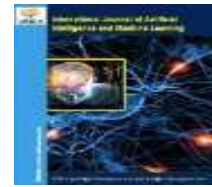




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

From Retrieval to Orchestration: Rethinking Agentic AI Integration in Enterprise HCM Platforms

Swetha Priya Sathiyam

Independent Researcher, USA. ORCID: 0009-0000-5248-8667, E-mail: swethapsathiyam@gmail.com

Abstract

Enterprise Human Capital Management platforms have integrated large language models primarily through retrieval-augmented generation, an architecture suited to information lookup but structurally unable to support goal-directed, multi-step workflow execution. This paper identifies the retrieval ceiling as a systematic limit on RAG performance in HCM and argues for an orchestration-first design in which reasoning agents coordinate tool use, memory management, and cross-system task execution across the full HCM workflow spectrum. The analysis synthesizes agentic AI research, enterprise HCM domain literature, and recent empirical studies of AI-assisted work to propose a five-layer orchestration architecture and a four-level maturity model. Comparative analysis across workforce planning, talent acquisition, and employee lifecycle management demonstrates that orchestration-first design produces qualitatively different outcomes than retrieval augmentation: decision artifacts versus content, coordinated action versus synthesized text, and compounding workflow automation versus discrete query response. The paper concludes with governance requirements for data privacy, algorithmic fairness, and audit compliance, positioning these as architectural constraints rather than post-deployment additions.

Keywords: agentic AI; enterprise HCM; orchestration architecture; retrieval-augmented generation; large language models; workflow automation; human capital management

1. Introduction

The integration of artificial intelligence into enterprise Human Capital Management has passed through a recognizable technology adoption pattern: early proof-of-concept deployments, rapid scaling of the most tractable applications, and a growing recognition that the first wave of implementations, while valuable, did not address the deeper operational challenges that motivated the investment. The first wave was retrieval. The next wave is orchestration.

Retrieval-augmented generation addressed a genuine problem. Large language models hallucinate. They confabulate facts, misrepresent policies, and fabricate citations when operating without grounding. RAG solved this by tethering model outputs to retrieved document content, giving HR professionals accurate answers to policy questions, correct benefits information, and coherent job description drafts (Lewis et al., 2020). These gains are real and have not been exhausted. Many organizations still benefit from better RAG implementations. The gains from retrieval optimization are, however, bounded by a structural constraint: retrieval cannot execute. An HR system that retrieves the correct annual review policy does not thereby complete the annual review process. An AI assistant that surfaces the relevant compensation benchmark does not thereby produce the compensation recommendation. The gap between retrieved information and executed workflow is precisely the gap that human HR professionals currently fill, and it is a large one. For example, according to Dasaklis et al. (2025), coordination and execution duties make up the larger portion of HR professionals' time at work, while information seeking and document preparation represent the smaller proportion of their activities.

Agentic AI provides the architectural substrate for closing this gap. An agent is not a better retriever; it is a different kind of system, one that maintains goal state, selects and executes tools, manages intermediate results across a reasoning chain, and adapts its plan in response to tool outputs (Yao et al., 2023; Wang et al., 2024). The distinction is categorical. A retriever searches; an agent acts. Enterprise HCM requires both, but has invested primarily in the first and underinvested in the second.

This paper makes three targeted contributions. First, it provides a structural characterization of the retrieval ceiling in HCM contexts, specifying the workflow categories where RAG performance reaches a systemic limit. Second, it proposes an orchestration-first architecture for enterprise HCM grounded in the agentic AI literature

and evaluated against specific workflow requirements. Third, it presents a four-level maturity model that gives practitioners a concrete progression path from current-state retrieval deployments to governed agentic orchestration. The paper does not argue that RAG should be replaced; it argues that RAG should be bounded to the workflow categories where it excels, and that orchestration should handle the rest.

2. Literature Review

2.1 HCM as a Domain for AI Integration

Human resource management functions sit at the intersection of organizational strategy, regulatory compliance, interpersonal dynamics, and administrative complexity. This combination creates both a highly valuable target for AI automation and highly risky scenario for AI to make mistakes. Strohmeier (2020) describes the current HR function as highly data-driven; however, the conversion of data abundance into quality decisions involves more than simply the capacity for analytics and legitimacy.

Tambe et al. (2019) provide an early systematic analysis of AI application opportunities and risks in HRM, organizing the landscape around four domains: recruiting and selection, performance management, learning and development, and workforce planning. Across all four, they identify a pattern in which AI improves decision quality as decision support but introduces bias and legitimacy risks as a decision replacement. This framing remains relevant for orchestration-first design: the architecture must preserve human decision authority at consequential junctures while automating the coordination and information integration tasks that do not require it. According to Marler & Parry (2016), there is a connection between the sophistication of HR technology and the degree of strategic HR participation, where the former implies higher levels of HR strategic contribution.

Analysis of the topic is also provided by Malik et al. (2020) in relation to the gig economy, in which the changing workforce demands a more dynamic set of HR tools. Their observation that AI-assisted HRM carries particular bias risks for underrepresented groups anticipates the governance requirements that orchestration-first architectures must address at a structural level rather than through post-deployment patches.

2.2 Foundations of Agentic AI

The theoretical foundations of agentic AI in the LLM era draw on three converging research streams. Chain-of-thought reasoning was the first step that made it possible to prove that LLMs were capable of multi-step reasoning when asked to describe their intermediary steps (Wei et al., 2022). The research was built upon the development of transformers by Vaswani et al. (2017), and it proved that the gap between one-step reasoning and multi-step reasoning could be overcome. The second stream of LLM developments is the use of tools. This concept was demonstrated by WebGPT by showing how LLMs could call outside tools and incorporate them in their answers (Nakano et al., 2021). LangChain (Chase, 2022) made use of tooling as a compositional design pattern.

The third stream is agent planning and memory. The ReAct paradigm (Yao et al., 2023) integrates reasoning and action with the model generating explicit reasoning traces interlaced with calls to tools. AutoGPT (Significant Gravitas, 2023) extended this to goal-directed autonomous operation. AutoGen (Wu et al., 2023) brought multi-agent coordination to be used as an approach to managing tasks with difficulty levels beyond the scope of single-agent systems. Park et al. (2023) found out that generative agents, which were endowed with memory and planning functionalities, could maintain consistent goal-directed behavior even in long sequences of actions. Wang et al. (2024) combine all of this literature into a taxonomy based on four factors.

2.3 The Productivity Evidence Base

The empirical evidence for AI-enabled work has mounted considerably. Noy and Zhang (2023) show in a randomized control trial that providing access to an LLM assistant boosts worker productivity by 37% in writing-heavy professional tasks, particularly those workers who have lower productivity initially. Brynjolfsson et al. (2023) reveal that access to generative AI boosts the performance of customer service workers by 14% on average. Eloundou et al. (2023) estimate that the jobs which will be most exposed to LLMs include those which are related to HR: HR specialists, compensation analysts, training coordinators, and recruiters.

An important point about this paper is that all three research works look at the use of RAG for AI support, and not agentic orchestration. The tasks studied are writing, information retrieval, and single-turn advisory interactions. Orchestration-first systems address a different category of tasks: coordination, sequential execution, cross-system integration, and workflow management. The productivity evidence base for orchestration remains underdeveloped precisely because orchestration-grade enterprise deployments are still nascent. This paper provides the architectural framework that will enable those deployments and the evidence base that follows from them.

3. The Retrieval Ceiling and an Orchestration-First Framework

3.1 Structural Analysis of the Retrieval Ceiling

The retrieval ceiling is not a performance plateau that additional optimization can raise. It is a structural boundary determined by what retrieval architectures can and cannot do. RAG systems operate on a query-response cycle: a user poses a query; the system retrieves relevant document chunks from a vector store; an LLM synthesizes the retrieved content into a response. The system state resets between queries. There is no persistent goal, no memory of prior actions, no ability to execute tools beyond retrieval itself, and no mechanism for coordinating actions across multiple systems.

Four HCM workflow categories reveal this ceiling most clearly. Complex processes like workforce planning involve retrieving information from HRIS systems, finance software, compensation data from outside sources, and business planning systems. This type of process cannot be performed by any individual retrieval process. Coordinated actions among recruitment and onboarding activities involve multiple actions to be taken within the recruitment software system, calendar system, HRIS system, LMS system, and payroll system. Each of these actions is based on the output of the previous action. Persistent goal management in annual review cycles or succession planning unfolds over weeks and requires an agent to maintain task state across sessions, resume interrupted workflows, and track progress against completion criteria. Policy-governed adaptive execution in compliance-sensitive workflows requires applying rule sets to specific factual situations and routing exceptions to human reviewers, which is rule-application-to-facts rather than document retrieval.

In both categories, the constraining element is not the retrieval of information, but the lack of execution that the agentic architecture offers. Improving embedding models or expanding vector stores will not close this gap.

3.2 Orchestration-First Design Principles

An orchestration-first HCM AI architecture places a reasoning agent at the center of workflow execution rather than at the periphery of information retrieval. Four principles govern this design. Decomposition of goals refers to the process whereby the agent receives task descriptions and converts them to structured goals, complete with subgoals, tool assignments and goal completion criteria, all through the task lifecycle. The first one is tool-nativeness, which suggests that in HCM systems, the APIs, data sources and modes of communication are top tools for use directly by the agent and not from the documentations retrieved afterwards (Jaidi, 2024; Chase, 2022). Layered memory management is a form of memory management system that operates in three levels of memory: the working memory, which contains the information on the context of tasks carried out by the agent; episodic memory, which contains the information on the history of tasks; and semantic memory, which contains the organization policies and domain knowledge (Wang et al., 2024).

3.3 Five-Layer Architecture

Enterprise HCM architecture, which is suggested to be used, contains five connected layers. Intention layer gets natural language instructions, then uses intention classification to direct information requests to the retrieval layer, tool use requests to the tool layer, and complex requests to the planning layer. Such approach makes sure that retrieval-related requests will not use unnecessary agent resources. Reasoning & planning layer uses ReAct reasoning loop (Yao et al., 2023) as its agentic core, which includes such steps as creating a plan, choosing a tool, observing results, adjusting the plan, and repeating until a goal state or human intervention is required. When tasks require more agents, the auto-generation technique (Wu et al., 2023) may be used.

The interface for integrating the tooling layer offers a set of HCM system interfaces that are both curated and authorized to be invoked as tools. An example of how to do this effectively is Jaidi's (2024) composable middleware architecture. The memory layer implements the three-tier model described above, using the context window plus a compact external store for working memory, a structured database for episodic memory, and a vector store for semantic memory. This last element effectively embeds RAG within the orchestration architecture for policy grounding rather than as the primary AI interaction pattern. The governance layer acts as a cross-cutting concern, intercepting all data access requests and tool calls, enforcing data access policies, logging all actions made by the agent, warning on actions nearing policy constraints, and producing compliance documentation for regulated HR processes.

Dimension	RAG (Retrieval-First)	Orchestration-First
Primary operation	Document retrieval and synthesis	Goal decomposition and tool execution
Agent state	Stateless (resets per query)	Persistent across multi-step workflows
Tool access	Vector database only	HCM APIs, data feeds, communication services

Memory management	Context window only	Working, episodic, and semantic memory tiers
Human checkpoints	Each query is a checkpoint	Defined policy-governed approval gates
Output type	Synthesized content	Decision artifacts and executed workflow states
Best suited to	Information lookup, FAQ, policy Q&A	Workforce planning, recruitment pipelines, lifecycle coordination
Ceiling condition	Reached when task requires execution	No ceiling within governed policy scope

Table 1. Structural comparison of RAG and orchestration-first architectures in enterprise HCM contexts.

4. Workflow Analysis: Orchestration Outcomes Across HCM Domains

4.1 Workforce Planning

Workforce planning in the orchestration-first model proceeds as a coordinated multi-source analytical workflow. The agent receives a planning specification, decomposes it into data acquisition, analytical modeling, scenario generation, and output formatting subtasks, and executes each in sequence with dependency management between steps. In the process of data collection, the HRIS API will be called to get the latest headcount based on roles, levels, and business units, the attrition risk model API to get the projection of the turnover rates at the role level, the compensation benchmarking service for getting market rates, and finally the finance system API to get the budget headcount envelopes.

A RAG system responding to the same request retrieves historical headcount reports, attrition rate summaries, and hiring plan templates. The HR leader receives synthesized text but must perform the analytical integration manually. The difference is not at the margins: the RAG system produces content while the orchestration agent produces a decision artifact. The labor involved in moving from content to decision artifact, covering data gathering, analytical integration, format compliance, and routing, is precisely the labor that enterprise AI investments should eliminate.

4.2 Talent Acquisition

Recruitment workflows present similar dynamics. A RAG system can generate job descriptions, answer candidate questions about benefits, and retrieve screening criteria. These are useful capabilities addressing the informational periphery of recruitment, not its operational core. Orchestration-first talent acquisition integrates the full pipeline: posting requisitions to configured job boards via API, monitoring application inflow, applying configurable screening criteria, scheduling interviews through calendar integration, collecting structured evaluation data, applying calibration norms, producing a ranked shortlist, and routing it to the hiring manager with documented evaluation rationale.

At each consequential step, the governance layer evaluates whether human approval is required under organizational policy. Sourcing decisions, screening criteria application, and scheduling execute autonomously within defined parameters. Candidate ranking and offer recommendation surface to human decision-makers with full reasoning traces. This checkpoint architecture preserves human authority over the decisions with the highest legal and organizational stakes while automating the coordination tasks that currently consume disproportionate recruiter time. Gurumurthy et al. (2023) demonstrate that AI-based HR functions have the highest efficiency impacts in workflow types that are characterized by data consolidation across systems and approval routing, which is exactly the case with enterprise recruitment.

4.3 Employee Lifecycle Management

The lifecycle of an employee comprises processes such as onboarding, performance management, learning and development, internal mobility, and offboarding. All these phases entail multi-step process flows that have cross-application dependencies and compliance aspects. During the performance management process, the orchestration engine launches review cycles as scheduled, allocates evaluation tools to relevant reviewers based on hierarchy data pulled from HRIS, tracks completion ratios against SLAs, issues notifications specific to completion status, identifies rating inconsistencies with respect to calibration standards, facilitates approval workflows for completed evaluations, and maintains employees' records post completion of the cycle. The role of the HR practitioner is thus transformed from process administrator to exception handler and decision maker, as noted by Strohmeier (2020).

In learning and development, the orchestration agent collects information regarding the gaps between skill sets through performance appraisals and job roles, creates customized learning suggestions using career path

information, registers the employees for courses based on LMS API, oversees course completion and performance assessments, and updates skill inventories in HRIS. This integration of multiple sources of information, which is usually done manually, occurs in the reason loop of the agent, throughout the entire process of development.

HCM Workflow	RAG Contribution	Orchestration Contribution	Primary Value Driver
Workforce Planning	Historical report synthesis	Multi-source data integration, scenario modeling	Decision artifact generation
Talent Acquisition	JD drafting, screening criteria lookup	End-to-end pipeline coordination across ATS, calendar, HRIS	Coordinator labor reduction
Onboarding	Policy and procedure Q&A	Cross-system task sequencing, completion tracking	Process speed and consistency
Performance Management	Review criteria and calibration norms retrieval	Cycle initiation, routing, anomaly flagging, record update	Compliance and cycle completion rate
Learning and Development	Course recommendations	Skill gap analysis, enrollment, completion monitoring, HRIS update	Personalization at scale
Compliance Monitoring	Policy text retrieval	Transaction monitoring, violation detection, corrective workflow initiation	Regulatory risk reduction

Table 2. RAG and orchestration contributions across core HCM workflow categories.

5. A Maturity Model for HCM AI Integration

There is a well-defined process model through which organizations can successfully introduce agentic AI within their Human Capital Management scenarios as there will be a calibration of investment according to organizational readiness and risk appetite. The following process model includes four levels from retrieval-first to orchestration.

Level 1 — Retrieval-Augmented Assistance. The company uses RAG AI for accessing HR information such as policy Q&A, benefits inquiry, FAQ automation for onboarding, and document generation. AI functions as an information intermediary. Workflow coordination, system integration, and decision execution remain entirely with human HR staff. Most enterprise HCM deployments currently operate at Level 1 or are completing Level 1 implementations. The value delivered at this level, reduced time to answer common HR questions and consistent policy application in AI-generated responses, is real, but the coordination burden on HR professionals remains unchanged.

Level 2 — Tool-Augmented Response. The organization extends RAG with limited tool calling. It has the capability to obtain live data based on a natural language question and initiate workflow actions based on predefined triggers for a particular request from the user. While level 2 makes HR transactions frictionless and answers data-dependent questions accurately, it neither maintains goal state during interaction nor performs multi-step workflows or multi-system actions. The demarcation between levels 2 and 3 is persistent agent state and multi-step planning.

Level 3 — Supervised Orchestration. The organization uses orchestration agents for workflow categories with mandatory human checkpoints. Agents decompose task specifications into subtask sequences, execute tool invocations in the required order, manage intermediate state, and surface reasoning at checkpoint junctures. Human approval gates are mandatory for data modification, system updates, and decision artifacts. The organization maintains audit logs of all agent actions. Brynjolfsson et al. (2023) find that AI assistance shows the strongest productivity effects when it handles task coordination while preserving human decision authority, precisely the model that Level 3 implements.

Level 4 — Governed Autonomous Orchestration. The organization applies supervised orchestration in high throughput standardized processes where compliance with policy can be validated algorithmically. Agents perform specific categories of workflows without individual instances needing human approval; this is monitored on an ongoing basis. Human supervision changes from per-action review to exception handling, policy enforcement, and monitoring of performance. Benefits enrollment processing, onboarding task sequencing, mandatory training compliance tracking, and routine reporting are candidate applications. High-

stakes decisions such as compensation changes, performance ratings, and employment status changes remain at Level 3 indefinitely.

Level	Label	AI Capability	Human Role	Key Limitation
1	Retrieval-Augmented Assistance	RAG for policy lookup, FAQ, document generation	Full workflow coordination and execution	No execution capability; coordination labor unchanged
2	Tool-Augmented Response	RAG plus limited tool calling (live data, simple triggers)	Coordinates tool outputs; approves all actions	No persistent goal state; no multi-step planning
3	Supervised Orchestration	Multi-step agents with mandatory human approval gates	Exception handler and decision authority at checkpoints	Per-action approval overhead for high-volume routine work
4	Governed Autonomous Orchestration	Agents execute defined routine workflows autonomously	Exception management, policy governance, performance monitoring	Requires robust compliance verification for autonomous scope

Table 3. Four-level HCM AI integration maturity model.

6. Governance and Compliance Considerations

6.1 Data Privacy and Access Control

The orchestrating agents operate on sensitive data about employees in different systems in a single reasoning thread. Accessing data from different systems in one reasoning thread creates privacy threats that point-in-time RAG queries cannot create. Access should be governed based on purpose: the agent uses only the minimum amount of data that is needed for the task, which is defined by the task description. Data minimization principle applies to each use of the tool, not only to the overall workflow. The obligation to minimize data follows from the GDPR and CCPA, among others. Tambe et al. (2019) and Malik et al. (2020) both identify regulatory compliance as a first-order concern for AI integration in HR, and this concern intensifies as AI moves from information retrieval to workflow execution.

6.2 Algorithmic Fairness

Orchestrating agents that participate in hiring, performance evaluation, and compensation workflows operate in legally sensitive territory. Disparate impact analysis, which applies to any selection procedure resulting in significantly different selection rates across protected groups, extends to AI-assisted selection procedures in most jurisdictions with meaningful anti-discrimination enforcement. Pre-deployment auditing for bias in relation to the various components of the agent used for consequential HR processes should be ensured for organizations belonging to levels 3 and 4, apart from the continuous monitoring of agent decision-making distributions and documented remediation measures if bias is found, and independent audit review of the monitoring process. Malik et al. (2020) point out that workforce AI bias has a disproportionate impact on minorities within the labor force, thus necessitating monitoring of the same.

6.3 Auditability and Explainability

Employment decisions have legal bearing, and there is growing regulation of AI-influenced decisions. The governance layer will need to create human-readable reasoning trails for every agent action that impacts significant HR decisions. These trails need to be stored in immutable audit logs that can be accessed by HR management, lawyers, and regulatory examiners. The explainability requirement is not merely about post-hoc documentation; it informs the architectural design of the reasoning layer. Agents must generate reasoning that humans can evaluate, not merely reasoning that produces correct outputs. The distinction matters because the audit requirement is triggered by process as well as outcome: an agent that produces a correct hiring recommendation through a reasoning process that cannot be explained may satisfy the accuracy requirement while failing the compliance requirement.

Governance Dimension	Key Requirement	Applicable Regulation / Standard	Implementation Mechanism
Data Privacy	Purpose-based access control; data minimization per tool invocation	GDPR Art. 5, 25; CCPA; sector equivalents	Governance layer: access policy enforcement, audit logging
Algorithmic Fairness	Pre-deployment bias audit; continuous disparate impact monitoring	EEOC Uniform Guidelines; EU AI Act (HR category)	Monitoring pipeline; remediation procedures; independent audit
Auditability	Human-readable reasoning traces for all consequential actions	Employment law (varies by jurisdiction); EU AI Act Art. 13	Tamper-evident audit log; trace retention policy
Explainability	Agent reasoning evaluable by humans at every checkpoint	EU AI Act Art. 13; NIST AI RMF 1.0	Checkpoint surfacing in governance layer

Table 4. Enterprise HCM orchestration governance requirements.

7. Discussion

7.1 Strategic Implications for HR Technology Leaders

The orchestration-first argument presented in this paper carries a direct strategic implication for HR technology investment: RAG platforms and orchestration frameworks are not interchangeable, and vendor-facing strategies that treat LLM integration as a retrieval upgrade will miss the orchestration dividend. The maturity model provides a framework for honest current-state assessment. For most organizations, the gap between Level 1 and Level 3 involves three specific investments that retrieval optimization does not address: a tool integration layer that exposes HCM system APIs to the reasoning layer, a memory management infrastructure that supports cross-session task continuity, and a governance layer that enforces compliance requirements at the action level. These are infrastructure investments, not incremental improvements to existing retrieval stacks.

HR technology leaders evaluating agentic AI platforms should distinguish between vendors offering tool-augmented chat interfaces, which represent Level 2 capability, and vendors offering genuine orchestration frameworks with persistent goal management, multi-step planning, and governance-layer integration. The architectural difference between these two categories is larger than the marketing language surrounding them typically suggests, and the operational implications of conflating them are significant.

7.2 Implications for HCM Research

Damayanti et al. (2026) document substantial growth in digital HR and workflow automation research since 2020, identifying it as one of the most active convergence zones in the HCM literature. This paper contributes an architectural framework to that literature, one that distinguishes between the retrieval contribution and the orchestration contribution to HCM AI value creation and provides testable propositions about the workflow categories where each contribution is largest. The four-level maturity model generates empirically testable hypotheses: organizations at Level 3 should show larger reductions in HR coordinator labor hours on coordination-intensive workflows than organizations at Level 1 or Level 2; organizations at Level 4 should show lower error rates on high-volume routine workflows than organizations with human-in-the-loop models for those workflows. These hypotheses can be tested as enterprise orchestration deployments mature and longitudinal data becomes available.

7.3 Labor Market Considerations

Eloundou et al. (2023) project that occupations with the highest LLM exposure include many HR-adjacent professional roles. Orchestration-first AI will concentrate this exposure on the coordination and execution tasks that currently differentiate HR professionals, not only the information tasks that RAG already addresses. Organizations implementing Level 3 and Level 4 orchestration should expect demand to shift from high-volume HR coordinators and administrators toward HR professionals with stronger analytical, governance, and vendor management capabilities. It is reported by Brynjolfsson et al. (2023) that the influence of AI-enabled skill development is stronger in case of new employees who have less experience, which means that the implementation of orchestration might help to develop skills of young HR professionals faster. These dynamics

warrant proactive workforce planning within the HR function itself, a somewhat ironic implication for the teams deploying the technology.

8. Conclusion

The case for orchestration-first agentic AI in enterprise HCM rests on a structural argument: retrieval-augmented generation cannot execute. It can inform; it cannot act. The coordination, planning, and multi-system integration tasks that represent the bulk of enterprise HR operational labor require a different architectural foundation, one in which a reasoning agent maintains goal state, selects and invokes tools, manages intermediate results across a workflow, and adapts to changing conditions throughout execution.

This paper has characterized the retrieval ceiling as a structural limit rather than a performance plateau, proposed a five-layer orchestration architecture grounded in the agentic AI literature, analyzed the differential outcomes of retrieval and orchestration across three core HCM workflow domains, presented a four-level maturity model for practitioner use, and articulated the governance requirements that enterprise HR compliance environments impose on orchestrating agents. The governance considerations are positioned as architectural requirements, not post-deployment additions, because embedding compliance at the design stage is both more effective and more defensible than retrofitting it after deployment.

The transition from retrieval to orchestration in enterprise HCM will not happen in a single architectural replacement. It will proceed through the maturity levels described here, paced by organizational risk tolerance, technical capacity, and regulatory readiness. Organizations that initiate this transition now, building tool integration infrastructure, developing orchestration governance frameworks, and piloting supervised orchestration in defined workflow categories, will accumulate architectural capital that compounds as agentic capabilities mature. Organizations that continue optimizing retrieval past its productive ceiling will face the architectural transition from a position of increasing competitive disadvantage.

Ethics Statement

Conflict of Interest

The author declares no conflict of interest.

Statement of Human and Animal Rights

This study did not involve human participants, animal subjects, or identifiable personal data requiring ethics board review. All analysis is conceptual and draws on published literature.

Informed Consent

Not applicable.

References

1. Anthropic. (2024). Model specification [Technical report]. <https://www.anthropic.com/model-spec>
2. Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. National Bureau of Economic Research Working Paper, No. 31161. <https://doi.org/10.3386/w31161>
3. Chase, H. (2022). LangChain: Building applications with LLMs through composability [Software repository]. GitHub. <https://github.com/langchain-ai/langchain>
4. Dasaklis, T. K., Ourani, H., & Salami, H. (2025). Agentic AI in enterprise systems: An empirical review of orchestration frameworks and workflow automation. *Technological Forecasting and Social Change*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5314976
5. Damayanti, N., Wulandari, F., & Santoso, I. (2026). Two decades in five years: Mapping digital HR, workflow automation, and innovation ecosystem research (2020–2025). *Daengku: Journal of Humanities and Social Sciences Innovation*, 6(1). <https://jurnal.ahmar.id/index.php/daengku/article/view/4894>
6. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*. <https://arxiv.org/abs/2303.10130>
7. Gurumurthy, R., Bhatt, N., & Maciejewski, A. (2023). From automation to augmentation: How AI is reshaping HR operations in large enterprises. *Journal of Business Research*, 162, 113907. <https://doi.org/10.1016/j.jbusres.2023.113907>
8. Jaidi, N. (2024). Composable middleware architecture for AI-ready enterprise platforms. *International Journal of Intelligent Systems and Applications in Engineering*, 12(3), 1–14. <https://ijisae.org/index.php/IJISAE/article/view/8342>

9. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
10. Malik, A., Budhwar, P., & Srikanth, N. R. (2020). Gig economy, social media, and HRM. *International Journal of Human Resource Management*, 31(19), 2420–2450. <https://doi.org/10.1080/09585192.2020.1756162>
11. Marler, J. H., & Parry, E. (2016). Human resource management, strategic involvement and e-HRM technology. *International Journal of Human Resource Management*, 27(19), 2233–2253. <https://doi.org/10.1080/09585192.2015.1091980>
12. Nakano, R., Hilton, J., Balwit, A., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Krueger, G., Hilton, K., & Schulman, J. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*. <https://arxiv.org/abs/2112.09332>
13. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192. <https://doi.org/10.1126/science.adh2586>
14. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, Article 2. <https://doi.org/10.1145/3586183.3606763>
15. Significant Gravititas. (2023). AutoGPT: An autonomous GPT-4 experiment [Software repository]. GitHub. <https://github.com/Significant-Gravititas/AutoGPT>
16. Strohmeier, S. (2020). Digital human resource management: A conceptual clarification. *German Journal of Human Resource Management*, 34(3), 345–365. <https://doi.org/10.1177/2397002220940936>
17. Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4), 15–42. <https://doi.org/10.1177/0008125619867910>
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
19. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18, 186345. <https://doi.org/10.1007/s11704-024-40231-1>
20. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
21. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-generation LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*. <https://arxiv.org/abs/2308.08155>
22. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X