Theoretical Framework for Completeness and Consistency of Knowledge in Generative AI Large Language Models

Thuan Luong Nguyen
University of North Texas, Denton, Texas, USA

Thuan.Nguyen@unt.edu

Abstract: Generative AI Large Language Models (LLMs) such as GPT-4, Claude, and Gemini are reshaping knowledge work across disciplines. Yet these systems exhibit a puzzling paradox: they can pass rigorous professional examinations while simultaneously failing elementary reasoning tasks. This paper presents a condensed theoretical framework explaining how knowledge is created, stored, and retrieved in LLMs through Stochastic Knowledge Aggregation (SKA) – a process fundamentally different from the Systematic Knowledge Scaffolding (SKS) that characterizes human learning. We introduce 19 foundational concepts, three formal theories, and a set of propositions collectively forming the Jagged Knowledge Frontier (JKF) framework. Empirical cases validate the framework and illuminate implications for Knowledge Management (KM). The paper argues that understanding AI knowledge creation is essential for bridging traditional human KM with the emerging discipline of AI Knowledge Management, and for designing governance structures that account for the inherent incompleteness and inconsistency of LLM knowledge.

Keywords: Large language models, Knowledge creation, AI knowledge management, Stochastic knowledge aggregation, Jagged knowledge frontier, AI knowledge consistency, Wharton paradox

1. Introduction

The rapid deployment of Generative AI Large Language Models has positioned these systems as pivotal actors in contemporary Knowledge Management ecosystems. LLMs trained on trillions of tokens exhibit capabilities that confound traditional cognitive benchmarks: GPT-4 passes the United States Bar Examination in the 90th percentile (Katz et al., 2023) and demonstrates competence across an astonishing breadth of academic and professional domains (Brown et al., 2020; Bommasani et al., 2021). Yet the same models fail elementary math (Hassabis, 2025), confabulate fabricated legal citations (Dahl et al., 2024), and produce self-contradictory outputs within a single conversation (Elazar et al., 2021).

This paradox – named as the Wharton Paradox by Nguyen (2026), following Terwiesch's (2023) landmark assessment of ChatGPT on a Wharton MBA examination – is not an engineering defect that improved training will simply resolve. It is a structural consequence of how knowledge is created within these systems. LLMs do not learn the way human learners do. Where human cognition builds on Systematic Knowledge Scaffolding (SKS) – hierarchically organised, causally grounded, and epistemically calibrated – LLMs acquire knowledge through Stochastic Knowledge Aggregation (SKA): a probabilistic, co-occurrence-based process that produces islands of dense competence interspersed with vast voids of ignorance, forming a characteristically Jagged Knowledge Frontier (JKF) (Nguyen, 2026).

This paper addresses three research questions:

- (RQ1) What mechanisms govern the creation of knowledge in generative AI LLMs?
- (RQ2) How do these mechanisms produce characteristic patterns of knowledge incompleteness and inconsistency?
- (RQ3) What are the implications for Knowledge Management theory and practice?

Answering these questions is urgent because organizations are integrating LLMs into their knowledge workflows at scale, often without a theoretical basis for anticipating where and why these systems will fail. The framework developed here provides that basis and charts a course toward a unified discipline of AI Knowledge Management that complements existing human-centred KM traditions.

2. Literature Review

Research on LLM capabilities has progressed rapidly since the introduction of the Transformer architecture (Vaswani et al., 2017) and the subsequent emergence of large-scale pre-trained models (Brown et al., 2020). Scaling laws established that model performance improves predictably with compute, data, and parameters (Kaplan et al., 2020), generating a research culture focused on benchmark achievement that obscures the uneven, jagged nature of LLM competence. The hallucination literature documents a pervasive tendency of LLMs to generate plausible-sounding but factually incorrect content (Ji et al., 2023). Hallucination arises from the statistical nature of next-token prediction, which optimises for linguistic fluency rather than factual

grounding. Closely related is the consistency problem: LLMs produce contradictory outputs when semantically equivalent questions are phrased differently (Elazar et al., 2021), and fail to retrieve relational knowledge bidirectionally – a phenomenon formalised as the Reversal Curse (Allen-Zhu & Li, 2023).

Knowledge representation theory distinguishes between declarative knowledge (knowing that), procedural knowledge (knowing how), and epistemic knowledge (knowing what one knows). LLMs appear to possess rich declarative-like associations but lack genuine epistemic calibration (Kadavath et al., 2022). The Symbol Grounding Problem (Harnad, 1990) illuminates a deeper issue: LLM token representations lack sensorimotor grounding, meaning linguistic competence does not imply semantic comprehension (Bisk et al., 2020). Knowledge Management scholarship emphasises embodied, socially situated knowing; LLMs engage exclusively with explicit, textual representations, rendering the tacit knowledge dimension structurally inaccessible. Theory-building methodology for applied disciplines (Whetten, 1989; Gregor, 2006) guides the framework's architecture, connecting AI research, cognitive science, and KM theory through interdisciplinary integration.

3. Technical Background

The dominant LLM architecture is the Transformer (Vaswani et al., 2017), which processes token sequences via multi-head self-attention. Attention allows each token to aggregate contextual information from all other tokens, producing context-sensitive representations that are then passed through feed-forward sub-layers and layer normalisation. The feed-forward layers function as distributed key-value memories storing factual associations (Geva et al., 2021), providing the primary substrate for LLM knowledge.

Pre-training constitutes the foundational knowledge-creation event. The model processes hundreds of billions to trillions of tokens from web pages, books, scientific articles, and code repositories, optimizing next-token prediction via gradient descent. Crucially, no explicit semantic representation is encoded; instead, billions of parameters converge on weight configurations that implicitly capture statistical regularities across the corpus. The resulting knowledge is distributed, sub-symbolic, and non-propositional: knowledge emerges from parameter interactions rather than discrete storage units (Elhage et al., 2021), and no single parameter stores a discrete fact.

Scaling pre-training yields emergent capabilities appearing at certain compute thresholds (Wei et al., 2022). However, many apparent emergent abilities may be artefacts of evaluation metrics rather than genuine phase transitions (Schaeffer et al., 2023), reinforcing the characterisation of LLM knowledge as fundamentally probabilistic.

Post-training processes critically shape deployed knowledge. Instruction fine-tuning and Reinforcement Learning from Human Feedback (RLHF; Christiano et al., 2017) align model outputs to human preferences but optimise for perceived quality rather than epistemic accuracy, introducing systematic bias toward confident, fluent responses regardless of factual grounding – a key hallucination mechanism (Lin et al., 2022). Training on model-generated data further compounds errors over generations, degrading knowledge fidelity in successive LLM iterations (Shumailov et al., 2023).

LLM knowledge is therefore characterised by four fundamental technical properties: (1) statistical rather than causal; (2) distributed rather than localized; (3) context-sensitive rather than invariant; and (4) cutoff-bounded, rendering the model epistemically static in a dynamically evolving world.

4. Methodology

To address the research questions, this paper employs **Constructive Theoretical Synthesis (CTS),** a rigorous, interdisciplinary research methodology that integrates disparate findings, concepts, and empirical observations into a unified explanatory structure (Jaccard & Jacoby, 2010). CTS employs a multi-paradigm theory-building methodology appropriate for constructing explanatory frameworks in applied disciplines where phenomena are complex and emergent (Gregor, 2006; Lynham, 2002). The methodology proceeds through four phases: conceptual synthesis, formal theory construction, empirical case triangulation, and criterion-based evaluation.

In the conceptual synthesis phase, we conducted an integrative literature review spanning AI/ML research, cognitive science, philosophy of mind, and knowledge management. We applied interdisciplinary integration techniques (Repko & Szostak, 2021) to reconcile conceptual frameworks from disparate disciplines, resolving tensions between information systems and cognitive science definitions of knowledge.

The formal theory construction phase follows Whetten's (1989) criteria and Dubin's (1978) theory-building model. We defined nineteen foundational concepts with operational precision, articulated relational propositions among them, provided causal logic to ground these relationships, and specified boundary conditions that delimit the theory's scope.

Empirical triangulation was applied across multiple data types to validate theoretical propositions: (1) empirical benchmark results documenting LLM capabilities and failures; (2) reported incidents of harmful LLM outputs in legal and journalistic contexts (Dahl et al., 2024; BBC, 2025a); (3) controlled experimental studies of knowledge inconsistency (Elazar et al., 2021; Allen-Zhu & Li, 2023); and (4) widely replicated demonstrations of LLM reasoning failures. This multi-source triangulation strengthens confidence that theoretical constructs capture genuine phenomena rather than isolated edge cases.

Theory evaluation employed Gregor's (2006) typology of IS theories alongside criteria of falsifiability, utility, and parsimony. The resulting framework is classified as a Theory of Explanation and Prediction, generating testable propositions – for example, that knowledge inconsistency is higher for knowledge acquired near the training cutoff than for well-represented historical knowledge – that future empirical work can investigate.

5. Conceptual Foundations

The framework rests on 19 foundational concepts that collectively characterize the epistemic structure of LLM knowledge. We present them in five thematic clusters, noting that the concepts are interdependent rather than orthogonal.

5.1 Knowledge Creation Mechanism

Concept 1 – Stochastic Knowledge Aggregation (SKA): The primary mechanism by which LLMs acquire knowledge during pre-training. SKA operates through gradient-based optimization across next-token prediction objectives applied to massive text corpora. Unlike human learning, SKA aggregates statistical associations without causal understanding, semantic grounding, or hierarchical organization. The result is a distributed parameter configuration that implicitly encodes vast but unevenly distributed knowledge.

Concept 2 – Systematic Knowledge Scaffolding (SKS): The contrasting mechanism by which humans acquire knowledge. SKS involves hierarchically organised conceptual structures built through developmental stages (Piaget, 1952; Bloom, 1968), prerequisite-driven learning sequences (Gagné, 1968), and iterative mastery (Carroll, 1963; Bloom, 1971). SKS produces knowledge that is causally coherent, epistemically calibrated, and capable of supporting genuine novel inference.

Concept 3 – Knowledge Materialization Event: The training process itself, which constitutes the moment at which statistical patterns in text corpora are crystallized into model parameters. Unlike human knowledge, which updates continuously through lived experience, LLM knowledge materialises at a discrete point in time and remains static thereafter, absent further training.

Concept 4 – Stochastic Activation Patterns (SAP): At inference time, knowledge retrieval is itself probabilistic. For a given query, the model's attention and feed-forward mechanisms activate different parameter subsets in response to subtle variations in phrasing, context length, and sampling temperature, producing non-deterministic outputs that reflect the stochastic nature of both acquisition (SKA) and retrieval (SAP).

5.2 Knowledge Topology

Concept 5 – Jagged Knowledge Frontier (JKF): The characteristic topography of LLM knowledge, marked by steep gradients between regions of competence and incompetence. The JKF is not random: it reflects the distribution of representational density in training data, whereby well-documented domains (e.g., English literature, mainstream mathematics) exhibit dense, reliable knowledge while niche, emerging, or underrepresented domains exhibit sparse, unreliable knowledge.

Concept 6 – Knowledge Islands: Localized regions within the JKF where representational density is sufficient to support consistent, accurate, and compositionally generalizable responses. Knowledge Islands are domains where training data is abundant, high-quality, diverse exposures.

Concept 7 – Knowledge Voids: Regions within the JKF where representational density is insufficient to support reliable knowledge retrieval. In Knowledge Voids, the model produces outputs by extrapolating from adjacent Knowledge Islands, resulting in plausible-sounding but faulty generations – the proximate mechanism of hallucination.

Concept 8 – Knowledge Density and Sparsity: The quantitative dimension of the JKF, expressing the frequency and diversity of training data exposures relevant to a given domain. High density corresponds to Knowledge Islands; low density corresponds to Knowledge Voids. Density is not equivalent to accuracy: high-density regions containing systematically biased training data produce dense but systematically erroneous knowledge.

5.3 Knowledge Quality Dimensions

Concept 9 – Knowledge Completeness: The degree to which an LLM's knowledge of a domain covers the domain's full conceptual space without gaps. Completeness is relative to a reference ontology and is inherently partial for all LLMs. Critically, the model itself cannot reliably identify its own completeness gaps – a property we term epistemic opacity.

Concept 10 – Knowledge Consistency: The degree to which an LLM produces logically coherent, non-contradictory outputs across semantically equivalent queries. Inconsistency arises from Stochastic Activation Patterns interacting with distributed knowledge representations; the same underlying knowledge is accessed through different parameter pathways depending on query phrasing.

Concept 11 – Semantic Coherence vs. Factual Accuracy: LLMs are optimized for semantic coherence – outputs that are grammatically fluent and contextually plausible – but not for factual accuracy. These dimensions are dissociable: highly coherent outputs can be factually wrong, as documented in hallucination studies (Ji et al., 2023; OpenAI, 2025).

Concept 12 – Knowledge Calibration: The alignment between an LLM's expressed confidence and its actual accuracy. Well-calibrated models express high confidence in correct outputs and low confidence in incorrect ones. Current LLMs are systematically overconfident (Guo et al., 2017), a property exacerbated by RLHF's optimization for human preference over epistemic truth.

5.4 Knowledge Boundary and Transfer Phenomena

Concept 13 – Knowledge Boundary Ambiguity: The inability of an LLM – and often its users – to reliably identify where Knowledge Islands end and Knowledge Voids begin. This ambiguity creates dangerous deployment scenarios in which users receive confident-sounding responses from within Knowledge Voids.

Concept 14 – The Reversal Curse: The empirically documented failure of LLMs to retrieve relational knowledge bidirectionally. A model trained on "A is related to B" does not automatically learn "B is related to A" (Allen-Zhu & Li, 2023). This failure reflects SKA's sensitivity to training data directionality – a fundamental asymmetry absent from human knowledge structures.

Concept 15 – Cross-Domain Knowledge Interference: The phenomenon whereby dense training on one domain degrades knowledge quality in adjacent domains through parameter competition. Because LLM knowledge is distributed across shared parameters, gains in one region of the knowledge space can induce losses in another.

Concept 16 – Knowledge Persistence vs. Volatility: The differential stability of LLM knowledge across time and query variations. Core Knowledge Islands exhibit persistence – consistent retrieval across diverse phrasings. Knowledge Voids and boundary regions exhibit volatility – retrieval outcomes, highly sensitive to minor input variations.

5.5 Epistemic and Metacognitive Properties

Concept 17 – Epistemic Opacity: The structural inability of an LLM to access or accurately report the boundaries, confidence levels, and provenance of its own knowledge. Unlike human knowers, who possess at least partial metacognitive awareness, LLMs lack a dedicated self-monitoring mechanism (Kadavath et al., 2022). Epistemic opacity is a direct consequence of the distributed, sub-symbolic nature of LLM knowledge.

Concept 18 – Intellectual Capability Level (ICL): A proposed metric for assessing the depth of an LLM's competence within a domain, analogous to Bloom's Taxonomy levels for human cognition (Bloom et al., 1956; Anderson & Krathwohl, 2001). ICL distinguishes surface retrieval (ICL-1), application (ICL-2), analysis (ICL-3), synthesis (ICL-4), and evaluation (ICL-5). LLMs characteristically exhibit high ICL on well-represented domains and near-zero ICL on Knowledge Voids.

Concept 19 – AI Knowledge Management (AI-KM): An emerging discipline concerned with understanding, governing, and improving the knowledge that AI systems create, store, retrieve, and communicate. AI-KM

bridges traditional human-centred KM (concerned with human knowledge creation, sharing, and utilization) and AI systems management, with the JKF framework providing its foundational theoretical architecture.

6. Theoretical Framework

The 19 concepts of Section 5 are integrated into three formal theories whose collective scope constitutes the JKF Framework for AI Knowledge Creation. Each theory is stated as a formal claim, followed by its foundational propositions.

6.1 Theory I: The Stochastic Knowledge Aggregation Theory (SKAT)

SKAT Claim: Knowledge in LLMs is created exclusively through Stochastic Knowledge Aggregation and cannot, by any amount of scaling, achieve the epistemic properties characteristic of Systematic Knowledge Scaffolding. The knowledge topology produced by SKA is necessarily jagged and structurally distinct from human knowledge.

- **P 1.1 – Knowledge Formation Proposition**: For any domain D, the quality of an LLM's knowledge of D is a monotonically increasing function of the density, diversity, and quality of D-relevant training data, and a non-monotonic function of model scale.
- **P 1.2 – Causal Opacity Proposition**: Because SKA does not require causal understanding as a training signal, LLMs cannot reliably distinguish causal from spurious correlations within Knowledge Islands.
- **P 1.3 – Grounding Deficit Proposition**: LLM knowledge is ungrounded in perceptual, sensorimotor, or social experience; all knowledge is derived from linguistic co-occurrence statistics alone (Harnad, 1990; Bisk et al., 2020).
- **P 1.4 – Epistemic Incommensurability Proposition**: Because SKA and SKS produce structurally different knowledge representations, human epistemic frameworks – including educational taxonomies (Anderson & Krathwohl, 2001), mastery learning (Bloom, 1968), and zone of proximal development (Vygotsky, 1978) – cannot be directly applied to LLM knowledge assessment without theoretical adaptation.

6.2 Theory II: The Knowledge Completeness–Consistency Theory (KCCT)

KCCT Claim: The inherent properties of SKA produce systematic and predictable incompleteness and inconsistency in LLM knowledge. These deficits are not random but structured by the JKF, and cannot be fully resolved through post-training alignment techniques alone.

- **P 2.1 – Incompleteness Proposition**: Every LLM knowledge base K is incomplete with respect to every non-trivial domain D; the incompleteness is maximal in Knowledge Voids and minimal in Knowledge Islands.
- **P 2.2 – Inconsistency Propagation Proposition**: Knowledge inconsistency is highest at JKF boundaries – transitions between Knowledge Islands and Knowledge Voids – where different phrasings probabilistically activate different parameter pathways, yielding contradictory outputs.
- **P. 2.3 – Confidence-Accuracy Decoupling Proposition**: LLM confidence (as expressed in linguistic markers such as "certainly," "clearly," or "it is known that") is not calibrated to factual accuracy and is elevated in Knowledge Voids by RLHF's preference for assertive-sounding outputs.
- **P. 2.4 – Reversal Asymmetry Proposition**: Relational knowledge in LLMs is directionally asymmetric; training on "A implies B" does not guarantee reliable retrieval of "B implies A" (Allen-Zhu & Li, 2023), constituting a structural inconsistency without a human analogue.

6.3 Theory III: The Jagged Knowledge Frontier Management Theory (JKFMT)

JKFMT Claim: Effective Knowledge Management of LLM-integrated systems requires explicit mapping of the JKF, governance protocols aligned with Knowledge Island and Knowledge Void topology, and institutional frameworks that bridge human KM and AI-KM.

- **P 3.1 – JKF Mapping Proposition**: Organizations deploying LLMs can systematically identify Knowledge Islands and Voids through structured domain-specific evaluation protocols, enabling risk-stratified governance.
- **P. 3.2 – Human-AI KM Complementarity Proposition**: Human KM processes (tacit knowledge externalization, socialization, organizational memory) and AI-KM processes (SKA-based knowledge creation, JKF topology, epistemic opacity management) are complementary rather than substitutive; effective knowledge organizations must coordinate both.

- **P. 3.3 – Governance Alignment Proposition**: LLM governance policies that do not account for JKF structure will systematically fail; policies must specify acceptable use conditions relative to known Knowledge Islands rather than treating LLM capability as uniform across domains.
- **P. 3.4 – Epistemic Humility Proposition**: Organizations and individual users must cultivate calibrated epistemic humility – a disposition to treat LLM outputs as probabilistic estimates from an unknown region of the JKF rather than as authoritative knowledge – as a primary KM competence in AI-integrated work environments.

7. Empirical Validations

The JKF Framework draws validation from four empirical cases: the Wharton Paradox, where LLMs pass professional examinations yet fail elementary tasks (Terwiesch, 2023; r/ChatGpt, 2023); legal hallucinations involving fabricated citations (Dahl et al., 2024); the Reversal Curse confirming directional knowledge asymmetry (Allen-Zhu & Li, 2023); and the BBC misinformation audit (BBC, 2025a).

Beyond these documented cases, we conducted a controlled experiment to directly test the framework's predictive power. Using the theories, we predicted that even the latest, most powerful LLMs would fail at a simple task – counting every word in a document individually – despite their capacity for sophisticated intellectual work such as writing complex code and discussing quantum computing at an expert level. This prediction follows from SKAT Proposition 1.3 (Grounding Deficit): LLMs process tokens and high-dimensional embedding vectors rather than human-recognizable words (Vaswani et al., 2017), so the semantic integrity required for basic counting is structurally unavailable.

To ensure rigorous control, we developed Green Prompting Language (GPL), a novel Structured Prompting Language that provides step-by-step algorithmic instructions to eliminate ambiguity and prevent errors stemming from prompt misunderstanding rather than genuine capability limitations.

The experiment, conducted in early April 2026, tested two leading LLMs: Google's Gemini 3.1 Pro and Anthropic's Claude Opus 4.6 Extended. Each received a GPL prompt containing a 4,158-word research paper with instructions to count every word – a task any elementary student could complete. Both failed: Gemini reported 1,820 words; Claude reported 4,850 words.

Follow-up explanations revealed structurally significant admissions. Gemini acknowledged working on tokens rather than words and admitted it "simulated" a plausible-looking result. Claude admitted its answer was not computed but a fabricated estimate.

These results validate three core claims: the Wharton Paradox is caused by structural limitations, not engineering defects; foundational technologies such as tokenization simultaneously enable powerful capabilities and cause elementary failures (SKAT P1.1, P1.3); and the theories successfully predict LLM behavior, confirming the framework's explanatory and predictive utility (Gregor, 2006). The controlled experiment, combined with preceding observational evidence, strongly validates the proposed theoretical framework.

8. Discussion, Implications, and Applications

The JKF Framework carries substantial implications for Knowledge Management theory and practice, particularly for the ECKM community's focus on Knowledge Creation and Sharing.

Theoretical Implications for Knowledge Management: The framework demands an expansion of KM's conceptual vocabulary to accommodate AI as an active knowledge creator – not merely a knowledge tool. Existing KM theories are designed for human epistemic agents whose knowledge creation follows SKS principles. As LLMs are integrated into organisational knowledge workflows, the field requires theories that account for SKA-structured knowledge: jagged, inconsistent, opaque, and static. The concept of AI-KM (Concept 19) represents a proposed theoretical expansion that preserves the insights of traditional KM while accommodating the structural differences of AI knowledge. In particular, the SECI model must be extended to address LLMs as knowledge externalization engines that lack the tacit knowledge dimension essential to the original framework.

Managerial and Organisational Implications: JKFMT P3.1 implies that organisations should invest in systematic JKF mapping – domain-specific evaluations identifying Knowledge Islands and Voids – before integrating LLMs into high-stakes knowledge processes. JKFMT P3.3 implies that governance frameworks must move beyond generic "human oversight" recommendations to domain-specific, JKF-calibrated oversight protocols. For

example, an organisation deploying LLMs in legal research should implement mandatory expert review of all citation-bearing outputs, while the same organisation might safely reduce oversight for LLM-assisted drafting of routine communications within well-documented Knowledge Islands.

Implications for AI Knowledge Creation and Sharing Policy: At a societal level, the framework supports calls for informed scepticism toward AI-generated information (BBC, 2025b). The Wharton Paradox demonstrates that surface-level expertise signals – the fluency and confidence of LLM outputs – are not reliable proxies for factual accuracy. Public AI literacy initiatives must therefore prioritise JKF awareness, training users to identify Knowledge Islands versus Voids in domains relevant to their decisions. The framework also informs AI regulatory design: effective governance must address the structural epistemic opacity of LLMs (Concept 17) and mandate disclosure mechanisms that enable users to appropriately calibrate their trust.

9. Limitations and Future Work

This paper presents a theoretical framework derived primarily from conceptual synthesis and empirical case triangulation. Several limitations warrant acknowledgement. First, the framework was developed primarily with reference to autoregressive, transformer-based LLMs of the GPT, Claude, and Gemini families. Future model families based on novel architectures may require revision. Second, empirical validation relies on publicly documented cases and controlled experiments rather than a systematic empirical study specifically designed to test the framework's propositions; future research should conduct targeted experiments to quantitatively measure Knowledge Island and Void boundaries. Third, the ICL metric (Concept 18) is proposed as a theoretical construct that needs a validated measurement instrument.

Future research directions arising from the framework are rich. Within AI research, the framework motivates: (1) mechanistic interpretability studies locating Knowledge Island and Void boundaries within model weight space (Elhage et al., 2021); (2) calibration methods reducing Confidence–Accuracy Decoupling (Kadavath et al., 2022; Guo et al., 2017); and (3) continual learning architectures enabling dynamic JKF updating without catastrophic forgetting (Jang et al., 2022; Behrouz et al., 2025). Within KM research, the framework motivates: (1) empirical studies of how JKF-aware governance affects organisational knowledge quality; (2) extension of the SECI model to AI knowledge creation; (3) development of KM maturity models incorporating AI-KM capabilities; and (4) cross-cultural studies of JKF topology variation across languages and cultural contexts. The ECKM community is uniquely positioned to advance this interdisciplinary research agenda.

10. Conclusion

This paper presents the Jagged Knowledge Frontier Framework – a theoretical architecture for understanding how knowledge is created, structured, and retrieved in Generative AI Large Language Models, and for managing the consequences of that knowledge structure across organisational and societal contexts. The framework's 19 foundational concepts and three formal theories – the Stochastic Knowledge Aggregation Theory (SKAT), the Knowledge Completeness–Consistency Theory (KCCT), and the Jagged Knowledge Frontier Management Theory (JKFMT) – collectively explain the Wharton Paradox and related phenomena as structural consequences of the SKA process rather than engineering defects amenable to simple correction.

The framework's most significant contribution to the Knowledge Management discipline is its articulation of AI Knowledge Management (AI-KM) as a theoretically grounded subdiscipline. As LLMs are integrated into organisational knowledge workflows at unprecedented speed and scale, KM practitioners and researchers require conceptual tools bridging the well-established literature on human knowledge creation and sharing with the fundamentally different epistemic architecture of AI systems. The JKF Framework provides those tools: explaining where LLMs will be reliable (Knowledge Islands), where they will fail (Knowledge Voids), how confidence and accuracy are decoupled (KCCT P2.3), and how governance must be structured responsibly (JKFMT).

For the ECKM community, whose mission is to advance the theory and practice of knowledge management in a rapidly evolving technological landscape, the emergence of AI as a knowledge-creating agent represents both a profound challenge and a generative research opportunity. Understanding how AI creates knowledge – through Stochastic Knowledge Aggregation, with all its attendant incompleteness, inconsistency, and opacity – is not a peripheral technical concern. It is the central Knowledge Management problem of the current era, and one that the KM community is uniquely equipped to address through its integrative, human-centered, and governance-aware theoretical traditions.

Ethics Declaration: Ethical clearance was not required for the research.

AI Declaration: AI tools were not used to create the paper.

References

Allen-Zhu, Z., and Li, Y. (2023) *The Reversal Curse: LLMs Trained on "A is B" Fail to Learn "B is A"*. Retrieved from https://arxiv.org/abs/2309.12288.

Anderson, L. W. and Krathwohl, D. R. (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.

BBC (2025a) *Representation of BBC News Content in AI Assistants*. Retrieved from https://www.bbc.co.uk/aboutthebbc/documents/bbc-researchinto-aiassistants.pdf.

BBC (2025b) *Don't Blindly Trust What AI Tells You, Says Google's Sundar Pichai*. Retrieved from https://www.bbc.com/news/articles/c8drzv37z4jo.

Behrouz, A., Zhong, P., Mirrokni, V. (2025) *Titans: Learning to Memorize at Test Time*. Retrieved from https://arxiv.org/abs/2501.00663.

Bisk, Y., Holtzman, A., Thomason, J., et al. (2020) *Experience Grounds Language*. In Proceedings of EMNLP (pp. 8718–8735).

Bloom, B. S. (1968) *Learning for Mastery*. Evaluation Comment, 1(2), 1–12.

Bloom, B. S. (1971) *Mastery Learning*. In J. H. Block (Ed.) *Mastery Learning: Theory and Practice* (pp. 47–63). Holt, Rinehart and Winston.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Handbook I: Cognitive Domain. New York: David McKay Company.

Bommasani, R., Hudson, D., Adeli, E., et al. (2021) *On the Opportunities and Risks of Foundation Models*. Retrieved from https://arxiv.org/abs/2108.07258.

Brown, T. B., Mann, B., Ryder, N., et al. (2020) *Language Models Are Few-Shot Learners*. Advances in Neural Information Processing Systems, 33, 1877–1901.

Carroll, J. B. (1963) *A Model of School Learning*. Teachers College Record, 64(8), 723–733.

Christiano, P. F., Leike, J., Brown, T., et al. (2017) *Deep Reinforcement Learning from Human Preferences*. Advances in NeurIPS, 30.

Dahl, M., Mrowca, V., Suzgun, M., and Ho, D. E. (2024) *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*. Journal of Legal Analysis, 16(1), 64–92.

Dubin, R. (1978) *Theory Building* (Rev. ed.). Free Press.

Elazar, Y., Kassner, N., Ravfogel, S., et al. (2021) *Measuring and Improving Consistency in Pretrained Language Models. Transactions of the ACL*, 9, 1012–1031.

Elhage, N., Nanda, N., Olsson, C., et al. (2021) *A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread*. Retrieved from https://transformer-circuits.pub/2021/framework/index.html.

Gagné, R. M. (1968) *Learning Hierarchies*. Educational Psychologist, 6(1), 1–9.

Geva, M., Schuster, R., Berant, J., and Levy, O. (2021) *Transformer Feed-Forward Layers Are Key-Value Memories*. Proceedings of EMNLP, 5484–5495.

Gregor, S. (2006) *The Nature of Theory in Information Systems*. MIS Quarterly, 30(3), 611–642.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017) *On Calibration of Modern Neural Networks*. In Proceedings of ICML (Vol. 70, pp. 1321–1330). PMLR.

Harnad, S. (1990) *The Symbol Grounding Problem*. Physica D: Nonlinear Phenomena, 42(1–3), 335–346.

Hassabis, D. (2025) *Demis Hassabis on the Frontiers of AI* [Video]. YouTube. Retrieved from https://www.youtube.com/watch?v=Uh-7YX8tkxI.

Jaccard, J., and Jacoby, J. (2010) Theory Construction and Model-Building Skills: A Practical Guide for Social Scientists. Guilford Press.

Jang, J., Ye, S., Yang, S., et al. (2022) *Towards Continual Knowledge Learning of Language Models*. In ICLR.

Ji, Z., Lee, N., Frieske, R., et al. (2023) *Survey of Hallucination in Natural Language Generation*. ACM Computing Surveys, 55(12), 1–38.

Kadavath, S., Conerly, T., Askell, A., et al. (2022) *Language Models (Mostly) Know What They Know*. Retrieved from https://arxiv.org/abs/2207.05221.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020) *Scaling Laws for Neural Language Models*. Retrieved from https://arxiv.org/abs/2001.08361.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. (2023) *GPT-4 Passes the Bar Exam*. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4389233.

Lin, S., Hilton, J., and Evans, O. (2022) *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. In Proceedings of ACL (Vol. 1, pp. 3214–3252).

Lynham, S. A. (2002) *The General Method of Theory-Building Research in Applied Disciplines*. Advances in Developing Human Resources, 4(3), 221–241.

Nguyen, T. L. (2026) *The Completeness and Consistency of Knowledge in Generative AI Large Language Models*. Retrieved from https://doi.org/10.6084/m9.figshare.31231300.

OpenAI. (2025) *Why Language Models Hallucinate*. Retrieved from https://arxiv.org/pdf/2509.04664.

Piaget, J. (1952) *The Origins of Intelligence in Children (M. Cook, Trans.)*. International Universities Press.

r/ChatGPT (2023). *Wife Is Always Correct*. Retrieved from https://www.reddit.com/r/ChatGPT/comments/10kw29n/wife_is_always_correct/

Repko, A. F., and Szostak, R. (2021) *Interdisciplinary Research: Process and Theory* (4th ed.). SAGE Publications.

Schaeffer, R., Miranda, B., and Koyejo, S. (2023) *Are Emergent Abilities of Large Language Models a Mirage?* Retrieved from https://arxiv.org/abs/2304.15004.

Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2023) *The Curse of Recursion: Training on Generated Data Makes Models Forget*. Retrieved from https://arxiv.org/abs/2305.17493.

Terwiesch, C. (2023) *Would Chat GPT Get a Wharton MBA?* Mack Institute for Innovation Management, University of Pennsylvania. Retrieved from https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) *Attention Is All You Need*. Advances in NeurIPS, 30, 5998–6008.

Vygotsky, L. S. (1978) *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wei, J., Tay, Y., Bommasani, R., et al. (2022) *Emergent Abilities of Large Language Models*. Transactions on Machine Learning Research.

Whetten, D. A. (1989) *What Constitutes a Theoretical Contribution?* Academy of Management Review, 14(4), 490–495.