

SCOPE: Prompt Evolution for Enhancing Agent Effectiveness

Zehua Pei¹ Hui-Ling Zhen² Shixiong Kai² Sinno Jialin Pan¹ Yunhe Wang² Mingxuan Yuan² Bei Yu¹

Abstract

Large Language Model (LLM) agents are increasingly deployed in environments that generate massive, dynamic contexts. However, a critical bottleneck remains: **while agents have access to this context, their static prompts lack the mechanisms to manage it effectively, leading to recurring Corrective and Enhancement failures.** To address this capability gap, we introduce **SCOPE** (Self-evolving Context Optimization via Prompt Evolution). **SCOPE frames context management as an *online optimization* problem, synthesizing guidelines from execution traces to automatically evolve the agent’s prompt.** We propose a **Dual-Stream mechanism** that balances tactical specificity (resolving immediate errors) with strategic generality (evolving long-term principles). Furthermore, we introduce **Perspective-Driven Exploration to maximize strategy coverage**, increasing the likelihood that the agent has the **correct strategy for any given task**. Experiments on the HLE benchmark show that SCOPE improves task success rates from 14.23% to 38.64% without human intervention. We make our code publicly available at <https://github.com/JarvisPei/SCOPE>.

1. Introduction

The core functionality of a Large Language Model (LLM) agent is to perceive and react to context. Whether interpreting a user’s ambiguous instruction, analyzing a screen full of code, or deciphering a traceback from a failed tool execution, the agent’s success depends entirely on its ability to process this incoming information and decide on the next action. As agents are tasked with increasingly complex problems, the volume and complexity of this context grow exponentially, such as in deep research (OpenAI, 2025; Google, 2025; Wei et al., 2025; Zhang et al., 2025b; Team et al., 2025) and agentic coding (Anthropic, 2025; Jimenez

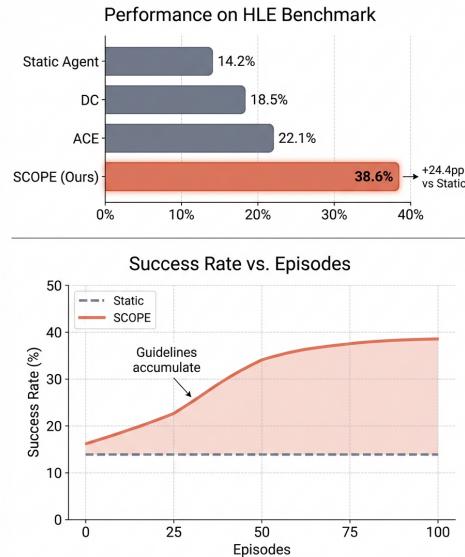


Figure 1. SCOPE enables prompt evolution for enhanced agent effectiveness. SCOPE significantly outperforms static agents and existing methods on the HLE benchmark; success rate improves as strategic guidelines accumulate over episodes.

et al., 2023). Consequently, the definition of a capable agent is shifting from one that simply knows facts to one that can effectively manage and navigate complex contexts.

However, current research in agentic systems has predominantly focused on the availability of context rather than how agents manage it. Innovations such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) or infinite context windows (Team et al., 2024) focus on feeding more data to the agent. We argue that there is a fundamental gap between possessing context and dealing with context. Even recent test-time learning methods like DC (Suzgun et al., 2025) or ACE (Zhang et al., 2025a), which attempt to bridge this gap, operate at task-level granularity and cannot adapt during execution. Without an evolved internal strategy to interpret these signals, an agent flooded with context is merely an agent flooded with noise. Figure 2 illustrates this gap: agents receive actionable information in their context but fail to leverage it effectively.

In this paper, we first substantiate this gap through an extensive analysis of over 1.5 million lines of execution logs from standard agents. Our observations reveal that static

¹The Chinese University of Hong Kong, Hong Kong SAR

²Noah’s Ark Lab, Huawei, Hong Kong SAR.

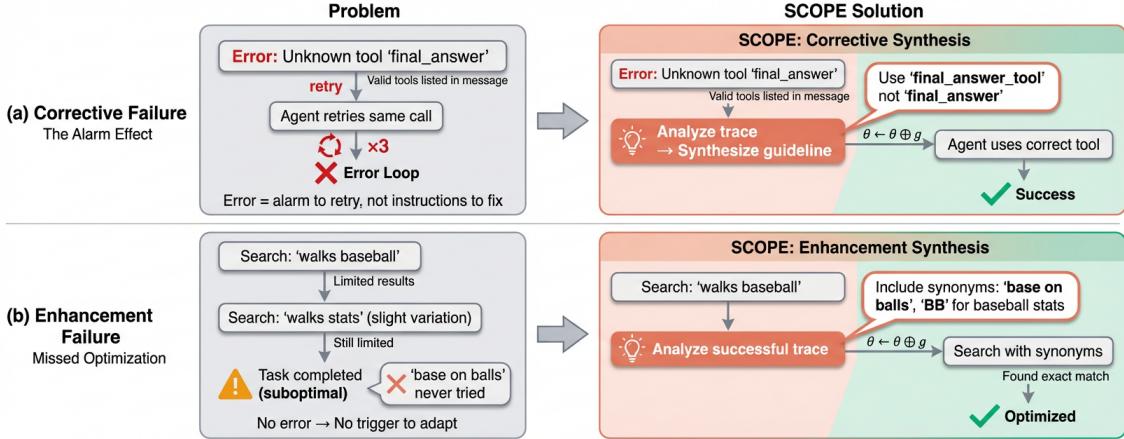


Figure 2. Two failure modes and how SCOPE addresses them. (a) **Corrective Failure**: The agent treats errors as generic alarms, entering error loops despite the error message containing the solution. SCOPE synthesizes a corrective guideline and integrates it into the prompt, enabling recovery. (b) **Enhancement Failure**: The agent persists with suboptimal strategies (e.g., single-term search) when no error is raised. SCOPE proactively analyzes successful traces to synthesize optimization guidelines.

prompts fail to adapt to dynamic context needs, leading to two distinct failure modes. First, agents suffer from *Corrective Failures*: when errors occur, agents treat error logs as generic alarms (signals to retry) rather than actionable feedback containing the solution. In severe cases, agents fabricate data to proceed rather than managing uncertainty. Second, agents exhibit *Enhancement Failures*: even when operating without errors, agents miss opportunities to optimize, persisting with suboptimal strategies because their static prompts lack the mechanism to learn from execution patterns.

These failures are magnified in modern agentic systems, which are composed of specialized agents (e.g., coder, browser, analyzer) engaging in long, multi-turn interactions. The complexity of these systems and the sheer volume of execution steps render static, one-size-fits-all prompts obsolete. Yet this structure also creates an opportunity: since agents are invoked repeatedly, their prompts can evolve online.

To address this gap, we propose **SCOPE** (Self-evolving Context Optimization via Prompt Evolution), a framework that transforms context management from a manual engineering task into an automatic optimization process. SCOPE operates on the insight that the agent's own execution trace serves as the perfect training signal. Through **Trace-Based Guideline Synthesis**, SCOPE analyzes these traces online to generate guidelines that teach the agent how to handle specific context patterns. A **Dual-Stream Routing** mechanism balances tactical updates (immediate error correction) with strategic updates (long-term principles), ensuring the agent not only survives the current task but becomes smarter for future ones. Finally, **Perspective-Driven Exploration** evolves multiple parallel prompts guided by distinct per-

spectives (e.g., Efficiency vs. Thoroughness), maximizing strategy coverage. As shown in Figure 1, SCOPE significantly outperforms static agents and existing methods, with success rate improving as guidelines accumulate.

Our contributions are as follows:

- We systematically study agent failures in modern agentic systems. We identify two failure modes, *i.e.* Corrective Failures and Enhancement Failures, supported by analysis of over 1.5 million lines of execution logs.
- We propose SCOPE, a framework that synthesizes guidelines from execution traces, routes them via dual-stream (tactical/strategic), and explores diverse strategies through multiple perspectives.
- We demonstrate that SCOPE significantly outperforms static baselines, raising task success rates on the challenging HLE benchmark from 14.23% to 38.64% and on GAIA from 32.73% to 56.97%.

2. Observations and Motivation

We analyzed over 1.5 million lines of execution logs from baseline agents on the GAIA (Mialon et al., 2023) and DeepSearch (Chen et al., 2025) benchmarks. Our findings reveal that agents often have the information needed to succeed, but their static prompts lack the mechanism to learn from execution feedback. We categorize failures into two modes based on their trigger: *Corrective Failures* occur when errors provide explicit signals that agents fail to act upon, while *Enhancement Failures* occur when agents miss opportunities to optimize even in the absence of errors. Figure 2 illustrates both failure modes and how SCOPE addresses them through trace-based guideline synthesis.

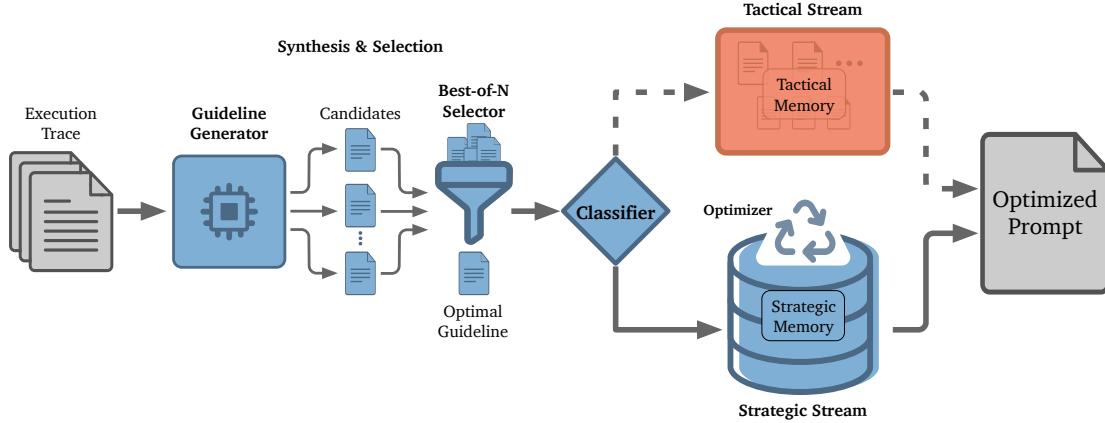


Figure 3. Overview of the SCOPE Framework. The **Generator** (π_ϕ) synthesizes candidate guidelines from execution traces. A **Selector** (π_σ) chooses the best guideline, which is then routed by the **Classifier** (π_γ) to either **Tactical** (task-specific) or **Strategic** (persistent) memory. The **Optimizer** (π_ω) consolidates the strategic memory before updating the prompt.

2.1. Corrective Failures

When errors occur, execution traces contain explicit signals (error messages, stack traces, valid argument lists) that should guide correction. However, static agents treat these as generic “alarms” rather than actionable feedback. We observed over 70 instances where agents misused tools despite the error message explicitly listing valid usage, creating error loops where agents acknowledge failures but repeat the same action. In severe cases, agents respond to errors by fabricating data (*e.g.*, “Let’s assume the file contains...”), *i.e.* a catastrophic failure to manage uncertainty.

2.2. Enhancement Failures

A subtler failure mode occurs when agents miss optimization opportunities even without explicit errors. Agents demonstrate rigid behavior: when search results are poor, context often implies synonyms would help (*e.g.*, “base on balls” vs. “walks”), yet agents persist with single keywords. We also observed silent quality issues, where agents misinterpreting domain concepts (*e.g.*, confusing Eulerian and Hamiltonian paths) because static prompts provide no mechanism for additional verification.

2.3. Complexity of Modern Agentic Systems

Modern benchmarks demand long, multi-role interactions: successful GAIA trajectories average 16.4 steps, with complex tasks exceeding 30 turns. Furthermore, systems comprise specialized sub-agents (*e.g.*, browser, analyzer) with distinct failure patterns, *e.g.* the Web Search agent drives suboptimal strategy errors while the Analyzer accounts for silent quality issues. This heterogeneity implies optimization must be agent-specific. However, this structure also creates an opportunity: agents are invoked repeatedly, forming a natural loop for online learning if lessons can be encoded

into their prompts.

2.4. From Limitations to Solution

Existing Methods and Their Limitations. Prior work on improving agent performance falls into three categories, each with fundamental limitations for modern agentic systems.

Memory-augmented methods accumulate strategies into a playbook incorporated into the solver’s prompt. Examples include Dynamic Cheatsheet (DC) (Suzgun et al., 2025), Agentic Context Engineering (ACE) (Zhang et al., 2025a), and ReasoningBank (Ouyang et al., 2025). These methods are designed for single-query question answering with custom solvers. They operate at task-level granularity, *i.e.* reflecting after each task attempt to update the playbook. This means an agent cannot adapt during a difficult task; if it fails at step 5 of a 30-step task, it must complete (or abandon) the entire task before learning. Furthermore, they maintain a one-for-all playbook that mixes strategies across diverse scenarios and agent roles.

In-context correction methods feed errors and self-generated feedback back into the agent’s context to guide future steps. Reflexion (Shinn et al., 2023) is a classical example, with related approaches including Self-Refine (Madaan et al., 2023) and ReAct (Yao et al., 2022). These mechanisms can be adapted to agent systems. However, feedback is appended to the conversation history rather than integrated into the agent’s instructions. The agent must infer corrections from this feedback, which often leads to error loops where the agent acknowledges mistakes but repeats them. We characterize this as an alarm-based approach: the agent is alerted to errors but not taught how to fix them.

Offline prompt optimization methods search for better prompts before deployment using training sets or

evolutionary algorithms. Classical examples include OPRO (Yang et al., 2023), DSPy (Khattab et al., 2023), and GEPA (Agrawal et al., 2025). While these methods create strong initial prompts, the resulting prompt remains static during inference. If the agent encounters novel contexts or error patterns not represented in the training set, it cannot adapt.

Our Insight: Prompt Evolution for Agents. We propose a fundamentally different approach that leverages the structural properties of agentic systems. Rather than accumulating external strategies (memory-augmented), relying on in-context inference (alarm-based), or optimizing offline (static), we treat the agent’s prompt as an evolvable parameter that improves from execution traces online. Since agents are invoked repeatedly, their prompts can evolve throughout execution, where each invocation is an opportunity to apply learned guidelines. Guidelines synthesized from traces are integrated directly into each agent’s prompt, teaching the agent how to handle specific situations rather than hoping it infers corrections. This approach naturally enables:

- **Step-level adaptation:** Update the prompt during execution, allowing recovery from mid-task failures.
- **Per-agent optimization:** Each agent role evolves its own prompt based on role-specific patterns.

Empirical validation of these design choices is provided in Appendix G.

Problem Formulation. We formalize this as a prompt optimization problem. Consider an agent whose behavior is governed by a prompt θ . The execution generates a trace τ_t containing actions and observations. We treat τ_t as a learning signal, *i.e.* similar to how gradients guide optimization, where errors and suboptimal behaviors indicate how θ should be updated.

Since the prompt space is discrete, we cannot compute gradients directly. Instead, we synthesize a natural language *guideline* g_t from the trace τ_t :

$$\theta_{t+1} = \theta_t \oplus g_t \quad (1)$$

where θ_t is the current prompt, g_t is the synthesized guideline, and \oplus denotes integration. This formulation shifts from static prompt engineering to self-evolving prompts that enhance agent effectiveness.

3. Methodology

We introduce **SCOPE** (Self-evolving Context Optimization via Prompt Evolution), a framework that implements the prompt optimization formulated in Section 2. SCOPE evolves the prompt θ in response to the execution trace τ , synthesizing guidelines that improve agent effectiveness.

Algorithm 1 SCOPE: Self-Evolving Prompt Optimization

```

input Task, Base Prompt  $\theta_{\text{base}}$ , Strategic Memory  $\mathcal{M}_{\text{strat}}$ ,  
Rubrics  $\mathcal{I}$ 
output Completed Task, Updated  $\mathcal{M}_{\text{strat}}$ 
1: Initialize  $\mathcal{M}_{\text{tact}} \leftarrow \emptyset$ ,  $\theta_t \leftarrow \theta_{\text{base}} \oplus \mathcal{M}_{\text{strat}}$ 
2: while Task not completed do
3:   Agent executes with prompt  $\theta_t$ , trace  $\tau_t$  is updated
4:   if Trigger Condition (Error or Sub-task completion) then
5:     // 1. Guideline Synthesis
6:      $G \leftarrow \pi_\phi(\tau_t, \theta_t, \mathcal{I}_{\text{corr/enh}})$  // Generate  $N$  candidates
7:      $g^* \leftarrow \pi_\sigma(G, \theta_t, \mathcal{I}_{\text{sel}})$  // Select best
8:     // 2. Classification & Routing
9:      $(c, \text{conf}) \leftarrow \pi_\gamma(g^*, \theta_t, \mathcal{I}_{\text{cls}})$ 
10:    if  $c = \text{Tactical OR } \text{conf} < c_{\text{thresh}}$  then
11:       $\mathcal{M}_{\text{tact}} \leftarrow \mathcal{M}_{\text{tact}} \cup \{g^*\}$ 
12:    else
13:       $\mathcal{M}_{\text{strat}} \leftarrow \pi_\omega(\mathcal{M}_{\text{strat}} \cup \{g^*\})$  // Optimize
14:    end if
15:    // 3. Update Prompt for next step
16:     $\theta_{t+1} \leftarrow \theta_{\text{base}} \oplus \mathcal{M}_{\text{strat}} \oplus \mathcal{M}_{\text{tact}}$ 
17:  end if
18: end while

```

Our framework, illustrated in Figure 3, consists of four components: (1) **Guideline Synthesis**, (2) **Dual-Stream Routing**, (3) **Memory Optimization**, and (4) **Perspective-Driven Exploration**. The complete procedure is outlined in Algorithm 1.

3.1. Guideline Synthesis

The core of SCOPE is synthesizing guidelines from the execution trace. We use a generator π_ϕ to produce candidate guidelines, followed by a selector π_σ to choose the best one.

Corrective Synthesis. When the agent encounters an error, the generator analyzes the trace using corrective rubrics $\mathcal{I}_{\text{corr}}$ (see Appendix F.1) to synthesize corrective guidelines:

$$g = \pi_\phi(\tau_t, \theta_t, \mathcal{I}_{\text{corr}}). \quad (2)$$

Conditioning on θ_t avoids redundant updates (*e.g.*, re-learning an existing guideline).

Enhancement Synthesis. When execution succeeds but appears suboptimal, the generator uses enhancement rubrics \mathcal{I}_{enh} (see Appendix F.1) to identify inefficiencies and synthesize enhancement guidelines:

$$g = \pi_\phi(\tau_t, \theta_t, \mathcal{I}_{\text{enh}}). \quad (3)$$

Best-of-N Selection. To reduce variance, the generator produces N candidates $G = \{g_1, \dots, g_N\}$. A selector π_σ then chooses the best guideline according to selection rubrics \mathcal{I}_{sel} (see Appendix F.1):

$$g^* = \pi_\sigma(G, \theta_t, \mathcal{I}_{\text{sel}}). \quad (4)$$

3.2. Dual-Stream Routing

After synthesis, each guideline g^* is routed to one of two memory streams based on its scope. We maintain a strategic memory $\mathcal{M}_{\text{strat}} = \{g_1, g_2, \dots\}$ for long-term guidelines and a tactical memory $\mathcal{M}_{\text{tact}}$ for task-specific guidelines.

A classifier π_γ evaluates the guideline’s generality and assigns a confidence score according to classification rubrics \mathcal{I}_{cls} (see Appendix F.2):

$$(c, \text{conf}) = \pi_\gamma(g^*, \theta_t, \mathcal{I}_{\text{cls}}), \quad (5)$$

where $c \in \{\text{Tactical, Strategic}\}$ and $\text{conf} \in [0, 1]$.

Tactical Stream. Guidelines classified as tactical, or strategic with low confidence ($\text{conf} < c_{\text{thresh}}$), are added to the tactical memory:

$$\mathcal{M}_{\text{tact}} \leftarrow \mathcal{M}_{\text{tact}} \cup \{g^*\}. \quad (6)$$

These guidelines are valid only for the current task.

Strategic Stream. High-confidence guidelines ($\text{conf} \geq c_{\text{thresh}}$) identifying general principles are added to the strategic memory:

$$\mathcal{M}_{\text{strat}} \leftarrow \mathcal{M}_{\text{strat}} \cup \{g^*\}. \quad (7)$$

These guidelines persist across tasks.

The prompt evolves by combining both memories:

$$\theta_{t+1} = \theta_{\text{base}} \oplus \mathcal{M}_{\text{strat}} \oplus \mathcal{M}_{\text{tact}}, \quad (8)$$

where θ_{base} is the initial prompt before any guidelines are added.

3.3. Memory Optimization

Unconstrained growth of $\mathcal{M}_{\text{strat}}$ can dilute the agent’s attention with redundant or conflicting guidelines. We apply an optimizer π_ω , which is a multi-step pipeline (see Appendix F.3 for complete rubrics), to consolidate the memory:

$$\mathcal{M}_{\text{strat}} \leftarrow \pi_\omega(\mathcal{M}_{\text{strat}} \cup \{g^*\}). \quad (9)$$

The pipeline consists of three steps: (1) **Conflict Resolution**: merging contradictory guidelines. (2) **Subsumption Pruning**: removing specific guidelines covered by general ones. (3) **Consolidation**: merging similar guidelines into comprehensive ones. See Appendix D.3 for a concrete example.

3.4. Perspective-Driven Exploration

A single evolved prompt may converge to a strategy that works well on some tasks but poorly on others. To increase coverage, we initialize K parallel streams with distinct “Perspectives” (e.g., Efficiency vs. Thoroughness), each evolving its own prompt θ^k . At test time, we select the best

result:

$$\text{Success} = \max_{k \in \{1..K\}} \text{Eval}(\theta^k, \text{task}), \quad (10)$$

where $\text{Eval}(\theta^k, \text{task})$ returns whether the agent with prompt θ^k successfully completes the task. This allows the system to leverage diverse strategies for different problem types.

4. Experiments

4.1. Experimental Setup

Baseline Agent System. We implement a hierarchical agent system with a planning agent that delegates to specialized subordinate agents (web search, analyzer, browser), each equipped with domain-specific tools. We use Gemini-2.5-Pro for the web search and analyzer agents, and GPT-4.1 for the planning and browser agents. Details are in Appendix B.

Benchmarks. We evaluate on three benchmarks requiring multi-step reasoning: **HLE** (Phan et al., 2025) (2,500 expert-level questions), **GAIA** (Mialon et al., 2023) (165 validation questions for General AI Assistants), and **DeepSearch** (Chen et al., 2025). Details are in Appendix C.

SCOPE Configuration. We configure SCOPE with $N = 2$ candidates for Best-of-N synthesis and $K = 2$ parallel streams (Efficiency and Thoroughness) for global exploration. The strategic memory is capped at 10 guidelines per domain, triggering consolidation when exceeded, with a confidence threshold of 0.85 for promotion. All SCOPE components (Generator, Selector, Classifier, Optimizer) utilize GPT-4.1. Additionally, we implement Agent-Specific Optimization, step-level updates, and place Synthesized guidelines in the system prompt (see Section 4.3 for ablation). Detailed hyperparameter settings are in Appendix D.

Prompt Optimization Baselines. We compare SCOPE against two recent prompt optimization methods: (1) **Dynamic Cheatsheet (DC)** (Suzgun et al., 2025): A test-time learning framework that endows LLMs with persistent, evolving memory. We implement the DC-Cu variant, which cumulatively updates memory after processing each input. (2) **Agentic Context Engineering (ACE)** (Zhang et al., 2025a): A playbook-based learning approach that maintains bullet-point strategies across predefined categories and updates them via a reflector-curator loop. We implement ACE adapting it to our baseline agent system with step-level granularity for fair comparison.

4.2. Overall Performance and Ablation

Table 1 and Table 2 present the main experimental results and component analysis. All methods are evaluated with two independent runs per task; a task is considered solved if either run succeeds (Pass@2). SCOPE establishes a new state-of-the-art across all benchmarks, significantly outperforming both the static baseline and recent optimization

Table 1. Performance comparison on three challenging agent benchmarks. We compare SCOPE against state-of-the-art LLMs and a strong hierarchical agent baseline.

| Method | HLE | GAIA | DeepSearch |
|---------------------------|--------------|--------------|--------------|
| <i>Base Models</i> | | | |
| GPT-4.1 | 6.00 | 8.48 | 4.00 |
| Gemini-2.5-Pro | 18.76 | 26.67 | 19.00 |
| <i>Agent Systems</i> | | | |
| Baseline Agent | 14.23 | 32.73 | 14.00 |
| DC (Suzgun et al., 2025) | 18.44 | 35.76 | 21.00 |
| ACE (Zhang et al., 2025a) | 23.72 | 38.18 | 23.00 |
| SCOPE | 38.64 | 56.97 | 32.00 |

methods like DC (Suzgun et al., 2025) and ACE (Zhang et al., 2025a). On the expert-level HLE benchmark, SCOPE more than doubles the baseline performance (38.64% vs. 14.23%).

The ablation study (Table 2) confirms that every component contributes to this success. While the basic Guideline Generator provides the initial boost (+4.85%), the largest gain comes from our Perspective-Driven Exploration ($K = 2$), which adds 10.91%. This underscores that a single context management strategy is insufficient; maintaining diverse optimization streams (Efficiency and Thoroughness) is critical for robust performance across heterogeneous tasks.

4.3. Guideline Placement: System Prompt vs. User Prompt

Table 3 compares four placement strategies on GAIA: (1) all in system prompt, (2) all in user prompt, (3) split (strategic in system, tactical in user), and (4) hybrid (saved in system, online in user). A counter-intuitive finding emerges: system prompt placement achieves the highest accuracy (46.06%) despite more tasks hitting the step limit than user prompt (227 vs. 130). We hypothesize that system prompt guidelines act as *implicit background guidance*, allowing the agent to internalize guidelines and explore more solution paths. Conversely, user prompt placement leads to *over-compliance*—the agent follows instructions too strictly, terminating early rather than continuing exploration, which results in fewer errors but lower task success. The split strategy performs worst (35.76%), as distributing guidelines across both prompts creates conflicting priorities. A detailed behavioral analysis is provided in Appendix H.

4.4. Model Choice for the Evolving Module

A practical deployment question is which model to use for SCOPE’s components (Generator, Classifier, Optimizer). Since our baseline agent uses both GPT-4.1 and Gemini-2.5-Pro for different roles, we evaluate three configurations: (1) all components use GPT-4.1, (2) all use Gemini-2.5-Pro, and (3) each meta-agent uses the same model as the base

Table 2. Ablation study on the GAIA benchmark. Each row represents the cumulative addition of a component to the previous system configuration.

| Configuration | Accuracy (%) | Δ Improvement |
|---------------------------|--------------|----------------------|
| Baseline Agent (Static) | 32.73 | - |
| + Guideline Generator | 37.58 | +4.85 |
| + Dual-Stream Routing | 41.21 | +3.63 |
| + Best-of-N Selection | 44.24 | +3.03 |
| + Memory Optimization | 46.06 | +1.82 |
| + Perspective Exploration | 56.97 | +10.91 |

Table 3. Impact of guideline placement on GAIA. Errors: general execution failures; Max Steps: tasks terminated by step limit; Steps: total agent actions. All placements reduce errors vs. baseline (15–42%), but system prompt achieves the best accuracy.

| Placement | Acc. (%) | Errors | Max Steps | Steps |
|-------------------|--------------|--------------|-----------|-------|
| Baseline | 32.73 | 1,714 | 255 | 9,824 |
| (1) System Prompt | 46.06 | 1,461 (-15%) | 227 | 7,430 |
| (2) User Prompt | 41.21 | 1,000 (-42%) | 130 | 7,408 |
| (3) Split | 35.76 | 1,204 (-30%) | 141 | 7,319 |
| (4) Hybrid | 43.64 | 1,109 (-35%) | 139 | 6,620 |

Table 4. Impact of model choice for SCOPE’s meta-agents on GAIA. Despite generating vastly different numbers of guidelines, all configurations achieve similar accuracy.

| Meta-Agent Model | Acc. (%) | Guidelines | Avg. Length |
|--------------------|--------------|------------|-------------|
| All GPT-4.1 | 46.06 | 111 | 380 chars |
| All Gemini-2.5-Pro | 46.67 | 163 | 426 chars |
| Same as Base Agent | 45.45 | 108 | 397 chars |

agent it optimizes.

Table 4 reveals a surprising finding: all three configurations achieve nearly identical performance (within 1.2%), despite Gemini generating 46% more guidelines than GPT-4.1. This suggests that guideline quality matters more than quantity, *i.e.* SCOPE’s selection and optimization mechanisms effectively filter for useful guidelines regardless of how many candidates are generated. This robustness to model choice simplifies deployment, allowing practitioners to select meta-agent models based on cost or latency rather than accuracy concerns. Further analysis of guideline distribution patterns is provided in Appendix J.

4.5. SCOPE Enhances Robustness in Long-Horizon Domains

Figure 4 breaks down performance by subcategory. On HLE, the largest gains are in knowledge-intensive domains: Biology/Medicine (14.9% → 43.2%) and Chemistry (14.1% → 50.3%), where specialized tool errors are common and SCOPE’s domain-specific guidelines enable recovery. On GAIA Level 3 tasks, *i.e.* the longest trajectories with most noise, SCOPE achieves 30.8% vs 23.1%, confirming that dynamic prompts prevent error propagation and maintain coherence in long-horizon scenarios.

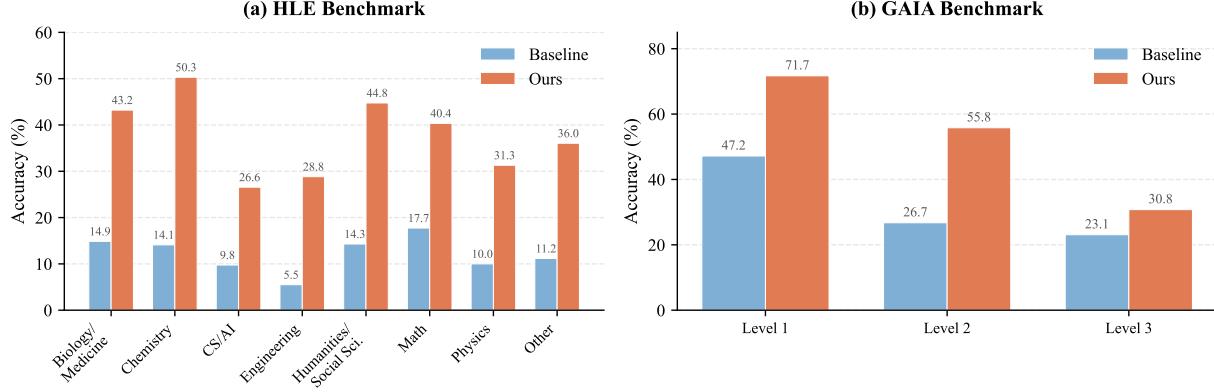


Figure 4. Performance breakdown by subcategory on HLE (a) and GAIA (b). SCOPE (coral) consistently outperforms the baseline (blue) across all domains. Notable gains are observed in domains requiring strict adherence to complex protocols (Biology/Medicine) and on the hardest GAIA tasks (Level 3), where context management is most critical.

5. Analysis

5.1. Mitigating Agentic Fragility via Dual-Mode Synthesis

A counter-intuitive finding in Table 1 is that the Baseline Agent underperforms raw Gemini-2.5-Pro on HLE (14.23% vs. 18.76%) and DeepSearch (14.00% vs. 19.00%). While agentic wrappers enable tool use, they introduce structural fragility: repeated tool errors or infinite loops can cause a capable model to fail tasks it could answer zero-shot. Table 3 quantifies this on GAIA: the baseline accumulates 1,714 errors across 165 tasks.

SCOPE addresses this fragility through two complementary synthesis modes: **Corrective** (reactive error correction) and **Enhancement** (proactive optimization from successful patterns). For instance, a `NameError` triggers a corrective guideline (“Define all variables in code snippets”), while a successful search triggers an enhancement guideline (“Try search term variants”). The corrective mode enables real-time “debugging”, where synthesizing guidelines from failure traces reduces errors by 15–42% depending on placement strategy (Table 3), explaining why SCOPE exceeds base model capabilities.

Crucially, Table 5 shows that enhancement guidelines constitute 61% of all synthesized guidelines, demonstrating that SCOPE is predominantly an optimizer rather than merely an error debugger. This prevalence of proactive guidelines explains why SCOPE continues to improve even when error rates are low, *i.e.* it codifies successful patterns to prevent potential failures before they occur.

5.2. Guideline Compliance: Do Agents Follow Synthesized Guidelines?

A critical question is whether synthesized guidelines actually influence agent behavior. Figure 5 illustrates what we

Table 5. Distribution of synthesized guidelines by mode and domain on GAIA. Enhancement guidelines (61%) dominate, showing SCOPE optimizes beyond error correction.

| Domain | Corrective | Enhancement | Total |
|----------------------|-----------------|-----------------|------------|
| Analysis Methodology | 9 | 22 | 31 |
| Tool Usage | 15 | 13 | 28 |
| Data Validation | 8 | 15 | 23 |
| Efficiency | 2 | 16 | 18 |
| Error Handling | 9 | 2 | 11 |
| Total | 43 (39%) | 68 (61%) | 111 |

term language adoption: after SCOPE synthesizes a guideline recommending “plausible label synonyms and phrase variants”, the agent’s subsequent output incorporates this exact phrasing verbatim. This direct linguistic transfer provides strong evidence that guidelines are actively integrated into the agent’s decision-making, not merely stored.

Beyond linguistic adoption, we observe immediate behavioral changes (typically within seconds) following guideline synthesis, demonstrating real-time adaptation within a single episode. This compliance propagates across all agent types in our hierarchy, with guidelines for the planning agent influencing delegation strategy and guidelines for subordinate agents improving specialized tool usage.

5.3. Perspective Divergence in Guideline Synthesis

Table 6 validates our Perspective-Driven Exploration design by analyzing the two perspectives. While both streams achieve similar total accuracy (44.85% vs 46.06%), their intersection is only 33.94%, *i.e.* approximately 23% of solved problems are unique to one perspective. Notably, Efficiency outperforms on Level 3 tasks (26.92% vs 11.54%), suggesting concise context management is more effective for complex, long-horizon tasks, while Thoroughness excels at Level 2. The Global Ensemble captures the union of these strengths.

This quantitative divergence stems from fundamentally dif-

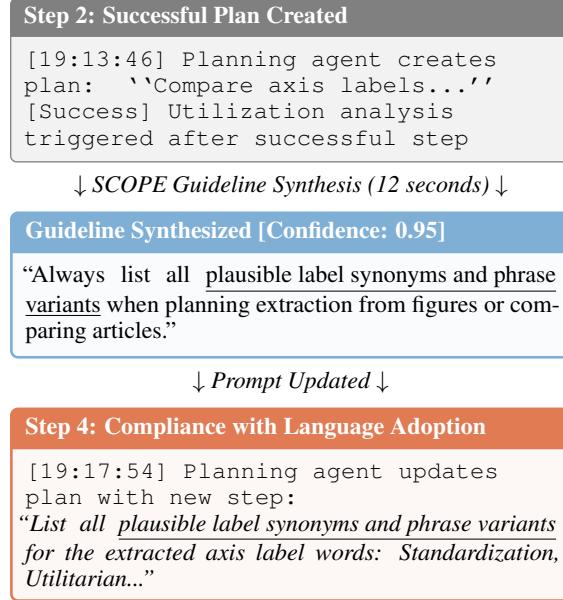


Figure 5. Compliance trace demonstrating *language adoption*. After SCOPE synthesizes a guideline (blue), the agent’s subsequent output (coral) incorporates the exact phrasing (underlined).

Table 6. Performance of individual perspectives vs. Global Ensemble on GAIA. Their low overlap (33.94%) validates the need for diverse optimization streams.

| Perspective | Level 1 | Level 2 | Level 3 | Total |
|--------------------------------|--------------|--------------|--------------|--------------|
| Efficiency Stream | 60.38 | 40.70 | 26.92 | 44.85 |
| Thoroughness Stream | 60.38 | 47.67 | 11.54 | 46.06 |
| Intersection (Overlap) | 49.06 | 32.56 | 7.69 | 33.94 |
| Global Ensemble (Union) | 71.70 | 55.81 | 30.77 | 56.97 |

ferent guideline strategies. As illustrated in Figure 6, when the browser agent encounters access denials (HTTP 403), the Efficiency stream learns to “fail-over fast” by escalating to the web search agent, while the Thoroughness stream synthesizes guidelines to “find alternate sources” (e.g., Archive.org). This duality allows SCOPE to handle both time-constrained and deep retrieval tasks. A detailed qualitative comparison of strategy patterns is provided in Appendix I.

6. Related Work

We position SCOPE at the intersection of context management, prompt optimization, and agentic learning.

Context Augmentation and Compression. The primary challenge in agentic workflows is handling extensive information. Traditional approaches address this via retrieval and compression. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) fetches relevant context, while methods like LLM Lingua (Jiang et al., 2023; Sun et al., 2025) optimize efficiency via compression. They focus on what context to provide, whereas SCOPE focuses on how

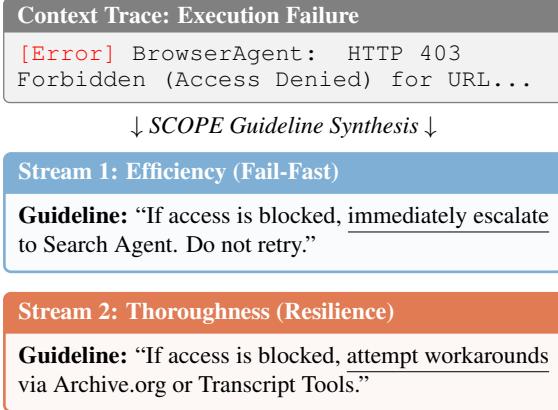


Figure 6. Divergence in evolved guidelines. From the same error, Efficiency (blue) optimizes for speed, while Thoroughness (coral) optimizes for success.

the agent processes context by evolving prompt.

Prompt Optimization. A significant body of work focuses on prompt engineering. Offline methods like OPRO (Yang et al., 2023), DSPy (Khattab et al., 2023), and GEPA (Agrawal et al., 2025) search for optimal prompts before deployment using training sets. While these create strong initializations, the resulting prompt remains static during inference. SCOPE complements these by enabling online adaptation: the prompt evolves during execution based on step-level feedback.

Agent Memory and Learning. Recent work gives agents persistent memory to accumulate experience. Memory-augmented methods like DC (Suzgun et al., 2025), ACE (Zhang et al., 2025a), and ReasoningBank (Ouyang et al., 2025) build external strategy libraries. In-context correction methods like Reflexion (Shinn et al., 2023) and Self-Refine (Madaan et al., 2023) feed feedback into the context. As discussed in Section 2.4, they face limitations in agentic settings: task-level granularity, one-for-all memory, and alarm-based correction. SCOPE addresses these with step-level adaptation, per-agent optimization, and direct prompt evolution.

7. Conclusion

We introduced **SCOPE**, a framework that enables prompt evolution for enhancing agent effectiveness. By treating execution traces as learning signals, SCOPE synthesizes guidelines that are integrated directly into agent prompts, enabling step-level adaptation and per-agent optimization. Experiments on HLE and GAIA demonstrate that SCOPE significantly outperforms static baselines and existing methods, more than doubling success rates in expert-level domains. Our findings suggest a new direction: rather than engineering static prompts, we should enable agents to evolve their own prompts online.

References

- Agrawal, L. A., Tan, S., Soylu, D., Ziems, N., Khare, R., Opsahl-Ong, K., Singhvi, A., Shandilya, H., Ryan, M. J., Jiang, M., et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*, 2025.
- Anthropic. Claude code. 2025.
- Chen, K., Ren, Y., Liu, Y., Hu, X., Tian, H., Xie, T., Liu, F., Zhang, H., Liu, H., Gong, Y., et al. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *arXiv preprint arXiv:2506.13651*, 2025.
- Google. Deep research is now available on gemini 2.5 pro experimental. <https://blog.google/products/gemini/deep-research-gemini-2-5-pro-experimental/>, 2025.
- Jiang, H., Wu, Q., Lin, C.-Y., Yang, Y., and Qiu, L. Llmling-gua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*, 2023.
- Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.
- Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhaman, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.
- Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.
- OpenAI. Deep research system card. <https://cdn.openai.com/deep-research-system-card.pdf>, 2025.
- Ouyang, S., Yan, J., Hsu, I., Chen, Y., Jiang, K., Wang, Z., Han, R., Le, L. T., Daruki, S., Tang, X., et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. *arXiv preprint arXiv:2509.25140*, 2025.
- Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*, 2025.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- Sun, W., Lu, M., Ling, Z., Liu, K., Yao, X., Yang, Y., and Chen, J. Scaling long-horizon llm agent via context-folding. *arXiv preprint arXiv:2510.11967*, 2025.
- Suzgun, M., Yuksekgonul, M., Bianchi, F., Jurafsky, D., and Zou, J. Dynamic cheatsheet: Test-time learning with adaptive memory. *arXiv preprint arXiv:2504.07952*, 2025.
- Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Team, T. D., Li, B., Zhang, B., Zhang, D., Huang, F., Li, G., Chen, G., Yin, H., Wu, J., Zhou, J., et al. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025.
- Wei, J., Sun, Z., Papay, S., McKinney, S., Han, J., Fulford, I., Chung, H. W., Passos, A. T., Fedus, W., and Glaese, A. Browsecamp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.
- Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.
- Zhang, Q., Hu, C., Upasani, S., Ma, B., Hong, F., Kamanuru, V., Rainton, J., Wu, C., Ji, M., Li, H., et al. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*, 2025a.
- Zhang, W., Cui, C., Zhao, Y., Hu, R., Liu, Y., Zhou, Y., and An, B. Agentorchestra: A hierarchical multi-agent framework for general-purpose task solving. *arXiv e-prints*, pp. arXiv–2506, 2025b.

Appendix Overview

This appendix provides supplementary materials organized as follows:

- **Appendix A: Detailed Analysis of Baseline Failure Modes** — Comprehensive taxonomy and examples of agent failures from over 1.5 million lines of execution logs.
- **Appendix B: Baseline Agent System Details** — Architecture and configuration of our hierarchical multi-agent system.
- **Appendix C: Benchmark Details** — Description of HLE, GAIA, and DeepSearch evaluation benchmarks.
- **Appendix D: Hyperparameter Settings** — Complete hyperparameters, domain categories, optimization pipeline details, and a concrete memory optimization example.
- **Appendix E: Perspective-Driven Exploration Details** — Implementation of Efficiency and Thoroughness optimization streams.
- **Appendix F: Complete System Prompts** — Full rubrics for all meta-agents (Generator, Selector, Classifier, Optimizer).
- **Appendix G: Per-Agent Specialization and Step-Level Updates** — Empirical validation of per-agent and step-level design choices.
- **Appendix H: Guideline Placement Analysis** — Deep dive into system vs. user prompt placement effects.
- **Appendix I: Qualitative Comparison of Perspective Streams** — Detailed strategy patterns for Efficiency vs. Thoroughness streams.
- **Appendix J: Model Choice for Meta-Agents** — Detailed analysis of how different model choices affect guideline generation patterns.

A. Detailed Analysis of Baseline Failure Modes

In this appendix, we provide a more granular analysis of the failure patterns observed in the baseline agent logs. Our dataset consists of over 1.5 million lines of execution traces from the GAIA and DeepSearch benchmarks. We categorize these failures into two modes based on their trigger: *Corrective Failures* (error-triggered) and *Enhancement Failures* (optimization opportunities when no error occurs). Table 7 provides a summary taxonomy.

A.1. Corrective Failures: Systematic Error Repetition

A.1.1. PERSISTENT TOOL MISUSE

A recurring pattern in our logs is the agent’s inability to internalize tool schema constraints even after explicit feedback.

- **Incorrect Tool Identifiers:** In 70 recorded instances, the agent attempted to call a tool using a generic name (*e.g.*, `final_answer`) instead of the specific identifier defined in the environment (*e.g.*, `final_answer_tool`). Despite the environment returning an error message listing the valid tools, the agent failed to update its behavior in subsequent steps.
- **Parameter Hallucination:** We observed 29 cases where the agent hallucinated parameters for planning tools. For example, it repeatedly passed a “notes” argument to a function that only accepted a task description, causing execution failures. The static prompt provided no mechanism to “remember” that this parameter was invalid.

A.1.2. IGNORED ENVIRONMENTAL CONSTRAINTS

The agent frequently encountered environment-specific restrictions (*e.g.*, forbidden libraries) but failed to learn from them.

- **Forbidden Imports:** When the agent attempted to import restricted Python libraries (*e.g.*, `pandas`), the environment returned a `ModuleNotFoundError` or a security exception, often accompanied by a whitelist of allowed libraries (*e.g.*, `collections`, `re`). In multiple traces, the agent ignored this whitelist and simply attempted to import a different forbidden library in the next step, demonstrating a complete lack of adaptation to the explicitly stated constraints.
- **Repeated Timeouts:** In 310 instances, the agent reached the maximum step limit. Analysis of these traces reveals that the agent often entered infinite loops of “try-fail-retry” using the exact same parameters, rather than switching strategies after a timeout warning.

A.2. Enhancement Failures: Suboptimal Behavioral Patterns

A.2.1. INEFFICIENCY AND REDUNDANCY

Even when the agent succeeded, its path to the solution was often highly inefficient.

- **Redundant Verification:** The agent’s plans consistently included steps to “cross-check results” or “ver-

Table 7. Taxonomy of agent failures observed in over 1.5 million lines of baseline agent logs. The agent often has the necessary information in its context but fails to translate it into correct actions due to a static instruction set.

| CATEGORY | FAILURE PATTERN | DESCRIPTION | COUNT |
|----------------------|-------------------------|--|-----------|
| Corrective Failures | Tool Misuse | Context shows instructions, but agent uses generic actions | >70 |
| | Constraint Neglect | Context lists allowed resources; agent ignores list and retries forbidden ones | Recurrent |
| | Parameter Hallucination | Context defines arguments, but agent invents non-existent ones | 29 |
| | Error Loop | Context shows error message, agent repeats action without modification | Multiple |
| Enhancement Failures | Redundant Verification | Agent re-verifies facts already present and confirmed in context | – |
| | Single-Term Bias | Context implies broad search needed; agent sticks to single keyword | Recurrent |
| | Generic Strategy | Agent persists with inefficient strategy despite context clues | Recurrent |
| Severe Cases | Data Fabrication | Agent invents data when retrieval context is empty/failed | Recurrent |
| | Domain Misconception | Agent misinterprets technical context due to lack of domain guidelines | Recurrent |

ify with an additional source”, even for simple retrieval tasks where the primary source was authoritative. This heuristic, likely derived from safety-focused pre-training, inflated token costs and execution time without improving accuracy.

- **Extreme Verbosity:** We observed a tendency for the agent to wrap simple numeric or boolean answers in verbose explanations. In extreme cases, a single-digit answer was accompanied by over 1,000 tokens of methodological description and hypothetical context, wasting computational resources.
- **Sequential Querying Anti-Pattern:** For tasks involving multiple similar entities (*e.g.*, “Find salaries for Players A, B, and C”), the agent consistently executed N separate tool calls rather than a single batched query. This multiplied the latency and token costs linearly with the number of entities.

A.2.2. SEARCH AND RETRIEVAL BIAS

We identified a prevalent “Single-Terminology Bias” in the agent’s search strategy.

- **Keyword Rigidity:** The agent frequently failed to use necessary synonyms or domain-specific variants. For example, when searching for baseball statistics, it used only “walks” and missed data indexed under “base on balls” or “BB”.
- **Monolingual Blindness:** In cross-lingual tasks, the agent often searched only in English, neglecting native-script terms (*e.g.*, using “Dai Wangshu” but not the native Chinese characters). This significantly reduced recall for region-specific queries.
- **Missing Fallback Protocols:** Upon encountering dead links or blocked content, the agent rarely attempted to

use archival services (*e.g.*, archive.org) or alternative reputable sources, instead treating the block as a hard failure condition.

A.2.3. CONTEXT CONTAMINATION

In multi-turn sessions, we observed issues with context management.

- **Token Explosion:** Without active context pruning, the agent carried the full history of long interaction chains, leading to input contexts exceeding 100,000 tokens. This not only increased latency but also degraded performance as relevant instructions became diluted in the context window.

A.3. Domain-Specific Knowledge Failures

Beyond mechanical errors, we identified a class of “knowledge-level” failures where the agent operated syntactically correctly but relied on flawed reasoning or incorrect domain knowledge.

- **Statistical Misconceptions:** In scientific tasks, the agent frequently misinterpreted statistical concepts. For example, one trace shows the agent conflating a p-value of 0.04 with a “4% chance of the result being wrong”, revealing a fundamental misunderstanding of hypothesis testing that led to incorrect conclusions.
- **Technical Ambiguity:** When faced with ambiguous technical terms, the agent often defaulted to incorrect definitions without verification. In a graph theory task, the agent confused “Eulerian circuits” (visiting every edge) with “Hamiltonian circuits” (visiting every vertex), leading to a mathematically valid but task-irrelevant solution.

- **Entity Conflation:** In humanities tasks, the agent demonstrated a tendency to conflate related historical figures. For instance, it attributed a specific poem to *Hu Shi* (a proponent of the vernacular movement) instead of *Dai Wangshu* (the actual author), likely due to their semantic proximity in the training data.
- **Jargon Hallucination:** Most deceptively, the agent occasionally fabricated plausible-sounding technical definitions. In one instance, it defined a “mean tangle” using sophisticated physics jargon (“average state of quantum entanglement in non-linear systems”), despite the term being non-existent in that context. These hallucinations are particularly dangerous as they can mislead non-expert users.

A.4. Efficiency Anti-Patterns

In addition to general verbosity, we observed specific operational patterns that needlessly consumed steps and tokens.

- **Sequential Querying:** When tasked with retrieving data for multiple entities (*e.g.*, salaries for three NBA players), the agent consistently performed sequential tool calls—waiting for one result before requesting the next—rather than batching them into a single parallel execution.
- **Unnecessary Tool Delegation:** The agent frequently offloaded trivial operations to external tools. For example, we observed the agent invoking a Python interpreter tool solely to perform simple arithmetic (*e.g.*, $1468.88 - 1430.08$) that it could have computed internally, wasting a full execution turn.

A.5. Severe Cases: Safety Risks and Hallucination

A.5.1. FABRICATION UNDER UNCERTAINTY

The most critical safety finding from our analysis is the agent’s tendency to fabricate information when tools fail.

- **Scenario:** When the agent encountered unreadable files (*e.g.*, corrupted PDFs or inaccessible URLs) and could not retrieve the necessary data.
- **Behavior:** Instead of reporting a failure, the agent often generated a “hypothetical” analysis. For example, logs show the agent stating: “*Since the automated tools cannot read the files... Let's assume the applicant data contains...*”
- **Outcome:** The agent then proceeded to produce a confident, formatted final answer based entirely on this hallucinated data. This behavior poses a significant risk in real-world deployment, as the final output appears legitimate despite being grounded in fiction.

A.6. Environmental Barriers

A.6.1. CAPTCHA AND BLOCKING LOOPS

The agent demonstrated poor adaptability to web access barriers.

- **Repetitive Retries:** Upon encountering CAPTCHAs or “Access Denied” pages, the agent frequently attempted to “click” the verify button or refresh the page using the browser tool.
- **Failure to Pivot:** In over 180 instances, the agent exhausted its retry limit on a blocked site without attempting to switch to an alternative data source (*e.g.*, an archived version or a different domain), resulting in task failure.

B. Baseline Agent System Details

We implement a hierarchical multi-agent system as our baseline. The architecture consists of a top-level planning agent that orchestrates task decomposition and delegates sub-tasks to specialized subordinate agents.

B.1. Agent Specifications

Table 8. Agent configurations in our hierarchical system.

| Agent | Model | Max Steps | Tools |
|------------------|----------------|-----------|----------------------------|
| Planning Agent | GPT-4.1 | 20 | Task delegation |
| Web Search Agent | Gemini-2.5-Pro | 3 | Search API, Python |
| Analyzer Agent | Gemini-2.5-Pro | 3 | Analysis, Python |
| Browser Agent | GPT-4.1 | 5 | Browser automation, Python |

Planning Agent. The planning agent serves as the top-level orchestrator. It receives the original task, decomposes it into sub-tasks, and delegates them to appropriate subordinate agents. The planning agent maintains a task queue and tracks completion status. It uses GPT-4.1 and operates with a maximum of 20 planning steps.

Web Search Agent. The web search agent specializes in information retrieval from the internet. It is equipped with search APIs for querying web content and a Python interpreter for processing search results. The agent uses Gemini-2.5-Pro and operates with a maximum of 3 steps per sub-task.

Analyzer Agent. The analyzer agent performs systematic, step-by-step reasoning and analysis. It handles tasks requiring mathematical computation, logical deduction, and structured analysis. The agent uses Gemini-2.5-Pro with access to analysis tools and a Python interpreter.

Browser Agent. The browser agent enables direct web page interaction and navigation. It can click elements, fill forms, scroll pages, and extract content from dynamic web pages.

The agent uses GPT-4.1 and operates with a maximum of 5 steps to handle complex web interactions.

C. Benchmark Details

C.1. Humanity’s Last Exam (HLE)

Humanity’s Last Exam (HLE) (Phan et al., 2025) is a comprehensive multi-modal benchmark designed to evaluate the reasoning capabilities of large language models at an expert level. The benchmark comprises 2,500 questions contributed by domain experts across diverse fields.

Subject Distribution. HLE covers a broad spectrum of disciplines including:

- **STEM:** Mathematics (algebra, calculus, topology), Physics (quantum mechanics, thermodynamics), Chemistry (organic, inorganic), Biology (molecular, evolutionary), Computer Science (algorithms, complexity theory), Engineering
- **Humanities:** History, Philosophy, Literature, Linguistics
- **Social Sciences:** Economics, Psychology, Political Science

Question Format. The benchmark includes both multiple-choice and short-answer questions. A significant portion of questions require multi-modal understanding, incorporating images, diagrams, equations, and tables. Questions are designed to test genuine understanding rather than pattern matching or memorization.

Difficulty Level. HLE questions are calibrated to be challenging even for domain experts. The benchmark specifically targets the “last exam” frontier—problems at the boundary of what current AI systems can reliably solve.

C.2. GAIA

GAIA (General AI Assistants) (Mialon et al., 2023) is a benchmark designed to evaluate AI assistants on real-world tasks requiring multiple fundamental capabilities.

Task Characteristics. The benchmark comprises 466 questions (301 test, 165 validation). In our experiments, we evaluate on the validation set since test set labels are not publicly available. The tasks have the following characteristics:

- **Multi-step reasoning:** Tasks require chaining multiple logical steps
- **Tool use:** Effective use of web search, calculators, and other tools
- **Multi-modality:** Some tasks involve images, audio, or video content

- **Web browsing:** Information retrieval from live websites
- **Real-world grounding:** Tasks based on actual facts and current information

Difficulty Levels. GAIA organizes tasks into three difficulty levels:

- **Level 1:** Simple tasks requiring 1-2 steps
- **Level 2:** Moderate tasks requiring 3-5 steps
- **Level 3:** Complex tasks requiring extensive reasoning and tool use

Design Philosophy. A key insight from GAIA is that tasks conceptually simple for humans can be remarkably challenging for AI systems. The benchmark emphasizes robustness and reliability over raw capability.

C.3. DeepSearch

DeepSearch (Chen et al., 2025) is a benchmark specifically designed to evaluate deep search and information synthesis capabilities of AI systems.

Task Types. The benchmark includes:

- **Multi-hop retrieval:** Finding information that requires traversing multiple sources
- **Cross-document synthesis:** Combining information from disparate documents
- **Temporal reasoning:** Answering questions about events across time
- **Comparative analysis:** Comparing entities across multiple dimensions

Evaluation Metrics. Tasks are evaluated on both accuracy of the final answer and the quality of the reasoning chain leading to the answer.

D. Hyperparameter Settings

Table 9 summarizes the hyperparameters used in our dynamic prompt evolution framework.

D.1. Guideline Domain Categories

We categorize learned guidelines into 7 semantic domains to facilitate organization and optimization:

1. **tool_usage:** Guidelines for correct tool invocation, argument formatting, and sequencing
2. **data_validation:** Guidelines for validating inputs, outputs, and intermediate data

Table 9. Hyperparameter settings for our method.

| Hyperparameter | Value | Description |
|------------------------------|---------|--|
| <i>Best-of-N Selection</i> | | |
| <i>N</i> (candidates) | 2 | Number of candidate guidelines generated |
| Candidate models | GPT-4.1 | Models for guideline generation |
| Selector model | GPT-4.1 | Model for selecting best candidate |
| <i>Classifier</i> | | |
| Classifier model | GPT-4.1 | Model for strategic vs tactical classification |
| <i>Guideline Management</i> | | |
| Optimization model | GPT-4.1 | Model for guideline consolidation/subsumption |
| Max guidelines per domain | 10 | Threshold triggering optimization |
| Target count (post-opt) | 8 | Target guidelines after optimization (80%) |
| Max iterations | 2 | Maximum optimization passes |
| Max guidelines per task | 20 | Limit on tactical guidelines per task |
| <i>Confidence Thresholds</i> | | |
| Strategic promotion | 0.85 | Min. confidence for strategic memory |
| Auto-accept threshold | 0.5 | Min. confidence to apply guideline |
| <i>Analysis Settings</i> | | |
| Quality analysis freq. | 1 | Analyze every <i>N</i> successful steps |

3. **error handling:** Strategies for error recovery, retries, and fallback mechanisms
4. **efficiency:** Optimizations for speed, cost, and resource utilization
5. **analysis methodology:** Core problem-solving strategies, verification methods, and scientific reasoning
6. **safety:** Guidelines for preventing harmful outcomes and maintaining reliability
7. **general:** High-quality guidelines that do not fit specific categories

D.2. Guideline Optimization Pipeline

When the number of guidelines in a domain exceeds the maximum threshold (10), we trigger an automatic optimization pipeline:

1. **Analysis:** Identify consolidation opportunities, subsumption relationships, and conflicts
2. **Conflict Resolution:** Resolve contradictory guidelines by synthesizing or selecting the better guideline
3. **Subsumption Pruning:** Remove specific guidelines that are covered by more general ones
4. **Consolidation:** Merge semantically similar guidelines into comprehensive ones

The optimization targets 80% of the maximum capacity (8 guidelines) to leave room for future learning while maintaining a compact, high-quality guideline set.

D.3. Memory Optimization Example

We illustrate the memory optimization pipeline with a realistic example from the `efficiency` domain.

Trigger Condition. Suppose the `efficiency` domain currently contains 10 guidelines (at the limit). When a new guideline g^* is synthesized and classified as strategic with

domain `efficiency`, the domain now has 11 guidelines, exceeding the maximum threshold of 10. This triggers the optimization pipeline.

Before Optimization (11 guidelines).

- R1:** Use batch operations instead of sequential queries when fetching multiple items.
- R2:** When searching for multiple entities, combine them into a single query.
- R3:** Avoid redundant API calls by caching results locally.
- R4:** Cache intermediate results to avoid re-computation.
- R5:** Prefer local computation over tool calls for simple arithmetic.
- R6:** Do not invoke Python interpreter for basic math like addition or subtraction.
- R7:** Limit verbose explanations in final answers to under 100 words.
- R8:** Keep final outputs concise; avoid unnecessary elaboration.
- R9:** When a search returns sufficient results, stop querying additional sources.
- R10:** Use early termination when the answer is found; do not continue searching.
- R11:** (new) Batch similar file reads into a single operation.

Step 1: Conflict Resolution. The analyzer identifies no direct conflicts in this set (guidelines give consistent guidance).

Step 2: Subsumption Pruning. The analyzer identifies:

- **R1** (general batch operations) subsumes **R2** (specific to search queries) and **R11** (specific to file reads).
- **R5** (prefer local computation) subsumes **R6** (specific to Python interpreter for math).

After verification, **R2**, **R6**, and **R11** are removed. Remaining: 8 guidelines.

Step 3: Consolidation. The analyzer identifies consolidation opportunities:

- **R3** and **R4** both address caching → merged into a single guideline.
- **R7** and **R8** both address output conciseness → merged.
- **R9** and **R10** both address early termination → merged.

After Optimization (5 guidelines).

- R1:** Use batch operations instead of sequential queries when fetching multiple items.
- R2:** Cache intermediate results and API responses to avoid redundant computation and calls.
- R3:** Prefer local computation over tool calls for simple operations.
- R4:** Keep final outputs concise (under 100 words); avoid unnecessary elaboration.
- R5:** Use early termination when sufficient information is found; do not continue searching.

The optimization reduced 11 guidelines to 5, well below the target of 8, while preserving all essential guidance. The remaining capacity allows continued learning without immediate re-triggering of optimization.

E. Perspective-Driven Exploration Details

This appendix details the implementation of our Perspective-Driven Exploration. We execute $K = 2$ parallel evolutionary streams, each guided by a distinct optimization persona defined via the system prompt.

E.1. Stream 1: Efficiency Persona

The *Efficiency Stream* utilizes our efficiency-focused synthesis prompt (see **Enhancement Synthesis Prompt (Efficiency)** in Appendix F), which optimizes for operational speed, token economy, and conciseness. It specifically targets redundant tool calls and verbose outputs.

E.2. Stream 2: Thoroughness Persona

The *Thoroughness Stream* utilizes our thoroughness-focused synthesis prompt (see **Enhancement Synthesis Prompt (Thoroughness)** in Appendix F), designed to foster self-evolving domain expertise. It prioritizes correctness, deep reasoning, and the acquisition of domain-specific heuristics over raw speed.

F. Complete System Prompts

This appendix provides the full system prompts for all meta-agents used in our framework, including the Generator, Classifier, and Guideline Optimizer agents.

F.1. Guideline Generator Prompts

Corrective Synthesis Prompt. This prompt is used to generate corrective guidelines when the agent encounters an error.

Corrective Synthesis Prompt

You are a prompt engineering expert analyzing agent execution errors.

Your task: Generate a SHORT, TARGETED system prompt addition (1-3 lines) that will help prevent this error in the future.

Context:

- Agent Name: {agent_name}
- Agent Role: {agent_role}
- Task: {task}
- Error Type: {error_type}
- Error Message: {error_message}

Previous actions taken:
{last_step_summary}

Current system prompt (for reference, to avoid duplication):
{current_system_prompt}

Already applied rules (DO NOT duplicate these):
{applied_rules}

Guidelines:

1. Be SPECIFIC and ACTIONABLE – target the exact error cause
2. Be BRIEF – max 1-3 lines
3. Use imperative language ("Always...", "Never...", "When X, do Y...")
4. Don't repeat what's already in the current system prompt
5. Focus on formatting, structure, or procedure constraints

Output ONLY valid JSON with this exact format:

```
{}  
"update_text": "The actual prompt  
addition text here",  
"rationale": "Brief 1-sentence why  
this helps",  
"confidence": "low|medium|high"  
}}
```

Enhancement Synthesis Prompt (Efficiency). This prompt is used for the Efficiency Stream to optimize operational speed and conciseness.

Enhancement Synthesis Prompt (Efficiency)

You are a prompt engineering expert analyzing agent execution quality.

Your task: Analyze this successful step and determine if there are inefficiencies or areas for improvement. If found, generate a **SHORT, TARGETED** system prompt addition (1-3 lines).

Context:

- Agent Name: {agent_name}
- Agent Role: {agent_role}
- Task: {task}
- Step Number: {step_number}

Step details:

{last_step_summary}

Current system prompt (for reference):

{current_system_prompt}

Already applied rules (DO NOT duplicate these):

{applied_rules}

Analyze for:

1. ****Inefficient tool usage**:**
Multiple calls when one would suffice, redundant operations
2. ****Verbose outputs**:**
Unnecessarily long reasoning or outputs
3. ****Missing best practices**:**
Not following domain-specific best practices
4. ****Suboptimal approaches**:** Using less efficient methods when better ones exist

Guidelines:

- Only suggest an update if there's a CLEAR, ACTIONABLE improvement
- Be SPECIFIC about what to improve
- Be BRIEF - max 1-3 lines
- Use imperative language ("Always...", "Prefer...", "When X, use Y instead of Z...")
- Don't repeat what's already in the current system prompt

Output ONLY valid JSON with this exact format:

```
{}  
"update_text": "The actual prompt  
addition text here (or empty string  
if no improvement needed)",  
"rationale": "Brief 1-sentence  
why this helps (or 'No improvement  
needed')",  
"confidence": "low|medium|high"  
}}
```

Enhancement Synthesis Prompt (Thoroughness). This prompt is used for the Thoroughness Stream to foster self-evolving expertise and correctness.

Enhancement Synthesis Prompt (Thoroughness)

You are a prompt engineering expert analyzing agent execution quality.

Your task: Analyze this successful step and determine if there are inefficiencies or areas for improvement. If found, generate a **SHORT, TARGETED** system prompt addition (1-3 lines).

[Context parameters same as above]

Analyze for improvements in these dimensions:

1. **Correctness & Logic**: Assumptions, edge cases, sound approach, validation checks.
2. **Domain-Specific Strategies**: Terminology variants, authoritative sources, domain heuristics.
3. **Strategic Planning & Approach**: Problem decomposition, simpler methods first, batch operations.
4. **Tool Usage Efficiency**: Consolidation, redundancy reduction, local computation.
5. **Information Preservation**: Context tracking, evidence citation, intermediate results.
6. **Robustness & Error Recovery**: Fallback strategies, retry logic, error anticipation.
7. **Output Quality**: Verbosity control, structure, parseability.

Guidelines:

- PRIORITIZE correctness over efficiency
- Only suggest if there's a CLEAR, ACTIONABLE, GENERALIZABLE improvement
- For domain rules: Include SPECIFIC terms/sources/values, not placeholders
- Be BRIEF - max 1-3 lines
- Use imperative language ("Always...", "Prefer...", "When X, do Y...")
- Don't repeat what's already in the current system prompt
- Look for PATTERNS that generalize beyond this single instance

Output ONLY valid JSON with this exact format:

```
{
  "update_text": "The actual prompt addition text here (or empty string if no improvement needed)",
  "rationale": "Brief 1-sentence why this helps (or 'No improvement needed')",
  "confidence": "low|medium|high"
}
```

Best-of-N Selector Prompt. This prompt is used to select the best candidate update from multiple generated options.

Selector Prompt

You are a prompt engineering expert evaluating multiple candidate prompt updates.

Your task: Select the **BEST** candidate update from the options below.

Context:

- Agent Name: {agent_name}
- Agent Role: {agent_role}
- Task: {task}
- Issue Type: {issue_type}

Issue Details:
{issue_details}

Current system prompt (for reference):
{current_system_prompt}

Candidate Updates:
{candidates}

Evaluation Criteria (in priority order):

1. **Specificity & Relevance**: Most directly addresses the actual issue/error
2. **Actionability**: Clear, implementable instructions that the agent can follow
3. **Generalizability**: Useful beyond just this instance, but not too vague
4. **Brevity**: Concise and clear without unnecessary words
5. **Non-duplication**: Doesn't repeat what's already in the system prompt

Select the candidate that best balances these criteria for the current situation.

Output ONLY valid JSON with this exact format:

```
{
  "selected_index": 0,
  "rationale": "Brief 1-2 sentence explanation of why this candidate is best"
}
```

F.2. Classifier Agent Prompt

This prompt is used to classify updates as Tactical or Strategic and check for duplicates.

Classifier Prompt

```
You are a rule classifier. Analyze
the proposed update and determine:

1. **Is it a DUPLICATE/REDUNDANT?**  

Check if it's already covered by
existing strategic or tactical rules.  

2. **What is its SCOPE?**  

- STRATEGIC: General best practice
that applies across different tasks
(e.g., "Always validate inputs", "Use
batch operations when possible")
- TACTICAL: Task-specific constraint
for current task only (e.g., "This
dataset has missing values in column
X", "API rate limit is 100/min")  

3. **Refined CONFIDENCE**: Assess
confidence (0.0-1.0) based on how
actionable and useful this rule is.  

4. **DOMAIN**: If strategic,
you MUST categorize it into ONE
of the following allowed domains:  

{allowed_domains}

==== PROPOSED UPDATE ====  

Update: {update_text}  

Rationale: {rationale}  

Initial Confidence:  

{initial_confidence}

{all_rules_context}

==== YOUR ANALYSIS ====  

Respond in JSON format:  

{{  

  "is_duplicate": true/false,  

  "scope": "strategic" or "tactical",  

  "confidence": 0.0-1.0,  

  "domain": "domain_name" (only if
  scope is strategic, otherwise ""),  

  "reason": "Brief explanation of your
  classification"
}}
```

Guideline Analyzer Prompt

You are a rule optimization analyzer.
Analyze these {count} rules and
identify optimization opportunities.

{rules_text}

Your task is to identify three types
of optimization opportunities:

1. **CONFLICTS**: Pairs of rules
 that give contradictory guidance.
 - PRIORITY: Highest - conflicts must
 be resolved first
2. **SUBSUMPTION**: Pairs where
 a specific rule is entirely covered
 by a more general rule.
 - Format: [general_rule_index,
 specific_rule_index]
 - PRIORITY: High - clear redundancy
 should be removed
3. **CONSOLIDATION**: Groups of
 rules that express similar concepts
 and can be merged into a single, more
 comprehensive rule.
 - PRIORITY: Medium - merge rules
 that address the same concern from
 different angles

IMPORTANT PRIORITY RULES:

1. If two rules CONFLICT, do NOT
 also mark them for consolidation or
 subsumption
2. If one rule SUBSUMES another, do
 NOT also mark them for consolidation
3. A rule pair should appear in AT
 MOST ONE category

You MUST respond with ONLY valid JSON
in this exact format:

```
{{  

  "consolidation": [[idx1, idx2],  

    [idx3, idx4]],  

  "subsumption": [[general_idx,  

    specific_idx]],  

  "conflicts": [[idx1, idx2]]  

}}
```

F.3. Guideline Optimization Prompts

Guideline Analyzer Prompt. Identifies optimization opportunities among existing guidelines.

Consolidation Prompt. Merges similar guidelines into a single comprehensive guideline.

Consolidation Prompt

You are merging similar rules into one comprehensive rule.

Original rules to merge:
`{rules.text}`

Create a single, comprehensive rule that:

1. Captures all the key guidance from the original rules
2. Is clear and actionable
3. Eliminates redundancy
4. Maintains the original intent

Also provide a brief rationale explaining what the merged rule accomplishes.

Return JSON:
`{}
"rule": "The merged rule text",
"rationale": "Brief explanation of what this rule accomplishes"
{}}`

Conflict Resolver Prompt

You are resolving a conflict between two rules.

Rule {idx1}:
Text: `{rule1.text}`
Rationale: `{rule1.rationale}`

Rule {idx2}:
Text: `{rule2.text}`
Rationale: `{rule2.rationale}`

These rules appear to conflict. Your task is to:

1. Verify the conflict is real
2. Synthesize a single rule that reconciles both OR pick the better rule with justification

Return JSON:
`{}
"rule": "The resolved rule text",
"rationale": "Explanation of how this resolves the conflict"
{}}`

Subsumption Verification Prompt. Verifies if a general guideline completely covers a specific guideline.

Subsumption Verification Prompt

Verify if the general rule subsumes the specific rule.

General Rule: `{general.rule}`
Specific Rule: `{specific.rule}`

A rule is subsumed if:

- Following the general rule AUTOMATICALLY means you follow the specific rule
- The specific rule adds NO additional constraints or guidance
- The specific rule is merely a special case or example of the general rule

Does the general rule completely subsume the specific rule? Answer with JSON:

`{}
"subsumed": true or false,
"reasoning": "Brief explanation"
{}}`

Conflict Resolver Prompt. Resolves contradictory guidelines.

G. Per-Agent Specialization and Step-Level Updates

This appendix provides empirical validation of two key design choices discussed in Section 2.4: per-agent optimization and step-level updates.

G.1. Per-Agent Specialization

Table 10 shows how SCOPE’s per-agent optimization produces meaningfully different guidelines across agent types. The browser agent accumulates the most guidelines (39), concentrated in *efficiency* guidelines for web-specific challenges (pagination, popups, batching). The planning agent’s guidelines focus on *tool usage* (10 guidelines), reflecting its orchestration role. This specialization matters: a one-size-fits-all approach would force conflicting guidelines across agents (*e.g.*, browser inheriting “limit steps” guidelines that conflict with its need for thoroughness). SCOPE’s step-level operation enables fine-grained attribution of guidelines to the specific agent that benefits.

Table 10. Distribution of guidelines across agents. Each agent develops distinct specialization patterns reflecting its role.

| Agent | Method. | Tool | Valid. | Effic. | Error | Total |
|------------|---------|------|--------|--------|-------|-------|
| Planning | 4 | 10 | 5 | 5 | 3 | 27 |
| Browser | 10 | 4 | 10 | 10 | 5 | 39 |
| Analyzer | 10 | 10 | 7 | 2 | 2 | 31 |
| Web Search | 7 | 4 | 1 | 1 | 1 | 14 |

G.2. Why Step-Level Updates Matter

Unlike task-level methods (DC, ACE) that update only after task completion, SCOPE updates prompts at each step. This matters because errors occur early: among 49 error events analyzed, over 40% occur at step 1. With task-level methods, an error at step 1 of a 15-step task means 14 steps execute with the uncorrected prompt. SCOPE synthesizes corrective guidelines within seconds (representative trace: error at 19:13:33, guideline accepted at 19:13:42, step 2 proceeds at 19:13:46). This step-level granularity enables mid-task recovery that task-level methods cannot achieve.

H. Guideline Placement: Detailed Analysis

This appendix provides a deeper analysis of guideline placement strategies, extending the results in Table 3. We examine behavioral patterns that explain why system prompt placement outperforms alternatives.

H.1. Quantitative Behavioral Metrics

Table 11 presents additional metrics beyond accuracy that reveal distinct behavioral patterns across placement strategies.

Table 11. Detailed behavioral metrics for guideline placement strategies on GAIA. System prompt achieves best accuracy despite higher error tolerance; Hybrid achieves best efficiency (accuracy/steps) with fewest total steps.

| Placement | Acc. (%) | Errors | Timeouts | Steps | Eff.* |
|--------------------------|--------------|--------------|------------|-------|-------------|
| Baseline (No Guidelines) | 32.73 | 1,714 | 255 | 9,824 | 0.33 |
| (1) System Prompt | 46.06 | 1,461 | 227 | 7,430 | 0.62 |
| (2) User Prompt | 41.21 | 1,000 | 130 | 7,408 | 0.56 |
| (3) Split | 35.76 | 1,204 | 141 | 7,319 | 0.49 |
| (4) Hybrid | 43.64 | 1,109 | 139 | 6,620 | 0.66 |

*Efficiency = Accuracy / Total Steps × 100

Key Observations.

- Baseline has highest errors (1,714) and timeouts (255):** Without guidance, agents get stuck more often and make more mistakes.
- User prompt has lowest errors (1,000) and timeouts (130):** This apparent “safety” does not translate to best accuracy, suggesting premature termination.
- System prompt tolerates more errors (1,461) but achieves best accuracy:** Controlled risk-taking leads to better outcomes.
- All guideline placements improve efficiency:** 24–33% reduction in total steps vs. baseline.

H.2. The “Background Guidance” vs. “Over-Compliance” Trade-off

Our analysis reveals a fundamental trade-off between how guidelines are internalized:

System Prompt: Constitutional Guidance. When guidelines are placed in the system prompt, they act as *implicit background context*:

- Guidelines are “loaded” silently and shape behavior without explicit acknowledgment
- Agent maintains flexibility in interpretation, enabling adaptive exploration
- Higher error tolerance allows discovering successful paths that conservative agents miss
- Fewer explicit guideline checks per step reduces cognitive overhead

User Prompt: Direct Instructions. When guidelines are placed in the user prompt, they act as *explicit orders*:

- Agent frequently acknowledges and re-states guidelines, creating verbose responses
- Guidelines accumulate over time, leading to longer contexts (up to 7,600+ tokens observed)
- Pressure for explicit compliance leads to conservative, risk-averse behavior
- Lower timeout rate (130 vs. 227) suggests premature task termination rather than thorough exploration

H.3. Guideline Dynamics Analysis

Table 12 shows how guideline management differs across placement strategies.

Table 12. Guideline dynamics across placement strategies. User prompt shows high guideline churn while system prompt maintains stable guidance.

| Placement | Guideline Updates | Strategic Guidelines Loaded |
|---------------|-------------------|-----------------------------|
| System Prompt | 1,379 | 2,962 |
| User Prompt | 1,648 | 1,990 |
| Split | 1,416 | 2,978 |
| Hybrid | 1,282 | 2,804 |

Interpretation. User prompt placement generates the most guideline updates (1,648) but loads fewer strategic guidelines (1,990). This indicates *guideline churn*—constant adjustment rather than stable guidance. In contrast, system prompt loads more persistent strategic guidelines (2,962) with fewer updates (1,379), suggesting more stable, internalized guidance.

H.4. Why Split Placement Fails

The split strategy (strategic guidelines in system prompt, tactical guidelines in user prompt) performs worst at 35.76%, even below user-prompt-only placement. We hypothesize this creates *cognitive dissonance*:

- Agent receives guidance from two sources with potentially conflicting priorities
- Ambiguity about whether to follow “background” strategic guidelines or “immediate” tactical guidelines
- Lowest error count (1,204) suggests over-caution from mixed signals, leading to missed opportunities

H.5. Theoretical Framework

Based on our analysis, we propose a framework for understanding guideline placement:

Table 13. Theoretical framework for guideline placement effects.

| Aspect | System Prompt (“Constitution”) | User Prompt (“Direct Orders”) |
|----------------|--------------------------------|-------------------------------|
| Integration | Implicit, internalized | Explicit, acknowledged |
| Flexibility | High (adaptive interpretation) | Low (literal compliance) |
| Risk Tolerance | Higher (exploratory) | Lower (conservative) |
| Cognitive Load | Low (background) | High (accumulating) |
| Stability | Persistent across tasks | Frequent updates |

Recommendation. Treating guidelines as *constitutional principles* (system prompt) rather than *direct orders* (user prompt) produces superior performance by: (1) reducing cognitive load per step, (2) enabling natural guideline internalization, (3) maintaining consistent guidance, and (4) permitting adaptive behavior within guidelines.

I. Qualitative Comparison of Perspective Streams

This appendix provides a detailed qualitative comparison of the guidelines evolved by the Efficiency and Thoroughness optimization streams.

I.1. Strategy Patterns by Scenario

Table 14 summarizes the distinct strategies that emerge from each perspective stream across common failure scenarios.

Table 14. Qualitative comparison of guidelines evolved by the two perspective streams. The Efficiency stream focuses on latency reduction and fast failure recovery, while the Thoroughness stream prioritizes exhaustiveness and resilience.

| Scenario | Efficiency Stream Strategy | Thoroughness Stream Strategy |
|----------------|----------------------------|--------------------------------------|
| Blocked Access | Switch Agent (Escalate) | Find Alternate Source (Archive.org) |
| Tool Failure | Switch Tool (Escalate) | Fix Input / Manual Fallback |
| Planning | Limit Steps (≤ 5) | Exhaustive Case Enumeration |
| Goal | Minimize Latency | Maximize Success Rate & Auditability |

Blocked Access (HTTP 403/404). When the browser agent

encounters access denials, the two streams diverge significantly:

- **Efficiency:** Synthesizes guidelines to “fail-over fast” by immediately escalating to the web search agent, avoiding wasted retries on blocked resources.
- **Thoroughness:** Synthesizes guidelines to “find alternate sources” such as Archive.org, Wayback Machine, or transcript extraction tools, prioritizing information recovery over speed.

Tool Failures. For general tool execution errors:

- **Efficiency:** Favors immediate tool switching—if one tool fails, escalate to a different tool or agent rather than debugging.
- **Thoroughness:** Attempts input repair or manual fallback strategies, such as reformatting queries or trying alternative parameters.

Planning Strategy. The streams also differ in their approach to task planning:

- **Efficiency:** Generates guidelines limiting plan complexity (e.g., “Keep plans under 5 steps”), favoring concise execution paths.
- **Thoroughness:** Generates guidelines encouraging exhaustive case enumeration and comprehensive validation before proceeding.

I.2. Complementary Strengths

This duality is why the Global Ensemble outperforms either individual stream. Efficiency excels at time-constrained tasks and long-horizon scenarios (Level 3) where context bloat is dangerous, while Thoroughness excels at tasks requiring deep retrieval and careful validation (Level 2). By maintaining both streams and selecting the best outcome, SCOPE adapts to heterogeneous task requirements without manual configuration.

J. Model Choice for Meta-Agents: Detailed Analysis

This appendix provides a detailed analysis of how different model choices for SCOPE’s meta-agents (Generator, Classifier, Optimizer) affect the characteristics of evolved guidelines. We compare three configurations on the GAIA benchmark: (1) all meta-agents use GPT-4.1, (2) all use Gemini-2.5-Pro, and (3) each meta-agent uses the same model as the base agent it optimizes.

J.1. Guideline Generation Statistics

Table 15 provides comprehensive statistics on the guidelines generated by each configuration.

Table 15. Detailed statistics of guidelines generated by different meta-agent model configurations.

| Metric | GPT-4.1 | Gemini-2.5-Pro | Same Model |
|-------------------------------|---------|----------------|------------|
| Total Guidelines | 111 | 163 | 108 |
| Avg. Guideline Length (chars) | 380.1 | 426.3 | 396.8 |
| Max Guideline Length (chars) | 2,895 | 2,282 | 2,340 |
| Min Guideline Length (chars) | 105 | 82 | 94 |
| Avg. Confidence | 0.939 | 0.940 | 0.942 |
| Accuracy (%) | 46.06 | 46.67 | 45.45 |

A key finding is that **Gemini generates 46% more guidelines than GPT-4.1**, yet achieves only marginally better performance (+0.61%). This strongly suggests that SCOPE’s effectiveness stems from guideline quality rather than quantity—the Best-of-N selection and memory optimization mechanisms effectively filter for useful guidelines regardless of how many candidates are initially generated.

J.2. Category Distribution Analysis

Table 16 shows the distribution of guidelines across semantic categories for each configuration.

Table 16. Guideline distribution by category. Each model exhibits distinct “personality” in guideline generation patterns.

| Category | GPT-4.1 | Gemini-2.5-Pro | Same Model |
|----------------------|------------|----------------|------------|
| Analysis Methodology | 31 | 37 | 33 |
| Data Validation | 23 | 24 | 25 |
| Tool Usage | 28 | 36 | 26 |
| Efficiency | 18 | 32 | 20 |
| Error Handling | 11 | 26 | 4 |
| General | 0 | 6 | 0 |
| Safety | 0 | 2 | 0 |
| Total | 111 | 163 | 108 |

Gemini generates the most guidelines across nearly all categories, with notably more *error_handling* guidelines (26 vs. 4–11 for others) and *efficiency* guidelines (32 vs. 18–20). It is also the only model to generate *safety*-related guidelines (2). This suggests Gemini’s synthesis tends toward comprehensive, verbose guidelines covering edge cases and operational robustness.

J.3. Confidence Score Distribution

All three configurations achieve remarkably similar average confidence scores:

- GPT-4.1: 0.939
- Gemini-2.5-Pro: 0.940
- Same Model: 0.942

This consistency (~0.94 across all settings) suggests that all

models are capable of generating guidelines they assess as reliable, regardless of their different generation patterns.

J.4. Implications for Deployment

The near-identical performance across configurations (within 1.2%) has important practical implications:

1. **Model-Agnostic Framework:** SCOPE’s effectiveness is not tied to a specific meta-agent model, simplifying deployment across different infrastructure constraints.
2. **Cost-Performance Trade-off:** Practitioners can choose meta-agent models based on cost or latency rather than accuracy. For example, using a cheaper model for the Generator may reduce costs without impacting performance.
3. **Quality Over Quantity:** The 46% guideline generation difference between Gemini and GPT-4.1 without proportional accuracy gains validates SCOPE’s guideline management mechanisms—Best-of-N selection and memory optimization effectively filter for high-quality guidelines.
4. **Consistent Guideline Quality:** The similar average confidence scores (~0.94) across all configurations suggest that both models are capable of generating guidelines they assess as reliable, regardless of their different generation patterns.