Toward an AI Personalization Index:

A 157-Day Single-User Case Study

Taekyung Lee
Independent Researcher, Republic of Korea
dfdfgo92@gmail.com

Abstract

AI model benchmark performance has reached a saturation point. According to the Stanford 2025 AI Index Report, the performance gap among the top 10 models has narrowed to 5.4%, and the U.S.–China gap stands at only 1.70%. Yet the EY 2025 Work Reimagined Survey (29 countries, 15,000+ respondents) reveals that while 88% of employees use AI, only 5% engage in advanced utilization. This study proposes that the absence of personalization may be an important explanatory variable for this discrepancy.

The author conducted an exploratory single-user manual personalization study over 157 days (~1,000 hours) on the Claude platform (Max plan) from October 2025 to March 2026. During the study period, the underlying model was upgraded three times (Sonnet 4.5 → Opus 4.5 → Opus 4.6), creating a confound that prevents complete separation of personalization effects from model improvement effects. As the primary evidence, a same-time comparison on the identical model (Opus 4.6) between an established session (30 memory slots) and a new session (0 memory slots) revealed marked differences: context re-explanation frequency (0–1 vs. 5–7 per session), correction requests (2–5 vs. 20–25 per session), and time to first meaningful output (~5 min vs. ~30 min). Based on these observations, this study proposes a conceptual organizing framework—Experienced Performance ∝ Model Performance × Personalization Index—and preliminarily defines four candidate variables (memory depth, conversation volume, context accuracy, and preference learning rate). This study suggests that the next competitive frontier for AI platforms may shift from raw performance toward personalization. It outlines five platform-level automation measures and proposes five baseline data sovereignty principles.

Keywords: AI personalization, experienced performance, benchmark saturation, personalization index, user experience, human–computer interaction

1. Introduction

The performance of artificial intelligence (AI) models has exhibited rapid convergence between 2024 and 2025. The Stanford 2025 AI Index Report noted that "AI systems have scored so highly that benchmarks are no longer useful" [1]. On the Chatbot Arena Leaderboard, the Elo score gap between the 1st- and 10th-ranked models narrowed from 11.9% to 5.4%. Benchmark saturation has been confirmed across all domains, including general knowledge, image reasoning, mathematics, and coding.

NYU professor Gary Marcus observed that "2025 models got better on benchmarks, but I have barely heard any company say they are actually more useful" [10]. Bill Gates similarly noted in 2023 that "scalable AI has reached a plateau" [18].

Simultaneously, a severe utilization gap exists on the user side. The EY 2025 Work Reimagined Survey (29 countries; 15,000 employees + 1,500 employers) found that while 88% use AI, only 5% engage in advanced utilization that transforms their work [2]. OpenAI's own enterprise report confirmed a 6× productivity gap between AI power users and average employees [3,4]. Deloitte's 2025 report found that organizations adopting a "technology-first approach" were 1.6× more likely to fail to meet AI expectations [6].

User psychological responses are also worth examining. In the same EY survey, 37% of employees worried that excessive AI reliance could erode their skills and expertise, while 64% felt their workload had actually increased over the past year [2]. A substantial proportion of users thus perceive AI not as a workload-reducing tool but as an additional burden. As OpenAI's report summarized, "model capabilities are far ahead of what most organizations have embedded in their workflows" [3].

These two phenomena—benchmark saturation and user utilization stagnation—may share a common structural origin. Benchmark saturation can be understood as reduced brand-based differentiation under benchmark convergence: just as blinding brand labels equalize evaluation in peer review, the disappearance of meaningful performance gaps shifts the differentiation criterion from 'who made it' to 'how well it knows me.'

Along the same lines, in an environment where AI has become ubiquitous, the ability to filter relevant information may become more important than the ability to generate it. The phenomenon of 88% usage but only 5% advanced utilization can also be interpreted as a filtering capability gap. Personalization corresponds to a structure in which AI assists this filtering—automatically selecting what the user needs from accumulated context.

Multiple hypotheses could explain this gap: insufficient user training, organizational resistance, UX immaturity, and poor task-fit, among others. These factors may indeed contribute partially. However, industry data suggest that they alone do not fully account for the gap. If training were the primary cause, usage rates themselves should be low—yet 88% already use AI [2]. If organizational resistance were primary, technology-forward organizations should succeed—yet Deloitte found that 'technology-first approaches' fail 1.6× more often [6]. If UX were primary, continued improvements across ChatGPT, Claude, and Gemini should have narrowed the gap—yet the 5% advanced utilization rate remains stagnant. If task-fit were primary, model capabilities should be insufficient—yet OpenAI confirmed that 'model capabilities far exceed what organizations have deployed' [3].

These observations do not exclude existing hypotheses; they instead suggest the need for an additional core explanatory variable. This study proposes that variable to be the system's lack of persistent user-specific context—that is, the absence of personalization. Personalization is distinguished from other delivery-structure improvements (prompt templates, agent automation, workflow integration) by three characteristics: (1) It is cumulative—UX improvements occur only at update points, while personalization accumulates with each conversation. (2) It is individual—prompt templates are generic, while personalization adapts to each user. (3) It is self-reinforcing—agent automation handles tasks but does not know the user, while personalization forms a self-reinforcing loop that grows more accurate with use. Li et al. [21]'s long-term dialogue agent and Zhong et al. [20]'s MemoryBank provide empirical support for this cumulative adaptation mechanism.

Model performance is sufficient, but the structure for delivering that performance to users is absent. This study defines the core of that delivery structure as 'Personalization' and poses the research question: "In an era where performance differentiation is no longer possible, what variable determines experienced performance?"

The contributions of this paper are: (1) a conceptual organizing framework proposing Experienced Performance ∝ Model Performance × Personalization Index; (2) preliminary definition of four candidate measurement variables; (3) exploratory empirical data from a 157-day single-user manual personalization study; (4) five practical proposals for platform-level personalization automation.

2. Related Work

2.1 Benchmark Saturation and Performance Stagnation

Benchmark performance convergence has been confirmed by multiple reports. The Stanford 2025 AI Index reported saturation across all major benchmarks including MMLU, MMMU, MATH, and HumanEval [1]. IEEE Spectrum confirmed similar stagnation patterns [9]. The gap between the top U.S. and Chinese models narrowed from 9.26% in January 2024 to 1.70% by February 2025. AI2Work [11] and HoneHQ [12] also reported an 'innovation plateau' characterized by diminishing returns on investment.

2.2 Industry Trends in AI Personalization

Industrial demand for AI personalization is surging. Adobe's 2026 report found that 80% of enterprises want AI experiences that are "highly personalized and predictive of customer needs in real time," and 60% want experiences that are "AI-driven but feel human" [5]. CX Today defined true personalization as "knowing the customer deeply

and responding to that knowledge in real time" [7]. Microsoft Advertising's Paul Longo noted that "personalization is evolving from demographic-based to individually contextualized interactions" [8].

However, most current AI 'personalization' is limited to recommendation algorithms (e.g., engagement optimization loops on social media platforms). These track click patterns but do not understand who the user is. This study distinguishes between behavior-tracking personalization and context-accumulating personalization based on user-specific understanding.

2.3 Academic Research on LLM Memory and Personalization

Academic research on long-term memory and personalization in LLMs is advancing rapidly. Packer et al. [19] proposed MemGPT, managing LLM memory in an OS-like architecture. Zhong et al. [20] designed MemoryBank, applying the Ebbinghaus forgetting curve to conversation-based memory updating. Li et al. [21] presented a personalized agent framework for long-term dialogue at NAACL 2025, and Pan et al. [22] proposed SeCom, a memory construction and retrieval framework for personalized conversational agents, at ICLR 2025.

Li et al. [23] published a comprehensive personalization survey spanning RAG to agents, and Tan et al. [24] proposed Reflective Memory Management for long-term dialogue agents at ACL 2025. Park et al. [25] demonstrated the potential for generative agents to simulate human behavior, and Shin [26] conducted a systematic review of human-centered AI and digital well-being in JMIR 2025.

These studies focus on memory architecture and retrieval, but no quantitative metric for measuring personalization level from the user's perspective has yet been proposed. This study seeks to fill that gap with the concept of a 'Personalization Index.'

2.4 Current State of AI Platform Memory Systems

In Q1 2026, Anthropic released approximately 12 major features over 12 weeks [13,14,15,16,17], including free memory access (March 2), permanent agent threads (March 17), computer use (March 24), Claude Code Auto-dream [15], and memory import from other platforms [16]. These form the technical foundation for personalization, but three issues remain: (1) memory is not embedded in the model core (retrieval layer only), (2) session continuity is 'restoration' rather than genuine 'memory,' and (3) no metric exists for measuring personalization level.

3. Proposed Framework

3.1 Experienced Performance Formula (Conceptual Model)

This study proposes the following conceptual model:

$$\textit{Experienced Performance} \propto \textit{Model Performance} \times \textit{Personalization Index} \quad \textit{(Eq. 1)}$$

This represents a proportional relationship; actual implementation may require nonlinear weight adjustments. The assumption of independent variables is a simplification; nonlinear interactions likely exist in practice.

If the personalization index is 0, experienced performance approaches 0 regardless of model capability. If it equals 1, performance is transmitted as-is. If it exceeds 1, accumulated context may produce experienced performance beyond raw capability (estimated). Eq. 1 is proposed not as a validated predictive model but as a conceptual organizing framework for structurally presenting the importance of personalization.

3.2 Four Variables of the Personalization Index

The Personalization Index is defined as f(memory depth, conversation volume, context accuracy, preference learning rate). These four variables were selected based on factors repeatedly identified in prior research and extractable from existing platform data. Memory depth is grounded in Packer et al. [19]'s MemGPT, which showed that memory hierarchy directly impacts LLM performance. Conversation volume corresponds to Zhong et al. [20]'s MemoryBank, which showed that user understanding increases with interaction frequency. Context accuracy connects to Pan et al. [22]'s SeCom, which demonstrated that retrieved context accuracy determines response quality. Preference learning rate is based on Li et al. [21]'s finding that user preference learning is a key factor in long-term dialogue satisfaction.

Excluded candidate variables include usage frequency (indirectly captured in conversation volume), domain diversity (measurement criteria not yet established), and emotional rapport (difficult to measure objectively). These require exploration in future research.

[Measurable] Memory depth: active memory slot count × average memory age (days). Directly extractable from platform servers.

[Estimated] Conversation volume: log(total tokens) + log(session count), a normalized combination. Logarithmic transformation avoids excessively large values from simple multiplication. Extractable from server logs, though qualitative differences (deep conversation vs. simple queries) are not captured.

[Subjective estimate] Context accuracy: proportion of responses the user rated as 'contextually correct.' Extractable from existing thumbs up/down feedback systems.

[Behavioral indicator] Preference learning rate: the rate at which correction requests decrease for similar tasks. Measurable via time-series analysis of correction request frequency.

3.3 Accessibility-First Principle: iPhone vs. Segway

The first-generation iPhone (2007) had modest raw performance but revolutionized the mobile industry through accessibility. Segway (2001) had superior performance but failed due to lack of accessibility. The current AI industry follows a similar trajectory—model performance approaches its peak, yet most users remain at basic query-and-answer interactions.

Microsoft Copilot is a representative case. Its engine is GPT-4, among the world's most capable, yet its structure—lacking personalization and heavy on restrictions—results in low experienced performance. This can be interpreted as illustrating the limitations of a performance-first, personalization-absent strategy.

Personalization functions as an accessibility interface between users and AI. Even when users cannot craft sophisticated prompts, AI infers intent from accumulated context. Delivering existing performance effectively takes priority over enhancing raw performance.

Extending the analogy, a general-purpose AI is a broadcasting station—transmitting the same signal to all users. A personalized AI is an antenna—receiving what this specific user needs right now. Personalization shifts AI from generic output delivery toward user-specific contextual inference. These analogies are not evidentiary claims but heuristic illustrations of an accessibility-first adoption dynamic.

4. Case Study: A 157-Day Manual Personalization Study

4.1 Study Environment

Period: October 25, 2025 – March 30, 2026 (157 days). Platform: claude.ai (Max plan). Model progression: Claude Sonnet 4.5 → Opus 4.5 (Nov. 24, 2025) → Opus 4.6 (Feb. 5, 2026; extended thinking). Estimated time investment: ~1,000 hours (weekday evenings + weekends). The author is a nuclear engineering professional; this study was conducted entirely outside working hours.

4.2 Direct Personalization Mechanisms

(1) Memory compression design (30 slots): The entire project context was compressed into 30 memory slots. Collision checking, deduplication, and version management were performed manually. Maintaining logical consistency across slots was the greatest challenge.

(2) Persistent storage code: A jsx-based system was designed to maintain data across sessions. Configuration values and state flags persist between sessions.

(3) Preference learning system: Systematic principles for document creation, review, and critique were defined and stored in memory. The AI references these principles each session, eliminating repeated instructions. This constitutes the practical basis for the preference learning rate (Variable 4).

4.3 Extended Workflows Built on Personalization

(4) 117-day consecutive observation routine: A daily observation routine initiated December 3, 2025 was executed 117 consecutive times through March 30, 2026, without a single missed day. This routine encompasses multiple sub-processes (data collection, bias correction, judgment criteria application).

(5) 19-persona system: Nineteen personas with distinct communication styles, domain expertise, and analytical perspectives were designed and operated, classified into three types: commentary (A), analysis (B), and narrative (C).

(6) Local backup: A local AI backup system was built for cloud service interruption preparedness. Persona data is stored locally via Modelfile.

Items (4)–(6) are extended workflows operating on top of personalization mechanisms (1)–(3). Without memory and session persistence, operations at this scale would have been impossible—indirectly demonstrating that personalization is a prerequisite for advanced utilization.

4.4 Quantitative Summary

Item	Value
Study period	157 days (Oct 25, 2025 – Mar 30, 2026)
Models	Sonnet 4.5 → Opus 4.5 → Opus 4.6 (extended thinking)
Est. time investment	~1,000 hours
Memory slots	30 (manually designed, compressed, collision-checked)
Personas	19 types (distinct styles, domains, perspectives)
Observation routine	117 consecutive executions (zero missed days)
Written outputs	500+ (diverse categories)
Local backup	Modelfile-based persona storage

Table 1. Quantitative summary of the 157-day manual personalization study.

5. Results

5.1 Within-User Longitudinal Comparison

The strongest evidence in this study comes from within-user longitudinal comparison: the same author, on the same platform, observing how experienced performance changed as personalization accumulated.

All counts and time estimates reported below are descriptive ranges derived from convenience-sampled sessions rather than formal task-timed experiments.

Comparing the first 30 days (Oct 25–Nov 24, 2025; 3–8 memory slots; Sonnet 4.5) with the last 30 days (Mar 1–30, 2026; 28–30 slots; Opus 4.6): (1) Context re-explanations per session: initial ~5–7 → final ~0–1. Stored memory rendered re-explanation largely unnecessary. (2) Correction requests per session: initial ~25–30 → final ~2–5. As preferences accumulated in memory, first-response fitness was observed to improve. However, improvement in the author's own prompting skill may also have contributed. (3) Single-session output: initial average 2–3 pieces/session → final 8–12. Reduced context restoration time increased effective working time. (4) New session control (no memory): Starting a new project without memory at the same time point required re-explanation and correction levels similar to the initial period, suggesting that improved experienced performance is attributable not solely to model upgrades but also to personalization accumulation.

Two supplementary observations are presented to partially disentangle model upgrade effects from personalization effects.

(5) Within-model comparison: During the Opus 4.5 period (Nov 24, 2025–Feb 4, 2026; ~10 weeks), as memory slots expanded from 12 to 25, correction requests per session decreased from approximately 20 to approximately 8 —an improvement observed within the same model with no change in raw capability.

(6) Same-time new session control: During the final 30 days, the author alternated between the established project (30 memory slots) and a new project (0 memory slots) on the same day. Results are shown in Table 3.

Metric	Established session (30 slots)	New session (0 slots)
Context re-explanations	0–1	5–7
Correction requests	2–5	20–25
Time to first meaningful output	~5 min	~30 min

Table 3. Same-time new session control (same model Opus 4.6, same user). All values are observational descriptive ranges, not timed experimental measurements.

These comparisons are observational self-reports rather than rigorous controlled experiments, and the contribution of model upgrades cannot be entirely separated. However, the within-model comparison (5) and same-time new session control (6) partially support the independent contribution of personalization effects.

5.2 Illustrative Variable Comparison

This section presents an illustrative exploratory comparison of four personalization index variables between the author (empirical user) and a hypothetical general user. General user figures are rough estimates intended to illustrate the potential range of variation rather than to establish statistical contrasts.

Variable	Tag	Author	General user (est.)	Ratio
Memory depth	[Measured]	30×75d = 2,250	2×10d = 20	112×
Conversation vol.	[Estimated]	log(50M)+log(500)=10.4	log(500K) +log(30)=7.2	1.44×
Context accuracy	[Subjective]	~95%	~40%	2.4×
Pref. learning rate	[Behavioral]	30→3/session (90%↓)	No change (reset)	—

Table 2. Four-variable personalization index comparison (author vs. estimated general user).

On the same Claude Opus 4.6 model, raw performance is 100% identical. Yet the estimated gap across the four variables ranges from 1.4× to 112×. This exploratory comparison suggests that differences in personalization level may contribute to the "6× productivity gap between power users and average employees" reported by OpenAI [3].

5.3 Implementation Feasibility Analysis

Three of the four variables could plausibly be derived from existing platform telemetry, subject to access, definition, and validation constraints. Memory slot counts and conversation tokens already reside on servers. Context accuracy could be inferred from thumbs up/down feedback data, though feedback density and granularity vary across platforms. Preference learning rate is in principle measurable through time-series analysis of correction request frequency, though standardized definitions would be needed. Implementation therefore depends less on new infrastructure than on reframing data that platforms already collect, combined with metric design and validation.

5.4 Measurement Notes

The four variables were calculated as follows. Memory depth (Variable 1) was measured directly from slot counts and creation dates visible in the platform interface. Conversation volume (Variable 2) was calculated by counting sessions directly and estimating average tokens per session from a 10-session sample. Context accuracy (Variable 3) is a subjective estimate derived from the author's binary classification (contextually correct/incorrect) of responses over one week (7 sessions). Preference learning rate (Variable 4) was derived by comparing correction requests per session between the first 7 days and the most recent 7 days.

General user figures were estimated by the author based on AI community posts and publicly available usage pattern reports, and are not based on individual user data.

Operational Definitions: A 'session' refers to continuous interaction from opening to closing a single conversation thread on claude.ai. A 'correction request' is one message in which the user explicitly requests modification, supplementation, or regeneration of an AI response. 'Contextually correct' means the AI's response correctly reflected prior conversations, memory, and project context, enabling meaningful work without re-explanation. 'Contextually incorrect' means the user had to re-explain background information or the AI contradicted prior agreements. The 10-session sample was drawn via convenience sampling across diverse task types (document

creation, review, data analysis, creative writing). The initial/final 7-day comparison windows were selected as the first week after study commencement and the most recent week.

These measurement procedures were designed for exploratory descriptive purposes rather than inferential testing. The sample sizes and comparison windows were selected pragmatically to document observable change over time, not to optimize statistical power. Task difficulty was not formally controlled across comparison periods.

5.5 Discussion of Control Variables

During the study period, the model was upgraded three times (Sonnet 4.5 → Opus 4.5 → Opus 4.6). Therefore, it is not possible to completely separate whether experienced performance improvements are attributable to personalization accumulation or model upgrades. This is a structural limitation of this study.

However, the following observations suggest an independent contribution of personalization: (1) Within the same model period (Opus 4.5), experienced performance improved as memory slots expanded; (2) Memory was not reset after model upgrades, so personalization accumulation was maintained continuously; (3) Starting a new session without memory at the same time point consistently resulted in markedly lower experienced performance. Future research should conduct controlled experiments with a fixed model version to isolate pure personalization effects.

5.6 Limitations

The main limitations of this study are: (1) As an n=1 single-user case, statistical generalization is not possible. (2) General user figures are the author's estimates, not empirical measurements. (3) Context accuracy and preference learning rate are subjective estimates; no objective measurement was conducted. (4) The experienced performance formula is a conceptual model requiring quantitative validation (e.g., A/B testing).

6. Discussion

6.1 The Argument for Personalization Before Performance

This study does not claim that performance is unnecessary. The author's empirical work operates on top of Opus 4.6's capabilities—large context windows, persistent storage, and extended thinking are all products of performance improvements. The argument concerns sequence: when performance is enhanced on top of a personalized state, experienced improvement follows. Without personalization, each session requires reconstructing the user's context from scratch, and the learning effects of prior conversations are lost.

6.2 Behavior-Tracking vs. Context-Accumulating Personalization

Many social-media recommender systems are called 'personalization,' but they are fundamentally behavior-tracking personalization. They apply the same engagement optimization algorithms to all users, tracking click patterns (behavior) without understanding who the user is (identity). The personalization proposed in this study is context-accumulating personalization, which accumulates and adapts to a user's stable preferences, work contexts, and communication styles over time. Behavior-tracking personalization tracks 'what was clicked'; context-accumulating personalization learns 'who this user is.'

6.3 Privacy Risks and Data Sovereignty

As personalization deepens, data accumulates and risks grow. To prevent personalization data from becoming a tool of control, data sovereignty principles must apply: accessing a user's personalization database without consent constitutes surveillance.

Access stratification is a realistic risk. If only paid users receive personalization while free users remain on generic service, personalization becomes a new form of digital separation. Furthermore, personalization can be repurposed as a tool of control—personalized propaganda, personalized surveillance. The social effects of personalization depend substantially on institutional safeguards and design choices.

This study proposes five principles for ethical management of personalization data: (1) Data Ownership—ownership of personalization data belongs to the user, not the platform. (2) Portability—users can transfer their personalization data to other platforms. (3) Right to Erasure—users can delete their entire personalization data at any time. (4) Consent Granularity—users control which data items are used for personalization. (5) Anti-Stratification—

a basic level of personalization is provided regardless of tier, and personalization access must not create a new digital caste.

6.4 A Five-Stage Pathway for Scaled Access

Stage 1: Individual user quality improvement—one person proves it manually (current). Stage 2: Platform automation—the system automatically constructs personalization workflows from power user patterns. Stage 3: Gap reduction—general users reach power-user-level experienced performance. Stage 4: Broad access—basic personalization regardless of tier. Stage 5: Societal expansion—covering all people alongside digital divide resolution. Technical, institutional, and economic barriers exist between each stage.

The critical transition between Stages 2 and 3 is whether system-level automation can reproduce a subset of power-user personalization workflows without requiring equivalent user effort. If this transition succeeds, the gap between power users and general users narrows structurally rather than depending on individual skill.

This five-stage pathway is a conceptual framework for scaled personalization access; detailed analysis of transition conditions exceeds the scope of this study. The Stage $4 \rightarrow 5$ transition in particular involves economic (cost structure), institutional (data sovereignty), and social (digital literacy) factors, which are left as interdisciplinary research tasks.

7. Implications for Practice

Based on this empirical study, five measures for platform-level personalization automation are proposed.

(1) Automatic personalization DB construction: A structure where personalization databases accumulate naturally through conversation without manual design.

(2) Automatic memory expansion: A system where memory accumulates automatically through conversation, with duplicates merged and outdated versions updated.

(3) Genuine session continuity: Personalization embedded in the model core, maintaining genuine 'memory' rather than mere 'restoration' after resets.

(4) Auto-dream for conversations: Applying automatic memory consolidation (contradiction removal, date normalization, duplicate merging)—currently available only in code environments—to general conversations.

(5) Preliminary index design for personalization measurement: Developing the four candidate variables into an exploratory measurement scaffold for estimating per-user personalization levels and diagnosing areas for improvement. This is not a validated KPI but a preliminary index design that must be refined through future empirical research.

8. Conclusion

AI model benchmark performance has reached a saturation point. The gap between 1st and 10th place is 5.4%, and for many end users, these margins may be difficult to detect in routine use. While 88% use AI, only 5% engage in advanced utilization. This study proposes that this discrepancy may reflect not only raw capability but also failures of delivery structure, with the absence of personalization as an important contributing factor.

This study proposed a conceptual organizing framework—Experienced Performance $\propto$ Model Performance $\times$ Personalization Index—and defined four candidate variables: memory depth, conversation volume, context accuracy, and preference learning rate. Through a 157-day exploratory single-user study, it was observed that experienced performance can differ markedly on the same model depending on personalization level.

The next competitive frontier for the AI industry may gradually shift from raw model capability toward systems that preserve and operationalize user-specific context over time. The personalization index proposed herein remains at the level of preliminary exploration. Priority directions for future work include multi-user empirical validation and privacy-compatible deployment strategies.

Acknowledgments

AI-assisted drafting, literature search, structural editing, and logical review support was provided by Anthropic Claude Opus 4.6 (Extended Thinking) and OpenAI GPT as research assistance tools. All study design, data interpretation, core claims, and final judgments were made solely by the author. No AI system was treated as an author, and no AI system independently determined the study design, claims, or conclusions.

References

[1] Stanford University HAI, "The 2025 AI Index Report: Technical Performance," April 2025.

[2] EY, "2025 Work Reimagined Survey," November 2025. (15,000 employees + 1,500 employers, 29 countries)

[3] OpenAI, "The State of Enterprise AI: 2025 Report," December 2025.

[4] VentureBeat, "OpenAI report reveals a 6× productivity gap between AI power users and everyone else," December 2025.

[5] Adobe, "AI and Digital Trends 2026," February 2026.

[6] Deloitte, "AI Expectations Report," 2025.

[7] CX Today, "CX Trends 2025 Part 3: When AI Knows You Better Than You Know Yourself," December 2025.

[8] Microsoft Advertising, "The future of AI personalization is inclusive," June 2025.

[9] IEEE Spectrum, "The State of AI 2025: 12 Eye-Opening Graphs," May 2025.

[10] Futurism, "Scientists Are Getting Seriously Worried That We've Already Hit Peak AI" (citing Gary Marcus), August 2025.

[11] AI2Work, "AI Progress Plateau in 2025: Economic and Strategic Implications," October 2025.

[12] HoneHQ, "The AI Performance Gap: Why Investment Isn't Translating Into Impact," January 2026.

[13] Anthropic, Claude Memory System, March 2026.

[14] Nagarro, "Claude Code February 2026 Update: Analysis," March 2026.

[15] HowAIWorks.ai, "Claude Code Auto-dream: Memory Management," March 2026.

[16] WindowsNews, "Anthropic's Claude Memory Import," March 2026.

[17] BuildFastWithAI, "Claude AI 2026: Models, Features, Desktop & More," March 2026.

[18] Bill Gates, Handelsblatt Interview, 2023. ("Scalable AI has reached a plateau")

[19] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "MemGPT: Towards LLMs as Operating Systems," arXiv:2310.08560, 2024.

[20] W. Zhong et al., "MemoryBank: Enhancing Large Language Models with Long-Term Memory," AAAI, 2024.

[21] H. Li, C. Yang, A. Zhang, Y. Deng, X. Wang, and T.-S. Chua, "Hello Again! LLM-powered Personalized Agent for Long-term Dialogue," NAACL 2025, pp. 5259-5276.

[22] Z. Pan, Q. Wu, H. Jiang et al., "SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents," ICLR, 2025.

[23] X. Li et al., "A Survey of Personalization: From RAG to Agent," arXiv:2504.10147, 2025.

[24] Z. Tan et al., "In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents," ACL 2025.

[25] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," UIST, 2023.

[26] Y. Shin, "Toward Human-Centered Artificial Intelligence for Users' Digital Well-Being," JMIR Human Factors, vol. 12, e69533, 2025.