A Capability–Maturity–Value Taxonomy for Agentic AI Systems: Bridging Technical Capabilities and Organizational Value

Boon Sing Thia[1]

Singapore Management University, Singapore
bs.thia.2024@engd.smu.edu.sg

Submitted to:
Prof. TA Nguyen Binh Duong and Prof. Zhu Feida

Abstract

Agentic AI systems – autonomous agents capable of reasoning, planning, tool utilization, and multi-agent coordination – are advancing rapidly. However, corporate adoption is constrained by the lack of frameworks that link technological capabilities to quantifiable organizational outcomes. This paper presents the Capability-Maturity-Value (CMV) Taxonomy, a structured classification framework developed using the systematic Nickerson et al. methodology. The taxonomy comprises nine dimensions (Reasoning Sophistication, Tool Integration Depth, Memory Persistence, Coordination parameter, Autonomy Level, Adaptation Mechanism, Human Collaboration Mode, Compliance & Alignment, and Safety Architecture). Each dimension has 4 characteristics that are linked to organizational value categories. We augment the capability taxonomy with a maturity model that assesses implementation quality irrespective of the capability type. By categorizing 25 reference objects sourced from academic research and industry experience, we illustrate the taxonomy's discriminative efficacy and practical applicability. The CMV framework gives researchers a systematic vocabulary for classifying agentic AI, provides practitioners with useful advice for designing and evaluating systems, and gives business leaders a basis for making decisions about where to spend and how to deploy. The artifacts of the paper can be found in: `https://github.com/alphathia/ERP1`.

Keywords: Agentic AI, Taxonomy Development, Capability Maturity, Enterprise AI, LLM Agents, Organizational Value

1 Introduction

Agentic artificial intelligence represents a paradigm shift from assistive AI systems to autonomous agents capable of independent reasoning, planning, tool orchestration, and goal-directed action [1, 43]. These systems move beyond single-turn interactions to execute complex, multi-step workflows with varying degrees of human oversight. Recent advances in large language models (LLMs) have accelerated this transition, enabling agents to decompose problems, invoke external tools, maintain context in extended interactions, and collaborate with other agents or humans [51, 33].

Despite rapid technical progress, the adoption of agentic AI in the enterprise remains constrained. Industry surveys indicate that while 88% of organizations report regular use of AI, only 39% attribute material business impact to AI deployments, with the majority remaining in the experimental or pilot stages [31]. This gap between capability and value realization reflects a fundamental evaluation challenge: organizations possess increasingly sophisticated technical options but lack frameworks for translating capability assessments into deployment decisions and value projections.

Existing evaluation approaches address this challenge incompletely. Technical benchmarks such as AgentBench [24] assess task performance across environments but do not connect results to organizational outcomes. Capability taxonomies [1, 33] catalog agent components but organize them by technical functions rather than value creation pathways. Maturity frameworks from adjacent domains (e.g., CMMI) do not capture agentic-specific characteristics such as autonomous tool use, long-term memory, or multi-agent coordination. The Agentic ROI concept [23] provides a valuable economic framework, but lacks the systematic capability classification needed for comparative assessment.

This research addresses this gap by developing the Capability-Maturity-Value (CMV) Taxonomy – a systematic classification framework that explicitly connects agentic AI capabilities to organizational value creation. Following the rigorous taxonomy development methodology of Nickerson et al. [38], we construct a framework grounded in systematic analysis of academic literature and industry practice.

Our contributions are fourfold:

1. **A nine-dimension capability taxonomy** with 36 characteristics, each grounded in literature and explicitly linked to value creation mechanisms. A capability dimension is a conceptually distinct category of agent functionality that influences organizational value creation—such as reasoning, tool use, or memory. Each dimension comprises four characteristics: mutually exclusive, hierarchically ordered levels of capability sophistication within that dimension, where higher levels generally subsume lower-level capabilities while adding new ones. Together, the dimensions

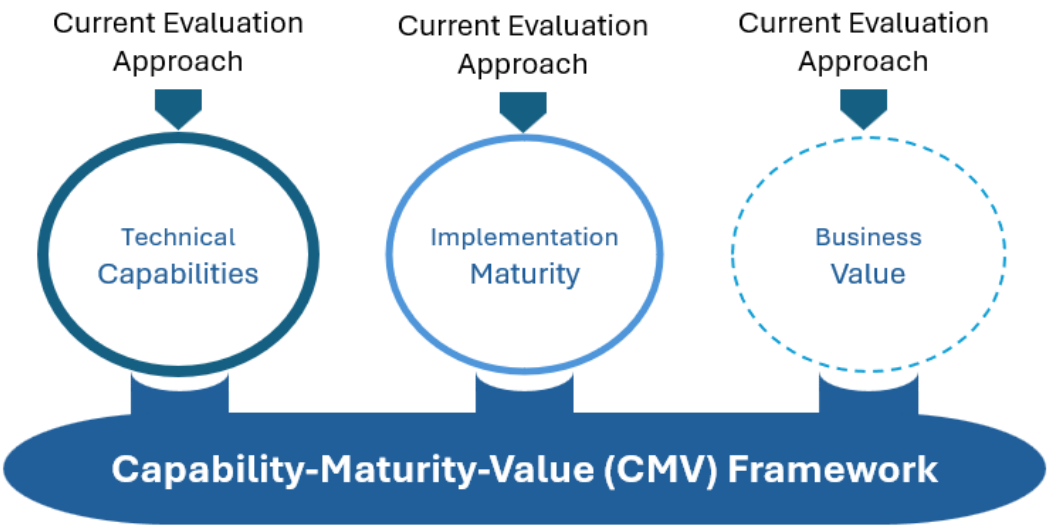

Figure 1: Current agentic AI evaluation approaches address technical capabilities, implementation maturity, and business value in isolation. The CMV Framework provides an integrated lens connecting all three perspectives.

define what category of capabilities matters for value creation, while characteristics specify how those capabilities manifest at increasing levels of advancement (Section 4.2; complete definitions in Appendix A).

2. **A maturity model** assessing implementation quality independent of capability type, enabling evaluation of "how well" systems deliver their capabilities (Section 4.3). Figure 1 illustrates the Capability-Maturity-Value (CMV) Framework.

3. **Value realization pathways** mapping dimensions to measurable organizational outcomes, including productivity, decision quality, risk mitigation, and innovation (Section 4.5).

4. **Empirical validation** through classification of 25 diverse reference objects (appendix B), demonstrating profile uniqueness (24 of 25 distinct classifications), confirming that the taxonomy meaningfully differentiates agentic AI systems. (Section 5).

The remainder of this paper is organized as follows. Section 2 reviews related work on agentic AI evaluation and positions our contribution. Section 3 details our methodology for the development of taxonomies. Section 4 presents the CMV taxonomy, including dimensions, characteristics, and maturity integration. Section 5 demonstrates empirical validation through object classification. Section 6 discusses implications and limitations. Section 7 concludes with directions for future research.

2 Background and Related Work

2.1 Agentic AI Systems

Agentic AI systems are distinguished from traditional AI by their capacity for autonomous, goal-directed behavior. Acharya et al. [1] define agentic AI as systems exhibiting autonomy (independent operation without continuous human intervention), goal-orientation (executing objectives through planning and execution), adaptability (adjustment to environmental changes), and proactivity (anticipation of needs and initiation of actions). These systems build upon LLM foundations, but extend them through architectural patterns enabling tool use, memory persistence, and multi-agent coordination [25].

The technical landscape encompasses diverse implementations. Single-agent architectures range from reactive patterns (e.g., ReAct [51]) to reflective systems with self-correction capabilities. Multi-agent systems introduce coordination paradigms including hierarchical delegation, peer-based collaboration, and dynamic coalition formation [29, 50]. Enterprise deployments span domains from software engineering [32] to IT operations [58] and integration of the Enterprise Resource Planning (ERP) system [49].

The rise of agentic AI has been accelerated by open-source and commercial AI agent development frameworks that lower the barrier to building sophisticated agentic AI systems. **LangChain** [18] provides a modular framework for composing LLM-powered applications with chains, agents, and retrieval-augmented generation (RAG) capabilities, establish-

ing foundational patterns for tool integration and memory management. **AutoGen** [57] introduces a multi-agent conversation framework enabling customizable agents to collaborate through structured dialog, supporting diverse coordination patterns from hierarchical to peer-based configurations. **CrewAI** [5] focuses on role-based multi-agent orchestration, allowing developers to define specialized agent roles (e.g., researcher, writer, analyst) that collaborate on complex tasks through structured workflows. Most recently, **Google Agent Development Kit (ADK)** [11] provides enterprise-grade infrastructure to build production-ready agents with built-in support for safety guardrails, compliance monitoring, and scalable deployment. These frameworks operationalize the theoretical capabilities discussed in academic surveys but do not themselves provide systematic taxonomies for capability assessment or value realization—a gap the CMV Framework addresses.

2.2 Capability Taxonomies

The systematic classification of agentic AI capabilities has emerged as a foundational research stream. Li et al. [19] synthesize fundamental capabilities in reasoning, planning, tool integration, and memory management, identifying capability dependencies where effective tool use presupposes adequate reasoning. Comprehensive surveys have mapped the capability landscape: Acharya et al. [1] identify four defining properties (autonomy, goal-orientation, adaptability, proactivity); Pati et al. [43] distinguish short-term working memory from long-term episodic stores; and Piccialli et al. [45] document domain-specific capability manifestations in healthcare, finance, and manufacturing.

Reasoning and planning capabilities enable agents to decompose complex goals into actionable steps. Sun et al. [51] categorize mechanisms from chain-of-thought prompting to tree-of-thought exploration and self-reflection loops. Multi-stage reasoning frameworks combine knowledge integration with multi-hop reasoning [55, 47], while the plan-then-execute paradigm affects user trust and team performance [13]. Goal reasoning can incorporate normative constraints [41], and Liu et al. [25] catalog 18 architectural patterns to implement these capabilities.

Tool integration and memory mechanisms extend agent functionality and distinguish sophisticated agents from reactive systems. Tool use encompasses discovery, selection, invocation, and orchestration [9], with enterprise applications demonstrating significant cost reductions in ERP [49], e-commerce [3], legal services [46], and infrastructure security [52]. Memory implementation in LLM-based agents varies across three dimensions: sources (inside-trial, cross-trial, external knowledge), forms (textual or parametric), and operations (writing, management, reading) [59]. Textual memory dominates due to its interpretability, yet current systems under-invest in parametric mechanisms, structure-enhanced representations, and lifelong learning [12]. The proposed solutions include MemoryBank [60] and designs inspired by the human brain [26].

Coordination and collaboration capabilities address both multi-agent and human-agent interaction. Multi-agent frameworks standardize deployment components [29] and address cooperative decision-making [50, 35]. Dynamic task decomposition generates specialized subagents [56], with applications in smart cities [28]. Human-agent collaboration requires transparent communication and trust calibration [15], with taxonomies distinguishing introspective from assistive explanations [53]. Transparency in planning improves the effectiveness of collaboration [13, 9].

Autonomy and adaptation span the spectrum from human-in-the-loop to fully autonomous operation. Taxonomies examine how autonomy manifests in application domains [1, 45], with frameworks to specify SLEEC requirements [10]. Autonomous visualization agents leverage multi-modal models [22], while generalist agents outperform specialists across diverse tasks [32]. Adaptive capabilities enable self-evolution [8], dynamic manufacturing decision-making [16], and feedback-driven refinement [14]. Deep learning and reinforcement learning contribute foundational techniques [30], with knowledge editing allowing efficient capability updates [54].

Governance and safety capabilities ensure agents conform to policies, prevent harm, and protect privacy. The frameworks address the SLEEC (Social, Legal, Ethical, Empathetic, and Cultural) requirements [10], value preference estimation [21], and reward alignment metrics [48, 27]. Safety research identifies AI hazards and responsibility gaps [6], surveys security challenges throughout the agent lifecycle [7, 37], and proposes secure lifecycle management [39]. Privacy frameworks address multi-agent applications [42], while domain-specific safety architectures protect vulnerable populations [36]. The OpenAI Preparedness Framework [40] exemplifies industry approaches with risk thresholds gating deployment.

Despite their comprehensiveness, existing capability taxonomies organize by *architectural function* rather than *value creation pathways*, limiting guidance for deployment decisions.

2.3 Maturity Models and Performance Levels

Maturity models provide structured frameworks for assessing implementation quality along defined progression paths. The Capability Maturity Model (CMM), originally developed by Paulk et al. [44] for software process improvement, established the paradigm of ordered maturity levels (Initial, Repeatable, Defined, Managed, Optimizing) that capture

increasing process sophistication and predictability. This framework has proven influential across domains, demonstrating that *how well* an organization implements capabilities matters as much as *what* capabilities it possesses.

In the AI domain, Morris et al. [34] propose AGI Levels (L0–L5) as a framework for classifying system performance relative to human capabilities. Level 0 (No AI) represents systems without AI components; Level 1 (Emerging) describes systems that match unskilled human performance on some tasks; Level 2 (Competent) matches median human performance; Level 3 (Expert) matches 90th-percentile human performance; Level 4 (Virtuoso) matches 99th-percentile performance; and Level 5 (Superhuman) exceeds all human performance. This benchmark-oriented approach provides clear performance thresholds but focuses on task-level competence rather than deployment-level implementation quality.

The OpenAI Preparedness Framework [40] takes a safety-oriented approach to capability progression, defining High and Critical thresholds in different risk domains, specifically cybersecurity, biological and chemical threats, and AI self-improvement. This framework distinguishes between task-level capability evaluations, which measure raw model potential (e.g., identifying zero-day exploits) and deployment-level safeguards, which mitigate residual risk through governance structures like monitoring and refusal mechanisms. Consequently, deployment decisions are gated by safety assessments, ensuring that advancement in capabilities are matched with increasing safeguards.

These maturity and performance frameworks advance the understanding of the progression of the AI system but exhibit limitations for enterprise evaluation. AGI Levels assess performance relative to human benchmarks without addressing organizational integration or value delivery. Safety frameworks focus on risk mitigation rather than value creation. Neither approach systematically connects the maturity assessment to the full spectrum of agentic capabilities or organizational outcomes.

2.4 Value Frameworks and Enterprise Impact

The evaluation of agentic AI systems has shifted from simple cost-benefit analysis to sophisticated frameworks that capture multidimensional value. Liu et al. [23] formalize this through Agentic ROI, arguing that widespread adoption hinges on balancing "Information Quality" against the "Interaction Cost" (user time and expense) required to achieve it (see Appendix D for a formal definition). This formulation reorients the evaluation from the raw capability toward practical utility. It recognizes that technical sophistication creates zero value if the output is poor or the user burden is too high. Enterprise data validates this view: while most organizations focus on efficiency, McKinsey's State of AI report [31] reports that "high performers" distinguish themselves by leveraging AI for transformational innovation, revenue growth, and competitive differentiation.

Domain-specific implementations demonstrate these value pathways in practice. For example, Sarferaz [49] documents ERP integration patterns in which agentic AI creates value through automated workflow execution, intelligent exception handling, and predictive maintenance scheduling. These enterprise deployments illustrate how capability configurations map to specific value outcomes –reasoning capabilities enable better decision support, while tool integration capabilities drive process automation.

However, a gap remains in the comparative assessment. Although frameworks like Agentic ROI provide economic logic and enterprise studies document the outcomes, the field lacks a systematic structure mapping specific technical capabilities to these value drivers. We need a framework that specifies exactly which capabilities improve information quality or reduce interaction costs to effectively prioritize deployment.

2.5 Existing Evaluation Approaches

Current evaluation frameworks approach agentic AI from different perspectives, each with limitations for enterprise deployment decisions.

Technical Benchmarks. AgentBench [24] provides multi-environment evaluation across eight domains, assessing reasoning and decision-making through task completion metrics. Although valuable for model comparison, such benchmarks do not address implementation quality or organizational value. Similarly, evaluation surveys [33] organize metrics by behavior, capabilities, reliability, and safety but focus on *what* to measure rather than *why* these capabilities matter for business outcomes.

Capability Taxonomies. Comprehensive surveys [1, 43, 45] catalog agentic capabilities including reasoning, tool use, memory, and coordination. These taxonomies excel at technical coverage but organize capabilities by architectural function rather than value creation pathways, limiting their utility for deployment prioritization.

Maturity and Performance Levels. Morris et al. [34] propose AGI Levels (L0–L5) based on performance benchmarks relative to human capabilities, from "No AI" to "Superhuman." The OpenAI Preparedness Framework [40] defines capability thresholds requiring different safeguards. Although these provide progression models, they do not systematically connect levels to organizational value or address the full spectrum of agentic capabilities.

Economic Frameworks. Liu et al. [23] introduce Agentic ROI, defining usability as the ratio of information quality to interaction costs. This economic perspective valuably reorients the evaluation toward practical utility but lacks the

Table 1: Comparison of Agentic AI Literature Streams

Category	Representative Work	Cap.	Mat.	Value	Key Limitation
Agent Capability Surveys					
	Li et al. [19]	Yes	No	No	No value linkage
	Acharya et al. [1]	Yes	No	No	Architecture-focused
	Pati et al. [43]	Yes	No	No	No deployment guidance
Evaluation Frameworks					
	AgentBench [24]	Partial	No	No	Task metrics only
	Mohammadi et al. [33]	Partial	No	Partial	Process-focused
Enterprise Deployment					
	McKinsey [31]	No	Partial	Yes	No capability structure
	Sarferaz [49]	Partial	No	Yes	Domain-specific
Safety/Governance					
	Morris et al. [34]	Partial	Yes	No	Human benchmark only
	OpenAI Prep. [40]	Partial	Yes	Partial	Safety-centric
Multi-Agent & Human-AI					
	Maldonado et al. [29]	Yes	No	No	Coordination only
	Iftikhar et al. [15]	Partial	No	Partial	Team dynamics only
CMV (Ours)	**Integrated Framework**	**Yes**	**Yes**	**Yes**	—

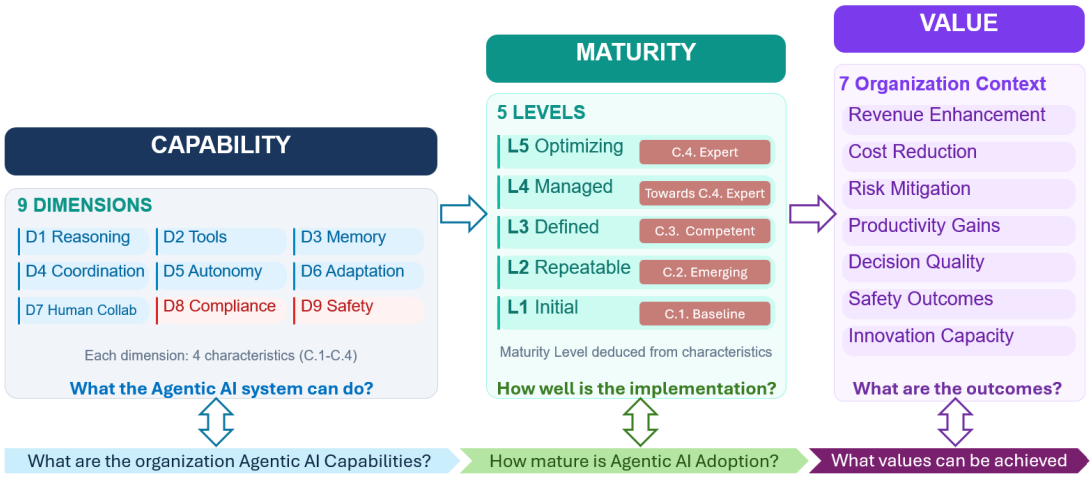

Figure 2: **The Capability–Maturity–Value (CMV) Framework**. The framework links nine capability dimensions, five maturity levels (L1–L5) through 4 characteristics, and seven value outcome categories, providing an integrated lens from technical design to enterprise impact.

systematic capability classification needed for comparative assessment across diverse types of agents.

2.6 Research Gap

The preceding literature review reveals a fragmented landscape in which each research stream addresses important aspects of agentic AI assessment while leaving critical dimensions unexplored. Table 1 synthesizes these gaps across five corpus categories.

Capability taxonomies (Section 2.2) provide comprehensive technical coverage but organize capabilities by architectural function rather than value creation pathways, offering limited guidance for deployment prioritization. **Maturity models** (Section 2.3) establish progression frameworks but focus on human-relative performance benchmarks or safety thresholds without connecting to organizational outcomes. **Value frameworks** (Section 2.4) capture economic logic and document the impact of the organization, but lack systematic capability structures that allow comparative evaluation of various types of agents.

No current framework simultaneously provides: (1) comprehensive capability classification that encapsulates the nine dimensions essential for agentic behavior, (2) independent maturity assessment of the type of capability that evaluates the quality of the implementation, and (3) explicit link between capabilities and the seven organizational value categories through which AI systems create impact. The CMV Framework (Figure 2) addresses this integration gap by providing a value-centric view for agentic AI systems that connects capability, maturity, and organizational results. On the capability

side, it specifies nine dimensions of agentic AI capabilities (D1–D9), each instantiated through four ordered characteristics (C.1–C.4) that describe what an agent can do and how sophisticated that capability is (for example, from the use of single basic tools to the integration of the enterprise ecosystem). For any given system, the selected characteristics across the nine dimensions form its capability profile. The maturity panel then evaluates how well each selected characteristic is implemented using a five-level scale — (L1 to L5, from Initial to optimization), producing an independent maturity score per dimension that separates the mere possession of a capability from its consistency, reliability, and optimization in practice. Finally, the maturity profile feeds into the value panel, which traces how mature capabilities activate specific value pathways across seven outcome categories: revenue enhancement, cost reduction, risk mitigation, productivity gains, decision quality, safety outcomes, and innovation capacity. Together, this structure answers three core questions: 1) what agentic AI capabilities the organization has, 2) how mature is its adoption, and 3) what value can be achieved—and provides a traceable mechanism from concrete capabilities, through their maturity, to measurable organizational outcomes that inform deployment and investment decisions.

3 Methodology

3.1 Taxonomy Development Approach

A *taxonomy* is a formal classification system that organizes objects of interest into mutually exclusive and collectively exhaustive categories based on shared characteristics [38]. In the context of this research, we develop a taxonomy to systematically classify agentic AI systems according to their capability configurations and value-creation potential. Unlike ad-hoc categorizations, a rigorous taxonomy provides a structured vocabulary that enables precise comparison, gap identification, and theory building, essential for an emerging domain where conceptual clarity is lacking.

We employ the taxonomy development methodology proposed by Nickerson, Varshney, and Muntermann [38], which has become the gold standard for classification system development in information systems research. This method is particularly well-suited to our research objectives for three reasons. First, it provides a systematic, iterative approach that accommodates the evolving understanding characteristic of nascent technological domains. Second, it balances theoretical rigor with practical applicability, ensuring that the resulting taxonomy serves both academia and practitioner. Third, it prescribes explicit ending conditions that provide objective criteria for determining when taxonomy development is complete.

The Nickerson et al. method formally defines a taxonomy T as a set of n dimensions D, where each dimension D_i consists of k_i mutually exclusive and collectively exhaustive characteristics C_{ij}:

$$T = \{D_i, i = 1, ..., n \mid D_i = \{C_{ij}, j = 1, ..., k_i\}\} \tag{1}$$

The method prescribes an iterative development process that alternates between two complementary approaches.

Empirical-to-Conceptual (E→C): This inductive approach begins with concrete objects. In our case, the documented agentic AI systems, architectural patterns, and deployment scenarios are drawn from the literature. By analyzing the observable characteristics of these objects, we derive dimensions and characteristics that capture meaningful distinctions. This approach is appropriate when there are rich empirical data, as it ensures that the taxonomy is grounded in real-world manifestations of the phenomenon.

Conceptual-to-Empirical (C→E): This deductive approach begins with theoretical constructs or conceptual frameworks, then validates whether the proposed dimensions and characteristics adequately classify the objects of interest. This approach is appropriate for refining initial structures, testing boundary conditions, and ensuring conceptual coherence.

In our research, "empirical" refers to the documented instances of agentic AI systems and patterns found in our literature corpus, i.e. the concrete manifestations from which we inductively derive classification dimensions. "Conceptual" refers to the abstract organizing principles — dimensions and characteristics — that we construct to systematically categorize these instances. Through iterating between "Empirical-to-Conceptual" and "Conceptual-to-Empirical" ensures that our taxonomy is simultaneously grounded in observable reality and conceptually coherent.

We start with defining the meta-characteristics (section 3.2) to derive 25 reference objects (appendix B). Then, applying the method of Nickerson et al. [38], we performed three iterations to arrive at the end conditions. In Iteration 1 (E→C), we inductively derived an initial capability taxonomy by classifying 25 reference objects and their documented capabilities, yielding eight preliminary dimensions. In Iteration 2 (C→E), we applied the initial capability taxonomy to the same objects to test classifications, surface definition ambiguities and refine dimension boundaries. This includes splitting the original governance dimension into separate dimensions "Compliance and Alignment" and "Safety Architecture" and formalizing value-linked characteristic definitions. Iteration 3 (C→E) then conducted a gap analysis and verified that all objective and subjective ending conditions of Nickerson et al. were satisfied, yielding the final nine-dimension, 36-characteristic CMV taxonomy. Figure 3 illustrates this process.

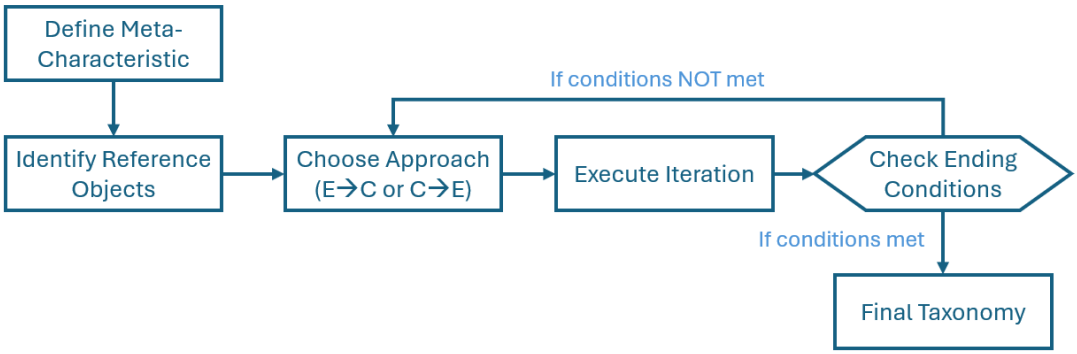

Figure 3: The Nickerson et al. taxonomy development process as applied in this research. Beginning with meta-characteristic definition, we alternated between empirical-to-conceptual (E→C) and conceptual-to-empirical (C→E) approaches across three iterations until all ending conditions were satisfied.

3.2 Meta-Characteristic Definition

The *meta-characteristic* is the foundational property that determines the scope and focus of the entire taxonomy [38]. It serves as a conceptual filter: every dimension and characteristic must demonstrably relate to this overarching property. Without a well-defined meta-characteristic, taxonomy development risks becoming an unbounded exercise that produces dimensions of varying relevance and conceptual coherence.

The meta-characteristic is critical for three reasons. First, it constrains the solution space by specifying which attributes of objects are relevant for classification and which are out of scope. Second, it provides decision criteria for dimension inclusion: a candidate dimension is included only if it influences the meta-characteristic. Third, it ensures that the taxonomy maintains focus on its intended purpose, preventing scope creep during iterative development.

We define our meta-characteristic as:

> *"The capacity of an agentic AI system to generate measurable organizational value — through revenue enhancement, cost reduction, risk mitigation, productivity improvement, decision quality enhancement, safety assurance, or innovation enablement — as determined by its technical capabilities, implementation maturity, and effective integration with human-organizational systems."*

This definition was derived through systematic analysis of our literature corpus and reflects three design principles:

Value-Centric: Unlike purely technical meta-characteristics (e.g., "architectural components" or "computational mechanisms"), our formulation centers on organizational value creation. This orientation ensures that the taxonomy addresses the practical question motivating enterprise adoption: "How do different agentic AI configurations translate into business outcomes?" The seven value categories were synthesized from enterprise AI studies [31, 23, 49] and represent the principal mechanisms through which AI systems create organizational impact.

Multi-Dimensional Scope: The meta-characteristic explicitly covers three factors that shape how value is realized: 1) the technical capabilities of the system (what it can do), 2) the maturity of its implementation (how reliably and consistently it does), and 3) the value it creates (which organizational outcomes it enables). Together, these three factors motivated the development of both a capability taxonomy and a separate, independent maturity model.

Constraining Effect: The meta-characteristic operates as an inclusion filter during dimension identification. For example, a capability like "reasoning sophistication" is included because it directly affects decision quality and task success rates. In contrast, purely technical attributes without demonstrated value linkage (e.g., parameter count) are excluded. This filtering ensures that every taxonomy element connects to practical outcomes.

The meta-characteristic thus guides the entire taxonomy exercise: dimensions capture *what* capabilities influence value creation; characteristics define *how* those capabilities manifest at different levels of sophistication; and the maturity model assesses *how well* systems implement those capabilities in practice.

3.3 Corpus and Reference Objects

Systematic taxonomy development requires two inputs: a *corpus* of source documents providing conceptual grounding and a set of *reference objects* serving as classification targets.

The **corpus** comprises the body of literature from which we extract capability concepts, value linkages, and architectural patterns. We assembled a corpus of 18 documents that span five categories: (1) comprehensive surveys of agentic AI capabilities [1, 43, 45], (2) evaluation and benchmarking frameworks [33, 24], (3) enterprise deployment

Table 2: Document Corpus and Reference Object Summary

Category	Count	Key Sources
Document Corpus		
Agent Capability Surveys	5	Acharya et al., Pati, Piccialli et al. [1, 43, 45]
Evaluation Frameworks	4	Mohammadi et al., Liu et al. (AgentBench) [33, 24]
Enterprise Deployment	3	McKinsey, Sarferaz, Zhang et al. [31, 49, 58]
Safety/Governance	3	OpenAI Preparedness, Morris et al. [40, 34]
Multi-Agent & Human-AI	3	Maldonado et al., Sun et al., Iftikhar et al. [29, 51, 15]
Reference Objects (25 total)		
Task-Executing Agents	14	ReAct, AutoGPT, AIOps, ERP-Integrated
Architectural Patterns	5	Hierarchical, Peer-Based, Cross-Reflection
Collaboration Models	2	Human-Agent Teams, Agent Workflows
Safety Configurations	2	Safeguarded Models, RBAC-Compliant
Evaluation/Infrastructure	2	Agent Evaluator, Tool Registry

studies [31, 49, 58], (4) safety and governance frameworks [40, 34], and (5) multi-agent and human-agent collaboration research [29, 50, 15]. Corpus selection prioritized recency (2023–2025), coverage breadth (technical capabilities through enterprise value), and methodological diversity (academic surveys, industry reports, technical frameworks).

Reference objectsare the concrete agentic AI systems, patterns, and configurations that the taxonomy must classify. Following Nickerson et al.'s guidance that reference objects should reflect the phenomenon's diversity, we selected 25 objects distributed across five object-type categories: 1) task-executing agents, 2) architectural patterns, 3) collaboration models, 4) safety configurations, and 5) evaluation/infrastructure components as summarized in Table 2:

- **Task Executing Agents** (14 objects): Specific agent frameworks and systems documented in the literature, including ReAct [51], AutoGPT [12], and domain-specific agents for AIOps [58], software engineering and ERP integration [49].
- **Architectural Patterns** (5 objects): Reusable design patterns for agent construction, including hierarchical coordination, peer-based collaboration, and cross-reflection patterns [25].
- **Collaboration Models** (2 objects): Human-agent team configurations [15] and workflow-based agent orchestration.
- **Safety Configurations** (2 objects): Deployment scenarios focusing on governance, including protected high-capability models [40] and RBAC-compliant enterprise agents.
- **Evaluation and Infrastructure Frameworks** (2 objects): Meta-patterns that either assess agents (e.g., benchmark harnesses and evaluator components) or provide shared services that enable other agents to function (e.g., tool/agent registries).

This diversity ensures that the taxonomy can classify both concrete system implementations and abstract architectural patterns – supporting theoretical research and practical evaluation alike.

3.4 Ending Conditions

Ending conditions specify when taxonomy development is complete [38]. Without explicit termination criteria, iterative development risks premature conclusion (yielding an incomplete taxonomy) or indefinite refinement (consuming resources without convergent improvement). The Nickerson et al. methodology prescribes both objective conditions (verifiable through inspection) and subjective conditions (requiring informed judgment).

Objective ending conditions are binary criteria verified through inspection of the taxonomy structure and classification results. Adhering to Nickerson et al.'s method [38], We adopted the eight conditions listed in Table 3. In our study these eight conditions are instantiated as: 1) all reference objects examined, 2) no objects merged or split in the final iteration, 3) at least one object classified under every characteristic, 4) no new dimensions added in the final iteration, 5) no dimensions or characteristics merged or split in the final iteration, 6) every dimension conceptually unique, 7) every characteristic unique within its dimension, and 8) each object having a unique combination of characteristics. Table 3 reports how each condition is satisfied for the final taxonomy..

Subjective ending conditions require qualitative assessment against a set of quality criteria to adhere to Nickerson et al.'s taxonomy-development method. We applied the five conditions from Nickerson et al. [38]: *Concise* (limited to 5–9 dimensions per methodological guidance), *Robust* (meaningfully differentiates objects), *Comprehensive* (classifies

Table 3: Ending Conditions Verification (Final Iteration)

Type	Condition	Status
Objective	All objects examined	✓
	No objects merged or split in final iteration	✓
	At least one object per characteristic	✓
	No new dimensions added in final iteration	✓
	No dimensions or characteristics merged or split in final iteration	✓
	Every dimension is unique	✓
	Every characteristic is unique within its dimension	✓
	Each cell combination (object profile) is unique	✓
Subjective	Concise: Limited to 5–9 dimensions	✓
	Robust: Sufficiently differentiates objects of interest	✓
	Comprehensive: All objects can be classified	✓
	Extendible: Accommodates new objects without restructuring	✓
	Explanatory: Explains capability-value relationships	✓

all relevant objects without residual categories), *Extendible* (accommodates adding of new objects without structural revision), and *Explanatory* (dimensions and characteristics explain the underlying obhect).

Our verification process involved systematic checking at each iteration's conclusion. Objective conditions were evaluated by inspecting the classification matrix and the change logs. Subjective conditions required judgment informed by dimension definitions, characteristic coverage distributions, and value-linkage documentation. Three iterations were required before all conditions were satisfied: Iteration 1 established the initial structure but revealed definitional ambiguities; Iteration 2 resolved these by splitting the original "Governance Posture Dimension" (original D8) into separate Compliance (D8) and Safety (D9) dimensions, and hardening definitional boundaries; in Iteration 3, we checked the taxonomy for gaps by testing three possible new dimensions: Planning Horizon, Domain Specialization, and Deployment Context. The review showed that each of these was already covered by the nine existing dimensions or was better handled in the maturity model, confirming that the nine-dimension structure was complete and stable.

4 The CMV Taxonomy

4.1 Framework Architecture

The CMV Framework combines three different views on agentic AI systems: *Capability* (what the system can accomplish), *Maturity* (how well those capabilities are put into action), and *Value* (the results that the technology helps the enterprise achieve). This three-perspectives framework fixes the problems with current methods that just look at technical capabilities, implementation quality, and business value on their own.

Value is intimately related to the agentic AI system capability, maturity and organization's context. Value is only high when the Agentic AI capability, maturity, and context are also high. A system having advanced features (such coordinating several agents) but low implementation maturity will usually be less useful to an organization than a basic feature that is implemented in a reliable and well-governed fashion.

Overall expected value V for a given deployment aggregates these per–dimension contributions and is additionally shaped by organizational context:

$$V = f(\text{Capability Profile}, \text{Maturity Profile}, \text{Context}), \tag{2}$$

where the capability profile specifies which value pathways are accessible, the maturity profile determines how much of that potential is actually realized, and the context term captures factors such as organizational readiness and use–case fit.

Figure 2 illustrates this integration. The capability taxonomy comprises nine dimensions (D1–D9), each with four ordered characteristics representing increasing sophistication. Maturity assessment (five levels, L1–L5) evaluates implementation quality independent from capability type—two systems classified at the same characteristic may differ substantially in reliability, consistency, and optimization. Value realization emerges from dimension-maturity combinations, with each pathway connecting to one or more of seven organizational value categories (formally defined in Section 4.5).

4.2 Capability Dimensions

The nine dimensions emerged through iterative analysis of 25 reference objects drawn from our literature corpus. Each dimension satisfies three criteria: (1) it influences organizational value creation per the meta-characteristic, (2) it is con-

Table 4: Nine Dimensions of the CMV Capability Taxonomy

ID	Dimension	Definition
D1	Reasoning Sophistication	Internal cognitive processes for analyzing information and generating solutions
D2	Tool Integration Depth	Extent of discovering, selecting, invoking, and orchestrating external tools
D3	Memory Persistence	Temporal scope and sophistication of information retention mechanisms
D4	Coordination Paradigm	Structural arrangement for multi-agent labor division and integration
D5	Autonomy Level	Locus of decision authority between humans and agents
D6	Adaptation Mechanism	Means by which agent improves performance over time
D7	Human Collaboration Mode	Nature and depth of interaction patterns with humans
D8	Compliance & Alignment	Degree of operating within policies, regulations, and ethical standards
D9	Safety Architecture	Comprehensiveness of mechanisms protecting against harmful outputs

Table 5: Maturity Level Definitions with Characteristic Baseline Mapping

Level	Label	General Description	Value Realization	Typical Char.
L1	Initial	Ad-hoc, inconsistent, unpredictable outcomes	Minimal; high variance; may be negative	Baseline
L2	Developing	Repeatable in limited contexts; emerging patterns	Emerging; positive ROI in narrow domains	Emerging
L3	Defined	Standardized, consistently applied, documented	Consistent; measurable ROI; enterprise-viable	Emerging–Competent
L4	Managed	Measured, optimized, quantitatively controlled	Significant; predictable outcomes	Competent–Expert
L5	Optimizing	Continuously improved; drives strategic advantage	Transformative; competitive differentiation	Expert

ceptually distinct from other dimensions, and (3) it admits ordered characteristics representing meaningful capability progression. Table 4 presents the dimensions with definitions and value creation logic.

Characteristics refers to the sophistication level within each dimension. Each dimension contains 4 hierarchical characteristics, i.e. C.1 to C.4, representing progressive capability levels, where higher characteristics subsume capabilities of lower levels. Systems are classified with exactly one characteristic per dimension. For example, for D1 (Reasoning), the characteristics are C1.1 Reactive Execution - Simple response, C1.2 Single-Path Deliberation - Chain-of-thought reasoning, C1.3 Multi-Path Exploration - Evaluates multiple alternatives and C1.4 Reflective Refinement - Self-reflection and iterative improvement

Design Decisions. Several boundary-hardening decisions ensure mutual exclusivity. D1 (Reasoning) captures *within-agent* cognition (how an individual agent reasons) while D4 (Cordination) captures *between-agent* coordination (how multiple agents organize); a highly coordinated system of simple reasoners differs from a single sophisticated reasoner. D5 (Autonomy) captures *authority distribution* (who decides) while D7 (Collaboration) captures *interaction patterns* (how they communicate); an autonomous agent may have minimal human interaction, while a tool-mode agent may engage in rich bidirectional dialogue. D8 (Compliance) and D9 (Safety) were separated during Iteration 2 to distinguish policy adherence (organizational/regulatory constraints) from protective mechanisms (harm prevention); an agent may be fully compliant with policies yet lack adversarial robustness, or vice versa.

4.3 Maturity Model Integration

The CMV Framework distinguishes *capability type* (what the system can do) from *implementation maturity* (how well it delivers that capability). This distinction is essential because capability possession differs fundamentally from capability reliability. Two systems both classified at C1.3 (Multi-Path Exploration) may exhibit vastly different performance: one generating multiple reasoning paths inconsistently with unreliable selection, another doing so systematically with measured optimization.

We define five maturity levels following established capability maturity model conventions [44] (Table 5). Maturity assessment occurs per-dimension, producing a nine-element maturity profile $(M_1, M_2, ..., M_9)$ alongside the nine-element capability profile $(C_1, C_2, ..., C_9)$.

Table 6: Characteristics by Dimension (Ordered from Basic to Advanced)

Dim	C_.1 (Basic)	C_.2	C_.3	C_.4 (Advanced)
D1	Reactive Execution	Single-Path Delib.	Multi-Path Explor.	Reflective Refine.
D2	No External Tools	Single-Tool Invoc.	Multi-Tool Orch.	Ecosystem Integr.
D3	Context-Window	Session-Extended	Long-Term Episodic	Comprehensive Arch.
D4	Single-Agent	Hierarchical Coord.	Peer-Based Collab.	Dynamic Coalition
D5	Tool Mode	Guided Execution	Supervised Auton.	Full Autonomy
D6	Static Config.	In-Context Adapt.	Feedback-Driven	Continuous Learning
D7	Output Delivery	Query-Response	Explan. Partnership	Bidirect. Teaming
D8	Unconstrained	Policy-Informed	Compliance-Enforced	Value-Aligned
D9	Unprotected	I/O Filtered	Robustness-Tested	Defense-in-Depth

4.3.1 Independence of Capability and Maturity.

The CMV Framework assesses capability characteristics (C_.1–C_.4) and maturity levels (L1–L5) independently. *Capability characteristics* answer "WHAT can the system do?". They classify the type and sophistication of a capability. *Maturity levels* answer "HOW WELL does it implement that capability?". They assess implementation quality, reliability, and optimization.

This independence is fundamental: two systems with identical capability classifications may differ substantially in maturity.

4.3.2 Baseline Guidance for Characteristic-Maturity Mapping.

Although capability and maturity are defined as independent dimensions, implementation experience shows that characteristic levels tend to correlate with maturity. In practice, Baseline (C.1) capability is most often observed in organizations at Initial (L1) maturity. As maturity increases, organizations are better able to implement higher-level Agentic AI capability characteristics, whereas organizations at lower maturity levels are unlikely to sustain higher characteristic levels. Table 5 summarizes these typical associations, which should be treated as heuristic guidance rather than deterministic mappings:

- **Baseline (C_.1)**: Basic capabilities typically manifest at L1–L2 maturity, reflecting nascent implementation.
- **Emerging (C_.2)**: Functional capabilities typically achieve L2–L3 maturity as implementations stabilize.
- **Competent (C_.3)**: Sophisticated capabilities generally require L3–L4 maturity for reliable operation.
- **Expert (C_.4)**: Advanced capabilities span L3–L5 depending on implementation investment.

These connections show a practical truth: advanced capabilities (C_.4) are not likely to reach their full potential at low maturity (L1-3), and basic capabilities (C_.1) rarely make the investment in optimization worth it to reach L5. Nevertheless, maturity should be deduced from implementation data and evaluated autonomously.

Important Findings. An examination of enterprise deployment patterns uncovers multiple threshold effects. D8 (Compliance) and D9 (Safety) at $\geq$L3 are required for enterprise deployment. Systems that do not meet this standard will face regulatory hurdles and unacceptable risk exposure, no matter what other features they have [33, 40]. Also, moving up in D5 (Autonomy) requires the same level of maturity in D8/D9—having a lot of autonomy without strong governance can create risks that cancel out or even reverse productivity benefits [25, 34]. These interdependencies affect the order in which deployments happen: before giving people more freedom, organizations should make sure they are following the rules and are safe.

Relationship between Maturity and Value. The correlation between maturity level and value realization is nonlinear. At levels 1 to 2, value is generally negative or very low because of the extra time and effort needed to fix mistakes. L3 is the "enterprise-ready" level at which consistent, positive ROI becomes possible. L4–L5 show value-multiplication effects because reliability makes it possible to use the technology in more strategic ways and on a larger scale. This pattern helps us understand why companies often say they do not get anything back from their AI investments [31]: many deployments operate below the L3 threshold required for consistent value capture.

4.4 Characteristic Definitions

Each dimension comprises four characteristics representing ordered levels of capability sophistication. Characteristics are mutually exclusive within dimensions and hierarchically ordered: higher levels generally subsume lower-level capabilities while adding new ones. Table 6 provides a condensed view; complete definitions with example objects appear in Appendix A.

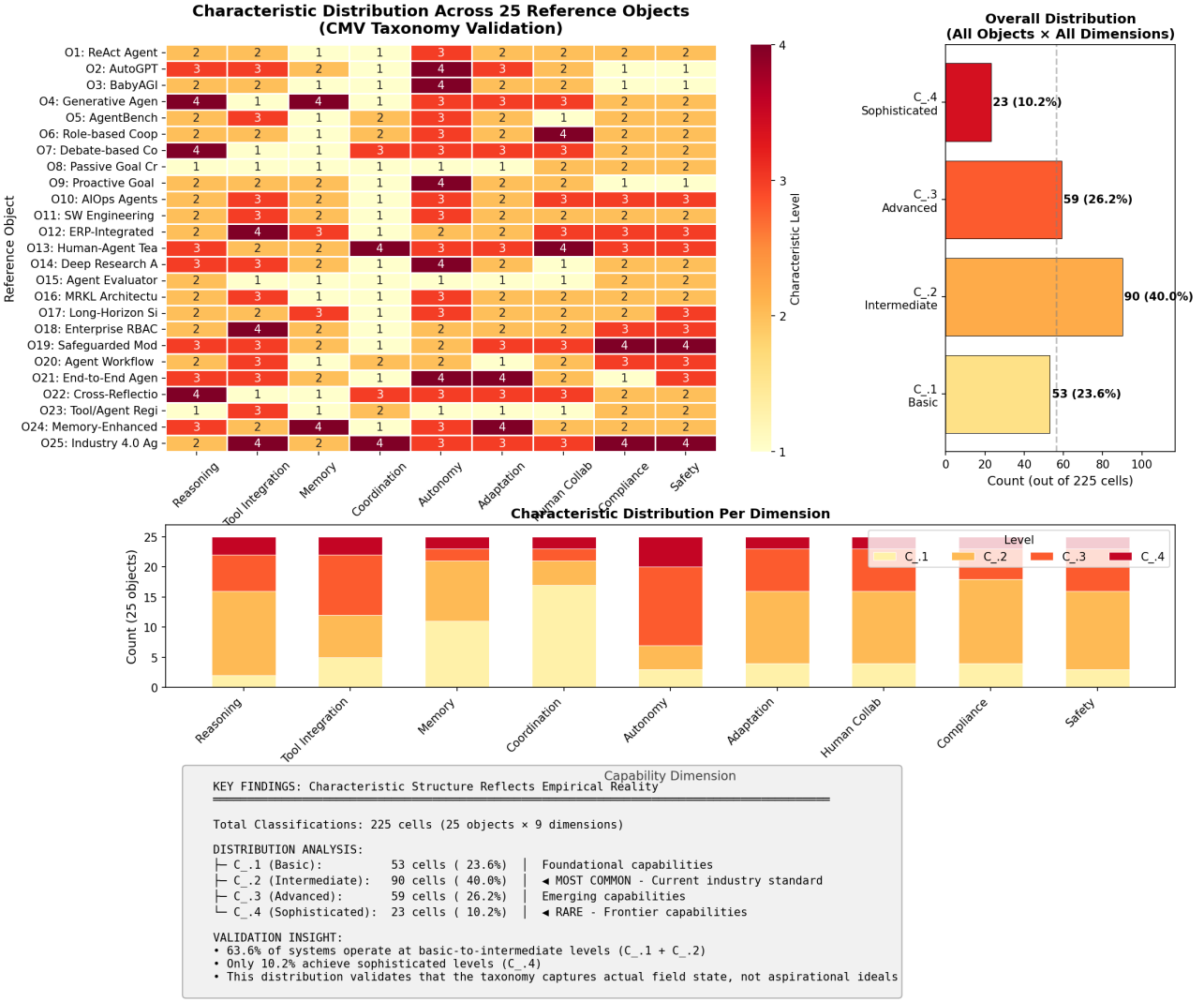

Figure 4: Distribution of characteristic levels across surveyed agentic AI systems.

The characteristic structure mirrors empirical observation: as illustrated in Figure 4, the majority of contemporary agentic AI systems form at C_.2 (intermediate) levels, while advanced characteristics (C_.4) are still uncommon. This distribution validates that the taxonomy captures the actual state of the field rather than aspirational ideals.

4.5 Value Realization Pathways

The CMV Framework grounds capability assessment in organizational value creation. This section defines the seven value categories that constitute the "V" in CMV, then establishes how each capability dimension connects to these value outcomes.

4.5.1 Organizational Value Categories.

We identified seven value categories through systematic synthesis of enterprise AI deployment studies [31, 23, 49], safety frameworks [40, 45], and agent design literature [25]. Table 7 presents each category with its definition, representative measurement indicators, and primary sources. Extended measurement frameworks including the Agentic ROI formula and enterprise benchmarks appear in Appendix D.

These categories are not mutually exclusive; a single agent capability may contribute to multiple value outcomes. For example, D1 (Reasoning) primarily drives Decision Quality, but it also helps Innovation Enablement by finding new solutions. The categories provide a way for different stakeholders to communicate about the values of Agentic AI. For example, technical teams look at capability indicators, while business leaders look at organizational results.

Table 7: Organizational Value Categories: Definitions and Measurement Indicators

Category	Definition	Representative Indicators
Revenue Enhancement	Agent enables new business models, expanded market reach, or new service capabilities through enterprise integration	API success rate; new service count; market reach metrics [17, 49]
Cost Reduction	Agent reduces operational expenses through automation, elimination of training costs, or resource optimization	Function-specific cost decrease; training cost savings; maintenance hours reduction [3]
Risk Mitigation	Agent operates within policies and compliance constraints to prevent harm and ensure regulatory adherence	Policy compliance %; audit pass rate; violation incident count [10, 4]
Productivity Gain	Agent automates workflows, reduces manual effort, accelerates task completion, or enables continuous operation	Time reduction %; human intervention rate; cycle time improvement [24, 52]
Decision Quality	Agent generates more accurate, well-reasoned solutions through improved analysis and pattern recognition	Task success rate; error reduction %; reasoning accuracy [50, 2]
Safety Assurance	Agent implements protective mechanisms against harmful outputs, adversarial attacks, and system failures	Harmful output rate; adversarial resistance %; incident response time [6, 7]
Innovation Enablement	Agent discovers non-obvious solutions through alternative reasoning, collaboration, or continuous capability expansion	Novel solution rate; capability growth rate; transfer success % [20, 8]

4.5.2 Dimension-Value Mapping.

Each capability dimension connects to organizational value through specific mechanisms. For example, Legal reasoning agents like LegalReasoner [55] use multi-hop reasoning to analyze complex cases, this reduces legal analysis errors and lead to better "Decision Quality". Table 8 summarizes these mappings, identifying the primary and secondary value categories for each dimension along with the mechanism by which the value is created.

4.5.3 Value Pathway Principles.

Three principles influence the realization of value:

Pathway Activation. For a dimension to activate its value pathway, it must attain a certain level of capacity. D4 (Coordination) at C4.1 (Single-Agent) has no coordination value by definition because the pathway only works at C4.2 or higher. D2 (Tools) in C2.1 (No External Tools) cannot also improve productivity by automating the tools.

Maturity Gating. For activated pathways to work, they need to be adequately mature to deliver positive returns. For example, D4 (Coordinating) at L1 maturity often yields a negative value due to overhead exceeding the benefits. The L3 threshold represents the point at which most pathways begin to generate consistent positive ROI [31].

Cross-Dimension Dependencies. Some value pathways need more features to work properly. For instance, D5 (Autonomy) needs enough D8/D9 (Compliance/Safety) to keep risk-adjusted values from going down. This reliance is why capability-focused agents, like AutoGPT with D5.4 (Full Autonomy) but D8.1 (Unconstrained)/D9.1 (Unprotected), have a hard time getting businesses to employ them, even though they have great autonomous abilities.

These mappings create testable hypotheses for empirical investigation: *Does the maturity of D7 (collaboration) correlate with user adoption rates? Does the D8 (Compliance)/D9 (Safety) threshold $\geq$ L3 predict the success of the deployment?* Such questions represent directions for future validation.

5 Empirical Validation

Taxonomy validation requires demonstrating that the developed classification structure satisfies both methodological rigor and practical utility [38]. We validate the CMV Taxonomy through three complementary analyzes: (1) verification that all the end conditions of Nickerson et al. are satisfied, (2) evaluation of discriminative power through the classification of

Table 8: Dimension-Value Mapping with Mechanisms

Dim	Primary Value	Secondary	Value Creation Mechanism
D1	Decision Quality	Innovation	Better analysis; reduced errors; novel solutions through multi-path reasoning
D2	Productivity	Revenue	Workflow automation; cross-system connectivity; enterprise integration
D3	Productivity	Decision Quality	Reduced repetition; personalization; historical pattern recognition
D4	Productivity	Innovation	Parallel execution; collective verification; complex problem decomposition
D5	Productivity	Cost Reduction	Reduced oversight burden; continuous 24/7 operation capability
D6	Cost Reduction	Innovation	Eliminated retraining; rapid adaptation; self-improvement over time
D7	Risk Mitigation	Productivity	Trust calibration; effective human-AI teaming; appropriate oversight
D8	Risk Mitigation	Revenue	Regulatory compliance; policy adherence; enterprise deployment enablement
D9	Safety Assurance	Risk Mitigation	Harm prevention; liability protection; deployment trust

Table 9: Ending Conditions Verification Summary

Condition Type	Condition	Verification Evidence
Objective Conditions (7/8 Satisfied)		
	All objects examined	25/25 reference objects classified
	No object mergers/splits	Stable object set across Iterations 2–3
	All characteristics populated	Each of 36 characteristics has $\geq$1 object
	No new dimensions	Dimension set stable in Iteration 3
	No dimension mergers/splits	Structure unchanged in Iteration 3
	Unique dimensions	9 dimensions with distinct conceptual scope
	Unique characteristics	4 mutually exclusive characteristics per dimension
	Unique cell combinations	24/25 unique profiles (96%)[†]
Subjective Conditions (5/5 Satisfied) - rated by a single rater		
	Concise	9 dimensions within 5–9 recommended range
	Robust	96% differentiation across diverse objects
	Comprehensive	All object types classifiable without forcing
	Extendible	Characteristic ranges accommodate new objects
	Explanatory	Value linkages established for all dimensions

[†]O7 (Debate-Based) and O22 (Cross-Reflection) share identical profiles

reference objects, and (3) evaluation of characteristic distributions to evaluate empirical grounding, confirming that the taxonomy reflects observable patterns in deployed systems rather than theoretical ideals.

5.1 Ending Conditions Verification

The Nickerson et al. methodology prescribes eight objective and five subjective ending conditions that collectively determine when iterative refinement should end. Table 9 summarizes the verification results after three development iterations.

The taxonomy achieved structural stability in Iteration 3. Although 24 of 25 objects received unique nine-dimensional profiles (96%), two architectural patterns—O7 (Debate-Based) and O22 (Cross-Reflection)—share identical classifications despite implementing distinct coordination mechanisms. This single overlap represents a potential refinement opportunity (discussed in Section 5.4) rather than a structural deficiency, as patterns differ in the implementation approach rather than in the capability type. A conceptual-to-empirical gap analysis evaluated three candidate dimensions that emerged during corpus analysis:

- **Planning Horizon**. The temporal scope of agent goal-setting (reactive vs. long-term strategic planning) was determined to be adequately captured by D1 (Reasoning).

- **Domain Specialization**. Whether agents are generalist or domain-specific (e.g., legal, medical, financial) was deemed independent to capability assessment; domain context affects *application* rather than *capability type*.

Table 10: Object Classification Matrix (Selected Objects. Complete matrix in Appendix B)

ID	Object	D1	D2	D3	D4	D5	D6	D7	D8	D9
O1	ReAct Agent	.2	.2	.1	.1	.3	.2	.2	.2	.2
O2	AutoGPT	.3	.3	.2	.1	.4	.3	.2	.1	.1
O4	Generative Agents	.4	.1	.4	.1	.3	.3	.3	.2	.2
O10	AIOps Agents	.2	.3	.2	.1	.3	.2	.3	.3	.3
O12	ERP-Integrated	.2	.4	.3	.1	.2	.2	.3	.3	.3
O13	Human-Agent Teams	.3	.2	.2	.4	.3	.3	.4	.3	.3
O19	Safeguarded Models	.3	.3	.2	.1	.2	.3	.3	.4	.4
O25	Industry 4.0	.2	.4	.2	.4	.3	.3	.3	.4	.4

- **Deployment Context**. Operational environment characteristics (i.e. cloud, edge, on-premise) belong to the maturity model's implementation quality assessment rather than the capability taxonomy.

This analysis confirmed that the nine-dimension structure provides sufficient granularity for the meta-characteristic without unnecessary complexity.

5.2 Reference Object Classification

To validate discriminative power, we categorized 25 reference objects into five categories: task-executing agents (14), architectural patterns (5), collaboration models (2), safety configurations (2), and evaluation frameworks (2). Table 10 shows some chosen classifications that illustrate how the taxonomy can distinguish systems based on their varied capability profiles.

Twenty-four of the twenty-five objects received unique nine-dimensional profiles, resulting in 96% discriminative power (see Table 9). Classification uncovered three experimentally substantiated patterns.

1. A business-ready profile emerged between O12 (ERP-Integrated), O19 (Safeguarded Models) and O25 (Industry 4.0), defined by enhanced D8/D9 (compliance and safety in C .3–C .4) paired with moderate D5 (autonomy in C .2–C .3).

2. There was a gap between capability and safety in O2 (AutoGPT). This gap shows great autonomy (D5.4) but low compliance and safety (D8.1, D9.1). This is an example of the risk profile that makes companies reluctant to utilize capability-focused agents.

3. The intensity of D7 (Collaboration) set O13 (Human-Agent Teams) apart from other "objects". It is the only "objects" that can do both D4.4 (Dynamic Coalition) and D7.4 (Bidirectional Teaming), which shows that it was designed to focus on human-AI integration.

The only overlapping pair (O7: Debate-Based Pattern and O22: Cross-Reflection Pattern) has the same nine-dimensional profiles, but they use different coordinating methods: debate-based deliberation and reflective self-improvement. Both accomplish sophisticated reasoning (C1.4), peer-based collaboration (C4.3), and feedback-driven adaptation (C6.3).

5.3 Distribution Analysis

Figure 5) shows the current maturity of agentic AI systems by looking at the characteristic distributions in the 25 reference objects. Two dimensions show notably skewed distributions that need to be looked at.

Memory Persistence (D3): 84% of objects (21/25) are grouped in C3.1 (Context-Window) or C3.2 (Session-Extended), with only four objects achieving long-term or comprehensive memory (C3.3–C3.4). This distribution reflects a genuine capability gap: strong cross-session memory with consolidation mechanisms is still technically challenging [59, 12] and is mostly used in research systems (O4: Generative Agents, O24: Memory-Enhanced Agent) or specialized enterprise contexts (O12: ERP-Integrated Agents, O17: Long-Horizon Simulation).

Coordination (D4): 68% of objects (17/25) function as individual agents (C4.1), while multi-agent coordination patterns (C4.2–C4.4) are exemplified by merely eight objects. This lack of coordination shows that most agentic AI systems deployed work on their own [25]; Complex coordination is still limited to architectural patterns (O6, O7, O22) and specific collaborative situations (O13, O25).

These distributions confirm that the taxonomy reflects practical reality rather than idealistic aspirations. The concentration at intermediate capacity levels (C .2) in D6 (Adaptation), D7 (Human Collaboration), D8 (compliant), and D9 (Safety) demonstrates that contemporary systems have gone beyond rudimentary implementations but have not yet achieved advanced continuously optimizing capabilities [33]. This trend is in line with what many in the business have seen: most AI installations are still in the experimental or pilot stages [31].

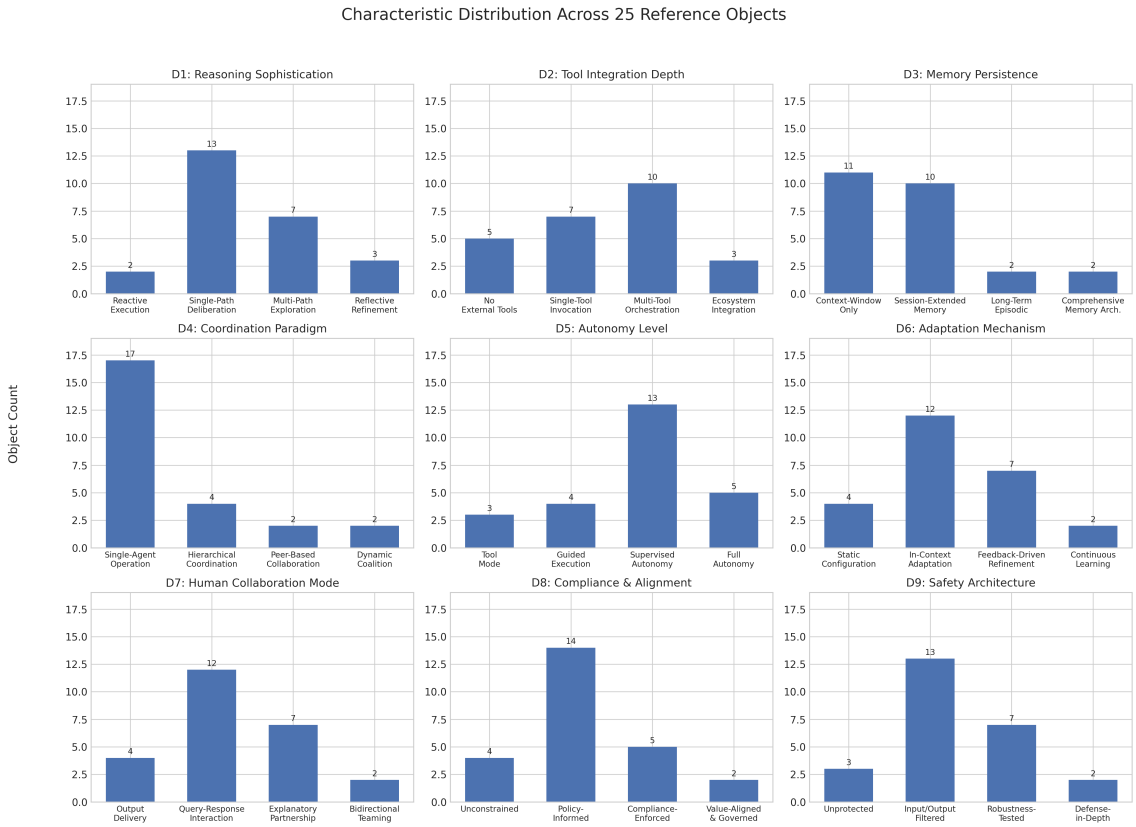

Figure 5: Characteristic distribution across 25 reference objects. Concentrations at C_.1–C_.2 in Memory (D3) and Coordination (D4) reflect nascent development of these capabilities. Advanced characteristics (C_.4) remain rare across most dimensions, consistent with the field's rapid but uneven progress.

5.4 Taxonomy Quality Metrics

In addition to ending conditions, we calculated quantitative metrics to evaluate taxonomy quality (see Appendix F). The *Profile uniqueness* of the profile shows that 24 of 25 objects received different nine-dimensional classifications (96% discrimination).

Table 11 shows how *Distribution entropy* measures the dispersion of characteristics in each dimension. D4 (Coordination) has the lowest entropy (H = 1.38 bits, 69% of maximum), which shows that most activity occurs in C4.1 (Single-Agent). D2 (Tool Integration) has the highest entropy (H = 1.87 bits, 94% of maximum), which means that the features are evenly spread out. These entropy measures support what we saw qualitatively: Memory (D3) and Coordination (D4) indicate authentic capability deficiencies in existing systems, while Tool Integration (D2) and Autonomy (D5) exhibit a more uniform adoption across the capability range.

We used Cramer's V, a statistical measure of the relationship between categorical variables that goes from 0 (total independence) to 1 (perfect association), to see how independent the dimensions were. We calculated Cramér's V for all 36 unique dimension pairs across the 25 reference objects to assess whether dimensions represent separate facets of agentic AI capabilities or demonstrate redundant overlap. Most dimension pairs show a weak to moderate relationship (V ¡ 0.50), which shows that they are conceptually different and that the taxonomy structure is correct. The most significant correlation exists between D8 (Compliance) and D9 (Safety), with V = 0.90 (p ¡ 0.001), reflecting their common emphasis on governance. This association affirms rather than contradicts their distinction: systems engineered with stringent compliance frameworks typically also incorporate resilient safety designs, while the constructs persist as conceptually separate (policy adherence versus harm prevention).

The Spearman correlation analysis indicates a statistically significant negative relationship between D5 (Autonomy) and D8 (Compliance): $\rho = -0.59$ ($p = 0.002$). This finding quantitatively substantiates the capability-safety gap pattern identified qualitatively in Section 5.2 — systems characterized by high autonomy (e.g., O2: AutoGPT at C5.4) generally demonstrate weaker compliance mechanisms (C8.1), whereas enterprise-ready systems (e.g., O19: Safeguarded Models) combine moderate autonomy with robust governance. This empirical association substantiates the CMV Framework's cross-dimension dependency principle: enhancing autonomy without corresponding governance investment incurs de-

Table 11: Distribution Entropy by Dimension (Maximum $H_{\max} = 2.0$ bits)

Dim	Name	H (bits)	$H/H_{\max}$	Interpretation
D1	Reasoning	1.62	0.81	Moderate
D2	Tool Integration	1.87	0.94	Balanced
D3	Memory	1.63	0.82	Moderate
D4	Coordination	1.38	0.69	Skewed
D5	Autonomy	1.75	0.87	Balanced
D6	Adaptation	1.74	0.87	Balanced
D7	Human Collaboration	1.74	0.87	Balanced
D8	Compliance	1.65	0.82	Moderate
D9	Safety	1.66	0.83	Moderate

ployment risk.

5.5 Reliability Considerations

In this research, reference objects were drawn from the academic literature and industry frameworks. This corpus-bounded scope establishes *structural validity* for the taxonomy to classify diverse agentic AI configurations but does not confirm whether the capability profiles accurately forecast organizational value outcomes. The phased approach to establish predictive validity through retrospective case studies and prospective field studies, detailed in Section 6.3, represents the essential direction for extended validation.

6 Discussion

6.1 Theoretical Contributions

The CMV Taxonomy propels agentic AI research with three primary contributions. First, the value-centric meta-characteristic shifts the focus of classification from technical functions to organizational results. Existing agentic AI Taxonomies group agents by architectural parts or technical procedures [1, 33]. By making it possible to group agentic AI based on their potential to create values, the CMV Taxonomy links technical assessment to business decision-making.

Second, independent assessment of capability type and implementation maturity offers practical analytics compared to current frameworks. This division recognizes a key difference: having a capability is not the same as being able to derive values from it.

Third, the links between dimensions and values make it possible to test hypotheses in real-world research: *Is there a link between D7 (Human Collaboration) maturity and how many people utilize it? Does the D8/D9 $\geq$ L3 a good way to estimate how well an enterprise will deploy?* These assumptions establish a study framework for correlating agentic AI capabilities with quantifiable organizational results.

6.2 Practical Implications

The CMV Framework offers practical guidance for practitioners and enterprise architects across the entire deployment lifecycle.

System Selection. Capability profiles allow for a structured comparison of agent options versus requirement specifications across nine specified dimensions. For example, when choosing between AutoGPT and a Safeguarded Model for customer service, the CMV profile shows that AutoGPT has a lot of freedom (D5.4) but not much compliance (D8.1). On the other hand, Safeguarded Models have moderate freedom (D5.2) and strong governance (D8.4, D9.4). This means that the choice should be made for the latter in regulated industries.

Gap Analysis. An assessment of maturity finds areas of capacity that need more investment to grow. The D8/D9 $\geq$L3 criterion gives clear directions for the order in which to make investments. For example, a company using a ReAct Agent (D8.2, D9.2) for financial processes finds a governance gap. The maturity model shows what particular changes need to be made to meet the L3 barrier before the enterprise may be deployed.

Risk Management. The capability-safety gap pattern (high D5 autonomy with low D8/D9 governance) might help you figure out how risky it is to deploy. For example, AutoGPT's profile (D5.4: Full Autonomy, D8.1: Unconstrained, D9.1: Unprotected) quickly shows that there is a risk of deployment; the organization can need extra safety measures or human monitoring before the product is used in production.

Value Forecasting. Dimension-value mappings help build business cases by linking technical decisions to projected organizational outcomes across seven value categories. For example, moving D2 (Tool Integration) from C2.2 to C2.4

is an investment that leads to more revenue through the connectivity of the enterprise ecosystem, making the costs of integration infrastructure worth it.

Value Projection. Dimension-value mappings help build business cases by linking technical decisions to projected organizational outcomes across seven value categories. For example, moving D2 (Tool Integration) from C2.2 to C2.4 is an investment that leads to more revenue through enterprise ecosystem connectivity, which makes the costs of integration infrastructure worth it.

Roadmap Planning. The maturity model lays out clear avenues for growth and sets clear standards on how to get there. For example, a company in L2 (Developing) can plan a phased transition to L3 (Defined) because they know that L3 needs standardized processes, documented procedures, and proactive risk management in all areas.

Systematic Assessment. Appendix E provides a structured Assessment Scorecard Template that professionals may use to systematically rate agentic AI systems on all nine dimensions, check maturity levels, confirm deployment thresholds, and figure out what measures should be taken first to make things better. For instance, a team using the scorecard to evaluate an ERP-Integrated Agent keeps a record of its D2.4 (Ecosystem Integration), D3.3 (Long-Term Memory), and D8.3/D9.3 governance posture, which can be checked by an outside party.

6.3 Limitations and Future Validation Directions

Several limitations constrain the current work, each suggesting specific directions for future research.

Corpus Scope. The document corpus primarily comprises sources from academic settings. For highly regulated areas such as healthcare, financial services, and critical infrastructure, specific guidance to that industry or sector may need to be expanded by talking to domain experts and analyzing the regulatory environment.

Classification Reliability. *Future validation should prioritize surveys to establish reliability through industry-specific classification by multiple domain experts.* Specifically, we recommend: (1) identify 3–5 independent raters with agentic AI expertise to classify a representative subset of reference objects; (2) compute Cohen's kappa (for pairwise agreement) or Fleiss' kappa (for multi-rater agreement) for each dimension; and (3) refine characteristic definitions where survey outcome falls below acceptable thresholds ($\kappa < 0.6$). This survey reliability assessment would strengthen confidence in the taxonomy's practical applicability and identify dimensions requiring definitional refinement.

Distribution Validation. The distribution analysis presented in Section 5.3 shows what agentic AI can do as of early 2025. These distributions will change as the field changes rapidly. Subsequent research should replicate the distribution study using the recent literature and newly launched systems to monitor the trajectories of capacity growth. Longitudinal tracking would show if current capacity gaps (Memory in C3.1–C3.2, Coordination in C4.1) are getting less and how quickly they are doing so. This would provide a solid basis for technological roadmap planning in real life.

7 Conclusion

This paper presents the Capability–Maturity-Value (CMV) Taxonomy, a structured framework for categorizing agentic AI systems based on their ability to create organizational value. Following Nickerson et al. methodology, we developed nine capability dimensions with 36 characteristics. We also developed a maturity model that measures the quality of implementation regardless of the type of capability. The classification of 25 reference objects has discriminative power and uncovers experimentally validated patterns, including enterprise-ready profiles and capability-safety gaps.

The CMV Framework gives researchers a defined vocabulary for classifying agentic AI, practitioners with meaningful advice for designing and testing systems, and business leaders with a basis for setting investment priorities. The taxonomy makes it possible to make deployment decisions based on evidence in a time when agentic AI capabilities are growing quickly by grouping capabilities by value creation instead of a technical function.

Future work will pursue three directions: (1) *Predictive validation* through retrospective case studies and prospective field studies correlating CMV profiles with organizational results, following the phased approach outlined in Section 6.3; (2) *Reliability enhancement* through multi-rater classification studies establishing inter-rater reliability (Cohen's/Fleiss' kappa) for each dimension; and (3) *Automated assessment* developing LLM-based evaluation tools ("CMV-Judge") that operationalize the Assessment Scorecard Template (Appendix E) for scalable capability and maturity scoring, complemented by domain extensions for healthcare, finance, and manufacturing contexts with sector-specific compliance and safety requirements.

8 Acknowledgments

The author acknowledges the guidance of Prof. TA Nguyen Binh Duong and Prof. Zhu Feida in shaping this research proposal. In addition, AI-assisted tools were used in the preparation of this manuscript. ChatGPT (OpenAI) and Claude (Anthropic) were used for LaTeX formatting assistance and language check. Google Gemini was used to generate the

figures presented in this paper. The authors assume full responsibility for the content and have verified all AI-generated outputs.

References

[1] Acharya, D.B., Kuppan, K., Divya, B.: Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. IEEE Access **13**, 18912–18936 (2025). https://doi.org/10.1109/ACCESS.2025.3532853

[2] Akin, M., Sir, G.D.B., Karadag, A.A., Cercioglu, H.: A Multi-Criteria Comparison of Large Language Model Powered Assistants in Pre-Research Studies for the Academia. IEEE Access pp. 1–1 (2025). https://doi.org/10.1109/ACCESS.2025.3586502

[3] Alecsoiu, O.R., Faruqui, N., Panagoret, A.A., Ionuț, C.A., Panagoret, D.M., Nitu, R.V., Mutu, M.A.: EcoptiAI: E-Commerce Process Optimization and Operational Cost Minimization Through Task Automation Using Agentic AI. IEEE Access **13**, 70254–70268 (2025). https://doi.org/10.1109/ACCESS.2025.3560549

[4] Bukar, U.A., Sayeed, M.S., Fatimah Abdul Razak, S., Yogarayan, S., Sneesl, R.: Decision-Making Framework for the Utilization of Generative Artificial Intelligence in Education: A Case Study of ChatGPT. IEEE Access **12**, 95368–95389 (2024). https://doi.org/10.1109/ACCESS.2024.3425172

[5] CrewAI, Inc.: CrewAI: Framework for orchestrating role-playing, autonomous AI agents. `https://github.com/crewAIInc/crewAI` (2024), accessed: 2025-08-01

[6] Cummings, M.L.: Identifying AI Hazards and Responsibility Gaps. IEEE Access **13**, 54338–54349 (2025). https://doi.org/10.1109/ACCESS.2025.3552200

[7] Deng, Z., Guo, Y., Han, C., Ma, W., Xiong, J., Wen, S., Xiang, Y.: AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. ACM Comput. Surv. **57**(7), 1–36 (Jul 2025). https://doi.org/10.1145/3716628

[8] Gao, H.a., Geng, J., Hua, W., Hu, M., Juan, X., Liu, H., Liu, S., Qiu, J., Qi, X., Wu, Y., Wang, H., Xiao, H., Zhou, Y., Zhang, S., Zhang, J., Xiang, J., Fang, Y., Zhao, Q., Liu, D., Ren, Q., Qian, C., Wang, Z., Hu, M., Wang, H., Wu, Q., Ji, H., Wang, M.: A Survey of Self-Evolving Agents: On Path to Artificial Super Intelligence (Aug 2025). https://doi.org/10.48550/arXiv.2507.21046

[9] Ge, S., Sun, Y., Cui, Y., Wei, D.: An Innovative Solution to Design Problems: Applying the Chain-of-Thought Technique to Integrate LLM-Based Agents With Concept Generation Methods. IEEE Access **13**, 10499–10512 (2025). https://doi.org/10.1109/ACCESS.2024.3494054

[10] Getir Yaman, S., Ribeiro, P., Cavalcanti, A., Calinescu, R., Paterson, C., Townsend, B.: Specification, validation and verification of social, legal, ethical, empathetic and cultural requirements for autonomous agents. Journal of Systems and Software **220**, 112229 (Feb 2025). https://doi.org/10.1016/j.jss.2024.112229

[11] Google Cloud: Google Agent Development Kit (ADK): Enterprise-Grade Infrastructure for Building Production-Ready AI Agents. `https://cloud.google.com/products/agent-development-kit` (2025), accessed: 2025-08-01

[12] Hatalis, K., Christou, D., Myers, J., Jones, S., Lambert, K., Amos-Binks, A., Dannenhauer, Z., Dannenhauer, D.: Memory Matters: The Need to Improve Long-Term Memory in LLM-Agents. Proceedings of the AAAI Symposium Series **2**(1), 277–280 (2023). https://doi.org/10.1609/aaaiss.v2i1.27688

[13] He, G., Demartini, G., Gadiraju, U.: Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp. 1–22. No. 414 in CHI Conference Proceedings, Association for Computing Machinery, New York, NY, USA (Apr 2025)

[14] Holter, S., El-Assady, M.: Deconstructing Human-AI Collaboration: Agency, Interaction, and Adaptation. Computer Graphics Forum **43**(3), e15107 (2024). https://doi.org/10.1111/cgf.15107

[15] Iftikhar, R., Chiu, Y.T., Khan, M.S., Caudwell, C.: Human–Agent Team Dynamics: A Review and Future Research Opportunities. IEEE Transactions on Engineering Management **71**, 10139–10154 (2024). https://doi.org/10.1109/TEM.2023.3331369

[16] Keskin, Z., Joosten, D., Klasen, N., Huber, M., Liu, C., Drescher, B., Schmitt, R.H.: LLM-Enhanced Human–Machine Interaction for Adaptive Decision-Making in Dynamic Manufacturing Process Environments. IEEE Access **13**, 44650–44661 (2025). https://doi.org/10.1109/ACCESS.2025.3549529

[17] Kostis, A., Lidström, J., Nair, S., Holmström, J.: Too Much AI Hype, Too Little Emphasis on Learning? Entrepreneurs Designing Business Models Through Learning-by-Conversing With Generative AI. IEEE Transactions on Engineering Management **71**, 15278–15291 (2024). https://doi.org/10.1109/TEM.2024.3484750

[18] LangChain, Inc.: LangChain: Building applications with LLMs through composability. `https://github.com/langchain-ai/langchain` (2024), accessed: 2025-08-01

[19] Li, J., Gao, Y., Yang, Y., Bai, Y., Zhou, X., Li, Y., Sun, H., Liu, Y., Si, X., Ye, Y., Wu, Y., Lin, Y., Xu, B., Ren, B., Feng, C., Huang, H.: Fundamental Capabilities and Applications of Large Language Models: A Survey. ACM Comput. Surv. p. 3735632 (May 2025). https://doi.org/10.1145/3735632

[20] Lin, X., Wang, T., Sheng, F.: The Relationship Between AI-Adoption Intensity and Firm Innovation Performance: The Role of AI Affordances. IEEE Transactions on Engineering Management **72**, 2267–2278 (2025). https://doi.org/10.1109/TEM.2025.3572042

[21] Liscio, E., Siebert, L.C., Jonker, C.M., Murukannaiah, P.K.: Value Preferences Estimation and Disambiguation in Hybrid Participatory Systems. Journal of Artificial Intelligence Research **82**, 819–850 (Feb 2025). https://doi.org/10.1613/jair.1.14958

[22] Liu, S., Miao, H., Li, Z., Olson, M., Pascucci, V., Bremer, P.T.: AVA: Towards Autonomous Visualization Agents through Visual Perception-Driven Decision-Making. Comput Graphics Forum **43**(3) (2024). https://doi.org/10.1111/cgf.15093

[23] Liu, W., Qin, J., Huang, X., Zeng, X., Xi, Y., Lin, J., Wu, C., Wang, Y., Shang, L., Tang, R., Lian, D., Yu, Y., Zhang, W.: The Real Barrier to LLM Agent Usability is Agentic ROI (May 2025). https://doi.org/10.48550/arXiv.2505.17767

[24] Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., Tang, J.: AgentBench: Evaluating LLMs as Agents (Oct 2025). https://doi.org/10.48550/arXiv.2308.03688

[25] Liu, Y., Lo, S.K., Lu, Q., Zhu, L., Zhao, D., Xu, X., Harrer, S., Whittle, J.: Agent design pattern catalogue: A collection of architectural patterns for foundation model based agents. Journal of Systems and Software **220**, 112278 (Feb 2025). https://doi.org/10.1016/j.jss.2024.112278

[26] Lo, J.H., Huang, H.P., Chen, Y.C., Chen, J.H.: Memory Robot Design: A New Perspective From Human Brain Model and Large Language Model. IEEE Access **13**, 28539–28549 (2025). https://doi.org/10.1109/ACCESS.2025.3538889

[27] Luo, Y., Zheng, J., Yang, Z., Chen, N., Wu, D.: Pleno-Alignment Framework for Stock Trend Prediction. IEEE Transactions on Neural Networks and Learning Systems pp. 1–15 (2025). https://doi.org/10.1109/TNNLS.2025.3561811

[28] Mahad Malik, M., Altamimi, A., Ali Abbas Kazmi, S., Khan, Z.A., Waleed Ansari, M., Mujahid, K., Gao, J.: A Full-Fledged, Multi-Agent System Representing the Architecture of Smart Cities by Balancing Energy With Optimal Electricity Forecasting, Integrating Individual Comfort, and Extracting Financial Gains. IEEE Access **12**, 172280–172296 (2024). https://doi.org/10.1109/ACCESS.2024.3497752

[29] Maldonado, D., Cruz, E., Abad Torres, J., Cruz, P.J., Gamboa Benitez, S.d.P.: Multi-Agent Systems: A Survey About Its Components, Framework and Workflow. IEEE Access **12**, 80950–80975 (2024). https://doi.org/10.1109/ACCESS.2024.3409051

[30] Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., Uchibe, E., Morimoto, J.: Deep learning, reinforcement learning, and world models. Neural Networks **152**, 267–275 (Aug 2022). https://doi.org/10.1016/j.neunet.2022.03.037

[31] McKinsey: The State of AI: Global Survey 2025 | McKinsey. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (2025)

[32] Mindom, P.S.N., Nikanjam, A., Khomh, F.: Harnessing pre-trained generalist agents for software engineering tasks. Empir Software Eng **30**(1), 1–53 (Jan 2025). https://doi.org/10.1007/s10664-024-10597-8

[33] Mohammadi, M., Li, Y., Lo, J., Yip, W.: Evaluation and Benchmarking of LLM Agents: A Survey. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. pp. 6129–6139 (Aug 2025). https://doi.org/10.1145/3711896.3736570

[34] Morris, M.R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., Legg, S.: Position: Levels of AGI for operationalizing progress on the path to AGI. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, vol. 235, pp. 36308–36321. JMLR.org, Vienna, Austria (Jul 2024)

[35] Mu, C., Guo, H., Chen, Y., Shen, C., Hu, D., Hu, S., Wang, Z.: Multi-agent, human–agent and beyond: A survey on cooperation in social dilemmas. Neurocomputing **610**, 128514 (Dec 2024). https://doi.org/10.1016/j.neucom.2024.128514

[36] Mujtaba, G., Khowaja, S.A., Dev, K.: EdgeAIGuard: Agentic LLMs for Minor Protection in Digital Spaces. IEEE Internet of Things Journal pp. 1–1 (2025). https://doi.org/10.1109/JIOT.2025.3574961

[37] Narula, S., Ghasemigol, M., Carnerero-Cano, J., Minnich, A., Lupu, E., Takabi, D.: Exploring Research and Tools in AI Security: A Systematic Mapping Study. IEEE Access **13**, 84057–84080 (2025). https://doi.org/10.1109/ACCESS.2025.3567195

[38] Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. European Journal of Information Systems **22**(3), 336–359 (May 2013). https://doi.org/10.1057/ejis.2012.26

[39] Omran Almagrabi, A., Khan, R.A.: Optimizing Secure AI Lifecycle Model Management With Innovative Generative AI Strategies. IEEE Access **13**, 12889–12920 (2025). https://doi.org/10.1109/ACCESS.2024.3491373

[40] OpenAI: Our updated Preparedness Framework. https://openai.com/index/updating-our-preparedness-framework/

[41] Pardo, P., Strasser, C.: The Goal after Tomorrow: Offline Goal Reasoning with Norms. Journal of Artificial Intelligence Research **80**, 1703–1759 (Aug 2024). https://doi.org/10.1613/jair.1.15566

[42] Park, J.H., Madisetti, V.K.: CAPRI: A Context-Aware Privacy Framework for Multi-Agent Generative AI Applications. IEEE Access **13**, 43168–43177 (2025). https://doi.org/10.1109/ACCESS.2025.3549312

[43] Pati, A.K.: Agentic AI: A Comprehensive Survey of Technologies, Applications, and Societal Implications. IEEE Access pp. 1–1 (2025). https://doi.org/10.1109/ACCESS.2025.3585609

[44] Paulk, M.C., Curtis, B., Chrissis, M.B., Weber, C.V.: Capability Maturity Model for Software, Version 1.1:. Tech. rep., Defense Technical Information Center, Fort Belvoir, VA (Feb 1993). https://doi.org/10.21236/ADA263403

[45] Piccialli, F., Chiaro, D., Sarwar, S., Cerciello, D., Qi, P., Mele, V.: AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0. Expert Systems with Applications **291**, 128404 (Oct 2025). https://doi.org/10.1016/j.eswa.2025.128404

[46] Raju, N.V.D.S.S.V.P., Faruqui, N., Patel, N., Alecsoiu, O.R., Thatoi, P., Alyami, S.A., Azad, AKM.: LegalMind: Agentic AI-Driven Process Optimization and Cost Reduction in Legal Services Using DeepSeek. IEEE Access pp. 1–1 (2025). https://doi.org/10.1109/ACCESS.2025.3586781

[47] Sampath, A.N., Thakur, A., Krishnan, S.: Agentic Reasoning for Social Event Extrapolation: Integrating Knowledge Graphs and Language Models. IEEE Access pp. 1–1 (2025). https://doi.org/10.1109/ACCESS.2025.3612015

[48] Sanneman, L., Shah, J.A.: Validating metrics for reward alignment in human-autonomy teaming. Computers in Human Behavior **146**, 107809 (Sep 2023). https://doi.org/10.1016/j.chb.2023.107809

[49] Sarferaz, S.: Implementing Generative AI Into ERP Software. IEEE Access **13**, 73342–73354 (2025). https://doi.org/10.1109/ACCESS.2025.3564133

[50] Sun, C., Huang, S., Pompili, D.: LLM-Based Multi-Agent Decision-Making: Challenges and Future Directions. IEEE Robotics and Automation Letters **10**(6), 5681–5688 (Jun 2025). https://doi.org/10.1109/LRA.2025.3562371

[51] Sun, J., Zheng, C., Xie, E., Liu, Z., Chu, R., Qiu, J., Xu, J., Ding, M., Li, H., Geng, M., Wu, Y., Wang, W., Chen, J., Yin, Z., Ren, X., Fu, J., He, J., Wu, Y., Liu, Q., Liu, X., Li, Y., Dong, H., Cheng, Y., Zhang, M., Heng, P.A., Dai, J., Luo, P., Wang, J., Wen, J.R., Qiu, X., Guo, Y., Xiong, H., Liu, Q., Li, Z.: A Survey of Reasoning with Foundation Models: Concepts, Methodologies, and Outlook. ACM Comput. Surv. p. 3729218 (Apr 2025). https://doi.org/10.1145/3729218

[52] Toprani, D., Madisetti, V.K.: LLM Agentic Workflow for Automated Vulnerability Detection and Remediation in Infrastructure-as-Code. IEEE Access **13**, 69175–69181 (2025). https://doi.org/10.1109/ACCESS.2025.3560911

[53] Wachowiak, L., Coles, A., Canal, G., Celiktutan, O.: A Taxonomy of Explanation Types and Need Indicators in Human–Agent Collaborations. Int J of Soc Robotics **16**(7), 1681–1692 (Jul 2024). https://doi.org/10.1007/s12369-024-01148-8

[54] Wang, S., Zhu, Y., Liu, H., Zheng, Z., Chen, C., Li, J.: Knowledge Editing for Large Language Models: A Survey. ACM Comput. Surv. **57**(3), 1–37 (Mar 2025). https://doi.org/10.1145/3698590

[55] Wang, X., Zhang, X., Hoo, V., Shao, Z., Zhang, X.: LegalReasoner: A Multi-Stage Framework for Legal Judgment Prediction via Large Language Models and Knowledge Integration. IEEE Access **12**, 166843–166854 (2024). https://doi.org/10.1109/ACCESS.2024.3496666

[56] Wang, Y., Wu, Z., Yao, J., Su, J.: TDAG: A multi-agent framework based on dynamic Task Decomposition and Agent Generation. Neural Networks **185**, 107200 (May 2025). https://doi.org/10.1016/j.neunet.2025.107200

[57] 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.: AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In: Proceedings of the Fourth International Conference on Automated Machine Learning (AutoML 2024). pp. 1–20. PMLR (2024)

[58] Zhang, L., Jia, T., Jia, M., Wu, Y., Liu, A., Yang, Y., Wu, Z., Hu, X., Yu, P., Li, Y.: A Survey of AIOps in the Era of Large Language Models. ACM Comput. Surv. **58**(2), 1–35 (Jan 2026). https://doi.org/10.1145/3746635

[59] Zhang, Z., Dai, Q., Bo, X., Ma, C., Li, R., Chen, X., Zhu, J., Dong, Z., Wen, J.R.: A Survey on the Memory Mechanism of Large Language Model based Agents. ACM Trans. Inf. Syst. p. 3748302 (Jul 2025). https://doi.org/10.1145/3748302

[60] Zhong, W., Guo, L., Gao, Q., Ye, H., Wang, Y.: MemoryBank: Enhancing Large Language Models with Long-Term Memory. Proceedings of the AAAI Conference on Artificial Intelligence **38**(17), 19724–19731 (Mar 2024). https://doi.org/10.1609/aaai.v38i17.29946

A Complete Characteristic Definitions

Table 12 provides complete definitions for all 36 characteristics.

Table 12: Complete Characteristic Definitions by Dimension

ID	Characteristic	Definition	Primary Value
D1: Reasoning Sophistication			
C1.1	Reactive Execution	Agent executes predefined responses or simple rule-based logic without explicit reasoning chains	Cost Reduction
C1.2	Single-Path Deliberation	Agent employs chain-of-thought or sequential reasoning to solve problems through one deliberative pathway	Decision Quality
C1.3	Multi-Path Exploration	Agent generates and evaluates multiple alternative reasoning paths, selecting optimal solutions through comparison	Innovation
C1.4	Reflective Refinement	Agent iteratively improves reasoning through self-reflection or cross-agent verification, correcting errors and refining outputs	Decision Quality
D2: Tool Integration Depth			

Continued on next page

Table 12 continued

ID	Characteristic	Definition	Primary Value
C2.1	No External Tools	Agent operates solely through language generation without invoking external tools or APIs	Cost Reduction
C2.2	Single-Tool Invocation	Agent can invoke individual tools or functions but does not coordinate multiple tools in sequence	Productivity
C2.3	Multi-Tool Orchestration	Agent coordinates multiple tools in planned sequences to complete complex workflows	Productivity
C2.4	Ecosystem Integration	Agent integrates with enterprise systems through standardized interfaces, managing authentication, access control, and cross-system workflows	Revenue
D3: Memory Persistence			
C3.1	Context-Window Only	Agent memory limited to current context window; no persistence across sessions	Cost Reduction
C3.2	Session-Extended Memory	Agent maintains working memory across extended interactions within a session but no cross-session persistence	Productivity
C3.3	Long-Term Episodic Memory	Agent stores and retrieves past experiences and interactions across sessions, enabling historical pattern recognition	Decision Quality
C3.4	Comprehensive Architecture	Agent implements multiple memory types (episodic, semantic, procedural) with consolidation mechanisms for sustained learning	Innovation
D4: Coordination Paradigm			
C4.1	Single-Agent Operation	Agent operates independently without coordination with other agents	Cost Reduction
C4.2	Hierarchical Coordination	Agents organized in role-based hierarchies with designated planners, workers, and evaluators	Productivity
C4.3	Peer-Based Collaboration	Agents collaborate as equals through voting, debate, or negotiation mechanisms	Decision Quality
C4.4	Dynamic Coalition	Agents dynamically form and dissolve coalitions based on task requirements and capability matching	Innovation
D5: Autonomy Level			
C5.1	Tool Mode	Agent acts only on explicit instructions, providing outputs for human decision-making	Risk Mitigation
C5.2	Guided Execution	Agent executes defined workflows with human checkpoints and approval gates	Productivity
C5.3	Supervised Autonomy	Agent makes independent decisions within bounded domains, with human oversight for exceptions	Productivity
C5.4	Full Autonomy	Agent independently sets and pursues goals with minimal human intervention	Productivity
D6: Adaptation Mechanism			
C6.1	Static Configuration	Agent operates with fixed capabilities; no adaptation during or across deployments	Cost Reduction

Continued on next page

Table 12 continued

ID	Characteristic	Definition	Primary Value
C6.2	In-Context Adaptation	Agent adapts behavior through prompt engineering, few-shot examples, or retrieval augmentation without parameter updates	Cost Reduction
C6.3	Feedback-Driven Refinement	Agent improves through verbal feedback, reflection, or environmental signals within deployment	Decision Quality
C6.4	Continuous Learning	Agent undergoes parameter updates or skill library expansion through fine-tuning, RLHF, or procedural memory accumulation	Innovation
D7: Human Collaboration Mode			
C7.1	Output Delivery	Agent produces outputs for human consumption without interactive collaboration	Productivity
C7.2	Query-Response Interaction	Agent engages in turn-based dialogue, responding to human queries and instructions	Productivity
C7.3	Explanatory Partnership	Agent provides explanations of its reasoning and decisions, enabling human understanding and oversight	Risk Mitigation
C7.4	Bidirectional Teaming	Agent and human engage in mutual transparency, with shared mental models and collaborative decision-making	Productivity
D8: Compliance & Alignment Posture			
C8.1	Unconstrained	Agent operates without explicit policy constraints or alignment mechanisms	Risk Exposure
C8.2	Policy-Informed	Agent configured with basic alignment guidance but no enforcement mechanisms	Risk Mitigation
C8.3	Compliance-Enforced	Agent operates within defined access controls, audit trails, and regulatory constraints	Risk Mitigation
C8.4	Value-Aligned & Governed	Agent demonstrates sustained alignment with organizational values through RLHF, constitutional AI, or ongoing governance oversight	Risk Mitigation
D9: Safety Architecture			
C9.1	Unprotected	No explicit safety mechanisms beyond base model training	Safety Risk
C9.2	Input/Output Filtered	Basic guardrails filtering harmful inputs and outputs	Safety Assurance
C9.3	Robustness-Tested	Agent tested for adversarial robustness with graceful degradation mechanisms	Safety Assurance
C9.4	Defense-in-Depth	Multi-layered safety including guardrails, robustness, interruptibility, monitoring, and human override capabilities	Safety Assurance

B Complete Object Classification Matrix

Table 13 presents the complete classification of all 25 reference objects with source citations.

Table 13: Complete Object Classification Matrix (25 Objects)

Type codes: TA=Task-Executing Agent, AP=Architectural Pattern, EF=Evaluation Framework, CM=Collaboration Model, SC=Safety Configuration, IC=Infrastructure Component

ID	Type	Object	D1	D2	D3	D4	D5	D6	D7	D8	D9	Source
O1	TA	ReAct Agent	.2	.2	.1	.1	.3	.2	.2	.2	.2	[51, 33]
O2	TA	AutoGPT	.3	.3	.2	.1	.4	.3	.2	.1	.1	[25, 12]
O3	TA	BabyAGI	.2	.2	.1	.1	.4	.2	.2	.1	.1	[25]
O4	TA	Gen. Agents	.4	.1	.4	.1	.3	.3	.3	.2	.2	[12]
O5	EF	AgentBench	.2	.3	.1	.2	.3	.2	.1	.2	.2	[24]
O6	AP	Role-Based	.2	.2	.1	.2	.3	.2	.4	.2	.2	[25]
O7	AP	Debate-Based	.4	.1	.1	.3	.3	.3	.3	.2	.2	[25]
O8	AP	Prompt Chain	.1	.1	.1	.1	.1	.1	.2	.2	.2	[25]
O9	TA	Goal Creator	.2	.2	.2	.1	.4	.2	.2	.1	.1	[25]
O10	TA	AIOps	.2	.3	.2	.1	.3	.2	.3	.3	.3	[58]
O11	TA	SW Eng.	.2	.3	.2	.1	.3	.2	.2	.2	.2	[31, 23]
O12	TA	ERP-Integr.	.2	.4	.3	.1	.2	.2	.3	.3	.3	[49]
O13	CM	Human-Agent	.3	.2	.2	.4	.3	.3	.4	.3	.3	[15]
O14	TA	Deep Research	.3	.3	.2	.1	.4	.2	.1	.2	.2	[23, 31]
O15	EF	Evaluator	.2	.1	.1	.1	.1	.1	.1	.2	.2	[25]
O16	TA	MRKL	.2	.3	.1	.1	.3	.2	.2	.2	.2	[12]
O17	TA	Long-Horizon	.2	.2	.3	.1	.3	.2	.2	.2	.3	[33]
O18	SC	RBAC	.2	.4	.2	.1	.2	.2	.2	.3	.3	[33]
O19	SC	Safeguarded	.3	.3	.2	.1	.2	.3	.3	.4	.4	[40]
O20	CM	Workflows	.2	.3	.1	.2	.2	.1	.2	.3	.3	[25]
O21	TA	End-to-End	.3	.3	.2	.1	.4	.4	.2	.1	.3	[25]
O22	AP	Cross-Refl.	.4	.1	.1	.3	.3	.3	.3	.2	.2	[25]
O23	IC	Tool Registry	.1	.3	.1	.2	.1	.1	.1	.2	.2	[25]
O24	TA	Mem-Enhanced	.3	.2	.4	.1	.3	.4	.2	.2	.2	[12, 33]
O25	TA	Industry 4.0	.2	.4	.2	.4	.3	.3	.3	.4	.4	[45]

C Expanded Dimension-Value Mapping

Table 14 provides complete dimension definitions with value creation logic.

Table 14: Complete Dimension Definitions with Value Creation Logic

ID	Dimension	Hardened Definition	Value Creation Logic
D1	Reasoning Sophistication	The internal cognitive processes an individual agent employs to analyze information, generate solutions, and refine outputs—independent of whether other agents or humans are involved	Higher reasoning sophistication enables better decision quality, reduced errors in complex analyses, and generation of novel solutions through multi-path exploration and reflective refinement
D2	Tool Integration Depth	The extent to which an agent can discover, select, invoke, and orchestrate external tools and services to extend its action space beyond language generation	Deeper tool integration expands agent capabilities, enables workflow automation, and creates cross-system connectivity that drives productivity gains and potential revenue through enterprise integration
D3	Memory Persistence	The temporal scope and sophistication of information retention mechanisms enabling context preservation, pattern recognition, and cumulative learning	Greater memory persistence reduces repetitive information gathering, enables personalization through historical context, and supports pattern recognition that improves decision quality over time

Continued on next page

Table 14 continued

ID	Dimension	Hardened Definition	Value Creation Logic
D4	Coordination Paradigm	The structural arrangement by which multiple agents organize to divide labor, share information, and integrate outputs—independent of how individual agents reason	Multi-agent coordination enables parallel execution for throughput gains, collective verification for quality improvement, and dynamic coalition formation for complex problem decomposition
D5	Autonomy Level	The locus of decision authority determining what goals the agent pursues and what actions it takes, ranging from full human control to independent goal-setting	Higher autonomy enables reduced oversight costs, continuous 24/7 operation, and faster execution cycles; managed autonomy balances productivity gains with risk mitigation through appropriate human oversight
D6	Adaptation Mechanism	The means by which an agent improves its performance or adjusts to new situations over time, from static configuration to continuous parameter updates	Effective adaptation reduces ongoing training costs, enables rapid adjustment to new requirements, and supports self-improvement that compounds value over the deployment lifecycle
D7	Human Collaboration Mode	The nature and depth of interaction patterns between the agent and human users, from passive output delivery to bidirectional teaming with shared mental models	Appropriate collaboration modes build calibrated trust, ensure effective human oversight, and enable productive human-AI teaming that maximizes combined capabilities while mitigating risks
D8	Compliance & Alignment Posture	The degree to which an agent operates within organizational policies, regulatory requirements, and ethical standards through explicit constraints and enforcement mechanisms	Robust compliance enables enterprise deployment by satisfying regulatory requirements, reducing policy violation risks, and maintaining organizational reputation—prerequisites for revenue-generating applications
D9	Safety Architecture	The comprehensiveness of mechanisms protecting against harmful outputs, adversarial attacks, and system failures, from unprotected operation to defense-in-depth	Comprehensive safety architecture is a prerequisite for scaled deployment, enabling harm prevention that protects users, reduces liability exposure, and builds trust required for high-stakes applications

D Value Measurement Frameworks

This appendix provides extended measurement frameworks for the seven organizational value categories, including the Agentic ROI formula and enterprise benchmarks.

D.1 Agentic ROI Formula

Liu et al. [23] propose a formal measurement framework for agent value:

$$\text{Agentic ROI} = \frac{\text{Information Gain}}{\text{Cost}} = \frac{(\text{Information Quality} - \tau) \times (\text{Human Time} - \text{Agent Time})}{\text{Interaction Time} \times \text{Expense}} \tag{3}$$

Table 15 defines each component of the formula.

Table 15: Agentic ROI Formula Components

Component	Definition	Measurement
Information Quality	Accuracy, usefulness, and completeness of agent output	Task-specific quality metrics (0–1 scale)
τ (tau)	Minimum threshold for acceptable quality	Domain-dependent baseline
Human Time	Time to complete task without agent assistance	Hours/minutes baseline
Agent Time	Time taken by agent to complete same task	System logs
Interaction Time	User-agent interaction overhead	Session duration
Expense	Monetary cost incurred (API fees, compute)	Cost tracking

Key Insight: Agentic ROI is only defined when Information Quality $> \tau$, explaining why agents are predominantly adopted in high-baseline-effort domains (scientific research, code generation) where Human Time is inherently high.

D.2 Enterprise Value Benchmarks

Table 16 presents empirical benchmarks from enterprise AI deployment studies.

Table 16: Enterprise Value Benchmarks from Industry Studies

Source	Metric	High Performers	Others
McKinsey State of AI 2025 [31]			
	EBIT Attribution	>5%	<5%
	Workflow Redesign	55% fundamental	20% fundamental
	AI Agent Scaling	3× more likely	Baseline
Function-Specific Cost Reduction [31]			
	Software Engineering	56% report decrease	7% report ≥20%
	Manufacturing	56% report decrease	4% report ≥20%
	IT Operations	54% report decrease	8% report ≥20%
	Strategy & Finance	53% report decrease	8% report ≥20%
Sarferaz ERP Implementation [49]			
	Tax Configuration	50–90% time/resource reduction	
	Natural Language ERP	Significant adoption increase; reduced training costs	

D.3 Complete Value Category Definitions

Table 17 provides complete definitions for all seven value categories with measurement indicators and primary sources.

Table 17: Complete Value Category Definitions with Sources

Category	Definition	Measurement Indicators	Primary Sources
Revenue Enhancement	Agent integrates with enterprise systems enabling new business models, expanded market reach, and new service capabilities	Process cycle time; API success rate; new service capabilities count; market reach metrics; cross-functional automation %	Sarferaz [49]; Liu et al. [25]

Continued on next page

Table 17 continued

Category	Definition	Measurement Indicators	Primary Sources
Cost Reduction	Agent reduces operational expenses through automation, elimination of recurring training costs, and optimization of resource utilization	EBIT attribution %; function-specific cost decrease; training cost savings; time-to-deploy; maintenance hours reduction	McKinsey [31]; Liu et al. [23]
Risk Mitigation	Agent operates within organizational policies and compliance constraints to prevent harm, maintain accountability, and ensure regulatory adherence	Policy compliance %; audit pass rate; violation incident count; alignment drift detection rate; escalation appropriateness %	OpenAI [40]; Mohammadi et al. [33]
Productivity Gain	Agent automates workflow execution, reduces manual effort, accelerates task completion, and enables 24/7 continuous operation	Time reduction % (50–90% in ERP cases); human intervention rate; workflow completion rate; cycle time reduction; autonomy operation hours	McKinsey [31]; Sarferaz [49]
Decision Quality	Agent generates more accurate, well-reasoned solutions through improved analysis, reflection, verification, and learning from historical patterns	Information Quality score (Liu ROI formula); reasoning accuracy %; error reduction %; task success rate; consistency score	Liu et al. [23]; Sun et al. [51]
Safety Assurance	Agent implements comprehensive protective mechanisms against harmful outputs, adversarial attacks, system failures, and unintended consequences	Harmful output rate; adversarial resistance %; incident response time; production deployment readiness score; liability exposure index	OpenAI [40]; Piccialli et al. [45]
Innovation Enablement	Agent discovers non-obvious solutions through alternative reasoning paths, multi-agent collaboration, and continuous capability expansion beyond original design	Novel solution discovery rate; solution diversity metrics; capability growth rate; transfer success %; long-term performance trajectory	Liu et al. [25]; Gao et al. [8]

E Assessment Scorecard Template

The following template provides a structured approach for evaluating agentic AI systems using the CMV Framework. Practitioners can use this scorecard to systematically classify capabilities, assess maturity levels, identify gaps, and prioritize improvements.

E.1 Capability Classification (Part A)

For each dimension, identify the characteristic that best describes the system's current capability.

Table 18: Capability Classification Scorecard

Dim	C_.1	C_.2	C_.3	C_.4	Class
D1	Reactive	Single-Path	Multi-Path	Reflective	C1.__
D2	No Tools	Single-Tool	Multi-Tool	Ecosystem	C2.__
D3	Context-Window	Session	Episodic	Comprehensive	C3.__
D4	Single-Agent	Hierarchical	Peer-Based	Dynamic	C4.__
D5	Tool Mode	Guided	Supervised	Full	C5.__
D6	Static	In-Context	Feedback	Continuous	C6.__
D7	Output	Query-Response	Explanatory	Bidirectional	C7.__
D8	Unconstrained	Policy-Informed	Enforced	Value-Aligned	C8.__
D9	Unprotected	Filtered	Robustness	Defense-in-Depth	C9.__

E.2 Maturity Assessment (Part B)

For each dimension, assess implementation maturity (L1–L5) based on reliability, measurement, documentation, and value evidence.

Table 19: Maturity Assessment Scorecard

Dimension	Current (1–5)	Target (1–5)	Gap	Evidence
D1: Reasoning	L__	L__	__	
D2: Tool Integration	L__	L__	__	
D3: Memory	L__	L__	__	
D4: Coordination	L__	L__	__	
D5: Autonomy	L__	L__	__	
D6: Adaptation	L__	L__	__	
D7: Human Collaboration	L__	L__	__	
D8: Compliance	L__	L__	__	
D9: Safety	L__	L__	__	

E.3 Threshold Checks (Part C)

Critical thresholds that must be met for enterprise deployment.

Table 20: Deployment Threshold Verification

Threshold	Requirement	Current	Status
Deployment Threshold	D8 $\geq$ L3 AND D9 $\geq$ L3	D8=L__ D9=L__	PASS / FAIL
Autonomy-Governance Balance	D5 $\leq$ D8 AND D5 $\leq$ D9	D5=L__ D8=L__ D9=L__	PASS / FAIL

E.4 Summary Metrics (Part D)

E.5 Assessment Outcome

Based on the scorecard results, classify the system's deployment readiness:

- **Production Ready**: All thresholds passed; average maturity $\geq$ L3
- **Pilot Suitable**: Deployment threshold passed; some gaps remain
- **Experimental Only**: Deployment threshold not met
- **Not Recommended**: Critical safety/compliance gaps (D8 $<$ L2 or D9 $<$ L2)

Table 21: Assessment Summary Metrics

Metric	Calculation	Score
Average Maturity Score	Sum of all maturity levels ÷ 9	___/5.0
Deployment Readiness	Min(D8, D9) maturity level	L___

F Taxonomy Quality Metrics

This appendix presents the quantitative analyses underpinning Section 5.4, providing detailed derivations for profile uniqueness, distribution entropy, dimension independence, and correlation analyses.

F.1 Profile Uniqueness Analysis

Profile uniqueness measures the taxonomy's discriminative power—its ability to assign distinct classification profiles to different objects. For $n = 25$ reference objects with 9-dimensional profiles, we computed pairwise comparisons using Hamming distance (the count of dimensions where two objects differ).

Results: 24 of 25 objects (96%) received unique profiles. Objects O7 (Debate-Based Pattern) and O22 (Cross-Reflection Pattern) share identical nine-dimensional profiles:

Object	D1	D2	D3	D4	D5	D6	D7	D8	D9
O7 (Debate-Based)	.4	.1	.1	.3	.3	.3	.3	.2	.2
O22 (Cross-Reflection)	.4	.1	.1	.3	.3	.3	.3	.2	.2

Both patterns implement reflective refinement (C1.4), operate without external tools (C2.1), use context-window memory (C3.1), employ peer-based collaboration (C4.3), and share identical characteristics across remaining dimensions. This overlap suggests a potential refinement opportunity: introducing a sub-characteristic distinguishing *debate* (adversarial verification) from *reflection* (iterative improvement) coordination mechanisms.

The average Hamming distance across all 300 pairwise comparisons is 5.90 dimensions, indicating substantial differentiation capacity.

F.2 Distribution Entropy by Dimension

Shannon entropy H quantifies how uniformly objects distribute across characteristics within each dimension:

$$H = -\sum_{i=1}^{4} p_i \log_2(p_i)$$

where p_i is the proportion of objects at characteristic i. Maximum entropy for 4 characteristics is $H_{\max} = \log_2(4) = 2.0$ bits (uniform distribution).

Table 22: Characteristic Distribution and Entropy by Dimension

Dim	Name	C_.1	C_.2	C_.3	C_.4	H	H/H$_{max}$	Interp.
D1	Reasoning	2	14	6	3	1.62	0.81	Moderate
D2	Tool Integration	5	7	10	3	1.87	0.94	Balanced
D3	Memory	11	10	2	2	1.63	0.82	Moderate
D4	Coordination	17	4	2	2	1.38	0.69	Skewed
D5	Autonomy	3	4	13	5	1.75	0.87	Balanced
D6	Adaptation	4	12	7	2	1.74	0.87	Balanced
D7	Human Collab.	4	12	7	2	1.74	0.87	Balanced
D8	Compliance	4	14	5	2	1.65	0.82	Moderate
D9	Safety	3	13	7	2	1.66	0.83	Moderate

Interpretation: D4 (Coordination) exhibits the lowest entropy (1.38 bits, 69% of maximum), with 68% of objects at C4.1 (Single-Agent). This reflects empirical reality: multi-agent coordination remains uncommon in deployed systems. D2 (Tool Integration) shows the highest entropy (1.87 bits, 94% of maximum), indicating balanced adoption across the capability spectrum.

F.3 Dimension Independence Analysis

A well-designed taxonomy should have dimensions that capture *distinct* concepts—if two dimensions always move together, one may be redundant. Cramér's V quantifies the strength of association between categorical variables, helping us assess whether dimensions are conceptually independent or redundant.

Cramér's V is derived from chi-square statistics:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$$

where $n = 25$ (number of objects), χ^2 is the chi-square statistic from a contingency table cross-tabulating two dimensions, and r, c are the number of characteristic levels observed in each dimension. Values range from 0 (complete independence—knowing one dimension tells you nothing about the other) to 1 (complete association—one dimension perfectly predicts the other).

Interpretation thresholds: $V < 0.1$ = negligible; $0.1 \leq V < 0.3$ = weak; $0.3 \leq V < 0.5$ = moderate; $V \geq 0.5$ = strong association.

Table 23: Dimension Pair Associations (Cramér's V, Top 10 by Strength)

Pair	V	χ^2	p-value	Assoc.	Interpretation
D8–D9	0.90	60.97	<0.001	Strong	Governance co-occurrence
D1–D6	0.66	32.30	<0.001	Strong	Reasoning-adaptation link
D5–D8	0.65	31.33	<0.001	Strong	Autonomy-compliance tradeoff
D5–D9	0.56	23.45	0.005	Strong	Autonomy-safety tradeoff
D1–D5	0.54	22.09	0.009	Strong	Sophisticated reasoning enables autonomy
D5–D6	0.53	20.94	0.013	Strong	Autonomous systems adapt more
D4–D7	0.51	19.73	0.020	Strong	Coordination requires collaboration
D1–D4	0.51	19.64	0.020	Strong	Advanced reasoning in multi-agent
D1–D2	0.51	19.35	0.022	Strong	Reasoning supports tool use
D4–D6	0.46	16.18	0.063	Moderate	Non-significant

Reading the table: Each row represents a pair of dimensions. The **V** column shows association strength (higher = stronger relationship). The χ^2 column shows the chi-square test statistic. The **p-value** indicates statistical significance ($p < 0.05$ means the association is unlikely due to chance). The **Assoc.** column categorizes the strength, and **Interpretation** provides the substantive meaning.

Key findings: The D8–D9 association ($V = 0.90$) is the strongest, reflecting intentional co-design: systems with robust compliance mechanisms (D8) typically also implement comprehensive safety measures (D9). This association validates rather than undermines dimension separation—the constructs remain conceptually distinct (policy adherence vs. harm prevention) while exhibiting correlated implementation patterns in practice. Similarly, D1–D6 ($V = 0.66$) shows that sophisticated reasoning capabilities tend to co-occur with advanced learning/adaptation mechanisms, an expected technical dependency. The D5–D8 and D5–D9 associations ($V = 0.65$ and 0.56) reveal the *capability-safety gap*: autonomy and governance tend to be inversely related in current systems.

F.4 Autonomy-Governance Correlation

While Cramér's V measures general association (whether dimensions are related), Spearman's rank correlation (ρ) specifically tests *directional* monotonic relationships: as one dimension increases, does the other consistently increase (positive ρ) or decrease (negative ρ)? This is particularly important for understanding whether high-autonomy systems systematically lack governance mechanisms.

Spearman's ρ ranges from -1 (perfect negative relationship) through 0 (no monotonic relationship) to $+1$ (perfect positive relationship). **Interpretation thresholds**: $|\rho| < 0.3$ = weak; $0.3 \leq |\rho| < 0.5$ = moderate; $|\rho| \geq 0.5$ = strong.

Table 24: Autonomy-Governance Correlations (Spearman's ρ)

Dimension 1	Dimension 2	ρ	p-value	Interpretation
D5 (Autonomy)	D8 (Compliance)	−0.59	0.002	Strong negative
D5 (Autonomy)	D9 (Safety)	−0.39	0.052	Moderate negative (marginal)
D8 (Compliance)	D9 (Safety)	+0.81	<0.001	Strong positive

Reading the table: Each row tests the relationship between two dimensions. The ρ column shows correlation strength and direction: negative values indicate inverse relationships (as one increases, the other decreases), positive values indicate

direct relationships (both increase together). The **p-value** tests whether the correlation is statistically significant ($p < 0.05$).

Key findings:

- **D5–D8** ($\rho = -0.59$, $p = 0.002$): Strong negative correlation—as Autonomy increases, Compliance decreases. This quantitatively confirms the *capability-safety gap*: systems designed for maximum autonomy (e.g., O2: AutoGPT at C5.4) systematically lack compliance mechanisms (C8.1), while enterprise-ready systems (e.g., O19: Safeguarded Models) deliberately constrain autonomy (C5.2) in favor of robust governance (C8.4).
- **D5–D9** ($\rho = -0.39$, $p = 0.052$): Moderate negative correlation that approaches but does not reach statistical significance at the 0.05 level. The trend suggests high-autonomy systems also tend to have weaker safety mechanisms, though this relationship is less pronounced than with compliance.
- **D8–D9** ($\rho = +0.81$, $p < 0.001$): Strong positive correlation—Compliance and Safety increase together. Organizations investing in compliance infrastructure also invest in safety mechanisms, reflecting a coherent governance strategy.

This analysis validates the CMV Framework's cross-dimension dependency principle: *autonomy advancement requires commensurate governance investment to avoid deployment risk*. The empirical gap between capability-focused and governance-focused systems represents a key barrier to enterprise adoption of agentic AI.