The MCP Quality Paradox: Enhancing Tool Effectiveness in Long-Horizon Reasoning through Phase-Aware State Abstraction

Tong Zhang[1], **Yiquan Wu**[2], **Wenlin Zhong**[2], **Rujing Yao**[1], **Yang Wu**[3], **Yufei Shi**[4], **Ang Li**[2], **Xiaozhong Liu**[3]

[1]Nankai University, Tianjin, China
[2]Zhejiang University, Hangzhou, China
[3]Worcester Polytechnic Institute, MA, USA
[4]The Hong Kong Polytechnic University, Hong Kong, China

Abstract

Large Language Model (LLM) agents are increasingly tasked with long-horizon deep research, utilizing an expanding ecosystem of heterogeneous tools via the Model Context Protocol (MCP). While integrating diverse information sources, ranging from structured databases (SQL) to unstructured retrieval systems (Web search, ArXiv), theoretically enhances capability, it introduces significant integration challenges. Specifically, the unweighted fusion of high-fidelity internal data with noisy external retrieval can compromise reasoning consistency, while the linear accumulation of intermediate tool outputs leads to context suffocation, where critical signals are diluted by redundant interaction history.

To address these challenges, we introduce Q-STEAM (**Q**uality-aware **St**a**t**e **A**bstraction for **M**ulti-Hop Reasoning), a framework that reformulates long-horizon reasoning as a Phase-Aware Decision Process. Unlike mono-contextual paradigms, Q-STEAM: **Firstly**, dynamically evaluates tool reliability within each reasoning phase: Acquisition (retrieval quality), Analysis (extraction accuracy), and Propagation (aggregation robustness), rather than applying uniform tool credibility scores; **Secondly**, implements Phase-Aware State Abstraction, which synthesizes reasoning history into evolved reports at phase boundaries, selectively discarding redundancy while preserving reasoning continuity. We validate Q-STEAM on HotpotQA (controlled multi-hop reasoning) and a novel Legal Case Synthesis dataset (high-stakes, real-world uncertainty). Experimental results demonstrate that Q-STEAM achieves improvement over baselines.

MCP tool quality, long-horizon reasoning, information heterogeneity

Introduction

The advent of the Model Context Protocol (MCP) (Chen et al. 2025a) has catalyzed a paradigm shift in autonomous agents, transitioning them from passive query responders to active researchers capable of navigating complex information ecosystems. By standardizing interfaces for tools ranging from high-fidelity databases to open-web search engines, MCP enables agents to perform **Deep Research** across extended horizons (Chen et al. 2025b). However, as agents are deployed in high-stakes domains such as legal analysis and scientific discovery, the sheer abundance of tools and information has exposed fundamental limitations in current reasoning architectures. We identify two critical bottlenecks that impede the reliability and scalability of long-horizon agents:

First, the Challenge of Information Quality Heterogeneity. Existing multi-hop reasoning frameworks, such as ReAct (Banks and Porcello 2017) and standard reinforcement learning pipelines (e.g., GRPO (Tong et al. 2025)), often treat all tool outputs as equipotent evidence. This "blind trust" is problematic in a heterogeneous MCP environment. A fact verified against a structured internal database carries significantly higher evidentiary weight than a claim retrieved from a noisy web snippet. Without a mechanism to dynamically weight evidence based on source reliability, distinguishing between Information Acquisition, Analysis, and Propagation—agents are prone to "hallucination accumulation," where low-fidelity noise corrupts high-fidelity facts during multi-step reasoning.

Second, the Challenge of Context Suffocation via Linear Accumulation. To maintain reasoning consistency, traditional agents typically append every search query, API call, and intermediate observation to a monolithic context window. Our analysis of deep-research trajectories reveals that Most of tokens are functionally redundant, often consisting of repeated verifications or irrelevant retrieval artifacts. This linear accumulation creates a "noise tunnel", where critical signals are drowned out by redundant history, leading to degraded performance and inflated computational costs. While recent works like IterResearch (Chen et al. 2025b) attempt to mitigate this via state reconstruction, they often apply uniform compression, failing to distinguish between critical reasoning pivots and dispensable exploration trails. To resolve these challenges, we propose Q-STEAM (**Q**uality-aware **St**a**t**e **A**bstraction for **M**ulti-Hop Reasoning). Our key insight is that effective long-horizon reasoning requires Phase-Aware State Abstraction, a strategy that does not simply "forget" history, but actively refines it based on the current reasoning phase. Q-STEAM introduces three primary contributions:

- Dynamic MCP Tool Quality Profiling: We construct a multi-dimensional quality assessment matrix that evaluates tools based on their role (Acquisition/Analysis/Propagation) and reliability. This allows the agent to prioritize high-trust sources (e.g., ArXiv) over lower-trust ones

Proceedings of Bridge between Artificial Intelligence and Law 2026 (Short papers), pages 215–218
Januray 21, 2026, AAAI 2026 Bridge Program, Singapore

during synthesis.

- Phase-Aware State Abstraction: Learns phase transition dynamics by training the agent to synthesize reasoning history into state reports S_k at phase boundaries. Environment feedback guides the agent to selectively compress within-phase trajectories while preserving cross-phase critical information.
- Comprehensive Validation: We evaluate Q-STEAM on HotpotQA and a Legal Case Synthesis dataset. Furthermore, user studies demonstrate that our transparent, quality-graded outputs significantly enhance user trust compared to black-box baselines.

Related Work

Tool-Augmented Autonomous Agents

The integration of external tools is fundamental to advancing LLM capabilities beyond parametric knowledge. Early frameworks like ReAct (Banks and Porcello 2017) and Toolformer (Schick et al. 2023). demonstrated that interleaving reasoning traces with action execution improves performance on knowledge-intensive tasks. Recent systems such as ToolLLM (Qin et al. 2023) have scaled this paradigm to handle thousands of real-world APIs via instruction tuning. However, these approaches generally operate under a "trusted tool" assumption, lacking mechanisms to critically evaluate the quality of retrieved information. Recent studies on hallucination propagation have begun to address this by introducing self-reflection and critique steps. Our work extends these efforts by introducing a Dynamic Quality Profiling module specifically designed for the heterogeneous MCP ecosystem, enabling agents to weigh evidence from different tools differently during long-horizon reasoning.

Long-Horizon Reasoning and Context Management

As interaction turns increase, maintaining a coherent reasoning state becomes a primary challenge due to the "lost-in-the-middle" phenomenon. Traditional context management relies on sliding windows or external memory banks, as seen in MemGPT (Packer et al. 2023), which manages context like an operating system. Similarly, Context-Folding (Sun et al.) introduces a mechanism to branch and fold sub-trajectories to manage context length. While these methods effectively compress history, they often treat compression uniformly across the trajectory. Q-STEAM differentiates itself through Phase-Aware Abstraction. We argue that compression strategies must be adaptive, preserving high-fidelity data during Analysis phases while aggressively compressing noisy exploration trails during Acquisition phases. This ensures that the agent's working memory is optimized for reasoning continuity rather than just token reduction.

Methodology

Our proposed framework, **Q-STEAM**, is designed to resolve the fundamental tension in long-horizon reasoning: the need for extensive information retrieval versus the cognitive bottleneck of limited context windows. We first systematize the landscape of Model Context Protocol (MCP) tools and evaluation benchmarks (in Table 1). We then formulate the long-horizon reasoning process as a Phase-Aware Decision Process, where the agent learns to synthesize and propagate information across distinct phases. Finally, we introduce our optimization algorithm, which extends Group Relative Policy Optimization (GRPO) with quality-aware state abstraction objectives.

Taxonomy of MCP Tools and Benchmark Landscape

Effective tool orchestration requires distinguishing tools not merely by function but by their role in the information lifecycle. We categorize MCP tools into three distinct classes based on their contribution to reasoning fidelity:

- **Information Acquisition Tools** ($\mathcal{T}_{acq}$): Tools that retrieve raw external data. These are high-recall but potentially noisy sources. Examples include Google Search (for real-time events) and arXiv Search (for academic literature). Their primary utility lies in expanding the agent's knowledge boundary.
- **Information Analysis Tools** ($\mathcal{T}_{ana}$): Tools that process, verify, or structure existing data. These are high-precision sources often used for internal reasoning (e.g., BioNext for biological pathway analysis).
- **Information Propagation Tools** ($\mathcal{T}_{prop}$): Tools that facilitate user interaction and workflow integration, enhancing the utility of derived insights. Examples include Email clients and Calendar scheduling APIs, which do not generate new knowledge but operationalize it.

To rigorously evaluate these capabilities, we survey existing benchmarks and map them to our taxonomy. As shown in Table 1, benchmarks vary significantly in their focus. For instance, *HotpotQA* primarily tests the agent's ability to filter and analyze multi-hop information, whereas our proposed *Legal Case Synthesis* task demands a seamless integration of acquisition (finding statutes) and analysis (applying precedents).

Phase-Aware State Abstraction via Progressive Refinement

Standard ReAct agents maintain a monotonically increasing context history $H_t = [o_1, a_1, \ldots, o_t, a_t]$, leading to quadratic complexity and noise accumulation. Instead of treating the reasoning process as a single continuous stream, we reformulate it as a sequence of **Phases** $\Phi = \{\phi_1, \phi_2, \ldots, \phi_K\}$, where each phase represents a distinct sub-goal (e.g., "Information Gathering" or "Hypothesis Verification").

The transition between phases is governed by a **Synthesized State Report** S_k, which bridges phases ϕ_{k-1} and ϕ_k. Instead of conditioning the policy π_θ on full history H_t, we introduce a learnable **State Abstraction Function** that maps the previous phase's trajectory to a refined state representation:

$$S_k = \Psi_\theta \left(\text{Filter}(H_{\phi_{k-1}}, \tau_k) \right) \tag{1}$$

Table 1: Analysis of Benchmarks for MCP Tool Quality Assessment. We categorize datasets based on their primary focus: Acquisition (Acq.), Analysis (Ana.), and Propagation (Prop.).

Benchmark	Domain	Primary MCP Focus	Key Capabilities Tested	Example Tools
HotpotQA-II	General	Ana. > Acq.	Multi-hop reasoning, Fact verification	WikiSearch, Retrieval
Legal Case Syn.	Legal	Acq. ≈ Ana.	Evidence synthesis, Statute application	CaseLaw Search
The tool decathlon	Scientific	Acq. > Ana.	Broad information gathering	Google Search, ArXiv
WikiTableQuestions	General	Acq.	Entity extraction, List generation	Google Search, SQL
ToolMind	General	Ana.	Complex relationship inference	Bio-KG

where the filtering operation is:

$$\text{Filter}(H_{\phi_{k-1}}, \tau_k) = \{(o_i, a_i) \in H_{\phi_{k-1}} \mid q_i > \tau_k\} \quad (2)$$

Here $q_i = \text{Quality}(o_i)$ combines tool reliability and verification signals. The function Ψ_θ is LLM-based and compresses the filtered trajectory into a concise report, actively filtering redundant observations to ensure only high-fidelity signals propagate to the next phase.

The policy for action selection in phase ϕ_k is conditioned on the synthesized state and local context:

$$a_t \sim \pi_\theta(a_t \mid S_k, h_t^{\phi_k}) \quad (3)$$

where $h_t^{\phi_k}$ is the short-term working memory within the current phase. This formulation enables the agent to learn high-level state transitions while maintaining local tactical awareness.

Optimization via Quality-Aware GRPO

To enable effective state abstraction and tool selection, we extend GRPO with a **compound reward** that evaluates phase-level transitions:

$$R_{\text{total}} = R_{\text{outcome}} + \lambda_1 R_{\text{efficiency}} + \lambda_2 R_{\text{quality}} \quad (4)$$

Reward Components **Outcome Reward:** Standard task completion reward (e.g., EM on HotpotQA):

$$R_{\text{outcome}} = \begin{cases} 1 & \text{if task solved} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Efficiency Reward: Penalizes semantic redundancy in acquisition tool calls:

$$R_{\text{efficiency}} = -\sum_{t=1}^{T} \max_{j<t} \text{CosSim}(E(o_t), E(o_j)) \cdot \mathbb{I}[a_t \in \mathcal{T}_{\text{acq}}] \quad (6)$$

This encourages the agent to rely on synthesized state S_k rather than re-acquiring similar information.

Tool Quality Reward: Prioritizes high-precision analysis tools and penalizes hallucinations:

$$R_{\text{quality}} = \sum_{t=1}^{T} (\alpha \cdot \mathbb{I}[a_t \in \mathcal{T}_{\text{ana}} \wedge \text{Verified}(o_t)] - \beta \cdot \mathbb{I}[a_t \in \mathcal{T}_{\text{acq}} \wedge \text{Hallucinated}(o_t)]) \quad (7)$$

where Verified/Hallucinated are determined by comparing against ground-truth references (e.g., supporting facts in HotpotQA).

Policy Optimization Objective The final objective incorporates phase-aware state representation:

$$\mathcal{L}_{Q\text{-}STREAM}(\theta) = \mathbb{E}_{q\sim D}\, \mathbb{E}_{\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \Big[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{T_i}\sum_{t=1}^{T_i} \min\Big\{ {}^{\pi_\theta(a_{i,t}|)}\big/_{\pi_{\theta_{\text{old}}}(a_{i,t}|)}\, A_{i,t} \;\Big\} - \beta_{\text{KL}} D_{\text{KL}}\Big] \quad (8)$$

where:

- $S_{k_{i,t}}$ is the active synthesized state at step t of trajectory i
- $h_{i,t}$ is the short-term history within the current phase
- $A_{i,t}$ is the advantage estimate incorporating R_{total}

By optimizing this objective, Q-STREAM jointly learns: (1) phase-appropriate tool selection via π_θ, (2) effective state synthesis via Ψ_θ, and (3) phase quality thresholds $\{\tau_k\}$, thereby addressing context suffocation while maintaining information fidelity.

Experiments

In this section, we empirically validate Q-STEAM on long-horizon reasoning tasks, with a specific focus on the efficacy ofInformation Analysis Tools within a multi-phase inference framework.

In this section, we empirically validate Q-STREAM on long-horizon reasoning tasks. Our primary focus is evaluating the efficacy of Information Analysis Tools within a multi-phase inference framework. We aim to answer three research questions:

- **RQ1 (Effectiveness):** Does Q-STREAM outperform standard mono-contextual baselines (e.g., ReAct, GRPO) in multi-hop reasoning tasks?
- **RQ2 (Robustness):** Does Dynamic Tool Quality Profiling prevent hallucination accumulation when integrating heterogeneous information sources?

Experimental Setup

Datasets and Evaluation Metrics We primarily utilize **HotpotQA-II**, an enhanced version of the HotpotQA dataset specifically curated to evaluate MCP tool integration.

- **Dataset Construction:** The dataset comprises 15,000 training samples and 1,000 test samples. Unlike the

original dataset, HotpotQA-II requires agents to utilize specific MCP tools, to resolve multi-hop queries. This setup simulates a realistic "Information Analysis" scenario where answers must be synthesized from disjoint sources.

- **Metrics:** We report Exact Match (EM) and F1 Score to evaluate answer correctness, alongside BLEU and METEOR to assess the semantic quality of the generated reasoning traces. To measure efficiency, we implicitly evaluate the ability to maintain performance under constrained context windows.

Baselines and Model Architectures We evaluate our framework across a diverse set of Large Language Models (LLMs) to demonstrate model-agnostic effectiveness.

- **Base Models:** We employ open-source models **Qwen2.5-3B-Instruct** and **Llama-3.2-3B-Instruct** for agile experimentation.
- **Comparison Methods:**
  1. **ReAct (Prompting):** The standard interleaved reasoning-action paradigm with a fixed context window, tested on both Llama-3.2-3B and Qwen2.5-3B.
  2. **SFT (Supervised Fine-Tuning):** We fine-tune Qwen2.5-3B on expert trajectories, teaching the model basic tool usage patterns without phase-aware abstraction.
  3. **GRPO:** We apply Group Relative Policy Optimization on the Qwen2.5-3B model using binary outcome rewards, representing the state-of-the-art RL baseline for tool use.
  4. **Q-STREAM (Ours):** Our full method applied to Qwen2.5-3B, trained with phase-aware state abstraction and tool quality rewards.

Implementation Details All training experiments are conducted on a cluster of **3 $\times$ NVIDIA A800 (80GB) GPUs**. For GRPO and Q-STREAM, we set the group size $G = 8$, learning rate $5e^{-6}$, and KL coefficient $\beta = 0.04$. The maximum context length is set to 8,192 tokens.

Main Results

Table 2 presents the performance comparison on HotpotQA-II.

Table 2: Performance comparison on HotpotQA-II. Q-STREAM significantly improves upon the Qwen2.5-3B and Llama-3.2-3B baselines (SFT), demonstrating the value of phase-aware optimization.

Method	EM	F1	BLEU	METEOR
Llama-3.2-3B ReAct	0.2650	0.3609	0.1072	0.3417
Llama-3.2-3B SFT	0.3454	0.4583	0.1667	0.3851
Llama-3.2-3B GRPO	0.2600	0.3584	0.1182	0.3009
Llama-3.2-3B SFT+GRPO	0.2800	0.3597	0.1375	0.3663
Qwen2.5-3B SFT	0.2200	0.3343	0.0839	0.2735
Q-STREAM (Qwen2.5-3B)	**0.2400**	**0.3635**	**0.0948**	**0.2604**

Superior Reasoning Accuracy (RQ1): As shown in Table 2, **Q-STREAM (Ours)** achieves an EM score of **0.2400** and an F1 score of **0.3635**.

- **Baseline Landscape.** Among the Llama-3.2-3B variants, supervised fine-tuning (SFT) secures the highest EM/F1, while ReAct and GRPO lag consistently; the spread hints that recipe matters more than scale at the 3-B level.
- **Phase-Aware Lift.** On Qwen2.5-3B, Q-STREAM quietly overtakes the strongest SFT baseline by 0.029 F1 and 0.02 EM—an upward nudge that keeps the model in the same weight class yet places it atop the internal leaderboard.

High-Fidelity Tool Usage (RQ2): Q-STREAM's *Tool Quality Reward* constrains the agent to prioritize high-precision MCP tools, leading to more reliable and verifiable reasoning.

Legal Case Study: Validation in High-Stakes Scenarios

To validate practical utility beyond academic benchmarks, we conducted a user study with 12 legal professionals analyzing complex cases requiring synthesis of statutes (SQL), precedents (DocRetrieval), and updates (Web Search).

References

Banks, A.; and Porcello, E. 2017. *Learning React: functional web development with React and Redux*. " O'Reilly Media, Inc.".

Chen, C.; Hao, X.; Liu, W.; Huang, X.; Zeng, X.; Yu, S.; Li, D.; Wang, S.; Gan, W.; Huang, Y.; et al. 2025a. ACEBench: Who Wins the Match Point in Tool Usage? *arXiv preprint arXiv:2501.12851*.

Chen, G.; Qiao, Z.; Chen, X.; Yu, D.; Xu, H.; Zhao, W. X.; Song, R.; Yin, W.; Yin, H.; Zhang, L.; et al. 2025b. IterResearch: Rethinking Long-Horizon Agents via Markovian State Reconstruction. *arXiv preprint arXiv:2511.07327*.

Packer, C.; Fang, V.; Patil, S.; Lin, K.; Wooders, S.; and Gonzalez, J. 2023. MemGPT: Towards LLMs as Operating Systems.

Qin, Y.; Liang, S.; Ye, Y.; Zhu, K.; Yan, L.; Lu, Y.; Lin, Y.; Cong, X.; Tang, X.; Qian, B.; et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551.

Tong, C.; Guo, Z.; Zhang, R.; Shan, W.; Wei, X.; Xing, Z.; Li, H.; and Heng, P.-A. 2025. Delving into RL for Image Generation with CoT: A Study on DPO vs. GRPO. *arXiv preprint arXiv:2505.17017*.