HippoCamp: Benchmarking Contextual Agents on Personal Computers

Zhe Yang[1], Shulin Tian[1], Kairui Hu[1,2], Shuai Liu[1,2], Hoang-Nhat Nguyen[1], Yichi Zhang[1], Zujin Guo[1], Mengying Yu[1], Zinan Zhang[1], Jingkang Yang[1], Chen Change Loy[1,2]✉, and Ziwei Liu[1]✉

[1] S-Lab, Nanyang Technological University, Singapore [2] Synvo AI, Singapore

Fig. 1: Overview of task in HippoCamp benchmark. HippoCamp is a benchmark designed to evaluate agents' ability to search, perceive, and reason over long-term, realistic, large-scale personal file systems. It instantiates profile-local multimodal file-system environments containing tens of gigabytes of profile-specific assets across diverse modalities. Through tasks that demand grounded retrieval, cross-file pattern inference, and long-horizon reasoning, the benchmark rigorously evaluates the personalized multimodal memory of contextual agents.

Abstract. We present HippoCamp, a new benchmark designed to evaluate agents' capabilities on multimodal file management. Unlike existing agent benchmarks that focus on tasks like web interaction, tool-use, or software automation in generic settings, HippoCamp evaluates agents in user-centric environments to model individual user profiles and search from massive personal files for context-aware reasoning. Our benchmark instantiates device-scale file systems over real-world profiles spanning diverse modalities, comprising 42.4 GB of data across over 2K heterogeneous files. Building upon the raw files, we construct 581 QA pairs to assess agents' capabilities in search, evidence perception, and multi-step reasoning. To facilitate fine-grained analysis, we provide 46.1K fine-grained trajectory annotations for step-wise failure diagnosis. We evaluate a wide range of state-of-the-art multimodal large

✉ Corresponding authors.

language models (MLLMs) and agentic methods on HippoCamp. Our comprehensive experiments reveal a significant performance gap: even the most advanced commercial models achieve merely a 48.3% accuracy in user profiling, struggling particularly with long-horizon retrieval and cross-modal reasoning within dense personal file systems. Furthermore, our step-wise failure diagnosis identifies multimodal perception and evidence grounding as the primary bottlenecks. Ultimately, HippoCamp exposes the critical limitations of current agents in realistic, user-centric environments and provides a robust foundation for developing next-generation personal AI assistants. Our dataset is publicly available at: https://huggingface.co/datasets/MMMem-org/HippoCamp.

Keywords: Multimodal Agents · File-System · Contextual Benchmarking · Personalized Memory

1 Introduction

The rise of large multimodal models (LMMs) and agentic systems has accelerated progress toward intelligent assistants capable of operating across perception, reasoning, and action [38, 51, 55]. A natural and impactful application of such systems lies in **personalized digital environments** [45], where agents manage, retrieve, and reason over users' heterogeneous digital assets [23]. Unlike web-based or task-specific agents [27, 47, 56], personalized agents operating in local digital environments must navigate continuously evolving devices composed of multimodal files, diverse formats, and long-term behavioral traces [43, 49]. Reasoning in such contexts requires integrating perception across modalities with retrieval and reflection over past interactions, paralleling the role of the human hippocampus in contextual memory consolidation and recall.

However, despite recent progress in long-context reasoning and multimodal retrieval [24], there remains no standardized benchmark for evaluating an agent's ability to understand, recall, and reason over massive, personalized, multimodal file systems. While recent efforts address document-level multimodal retrieval [9] or personalized tool-use planning [52], none target the scale and heterogeneity of personal computing environments. Existing evaluations predominantly target general domains such as web automation [15, 19, 36], code generation [58], document understanding [31], or embodied planning [35], while neglecting the rich information encoded in personalized file systems. Although these benchmarks have advanced perception and tool-use capabilities, they operate in isolated, goal-specific scenarios detached from the user's personal context. Consequently, they fail to capture the long-term continuity, complex cross-referencing and validation across interrelated multimodal files over time, and the personalized reasoning required for realistic personal computing.

To fill this gap, we introduce **HippoCamp**, a benchmark for evaluating memory-augmented agents in realistic personal computing environments, as illustrated in Fig. 1. HippoCamp constructs *three* high-density archetypal personal file systems from real-world personal-device artifacts. Rather than serving as demographic samples or a literal three-way categorization of personal-computing

users, these environments are *archetypal* instantiations of a high-dimensional profile space. Together, they expose device-scale file-system challenges such as deep hierarchical organization, heterogeneous long-tail file formats, and cross-modal evidence dependencies. The benchmark encompasses two diagnostic task categories that mirror authentic computer usage: *factual retention* and *profiling*. Each task demands agentic behaviors integrating **search, perception, and reasoning**, enabling systematic evaluation of personalized multimodal intelligence. In summary, HippoCamp makes three main contributions:

1. **Personal file-system environments.** We construct profile-local, device-scale file systems through three distinct, file-intensive profiles. This design captures long-term continuity, idiosyncratic folder structures, and the interconnected nature of real-world personal file systems.
2. **Device-scale corpus with dense supervision.** We provide a massive dataset comprising over 2000 heterogeneous files (totaling 42.4 GB), accompanied by 581 user-need-driven, evidence-grounded queries. To support diagnostic analysis, we provide 46.1K fine-grained annotations, including file grounding, localized evidence, rationale steps, capability labels, and atomic units; these fields support evidence-level evaluation and step-wise failure diagnosis rather than constituting additional evaluation questions.
3. **Diagnostic agent capability evaluation.** We formulate a rigorous question-answering benchmark built on two core task categories: *factual retention*: retrieving specific information, and *profiling*: inferring user-level patterns, constraints, routines, and preferences. These tasks go beyond simple retrieval, requiring agents to combine file-system search, multimodal evidence perception, and personalized reasoning across interrelated files over time.

2 Related Works

Benchmarks for Multimodal Contextual Agents. Contextual retrieval benchmarks span from text-centric datasets [11, 32, 40, 54] to multimodal and agentic settings driven by richer evidence needs. MultimodalQA [39] supports cross-modal grounding over text, images, and tables, while WebQA [4] studies retrieval from distractor-containing candidate pools. Document benchmarks such as M3DocRAG [6] and MMDocRAG [10] emphasize fine-grained selection under noise. Agentic benchmarks move from static corpora to interactive environments: WebShop [56] and PaperBench [37] evaluate goal-driven action sequences, and MetaTool [16], MINT [44], InfoDeepSeek [50], and BrowseComp [47] probe tool use, multi-turn reasoning, and web exploration. These benchmarks largely assume public data and fully observable states, leaving long-lived personalized context and device-resident multimodal evidence underexplored. Real personal environments contain identity cues, behavioral histories, and longitudinal records across diverse file types—signals largely absent from current evaluations.

Agentic Systems with Memory and Personalization. Agentic systems perform multi-step workflows in interactive environments and thus require

Table 1: Comparison of Benchmarks Related to Multimodal Context and Agentic Reasoning. For modalities, T denotes text, denotes images, denotes documents, denotes videos, and denotes audio. HippoCamp introduces a personalized multimodal file-system environment that spans all these modalities, enabling cross-file and cross-modal reasoning tasks at real-world scale.

Benchmark	Source	Modalities	#Samples	Multimodal	User Profile	File-System
HotpotQA [54] / KILT [32]	Wikipedia	T	∼100k	✗	✗	✗
BrowseComp [47]	Web	T	1,266	✗	✗	✗
MetaTool [16] / MINT [44]	Web + Tools	T (w/ code)	∼3k–20k	✗	✗	✗
WebQA [4]	Web	T	7.5k	✓	✗	✗
GAIA [27] / WebShop [56] / PaperBench [37]	Web	T	∼1k–10k	✓	✗	✗
MultiModalQA [39]	Web	T (Table-Only)	29k	✓	✗	✗
MMDocRAG [10] / M3DocRAG [6]	Documents	T	∼4k–10k	✓	✗	✗
LoCoMo [26]†	Personal Lifelog	T	300	✗	✓	✗
EgoLifeQA [53] / Ego-R1-Bench [42]‡	Personal Lifelog	T	∼5k	✓	✓	✗
HippoCamp (Ours)	**Personal File Systems**	T	**581 QA pairs**	✓	✓	✓

† LoCoMo contains 300 text-only personal questions;
‡ EgoLifeQA/Ego-R1-Bench provides approximately 20% personalized samples, and is limited to videos/audio single-modal context.

context integration; memory is central to long-horizon coherence. Prior work explores (i) *trajectory-based* memory, e.g., internalizing experience via finetuning or reinforcement learning [12, 61]; (ii) *retrieval-based* memory, e.g., storing and retrieving past episodes or tool traces [25, 63]; and (iii) *skill distillation*, compressing reusable skills into inference-time modules [46, 64]. Other lines build structured episodic/semantic memories to organize interaction histories and improve multi-step reliability [53, 60]. Building on these mechanisms, personalization operationalizes memory at the user level: PersonaAgent [62] maintains user-specific episodic/semantic records, while recent systems learn user knowledge graphs, preference embeddings, or long-term histories to adapt without parameter updates [20, 45]. Complementarily, Telemem [5] consolidates user-grounded interactions into narrative and multimodal episodic memories, enabling personalized retrieval over dialogue and visual experience. However, existing evaluations are often small-scale or synthetic and are typically built around curated interaction histories or narrow modalities (e.g., text/webpages), rather than unstructured personal file systems where evidence is dispersed across text, images, documents, videos, and audio. This motivates benchmarks that test whether agents can recover, ground, and reason over user-specific evidence at device scale.

Multimodal Contextual Retrieval. Retrieval-augmented generation (RAG) uses retrieval as external memory to surface evidence beyond the context window, while reasoning-centric models such as DeepSeek-R1 [13] refine intermediate traces and invoke tools during multi-step inference. Multimodal RAG generalizes text retrieval to heterogeneous evidence: document systems [6, 9] encode layout and mixed text–image content for fine-grained grounding; image–text frameworks [14, 59] fuse visual and textual features for cross-modal alignment; and video methods such as VideoRAG [34] index temporal clips for long-range visual retrieval. Despite broader coverage, top-k retrieval remains brittle when cues are dispersed across many files or long time spans. Reasoning models partially mitigate this by generating intermediate traces to coordinate search: Search-R1 [18]

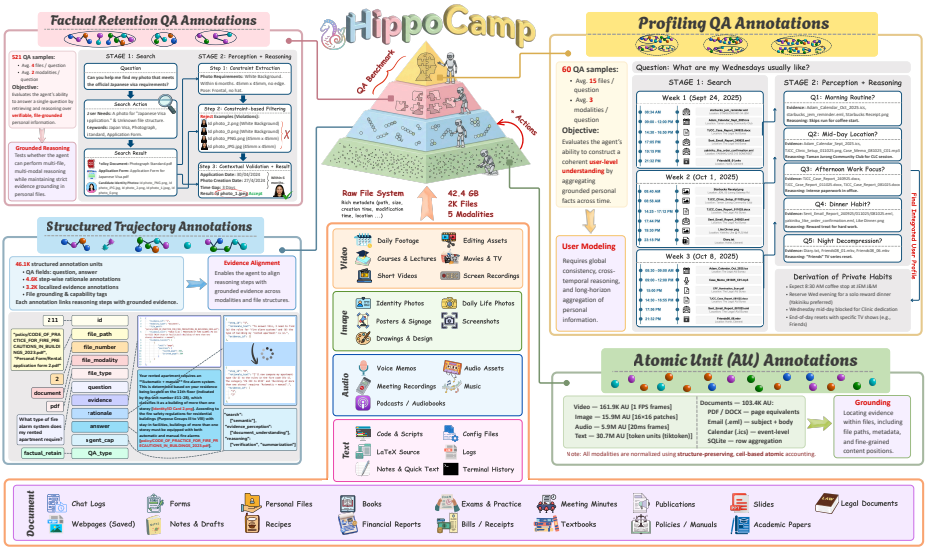

Fig. 2: HippoCamp hierarchical annotation schema. The pyramid organizes supervision from atomic grounding and action traces to structured trajectories and task-level QA, with increasing **abstraction and aggregation** toward user-level memory. *Factual retention* relies on localized evidence and intermediate reasoning, while *profiling* requires long-horizon, cross-modal synthesis into user-level inference.

and MMSearch-R1 [48] iteratively refine queries, and Ego-R1 [42] extends this paradigm to egocentric streams. However, both RAG and reasoning models primarily assume public or task-bounded retrieval spaces, rather than the personalized, long-lived multimodal context of real user ecosystems. Our benchmark targets this setting directly (see Tab. 1), evaluating models where relevant signals accumulate across all modalities, and where both retrieval and reasoning must operate within authentic, user-specific digital environments.

3 The HippoCamp Benchmark

HippoCamp is a benchmark for evaluating memory-augmented agents in realistic, device-resident personal file systems. It models personal computing as a multimodal, long-tail information space organized by file-system structure and temporal metadata, and formulates personalized file understanding as open-ended, evidence-grounded question answering. Fig. 2 specifies the benchmark's *supervision hierarchy*, progressing from localized evidence and action traces to structured trajectories and task-level evaluation, with increasing abstraction toward user-level memory. Within this structure, **factual retention** requires retrieving and reasoning over verifiable file-grounded facts, whereas **profiling** synthesizes grounded facts across time into user-level inferences such as preferences, behavioral patterns, scheduling information, retrospective reflections, and workflows.

Both tasks require **search**, **perception**, and **reasoning** to locate files, interpret multimodal contents, and integrate cross-file, cross-temporal evidence into context-aware answers.

3.1 Dataset Construction

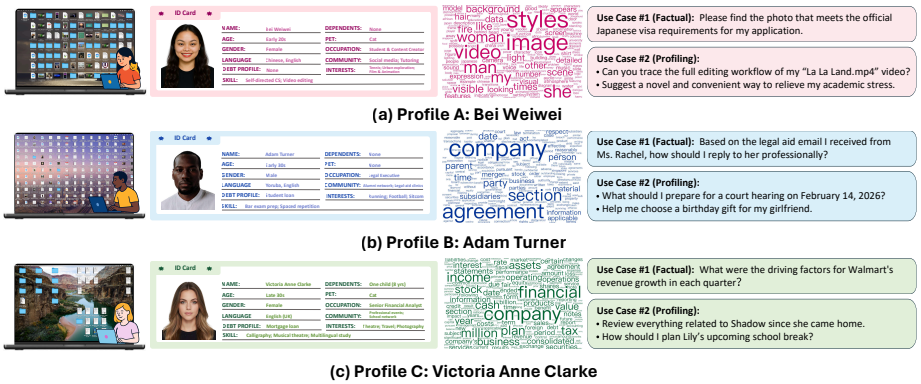

Fig. 3: Archetypal user profiles in HippoCamp. Each profile instantiates a distinct personal-device environment characterized along multiple dimensions, and is paired with a *word-cloud summary of text-converted file contents*, together with *factual retention* and *profiling* QA examples for (a) Bei, (b) Adam, and (c) Victoria.

Source. HippoCamp is derived from a coverage-oriented source pool formed through interviews with 100+ personal-device users. Source selection retained candidates whose file systems exhibited stable routines, heterogeneous organization, longitudinal traces, and cross-file evidence sufficient for auditable user-level inference. We aggregated selected contributors' files into three coherent high-density archetypal profiles by matching modality/file-type distributions, directory organization, temporal scope, and recurring workflows. The released profiles therefore function as profile-local diagnostic environments rather than raw single-user archives (see Fig. 3 for detailed specifications): Bei emphasizes a multimodal-heavy student/content-creation setting, Adam a document-centric legal/professional setting, and Victoria a finance/workplace setting with analytical, scheduling, and administrative records. Coherence checks, privacy filtering, and consistent pseudonymization produce unindexed "haystack" file systems that preserve task-relevant messiness such as native folder hierarchies, irregular naming, long-tail formats, metadata cues, and user-generated redundancy. All contributor data were processed under opt-in consent, privacy filtering, pseudonymization, and participant review before release (Supplementary Secs. A.1–A.5). To cover underrepresented professional document forms while preserving privacy, we use limited FinanceBench [17] and LegalBench-RAG [33] materials as document-form enrichment. Imported materials are sanitized to

match the fictional profile identities and re-annotated under the HippoCamp schema; details are provided in Supplementary Sec. A.6.

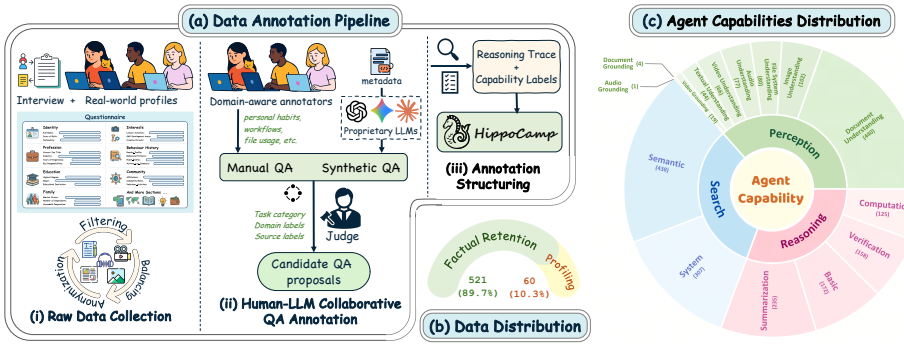

Fig. 4: Benchmark overview. (a) Human-led QA construction pipeline for grounded trajectory construction. The LLM-assisted branch denotes proposal cues for coverage checking, not finalized synthetic QA. (b) Distribution of *factual retention* and *profiling* tasks. (c) Distribution of annotated agent capability labels over *search*, *perception*, and *reasoning*.

Annotation. As shown in Fig. 4(a), QA construction is human-led, with LLMs used only as bounded proposal tools for coverage checking. Domain-aware annotators, drawn from the contributor groups underlying each archetypal profile, manually author user-driven questions grounded in their own files, routines, and realistic personal-computing needs. LLMs [7] may suggest candidate directions from restricted metadata and seed examples to surface coverage blind spots, but they do not author benchmark records: 0% of released QA/trajectory records are directly model-generated or accepted verbatim from LLM output. Retained items are human-finalized, file-verified, deduplicated, and balanced over task family, modality combination, and evidence-set size. For each finalized QA, annotators construct a grounded trajectory with step-wise rationale, explicit file-grounded evidence, and agent capability labels. Further details are provided in Supplementary Sec. B.

3.2 Tasks

Each QA pair is annotated as a structured trajectory stored as a JSON record, including the question and ground-truth answer, a step-wise rationale, and fine-grained localized evidence with atomic pointers into file content (e.g., page indices, table cells, or textual spans), together with file-level grounding metadata. Each trajectory is further labeled with **agent capability** tags (Fig. 4(c)) that decompose the required behaviors into three stages: **search** (system-level navigation, semantic retrieval), **perception** (file-system understanding, modality-specific understanding and grounding for text, documents, images, videos, and

audio), and **reasoning** (basic inference, computation, summarization, and verification). All trajectories are authored by human annotators following a minimalist annotation principle that explicitly links these stages, enabling interpretable diagnosis of multimodal agent behavior. This schema supports fine-grained capability analysis, evidence-level benchmarking, and evaluation of long-horizon, multi-step reasoning in realistic personalized file ecosystems.

Profiling. The profiling tasks evaluate whether an agent can construct a coherent user-level model from device-resident files by aggregating grounded personal facts across time. Each query targets high-level persona attributes, including preferences and routines, scheduling constraints, retrospective accounts, and workflow patterns spanning life, work, and study. Unlike single-fact queries, profiling requires synthesizing heterogeneous cues from multimodal content, file-system organization, and temporal regularities, and producing a globally consistent, actionable response. For example, "For the afternoon on October 27, 2025, schedule a good plan for me." tests whether the agent can integrate evidence from relevant files (e.g., calendars and communications) with historical routines and stated preferences to produce a feasible, personalized plan consistent with existing commitments.

Factual retention. The factual retention tasks focus on evaluating an agent's capability to retrieve, comprehend, and reason over factual information distributed across multimodal files within the user's device. Unlike conventional open-domain QA, these tasks are grounded in user-specific, device-scale contexts, requiring the agent to accurately locate relevant files, interpret heterogeneous content types, and integrate information to produce precise, context-aware answers. For example, a query such as "I have notes on the maximum flow problem. Which class is the notes record from, and what is the course duration?" tests whether the agent can identify, understand, and reason over content stored in diverse formats. Solving such tasks does not only require fine-grained retrieval and multimodal comprehension but also long-horizon contextual reasoning across temporally and semantically related files.

4 Experiment

4.1 Experimental Setup

All evaluations are conducted in a controlled, profile-isolated setting. For each test case, a method is given access only to the corresponding profile's file system in HippoCamp, with no external retrieval, web access, or auxiliary side-channel metadata. Agents receive the natural-language query as input and may freely explore the environment under full access permissions, using their native mechanisms to search, perceive, and reason over the multimodal file corpus. Model-system results in Tab. 2 follow a **max-budget protocol**: each method runs up to its predefined step, token, or platform budget rather than being truncated by an additional fixed wall-clock cutoff; human solvers receive no artificial wall-clock cap and serve only as a solvability reference under the same answer-plus-evidence protocol. We evaluate three method regimes under the

Table 2: Main results on HippoCamp. Representative MLLMs and agent methods are evaluated on *profiling* and *factual retention*; F1/Acc are reported per archetypal profile and as a profile-balanced average over the three profiles. Values are percentages (one decimal; % omitted). Best/second-best results among model systems are highlighted/underlined; human results provide a solvability reference.

Method	Profiling								Factual Retention							
	(a) Bei		(b) Adam		(c) Victoria		Overall		(a) Bei		(b) Adam		(c) Victoria		Overall	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
RAG Methods																
Standard RAG [21]	13.7	10.0	20.8	35.0	**20.6**	35.0	18.4	26.7	29.7	24.2	39.7	42.7	20.5	23.6	30.0	30.2
Self RAG [2]	13.8	5.0	16.0	25.0	15.9	0.0	15.2	10.0	33.9	26.1	41.5	38.8	20.2	17.7	31.9	27.5
Search Agent Methods																
ReAct [57] (Qwen3-30B-A3B [41])	5.5	5.6	17.8	25.0	12.2	10.0	11.8	13.5	**42.4**	26.1	**60.4**	37.9	26.5	21.7	**43.1**	28.5
ReAct [57] (Gemini-2.5-flash [7])	13.7	10.0	21.4	25.0	20.5	25.0	18.5	20.0	26.9	24.2	35.7	55.3	17.0	36.4	26.5	38.7
Search-R1 [18]	6.6	0.0	16.5	15.0	9.4	0.0	10.8	5.0	38.7	23.7	58.0	28.2	26.4	24.1	41.0	25.3
Autonomous Agent Systems																
Terminal Agent (Qwen3-VL-8B-Instruct [3])	5.4	0.0	16.3	25.0	13.2	25.0	11.6	16.7	14.6	10.7	21.6	13.6	15.7	10.3	17.3	11.5
Terminal Agent (Gemini-2.5-flash [7])	9.0	5.0	17.0	45.0	19.0	25.0	15.0	25.0	21.8	18.1	33.1	31.1	24.4	20.7	26.4	23.3
Terminal Agent (GPT-5.2 [29])	8.1	15.0	14.9	45.0	10.5	30.0	11.1	30.0	13.0	29.8	31.6	59.2	29.0	55.7	24.6	48.2
ChatGPT Agent Mode [28]	**23.8**	**35.0**	**22.7**	**55.0**	16.7	**55.0**	**21.0**	**48.3**	20.4	**31.2**	56.2	**90.3**	**29.3**	**67.0**	35.3	**62.8**
Human Solvability Reference																
Human Solvers (3 non-author)	82.0	87.0	89.0	90.0	93.0	84.0	88.0	87.0	95.0	88.0	87.0	78.0	84.0	79.0	89.0	82.0

same profile-local file structures and without additional supervision, external plug-ins, or manually curated signals: retrieval-augmented baselines index the full corpus with method-required reformatting, search agents interleave retrieval with iterative reasoning/tool use, and autonomous agents operate either in a Dockerized Ubuntu environment replicating the full file system or in non-vacuum product-grade agent modes. Supplementary Secs. D.1–D.3 provide implementation, tooling, and budget details.

4.2 Baseline Methods

RAG methods. Standard RAG [21] and Self-RAG [2] represent classical retrieval-and-generation pipelines. They rely on semantic retrieval over multimodal embeddings and assume that relevant evidence can be surfaced within a retrieved candidate set. These methods do not explicitly model the multi-step inspection, adaptive search, cross-file synthesis, or persistent-memory capabilities targeted by HippoCamp.

Search agent methods. Search-agent baselines embed retrieval within an explicit multi-turn search–reason loop. Following the ReAct framework [57], we instantiate ReAct with both Gemini 2.5 Flash [7] and Qwen3-30B-A3B [41] models, where the system alternates between reasoning steps and explicit search actions, conditioning subsequent generation on retrieved observations. Search-R1 [18] similarly interleaves reasoning and search through structured tags, enabling dynamic evidence acquisition across multiple steps.

Autonomous agent systems. We evaluate Docker-based autonomous agents instantiated with multiple model backends, including Gemini 2.5 Flash [7], GPT-5.2 [29], Claude Sonnet 4.5 [1], and Qwen3-VL-8B-Instruct [3], alongside proprietary product-grade agent modes. In the main table, we report ChatGPT

Agent Mode [28]; Claude-based agent settings are discussed separately due to severe file-system and long-document processing failures. These systems operate through recursive tool use, including issuing queries, browsing file previews, interpreting intermediate results, and updating hypotheses. They represent widely used publicly accessible paradigms for grounded reasoning, though their tool policies and architectural assumptions differ significantly.

4.3 Evaluation Protocol

Under the shared profile-local interface described above, HippoCamp evaluates two core dimensions: **question answering quality** and **evidence retrieval accuracy**. Both are computed from the same submitted answer and evidence outputs, so answer correctness and file grounding are measured under identical access constraints. The corresponding LLM-judge and file-set metrics are specified below; detailed execution-regime, budget, and robustness specifications are provided in Supplementary Secs. D.1–D.5.

Question answering evaluation. For all tasks, we adopt an **LLM-as-a-judge** [22] paradigm. A strong reference LLM, GPT-4o [30], is prompted with the question, ground-truth answer, and the model's generated response, and instructed to produce a binary correctness judgment (yes/no) together with a quality score on a standardized 0–5 scale. The prompt explicitly instructs the judge to consider factual alignment, reasoning soundness, and contextual personalization. We report overall accuracy, measured as the fraction of responses judged correct. We further calibrate the default GPT-4o judge on matched model outputs using Claude Sonnet 4.5 [1], DeepSeek V3.2 [8], and human verification. The calibration shows no systematic provider-specific shift in the relative conclusions: model rankings and group-level trends remain stable across judges, so the main findings are not driven by the choice of GPT-4o as the default judge. **Evidence retrieval evaluation.** For tasks requiring document or file retrieval, we assess retrieval quality using recall hit rate and F1 score based on the ground-truth evidence file set. F1 captures the balance between retrieving relevant files and avoiding spurious ones, while recall measures the agent's ability to identify all necessary evidence supporting correct reasoning.

4.4 Experimental Results

Tab. 2 reports results across all profiles and both task types. Human solvers achieve 88.0/87.0 File-F1/Acc on profiling and 89.0/82.0 on factual retention, indicating that the tasks are solvable under profile-local file access. In contrast, the best-performing evaluated agent setting reaches only 21.0/48.3 and 35.3/62.8, respectively, revealing substantial remaining gaps in discovery, verification, and evidence binding. We next analyze performance by method family.

RAG methods. RAG pipelines perform poorly overall, especially on profiling (Standard RAG: 18.4 F1 / 26.7 Acc; Self-RAG: 15.2 F1 / 10.0 Acc). Retrieval is brittle and frequently returns shallow or contextually irrelevant files, while generation lacks robust cross-file aggregation. This is most evident on Profile

(c) Victoria, where Self-RAG fails to produce any correct profiling answer (0.0 Acc). Factual retention improves only modestly (30.0–31.9 overall F1), largely in direct-lookup-like cases, but models still often overfit to filenames or directory strings rather than grounding in file content.

Search agent methods. Iterative search agents improve factual retention through multi-step exploration and file inspection, reaching up to 55.3 Acc on Profile (b) Adam with ReAct (Gemini-2.5-flash). Search-R1 obtains strong factual-retention F1 in the document-heavy Adam profile (58.0), suggesting benefits in dense document environments. However, these gains do not transfer to profiling: Search-R1 reaches only 10.8 overall profiling F1 with 5.0 Acc and 0.0 Acc on Profiles (a) and (c), showing that locating candidate files alone is insufficient for longitudinal user-level inference.

Autonomous agent systems. Autonomous agents yield moderate gains over search-based agents but remain far below the human solvability reference. ChatGPT Agent Mode performs best among evaluated non-human systems, reaching 55.0% profiling accuracy on Profiles (b) Adam and (c) Victoria and the highest overall scores among evaluated methods (profiling: 21.0 F1 / 48.3 Acc; factual retention: 35.3 F1 / 62.8 Acc). Despite these gains, remaining errors include imprecise file localization, inconsistent metadata interpretation, hallucinated file references, and unsupported metadata, indicating brittle grounding under constrained access. The strongest hosted setting is also computationally expensive and operationally unstable in long-horizon cases, often requiring 10–15 minutes per query and sometimes producing incomplete outputs or missing file references that trigger re-execution. Detailed retry handling, budget limits, and Claude-style interface boundary cases are provided in Supplementary Secs. D.2–D.3.

5 Analysis

Our main results (Tab. 2) and capability decomposition (Tab. 3) show that failures arise not only from evidence retrieval, but also from post-retrieval discrimination, grounding, integration, and verification under profile-local, cross-modal, and temporally extended conditions. We first localize these gaps through capability-wise metrics (Sec. 5.1), then examine representative failure modes (Sec. 5.2) and their implications for file-system agent design (Sec. 5.3).

5.1 Capability-wise Decomposition

Profiling requires a different capability composition than factual retention. Across methods, profiling accuracy is systematically lower than factual retention, and the gap is largest for search-centric methods. For example, Search-R1 drops from 25.3% factual accuracy to 5.0% on profiling, while ChatGPT Agent Mode narrows but does not close the gap (62.8% vs. 48.3%). Factual retention rewards localized, file-grounded lookup; profiling requires weak-signal aggregation across temporally extended traces, referent disambiguation among co-occurring entities, and longitudinal abstraction into stable user-level patterns.

Table 3: Agent capability-wise analysis on HippoCamp. For the methods in Tab. 2, we report F1 and LLM-judge accuracy (Acc) aggregated by agent capability labels, decomposed into *search*, *perception*, and *reasoning*, for *profiling* and *factual retention* as well as the overall average. Values are percentages (one decimal; % omitted). Best is highlighted; second-best is underlined.

Method	Profiling								Factual Retention							
	Search		Perception		Reasoning		Overall		Search		Perception		Reasoning		Overall	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
RAG Methods																
Standard RAG [21]	26.4	26.2	21.8	13.8	26.7	25.5	25.0	21.8	19.0	26.2	21.1	28.7	13.0	19.1	17.7	24.7
Self RAG [2]	27.8	23.1	23.1	13.2	27.7	22.4	26.2	19.6	15.2	8.9	16.4	12.2	10.7	7.0	14.1	9.4
Search Agent Methods																
ReAct [57] (Qwen3-30B-A3B [41])	**36.3**	26.1	27.9	16.7	**36.3**	23.9	33.5	22.2	11.9	13.4	14.5	18.7	8.6	10.6	11.7	14.2
ReAct [57] (Gemini-2.5-flash [7])	23.5	34.9	20.5	20.4	23.4	33.0	22.5	29.4	19.4	19.1	**21.3**	23.4	13.2	14.8	18.0	19.1
Search-R1 [18]	34.8	24.9	**32.8**	15.7	35.1	25.8	**34.2**	22.1	10.2	3.8	12.3	7.2	7.8	3.9	10.1	5.0
Autonomous Agent Systems																
Terminal Agent (Qwen3-VL-8B-Instruct [3])	15.8	11.1	14.6	5.7	16.9	11.4	15.8	9.4	12.2	18.6	14.0	24.2	8.1	12.2	11.4	18.3
Terminal Agent (Gemini-2.5-flash [7])	24.5	21.0	26.9	13.7	25.2	21.1	25.5	18.6	14.9	23.3	14.4	24.7	12.2	17.2	13.8	21.7
Terminal Agent (GPT-5.2 [29])	23.6	46.3	15.4	27.3	22.6	44.1	20.5	39.2	10.0	27.2	10.2	29.7	15.5	**36.4**	11.9	31.1
ChatGPT Agent Mode [28]	28.9	**56.5**	15.1	**28.5**	30.2	**55.8**	24.7	**46.9**	**19.5**	**49.1**	17.3	**55.5**	**30.4**	33.8	**22.4**	**46.1**

Methods built around retrieval strength alone therefore lack the full capability composition required for user-level inference.

Profile difficulty is more consistent with structural explicitness and entity ambiguity than with domain label alone. Performance varies consistently across profiles. Adam's document-centric legal environment yields the highest scores, with ChatGPT Agent Mode reaching 90.3% factual accuracy and 55.0% profiling accuracy, whereas Bei's socially entangled, media-rich college environment is hardest (31.2% factual and 35.0% profiling for the same system). Victoria falls between these extremes. This pattern is more consistent with structural explicitness, role boundaries, and entity ambiguity than with domain label alone: informal, multi-entity environments amplify referential ambiguity and weaken the signal-to-noise ratio of individual traces.

F1 and accuracy decouple in two directions. Tabs. 2 and 3 show that retrieval quality and answer quality are separable. When F1 exceeds accuracy, as with ReAct (Qwen3) on factual retention (43.1% F1 vs. 28.5% Acc), the method retrieves relevant files but fails to convert them into correct answers, indicating a breakdown in evidence discrimination or synthesis. When accuracy exceeds F1, as with Terminal Agent (GPT-5.2) on factual retention (24.6% F1 vs. 48.2% Acc), correct outputs may occur without retrieving the annotated support files, suggesting incomplete grounding or reliance on parametric knowledge. Capability-level scores reinforce this pattern: search is necessary but not decisive, perception is the most consistent bottleneck, and reasoning quality depends on prior evidence selection.

To localize where in the agent pipeline these gaps originate, we turn to the capability-wise decomposition in Tab. 3.

Search is necessary but not decisive. Search-centric agents achieve the highest retrieval F1 on profiling: ReAct (Qwen3) reaches 36.3% and Search-R1 reaches 34.8%. Their overall profiling accuracy, however, lags far behind—22.2%

and 22.1%, respectively. ChatGPT Agent Mode inverts this pattern: despite a lower search F1 of 28.9%, it achieves the highest search accuracy (56.5%) and the strongest capability-wise overall profiling accuracy (46.9%). Locating candidate files is a prerequisite, but remains insufficient for end-task correctness.

Perception is the most consistent bottleneck. Across all method families, perception accuracy on profiling is uniformly low, ranging from 13.2% (Self-RAG) to 28.5% (ChatGPT Agent Mode). Even the best-performing evaluated system achieves perception accuracy roughly half of its search accuracy (28.5% vs. 56.5%). This gap reflects more than OCR or vision failures. It captures the broader difficulty of converting heterogeneous file content—PDFs, calendar entries, photos, voice memos—into evidence units that can support downstream reasoning.

Reasoning quality is contingent on prior evidence discrimination. Reasoning is not an independent capability; it inherits errors from earlier stages. Terminal Agent (GPT-5.2) achieves 44.1% profiling reasoning accuracy despite only 27.3% perception accuracy, raising the possibility that some correct answers stem from parametric knowledge rather than grounded evidence—consistent with the Acc-exceeds-F1 pattern observed in Tab. 2. Search-R1 shows the converse: strong retrieval signals (35.1% reasoning F1) but weak answer commitment (25.8% accuracy). Among evaluated systems, ChatGPT Agent Mode shows the comparatively strongest and most balanced capability-wise accuracy across search, perception, and reasoning, consistent with its iterative exploration advantage, although all three axes remain far below human-level reliability. Strong reasoning cannot compensate for weak evidence selection, and high retrieval F1 does not guarantee that the evidence reaching the reasoning stage is correct.

5.2 Systematic Failure Analysis

The capability gaps identified above manifest as concrete error patterns in agent outputs. Fig. 5 illustrates these patterns on a representative profiling query whose ground truth requires cross-modal synthesis over calendar events, running logs, photos, voice memos, text notes, and emails. We identify four primary failure modes. *Retrieval mismatch* occurs when a method retrieves semantically related but profile-irrelevant files, as in Standard RAG. *Grounding avoidance* occurs when a method surfaces partially relevant candidates but answers generically instead of committing to file-grounded evidence, as in Search-R1. *Hard evidence hallucination* occurs when a method fabricates file paths, metadata, or support chains, as in Terminal Agent (GPT-5.2). *Entity misattribution* occurs when a method finds real evidence but binds it to the wrong profile-local referent, as in ChatGPT Agent Mode. Complementarily, the main success pattern is iterative discovery: ChatGPT Agent Mode repeatedly explores directories, reads candidate files, and revises hypotheses, which helps explain its stronger behavior in structured profiles but remains insufficient for weak, distributed, cross-modal profiling evidence.

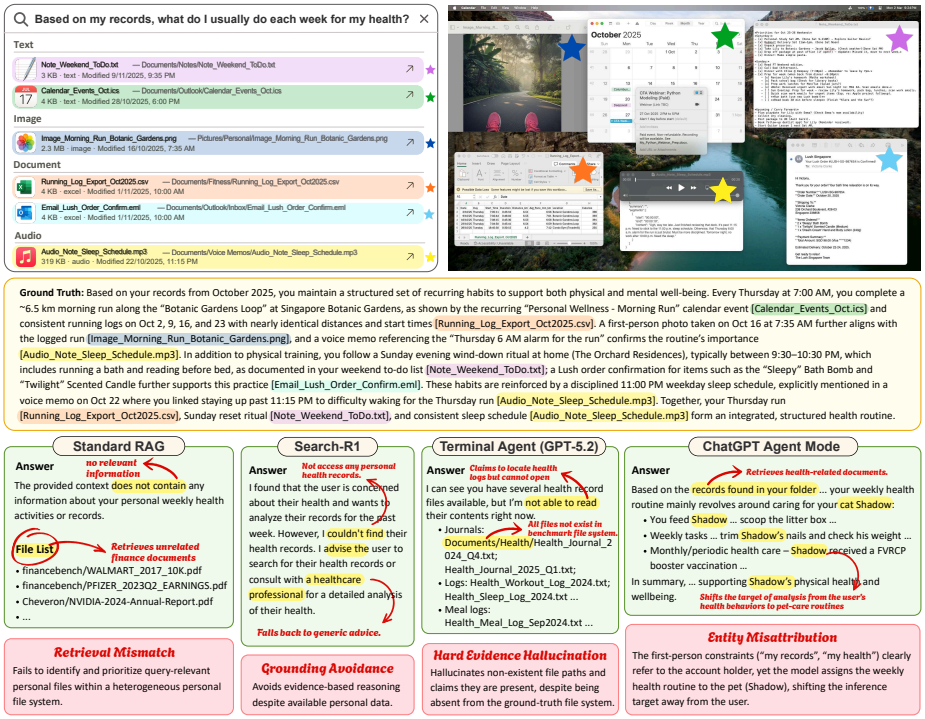

Fig. 5: Representative failure and success patterns on a cross-modal profiling query. The query requires aligning evidence across heterogeneous personal files and modalities. Standard RAG exhibits *retrieval mismatch* (retrieving unrelated finance documents), Search-R1 shows *grounding avoidance* (generic advice despite available evidence), Terminal Agent (GPT-5.2) triggers *hard evidence hallucination* (fabricated file paths/metadata), and ChatGPT Agent Mode makes an *entity misattribution* error (shifting the referent from the user to the pet), while iterative exploration yields the strongest overall behavior among evaluated methods.

These failures indicate task-dependent bottlenecks rather than a single uniform weakness: RAG systems often fail at early evidence discovery, search agents often fail at evidence commitment, terminal agents are vulnerable to fabricated support, and hosted autonomous agents benefit from iterative exploration but still suffer from entity binding and verification errors. Across methods, the common missing component is an explicit final-stage check that binds the answer to a minimal, coherent evidence set. Supplementary Sec. D.6 provides the complete case-level interpretation, audited primary-label failure coding, and stratification by modality breadth and file-hierarchy depth.

Beyond the representative case, the audited failure labels show that retrieval mismatch becomes increasingly concentrated as evidence spans more modalities and deeper folders: among failed outputs, its share rises from 40.7% in single-modality cases to 87.0% when three or more modalities are required, and from

42.6% at one-folder depth to 92.6% at depth three or more. This indicates that multimodal and deeply nested evidence often fails before downstream grounding or synthesis can begin.

5.3 Implications for File-System Agents

HippoCamp indicates that file-based personalization is not reducible to retrieving explicit, static persona descriptions. Instead, agents must infer user-level context from implicit behavioral signals distributed across heterogeneous files, temporal traces, and profile-local entities, while binding each inference to concrete evidence. The observed failures suggest four operational requirements. **Structure-aware search** should use folder hierarchy, temporal metadata, attachment relations, and version history as search priors, since flat semantic retrieval often confuses topically similar but profile-irrelevant files. **Evidence narrowing** should precede answer generation: the gap between retrieval F1 and answer accuracy in Tabs. 2 and 3 shows that broad candidate pools must be pruned into a minimal sufficient support set before synthesis. **Profile-local entity modeling** is required because personal file systems contain multiple co-occurring referents; the entity-misattribution errors in Fig. 5 show that correct retrieval can still fail when the account holder is confused with family members, pets, or colleagues. **Explicit verification** should re-bind final answers to localized file evidence and check for contradictions, closing the gap between finding plausible evidence and producing a traceable, coherent response.

6 Conclusion

This work introduces HippoCamp, a benchmark evaluating agents' ability to search, perceive, and reason over realistic, multimodal personal file systems. Our analysis shows that the bottlenecks are task-dependent: profiling is primarily constrained by hierarchical and temporal discovery of weak, distributed evidence, whereas factual retention additionally exposes post-retrieval failures in evidence discrimination, extraction, cross-file synthesis, and verification. Across both task families, reliable personalized memory requires grounding retrieved evidence to concrete files, resolving profile-local entities, and verifying final answers against a coherent support set. These results show that personalized multimodal memory remains difficult even for current agentic systems. HippoCamp provides a benchmark for measuring these failures and comparing future file-system agents under controlled, profile-local conditions.

Acknowledgment. This research is supported by the Ministry of Education, Singapore, through MOE AcRF Tier 2 grants MOE-T2EP20223-0002 and MOE-T2EP20224-0003. This work is also supported by cash and in-kind funding from NTU S-Lab and industry partner(s), with additional support from Synvo AI.

References

1. Anthropic: Claude Sonnet 4.5 System Card (Sep 2025), https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd/Claude%20Sonnet%204.5%20System%20Card.pdf, accessed 27 Jun 2026
2. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-rag: Learning to retrieve, generate, and critique through self-reflection (2023), https://arxiv.org/abs/2310.11511, accessed 27 Jun 2026
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631, accessed 27 Jun 2026
4. Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., Bisk, Y.: Webqa: Multihop and multimodal qa (2022), https://arxiv.org/abs/2109.00590, accessed 27 Jun 2026
5. Chen, C., Guan, M., Lin, X., Li, J., Lin, L., Wang, Q., Chen, X., Luo, J., Sun, C., Zhang, D., Li, X.: Telemem: Building long-term and multimodal memory for agentic ai (2026), https://arxiv.org/abs/2601.06037, accessed 27 Jun 2026
6. Cho, J., Mahata, D., Irsoy, O., He, Y., Bansal, M.: M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding (2024), https://arxiv.org/abs/2411.04952, accessed 27 Jun 2026
7. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261, accessed 27 Jun 2026
8. DeepSeek-AI, Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., et al.: DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models (2025), https://arxiv.org/abs/2512.02556, accessed 27 Jun 2026
9. Dong, K., Chang, Y., Goh, X.D., Li, D., Tang, R., Liu, Y.: Mmdocir: Benchmarking multimodal retrieval for long documents (2025), https://arxiv.org/abs/2501.08828, accessed 27 Jun 2026
10. Dong, K., Chang, Y., Huang, S., Wang, Y., Tang, R., Liu, Y.: Benchmarking retrieval-augmented multimodal generation for document question answering (2025), https://arxiv.org/abs/2505.16470, accessed 27 Jun 2026
11. Friel, R., Belyi, M., Sanyal, A.: Ragbench: Explainable benchmark for retrieval-augmented generation systems (2025), https://arxiv.org/abs/2407.11005, accessed 27 Jun 2026
12. Fu, D., He, K., Wang, Y., Hong, W., Gongque, Z., Zeng, W., Wang, W., Wang, J., Cai, X., Xu, W.: Agentrefine: Enhancing agent generalization through refinement tuning (2025), https://arxiv.org/abs/2501.01702, accessed 27 Jun 2026
13. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J.,

Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature **645**(8081), 633–638 (September 2025). https://doi.org/10.1038/s41586-025-09422-z, http://dx.doi.org/10.1038/s41586-025-09422-z, accessed 27 Jun 2026

14. Guo, Z., Ren, X., Xu, L., Zhang, J., Huang, C.: Rag-anything: All-in-one rag framework (2025), https://arxiv.org/abs/2510.12323, accessed 27 Jun 2026
15. Gur, I., Furuta, H., Huang, A., Safdari, M., Matsuo, Y., Eck, D., Faust, A.: A real-world webagent with planning, long context understanding, and program synthesis (2024), https://arxiv.org/abs/2307.12856, accessed 27 Jun 2026
16. Huang, Y., Shi, J., Li, Y., Fan, C., Wu, S., Zhang, Q., Liu, Y., Zhou, P., Wan, Y., Gong, N.Z., Sun, L.: Metatool benchmark for large language models: Deciding whether to use tools and which to use (2024), https://arxiv.org/abs/2310.03128, accessed 27 Jun 2026
17. Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., Vidgen, B.: Financebench: A new benchmark for financial question answering (2023), https://arxiv.org/abs/2311.11944, accessed 27 Jun 2026
18. Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., Han, J.: Search-r1: Training llms to reason and leverage search engines with reinforcement learning (2025), https://arxiv.org/abs/2503.09516, accessed 27 Jun 2026
19. Koh, J.Y., Lo, R., Jang, L., Duvvur, V., Lim, M.C., Huang, P.Y., Neubig, G., Zhou, S., Salakhutdinov, R., Fried, D.: Visualwebarena: Evaluating multimodal agents on realistic visual web tasks (2024), https://arxiv.org/abs/2401.13649, accessed 27 Jun 2026
20. Lee, C.P., Choi, J., Mutlu, B.: Map: Multi-user personalization with collaborative llm-powered agents. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. CHI EA '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3706599.3719853, https://doi.org/10.1145/3706599.3719853, accessed 27 Jun 2026
21. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021), https://arxiv.org/abs/2005.11401, accessed 27 Jun 2026

22. Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., Liu, H.: From generation to judgment: Opportunities and challenges of llm-as-a-judge (2025), https://arxiv.org/abs/2411.16594, accessed 27 Jun 2026
23. Li, Y., Wen, H., Wang, W., Li, X., Yuan, Y., Liu, G., Liu, J., Xu, W., Wang, X., Sun, Y., Kong, R., Wang, Y., Geng, H., Luan, J., Jin, X., Ye, Z., Xiong, G., Zhang, F., Li, X., Xu, M., Li, Z., Li, P., Liu, Y., Zhang, Y.Q., Liu, Y.: Personal llm agents: Insights and survey about the capability, efficiency and security (2024), https://arxiv.org/abs/2401.05459, accessed 27 Jun 2026
24. Liu, J., Zhu, D., Bai, Z., He, Y., Liao, H., Que, H., Wang, Z., Zhang, C., Zhang, G., Zhang, J., Zhang, Y., Chen, Z., Guo, H., Li, S., Liu, Z., Shan, Y., Song, Y., Tian, J., Wu, W., Zhou, Z., Zhu, R., Feng, J., Gao, Y., He, S., Li, Z., Liu, T., Meng, F., Su, W., Tan, Y., Wang, Z., Yang, J., Ye, W., Zheng, B., Zhou, W., Huang, W., Li, S., Zhang, Z.: A comprehensive survey on long context language modeling (2025), https://arxiv.org/abs/2503.17407, accessed 27 Jun 2026
25. Luo, H., Dai, S., Ni, C., Li, X., Zhang, G., Wang, K., Liu, T., Salam, H.: Agentauditor: Human-level safety and security evaluation for llm agents (2026), https://arxiv.org/abs/2506.00641, accessed 27 Jun 2026
26. Maharana, A., Lee, D.H., Tulyakov, S., Bansal, M., Barbieri, F., Fang, Y.: Evaluating very long-term conversational memory of llm agents (2024), https://arxiv.org/abs/2402.17753, accessed 27 Jun 2026
27. Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., Scialom, T.: Gaia: a benchmark for general ai assistants (2023), https://arxiv.org/abs/2311.12983, accessed 27 Jun 2026
28. OpenAI: ChatGPT Agent System Card (Jul 2025), https://cdn.openai.com/pdf/839e66fc-602c-48bf-81d3-b21eacc3459d/chatgpt_agent_system_card.pdf, accessed 27 Jun 2026
29. OpenAI: GPT-5 System Card (Aug 2025), https://cdn.openai.com/gpt-5-system-card.pdf, accessed 27 Jun 2026
30. OpenAI, Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., et al.: GPT-4o System Card (2024), https://arxiv.org/abs/2410.21276, accessed 27 Jun 2026
31. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., Shi, J., Wu, F., Chu, P., Liu, M., Li, Z., Xu, C., Zhang, B., Shi, B., Tu, Z., He, C.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations (2025), https://arxiv.org/abs/2412.07626, accessed 27 Jun 2026
32. Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., Cao, N.D., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., Riedel, S.: Kilt: a benchmark for knowledge intensive language tasks (2021), https://arxiv.org/abs/2009.02252, accessed 27 Jun 2026
33. Pipitone, N., Alami, G.H.: Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain (2024), https://arxiv.org/abs/2408.10343, accessed 27 Jun 2026
34. Ren, X., Xu, L., Xia, L., Wang, S., Yin, D., Huang, C.: Videorag: Retrieval-augmented generation with extreme long-context videos (2025), https://arxiv.org/abs/2502.01549, accessed 27 Jun 2026
35. Sadhu, T., Chen, Y., Pesaranghader, A.: VestaBench: An embodied benchmark for safe long-horizon planning under multi-constraint and adversarial settings. In: Potdar, S., Rojas-Barahona, L., Montella, S. (eds.) Proceedings of the 2025

Conference on Empirical Methods in Natural Language Processing: Industry Track. pp. 2122–2145. Association for Computational Linguistics, Suzhou (China) (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-industry.149, https://aclanthology.org/2025.emnlp-industry.149/, accessed 27 Jun 2026
36. Song, Y., Thai, K., Pham, C.M., Chang, Y., Nadaf, M., Iyyer, M.: Bearcubs: A benchmark for computer-using web agents (2025), https://arxiv.org/abs/2503.07919, accessed 27 Jun 2026
37. Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J.S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W., Heidecke, J., Glaese, A., Patwardhan, T.: Paperbench: Evaluating ai's ability to replicate ai research (2025), https://arxiv.org/abs/2504.01848, accessed 27 Jun 2026
38. Sumers, T.R., Yao, S., Narasimhan, K., Griffiths, T.L.: Cognitive architectures for language agents (2024), https://arxiv.org/abs/2309.02427, accessed 27 Jun 2026
39. Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Ilharco, G., Hajishirzi, H., Berant, J.: Multimodalqa: Complex question answering over text, tables and images (2021), https://arxiv.org/abs/2104.06039, accessed 27 Jun 2026
40. Tang, Y., Yang, Y.: Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries (2024), https://arxiv.org/abs/2401.15391, accessed 27 Jun 2026
41. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388, accessed 27 Jun 2026
42. Tian, S., Wang, R., Guo, H., Wu, P., Dong, Y., Wang, X., Yang, J., Zhang, H., Zhu, H., Liu, Z.: Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning (2025), https://arxiv.org/abs/2506.13654, accessed 27 Jun 2026
43. Vianna, D., Kalokyri, V., Borgida, A., Nguyen, T.D., Marian, A.: Searching heterogeneous personal digital traces (2019), https://arxiv.org/abs/1904.05374, accessed 27 Jun 2026
44. Wang, X., Wang, Z., Liu, J., Chen, Y., Yuan, L., Peng, H., Ji, H.: Mint: Evaluating llms in multi-turn interaction with tools and language feedback (2024), https://arxiv.org/abs/2309.10691, accessed 27 Jun 2026
45. Wang, Z., Li, Z., Jiang, Z., Tu, D., Shi, W.: Crafting personalized agents through retrieval-augmented generation on editable memory graphs (2024), https://arxiv.org/abs/2409.19401, accessed 27 Jun 2026
46. Wang, Z., Xu, H., Wang, J., Zhang, X., Yan, M., Zhang, J., Huang, F., Ji, H.: Mobile-agent-e: Self-evolving mobile assistant for complex tasks (2025), https://arxiv.org/abs/2501.11733, accessed 27 Jun 2026
47. Wei, J., Sun, Z., Papay, S., McKinney, S., Han, J., Fulford, I., Chung, H.W., Passos, A.T., Fedus, W., Glaese, A.: Browsecomp: A simple yet challenging benchmark for browsing agents (2025), https://arxiv.org/abs/2504.12516, accessed 27 Jun 2026
48. Wu, J., Deng, Z., Li, W., Liu, Y., You, B., Li, B., Ma, Z., Liu, Z.: Mmsearch-r1: Incentivizing lmms to search (2025), https://arxiv.org/abs/2506.20670, accessed 27 Jun 2026
49. Wu, Z., Han, C., Ding, Z., Weng, Z., Liu, Z., Yao, S., Yu, T., Kong, L.: Os-copilot: Towards generalist computer agents with self-improvement (2024), https://arxiv.org/abs/2402.07456, accessed 27 Jun 2026
50. Xi, Y., Lin, J., Zhu, M., Xiao, Y., Ou, Z., Liu, J., Wan, T., Chen, B., Liu, W., Wang, Y., Tang, R., Zhang, W., Yu, Y.: Infodeepseek: Benchmarking agentic information

seeking for retrieval-augmented generation (2025), https://arxiv.org/abs/2505.15872, accessed 27 Jun 2026
51. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., Gui, T.: The rise and potential of large language model based agents: A survey (2023), https://arxiv.org/abs/2309.07864, accessed 27 Jun 2026
52. Xiu, Z., Sun, D.Q., Cheng, K., Patel, M., Date, J., Zhang, Y., Lu, J., Attia, O., Vemulapalli, R., Tuzel, O., Cao, M., Bengio, S.: Astra-bench: Evaluating tool-use agent reasoning and action planning with personal user context (2026), https://arxiv.org/abs/2603.01357, accessed 27 Jun 2026
53. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., Ouyang, B., Lin, Z., Cominelli, M., Cai, Z., Zhang, Y., Zhang, P., Hong, F., Widmer, J., Gringoli, F., Yang, L., Li, B., Liu, Z.: Egolife: Towards egocentric life assistant (2026), https://arxiv.org/abs/2503.03803, accessed 27 Jun 2026
54. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D.: Hotpotqa: A dataset for diverse, explainable multi-hop question answering (2018), https://arxiv.org/abs/1809.09600, accessed 27 Jun 2026
55. Yao, H., Zhang, R., Huang, J., Zhang, J., Wang, Y., Fang, B., Zhu, R., Jing, Y., Liu, S., Li, G., Tao, D.: A survey on agentic multimodal large language models (2025), https://arxiv.org/abs/2510.10991, accessed 27 Jun 2026
56. Yao, S., Chen, H., Yang, J., Narasimhan, K.: Webshop: Towards scalable real-world web interaction with grounded language agents (2023), https://arxiv.org/abs/2207.01206, accessed 27 Jun 2026
57. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models (2023), https://arxiv.org/abs/2210.03629, accessed 27 Jun 2026
58. Yu, H., Shen, B., Ran, D., Zhang, J., Zhang, Q., Ma, Y., Liang, G., Li, Y., Wang, Q., Xie, T.: Codereval: A benchmark of pragmatic code generation with generative pre-trained models. In: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. ICSE '24, Association for Computing Machinery, New York, NY, USA (2024), https://doi.org/10.1145/3597503.3623316, accessed 27 Jun 2026
59. Zhan, Z., Wang, J., Zhou, S., Deng, J., Zhang, R.: Mmrag: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning (2025), https://arxiv.org/abs/2502.15954, accessed 27 Jun 2026
60. Zhang, G., Fu, M., Wan, G., Yu, M., Wang, K., Yan, S.: G-memory: Tracing hierarchical memory for multi-agent systems (2025), https://arxiv.org/abs/2506.07398, accessed 27 Jun 2026
61. Zhang, J., Lan, T., Murthy, R., Liu, Z., Yao, W., Zhu, M., Tan, J., Hoang, T., Liu, Z., Yang, L., Feng, Y., Kokane, S., Awalgaonkar, T., Niebles, J.C., Savarese, S., Heinecke, S., Wang, H., Xiong, C.: Agentohana: Design unified data and training pipeline for effective agent learning (2024), https://arxiv.org/abs/2402.15506, accessed 27 Jun 2026
62. Zhang, W., Zhang, X., Zhang, C., Yang, L., Shang, J., Wei, Z., Zou, H.P., Huang, Z., Wang, Z., Gao, Y., Pan, X., Xiong, L., Liu, J., Yu, P.S., Li, X.: Personaagent: Bridging memory and action for personalized llm agents (2026), https://arxiv.org/abs/2506.06254, accessed 27 Jun 2026

63. Zhao, A., Huang, D., Xu, Q., Lin, M., Liu, Y.J., Huang, G.: Expel: Llm agents are experiential learners (2024), https://arxiv.org/abs/2308.10144, accessed 27 Jun 2026
64. Zheng, B., Fatemi, M.Y., Jin, X., Wang, Z.Z., Gandhi, A., Song, Y., Gu, Y., Srinivasa, J., Liu, G., Neubig, G., Su, Y.: Skillweaver: Web agents can self-improve by discovering and honing skills (2025), https://arxiv.org/abs/2504.07079, accessed 27 Jun 2026