arXiv:2608.00680v1 [cs.AI] 1 Aug 2026

Multi-Dimensional Assessment for AI Cognition (MAAC): A Theoretical Framework for Process-Oriented Cognitive Evaluation of Text-Based AI Systems

Abdalla Doleh[a,*], Ratna Babu Chinnam[a]

[a]*Department of Industrial & Systems Engineering, Wayne State University, Detroit, MI 48202, USA*

Abstract

Evaluating artificial intelligence systems has historically relied on outcome-based benchmarks that measure task accuracy, robustness, or fairness. While indispensable, these benchmarks provide limited diagnostic insight into the underlying cognitive processes that generate performance—leaving critical questions unanswered about how AI systems reason, integrate memory, manage complexity, or avoid generating false information. This paper introduces the Multi-Dimensional Assessment for AI Cognition (MAAC), a theoretically grounded framework for shifting evaluation from *what* text-based AI systems produce to *how* they think.

MAAC defines nine cognitively motivated dimensions: Cognitive Load, Tool Execution, Content Quality, Memory Integration, Complexity Handling, Hallucination Control, Knowledge Transfer, Processing Efficiency, and Process-Outcome Alignment. Each dimension is grounded in established cognitive science theory—drawing on Marr's tri-level hypothesis, Baddeley's working memory model, Sweller's cognitive load theory, and unified theories of cognition.

Five theoretical analyses provide initial support for the framework's co-

*Corresponding author. ORCID: 0009-0008-5192-2167

Email addresses: ai5145@wayne.edu (Abdalla Doleh), ai2396@wayne.edu (Ratna Babu Chinnam)

herence and empirical testability: dimension-to-theory mapping; a coverage matrix assessing breadth and non-redundancy; a formal gap analysis relative to current evaluation practice; a worked diagnostic illustration; and a set of *a priori* interdependency predictions for future empirical testing.

MAAC provides a theoretical and operational framework for principled process-level cognitive assessment of text-based AI systems, complementing existing outcome-based benchmarks with cognitively grounded, multi-dimensional evaluation.

Keywords: artificial intelligence evaluation, cognitive assessment framework, process-oriented evaluation, AI benchmarking, machine cognition, multi-dimensional measurement

---

1. Introduction

The central measurement problem in AI evaluation is no longer only whether a system produces correct outputs, but whether its underlying cognitive processes can be characterized in a principled, theory-grounded, and diagnostically useful way. Outcome-based evaluation can reveal that a model succeeds or fails on a task, but it does not specify which cognitive capabilities produced that performance or which internal limitations make that performance brittle. The present paper addresses that problem by asking what should be measured when evaluating the cognitive behavior of text-based AI systems, and how those measurements should be organized into a coherent framework.

Current AI evaluation paradigms are not equipped to answer that question. State-of-the-art benchmarks—MMLU (Hendrycks et al., 2021), BIG-bench (Srivastava et al., 2022), HELM (Liang et al., 2022)—measure what AI systems produce. They are silent on the cognitive processes that generate those outputs. A system that achieves 90% accuracy through sophisticated pattern matching and one that achieves 90% through genuine multi-step inference are indistinguishable under outcome-based evaluation (Mitchell, 2021; Bender et al., 2021). As AI systems are deployed in high-stakes contexts—clinical decision support, legal reasoning, autonomous planning—this distinction determines whether performance is robust or brittle under distribution shift.

This paper introduces MAAC, a theoretically grounded framework for evaluating AI cognitive processes rather than task outcomes. MAAC defines

nine dimensions—each grounded in cognitive science theory—that together form a diagnostic cognitive profile. The choice of nine dimensions is justified primarily by construct-domain coverage, conceptual distinctiveness, and the need to balance comprehensiveness against interpretability. It is not intended as a literal application of human short-term memory limits.

1.1. Five Critical Gaps in Current AI Evaluation

Gap 1 — Outcome dominance. Existing benchmarks are predominantly outcome-focused, measuring what AI systems produce rather than how they produce it (Rogers et al., 2020; Mitchell, 2021).

Gap 2 — Limited dimensionality. Even holistic frameworks such as HELM (Liang et al., 2022) remain constrained in coverage of cognitive processes including memory integration, complexity handling, and knowledge transfer (Bommasani et al., 2021).

Gap 3 — Faithfulness concerns. Chain-of-thought and process supervision approaches face a fundamental validity challenge: externally generated reasoning traces may not faithfully reflect internal computation (Turpin et al., 2024; Saparov and He, 2023).

Gap 4 — Validity threats. Current benchmarks suffer from data contamination, leaderboard overfitting, and construct-irrelevant variance (Magar and Schwartz, 2022; Ethayarajh and Jurafsky, 2020). Recent evidence indicates that nearly half of 60 surveyed LLM benchmarks exhibit saturation (Akhtar et al., 2026), with interdisciplinary review work identifying broader benchmark-trust problems involving documentation failures and construct-validity weaknesses (Eriksson et al., 2025).

Gap 5 — Absence of theoretical grounding. Most AI evaluation approaches lack grounding in cognitive science theory (Mitchell, 2021; Bender et al., 2021), limiting the conceptual legitimacy and generalizability of their dimensions.

1.2. Contributions

This paper makes four primary contributions:

1. **Framework architecture.** Nine theoretically grounded cognitive dimensions providing comprehensive, non-redundant coverage of essential AI cognitive processes.
2. **Falsifiable empirical predictions.** Seven *a priori* directional predictions with magnitude thresholds for dimensional correlations in future empirical validation.

3. **Theoretical validation.** Five complementary analyses: content validity, coverage sufficiency, gap closure, diagnostic utility, and a theoretically motivated interdependency network.
4. **Gap closure infrastructure.** Demonstration that all five critical limitations in current AI evaluation practice are addressed by specific MAAC dimensions through falsifiable mechanisms.

2. Background and Related Work

2.1. Large-Scale AI Benchmarking

The dominant paradigm in AI evaluation has been the large-scale task suite. SuperGLUE (Wang et al., 2019) established multi-task evaluation as standard practice; MMLU (Hendrycks et al., 2021) extended this to 57 academic domains; BIG-bench (Srivastava et al., 2022) assembled over 200 tasks specifically designed to challenge models beyond training data.

These benchmarks have produced genuine scientific value but share a structural limitation: outcome focus by design. A model's MMLU score reflects the proportion of correct answers—it reveals nothing about whether those answers emerged from structured reasoning, pattern completion, or sophisticated guessing (Mitchell, 2021). Benchmark contamination further limits interpretability: Magar and Schwartz (2022) demonstrated that performance improvements often reflect memorization of test items. Ethayarajh and Jurafsky (2020) documented systematic performance inflation from construct-irrelevant variance. Saturation problems and broader benchmark-trust issues have been extensively documented (Akhtar et al., 2026; Eriksson et al., 2025).

2.2. Reasoning-Focused and Process-Oriented Evaluation

Recognizing limitations of outcome-only assessment, a parallel literature has attempted to evaluate AI reasoning more directly. Chain-of-thought prompting (Wei et al., 2022) elicits step-by-step reasoning alongside final answers. Self-consistency (Wang et al., 2022) aggregates reasoning paths to assess process stability. Tree-of-thoughts (Yao et al., 2024) structures deliberate problem-solving across branching reasoning trees. Applied and agentic-system evaluation has also called for process-oriented frameworks (Ma et al., 2026; Kapoor et al., 2024; Mehta, 2025).

These contributions represent genuine progress. However, they face a fundamental validity challenge: externally generated reasoning traces may

not faithfully reflect internal computation (Turpin et al., 2024). Saparov and He (2023) demonstrated that chain-of-thought models often behave as greedy reasoners that exploit heuristics rather than executing systematic inference.

2.3. Holistic Evaluation Frameworks

HELM (Liang et al., 2022) represents the most comprehensive attempt at holistic AI evaluation, assessing models across seven dimensions including accuracy, robustness, fairness, bias, toxicity, efficiency, and disinformation resistance. However, HELM's dimensions are selected on practical grounds rather than grounded in cognitive theory. Important cognitive constructs including memory integration, complexity handling, and knowledge transfer receive limited coverage (Bommasani et al., 2021). Related recent work on general scales seeks more explanatory and predictive evaluation beyond benchmark totals (Zhou et al., 2026), but does not specify a cognitively grounded process architecture.

Interpretability studies represent the closest analogue to process-oriented evaluation at the mechanistic level (Clark et al., 2019; Doshi-Velez and Kim, 2017). These approaches operate at Marr's (1982) implementational level rather than the algorithmic level where cognitive constructs reside. MAAC deliberately operates at the algorithmic level, complementing rather than competing with interpretability research.

2.4. Cognitive Science Foundations

Marr's Tri-Level Hypothesis. Marr (1982) proposed that intelligent systems must be understood at three levels: computational (what problem is solved), algorithmic (how it is solved), and implementational (physical realization). MAAC operates explicitly at the algorithmic level—currently unaddressed by benchmarks and interpretability studies. Recent work argues directly that Marr's levels provide a useful framework for understanding large language models (Ku et al., 2025).

Working Memory and Cognitive Load Theory. Baddeley's (1992; 2003) working memory model demonstrates that cognitive processing is constrained by limited-capacity systems. Sweller's (1988) cognitive load theory extends this to performance under varying task complexity, distinguishing intrinsic, extraneous, and germane load.

Unified Theories of Cognition. Newell's (1990) unified theories emphasized that intelligent behavior emerges from the interaction of specialized cognitive subsystems. Anderson's ACT-R architecture (2004) operationalizes

this through distinct modules for declarative memory, procedural memory, and goal management.

Transfer Learning and Analogical Reasoning. Barnett and Ceci (2002) established a taxonomy distinguishing near from far transfer. Gentner (1983) and Holyoak and Morrison (2012) identified analogical reasoning as the primary cognitive mechanism underlying far transfer.

Dual-Process Theory and Hallucination. Kahneman's (2011) dual-process framework distinguishes System 1 (fast, automatic, heuristic) from System 2 (slow, deliberate, rule-governed) processing. This maps directly onto AI hallucination: the generation of plausible but false information reflects fluency-maximizing heuristics in the absence of robust uncertainty calibration (Ji et al., 2023; Huang et al., 2023).

2.5. Positioning MAAC in the Evaluation Landscape

Table 1 positions MAAC relative to existing evaluation frameworks. MAAC occupies a distinct position: the only framework combining process orientation, cognitive science grounding, and multi-dimensional independence—complementing outcome benchmarks rather than replacing them.

Table 1: Comparison of MAAC with Existing AI Evaluation Frameworks

Framework	Primary Focus	Dims.	Grounding	Key Limitations	MAAC Contribution
MMLU (Hendrycks et al., 2021)	Factual recall across 57 domains	1	Minimal	Outcome-only; no process insight; contamination risk	MI, CH, KT add process-level assessment
BIG-bench (Srivastava et al., 2022)	Novel task performance	Task-specific	Limited	Emergence without explanation; atheoretical	POA validates that claimed reasoning reflects actual processing
HELM (Liang et al., 2022)	Holistic outcomes: accuracy, robustness, fairness	7	Practical	Outcome-focused; cognitive constructs absent	Cognitive science grounding; 9 dimensions covering constructs HELM omits
Chain-of-Thought (Wei et al., 2022)	Step-by-step reasoning traces	Process-oriented	Cognitive psych.	Faithfulness concerns	Process-Outcome Alignment provides cross-validated alignment
TruthfulQA (Lin et al., 2022)	Truthfulness of outputs	1	Epistemological	Outcome-only factual accuracy; no calibration	Hallucination Control as process-oriented dimension
MAAC (Present)	**How AI systems think: 9 cognitive process dimensions**	**9**	**Marr (1982), Newell (1990), Baddeley (1992), Sweller (1988)**	Requires future empirical validation; higher implementation overhead	Addresses all 5 gaps; complementary to all frameworks above

Note. MAAC Contribution column highlights the specific diagnostic value added relative to each existing approach.

2.6. Scoping Review Methodology

Framework development was informed by a structured scoping review spanning AI evaluation, cognitive psychology, psychometrics, and reasoning assessment, conducted between January and June 2025 (Arksey and O'Malley, 2005; Peters et al., 2020). Searches were conducted across arXiv.org, ACL Anthology, major machine learning conference proceedings (NeurIPS, ICLR, ICML, AAAI), ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, and Google Scholar.

After screening approximately 150 records against predefined inclusion criteria focused on process-oriented evaluation, cognitive constructs, and measurement theory, 108 sources were retained for construct mapping (Prinsen et al., 2018). Two coders independently assigned sources to provisional construct categories, with disagreements resolved through discussion; inter-rater agreement was $\kappa = .78$, 95% CI [.65, .91]. The resulting construct map informed both the consolidation of nine retained dimensions and the exclusion of overlapping candidate dimensions.

2.7. Literature Review Synthesis

Table 2 provides a systematic summary of how literature categories informed the development of specific MAAC dimensions.

Table 2: Literature Review Synthesis Supporting MAAC Framework Development

Literature Category	**Period**	**Representative Studies**	**MAAC Dims.**	**Key Contributions**	**Gaps Addressed**
Large-Scale Task Suites	2019–2022	Wang et al. (2019); Hendrycks et al. (2021); Srivastava et al. (2022)	All dimensions	Established need for process-oriented evaluation	Outcome dominance
Reasoning Evaluation	2020–2024	Wei et al. (2022); Yao et al. (2024); Wang et al. (2022)	CH, POA	Demonstrated importance of step-by-step reasoning	Faithfulness concerns
Memory & Retrieval	2020–2023	Lewis et al. (2020); Borgeaud et al. (2022); Shi et al. (2023)	MI, KT	Showed role of information persistence	Memory coherence gaps
Hallucination & Factuality	2020–2024	Maynez et al. (2020); Ji et al. (2023); Huang et al. (2023)	HC, CQ	Identified need for process-level error prevention	Lack of prevention-focused measurement
Efficiency & Scaling	2019–2024	Kaplan et al. (2020); Hoffmann et al. (2022); Sardana et al. (2024)	PE, CL	Established efficiency as cognitive concern	Missing cognitive interpretation
Interpretability & Validity	2017–2024	Doshi-Velez and Kim (2017); Clark et al. (2019); Bommasani et al. (2021)	POA, CH	Highlighted process-outcome alignment gap	Process-outcome validation
Cognitive Science	1956–2012	Marr (1982); Newell (1990); Baddeley (1992)	Framework-wide	Provided theoretical anchor for multi-dimensional assessment	Absence of theory

Note. Categories are not mutually exclusive. Synthesis is based on 108 core references retained after screening.

3. Gap Closure Analysis

This section formalizes the relationship between the five critical limitations and the MAAC framework, demonstrating precisely how each gap is closed—which dimensions address it, through what mechanism, and with what theoretical warrant. Two principles govern closure claims. First, closure is mechanistic, not nominal. Second, closure is partial where warranted: where a gap is addressed in principle but requires empirical confirmation, this is acknowledged explicitly.

Table 3: Gap-Closure Analysis: Five Critical Limitations and Their Resolution Through MAAC Dimensions

Gap	Limitation	MAAC Dimension(s)	Closure Mechanism	Key Citations	Status
1	Outcome Dominance	CH, MI, KT, CL	MAAC assesses process signatures directly: resource allocation under load, context coherence, cross-domain knowledge application, multi-step reasoning	Mitchell (2021); Bender et al. (2021)	Theoretically closed; empirical closure in future empirical work
2	Limited Dimensionality	All 9 dimensions	Nine-dimensional structure derived from systematic cognitive science theory rather than practical convenience	Liang et al. (2022); Bommasani et al. (2021)	Theoretically closed
3	Faithfulness Concerns	POA, CH	D9 measures process-outcome alignment across multiple problem instances, replacing single-instance trace inspection with cross-validated behavioral consistency	Turpin et al. (2024); Saparov and He (2023)	Theoretically closed; empirical closure in future empirical work
4	Validity Threats	KT, MI, HC	MAAC dimensions designed for use with dynamically generated, complexity-validated scenarios; KT requires novel cross-domain application that memorization cannot satisfy	Magar and Schwartz (2022); Ethayarajh and Jurafsky (2020)	Theoretically closed; empirical closure in future empirical work
5	Absence of Theory	Framework-wide	Every MAAC dimension is formally anchored in an established cognitive science theory. CL → Sweller (1988); MI → Baddeley (1992); KT → Barnett & Ceci (2002)	Mitchell (2021); Marr (1982)	Theoretically closed

Note. Citations refer to sources establishing each limitation and those motivating the corresponding MAAC closure mechanism.

3.1. Scope of Closure: Theoretical vs. Empirical

Table 3 demonstrates theoretical closure: for each gap, a principled mechanism exists within the MAAC framework that addresses the limitation. Empirical closure—demonstrating that MAAC scores actually behave as the closure mechanisms predict— requires future empirical validation. Future empirical studies should test whether MAAC dimensions discriminate cognitive profiles that outcome-based benchmarks conflate, whether dimensional scores

exhibit the predicted correlation structure, and whether Process-Outcome Alignment scores detect process-outcome misalignment in practice.

This distinction between theoretical and empirical closure is not a weakness—it is standard measurement science practice (DeVellis, 2017; Furr, 2018). The empirical validation roadmap constitutes the falsifiable predictions that make MAAC a scientific framework rather than a descriptive taxonomy.

4. The MAAC Framework

The MAAC framework comprises nine theoretically grounded dimensions that collectively capture the essential aspects of artificial cognitive processing. This section presents the complete framework architecture in five parts: (A) dimension-to-theory mapping; (B) coverage matrix; (C) formal definitions and operationalizations; (D) the Process-Outcome Alignment firewall; and (E) a worked diagnostic example.

4.1. Dimension-to-Theory Mapping

Each MAAC dimension is formally grounded in an established body of cognitive science theory operating at Marr's (1982) algorithmic level.

Table 4: MAAC Dimension-to-Theory Mapping

#	Dimension	Cognitive Theory	Key Construct	Primary Citations	Measurement Approach	Abbr.
D1	Cognitive Load	Sweller (1988); Baddeley (1992, 2003)	Working memory capacity; intrinsic vs. extraneous load	Sweller (1988)	Performance degradation curves across context length and constraint density	CL
D2	Tool Execution	Clark & Chalmers (1998); Hutchins (1995); Nakano et al. (2021)	Extended cognition; distributed cognitive processing	Clark and Chalmers (1998)	Tool selection accuracy and multi-step orchestration success rates	TE
D3	Content Quality	McNamara et al. (2010); Crossley et al. (2016); Halliday & Hasan (1976)	Semantic richness; discourse coherence	McNamara et al. (2010)	Semantic similarity, coherence scoring, register appropriateness	CQ
D4	Memory Integration	Baddeley (1992, 2000, 2003); Lewis et al. (2020)	Working memory persistence; episodic buffer; consolidation	Baddeley (1992)	Context coherence scores across turn depth and information persistence tests	MI
D5	Complexity Handling	Newell & Simon (1972); Halford et al. (2005); Wood (1986)	Problem decomposition; constraint satisfaction; goal mgmt	Newell and Simon (1972)	Multi-step reasoning accuracy across Simple/Moderate/Complex tiers	CH
D6	Hallucination Control	Kahneman (2011); Tversky & Kahneman (1974); Guo et al. (2017)	Uncertainty calibration; knowledge boundary recognition	Kahneman (2011)	Hallucination rate, uncertainty calibration curves, consistency indices	HC
D7	Knowledge Transfer	Barnett & Ceci (2002); Perkins & Salomon (1992); Gentner (1983)	Near and far transfer; analogical reasoning; abstraction	Barnett and Ceci (2002)	Cross-domain transfer accuracy across near-to-far transfer distance gradient	KT
D8	Processing Efficiency	Simon (1956, 1972); Griffiths et al. (2015); Strubell et al. (2019)	Bounded rationality; resource-rational computation	Simon (1956)	Quality-adjusted latency and token efficiency across complexity tiers	PE
D9	Process-Outcome Alignment	Cronbach & Meehl (1955); Messick (1995); Turpin et al. (2024)	Process-outcome alignment; convergent and discriminant validity	Cronbach and Meehl (1955)	Process-outcome alignment coefficients across paraphrased problem variants	POA

Note. Each dimension is anchored in an established cognitive science theory at Marr's (1982) algorithmic level. D9 has been renamed from "Construct Validity" to "Process-Outcome Alignment" to distinguish the AI system property being measured from the psychometric property of the MAAC instrument itself (Section 4.4). Measurement operationalization is developed in future empirical work.

4.2. Coverage Matrix: Exhaustiveness and Non-Redundancy

A framework claiming comprehensive cognitive coverage must demonstrate two properties: exhaustiveness (all essential cognitive construct categories represented) and non-redundancy (no two dimensions measure the same construct).

Table 5: MAAC Coverage Matrix

Cognitive Construct Category	CL	TE	CQ	MI	CH	HC	KT	PE	POA	Dims.
Working Memory & Capacity Constraints	●	○	○	●	●	○	○	●	○	4
Long-Term Memory & Retrieval	○	○	○	●	○	○	●	○	○	2
Problem Solving & Goal Management	○	○	○	○	●	○	○	○	●	2
Transfer & Generalization	○	○	○	○	○	○	●	○	○	1
Uncertainty & Calibration	○	○	○	○	○	●	○	○	●	2
Linguistic & Discourse Quality	○	○	●	○	○	○	○	○	○	1
Distributed & Extended Cognition	○	●	○	○	○	○	○	●	○	2
Process-Outcome Validation	○	○	○	○	○	○	○	○	●	1
Resource Rationality & Efficiency	●	○	○	○	○	○	○	●	○	2
Categories per Dimension	**2**	**1**	**1**	**2**	**2**	**1**	**2**	**3**	**3**	

Note. Filled cells (●) indicate primary construct coverage. Each construct category is covered by at least one dimension (exhaustiveness); no two dimensions cover identical category profiles (non-redundancy). Bottom row shows the number of construct categories covered per dimension.

4.3. Formal Dimension Definitions and Operationalizations

4.3.1. D1 — Cognitive Load (CL)

Definition: Cognitive Load assesses how AI system performance degrades as task complexity, context length, or simultaneous processing constraints increase—revealing capacity limitations analogous to working memory bottlenecks in human cognition.

Theoretical Foundation: Sweller (1988); Sweller et al. (2019); Baddeley (1992, 2003). Working memory capacity constraints; intrinsic vs. extraneous load.

Justification: Current benchmarks test systems under optimal conditions but ignore performance under strain. Cognitive Load patterns reveal fundamental capacity limitations critical for deployment reliability (Miller, 1956; Simon, 1972).

Measurement Constructs: (1) Performance degradation rate across increasing context length; (2) multi-constraint task management accuracy; (3) resource allocation consistency under competing demands; (4) capacity limitation threshold identification.

Framework note: Predicts positive correlation with Processing Efficiency ($r > .60$) via shared resource constraint mechanisms.

4.3.2. D2 — Tool Execution (TE)

Definition: Tool Execution evaluates an AI system's ability to coordinate with external tools and resources, including function calling, API in-

tegration, error recovery, and orchestration of multi-step tool-mediated processes.

Theoretical Foundation: Clark and Chalmers (1998); Hutchins (1995); Nakano et al. (2021). Extended cognition; distributed cognitive processing; meta-cognitive tool awareness.

Justification: As AI systems increasingly operate in tool-rich environments, effective external resource coordination becomes a critical cognitive capability (Schick et al., 2024).

Measurement Constructs: (1) Tool selection appropriateness and efficiency; (2) multi-step process orchestration accuracy; (3) error detection and recovery in tool interactions; (4) meta-cognitive awareness of tool limitations.

Framework note: Predicts moderate positive correlation with Processing Efficiency ($r > .40$) via shared operational efficiency mechanisms.

4.3.3. D3 — Content Quality (CQ)

Definition: Content Quality measures the semantic richness, discourse coherence, and communicative appropriateness of AI-generated content, focusing on linguistic and structural quality independent of factual accuracy.

Theoretical Foundation: McNamara et al. (2010); Crossley et al. (2016); Halliday and Hasan (1976). Semantic richness; discourse coherence; communicative effectiveness.

Justification: Factual accuracy captures only one dimension of communication quality. Coherence, register appropriateness, and organizational clarity are essential for effective deployment in professional contexts (Zhang et al., 2020).

Measurement Constructs: (1) Semantic richness and vocabulary sophistication; (2) discourse coherence across multi-sentence outputs; (3) register and style appropriateness; (4) clarity and communicative effectiveness.

Framework note: Conceptually complementary to Hallucination Control: CQ assesses linguistic product quality; HC assesses factual process integrity.

4.3.4. D4 — Memory Integration (MI)

Definition: Memory Integration assesses how effectively an AI system maintains, updates, and utilizes contextual information across multi-turn interactions, including coherence preservation, information persistence, and integration of new information with prior context.

Theoretical Foundation: Baddeley (1992, 2000, 2003); Lewis et al. (2020). Working memory persistence; episodic buffer; information consolidation.

Justification: Single-turn evaluation misses critical aspects of coherent extended behavior. Memory integration failures produce context drift, contradictions, and loss of established facts—failure modes invisible to outcome-based assessment (Borgeaud et al., 2022).

Measurement Constructs: (1) Context coherence across multiple interaction turns; (2) information persistence and retrieval accuracy over extended exchanges; (3) integration of new information without contradiction; (4) appropriate updating of established context.

Framework note: Predicts positive correlation with Knowledge Transfer ($r > .50$) via shared information retrieval and consolidation mechanisms (Baddeley, 1992).

4.3.5. D5 — Complexity Handling (CH)

Definition: Complexity Handling evaluates an AI system's ability to manage multi-step reasoning, hierarchically decompose problems, coordinate multiple simultaneous constraints, and integrate information from diverse sources toward a coherent solution.

Theoretical Foundation: Newell and Simon (1972); Halford et al. (2005); Wood (1986); Campbell (1988). Problem decomposition; constraint satisfaction; hierarchical goal management.

Justification: The ability to handle structurally complex problems is a hallmark of sophisticated intelligence. Complexity Handling goes beyond task completion to examine the processes by which systems manage cognitive complexity across Wood's (1986) component and coordinative task dimensions.

Measurement Constructs: (1) Multi-step reasoning coordination accuracy; (2) hierarchical problem decomposition quality; (3) constraint satisfaction across multiple simultaneous demands; (4) integration of diverse information sources toward coherent conclusions.

Framework note: Applicable to complexity-validated scenarios spanning Simple, Moderate, and Complex tiers.

4.3.6. D6 — Hallucination Control (HC)

Definition: Hallucination Control measures an AI system's ability to avoid generating false, fabricated, or inconsistent information—particularly in high-uncertainty scenarios—by assessing uncertainty awareness, consistency, and knowledge boundary recognition.

Theoretical Foundation: Kahneman (2011); Tversky and Kahneman (1974); Gal and Ghahramani (2016); Guo et al. (2017). Uncertainty calibration; knowledge boundary recognition; heuristic error suppression.

Justification: Unlike post-hoc fact-checking, this dimension examines the cognitive processes that lead to hallucination (Ji et al., 2023; Huang et al., 2023).

An important boundary follows from this theoretical choice. D6 targets *System 1 hallucination*—fabrication arising when fluency-maximizing heuristics outrun calibrated uncertainty control. A distinct failure mode—*System 2 hallucination*—exists in which deliberate reasoning is executed coherently but built on a fabricated baseline premise. D6's current constructs assess epistemic signaling quality, not the truth status of the initiating premise itself. Premise verification operates at Marr's (1982) computational level, whereas D6 is intentionally scoped to the algorithmic level. Detecting System 2 hallucinations requires an additional premise-grounding validation layer reserved for future framework extension.

Measurement Constructs: (1) False information generation frequency across domains; (2) uncertainty expression appropriateness and calibration; (3) consistency of claims across related query instances; (4) appropriate knowledge boundary recognition.

Framework note: Predicts negative correlation with Knowledge Transfer under high-uncertainty conditions ($r < -.30$).

4.3.7. D7 — Knowledge Transfer (KT)

Definition: Knowledge Transfer examines an AI system's ability to apply learned concepts, patterns, and structural relationships across different domains, contexts, and problem types—distinguishing genuine generalization from domain-specific memorization.

Theoretical Foundation: Barnett and Ceci (2002); Perkins and Salomon (1992); Gentner (1983); Holyoak and Morrison (2012). Near and far transfer; analogical reasoning; cross-domain abstraction.

Justification: Existing benchmarks implicitly reward specialization over generalization. Real-world deployment requires flexible cross-domain application. Barnett and Ceci's (2002) near-to-far transfer taxonomy provides the theoretical scaffolding for graded transfer assessment.

Measurement Constructs: (1) Zero-shot and few-shot cross-domain transfer accuracy; (2) analogical reasoning and structural mapping quality; (3) con-

ceptual abstraction and novel application performance; (4) performance degradation gradient as transfer distance increases.

Framework note: Predicts positive correlation with Memory Integration ($r > .50$) and negative correlation with Hallucination Control ($r < -.30$) under high-uncertainty transfer conditions.

4.3.8. D8 — Processing Efficiency (PE)

Definition: Processing Efficiency evaluates the computational economy of AI cognitive operations, including the relationship between resource expenditure (latency, token generation, computational cost) and output quality across tasks of varying complexity.

Theoretical Foundation: Simon (1956, 1972); Griffiths et al. (2015); Just and Carpenter (1992); Strubell et al. (2019). Bounded rationality; resource-rational computation; cognitive economy.

Justification: Inefficiency often signals cognitive brittleness rather than robust understanding—excessive computation may reflect brute-force search rather than structured reasoning (Schwartz et al., 2020). Simon's bounded rationality framework establishes efficiency as a cognitive property, not merely an engineering concern.

Efficiency expectations are interpreted relative to task complexity—the relevant question is not whether a response is brief in absolute terms, but whether its reasoning economy is proportional to the demand profile of the scenario.

Measurement Constructs: (1) Quality-adjusted computational cost per task; (2) scaling behavior of resource use across complexity tiers; (3) consistency of efficiency across domain types; (4) resource allocation rationality under constrained conditions.

Framework note: Predicts positive correlation with Cognitive Load ($r > .60$) via shared resource constraint mechanisms.

4.3.9. D9 — Process-Outcome Alignment (POA)

Definition: Process-Outcome Alignment serves as the meta-evaluative dimension, assessing whether an AI system's demonstrated reasoning processes are consistent with its outputs—ensuring that cognitive claims are empirically grounded rather than post-hoc rationalizations.

Theoretical Foundation: Cronbach and Meehl (1955); Messick (1995); Turpin et al. (2024); Saparov and He (2023). Process-outcome alignment; convergent and discriminant validity; nomological coherence.

Justification: Without process-outcome alignment validation, evaluations risk circular reasoning: declaring systems 'reason' simply because they produce correct answers (Turpin et al., 2024).

Measurement Constructs: (1) Process-outcome consistency across varied problem instances; (2) reasoning trace alignment with final answer patterns; (3) cross-validation of cognitive claims across measurement approaches; (4) stability of process signatures under problem paraphrasing.

Framework note: This dimension assesses the AI system's internal process-outcome alignment—NOT the validity of the MAAC framework itself. See Section 4.4.

4.4. The Process-Outcome Alignment Dimension: A Critical Conceptual Distinction

Dimension 9 requires explicit clarification to prevent a potential circular reasoning concern. The name 'Process-Outcome Alignment' in this context refers to a property of the AI system being evaluated—not a property of the MAAC framework itself.

What D9 measures (AI system property): The degree to which the AI system's demonstrated reasoning processes are consistent with its final outputs across varied problem instances. A system that produces coherent reasoning traces but arrives at conclusions inconsistent with those traces scores low on D9. This is a behavioral, empirically assessable property.

What D9 does NOT measure (framework property): The validity of the MAAC framework itself—whether MAAC scores are theoretically grounded, psychometrically sound, or empirically defensible. Framework-level construct validity is established through the theoretical analyses in this paper and the empirical validation studies.

This distinction follows directly from Messick's (1995) unified validity framework, which distinguishes between the validity of an assessment instrument and the construct properties it is designed to measure. D9 was labeled "Construct Validity" in earlier framework versions. The rename better reflects what is being measured and avoids conflation with the psychometric concept of construct validity as applied to the MAAC instrument itself.

4.5. Worked Diagnostic Example: What MAAC Reveals That Benchmarks Cannot

To illustrate the diagnostic value of multi-dimensional cognitive profiling, consider two hypothetical AI systems—Model A and Model B—that

achieve identical accuracy on a standard benchmark (84% on MMLU). Table 6 presents their MAAC cognitive profiles.

Table 6: Worked Diagnostic Example: Identical Benchmark Accuracy, Divergent Cognitive Profiles

Model / Metric	**CL**	**TE**	**CQ**	**MI**	**CH**	**HC**	**KT**	**PE**	**POA**
Benchmark Accuracy (MMLU)	84%	84%	84%	84%	84%	84%	84%	84%	84%
Model A (fast, high-throughput)	82	91	78	45	52	38	43	88	47
Model B (deliberate reasoning)	61	74	82	79	84	81	77	54	83

Note. Scores range 0–100. ≥ 75 = strong; 50–74 = moderate; < 50 = weak. Both models score 84% on MMLU. MAAC reveals fundamentally different cognitive architectures. Illustrative hypothetical example; not empirical evidence.

Despite identical benchmark accuracy, Model A and Model B exhibit fundamentally different cognitive architectures. Model A scores strongly on Cognitive Load (82), Tool Execution (91), and Processing Efficiency (88) but weakly on Memory Integration (45), Complexity Handling (52), Hallucination Control (38), Knowledge Transfer (43), and Process-Outcome Alignment (47)— indicating accuracy via pattern completion rather than structured reasoning. Model B shows the inverse pattern: strong Memory Integration (79), Complexity Handling (84), Hallucination Control (81), Knowledge Transfer (77), and Process-Outcome Alignment (83), but weaker Cognitive Load (61) and Processing Efficiency (54). A benchmark score of 84% provides none of this information. MAAC provides all of it.

4.6. Considered and Excluded Dimensions

Rigorous framework development requires transparent documentation of alternatives considered but excluded. Five criteria governed exclusions: (1) theoretical relevance; (2) empirical tractability; (3) diagnostic utility; (4) framework parsimony; and (5) generalizability.

Table 7: Systematically Considered and Excluded Elements in MAAC Development

Category	**Considered Element**	**Exclusion Rationale**	**MAAC Alternative**
Additional Dimensions	Emotional Intelligence / Affective Processing	Limited applicability to current AI; insufficient behavioral evidence for reliable measurement	Covered implicitly in Content Quality (contextual appropriateness)
	Creativity / Generative Novelty	Difficult to operationalize objectively; overlap with CH and KT	Integrated within existing constructs
	Social Cognition / Theory of Mind	Specialized domain; limited relevance to domain-general cognitive assessment	Candidate for future framework extension
	Meta-Cognitive Awareness	Partially captured in POA; high risk of dimensional overlap	Integrated within Process-Outcome Alignment
Framework Approaches	Hierarchical Factor Model (positing a g-factor)	Assumes general intelligence factor inappropriate for modular AI architecture	Multi-dimensional independent assessment adopted
	Process-Outcome Integration Scoring	Conflates process and outcome, reducing diagnostic specificity	Separate process-oriented framework maintained
	Competency-Based Framework	Task-specific focus conflicts with goal of domain-general assessment	Prioritized measurement of cognitive constructs
Methodological Alternatives	Single-Score Aggregation	Loses diagnostic specificity central to the framework's purpose	Multi-dimensional profiles preserved
	Binary Classification	Oversimplifies the continuous nature of complex cognitive constructs	Continuous dimensional scoring adopted
	Comparative Ranking (e.g., Elo)	Lacks absolute measurement properties needed for tracking individual system development	Absolute cognitive measurement maintained

Note. Exclusion decisions evaluated against five criteria: theoretical relevance, empirical tractability, diagnostic utility, framework parsimony, and generalizability.

5. Theoretical Validation Analyses

Framework validation at the theoretical level requires demonstrating that the framework satisfies established measurement science standards prior to empirical testing (Mokkink et al., 2010; Terwee et al., 2018).

5.1. Content Validity (H1): Dimension-to-Theory Mapping

Content validity requires that framework dimensions comprehensively capture essential aspects of the construct being assessed (Messick, 1995; Mokkink et al., 2010). The dimension-to-theory mapping in Table 4 provides this evidence across three criteria.

Theoretical warrant. Every MAAC dimension is anchored in at least one established cognitive science theory. The theoretical lineages span five traditions: capacity and load theory (Sweller, Baddeley, Miller), unified cognitive architecture (Newell, Anderson), transfer and generalization (Barnett & Ceci, Gentner), validity theory (Cronbach & Meehl, Messick), and extended and distributed cognition (Clark & Chalmers, Hutchins).

Construct distinctiveness. The 108 papers reviewed were mapped against the nine dimensions using stratified inter-rater coding ($\kappa = 0.78$, 95% CI [0.65, 0.91]). No two dimensions share an identical theoretical lineage, and the coverage matrix (Table 5) confirms no two dimensions cover identical construct category profiles.

Algorithmic-level focus. All nine dimensions operate at Marr's (1982) algorithmic level rather than the computational or implementational levels.

H1 (content validity) is supported at the theoretical level. Empirical content validity is a target for future empirical work.

5.2. Coverage Comprehensiveness (H1c)

H1c predicts that nine dimensions provide theoretically sufficient coverage while maintaining practical interpretability. The coverage matrix (Table 5) supports a nine-dimension solution as a parsimonious configuration covering nine essential cognitive construct categories without redundancy. Fewer dimensions would leave construct categories unaddressed; more dimensions would either duplicate existing coverage or introduce constructs excluded on principled grounds (Table 7). The nine-dimensional structure is therefore presented as a consequence of construct-domain coverage requirements rather than as a fixed design target.

5.3. Gap Closure (H3)

H3 predicts that MAAC addresses all five critical limitations. Section III presented the full mechanistic gap-closure analysis (Table 3). Three dimensions carry disproportionate gap-closure weight. Knowledge Transfer and Memory Integration each close two gaps. Process-Outcome Alignment closes

Gap 3 (faithfulness concerns) uniquely—no other dimension provides process-outcome alignment validation—establishing it as MAAC's most theoretically distinctive contribution. H3 is supported at the theoretical level for all five gaps; empirical closure requires future empirical studies.

5.4. Diagnostic Utility (H1b)

H1b predicts that multi-dimensional cognitive profiles provide diagnostic insights unavailable through aggregate performance scores. The worked example (Table 6) provides the theoretical demonstration: two models with identical 84% MMLU accuracy exhibit fundamentally different MAAC profiles with opposite deployment implications.

The diagnostic utility argument has a specific falsifiability condition: if MAAC dimensional scores were perfectly collinear with benchmark accuracy, MAAC would add no diagnostic value. Future empirical work should directly test this condition. The framework predicts correlations between a given AI system's benchmark accuracy and its nine MAAC dimensional scores will be moderate ($r < .70$) for most dimensions.

5.5. Theoretical Validation Summary

Table 8: Theoretical Validation Summary

Hypothesis	Analysis Type	Evidence Source	Validation Criterion	Outcome
H1 — Content Validity	Dim-to-theory mapping vs. 108 literature constructs	Table 4 + Section 2	Every dimension anchored in established cognitive science theory	All 9 dimensions mapped to distinct theoretical lineages across 5 traditions ✓ Supported
H1c — Coverage	Coverage matrix: 9 construct categories × 9 dimensions	Table 5 + Table 7	Exhaustiveness + non-redundancy; excluded dims documented with principled rationale	9/9 construct categories covered; all profiles distinct ✓ Supported
H2 — Structural Validity	Interdependency network with *a priori* directional predictions	Table 9 + Section 5.6	7 directional predictions with magnitude thresholds; $r < .85$ discriminant bound	7 predictions specified (6 positive, 1 negative); empirical testing in future work ○ Predictions Specified
H3 — Gap Closure	Mechanistic gap-closure across 5 limitations	Table 3	Each gap addressed by at least one dimension through a specific, falsifiable mechanism	All 5 gaps supported theoretically ✓ Supported
H1b — Diagnostic Utility	Worked example: identical accuracy, divergent profiles	Table 6	Multi-dimensional profiles reveal deployment-relevant distinctions invisible to unidimensional benchmarks	Model A vs. Model B: identical 84% MMLU, opposite MAAC profiles ✓ Supported

Note. ✓ = supported at the theoretical level. ○ = predictions specified, empirical testing in future work.

5.6. A Priori Interdependency Predictions for Future Empirical Validation (H2)

Structural validity requires that theoretically predicted interdependencies among dimensions be specified prior to empirical testing. Table 9 presents seven *a priori* directional predictions derived from the cognitive science theories underlying each dimension pair. Six predictions are positive; one is negative, reflecting a theoretically motivated tension between generalization drive and calibration constraints (Kovacs and Conway, 2016).

Table 9: *A Priori* Interdimensional Correlation Predictions for Future Empirical Validation

Dimension Pair	Direction	Magnitude	Theoretical Basis	Discriminant Bound	Tested In
CL (D1) ↔ PE (D8)	Positive	$r > .60$	Shared resource constraint mechanisms; systems with efficient resource allocation exhibit less performance degradation under load (Simon, 1972; Baddeley, 1992)	$r < .85$	Future work
MI (D4) ↔ KT (D7)	Positive	$r > .50$	Shared information retrieval and consolidation mechanisms; effective cross-domain transfer requires robust storage and retrieval (Baddeley, 1992; Barnett and Ceci, 2002)	$r < .85$	Future work
CH (D5) ↔ POA (D9)	Positive	$r > .40$	Systems that genuinely engage with problem structure are more likely to produce process traces consistent with outputs (Newell and Simon, 1972; Turpin et al., 2024)	$r < .85$	Future work
CQ (D3) ↔ HC (D6)	Positive	$r > .35$	Complementary output reliability mechanisms; systems with strong discourse coherence tend toward better calibration (Ji et al., 2023)	$r < .85$	Future work
TE (D2) ↔ PE (D8)	Positive	$r > .40$	Shared operational efficiency mechanisms; effective tool coordination reduces redundant computation (Hutchins, 1995; Schick et al., 2024)	$r < .85$	Future work
CL (D1) ↔ CH (D5)	Positive	$r > .45$	Capacity-complexity interaction; systems with higher effective working memory capacity handle structurally complex tasks more effectively (Baddeley, 2003; Halford et al., 2005)	$r < .85$	Future work
KT (D7) ↔ HC (D6)	Negative	$r < -.30$	Under high-uncertainty transfer conditions, the generalization drive enabling cross-domain application conflicts with calibration constraints suppressing confident fabrication (Barnett and Ceci, 2002; Kahneman, 2011; Ji et al., 2023)	N/A	Future work

Note. All seven predictions are specified prior to data collection. We adopt $r < .85$ as a conservative discriminant-validity heuristic (Terwee et al., 2018); any dimensional pair exceeding this bound would indicate construct redundancy requiring framework revision. The single negative prediction (KT↔HC) reflects a theoretically motivated tension rather than general antagonism. Unpredicted pairs will be reported descriptively in future empirical work without confirmatory interpretation.

6. Discussion

6.1. Theoretical Contributions

Process-oriented evaluation as a scientific program. The most significant contribution of MAAC is not any individual dimension but the demonstration that process-oriented cognitive assessment of AI systems is theoretically coherent, practically implementable, and scientifically falsifiable. Prior process-oriented approaches suffered from the faithfulness problem (Turpin et al., 2024). MAAC addresses this directly through the Process-Outcome Alignment dimension, which treats process-outcome alignment as an empirically assessable behavioral property rather than an assumption.

Bridging cognitive science and AI evaluation. MAAC demonstrates that classical cognitive science frameworks translate productively to artificial cognitive assessment when applied at Marr's (1982) algorithmic level. Sweller's (1988) cognitive load theory, Baddeley's (1992) working memory model, Barnett and Ceci's (2002) transfer taxonomy, and Newell and Simon's (1972) problem-solving architecture each find direct operationalization in MAAC dimensions.

The diagnostic profile as a unit of analysis. MAAC introduces the nine-dimensional cognitive profile as a new unit of analysis in AI evaluation. The worked example (Table 6) demonstrates that identical benchmark accuracy can coexist with fundamentally different cognitive architectures—a finding with direct implications for deployment decisions, architectural development, and safety assessment.

6.2. Practical Implications

For AI developers. MAAC dimensional scores provide targeted development guidance that aggregate benchmarks cannot. A model scoring poorly on Memory Integration but strongly on Complexity Handling points to retrieval system limitations rather than reasoning architecture deficits. Table 10 maps each dimension to its corresponding development target.

Table 10: MAAC Dimensions and Corresponding Development Targets

Dimension	**Poor Performance Indicators**	**Architectural Implications**	**Development Recommendations**
Cognitive Load	Context length degradation, multi-constraint failures	Attention mechanism limitations; insufficient working memory analog	Hierarchical attention, memory compression strategies
Tool Execution	Tool selection errors, orchestration failures, poor recovery	Poor meta-cognitive awareness of tool capabilities	Tool selection models, execution monitoring, error recovery protocols
Content Quality	Semantic incoherence, register mismatches	Generation control weaknesses, planning deficits	Content planning, discourse coherence mechanisms, style control
Memory Integration	Cross-turn inconsistency, information loss, contradictions	Memory management deficits, insufficient episodic buffering	Retrieval systems, context management, information persistence
Complexity Handling	Multi-step reasoning failures, decomposition errors	Problem decomposition limits, insufficient goal management	Hierarchical reasoning, constraint satisfaction, goal tracking
Hallucination Control	High fabrication rates, overconfident assertions	Uncertainty estimation deficits, poor calibration mechanisms	Uncertainty quantification, fact verification, boundary recognition
Knowledge Transfer	Domain adaptation failures, poor analogical reasoning	Representation inflexibility, insufficient abstraction	Abstraction capabilities, meta-learning, analogical mapping
Processing Efficiency	High computational costs relative to output quality	Algorithmic inefficiencies, brute-force search patterns	Adaptive computation, efficiency-quality balancing
Process-Outcome Alignment	Process-output inconsistency, unstable reasoning traces	Internal representation issues, post-hoc rationalization	Interpretable architectures, process monitoring, consistency training

Note. Poor performance indicators and development recommendations are theoretical; empirical validation of their predictive utility is a target for future empirical work.

For deployment decisions. The cognitive profile enables principled model-to-task matching. High-stakes applications requiring reliability under uncertainty (clinical decision support, legal analysis) should prioritize Hallucination Control, Process-Outcome Alignment, and Knowledge Transfer. High-throughput applications may tolerate lower Memory Integration and Complexity Handling in exchange for Processing Efficiency gains.

For AI governance. Policymakers assessing AI capabilities and risks currently lack principled tools for evaluating cognitive processes. MAAC dimensions map directly onto governance concerns: Hallucination Control addresses trustworthiness requirements in regulated domains; Processing Ef-

ficiency relates to environmental sustainability mandates; Process-Outcome Alignment provides an empirical basis for claims about AI reasoning (Bommasani et al., 2021; Raji et al., 2022).

One concrete use case clarifies what Process-Outcome Alignment adds in practice. Consider a regulator auditing a clinical decision-support model. A high POA score indicates that the model's reasoning behavior remains structurally consistent across paraphrased variants of the same clinical problem. A low POA score despite acceptable accuracy indicates that the system may reach correct answers through unstable or weakly grounded reasoning processes—creating deployment risk under paraphrase or distribution shift. That distinction affects whether the model should be approved, approved only for bounded use, or subjected to additional review.

6.3. Limitations

Four limitations require acknowledgment. First, all validation analyses in this paper are theoretical. The framework's scientific standing depends on future empirical studies confirming that these theoretical properties hold in practice (Messick, 1995).

Second, the LLM judge scoring architecture introduces a methodological dependency: dimensional scores are produced by LLM judges evaluating LLM outputs. While near-perfect inter-judge agreement is achievable for structural complexity scoring under tightly constrained rubric conditions, cognitive dimension scoring is more interpretively demanding and may exhibit lower agreement. This dependency should be treated as a future validation target rather than as an assumption already resolved by the present paper. The theoretical contribution of MAAC does not depend on having already proven that LLM judges are valid scorers; it depends on specifying what must be scored, why those dimensions belong together, and what empirical patterns would support or disconfirm the scoring architecture.

Third, the framework is currently validated for natural language AI systems producing text outputs. Multimodal systems, embodied agents, and systems with non-linguistic outputs may require dimension-specific adaptation.

Fourth, the nine-dimensional structure assumes cognitive modularity. If AI cognitive processing is highly integrated, factor analysis in future empirical work may reveal a dominant general factor rather than nine discriminable dimensions.

6.4. Framework Vulnerability and Safeguards

MAAC's diagnostic utility creates potential gaming risks: (1) superficial optimization for dimensional scores without genuine cognitive improvement; (2) selective reporting of favorable dimensional profiles; and (3) prompt engineering to exploit specific measurement scenarios.

Mitigation strategies address each risk. Dynamic scenario generation—regular updating of assessment content using complexity-controlled scenario design principles—prevents memorization, as the complexity-validated scenario space is too large to memorize. Cross-validation requirements mandate that dimensional claims be validated across multiple measurement approaches. Longitudinal consistency checks detect superficial score optimization. Regulatory considerations should emphasize diagnostic rather than comparative use, preventing MAAC from becoming a competitive ranking system that incentivizes gaming over genuine cognitive development.

6.5. Future Directions

Four directions for future research are identified. First, scenario-interactivity effects should be tested as independent predictors of AI performance on generated scenarios. Second, the model panel should be expanded to 10+ architectures spanning open-source, commercial, and multimodal systems. Third, longitudinal studies examining how MAAC dimensions evolve during training and scale with model size will provide insights into the development of artificial cognitive capabilities. Fourth, connections between MAAC dimensional scores and circuit-level findings from mechanistic interpretability research would provide convergent validity evidence (Clark et al., 2019; Doshi-Velez and Kim, 2017).

7. Conclusion

This paper introduced the Multi-Dimensional Assessment for AI Cognition (MAAC), a theoretically grounded framework for evaluating AI systems through the lens of cognitive processes rather than task outcomes. MAAC defines nine cognitive dimensions—Cognitive Load, Tool Execution, Content Quality, Memory Integration, Complexity Handling, Hallucination Control, Knowledge Transfer, Processing Efficiency, and Process-Outcome Alignment—each anchored in established cognitive science theory at Marr's (1982) algorithmic level.

Five theoretical analyses provide preliminary support for the framework's conceptual coherence. Content validity is argued through systematic dimension-to-theory mapping against 108 retained sources. Coverage breadth is examined through a construct matrix demonstrating exhaustiveness and non-redundancy. Diagnostic utility is illustrated through a worked example showing that identical benchmark accuracy can mask fundamentally different cognitive architectures. Structural validity remains a programmatic objective, with empirical confirmation deferred to future work.

Three theoretical contributions follow. First, process-oriented cognitive assessment of AI systems is theoretically coherent and scientifically falsifiable—the faithfulness concern that undermines chain-of-thought evaluation is addressed by treating process-outcome alignment as a measurable dimension rather than an assumption. Second, classical cognitive science frameworks translate productively to artificial cognitive assessment when applied at the algorithmic level. Third, the nine-dimensional cognitive profile constitutes a new unit of analysis in AI evaluation—one that contextualizes benchmark accuracy by revealing the cognitive architecture that produced it.

By shifting evaluation focus from what AI systems produce to how they think, MAAC advances the field toward more rigorous, trustworthy, and diagnostically useful assessment of artificial intelligence.

Funding

No funding was received for this research.

Declaration on Use of AI-Assisted Writing Tools

Large language model (LLM) tools were used to assist with manuscript preparation, including language editing and structural refinement. All intellectual content, theoretical development, and analytical conclusions are the work of the human authors, who take full accountability for the final version of the manuscript.

Data Availability

This paper presents a purely theoretical framework. No datasets were generated or analyzed during the preparation of this work. The framework definitions, dimension specifications, and theoretical validation analyses are fully described herein and require no supplementary data file.

References

Akhtar, M., Reuel, A., Soni, P., et al., 2026. When AI benchmarks plateau: A systematic study of benchmark saturation. arXiv preprint arXiv:2602.16763 doi:10.48550/arXiv.2602.16763.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of the mind. Psychological Review 111, 1036–1060. doi:10.1037/0033-295X.111.4.1036.

Arksey, H., O'Malley, L., 2005. Scoping studies: Towards a methodological framework. International Journal of Social Research Methodology 8, 19–32. doi:10.1080/1364557032000119616.

Baddeley, A., 1992. Working memory. Science 255, 556–559. doi:10.1126/science.1736359.

Baddeley, A., 2000. The episodic buffer: A new component of working memory? Trends in Cognitive Sciences 4, 417–423. doi:10.1016/S1364-6613(00)01538-2.

Baddeley, A., 2003. Working memory: Looking back and looking forward. Nature Reviews Neuroscience 4, 829–839. doi:10.1038/nrn1201.

Barnett, S.M., Ceci, S.J., 2002. When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin 128, 612–637. doi:10.1037/0033-2909.128.4.612.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. doi:10.1145/3442188.3445922.

Bommasani, R., Hudson, D.A., Adeli, E., et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 doi:10.48550/arXiv.2108.07258.

Borgeaud, S., Mensch, A., Hoffmann, J., et al., 2022. Improving language models by retrieving from trillions of tokens, in: Proceedings of the 39th International Conference on Machine Learning, pp. 2206–2240. URL: https://proceedings.mlr.press/v162/borgeaud22a.html.

Campbell, D.J., 1988. Task complexity: A review and analysis. Academy of Management Review 13, 40–52. doi:10.5465/amr.1988.4306775.

Clark, A., Chalmers, D., 1998. The extended mind. Analysis 58, 7–19. doi:10.1093/analys/58.1.7.

Clark, K., Khandelwal, U., Levy, O., Manning, C.D., 2019. What does BERT look at? an analysis of BERT's attention, in: Proceedings of the 2019 ACL Workshop BlackboxNLP, pp. 276–286. doi:10.18653/v1/W19-4828.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychological Bulletin 52, 281–302. doi:10.1037/h0040957.

Crossley, S.A., Kyle, K., McNamara, D.S., 2016. The tool for the automatic analysis of text cohesion (TAACO). Behavior Research Methods 48, 1227–1237. doi:10.3758/s13428-015-0651-7.

DeVellis, R.F., 2017. Scale Development: Theory and Applications. 4th ed., SAGE Publications.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 doi:10.48550/arXiv.1702.08608.

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., Fernandez-Llorca, D., 2025. Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. arXiv preprint arXiv:2502.06559 doi:10.48550/arXiv.2502.06559.

Ethayarajh, K., Jurafsky, D., 2020. Utility is in the eye of the user: A critique of NLP leaderboards, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4846–4853. doi:10.18653/v1/2020.emnlp-main.393.

Furr, R.M., 2018. Psychometrics: An Introduction. 3rd ed., SAGE Publications.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: Proceedings of the 33rd International Conference on Machine Learning, pp. 1050–1059. URL: https://proceedings.mlr.press/v48/gal16.html.

Gentner, D., 1983. Structure-mapping: A theoretical framework for analogy. Cognitive Science 7, 155–170. doi:10.1207/s15516709cog0702_3.

Griffiths, T.L., Lieder, F., Goodman, N.D., 2015. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Topics in Cognitive Science 7, 217–229. doi:10.1111/tops.12142.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html.

Halford, G.S., Baker, R., McCredden, J.E., Bain, J.D., 2005. How many variables can humans process? Psychological Science 16, 70–76. doi:10.1111/j.0956-7976.2005.00782.x.

Halliday, M.A.K., Hasan, R., 1976. Cohesion in English. Longman.

Hendrycks, D., Burns, C., Basart, S., et al., 2021. Measuring massive multitask language understanding, in: Proceedings of the International Conference on Learning Representations. doi:10.48550/arXiv.2009.03300.

Hoffmann, J., Borgeaud, S., Mensch, A., et al., 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 doi:10.48550/arXiv.2203.15556.

Holyoak, K.J., Morrison, R.G. (Eds.), 2012. The Oxford Handbook of Thinking and Reasoning. Oxford University Press.

Huang, L., Yu, W., Ma, W., et al., 2023. A survey on hallucination in large language models. arXiv preprint arXiv:2311.05232 doi:10.48550/arXiv.2311.05232.

Hutchins, E., 1995. Cognition in the Wild. MIT Press.

Ji, Z., Lee, N., Frieske, R., et al., 2023. Survey of hallucination in natural language generation. ACM Computing Surveys 55, 1–38. doi:10.1145/3571730.

Just, M.A., Carpenter, P.A., 1992. A capacity theory of comprehension: Individual differences in working memory. Psychological Review 99, 122–149. doi:10.1037/0033-295X.99.1.122.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux.

Kaplan, J., McCandlish, S., Henighan, T., et al., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 doi:10.48550/arXiv.2001.08361.

Kapoor, S., Stroebl, B., Siegel, Z.S., Nadgir, N., Narayanan, A., 2024. AI agents that matter. arXiv preprint arXiv:2407.01502 doi:10.48550/arXiv.2407.01502.

Kovacs, K., Conway, A.R.A., 2016. Process overlap theory: A unified account of the general factor of intelligence. Psychological Inquiry 27, 151–177.

Ku, A.Y., Campbell, D., Bai, X., et al., 2025. Levels of analysis for large language models. arXiv preprint arXiv:2503.13401 doi:10.48550/arXiv.2503.13401.

Lewis, P., Perez, E., Piktus, A., et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, pp. 9459–9474. doi:10.48550/arXiv.2005.11401.

Liang, P., Bommasani, R., Lee, T., et al., 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 doi:10.48550/arXiv.2211.09110.

Lin, S., Hilton, J., Evans, O., 2022. TruthfulQA: Measuring how models mimic human falsehoods, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 3214–3252. doi:10.18653/v1/2022.acl-long.229.

Ma, Z.R., Guo, Y.X., Xiao, Y., 2026. Beyond accuracy scores: Toward process-oriented evaluation of artificial intelligence clinical reasoning in clinical workflow integration. International Journal for Quality in Health Care 38, mzag034. doi:10.1093/intqhc/mzag034.

Magar, I., Schwartz, R., 2022. Data contamination: From memorization to exploitation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 157–165. doi:10.18653/v1/2022.acl-short.18.

Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Co.

Maynez, J., Narayan, S., Bohnet, B., McDonald, R., 2020. On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906–1919. doi:10.18653/v1/2020.acl-main.173.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., Graesser, A.C., 2010. Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes 47, 292–330. doi:10.1080/01638530902959943.

Mehta, S., 2025. Beyond accuracy: A multi-dimensional framework for evaluating enterprise agentic AI systems. arXiv preprint arXiv:2511.14136 doi:10.48550/arXiv.2511.14136.

Messick, S., 1995. Validity of psychological assessment. American Psychologist 50, 741–749. doi:10.1037/0003-066X.50.9.741.

Miller, G.A., 1956. The magical number seven, plus or minus two. Psychological Review 63, 81–97. doi:10.1037/h0043158.

Mitchell, M., 2021. Why AI is harder than we think, in: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 4–10. doi:10.1145/3449639.3465421.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., et al., 2010. The COSMIN checklist for assessing the methodological quality of studies on measurement properties. Quality of Life Research 19, 539–549. doi:10.1007/s11136-010-9606-8.

Nakano, R., Hilton, J., Balaji, S., et al., 2021. WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332 doi:10.48550/arXiv.2112.09332.

Newell, A., 1990. Unified Theories of Cognition. Harvard University Press.

Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall.

Perkins, D.N., Salomon, G., 1992. Transfer of learning, in: International Encyclopedia of Education. 2nd ed.. Pergamon Press.

Peters, M.D.J., Godfrey, C., McInerney, P., et al., 2020. Chapter 11: Scoping reviews, in: Aromataris, E., Munn, Z. (Eds.), JBI Manual for Evidence Synthesis. JBI. doi:10.46658/JBIMES-20-12.

Prinsen, C.A.C., Mokkink, L.B., Bouter, L.M., et al., 2018. COSMIN guideline for systematic reviews of patient-reported outcome measures. Quality of Life Research 27, 1147–1157. doi:10.1007/s11136-018-1798-3.

Raji, I.D., Kumar, I.E., Horowitz, A., Selbst, A., 2022. The fallacy of AI functionality, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 959–972. doi:10.1145/3531146.3533158.

Rogers, A., Kovaleva, O., Rumshisky, A., 2020. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics 8, 842–866. doi:10.1162/tacl_a_00349.

Saparov, A., He, H., 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought, in: Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023). URL: https://openreview.net/forum?id=qFVVBzXxR2V.

Sardana, N., Portes, J., Doubov, S., Frankle, J., 2024. Beyond Chinchilla-Optimal: Accounting for inference in language model scaling laws, in: Proceedings of the 41st International Conference on Machine Learning (ICML 2024). doi:10.48550/arXiv.2401.00448.

Schick, T., Dwivedi-Yu, J., Dessì, R., et al., 2024. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems 36. doi:10.48550/arXiv.2302.04761.

Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O., 2020. Green AI. Communications of the ACM 63, 54–63. doi:10.1145/3381831.

Shi, W., Min, S., Yasunaga, M., et al., 2023. REPLUG: Retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652 doi:10.48550/arXiv.2301.12652.

Simon, H.A., 1956. Rational choice and the structure of the environment. Psychological Review 63, 129–138. doi:10.1037/h0042769.

Simon, H.A., 1972. Theories of bounded rationality. Decision and Organization 1, 161–176.

Srivastava, A., Rastogi, A., Rao, A., et al., 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 doi:10.48550/arXiv.2206.04615.

Strubell, E., Ganesh, A., McCallum, A., 2019. Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650. doi:10.18653/v1/P19-1355.

Sweller, J., 1988. Cognitive load during problem solving: Effects on learning. Cognitive Science 12, 257–285. doi:10.1207/s15516709cog1202_4.

Sweller, J., van Merriënboer, J.J.G., Paas, F., 2019. Cognitive architecture and instructional design: 20 years later. Educational Psychology Review 31, 261–292. doi:10.1007/s10648-019-09465-5.

Terwee, C.B., Prinsen, C.A.C., Chiarotto, A., et al., 2018. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A delphi study. Quality of Life Research 27, 1159–1170. doi:10.1007/s11136-018-1829-0.

Turpin, M., Michael, J., Perez, E., Bowman, S.R., 2024. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems 36. doi:10.48550/arXiv.2305.04388.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. Science 185, 1124–1131. doi:10.1126/science.185.4157.1124.

Wang, A., Pruksachatkun, Y., Nangia, N., et al., 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems, in: Advances in Neural Information Processing Systems. doi:10.48550/arXiv.1905.00537.

Wang, X., Wei, J., Schuurmans, D., et al., 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 doi:10.48550/arXiv.2203.11171.

Wei, J., Wang, X., Schuurmans, D., et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35, 24824–24837. doi:10.48550/arXiv.2201.11903.

Wood, R.E., 1986. Task complexity: Definition of the construct. Organizational Behavior and Human Decision Processes 37, 60–82. doi:10.1016/0749-5978(86)90044-0.

Yao, S., Yu, D., Zhao, J., et al., 2024. Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems 36. doi:10.48550/arXiv.2305.10601.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2020. BERTScore: Evaluating text generation with BERT, in: Proceedings of the International Conference on Learning Representations. doi:10.48550/arXiv.1904.09675.

Zhou, L., Pacchiardi, L., Martínez-Plumed, F., et al., 2026. General scales unlock AI evaluation with explanatory and predictive power. Nature 652, 58–67. doi:10.1038/s41586-026-10303-2.