

Don’t Break the Cache: An Evaluation of Prompt Caching for Long-Horizon Agentic Tasks

Elias Lumer¹ Faheem Nizar¹ Akshaya Jangiti¹ Kevin Frank¹ Anmol Gulati¹
 Mandar Phadate¹ Vamse Kumar Subbiah¹

PricewaterhouseCoopers, U.S.

Abstract

Recent advancements in Large Language Model (LLM) agents have enabled complex multi-turn agentic tasks requiring extensive tool calling, where conversations can span dozens of API calls with increasingly large context windows. However, although major LLM providers offer prompt caching to reduce cost and latency, its benefits for agentic workloads remain underexplored in the research literature. To our knowledge, no prior work quantifies these cost savings or compares caching strategies for multi-turn agentic tasks. We present a comprehensive evaluation of prompt caching across three major LLM providers (OpenAI, Anthropic, and Google) and compare three caching strategies, including full context caching, system prompt only caching, and caching that excludes dynamic tool results. We evaluate on DeepResearchBench, a multi-turn agentic benchmark where agents autonomously execute real-world web search tool calls to answer complex research questions, measuring both API cost and time to first token (TTFT) across over 500 agent sessions with 10,000-token system prompts. Our results demonstrate that prompt caching reduces API costs by 45–80% and improves time to first token by 13–31% across providers. We find that strategic prompt cache block control, such as placing dynamic content at the end of the system prompt, avoiding dynamic traditional function calling, and excluding dynamic tool results, provides more consistent benefits than naive full-context caching, which can paradoxically increase latency. Our analysis reveals nuanced variations in caching behavior across providers, and we provide practical guidance for implementing prompt caching in production agentic systems.

¹PricewaterhouseCoopers, U.S.. Correspondence to: Elias Lumer <elias.lumer@pwc.com>.

Preprint. January 12, 2026.

1. Introduction

Recent advancements in Large Language Model (LLM) agents have enabled complex, long-horizon agentic tasks that require extensive tool calling across multi-turn conversations (Ji, 2025). Through function calling, LLM agents can invoke external APIs, execute web searches, interact with databases, and perform domain-specific actions on behalf of users. As these agentic workloads grow in complexity, conversations can span dozens of API calls with context windows accumulating tens of thousands of tokens, leading to significant costs and latency overhead. To address this, major LLM providers including OpenAI, Anthropic, and Google now offer prompt caching, a feature that reuses previously computed key-value (KV) tensors from attention layers to avoid redundant computation on repeated prompt prefixes (OpenAI, 2026; Anthropic, 2026; Google Cloud, 2026b).

While providers offer reduced pricing for cached input tokens, the benefits of prompt caching in real-world agentic workloads remain under-explored in the research literature. Existing work on KV cache optimization focuses primarily on inference-level memory management and compression (Kwon et al., 2023; Ge et al., 2023; Shi et al., 2024), rather than evaluating the enterprise-grade prompt caching features offered through provider APIs. Concurrent work has audited prompt caching across providers to detect timing side-channel vulnerabilities (Gu et al., 2025), but to our knowledge, no prior work has quantified the cost benefits of prompt caching or compared caching strategies for agentic workloads. This gap is particularly significant given the recent proliferation of long-running agents for deep research, coding assistance, and autonomous task completion, where prompt caching could substantially reduce operational costs and improve user experience through faster response times.

In this paper, we present the first comprehensive evaluation of prompt caching strategies for long-horizon agentic tasks across three major LLM providers (OpenAI, Anthropic, and Google) using four flagship models (Figure 1). We compare three caching strategies, including full context caching,

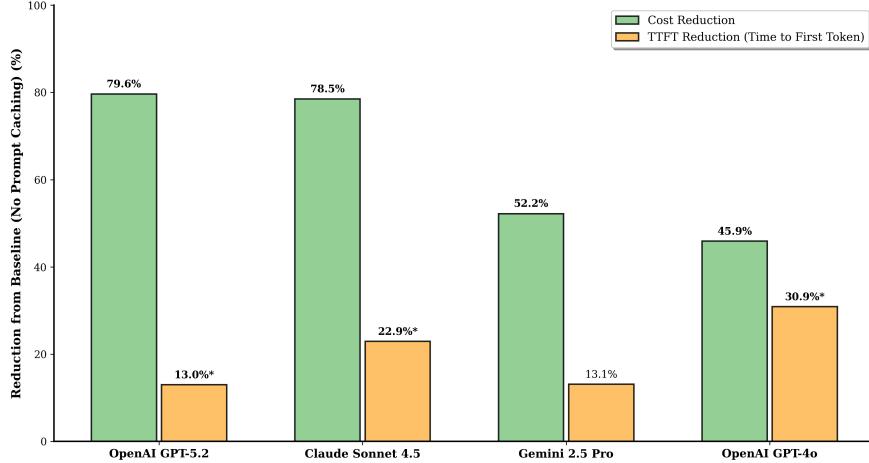


Figure 1. Prompt caching benefits (best cache mode per model). Percentage reduction in API cost and time to first token (TTFT) relative to a no-cache baseline. For each model, results correspond to the best-performing cache strategy. Asterisks denote statistically significant TTFT improvements ($p < 0.05$).

system prompt only caching, and caching that excludes dynamic tool results. We evaluate on DeepResearchBench (Du et al., 2025), a multi-turn agentic benchmark where agents autonomously execute web search tool calls to answer complex research questions. Our evaluation spans over 500 agent sessions with 10,000-token system prompts, measuring both API cost and time to first token (TTFT) across all conditions.

Our evaluation reveals three key findings:

Prompt caching delivers substantial and consistent cost savings across all providers: All four models tested show statistically significant cost reductions when prompt caching is enabled. Cost savings range from 45% to 80% across providers. These savings are consistent across all three caching strategies, demonstrating that prompt caching provides reliable cost benefits regardless of the specific caching approach employed.

Latency improvements vary significantly across providers and require careful strategy selection: Time to first token improvements range from 13% to 31% across providers, though latency variance differs substantially between providers. Notably, the cache strategy that maximizes cost savings does not always maximize latency improvement, highlighting the importance of strategy selection based on optimization goals.

Strategic cache boundary control outperforms naive full-context caching: Providers abstract much of the caching mechanism, automatically triggering cache creation when token thresholds are exceeded. However, naively enabling full-context caching can paradoxically increase latency, as dynamic tool calls and results may trigger cache writes for content that will not be reused across sessions. By

strategically controlling cache boundaries, such as caching only the system prompt or explicitly excluding tool results, practitioners can ensure that only stable, reusable content is cached. Our results show that system prompt only caching provides the most consistent benefits across both cost and latency dimensions.

2. Background

2.1. KV Cache and LLM Inference

Large Language Model inference consists of two distinct phases: the prefill phase, where the model processes the input prompt and generates attention key-value (KV) tensors, and the decode phase, where the model autoregressively generates output tokens (Pope et al., 2022). During prefill, the model computes attention over the entire input sequence, producing KV tensors that capture the contextual representations needed for subsequent generation. These KV tensors are stored in the KV cache and reused during decoding to avoid redundant computation, enabling efficient token-by-token generation (Not Lain, 2025).

As context windows have grown from thousands to millions of tokens, KV cache management has become a critical bottleneck in LLM serving (Shi et al., 2024). The memory footprint of KV caches scales linearly with sequence length and batch size, often consuming more GPU memory than the model weights themselves for long-context workloads. This has motivated extensive research on KV cache optimization, including memory management techniques such as PagedAttention (Kwon et al., 2023), which applies paging-style memory management to reduce fragmentation and waste, achieving 2-4x throughput improvements. Other approaches focus on KV cache compression through selec-

tive retention of important tokens (Ge et al., 2023), storage-compute tradeoffs that balance recomputation against cache loading (Jin et al., 2024), and shared prefix optimization for high-throughput inference (Juravsky et al., 2024; Wu et al., 2024a; Sun et al., 2025; Zhou et al., 2024).

2.2. Prompt Caching in Provider APIs

While KV caching is a general inference optimization technique, prompt caching refers to the productized, provider-managed features that reuse KV tensors across API requests when prompts share common prefixes (OpenAI, 2026; Gim et al., 2024). By caching the KV tensors from the prefill phase, providers can skip redundant computation when subsequent requests begin with the same content, reducing both latency and cost for users.

Major LLM providers have implemented prompt caching with varying approaches. OpenAI offers automatic prompt caching on GPT-4o and newer models, where caching activates automatically for prompts exceeding a minimum token threshold, with cache hits occurring only for exact prefix matches (OpenAI, 2026; 2024b). Anthropic provides developer-controlled caching through explicit cache breakpoints, allowing users to specify which portions of their prompt should be cached, with configurable time-to-live (TTL) options (Anthropic, 2026; 2025c). Google offers both implicit caching, which activates automatically with no guaranteed cost savings, and explicit context caching, where developers create and reference caches with guaranteed discounts (Google Cloud, 2026b; Kilpatrick, 2025).

Implementation details such as minimum token thresholds (typically 1,024-4,096 tokens depending on model, see Table 4), TTL durations (ranging from 5 minutes to 24 hours), and pricing structures vary across providers and are subject to change (PromptHub, 2025; Microsoft Azure AI, 2025). These differences have practical implications for cache hit rates and cost optimization. Recent work has audited prompt caching across 17 providers, demonstrating that cache hits produce measurable TTFT reductions and identifying security vulnerabilities from timing side-channels (Gu et al., 2025). However, their focus on security auditing using synthetic prompts and smaller, older generation models does not compare caching strategies or evaluate cost and latency benefits for long-running agentic tasks on modern flagship models.

2.3. Agentic Workloads and Context Engineering

Recent advances in LLM agents have enabled complex, long-horizon tasks that extend far beyond single-turn question answering. Modern agentic applications including deep research assistants, coding agents such as Claude Code and Cursor, and autonomous task completion systems like Manus routinely execute 30-50 or more tool calls within

a single session (Ji, 2025; Du et al., 2025; Mialon et al., 2023; Zhou et al., 2023; Drouin et al., 2024; Wei et al., 2025). Each tool call adds content to the conversation context, including the tool invocation, execution results, and the model’s subsequent reasoning, causing context windows to grow rapidly throughout the session.

This growth presents unique challenges for prompt caching (Guan et al., 2026; Laban et al., 2025). Recent work has proposed solutions for multi-turn caching scenarios (Jeong & Ahn, 2025; Yan et al., 2025). Unlike static question-answering scenarios where prompts are largely predetermined, agentic workloads feature dynamic, session-specific content that accumulates unpredictably. Tool results often contain user-specific data that will not benefit other sessions, and the interleaving of static system prompts with dynamic tool outputs complicates cache reuse. Context engineering strategies have emerged to manage these challenges, including treating external storage as extended memory and carefully structuring prompts to maximize cache efficiency (Ji, 2025; Lumer et al., 2025a). However, the effectiveness of prompt caching across different caching strategies in agentic workloads has not been comprehensively evaluated. Our work addresses this gap by measuring cost and latency benefits across controlled caching strategies on a multi-turn agentic benchmark.

3. Methodology

3.1. Experimental Setup

We evaluate prompt caching across three major LLM providers: OpenAI, Anthropic, and Google. For each provider, we select a flagship model: GPT-4o and GPT-5.2 from OpenAI, Claude Sonnet 4.5 from Anthropic, and Gemini 2.5 Pro from Google. These models represent the current state-of-the-art for agentic workloads and all support prompt caching through their respective APIs.

We use DeepResearchBench (Du et al., 2025) as our evaluation benchmark, a multi-turn agentic benchmark where agents autonomously execute web search tool calls to answer complex research questions. We selected this benchmark over alternatives such as other deep research benchmarks (FutureSearch: Bosse et al., 2025; Li et al., 2025) due to its focus on tool-intensive agentic workflows and real-world 100 PhD-level research tasks, each meticulously crafted by domain experts across 22 distinct fields. We implement our research agent using Deep Agents (LangChain, 2025), one of various open source libraries for creating long-running agents (Anthropic, 2025b; OpenAI, 2025; Google, 2025; Microsoft, 2023a; 2026; 2023b; CrewAI, 2023; LlamaIndex, 2022; Hugging Face, 2024; OpenAI, 2024a; Agno, 2026). Each agent session begins with a research question and the agent iteratively calls a web search tool to gather

information before synthesizing a comprehensive response. This benchmark reflects realistic agentic workloads where context windows grow dynamically through tool invocations and results.

For each model, we conduct 40 independent agent sessions per cache condition, with each session answering a unique research question from the benchmark. Sessions use a 10,000-token system prompt containing agent instructions for deep research, including guidance on tool usage, question decomposition, and report synthesis. Each session starts with a fresh context, ensuring that cache benefits are measured within individual multi-turn conversations rather than across sessions.

3.2. Latency Improvement

3.3. Cache Mode Implementation

We implement four cache conditions to systematically evaluate prompt caching strategies (see Appendix C, Figures 5–8 for visual illustrations). To control cache boundaries precisely, we use unique identifiers (UUIDs) to break the cache at specific points in the prompt, ensuring that content after the UUID is not cached from previous requests.

No Cache (Baseline): A UUID is prepended to the beginning of the system prompt, breaking the cache immediately and forcing the model to recompute all tokens. This serves as our baseline condition where no caching benefits are realized. In real-world agentic tasks, this symbolizes including dynamic content, such as timestamps and user information, to the system prompt on inference time (Ji, 2025).

Full Context Caching: No UUIDs are added, allowing the provider's caching mechanism to operate automatically. OpenAI and Google enable prompt caching automatically for eligible requests, while Anthropic requires explicit cache breakpoints in the API request. This condition represents naive caching where practitioners enable the feature without additional optimization.

System Prompt Only Caching: A UUID is appended to the end of the system prompt, breaking the cache at this boundary. This ensures that only the static system prompt is cached, while the dynamic conversation history, tool calls, and tool results are recomputed on each request.

Exclude Tool Results Caching: UUIDs are appended both after the system prompt and after each tool result. This strategy ensures that tool results, which are dynamic and session-specific, do not contribute to the cache. We found this dual-UUID approach necessary because provider-level KV cache handling can vary, and explicit boundaries provide more predictable caching behavior.

Table 1. Prompt caching benefits by model using the best-performing cache mode for each. Cost savings and TTFT improvement are relative to the no-cache baseline. Bold indicates highest value per metric.

Model	Cache Mode	Cost ↓	TTFT ↓
OpenAI GPT-5.2	Excl. Tool Results	79.6%	13.0%
Claude Sonnet 4.5	System Prompt	78.5%	22.9%
Gemini 2.5 Pro	System Prompt	41.4%	6.1%
OpenAI GPT-4o	System Prompt	45.9%	30.9%

3.4. Evaluation Protocol

We measure two primary metrics across all conditions: API cost and time to first token (TTFT).

Cost: We calculate cost using token counts reported in API responses, distinguishing between standard input tokens, cached input tokens (cache reads), and cache creation tokens (cache writes). Each token type is multiplied by the corresponding provider pricing (see Appendix A, Table 3) to compute total cost per session. Cost is aggregated across all API calls within a session.

Time to first token (TTFT): We measure TTFT using streaming responses, recording the time from request initiation to receipt of the first response chunk. TTFT captures the latency improvement from skipping prefill computation on cached tokens, making it the most relevant latency metric for prompt caching evaluation.

Prior to each experimental condition, we execute warmup calls to prime the cache and record cache creation tokens separately from evaluation runs. Between conditions for different cache modes, we wait sufficient time (exceeding 24 hours) to ensure cache entries expire based on provider TTL policies, preventing cross-condition cache contamination.

3.5. Statistical Analysis

We compare each cache condition against the no-cache baseline using independent samples t-tests. Statistical significance is determined at $\alpha = 0.05$. For each model and cache mode, we report mean cost, mean TTFT, percentage improvement over baseline, and p-values. Sample sizes are $n = 40$ per condition for all models.

4. Results

4.1. Overall Results

Table 1 summarizes the prompt caching benefits across all four models using the best-performing cache mode for each model. All experiments show statistically significant improvements ($p < 0.05$). Cost savings range from 41% to 80% across models, while time to first token improvements range from 6% to 31%.

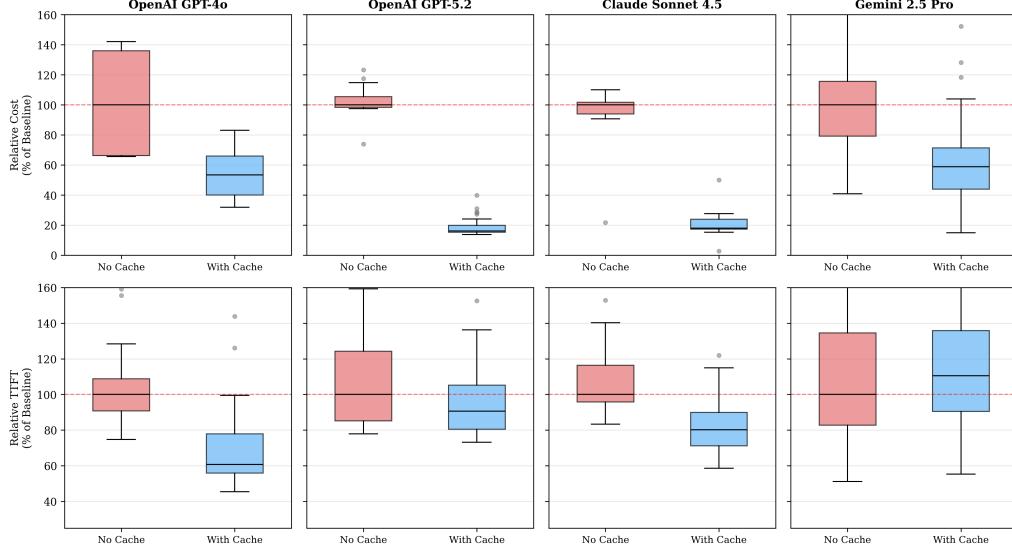


Figure 2. Prompt caching impact for normalized cost and time to first token (TTFT). Results use the system prompt only caching strategy. The no-cache baseline is normalized to 100% and lower values indicate better performance.

4.2. Cost Reduction

Prompt caching delivers substantial cost reductions across all providers and cache strategies. As shown in Table 2 and Figure 2, all cache modes achieve cost savings compared to the no-cache baseline for all four models. Cost reductions range from 79-81% for GPT-5.2, 78-79% for Claude Sonnet 4.5, 46-48% for GPT-4o, and 28-41% for Gemini 2.5 Pro depending on the cache mode selected. The consistency of cost savings across cache strategies suggests that the primary driver of cost reduction is caching the large system prompt, which remains stable across all requests within a session. Additional caching of conversation history and tool calls provides marginal incremental benefit for cost, as these components are smaller relative to the system prompt in our experimental setup.

Time to first token improvements show greater variation across providers compared to cost savings. GPT-4o shows 28-31% improvement with system prompt only and exclude tool results strategies, while full context caching exhibits a slight regression of 8.8%, suggesting that caching dynamic content can introduce overhead that negates latency benefits. Claude Sonnet 4.5 demonstrates consistent TTFT improvements across all cache strategies, ranging from 20.9% to 22.9%, including full context caching. This indicates that provider implementations differ in how they handle dynamic content caching. GPT-5.2 shows 13.0% improvement with the exclude tool results strategy, while Gemini 2.5 Pro shows 6.1% improvement with system prompt only caching. TTFT measurements exhibit natural variance due to factors including server load, network conditions, and provider infrastructure. This variance is reflected in the box

plot distributions in Figure 3.

4.3. Cache Strategy Comparison

Figure 3 presents normalized cost and TTFT distributions across all four cache strategies for each model. The results reveal important differences in how cache strategies perform across providers.

For cost optimization, all three caching strategies (full context, system prompt only, and exclude tool results) provide similar benefits within each model. The differences between strategies are typically within 2-4 percentage points, indicating that the system prompt, which is cached in all strategies, drives the majority of cost savings.

For latency optimization, the choice of cache strategy has a more pronounced impact. System prompt only caching and exclude tool results caching consistently outperform full context caching for TTFT improvement. For some models, full context caching shows no improvement or slight regression, while other strategies achieve 28-31% improvement. The likely explanation is that full context caching triggers cache writes for dynamic tool calls and results, introducing overhead that offsets the benefits of cache reads.

5. Discussion

5.1. Strategic Cache Boundary Control

Our results demonstrate that strategic control over cache boundaries is essential for maximizing prompt caching benefits in agentic workloads. The key insight is that providers abstract much of the underlying caching mechanism, auto-

Table 2. Full comparison of cache modes across all models. Cost savings and TTFT improvement are relative to the no-cache baseline. Negative TTFT values indicate regression.

Model	Cache Mode	Cost ↓	TTFT ↓
OpenAI GPT-5.2	No Cache (Baseline)	—	—
	Full Context	79.3%	9.5%
	System Prompt	81.4%	10.5%
	Excl. Tool Results	79.6%	13.0%
Claude Sonnet 4.5	No Cache (Baseline)	—	—
	Full Context	77.8%	21.8%
	System Prompt	78.5%	22.9%
	Excl. Tool Results	78.1%	20.9%
Gemini 2.5 Pro	No Cache (Baseline)	—	—
	Full Context	38.3%	6.0%
	System Prompt	41.4%	6.1%
	Excl. Tool Results	27.8%	-2.9%
OpenAI GPT-4o	No Cache (Baseline)	—	—
	Full Context	47.8%	-8.8%
	System Prompt	45.9%	30.9%
	Excl. Tool Results	46.8%	28.1%

matically triggering cache creation when token thresholds are exceeded. Without explicit boundary control, this automatic behavior can cache dynamic, session-specific content that will not be reused, leading to cache write overhead without corresponding read benefits.

The most effective strategy is to ensure that only stable, reusable content is cached. In agentic applications, the system prompt is the most stable component, containing agent instructions, tool definitions, and persona guidelines that remain constant across sessions. Conversation history, tool calls, and tool results are inherently dynamic and session-specific, making them poor candidates for cross-session caching. Practitioners should avoid including dynamic values in the system prompt itself. Common patterns that inadvertently break the cache include timestamps, datetime strings, session identifiers, or user-specific information embedded in the system prompt. If such dynamic information is necessary, it should be placed at the end of the system prompt to maximize the cacheable prefix. This ensures that the majority of the system prompt benefits from cache hits while only the dynamic suffix requires recomputation (Ji, 2025).

Similarly, dynamic function calling can break the cache when tool definitions change between requests. Modern agentic systems increasingly leverage dynamic tool discovery and registration through protocols such as the Model Context Protocol (MCP) (Model Context Protocol, 2026), where available tools may vary based on connected servers or runtime context (Lumer et al., 2025a). When tool definitions are included in the prompt, any change to the available tool set invalidates the cached prefix. A practical strategy is to maintain a fixed set of general-purpose, reusable functions (such as code execution, file operations, and bash/shell com-

mands), while implementing dynamic capabilities through code generation rather than traditional function calling (Ji, 2025; Wang et al., 2024; Jones & Kelly, 2025; Anthropic, 2025a; Varda & Pai, 2025; Hacker News, 2025). Those prior methods of dynamic function calling, while achieving strong retrieval and execution accuracy, can prevent prompt caching usage (Lumer et al., 2024; 2025b; Chen et al., 2024; Zheng et al., 2024; Chen et al., 2024; Wu et al., 2024b).

5.2. Tool Call Caching Considerations

For long-running agentic sessions with 30-50 or more tool calls, practitioners may consider caching tool calls and results to further reduce costs. However, this approach involves tradeoffs. Cache creation incurs a cost and latency overhead on the first request, which is only amortized if subsequent requests benefit from cache reads. For tool calls that produce highly variable results or that are unlikely to be repeated, caching provides no benefit and may introduce unnecessary overhead.

Furthermore, common context management strategies in agentic systems can interact poorly with tool call caching. Techniques such as summarizing or pruning old tool calls to manage context length (Ji, 2025) inherently modify the conversation history, breaking any cached representations of that content. If an application employs such strategies, caching tool calls becomes counterproductive. The emerging pattern for agentic applications is to maintain a large, stable system prompt that benefits from caching, while treating tool calls and results as dynamic content that may be summarized, pruned, or otherwise managed throughout the session.

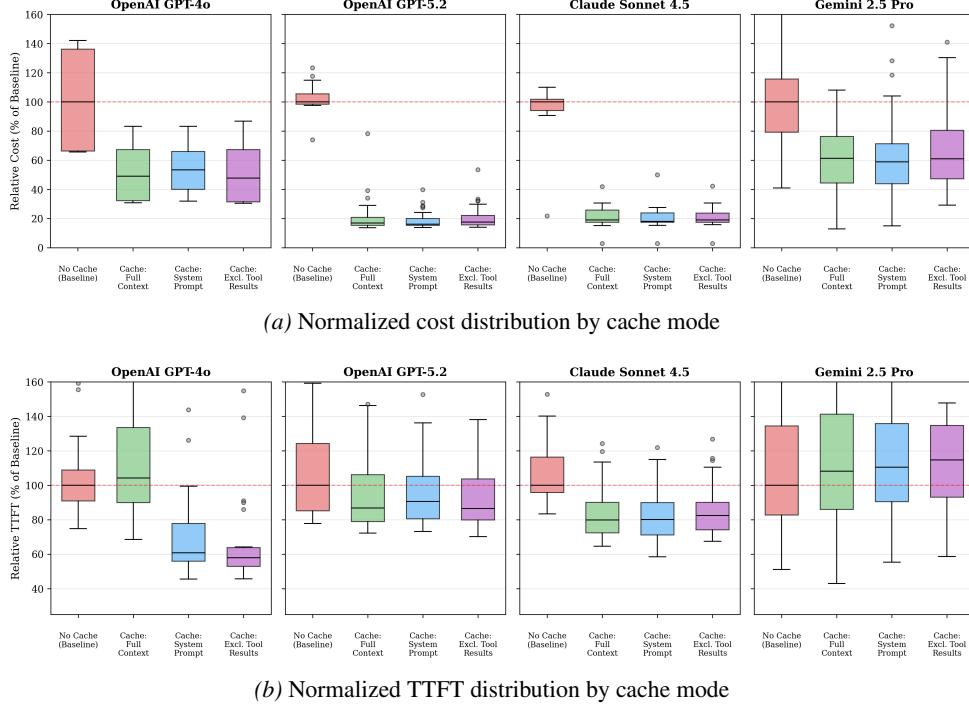


Figure 3. Normalized cost and time to first token (TTFT) distributions by model and cache strategy. The no-cache baseline is normalized to 100% and lower values indicate better performance. Cost reductions are consistent across cache strategies, while TTFT improvements vary significantly, with full-context caching sometimes underperforming more selective strategies.

5.3. Provider Implementation Variability

Provider implementations of prompt caching differ in important ways that affect practical deployment. Minimum token thresholds for cache eligibility range from 1,024 to 4,096 tokens depending on the provider and model (see Appendix A). Time-to-live (TTL) durations vary from 5 minutes to 24 hours, affecting whether cached content remains available across user sessions. Some providers offer automatic caching that activates without developer intervention, while others require explicit API parameters to enable caching. Enterprise deployments may also leverage dedicated caching infrastructure (Liu et al., 2024; Cheng et al., 2024; Yao et al., 2025; Cheng et al., 2025) to further optimize performance. These implementation details are subject to change and practitioners should consult current provider documentation when designing caching strategies.

Our results also reflect natural variance in API response times due to factors including server load, geographic distribution, and infrastructure differences across providers. When evaluating prompt caching benefits, practitioners should conduct experiments representative of their specific workloads and usage patterns rather than relying solely on published benchmarks. Practitioners should also be aware of security considerations, as recent work has demonstrated that prompt caching can introduce timing side-channels that

may leak information about cached content (Wu et al., 2025; Gu et al., 2025).

6. Conclusion

Recent advancements in Large Language Model (LLM) agents have enabled complex multi-turn agentic tasks requiring extensive tool calling, where conversations can span dozens of API calls with increasingly large context windows. While major LLM providers offer prompt caching to reduce costs and latency, the benefits of these features remain under-explored in the research literature. In this work, we present a comprehensive evaluation of prompt caching across three major LLM providers (OpenAI, Anthropic, and Google) using four flagship models. We compare three caching strategies, including full context caching, system prompt only caching, and caching that excludes dynamic tool results. We evaluate on DeepResearchBench, a multi-turn agentic benchmark where agents autonomously execute web search tool calls to answer complex research questions, measuring both API cost and time to first token (TTFT) across over 500 agent sessions with 10,000-token system prompts. Our results demonstrate that prompt caching reduces API costs by 45-80% and improves time to first token by 13-31% across providers. We find that strategic cache boundary control, such as excluding dynamic tool results,

provides more consistent benefits than naive full-context caching, which can paradoxically increase latency. As agentic systems continue to trend toward longer-running sessions with dozens of tool calls, strategic prompt caching becomes critical for reducing operational costs and improving user experience, and our findings provide practitioners with guidance for implementing prompt caching in production agentic systems.

References

- Agno. Agno: Framework for building ai agents and agentic workflows, 2026. URL <https://github.com/agno-agi/agno>. Accessed: 2026-01-08.
- Anthropic. Introducing advanced tool use on the Claude developer platform. <https://www.anthropic.com/engineering/advanced-tool-use>, 2025a. Engineering blog.
- Anthropic. Claude agent sdk (python): Sdk for building claude-powered agents, 2025b. URL <https://github.com/anthropics/clause-agent-sdk-python>. Accessed: 2026-01-08.
- Anthropic. Prompt caching with Claude. Claude Blog, August 2025c. URL <https://claude.com/blog/prompt-caching>.
- Anthropic. Prompt caching. Claude Docs, 2026. URL <https://platform.claude.com/docs/en/build-with-claude/prompt-caching>. Accessed 2026-01-07.
- Chen, Y., Yoon, J., Sachan, D. S., et al. Re-invoke: Tool invocation rewriting for zero-shot tool retrieval. *arXiv preprint arXiv*, 2024.
- Cheng, Y., Du, K., Yao, J., and Jiang, J. Do large language models need a content delivery network? *arXiv preprint arXiv:2409.13761*, 2024.
- Cheng, Y., Liu, Y., Yao, J., An, Y., Chen, X., Feng, S., Huang, Y., Shen, S., Du, K., and Jiang, J. Lmcache: An efficient kv cache layer for enterprise-scale llm inference. *arXiv preprint arXiv:2510.09665*, 2025.
- CrewAI. Crewai: Framework for orchestrating collaborative, autonomous ai agents, 2023. URL <https://github.com/crewAIInc/crewAI>. Accessed: 2026-01-08.
- Drouin, A., Gasse, M., Caccia, M., Laradji, I. H., Del Verme, M., Marty, T., Boisvert, L., Thakkar, M., Cappart, Q., Vazquez, D., Chapados, N., and Lacoste, A. WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks?, 2024. URL <https://arxiv.org/abs/2403.07718>.
- Du, M., Xu, B., Zhu, C., Wang, X., and Mao, Z. Deep-research bench: A comprehensive benchmark for deep research agents, 2025. URL <https://arxiv.org/abs/2506.11763>.
- FutureSearch: Bosse, N. I., Evans, J., Gambee, R. G., Hnyk, D., Mühlbacher, P., Phillips, L., Schwarz, D., and Wildman, J. Deep Research Bench: Evaluating AI Web Research Agents, 2025. URL <https://arxiv.org/abs/2506.06287>.
- Ge, S., Zhang, Y., Liu, L., Zhang, M., Han, J., and Gao, J. Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs. *arXiv preprint arXiv:2310.01801*, October 2023. doi: 10.48550/arXiv.2310.01801. URL <https://arxiv.org/abs/2310.01801>. ICLR 2024 (as listed on arXiv).
- Gim, I., Chen, G., seob Lee, S., Sarda, N., Khandelwal, A., and Zhong, L. Prompt Cache: Modular Attention Reuse for Low-Latency Inference. *Proceedings of Machine Learning and Systems (MLSys)*, 6, 2024. arXiv:2311.04934.
- Google. Agent development kit (adk): Open-source framework for building multi-agent applications, 2025. URL <https://github.com/google/adk-docs>. Accessed: 2026-01-08.
- Google Cloud. Prompt caching (Anthropic Claude models on Vertex AI). Vertex AI Documentation, 2026a. URL <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/partner-models/clause/prompt-caching>. Last updated 2026-01-02 UTC (accessed 2026-01-07).
- Google Cloud. Context caching overview. Generative AI on Vertex AI Documentation, 2026b. URL <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/context-cache/context-cache-overview>. Accessed 2026-01-07.
- Gu, C., Li, X. L., Kuditipudi, R., Liang, P., and Hashimoto, T. Auditing prompt caching in language model apis, 2025. URL <https://arxiv.org/abs/2502.07776>.
- Guan, S., Wang, J., Bian, J., Zhu, B., guang Lou, J., and Xiong, H. Evaluating llm-based agents for multi-turn conversations: A survey, 2026. URL <https://arxiv.org/abs/2503.22458>.
- Hacker News. Discussion: Code mode — the better way to use mcp. <https://news.ycombinator.com/item?id=45830318>, 2025. Accessed: 2025-11-08.
- Hugging Face. smolagents: A simple, code-first library for building agents, 2024. URL <https://github.com/huggingface/smolagents>. Accessed: 2026-01-08.
- Jeong, J. and Ahn, J. *Accelerating LLM Serving for Multi-turn Dialogues with Efficient Resource Management*, pp. 1–15. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400710797. URL <https://doi.org/10.1145/3676641.3716245>.
- Ji, Y. P. Context Engineering for AI Agents: Lessons from Building Manus. Manus Blog, July 2025. URL <https://manus.im/blog/Context-Engineering>

- for-AI-Agents-Lessons-from-Building-Manus.
- Jin, S., Liu, X., Zhang, Q., and Mao, Z. M. Compute Or Load KV Cache? Why Not Both? *arXiv preprint arXiv:2410.03065*, October 2024. doi: 10.48550/arXiv.2410.03065. URL <https://arxiv.org/abs/2410.03065>.
- Jones, A. and Kelly, C. Code execution with mcp: Building more efficient agents. <https://www.anthropic.com/engineering/code-execution-with-mcp>, November 2025. Published November 4, 2025; accessed 2025-11-06.
- Juravsky, J., Brown, B., Ehrlich, R., Fu, D. Y., Ré, C., and Mirhoseini, A. HydraGen: High-Throughput LLM Inference with Shared Prefixes. *arXiv preprint arXiv:2402.05099*, 2024. URL <https://arxiv.org/abs/2402.05099>.
- Kilpatrick, L. Gemini 2.5 Models now support implicit caching. Google Developers Blog, May 2025. URL <https://developers.googleblog.com/en/gemini-2-5-models-now-support-implicit-caching/>.
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient Memory Management for Large Language Model Serving with PagedAttention. *arXiv preprint arXiv:2309.06180*, September 2023. doi: 10.48550/arXiv.2309.06180. URL <https://arxiv.org/abs/2309.06180>.
- Laban, P., Hayashi, H., Zhou, Y., and Neville, J. Llms get lost in multi-turn conversation, 2025. URL <https://arxiv.org/abs/2505.06120>.
- LangChain. Deepagents: A library for building complex multi-step agents, 2025. URL <https://github.com/langchain-ai/deepagents>. Accessed: 2026-01-08.
- Li, M., Zeng, Y., Cheng, Z., Ma, C., and Jia, K. Report-Bench: Evaluating Deep Research Agents via Academic Survey Tasks, 2025. URL <https://arxiv.org/abs/2508.15804>.
- Liu, Y., Li, H., Cheng, Y., Ray, S., Huang, Y., Zhang, Q., Du, K., Yao, J., Lu, S., Ananthanarayanan, G., et al. Cachegen: Kv cache compression and streaming for fast large language model serving. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pp. 38–56, 2024.
- LlamaIndex. Llamaindex: Framework for building llm-powered agents over your data, 2022. URL https://github.com/run-llama/llama_index. Accessed: 2026-01-08.
- Lumer, E., Subbiah, V. K., Burke, J. A., Basavaraju, P. H., and Huber, A. Toolshed: Scale Tool-Equipped Agents with Advanced RAG-Tool Fusion and Tool Knowledge Bases, 2024. URL <https://arxiv.org/abs/2410.14594>.
- Lumer, E., Gulati, A., Nizar, F., Hedroits, D., Mehta, A., Hwangbo, H., Subbiah, V. K., Basavaraju, P. H., and Burke, J. A. Tool and Agent Selection for Large Language Model Agents in Production: A Survey. *Preprints*, December 2025a. doi: 10.20944/preprints202512.1050.v1. URL <https://doi.org/10.20944/preprints202512.1050.v1>.
- Lumer, E., Gulati, A., Subbiah, V. K., Basavaraju, P. H., and Burke, J. A. ScaleMCP: Dynamic and Auto-Synchronizing Model Context Protocol Tools for LLM Agents, 2025b. URL <https://arxiv.org/abs/2505.06416>.
- Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., and Scialom, T. GAIA: a benchmark for General AI Assistants, 2023. URL <https://arxiv.org/abs/2311.12983>.
- Microsoft. Autogen: A programming framework for agentic ai, 2023a. URL <https://github.com/microsoft/autogen>. Accessed: 2026-01-08.
- Microsoft. Semantic kernel: Model-agnostic sdk for building ai agents and multi-agent systems, 2023b. URL <https://github.com/microsoft/semantic-kernel>. Accessed: 2026-01-08.
- Microsoft. Agent framework: Sdk for building, orchestrating, and deploying ai agents and multi-agent workflows, 2026. URL <https://github.com/microsoft/agent-framework>. Accessed: 2026-01-08.
- Microsoft Azure AI. Prompt caching in Azure OpenAI Service (Foundry Models). Microsoft Learn Documentation, November 2025. URL <https://learn.microsoft.com/en-us/azure/ai-foundry/openai/how-to/prompt-caching>.
- Model Context Protocol. What is the model context protocol (mcp)? <https://modelcontextprotocol.io/docs/getting-started/intro>, 2026. Accessed: 2025-10-15.
- Not Lain. KV Caching Explained: Optimizing Transformer Inference Efficiency. Hugging Face Blog (Community Article), January 2025. URL <https://huggingface.co/blog/not-lain/kv-caching>.
- OpenAI. Swarm: Educational framework exploring ergonomic, lightweight multi-agent orchestration, 2024a. URL <https://github.com/openai/swarm>. Accessed: 2026-01-08.

- OpenAI. Prompt Caching in the API. OpenAI Blog, October 2024b. URL <https://openai.com/index/api-prompt-caching/>.
- OpenAI. Openai agents sdk (python): A lightweight framework for multi-agent workflows, 2025. URL <https://github.com/openai/openai-agents-python>. Accessed: 2026-01-08.
- OpenAI. Prompt caching. OpenAI API Documentation, 2026. URL <https://platform.openai.com/docs/guides/prompt-caching>. Accessed 2026-01-07.
- Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Levskaya, A., Heek, J., Xiao, K., Agrawal, S., and Dean, J. Efficiently Scaling Transformer Inference. *arXiv preprint arXiv:2211.05102*, November 2022. doi: 10.48550/arXiv.2211.05102. URL <https://arxiv.org/abs/2211.05102>.
- PromptHub. Prompt Caching with OpenAI, Anthropic, and Google Models. PromptHub Blog, August 2025. URL <https://www.promphub.us/blog/prompt-caching-with-openai-anthropic-and-google-models>. Last updated Aug 12, 2025 (accessed 2026-01-07).
- Shi, L., Zhang, H., Yao, Y., Li, Z., and Zhao, H. Keep the Cost Down: A Review on Methods to Optimize LLM's KV-Cache Consumption. *arXiv preprint arXiv:2407.18003*, July 2024. doi: 10.48550/arXiv.2407.18003. URL <https://arxiv.org/abs/2407.18003>. COLM 2024 (as listed on arXiv).
- Sun, H., Chang, L.-W., Bao, W., Zheng, S., Zheng, N., Liu, X., Dong, H., Chi, Y., and Chen, B. Shadowkv: Kv cache in shadows for high-throughput long-context llm inference, 2025. URL <https://arxiv.org/abs/2410.21465>.
- Varda, K. and Pai, S. Code mode: the better way to use mcp. <https://blog.cloudflare.com/code-mode/>, September 2025. Accessed: 2025-11-08.
- Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., and Ji, H. Executable code actions elicit better llm agents, 2024. URL <https://arxiv.org/abs/2402.01030>.
- Wei, J., Sun, Z., Papay, S., McKinney, S., Han, J., Fulford, I., Chung, H. W., Passos, A. T., Fedus, W., and Glaese, A. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents, 2025. URL <https://arxiv.org/abs/2504.12516>.
- Wu, G., Zhang, Z., Zhang, Y., Wang, W., Niu, J., Wu, Y., and Zhang, Y. I Know What You Asked: Prompt Leakage via KV-Cache Sharing in Multi-Tenant LLM Serving. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, February 2025. doi: 10.14722/ndss.2025.241772. URL <https://dx.doi.org/10.14722/ndss.2025.241772>.
- Wu, H. et al. Layer-Condensed KV Cache for Efficient Inference of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a. URL <https://aclanthology.org/2024.acl-long.602/>.
- Wu, M., Zhu, T., Han, H., et al. Seal-tools: Self-instruct tool learning dataset for agent tuning and detailed benchmark. *arXiv preprint*, 2024b.
- Yan, J., Ni, W., Chen, L., Lin, X., Cheng, P., Qin, Z., and Ren, K. Contextcache: Context-aware semantic cache for multi-turn queries in large language models, 2025. URL <https://arxiv.org/abs/2506.22791>.
- Yao, J., Li, H., Liu, Y., Ray, S., Cheng, Y., Zhang, Q., Du, K., Lu, S., and Jiang, J. Cacheblend: Fast large language model serving for rag with cached knowledge fusion. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 94–109, 2025. doi: 10.1145/3689031.3696098. URL <https://doi.org/10.1145/3689031.3696098>.
- Zheng, Y., Li, P., Liu, W., et al. Toolrerank: Adaptive and hierarchy-aware reranking for tool retrieval. *arXiv preprint*, 2024.
- Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., and Neubig, G. WebArena: A Realistic Web Environment for Building Autonomous Agents, 2023. URL <https://arxiv.org/abs/2307.13854>. v4 last revised Apr 16, 2024.
- Zhou, Z., Ning, X., Hong, K., Fu, T., Xu, J., Li, S., Lou, Y., Wang, L., Yuan, Z., Li, X., Yan, S., Dai, G., Zhang, X.-P., Dong, Y., and Wang, Y. A survey on efficient inference for large language models, 2024. URL <https://arxiv.org/abs/2404.14294>.

A. Prompt caching pricing at time of evaluation

Table 3. Token pricing as of early January 2026, used in our cost analysis (USD per 1M tokens unless otherwise noted). “Cached input” refers to cache-hit tokens. “Cache write” denotes cache creation costs when explicitly priced. Google additionally charges for context cache storage (\$4.50 per million tokens per hour), which is accounted for separately in our analysis. Pricing reflects public provider documentation at the time of evaluation. ([OpenAI, 2026](#); [Anthropic, 2026](#); [Google Cloud, 2026a](#))

Provider	Model	Input	Output	Cached	Write
OpenAI	GPT-4o	2.50	10.00	1.25	—
OpenAI	GPT-5.2	1.75	14.00	0.175	—
Anthropic	Claude Sonnet 4.5	3.00	15.00	0.30	3.75
Google	Gemini 2.5 Pro ($\leq 200K$)	1.25	10.00	0.125	—
Google	Gemini 2.5 Pro ($> 200K$)	2.50	15.00	0.250	—

Table 4. Minimum prompt length (in tokens) required for prompt caching to apply, as of early January 2026. Prompts shorter than these thresholds cannot benefit from caching, even when caching features are enabled. Thresholds reflect public provider documentation at the time of evaluation. ([OpenAI, 2026](#); [Anthropic, 2026](#); [Google Cloud, 2026b](#))

Provider	Model	Min. Tokens
OpenAI	GPT-4o	1,024
OpenAI	GPT-5.2	1,024
Anthropic	Claude Sonnet 4.5	1,024
Google	Gemini 2.5 Pro	4,096

B. Prompt Caching Mechanism

Figure 4 illustrates the fundamental mechanism underlying prompt caching. When a request is processed, the system checks whether the prompt prefix matches previously cached content. A cache hit occurs when the entire prefix matches exactly, allowing the system to reuse previously computed KV tensors (shown in green). A cache miss occurs when any token differs from the cached content, even at the very beginning (shown with an orange indicator), forcing complete recomputation of all tokens (shown in gray).

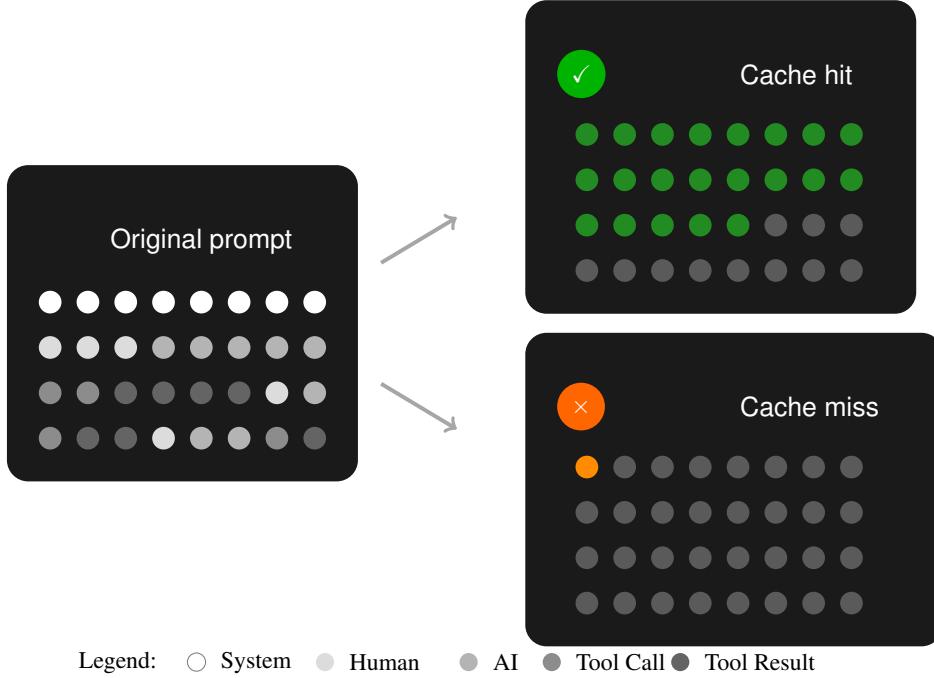
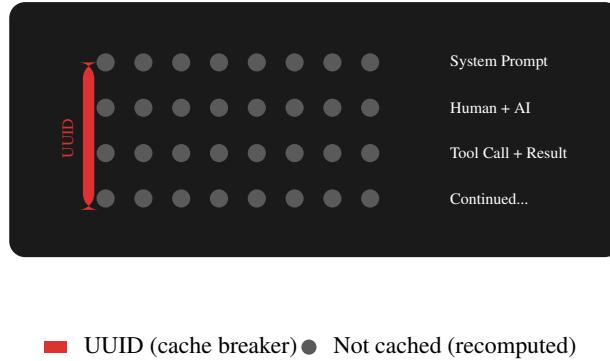


Figure 4. Prompt caching requires exact prefix matches. Different shades represent message types in agentic conversations: brightest (system prompt), light gray (human messages), medium gray (AI messages), darker gray (tool calls), and darkest (tool results). Cache hit: The prompt prefix matches a previously seen request exactly, so cached KV tensors are reused (green) and only new tokens appended at the end require computation (gray). Cache miss: Any difference in the prefix—even a single token at the beginning (orange)—prevents cache reuse, forcing full recomputation of all tokens (gray).

C. Cache Strategy Implementations

The following figures illustrate the four cache strategies evaluated in this work. Since prompt caching operates on exact prefix matches, we use UUIDs (indicated by red bars) to control cache boundaries. Static content placed before the UUID forms the cacheable prefix; content after the UUID varies between requests and prevents prefix matches beyond that point.

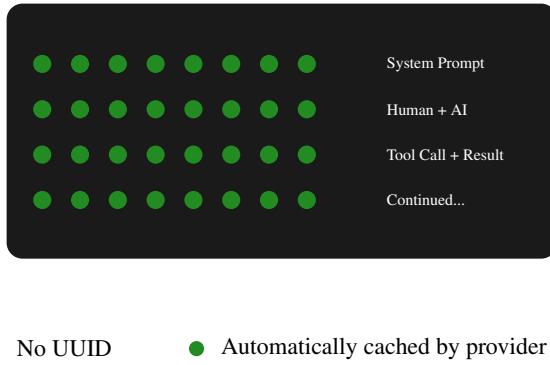
C.1. No Cache (Baseline)



■ UUID (cache breaker) ● Not cached (recomputed)

Figure 5. No Cache (Baseline): A unique UUID prepended to the start of the system prompt ensures no prefix match is possible with any prior request, forcing full recomputation of all tokens every time.

C.2. Full Context Caching



No UUID ● Automatically cached by provider

Figure 6. Full Context Caching: No UUIDs are added, allowing the provider to automatically cache the entire prompt prefix. However, this may cache dynamic content (e.g., tool results) that varies between sessions, potentially triggering cache writes without corresponding cache hits.

C.3. System Prompt Only Caching

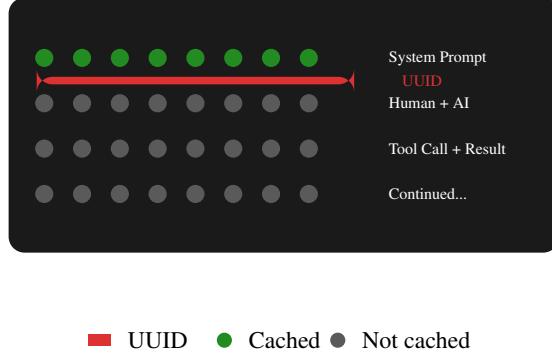


Figure 7. System Prompt Only Caching: A UUID appended after the system prompt breaks the cacheable prefix at this boundary. The static system prompt (placed at the beginning) benefits from prefix caching, while dynamic conversation content (placed after) is recomputed each request.

C.4. Exclude Tool Results Caching

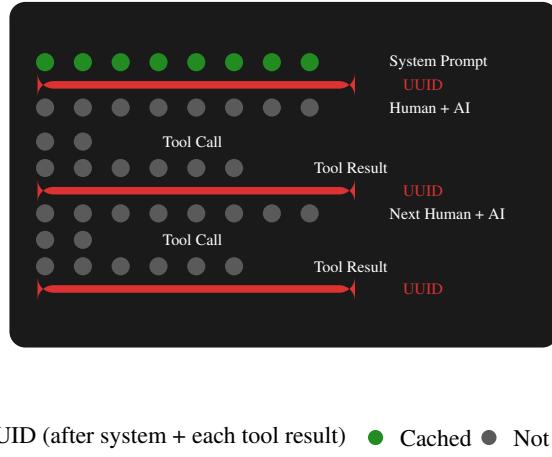


Figure 8. Exclude Tool Results Caching: UUIDs are appended after the system prompt and after each tool result to break the cacheable prefix at these boundaries. This prevents session-specific tool results from being cached, avoiding cache writes for content unlikely to produce future cache hits. Furthermore, this mirrors cache-breaking context engineering strategies that prune or summarize past tool calls.