, and checks the nonlinear image in the target space before selecting the answer. Together, these examples illustrate how explicit auditing and localized skill revision can expose hidden assumptions, preserve valid reasoning prefixes, and correct specific procedural failures. They remain qualitative case studies and should not be interpreted as population-level causal evidence.

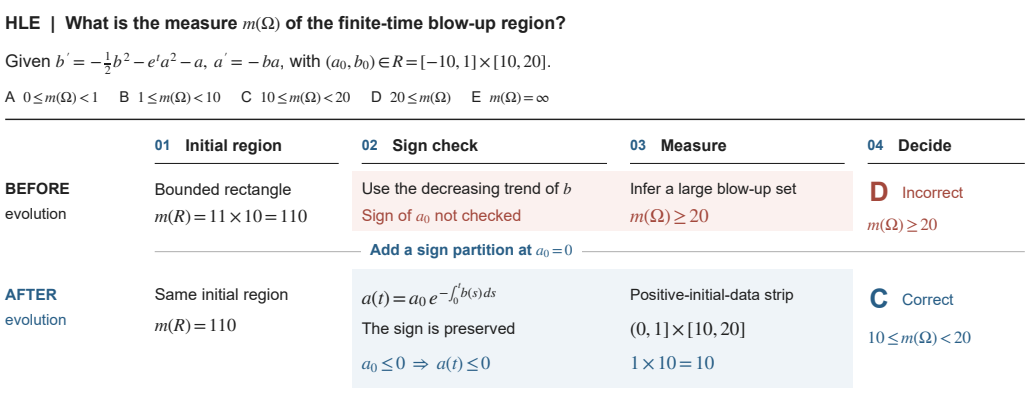


Figure 4: An HLE trajectory before and after skill evolution. The evolved skill preserves the valid analysis and inserts the missing sign-based case distinction before estimating the blow-up region.

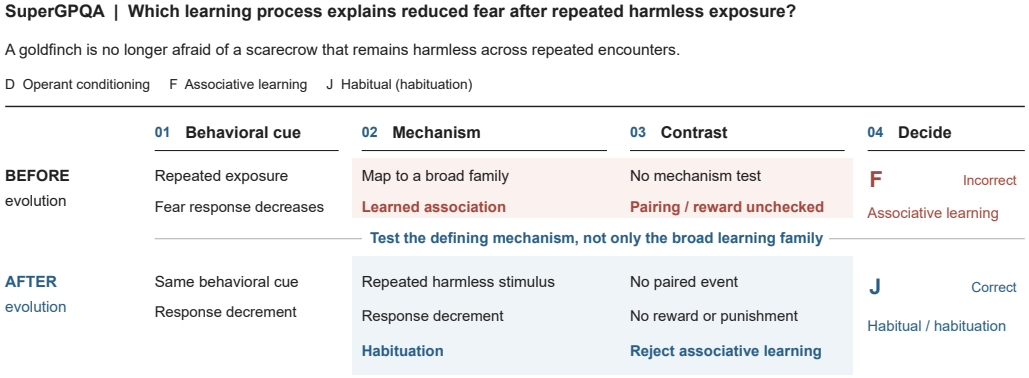


Figure 5: A SuperGPQA-Education trajectory in which evolution replaces a broad learning-family association with a defining-mechanism test, correcting associative learning to habituation.

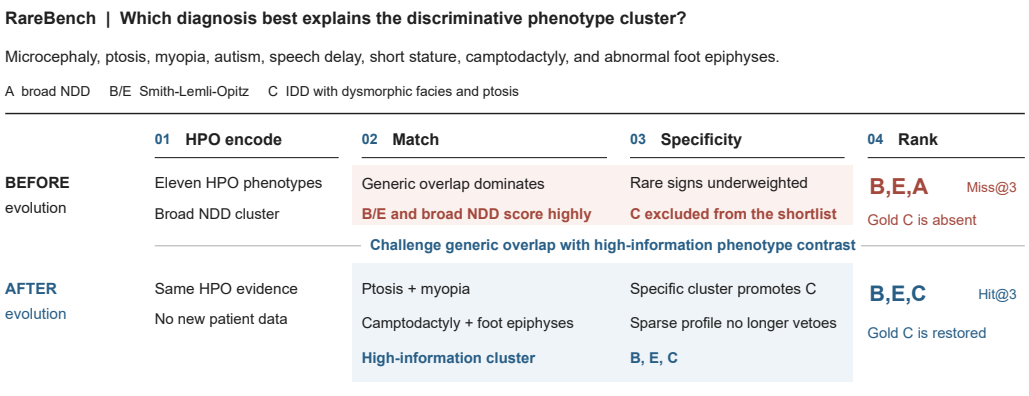


Figure 6: A RareBench trajectory in which evolution challenges generic phenotype overlap with a specificity contrast and restores the diagnosis supported by the discriminative phenotype cluster.

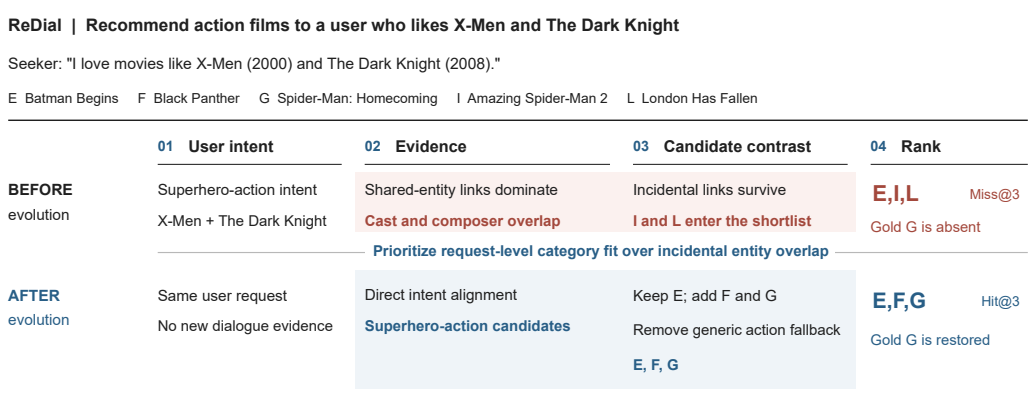


Figure 7: A ReDial trajectory in which evolution prioritizes request-level category fit over incidental shared-entity links, restoring a candidate aligned with the expressed superhero-action preference.

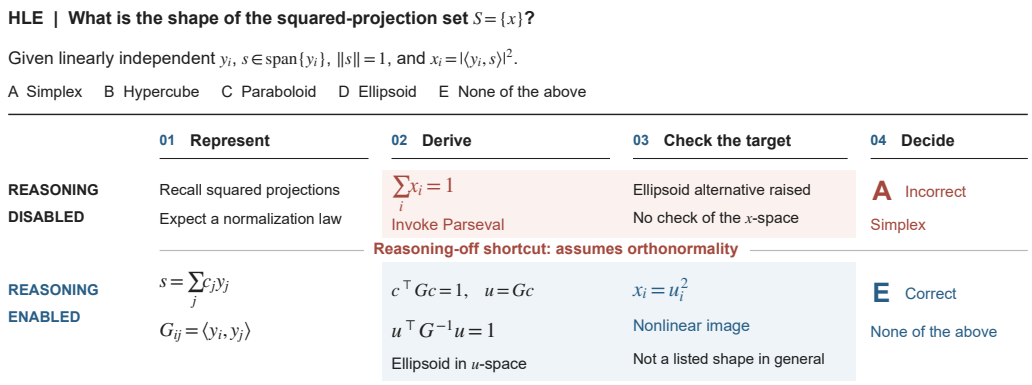


Figure 8: A representative HLE case illustrating the agent trajectories with reasoning disabled (top) and enabled (bottom).

F Dataset Construction and Evidence Sources

ReDial and TG-ReDial. For ReDial (Li et al., 2018), we retain 1,140 of the 1,342 dialogues in the official test split that contain a valid, previously unseen recommendation target. Each instance exposes only the dialogue prefix preceding the final recommendation, with the last newly mentioned movie treated as the gold candidate. Applying the same rule to TG-ReDial (Zhou et al., 2020) yields 747 test instances. Previously mentioned movies are excluded from both candidate sets. Topic and movie co-mention statistics for TG-ReDial are estimated exclusively from its training split.

RareBench. For RareBench (Chen et al., 2024b), we retain 1,099 cases from RAMEDIS, LIRICAL, HMS, and MME after removing instances containing explicit answer cues in public text or ambiguous mappings to multiple gold diagnoses.

HLE Math. We derive an 85-instance mathematical-reasoning subset from Humanity's Last Exam (Phan et al., 2025). It contains text-only, parseable multiple-choice problems and excludes image-dependent, free-form, and answer-revealing instances.

SuperGPQA-Education. We select 477 questions from the Education discipline of SuperGPQA (M-A-P Team et al., 2025), covering Education, Physical Education, and Psychology. Questions containing explicit answer cues are removed.

Every processed instance stores the complete candidate set used for evaluation, including stable candidate identifiers and a deterministic label order fixed before inference. Agents may rank only these candidates and cannot expand the set at inference time.

Evidence construction and leakage control. Each instance is represented as $(q, \mathcal{C}, \mathcal{D}, \mathcal{G})$, combining textual evidence with typed graph relations. The document view contains descriptive evidence, while the graph view contains typed relations for compositional and multi-hop verification. Evidence is collected from DBpedia for ReDial; Wikidata and public movie encyclopedias for TG-ReDial; HPO and HPOA for RareBench; retrieved Wikipedia mathematics pages for HLE Math; and ERIC, ISCED-F, and Wikipedia for SuperGPQA-Education. Gold labels are used only to define evaluation targets and apply the stated filters; they are not exposed to the retriever, evidence constructor, verifier, or inference-time agents. Evidence-acquisition rules are fixed before evaluation, and all split-dependent statistics are estimated exclusively from training data. The retained test-derived examples form a label-delayed online stream rather than a conventional untouched test set. Predictions for each batch are sealed before labels are revealed, and those labels can affect only later batches. We report prequential performance over the sealed predictions and never retrospectively rescore earlier instances with the final evolved skill.

G Evaluation Metrics and Online Protocol

Evaluation metrics. For a ranking instance x, let r_x be the position of the gold candidate in the returned evidence-ranked prefix and set $r_x = \infty$ if it is absent. The evaluated prefix has depth three; consequently, the reported MRR is the truncated reciprocal-rank metric MRR@3, denoted by MRR in the result tables for brevity. With N_{eval} instances and one relevant candidate per instance, we report

$$\begin{aligned} \text{H@}k &= \frac{1}{N_{\text{eval}}} \sum_x \mathbf{1}[r_x \le k], \qquad k \in \{1, 2, 3\}, \\ \text{MRR} &= \frac{1}{N_{\text{eval}}} \sum_x \mathbf{1}[r_x < \infty] \frac{1}{r_x}, \\ \text{NDCG@3} &= \frac{1}{N_{\text{eval}}} \sum_x \frac{\mathbf{1}[r_x \le 3]}{\log_2(r_x + 1)}. \end{aligned} \tag{42}$$

HLE Math and SuperGPQA-Education are single-answer tasks evaluated by accuracy. Unfilled ranks receive zero credit under every metric. A baseline that natively emits only one label is evaluated on top-1 accuracy; rank-sensitive metrics are reported only when it is rerun to return a ranked list through depth three over the same candidate set.

Audit and evolution diagnostics. All diagnostics are computed from records sealed before the corresponding label release. For RQ2, provenance validity is the micro-average of $\nu(e)$ over retained evidence units, candidate coverage is the fraction of decisions whose audited context contains every member of $\mathcal{C}$, and citation resolution is the fraction of cited identifiers that map to a retained evidence record. KG-path validity is the micro-average of $\text{Path}_{\mathcal{G}}$ over displayed graph paths, and dual-view coverage is the fraction of decisions containing at least one document span and one KG path. Audit coverage is the fraction of returned decisions for which every displayed candidate has both support and contradiction scores. The positive-margin rate is the fraction whose selected candidate has mean support minus mean contradiction greater than zero across its retained candidate-level evidence assertions. Aggregation consistency measures rank concordance rather than compliance with the deterministic aggregation rule. Let $\sigma_{J,x}$ be the complete permutation induced by L_J and let τ_{K} be Kendall's rank correlation on complete permutations; the metric averages $(1 + \tau_{\text{K}}(\sigma_{J,x}, \sigma_{i,x}^{T_{\max}}))/2$ over agents and then over returned decisions. Agent corroboration is the fraction of decisions supported by at least two agents with distinct evidence signatures. Rationale traceability is the fraction retaining both a final rationale and its linked evidence chain. Unit-level rates are micro-averaged over their applicable records; a decision with no applicable unit contributes zero to the corresponding coverage rate.

For the claim–evidence manual audit reported in Section 4.3, an item contains a final candidate-level claim and its cited evidence, with provider identity hidden. The rubric assigns *supported*

when the cited record establishes the stated relation in its displayed context, *contradicted* when it directly conflicts with the claim, and *insufficient* otherwise. The resulting rates are descriptive sample proportions: they measure alignment between claims and citations, not the external truth or completeness of the cited sources.

For RQ3, let $\mathcal{E}_{\text{match}}$ be the matched evolved–unevolved comparisons and $\Delta_x = s_{\text{evo}}(x) - s_{\text{base}}(x)$. Evolution success and non-regression are

$$\text{Succ}_{\text{evo}} = \frac{\sum_{x \in \mathcal{E}_{\text{match}}} \mathbf{1}[\Delta_x > 0]}{\sum_{x \in \mathcal{E}_{\text{match}}} \mathbf{1}[\Delta_x \neq 0]}, \qquad \text{NonReg}_{\text{evo}} = \frac{1}{|\mathcal{E}_{\text{match}}|} \sum_{x \in \mathcal{E}_{\text{match}}} \mathbf{1}[\Delta_x \geq 0]. \tag{43}$$

Both ratios are defined as zero when their denominators are zero. The reported success interval is a descriptive Wilson interval for the non-tied comparisons. Output validity requires successful parsing, schema compliance, and membership of every predicted label in $\mathcal{C}$. Behavioral diversity is the fraction of matched examples on which at least one evolved challenger differs from its unevolved counterpart.

A frozen coding rule assigns each accepted edit one of seven predeclared error codes: constraint, grounding, quantifier, relation, alternatives, definition, or computation. If p_r is the empirical share of code r, pattern entropy is

$$H_{\text{pat}} = -\frac{1}{\log 7} \sum_{r=1}^{7} p_r \log p_r, \qquad 0 \log 0 := 0. \tag{44}$$

We set $H_{\text{pat}} = 0$ when no accepted edit is available. Clause distance is the mean normalized token-level Levenshtein distance over unordered pairs of edited clauses, where $d_{\text{cl}}(u, v) = \text{Lev}(u, v)/\max\{1, |u|, |v|\}$ and $|u|$ is the token count; it is defined as zero when fewer than two clauses are available. For each reported domain or backbone setting s, let $\mathbf{p}_s$ be its seven-code distribution. Cross-setting coverage is the fraction of codes having nonzero frequency in every reported setting, and Jensen–Shannon similarity is

$$\text{JSSim} = \frac{1}{\binom{|\mathcal{S}|}{2}} \sum_{s<t} \left(1 - \frac{\text{JSD}(\mathbf{p}_s, \mathbf{p}_t)}{\log 2}\right). \tag{45}$$

We define this similarity as zero when $|\mathcal{S}| < 2$. These diagnostics measure structural validity, observed variation, and pattern recurrence; they do not establish semantic diversity or direct source-to-target skill transfer.

Label-delayed online evaluation. The evaluation stream is partitioned into a fixed sequence of batches. At the beginning of each batch, the deployed and shadow skills and all evolution states are frozen. Predictions and execution traces for the entire batch are finalized and sealed before its labels are revealed. Feedback from a batch can therefore affect only subsequent batches, and all reported metrics are computed from predictions made before label disclosure.

Default operating point. Unless a sensitivity experiment varies the corresponding factor, we use two debate rounds, a routed-agent budget of eight, 32 evolution batches, and a generation temperature of 0.3. The stream order and all remaining retrieval, verification, localization, and feedback-gate parameters are fixed before evaluation.

Frozen No-Evolve control. Frozen No-Evolve uses the same data stream, batch order, backbone LLMs, agent team, evidence, verifier, decoding configuration, and debate procedure as the full method. It differs only in disabling feedback-driven skill record updates, trace-scorer updates, edit generation, shadow acceptance and deployment, and rollback; its deployed skill therefore remains fixed throughout the stream. The champion-side inference procedure is otherwise unchanged.

Scope of the online comparison. Static baselines do not consume delayed labels and remain fixed, whereas EVESKILLDEBATE may use labels from completed batches to update later policies. Cross-method results therefore compare deployed procedures under their native update rules and do not, by themselves, isolate the effect of delayed feedback. The matched evolved–unevolved comparisons and Frozen No-Evolve control are the relevant comparisons for that question. Because the main experiments use one fixed stream order, sensitivity to stream order remains a limitation.

H BASELINE DETAILS

Classic MAD Du et al. (2024b): It first elicits independent solutions from multiple agents. In subsequent rounds, each agent observes its peers' arguments and revises its reasoning. The final prediction aggregates the agents' last-round answers, allowing interaction to correct errors in individual solutions.

Debate-or-Vote (Choi et al., 2025): It examines whether communication improves upon independent voting. Its decentralized variant permits all-to-all communication, the sparse variant restricts communication to neighboring agents, and the centralized variant uses a hub agent to synthesize the remaining agents' arguments. We report these variants as Dec., Sparse, and Cent., respectively.

MATTRL Hu et al. (2026): It introduces collaborative test-time reinforcement learning for multi-agent reasoning. Agents share task-solving experience through a common workspace and retrieve relevant experience for subsequent instances. Feedback converts successful behavior into reusable guidance and identifies unsuccessful behavior to avoid, after which a coordinating agent integrates the specialist responses.

MADKE Wang et al. (2025): It augments multi-agent debate with a shared external knowledge pool (Wang et al., 2025). Debaters select evidence relevant to their arguments, reason independently, and subsequently exchange evidence-based responses. A judge assesses convergence, and a summarizer produces the final answer.

Adaptive Judge Hu et al. (2025): It replaces a fixed debate length with iterative deliberation among multiple judges. Each judge updates its assessment after observing the others' opinions, while posterior candidate support and cross-round agreement determine when deliberation terminates. The final answer aggregates the stabilized judgments.

iMAD Fan et al. (2026): It uses interpretable uncertainty signals to determine whether an instance requires debate. An initial self-critique identifies confidence, uncertainty, and plausible alternatives, after which a learned gate estimates the value of further interaction. Confident instances retain their initial answers, whereas uncertain instances invoke multi-agent debate.

Identity Anonymization Choi et al. (2026): It studies how model-identity disclosure affects multi-agent deliberation (Choi et al., 2026). Agents use the same debate structure while observing either anonymous participant labels or explicit model identities. The anonymized and identity-exposed variants therefore isolate conformity and authority effects induced by identity disclosure.

ED2D Han et al. (2026): It structures debate around explicit evidence and opposing positions. An evidence detector ranks relevant information and classifies it as supporting, contradicting, or neutral toward a claim. Agents assigned different stances conduct argument and rebuttal rounds, after which multiple judges assess the debate and a final judge synthesizes the prediction.

MAD-Logic Yang et al. (2026b): It combines natural-language debate with symbolic reasoning and verification. Agents translate a problem into predicates and constraints, a critic checks the resulting representation, and independent reasoners solve the structured problem. A verifier then compares their reasoning traces and selects the final answer.

DoM Liu et al. (2025a): It separates reasoning over textual evidence from reasoning over graph-structured knowledge. Specialized agents answer subquestions using their respective evidence sources and cross-examine unsupported or conflicting claims. An arbiter reconciles the resulting arguments and returns the final prediction.

I DETAILED EFFICIENCY ANALYSIS

Table 4 reports task performance and LLM-call cost averaged across backbones. Because the ranking tasks use NDCG@3 whereas HLE Math and S-GPQA use accuracy, we examine the

trade-off separately for each task rather than combining them into a single performance score. EVESKILLDEBATE achieves the highest point estimates on four of the five tasks. Relative to the strongest baseline for each task, it improves ReDial, TG-ReDial, RareBench, and S-GPQA by 4.32, 23.94, 7.53, and 2.75 points, respectively. On HLE Math, it ranks second at 22.94%, trailing Debate-or-Vote (Sparse) by 1.47 points.

EVESKILLDEBATE requires 13.77 inference calls and 13.88 total calls per example, so non-inference operations introduce only 0.11 additional calls on average. Its total cost is 46.28% of the 30-call Debate-or-Vote budget, representing a 53.72% reduction. Compared with the three Debate-or-Vote variants, our method performs better on four tasks and remains competitive on HLE Math. It also strictly outperforms ED2D in the reported averages: EVESKILLDEBATE uses fewer total calls (13.88 versus 16.00) while achieving higher performance on every task.

Several baselines require fewer calls. MATTRL and MADKE use approximately 8.3 calls per example, the two Identity variants and DoM use 9.0, and Adaptive Judge uses 12.32. Nevertheless, each produces lower point estimates than EVESKILLDEBATE on all five tasks. Adaptive Judge provides the closest cost comparison: for 1.56 additional total calls per example, our method improves ReDial, TG-ReDial, RareBench, HLE Math, and S-GPQA by 4.32, 24.05, 7.89, 2.79, and 3.74 points, respectively. Thus, the observed performance gains are not merely a consequence of using the largest inference budget. Although EVESKILLDEBATE is not the least expensive method, no lower-cost baseline dominates it in the reported comparisons. These results place the method at a favorable point on the empirical performance–efficiency trade-off, particularly for evidence-grounded ranking tasks.

J Detailed Hyperparameter Sensitivity

Figure 3 presents one-factor-at-a-time sensitivity around the default configuration of two debate rounds, eight routed agents, temperature 0.3, and 32 evolution batches. During each sweep, all remaining hyperparameters are fixed. The macro score is the average of the five task-specific scores.

Debate rounds. Increasing debate depth from one to two rounds raises the macro score from 41.5% to 43.2%. ReDial and RareBench benefit most from this additional round, whereas the changes on the other tasks are smaller. Increasing the depth to three rounds produces mixed task-specific effects and slightly reduces the macro score to 43.0%. Thus, two rounds provide sufficient opportunity for critique and revision without adding a round that offers no aggregate improvement.

Agent budget. The agent budget has the clearest monotonic effect. Increasing the number of routed agents from four to six improves the macro score from 39.2% to 42.2%, and using eight agents further raises it to 43.2%. The improvement appears across the task curves rather than being driven by a single benchmark. This pattern is consistent with broader agent coverage providing more diverse candidate assessments and evidence paths.

Generation temperature. Temperature 0.3 achieves the highest macro score (43.2%), compared with 42.9% at 0.0 and 41.8% at 0.7. The deterministic setting remains competitive, but slightly underperforms the moderate-temperature configuration. A temperature of 0.7 improves SuperGPQA but reduces performance on most other tasks, resulting in the lowest macro score. Moderate stochasticity therefore appears useful for maintaining diversity without substantially weakening reasoning consistency.

Evolution batches. Performance is least sensitive to the number of evolution batches. The macro scores for 16, 32, and 64 batches are 43.1%, 43.2%, and 42.8%, respectively. Most task scores remain nearly unchanged, although RareBench declines at 64 batches. This narrow range suggests that the method is stable under moderate changes to evolution granularity.

K Representative Agent Skills

All skills follow a common schema while specializing in different reasoning roles and evidence. Given a task, a skill router composes a small team from the versioned policy registry. This router selects inference-visible procedural policies and is distinct from the evidence retriever in Eq. (1),

which retrieves task evidence from $(\mathcal{D}, \mathcal{G})$. Table 8 describes the primary skill and other registered skills available to each domain. The router activates a task-conditioned subset of this portfolio. These stable skills define the initial agent team, while online evolution refines their reasoning principles from prior feedback without changing their domain roles.

Table 8: Primary and other registered skills by domain. ReDial and TG-ReDial use the same CRS skill portfolio with different evidence collections.

Domain	Primary skill	Other registered skills
CRS	Preference-grounded recommendation	Compact single-choice commitment, compact ranked shortlist, evidence-skeptic recommendation, pairwise preference discrimination.
RareBench	Phenotype-grounded diagnosis	Differential-diagnosis critic, phenotype-specificity contrast.
HLE Math	Formal mathematical reasoning	*General:* independent query solving, semantic-constraint decomposition, option falsification. *Math:* solution critic, task decomposition, computation sanity check, theorem recall, option elimination, counterexample construction, proof synthesis.
SuperGPQA-Education	Metadata-conditioned expert reasoning	*General:* independent query solving, semantic-constraint decomposition, option falsification, deliberative reasoning, concept normalization. *Domain:* education-domain reasoning, assessment critic, evidence-grounded option verification.
All domains	Claim–evidence alignment verifier	KG-path verification, traceable evidence-chain construction, moderator judgement, independent disagreement resolution.

ReDial/TG-ReDial: Preference-grounded recommendation

DOMAIN-SPECIFIC SKILL

Purpose. Rank movie candidates against the seeker's current request while retaining compatible dialogue history.

Interface. *Input:* dialogue context, candidate movies, and document and graph evidence. *Output:* an evidence-grounded ranking with cited support.

Reasoning focus. Compare genre, tone, premise, creator, cast, and franchise evidence. Retrieved profiles and graph paths strengthen a match, but missing evidence is treated as unknown rather than negative.

RareBench: Phenotype-grounded diagnosis

DOMAIN-SPECIFIC SKILL

Purpose. Rank candidate diseases from patient phenotypes and external disease-profile evidence.

Interface. *Input:* patient phenotypes, candidate diseases, and ontology and profile evidence. *Output:* a differential ranking with supporting and contradicting findings.

Reasoning focus. Prefer exact patient–HPO matches and discriminative biochemical markers over counts of generic symptoms. Use disease annotations to compare alternatives and merge database aliases before comparison.

HLE: Formal mathematical reasoning

DOMAIN-SPECIFIC SKILL

Purpose. Select the option supported by an explicit derivation rather than wording overlap or debate popularity.

Interface. *Input:* the problem, answer choices, retrieved principles, and available tool results. *Output:* a selected option with an explicit derivation and option checks.

Reasoning focus. Formalize objects, assumptions, quantifiers, and the target. Then test every option using theorem conditions, counterexamples, computation checks, and simple boundary cases.

SuperGPQA-Education: Metadata-conditioned expert

DOMAIN-SPECIFIC SKILL

Purpose. Construct a subject expert from public field and subfield metadata, then solve independently before social aggregation.

Interface. *Input:* the question, answer choices, subject metadata, and retrieved evidence. *Output:* a selected option with a concept-grounded comparison of alternatives.

Reasoning focus. Recover canonical textbook terminology, preserve polarity and relation direction, and compare the closest distractor. For sentence completion, substitute each option into the original stem.

Shared Verification and Provenance Skills — ALL DOMAINS

Claim–evidence alignment → Check support, contradiction, and citation validity.

KG-path verification → Connect evidence to task and candidate nodes.

Traceable evidence chain — Preserve document and graph provenance for the final answer.

L EXAMPLES OF EVOLVED SKILLS

Skill evolution preserves the input and output interface while revising the decision procedure. The examples below are abridged from active skill policies in the main experiments. Each evolved skill combines its original policy with a retained rule learned from contrasting successful and failed trajectories.

ReDial: Compact single-choice commitment

Interface. Dialogue request, candidate set, and concise evidence → one candidate label with a rationale.

BEFORE EVOLUTION	AFTER EVOLUTION
Read the active request and complete candidate set, then choose one label independently of vote counts. Preserve the latest explicit request. Do not use candidate order, popularity, or missing evidence as answer support. ⟹	Keep the original interface. First reconstruct the operator, polarity, assumptions, and every constraint. Derive a candidate-independent target before comparing choices. Contrast the leading candidate with its closest alternative and expose missing premises. Commit only after binding the result to the exact candidate text. Revise and explicitly flag insufficient evidence when decisive support is missing.

Contrastive credit assignment. The failed trajectory accepted a broadly plausible candidate without checking every required constraint. The successful trajectory verified all hard constraints before commitment.

TG-ReDial: Preference-grounded recommendation

Interface. Chinese dialogue request, candidate movies, and document–graph evidence → an evidence-grounded ranked recommendation.

BEFORE EVOLUTION	AFTER EVOLUTION
Extract the seeker's stated preferences from the dialogue, rank candidates by request fit, and use retrieved profiles and graph paths as support. Treat missing evidence as unknown rather than negative. ⟹	Retain preference-grounded ranking. Separate the latest request from older context, translate it into explicit constraints, and compare the leading candidates under the same constraints. Require candidate-specific document or graph support for decisive claims, reject contradictions, and commit only after the closest alternative has been tested.

Contrastive credit assignment. The failed trajectory favored a movie associated with earlier preferences and broad co-mention signals. The successful trajectory first isolated the latest request, then directly compared candidate-specific support for each active constraint.

RareBench: Phenotype-specificity contrast

Interface. Patient phenotypes, candidate diseases, and ontology evidence → a differential ranking with supporting and conflicting findings.

BEFORE EVOLUTION	AFTER EVOLUTION
Compare each disease with the observed phenotype set. Prefer candidates matching rare or characteristic findings, account for explicit conflicts, and explain why the leading diagnosis fits better than alternatives. ⟹	Retain phenotype-grounded diagnosis. Distinguish shared findings from discriminative phenotypes, trace each decisive claim to candidate-specific evidence, and check absent or conflicting findings. Contrast the top two diagnoses on the same evidence before producing the differential ranking.

Contrastive credit assignment. The failed trajectory accumulated generic symptom overlap without tracing which findings were diagnostically specific. The successful trajectory used candidate-level ontology evidence and a discriminative finding to separate the leading disease from its closest alternative.

HLE: Mathematical option elimination

Interface. Problem, answer choices, and retrieved principles → selected option, eliminations, and citations.

BEFORE EVOLUTION	AFTER EVOLUTION
Inspect every option and eliminate it using a violated assumption, wrong quantifier, wrong scale or sign, impossible boundary case, theorem mismatch, or contradiction with proof evidence. Require a constraint-level reason rather than lexical overlap. ⟹	Retain option-by-option elimination. When a quantitative expression is present, require an explicit derivation or evidence trace. Apply the same constraints to competing options, then accept only the candidate that survives both derivation and direct comparison.

Contrastive credit assignment. The failed trajectory accepted a plausible quantitative option without an explicit derivation. The successful trajectory required the derivation before candidate comparison.

SuperGPQA-Education: Independent option falsification

Interface. Question, answer choices, and retrieved evidence → selected option with a falsification trace.

BEFORE EVOLUTION	AFTER EVOLUTION
For every plausible option, seek one decisive failure such as a definition mismatch, reversed relation, wrong sequence, missing condition, counterexample, or numerical boundary violation. Compare the final two candidates and prefer the one that survives.	Retain independent falsification. First map the stem and each leading option to the relevant definition or taxonomy relation. Then test task constraints and relation direction. Compare the final two only after this mapping, and select the option that survives both structural and evidential checks.

Contrastive credit assignment. The failed trajectory compared close options without an explicit definition or taxonomy mapping. The successful trajectory established that mapping before falsification.

M REPRESENTATIVE PROMPT TEMPLATES

We report the four prompt templates that implement the main reasoning stages of EVESkillDebate[1]. Each card combines the persistent instruction, runtime input, and output contract seen by that component. Runtime fields are enclosed in braces.

M.1 SKILL-WRAPPED DEBATE AGENT

Debate Proposal and Rebuttal

```
SYSTEM INSTRUCTION
You are a skill-wrapped, evidence-grounded debate agent.

Follow the active skill. Compare all candidates listwise and assess
    every
candidate. Cite document identifiers when retrieved evidence supports
    an
assessment. An empty evidence list means unknown, not false.
    Candidate labels
and presentation order are arbitrary and carry no semantic signal. Do
    not
invent answers outside the candidate set.

{task_conditioned_working_clause}

RUNTIME INPUT
Skill: {skill_name}
Role: {role_assignment}
Active skill policy: {evolved_or_vanilla_skill}
Selected procedural policy: {relevant_skill_procedure}
Phase: {proposal_or_rebuttal}
Requested output depth: {k_out}

Allowed labels in a deterministic, gold-blind order:
{candidate_labels}

Task instance:
{query_and_complete_candidate_texts}

Retrieved documents and graph-derived evidence:
{evidence_documents}

Debate history and verifier feedback:
{prior_agent_turns_and_verifier_signals}

Score every allowed label exactly once in [0,1]. Bind each decisive
    factual
```

[1]Transport-level instructions for API retries and serialization repair are omitted

```
claim to a retrieved document span, KG edge, or replayable tool
    output.
Candidate membership alone is not semantic support. During rebuttal,
    inspect
your prior turn, competing proposals, and verifier feedback. Revise
unsupported claims or citations and address the strongest conflict
    instead of
repeating the first-round answer.

OUTPUT CONTRACT
Return candidate_scores, candidate_evidence, answer_labels,
    rationale, cited
document identifiers, and confidence.
```

M.2 Evidence and Graph Verifier

Candidate-Wise Evidence Verification

```
SYSTEM INSTRUCTION
You are a verifier in an evidence-grounded multi-agent debate.

Judge whether the proposal is substantively supported and whether it
contradicts task evidence. Compare every candidate independently
    rather than
copying the proposed ranking. Keep substantive support separate from
    citation
validity. Support must come from claim-bound retrieved evidence or a
    replayable
registered derivation/tool output; unaudited parametric knowledge is
    not
sufficient.

Candidate membership, question-to-option paths, taxonomy membership,
    and
agent agreement establish provenance only; they do not establish
    answer
truth. A structural graph path can be perfectly traceable while
    providing
zero semantic support. Apply both the claim-evidence verification
    skill and
the KG-path verification skill.

RUNTIME INPUT
Verifier skill: {active_verification_procedure}
KG-path skill: {active_graph_tracing_procedure}
Instance: {query_and_candidate_options}
Proposal: {answer_scores_rationale_and_citations}
Evidence: {cited_and_candidate_conditioned_documents}
KG paths: {deterministically_traced_multihop_paths}
Prior checks: {support_contradiction_and_citation_checks}

Score support and contradiction for every candidate. Penalize
    conflicting
definitions, relation directions, orders, units, thresholds, user
constraints, or clinical findings. Do not convert traceability into
    semantic
support.

OUTPUT CONTRACT
Return candidate-wise and aggregate support and contradiction,
    citation
validity, graph traceability, and a concise verifier note.
```

M.3 Evidence-Aware Debate Moderator

Conflict Resolution and Final Ranking

```
SYSTEM INSTRUCTION
You are the judge in an evidence-grounded multi-agent debate.

Aggregate debater proposals and verifier scores, but do not reward
    consensus
without independent support. Prefer the candidate backed by the
    strongest
valid derivation or candidate-specific evidence, even when it is a
    minority
proposal. Traceability, candidate membership, taxonomy proximity, and
    repeated
agreement do not by themselves prove a candidate. Use only candidate
    labels
and document identifiers present in the input.

{task_conditioned_judgement_clause}

RUNTIME INPUT
Moderator skill: {active_moderator_procedure}
Task type: {task_type}
Requested output depth: {k_out}
Candidate labels: {candidate_labels}
Instance: {query_and_candidates}
Debate: {all_proposals_and_rebuttals}
Verification: {candidate_wise_support_contradiction_and_provenance}
Evidence: {documents_and_graph_derived_evidence}
Prior scores: {calibrated_initial_candidate_scores}

Score every candidate in [0,1], then judge listwise. Count distinct
    reasoning
methods rather than raw turns: rebuttals update an initial view, and
    agents
sharing a backbone are correlated. Parametric knowledge is not
    auditable
evidence. If support or traceability is insufficient, retain the
    complete
ranking but explicitly flag the affected candidates rather than
    treating
agreement as evidence.

OUTPUT CONTRACT
Return final candidate scores, the complete evidence-ranked labels,
    support
flags, the final rationale, supporting evidence identifiers, and
    concise
judgement notes.
```

M.4 Contrastive Skill Evolution

Trajectory-to-Skill Update

```
SYSTEM INSTRUCTION
You optimize one reusable reasoning skill from sealed earlier-batch
successful-versus-failed trajectories.

Infer why the successful reasoning was better, assign the repeated
    error to a
reasoning operation, and rewrite the procedure used on future tasks.
    Each step
```

```
must define an observable operation and a criterion that can reject a
candidate. The update must be domain-general and executable using
    only a new
query, its candidates, retrieved evidence, and current hypotheses.

Do not retain past entities, facts, theorem instances, numerical
    answers,
candidate phrases, option labels, dataset names, answer positions, or
    an
instruction to trust consensus.

SEALED PRIOR-BATCH INPUT
Skill: {skill_name}
Current policy: {vanilla_or_deployed_skill_policy}
Deployment scope: {routing_terms}

Trajectory contrasts: [
  { query: {past_query}; candidates: {past_candidates};
    successful: {higher_utility_trajectory};
    failed: {lower_utility_trajectory};
    failure_mode: {credit_assigned_error} }, ...
]

Use every supplied trajectory contrast. If the localized evidence is
insufficient or inconsistent, set apply_update to false; otherwise,
    produce a
complete replacement patch:
  INTERPRET -- reconstruct operator, polarity, assumptions, and
    target;
  DERIVE    -- obtain a candidate-independent result when possible;
  CONTRAST  -- falsifiably compare it with the closest alternative;
  COMMIT    -- bind the result to candidate text and state when the
    available
               evidence remains insufficient.

Do not mention feedback, history, successful/failed markers, or prior
    outcomes
inside the resulting skill.

OUTPUT CONTRACT
Return apply_update, the recurring failure pattern, four procedural
    steps
(interpret, derive, contrast, and commit), and the supporting
    trajectory
indices. The resulting procedure is first evaluated as a shadow skill
    and is
accepted for deployment only through subsequent lifecycle evaluation.
```

M.5 Task-Conditioned Working Clauses

Task-Conditioned Comparison Criteria

```
The shared templates remain fixed. Append one comparison criterion
    according
to the task:

ReDial/TG-ReDial: Active recent preferences dominate popularity and
    broad
genre overlap.
RareBench: Discriminative phenotypes and contradictions dominate raw
    overlap
counts.
```

```
HLE Math: Derivation, theorem conditions, counterexamples, or
    numerical checks
are required.
SuperGPQA-Education: Operator, polarity, grammatical slot, and
    relation direction are
preserved.
```

N EXAMPLES OF DOCUMENT AND GRAPH EVIDENCE

Figures 9–12 illustrate complementary views of the same evidence. Documents preserve descriptive context and definitions, while graphs expose entity connections and concept hierarchies that can be followed across multiple steps. The snippets and subgraphs below are selected from the released instances and shortened for presentation. Please note that gold answers are shown only to orient the reader, and are not part of the inference-time evidence.

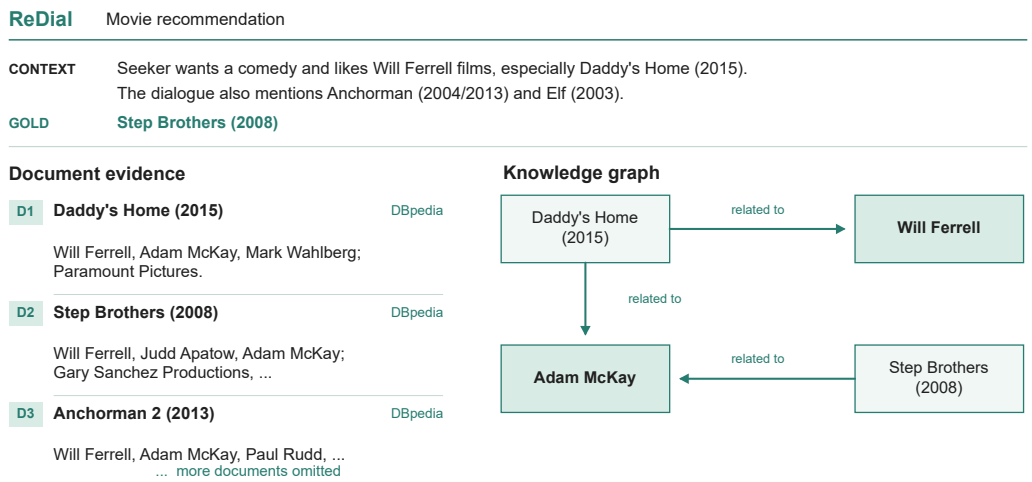


Figure 9: **Movie-recommendation evidence.** The dialogue preserves the user's comedy preference and previously discussed films. Document profiles describe people and studios, while labeled DBpedia relations expose shared-entity paths to candidates.

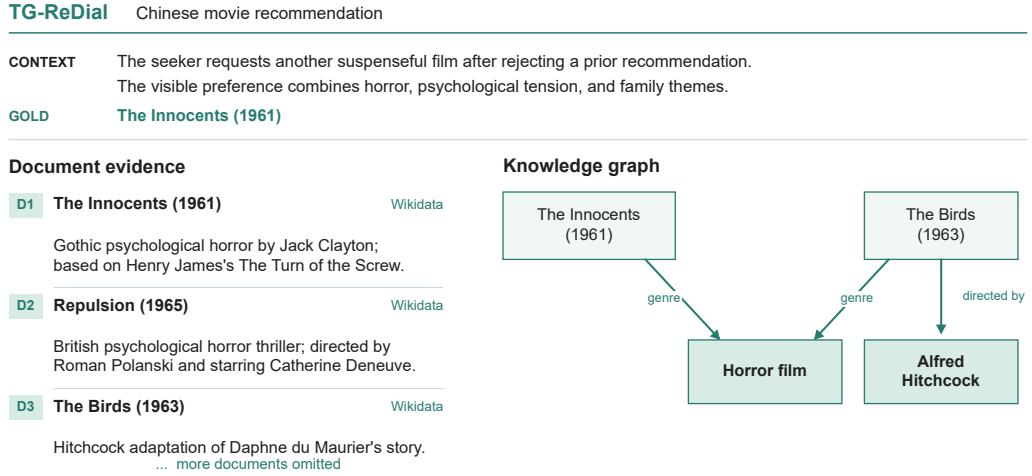


Figure 10: **Chinese movie-recommendation evidence.** External profiles describe genre, creators, cast, and source material, while the graph links films through shared Wikidata entities and a director relation.

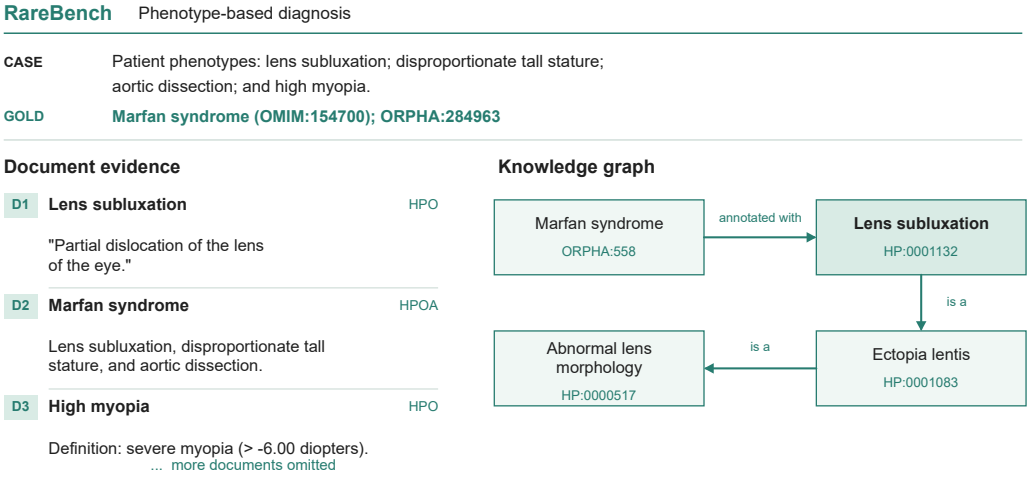


Figure 11: **Rare-disease evidence.** Documents provide a phenotype definition and external disease annotations. The three-hop graph path connects a disease annotation to successively broader phenotype concepts through the HPO hierarchy.

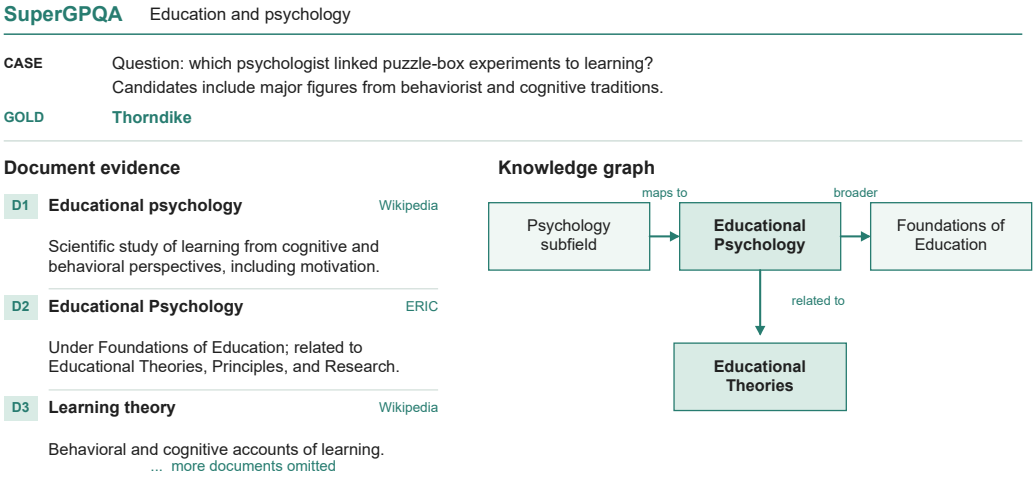


Figure 12: **Educational evidence.** Wikipedia documents provide substantive learning-theory context, while ERIC supplies typed descriptor relations connecting psychology, educational psychology, theories, and foundations of education.

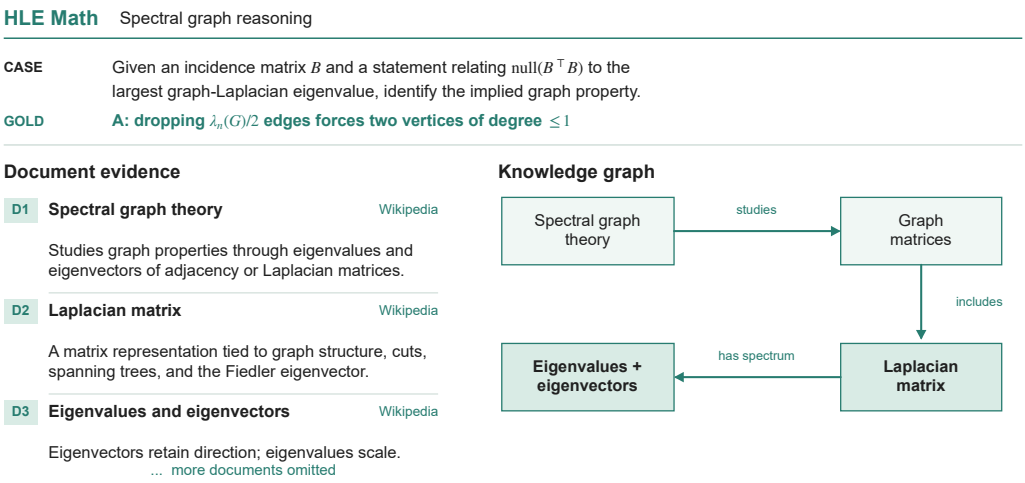


Figure 13: **Mathematical evidence.** Retrieved Wikipedia documents supply definitions and structural facts about graph spectra. The concept chain organizes these facts from spectral graph theory through graph matrices to Laplacian eigenpairs.