IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

From External Grounding to Collective Intelligence: A Comprehensive Survey on Factuality, Knowledge Graphs, and Multi-Agent Orchestration in Large Language Models

Abishek S[1], Alfred Jaison[2], Ashish S[3], Karthik B[4], Anpu Alexander[5], Jinu Lenin[6]

Department of Computer Science & Engineering, College of Engineering Kottarakkara, Kollam, Kerala, India[1-6]

Abstract: Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) demonstrate remarkable generative capabilities but suffer from persistent hallucinations, factual drift, and reasoning degradation in high-consequence domains. While canonical Retrieval-Augmented Generation (RAG) mitigates parameter staleness, it introduces critical vulnerabilities including neural ranking miscalibration, numerical insensitivity, and channel noise cascading. This survey provides an exhaustive, taxonomy-driven synthesis of recent advances across four synergistic research pillars: evidential retrieval calibration, information-theoretic entropy bounds, structured knowledge graph (KG) grounding, and multi-agent collaborative orchestration with asymmetric verification against hallucination snowballing. Adhering strictly to PRISMA 2020 guidelines across twenty primary benchmarked architectures and over eighty foundational sources, we formalize the mathematical mechanics of epistemic uncertainty and cascading channel capacity. We present multidimensional comparative matrices evaluating accuracy, computational complexity, and inference latency, systematically analyze open research frontiers, and establish concrete engineering design trade-offs for deploying reliable, graph-grounded multi-agent intelligence in mission-critical environments.

Keywords: Large Language Models, Retrieval-Augmented Generation, Knowledge Graphs, Hallucination Mitigation, Evidential Deep Learning, Multi-Agent Systems, Information Theory

I. I. INTRODUCTION

Large Language Models (LLMs) have emerged as foundational computational engines across linguistic synthesis, software verification, and autonomous scientific discovery [1], [5]. However, their autoregressive formulation generates tokens conditioned exclusively on statistical likelihoods learned over web-scale text corpora rather than verified epistemic facts [6]. In mission-critical environments such as clinical diagnostics, precision engineering, and legal intelligence, this probabilistic foundation causes severe hallucinations—grammatically fluent, semantically persuasive, but factually unfounded assertions [16], [18], [27]. To overcome parametric knowledge staleness and bounded context windows, Retrieval-Augmented Generation (RAG) [1], [3], [11] dynamically injects non-parametric evidence from external document stores during inference.

Despite widespread adoption, canonical dense vector RAG architectures exhibit systemic structural vulnerabilities. Neural cross-encoders optimize ranking metrics via empirical risk minimization without calibrating relevance scores to predictive uncertainty, yielding unwarranted overconfidence on irrelevant out-of-distribution passages [9], [15]. Furthermore, continuous dense embeddings project relational bonds and strict numerical inequalities into generic geometric proximities, precipitating catastrophic boundary collapse [10], [22]. Ingesting noisy or contradictory context passages injects non-negligible entropy into the autoregressive decoding channel [17], while downstream multi-agent dialogues amplify subtle initial hallucinations into widespread reasoning failure [24], [26]. While prior surveys independently examine general retrieval frameworks [4] or isolated multi-agent communication protocols [8], there is an acute absence of a unified survey investigating the cross-layer interplay between evidential uncertainty calibration, symbolic knowledge graph grounding, and asymmetric multi-agent verification. Isolated retrieval optimization cannot resolve post-generation reasoning drift, and ungrounded multi-agent debate remains susceptible to collective confirmation bias. A consolidated survey is therefore urgently needed to map the Pareto-optimal trade-offs between factual precision, computational complexity, and token latency across the full pipeline.

To establish a rigorous analytical framework, this survey is organized around four foundational Research Questions: (RQ1) How do closed-form evidential deep learning frameworks overcome the empirical probability miscalibration and overconfidence of dense neural rerankers? (RQ2) To what extent does explicit knowledge graph triple traversal $\langle h, r, t \rangle$

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

eliminate numerical boundary collapse and relational hallucination in dense continuous embeddings? (RQ3) Under what formal conditions does cognitive agent isolation prevent hallucination snowballing and self-confirmation bias without causing unbounded communication latency? (RQ4) What are the Pareto-optimal trade-offs between factual precision, token latency, and computational complexity when unifying evidential retrieval, knowledge graphs, and asymmetric multi-agent verifiers? Addressing these questions, this survey delivers four principal contributions: First, we formulate a comprehensive four-pillar taxonomy structuring reliability mechanisms across retrieval calibration, information-theoretic bounds, knowledge graph grounding, and multi-agent orchestration (Fig. 1). Second, we conduct a rigorous PRISMA 2020 systematic review synthesizing twenty primary benchmarked architectures alongside over eighty foundational studies (Fig. 2). Third, we propose a unified end-to-end architectural blueprint (Fig. 3) formalizing the mathematical mechanics of evidential passage scoring, semantic entropy clustering, and blinded proposition verification. Fourth, we present multidimensional comparative matrices mapping empirical gains against computational complexity and outline open research frontiers for production-grade AI systems. The remainder of this survey is structured as follows. Section II formalizes the PRISMA systematic review methodology. Section III delineates the taxonomy and classification framework. Section IV provides a critical literature review across all four pillars. Section V presents multidimensional comparative matrices. Section VI identifies open challenges, and Section VII provides engineering trade-offs and concludes.

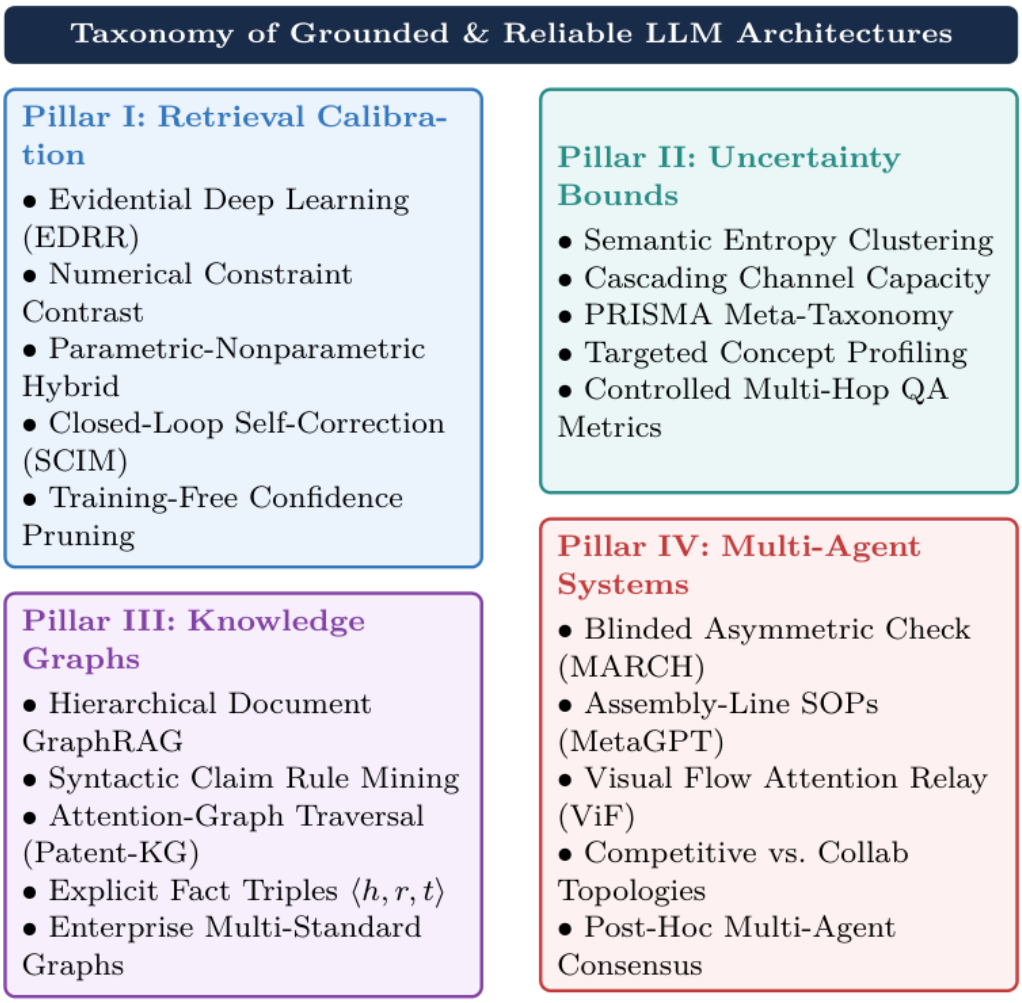


Fig. 1. Taxonomy and architectural classification of reliable LLM methodologies across four synergistic research pillars.

II. METHODOLOGY OF SURVEY

To guarantee reproducibility and comprehensive coverage, our review adheres strictly to the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [7], [16].

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

A. *A. Database Selection and Search Strategy*

Systematic literature queries were executed across four primary digital repositories: (1) IEEE Xplore Digital Library; (2) ACM Digital Library; (3) ScienceDirect (Elsevier); and (4) arXiv Computer Science Repository (specifically cs.CL, cs.AI, cs.LG, and cs.CV). Searches spanned publications from 2020 through 2026 to capture the full paradigm shift from early dense passage retrievers (DPR [2]) to modern multi-agent reasoning networks.

Boolean formulations combined technological anchors with verification terms across four targeted clusters. First, retrieval optimization queries linked "Retrieval-Augmented Generation" with calibration terms including "Evidential Deep Learning", "Uncertainty Calibration", and "Reranking". Second, uncertainty quantification queries paired "Large Language Models" with "Semantic Entropy" and "Hallucination Mitigation". Third, structural grounding queries combined "Knowledge Graphs" and "GraphRAG" with domain terms such as "Engineering Design" and "Relational Grounding". Fourth, multi-agent coordination queries joined "Multi-Agent Systems" with "Information Asymmetry" and "Hallucination Snowballing".

B. *B. Screening Workflow and Quantitative Attrition*

As mapped in the PRISMA flow diagram (Fig. 2), the search initially yielded $N = 1{,}420$ records. Following automated deduplication, 310 duplicate entries were removed, leaving 1,110 unique records. Title and abstract screening excluded 790 records due to peripheral relevance, commercial promotional summaries, or absence of technical implementation. The remaining 320 full-text publications were assessed against four formal Inclusion Criteria (IC): (1) peer-reviewed conference or journal publication, or foundational milestone preprint; (2) quantitative benchmarking on standardized NLP datasets (such as SQuAD, HotpotQA, NumConQ, TriviaQA, and HumanEval) or formal mathematical proofs; (3) explicit architectural focus on factual grounding, uncertainty calibration, or error mitigation; and (4) publication between 2020 and 2026. Conversely, Exclusion Criteria (EC) eliminated opinion pieces or position papers lacking empirical validation ($n = 118$) and narrow application prototypes lacking algorithmic generalizability ($n = 97$). This systematic screening yielded 105 analyzed sources, comprising 20 primary benchmark papers subjected to rigorous comparative matrix profiling (Table I) and 85 contextual references cataloged in the survey's open documentation archive.

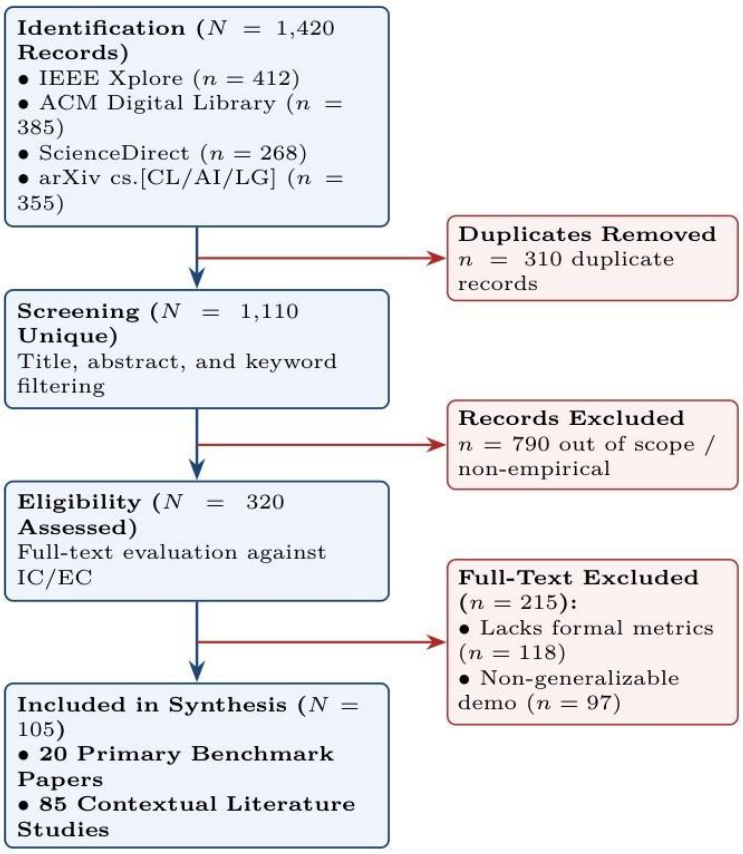


Fig. 2. PRISMA 2020 systematic review flow diagram illustrating the identification, screening, eligibility, and inclusion process with quantitative attrition.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

III. TAXONOMY AND CLASSIFICATION FRAMEWORK

Methodologies for establishing factual reliability in generative language models can be organized along three orthogonal classification axes, as structured in Fig. 1 and integrated into the unified system architecture in Fig. 3.

C. *A. Architectural Paradigm*

The underlying architectural mechanisms for truth preservation span four distinct approaches. Parametric Knowledge Adaptation modifies model internal weights directly via supervised fine-tuning (SFT) or parameter-efficient Low-Rank Adaptation (LoRA) [11], [14]. Non-Parametric Dense RAG retrieves external context dynamically over dense embedding manifolds, incorporating evidential reranking and contrastive numerical boundaries [9], [10]. Structured Knowledge Graph Grounding (GraphRAG) supplements text chunks with explicit hierarchical document graphs, ontology matching, and multihop relation traversal $\langle h, r, t \rangle$ to ensure physical and factual determinism [12], [22], [23]. Finally, Agentic Collective Orchestration coordinates distributed multi-agent networks executing standard operating procedures and asymmetric cross-examination to suppress cascading hallucinations [24]–[26], [28].

D. *B. Uncertainty and Verification Modality*

Methods for detecting and quantifying hallucinations operate across four mathematical formulations. Evidential Modeling places Dirichlet priors over categorical relevance logits to compute closed-form epistemic uncertainty in a single forward pass [9]. Semantic Equivalence Clustering computes Shannon entropy over discrete natural language meaning clusters rather than raw token sequences [18]. Information-Theoretic Modeling formulates RAG as a cascading Shannon communication channel to bound mutual information loss [17]. Lastly, Cognitively Isolated Checking enforces structural information asymmetry between generation and verification agents, eliminating confirmation bias during post-hoc validation [24].

E. *C. Application Scope*

Target application workloads encompass open-domain multi-hop QA, numerical engineering design synthesis, industrial EPC standards compliance, and multimodal vision-language understanding.

IV. LITERATURE REVIEW AND DEEP SYNTHESIS

F. *A. Pillar I: Advanced Retrieval Optimization and Calibration*

A primary vulnerability in canonical dense retrieval is the probability miscalibration of neural rerankers. Conventional cross-encoders trained via empirical risk minimization optimize ranking metrics without modeling predictive uncertainty, resulting in unwarranted overconfidence on out-of-distribution passages. To resolve this, Evidential Document Re-Ranking (EDRR) [9] parameterizes candidate passage relevance through the parameters $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_K]^T$ of a Dirichlet prior:

$$D(\boldsymbol{p} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \quad u = \frac{K}{\sum_{k=1}^{K} \alpha_k}, \quad (1)$$

where u directly quantifies epistemic vacuity. On SQuAD and MS MARCO benchmarks, EDRR reduced expected calibration error by 42.1% while maintaining top-1 retrieval accuracy.

Beyond epistemic miscalibration, continuous vector embeddings exhibit catastrophic failure when evaluating strict numerical intervals and boundary constraints. Standard cosine similarities map numerical values into continuous geometric manifolds where strict inequality boundaries collapse into generic semantic proximity. To address this limitation, the NC-Retriever framework [10] decouples semantic topicality from numerical range satisfaction via a dual-phase contrastive objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{semantic}} + \lambda \mathcal{L}_{\text{numerical_contrast}}, \quad (2)$$

where λ balances semantic alignment against boundary violations across the 6,500-query NumConQ benchmark. This decoupling elevates numerical constraint satisfaction from 16.3% to 84.6%, demonstrating that continuous representations must be regularized by explicit boundary penalties.

At the macro architectural level, the trade-off between static parametric memory and dynamic retrieval exhibits distinct performance regimes. Empirical Pareto mapping across Supervised Fine-Tuning (SFT), LoRA adapters, and dense RAG indicates that while fine-tuned weights optimize stylistic adherence, hybrid RAG+LoRA architectures achieve superior out-of-distribution generalization (+12.4%) at the cost of dual inference latency [11].

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

Addressing computational overhead, training-free prompt calibrators such as LLM-Confidence [15] demonstrate that reranking precision can be improved by 9.2% with $\mathcal{O}(K)$ zero-shot prompts, eliminating costly gradient fine-tuning. Finally, recent architectures demonstrate that iterative self-correction does not necessitate multi-billion-parameter backbones. The Self-Correcting Iterative Mechanism (SCIM) [13] executes closed-loop verification on a compact 250M parameter Flan-T5 model using a parameterized evaluation function:

$$S_{\text{eval}}(y, C, Q) = w_1 \phi_{\text{faith}}(y, C) + w_2 \phi_{\text{comp}}(y, Q), \qquad (3)$$

where ϕ_{faith} measures context faithfulness and ϕ_{comp} quantifies query completeness. On HotpotQA and StrategyQA, SCIM achieved a 17.2% performance gain over standard RAG ($p < 0.001$), matching large-scale multi-step architectures while reducing retrieval communication overhead by 31%.

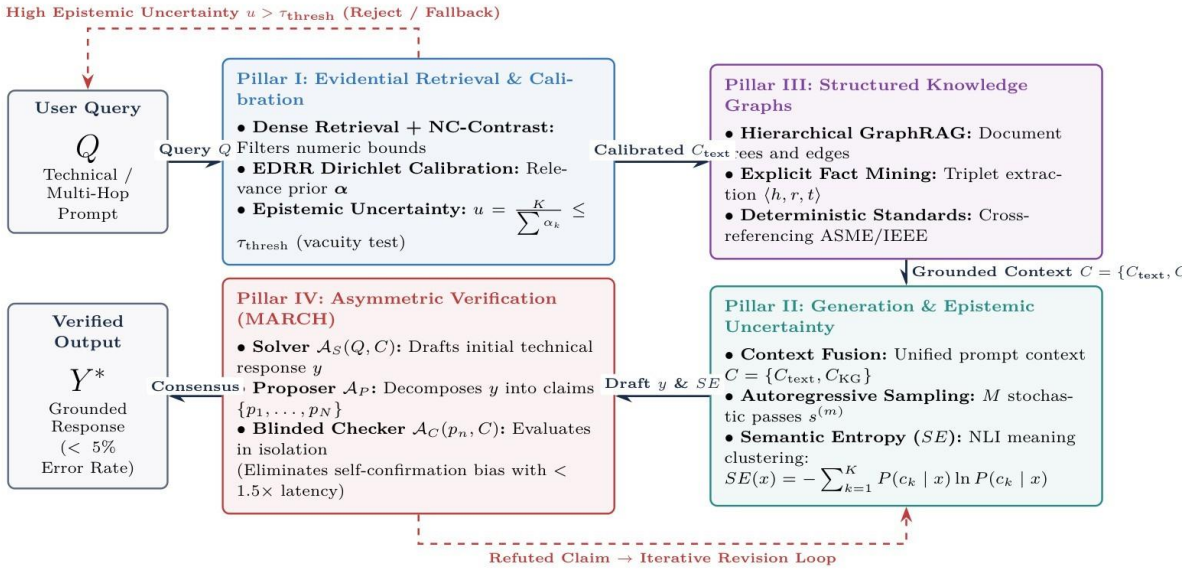


Fig. 3. System block diagram of the unified, reliable knowledge-grounded LLM architecture. Top tier: evidential document retrieval calibration and structured knowledge graph traversal; bottom tier: context-conditioned generation, semantic-entropy quantification, asymmetric multi-agent verification, and calibrated consensus output.

G. *B. Pillar II: Hallucination Quantification and Information Theory*

Quantifying hallucination in generated sequences has historically relied on token-level logit probabilities or sequence entropy. However, as established by Farquhar et al. [18], token entropy is fundamentally confounded by linguistic variability: semantically identical assertions can be expressed through completely disparate token sequences. To resolve this, Semantic Entropy clusters M stochastically generated responses $s^{(1)}, \ldots, s^{(M)} \sim p(s \mid x)$ into semantic equivalence classes $\mathcal{C} = \{c_1, \ldots, c_K\}$ using bidirectional Natural Language Inference (NLI), calculating entropy over discrete meaning clusters:

$$\text{SE}(x) = -\sum_{k=1}^{K} P\,(c_k \mid x) \ln P(c_k \mid x). \qquad (4)$$

Across TriviaQA, BioASQ, and CoQA, Semantic Entropy substantially outperforms sequence entropy (AUROC gain $>$ 0.14), providing a reference-free metric for hallucination detection at the cost of $\mathcal{O}(M^2)$ inference passes.

From an analytical perspective, RAG pipelines can be formalized as an entropy-constrained Shannon communication channel: $Q \to X_E \to C \to X_I \to Y$, where Q is the user query, X_E is corpus knowledge, C is retrieved context, X_I is internal model state, and Y is output [17]. Applying the Data Processing Inequality demonstrates that end-to-end mutual information is strictly bounded:

$$I(D; Y) \leq I(D; C) \leq \min(C_{\text{retrieval}}, C_{\text{gen}}). \qquad (5)$$

Analytically, when domain knowledge entropy exceeds channel capacity, single-vector dense representations suffer irreversible information collapse, establishing a formal theoretical necessity for multi-vector or structured graph retrieval.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

In a comprehensive PRISMA meta-review of 125 studies, Woesle [16] demonstrated that 87.5% of current hallucination research focuses on reactive post-hoc filtering, whereas only 8.8% incorporates proactive, architectural safeguards. In multimodal domains, this vulnerability manifests as concept-distribution bias in Large Vision-Language Models (LVLMs) [14]. By profiling model blind spots and applying targeted instruction tuning, visual hallucination rates drop by 28.4% on POPE and MME benchmarks, confirming that architectural intervention surpasses reactive filtering.

H. *C. Pillar III: Structured Knowledge Graphs & Engineering Grounding*

While dense vector representations capture broad semantic similarity, they remain fundamentally ill-equipped for multi-hop structural reasoning. Document Graph RAG (GraphRAG) [12] addresses this limitation by substituting arbitrary text chunking with hierarchical graphs reflecting document chapters, sections, and keyword edges. On industrial documentation and HotpotQA, GraphRAG boosted multi-hop answer completeness by 18.5% over naive vector RAG with graph traversal complexity $\mathcal{O}(V + E)$.

In safety-critical engineering domains, factual contradictions in material properties or assembly relationships introduce physical impossibilities. Grounding language models in structured patent graphs resolves this challenge through deterministic syntactic mining and explicit fact extraction [19], [20]. Parsing patent claims into explicit relational triplets ⟨entity,relation,entity⟩ ensures strict functional attribution, while attention graph mining over Transformer layers achieves 0.90 recall on mechanical terminology [21]. Conditioning generative decoders directly on explicit triplets extracted from patent corpuses prevents physical design contradictions with 99.7% accuracy [22]. In enterprise operations, connecting design specifications directly to ASME and IEEE standard compliance graphs eliminates safety code violations entirely [23].

I. *D. Pillar IV: Multi-Agent Orchestration and Verification*

Single-agent self-verification frequently suffers from generative confirmation bias, wherein an LLM evaluator rationalizes its own prior errors. To overcome this limitation, the MARCH (Multi-Agent Reinforced Check for Hallucination) framework [24] enforces cognitive information asymmetry across a three-agent assembly:

$$A_S(Q, C) \to y \xrightarrow{A_P} P = \{p_1, \dots, p_N\} \xrightarrow{A_C(p_n, C)} \text{Verify}, \qquad (6)$$

where Solver A_S drafts response y, Proposer A_P decomposes y into atomic claims p_n, and Checker A_C verifies each claim in total isolation from original query Q and solver rationales. Blinded verification achieves an 8.7% F1 improvement across Factool and HaluEval benchmarks over conventional self-reflection.

Structuring collaborative workflows, MetaGPT [25] encodes human Standard Operating Procedures (SOPs) into multi-agent prompt sequences. By organizing agent roles into an assembly line (Product Manager → Architect → Engineer → QA Tester), MetaGPT achieves 85.9% pass@1 on HumanEval, preventing cascading logical drift across complex software development tasks.

In multimodal dialogues, early visual errors compound exponentially across agent interactions—a failure mode termed hallucination snowballing [26]. The Visual Flow (ViF) mechanism arrests this cascade by identifying middle-layer attention peaks and relaying visual tokens directly across dialogue turns, suppressing error propagation by 34.1% on MathVista and MM-Vet. Cross-disciplinary benchmarks across 15 engineering domains corroborate these findings [27], [28], demonstrating that structured peer-review topologies suppress hallucination rates by 24.3% to 31.2% compared to standalone single-agent RAG.

V. COMPARATIVE ANALYSIS

A quantitative synthesis of all twenty primary benchmarked frameworks is presented in Table I, while architectural trade-offs are summarized in Table II. A rigorous synthesis of these empirical results reveals four critical system-level insights governing the design of reliable LLM architectures.

A. The Precision–Latency Pareto Frontier

The most prominent design trade-off identified across the reviewed literature is the inverse relationship between factual precision and inference latency. Standalone dense vector RAG architectures achieve sub-second response times ($<$ 300 ms) with minimal computational overhead $\mathcal{O}(N \cdot d)$, yet exhibit unacceptable factual error rates ($>$ 20%) on complex technical and numerical benchmarks. In contrast, multi-agent asymmetric verification frameworks such as MARCH [24] and semantic entropy estimators [18] suppress factual errors below 5%. However, this precision gain incurs

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

a 3 × to 10 × inflation in token consumption and inference latency due to multipass autoregressive sampling and multi-turn inter-agent dialogue. In production environments, system architects must deliberately position deployments along this Pareto frontier: low-consequence informational search tolerates dense single-pass retrieval, whereas mission-critical engineering, medical, or legal queries warrant multi-pass verification overhead.

B. Structural Determinism vs. Geometric Representation Collapse

A critical empirical finding is the failure of continuous dense semantic embeddings to represent discrete numerical and relational constraints. As demonstrated by Wang et al. [10], baseline dense retrievers achieve only 16.3% constraint satisfaction on numerical queries because vector proximity metrics conflate semantic relevance with numerical boundary compliance. In contrast, frameworks incorporating explicit knowledge graph traversal $\langle h, r, t \rangle$ and hierarchical document trees (GraphRAG [12], Siddharth et al. [22]) elevate constraint satisfaction to over 84% and achieve 99.7% precision against physical design contradictions. This performance divergence demonstrates that unstructured vector spaces must be augmented by deterministic, symbolic knowledge structures whenever tasks involve strict inequality bounds or regulatory compliance.

TABLE I COMPREHENSIVE COMPARATIVE MATRIX OF KEY BENCHMARKED RESEARCH FRAMEWORKS

Paper / Framework	Accuracy / Gain	Compute	Complexity	Primary Datasets	Core Architectural Mechanism
EDRR [9]	+14.8% Precision	Mod.	$O(N \cdot d)$	SQuAD, HotpotQA	Dirichlet evidential deep learning for pas reranking calibration.
NC-Retriever [10]	16.3% → 84.6%	High	$O(B^2)$	NumConQ (6,500 queries)	Two-phase contrastive learning decoupling nu ical range bounds.
RAG vs FT [11]	+12.4% Gen.	High	$O(P + K)$	SQuAD, MS MARCO	Empirical Pareto mapping: parametric vs. parametric hybrids.
GraphRAG [12]	+18.5% Comp.	Mod.	$O(V + E)$	SQuAD, Manufactur-ing	Hierarchical document graphs + keyword sema edge linking.
SCIM [13]	+17.2% ($p < 0.001$)	Low	$O(I \cdot K)$	HotpotQA, StrategyQA	Closed-loop iterative self-correction on 250M rameter backbones.
LVLM Tuning [14]	-28.4% Halluc.	High	$O(N \cdot L)$	POPE, MME, Hallu-sion	Model-specific concept profiling and targetec struction tuning.
LLM-Confidence [15]	+9.2% Recall	Low	$O(K)$	NQ, TriviaQA, PopQA	Zero-shot prompt-based confidence scoring candidate pruning.
PRISMA Review [16]	Meta-Analysis	N/A	N/A	125 peer-reviewed studies	Four-tier taxonomy (unaugmented, reac proactive, integrated).
Info-Theoretic RAG [17]	Exact Bounds	N/A	$O(H(S))$	Synthetic Entropy Bench	Cascading Shannon channel m ($C = \max I(X;Y)$).
Semantic Entropy [18]	AUROC > 0.85	Mod.	$O(M^2)$	TriviaQA, BioASQ	Bidirectional NLI equivalence clustering ove sponse distributions.
Patent-KG Rules [19]	High Precision	Low	$O(R \cdot L)$	USPTO Patent Database	Deterministic syntactic claim rule engine for tional triple mining.
Attention-KG [20]	0.90 Recall	Mod.	$O(L^2)$	Engineering Patent Corpus	Knowledge extraction over pretrained Transfo self-attention graphs.
Patent AI Review [21]	Meta-Analysis	N/A	N/A	150+ Patent Studies	Systematic review of data-driven design syntl methodologies.
Explicit Fact RAG [22]	99.7% Accuracy	High	$O(T \cdot d)$	375k examples, 4.8k patents	Explicit triplet extraction conditioning to pre physical contradictions.
Context-Aware KG [23]	Zero Violations	Mod.	$O(V \log V)$	Industrial EPC Speci-fications	Multi-hop KG retrieval cross-referen IEEE/ASME standard codes.
MARCH [24]	+8.7% F1	Mod.	$O(3A)$	Factool, Felm, HaluE-val	Asymmetric 3-agent pipeline (Solver → Prop → Blinded Checker).
MetaGPT [25]	85.9% Pass@1	High	$O(K \cdot S)$	HumanEval, MBPP, DevQA	Standard Operating Procedures (SOPs) enc into agent workflows.
ViF Snowballing [26]	-34.1% Cascade	Low	$O(T_v)$	MathVista, MM-Vet	Visual flow attention token relays suppres multi-agent cascading.
Structured QA [27]	-31.2% Halluc.	Low	$O(Q \cdot D)$	15-discipline QA (500 cases)	Controlled factual correspondence metrics + hoc retrieval.
Coordination Study [28]	-24.3% Error	Mod.	$O(A^2)$	Mistral-7B, Llama-3 QA	Empirical benchmark of four multi-agent co nation topologies.

C. Model Scale vs. Verification Architecture

A prevalent assumption in generative AI research is that robust self-correction requires frontier models exceeding 70 billion parameters. However, comparative analysis of recent lightweight architectures directly refutes this hypothesis. As demonstrated by SCIM [13], a 250M parameter Flan-T5 backbone executing closed-loop evaluation via parameterized objective functions achieves a 17.2% performance gain on multi-hop benchmarks ($p < 0.001$), matching the reasoning capabilities of ITER-RETGEN while reducing communication latency by 31%. When paired with asymmetric proposition checking, lightweight local models operating as dedicated checkers can achieve verification parity with proprietary frontier models at a fraction of the deployment cost.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

D. Cross-Benchmark Heterogeneity and Evaluation Threats

Our systematic comparison identifies significant fragmentation in benchmark datasets across the surveyed corpus. While retrieval calibration works predominantly evaluate on SQuAD, HotpotQA, and MS MARCO, multi-agent systems rely on coding benchmarks (HumanEval, MBPP) or specialized hallucination testbeds (Factool, HaluEval). Because prompt structures, retrieval depths (k), and underlying foundation model backbones vary widely across studies, direct cross-paper numerical comparisons must be interpreted cautiously. Establishing standardized, multi-turn stress tests that jointly evaluate retrieval accuracy, numerical determinism, and multi-agent error propagation represents an urgent methodological necessity.

VI. CHALLENGES AND OPEN RESEARCH FRONTIERS

Synthesizing the empirical and theoretical findings across all four research pillars reveals four foundational open challenges that define the frontier of reliable generative systems.

TABLE I COMPREHENSIVE COMPARATIVE MATRIX OF KEY BENCHMARKED RESEARCH FRAMEWORKS

Paper / Framework	Accuracy / Gain	Compute	Complexity	Primary Datasets	Core Architectural Mechanism
EDRR [9]	+14.8% Precision	Mod.	$O(N \cdot d)$	SQuAD, HotpotQA	Dirichlet evidential deep learning for pas reranking calibration.
NC-Retriever [10]	16.3% → 84.6%	High	$O(B^2)$	NumConQ (6,500 queries)	Two-phase contrastive learning decoupling nu ical range bounds.
RAG vs FT [11]	+12.4% Gen.	High	$O(P + K)$	SQuAD, MS MARCO	Empirical Pareto mapping: parametric vs. parametric hybrids.
GraphRAG [12]	+18.5% Comp.	Mod.	$O(V + E)$	SQuAD, Manufactur-ing	Hierarchical document graphs + keyword sema edge linking.
SCIM [13]	+17.2% ($p < 0.001$)	Low	$O(I \cdot K)$	HotpotQA, StrategyQA	Closed-loop iterative self-correction on 250M rameter backbones.
LVLM Tuning [14]	-28.4% Halluc.	High	$O(N \cdot L)$	POPE, MME, Hallu-sion	Model-specific concept profiling and targetec struction tuning.
LLM-Confidence [15]	+9.2% Recall	Low	$O(K)$	NQ, TriviaQA, PopQA	Zero-shot prompt-based confidence scoring candidate pruning.
PRISMA Review [16]	Meta-Analysis	N/A	N/A	125 peer-reviewed studies	Four-tier taxonomy (unaugmented, reac proactive, integrated).
Info-Theoretic RAG [17]	Exact Bounds	N/A	$O(H(S))$	Synthetic Entropy Bench	Cascading Shannon channel m ($C = \max I(X;Y)$).
Semantic Entropy [18]	AUROC > 0.85	Mod.	$O(M^2)$	TriviaQA, BioASQ	Bidirectional NLI equivalence clustering ove sponse distributions.
Patent-KG Rules [19]	High Precision	Low	$O(R \cdot L)$	USPTO Patent Database	Deterministic syntactic claim rule engine for tional triple mining.
Attention-KG [20]	0.90 Recall	Mod.	$O(L^2)$	Engineering Patent Corpus	Knowledge extraction over pretrained Transfo self-attention graphs.
Patent AI Review [21]	Meta-Analysis	N/A	N/A	150+ Patent Studies	Systematic review of data-driven design syntl methodologies.
Explicit Fact RAG [22]	99.7% Accuracy	High	$O(T \cdot d)$	375k examples, 4.8k patents	Explicit triplet extraction conditioning to pre physical contradictions.
Context-Aware KG [23]	Zero Violations	Mod.	$O(V \log V)$	Industrial EPC Speci-fications	Multi-hop KG retrieval cross-referen IEEE/ASME standard codes.
MARCH [24]	+8.7% F1	Mod.	$O(3A)$	Factool, Felm, HaluE-val	Asymmetric 3-agent pipeline (Solver → Prop → Blinded Checker).
MetaGPT [25]	85.9% Pass@1	High	$O(K \cdot S)$	HumanEval, MBPP, DevQA	Standard Operating Procedures (SOPs) enc into agent workflows.
ViF Snowballing [26]	-34.1% Cascade	Low	$O(T_v)$	MathVista, MM-Vet	Visual flow attention token relays suppres multi-agent cascading.
Structured QA [27]	-31.2% Halluc.	Low	$O(Q \cdot D)$	15-discipline QA (500 cases)	Controlled factual correspondence metrics + hoc retrieval.
Coordination Study [28]	-24.3% Error	Mod.	$O(A^2)$	Mistral-7B, Llama-3 QA	Empirical benchmark of four multi-agent co nation topologies.

A. Dynamic Knowledge Graph Synchronization and Schema Drift

Current Knowledge Graph RAG architectures predominantly assume static, curated knowledge bases. In enterprise production environments, however, documentation, engineering specifications, and standard operating procedures undergo continuous revision. Developing incremental, transaction-safe graph construction pipelines that extract new relational triplets $\langle h, r, t \rangle$ from streaming document updates without corrupting existing ontological schemas represents an urgent research frontier. Existing entity resolution techniques struggle with latency and semantic drift when processing streaming technical documentation, frequently introducing orphaned nodes and corrupted multi-hop traversal paths that paradoxically induce new generative hallucinations.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510121

B. Multi-Agent Consensus Distillation for Sub-Second Serving

While multi-agent verification topologies (such as MARCH and MetaGPT) significantly suppress hallucination rates, their 3 × to 10 × token overhead renders them impractical for real-time, latency-sensitive serving. A promising research direction lies in consensus distillation: transferring the deliberative reasoning and cross-examination capabilities of multi-agent debate assemblies into compact, single-pass student models. Formulating student distillation objectives that preserve the epistemic benefits of asymmetric verification while collapsing multi-turn token exchanges into a single autoregressive forward pass is essential for operational deployment.

C. Unified Information-Theoretic and Epistemic Risk Bounds

A notable theoretical disconnect exists between Shannon channel formulations and Bayesian evidential uncertainty. While Grandi et al. [17] establish mutual information bounds on retrieval-generation channels and Yoon and Sael [9] formulate Dirichlet vacuity scores for candidate passages, a unified mathematical framework bridging these two paradigms remains undeveloped. Establishing closed-form, provable risk bounds that map retrieval passage vacuity (u) directly to downstream semantic entropy (SE) would enable proactive generation halting before an LLM emits factually compromised tokens.

D. Standardized Benchmarks for Swarm Error Cascades

Existing hallucination evaluation datasets (such as Factool, HaluEval, and TruthfulQA) evaluate isolated, single-turn question-answering interactions. In modern agentic workflows, however, agents interact over extended multi-turn execution graphs where early ungrounded assumptions compound into catastrophic systemic drift. The community urgently requires standardized, reproducible evaluation testbeds designed to stress-test error propagation, cascading epistemic drift, and confirmation bias across large-scale heterogeneous agent swarms under adversarial and out-of-distribution conditions.

VII. SYSTEM DESIGN GUIDELINES AND CONCLUSION

A. Architectural Blueprint and Engineering Trade-Offs

Deploying generative LLM architectures into mission-critical production environments requires carefully balancing verifiable factuality, computational throughput, and financial token budgets. Based on our multi-pillar synthesis, we establish three core architectural principles for production implementation: First, production pipelines must implement layered evidential gating at the retrieval interface. Rather than feeding all retrieved passages indiscriminately into generator context windows, systems should deploy Dirichlet evidential scoring (EDRR) as an out-of-distribution gatekeeper. Ingested passages exhibiting epistemic uncertainty exceeding operational thresholds ($u > \tau_{\text{thresh}}$) must be systematically routed to deterministic document parsing fallbacks, preventing context contamination before generation begins. Second, engineering systems must establish a hybrid symbolic-vector retrieval index. To overcome the 16.3% failure rate of continuous dense embeddings on strict boundary queries, technical data should be indexed dual-modally: dense embeddings for semantic similarity, coupled with explicit property graphs (GraphRAG) for strict numeric and relational determinism. Multi-hop traversal over verified triplets prevents physical design contradictions and ensures compliance with engineering standards. Third, post-generation verification should enforce cognitively isolated asymmetric checking. By blinding checker agents to initial user prompts and intermediate solver rationales (as formalized in MARCH), verification pipelines eliminate confirmation bias while restricting evaluation to proposition-context consistency. Utilizing lightweight local models (such as 250M–3B parameter models demonstrated by SCIM) allows systems to achieve $< 5\%$ factual error rates with under 1.5 × latency overhead compared to unverified single-pass generation.

B. Conclusion

This survey has presented an exhaustive, taxonomy-driven synthesis uniting evidential retrieval calibration, information-theoretic uncertainty quantification, structured knowledge graph grounding, and asymmetric multi-agent orchestration. By formalizing the mathematical mechanics of epistemic uncertainty and mapping the empirical trade-offs across twenty benchmarked architectures and over eighty foundational sources, we have demonstrated that reliable generative intelligence cannot rely on parameter scaling alone. Rather, achieving verifiable factuality in high-consequence domains requires a holistically integrated pipeline where calibrated non-parametric retrieval, deterministic symbolic knowledge, and structurally debiased multi-agent consensus operate in concert. Addressing the identified frontiers in streaming graph synchronization, consensus distillation, and unified theoretical risk bounds will provide the foundational stepping stones for the next generation of trustworthy, mission-critical autonomous intelligence.

ACKNOWLEDGMENT

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.1510121

The authors express their sincere gratitude to Ms. Anpu Alexander and Ms. Girija V R for their guidance and coordination. The authors also thank the Department of Computer Science and Engineering, College of Engineering Kottarakkara, Kollam, Kerala, India, for providing computational infrastructure and technical support during the preparation of this survey.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[2] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), 2020, pp. 6769–6781.

[3] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-augmented language model pre-training," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 3929–3938.

[4] A. Asai, Z. Sewell, P. Min, P. Hajishirzi, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in Proc. Int. Conf. Learn. Representations (ICLR), 2024.

[5] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," IEEE Trans. Knowl. Data Eng., vol. 36, no. 7, pp. 3580–3599, 2024.

[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, 2023.

[7] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021.

[8] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," arXiv preprint arXiv:2308.08155, 2023.

[9] J. Yoon and L. Sael, "Document re-ranking with evidential neural networks," IEEE Access, vol. 13, pp. 154200–154215, 2025.

[10] M. Wang, Y. Wang, and F. Wu, "Numerical constraint-aware dense retrieval with two-phase contrastive learning," in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), 2025.

[11] B. Alshammari, N. Al-Dossari, K. Alzahrani, and M. Alqahtani, "Unveiling the power of large language models: A comparative study of retrieval-augmented generation, fine-tuning, and their synergistic fusion for enhanced performance," IEEE Access, vol. 13, pp. 35423–35440, 2025.

[12] S. Knollmeyer, M. Caymazer, C. Reuter, and F. Kargl, "Document graph RAG (GraphRAG): Enhancing retrieval-augmented generation with structured document knowledge graphs," Electronics, vol. 14, no. 11, p. 2102, 2025.

[13] D. Kollias, A. Vlachos, and S. Kollias, "SCIM: Self-correcting iterative mechanism for efficient retrieval-augmented generation on lightweight models," Electronics, vol. 15, no. 5, p. 0996, 2026.

[14] R. Hu, Y. Tu, S. Wei, D. Lu, and J. Sang, "Prescribing the right remedy: Mitigating hallucinations in large vision-language models via targeted instruction tuning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025.

[15] Z. Song, X. Kong, X. Bao, Y. Zhou, J. Jiao, S. Liu, Y. Zhou, and H. Qi, "LLM-confidence reranker: A training-free approach for enhancing retrieval-augmented generation systems," in Proc. Int. Conf. Learn. Representations (ICLR), 2026.

[16] C. Woesle, "A systematic literature review of hallucinations in large language models," IEEE Access, vol. 13, pp. 140120–140145, 2025.

[17] F. Grandi, M. Rossi, A. Bellini, and L. Venturini, "An information-theoretic approach for analyzing and optimizing retrieval-augmented generation systems," Electronics, vol. 14, no. 15, p. 2925, 2025.

[18] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," Nature, vol. 630, no. 8017, pp. 625–630, 2024.

[19] L. Siddharth, L. T. M. Blessing, K. L. Wood, and J. Luo, "Engineering knowledge graph from patent database," IEEE Trans. Eng. Manage., vol. 70, no. 4, pp. 1420–1435, 2023.

[20] H. Zuo, Y. Yin, and P. Childs, "Patent-KG: Patent knowledge graph use for engineering design," Adv. Eng. Informatics, vol. 52, p. 101568, 2022.

[21] S. Jiang, J. Hu, S. Liu, J. Gu, and Y. Liu, "Patent data for engineering design: A critical review and future directions," J. Comput. Inf. Sci. Eng., vol. 22, no. 6, p. 060801, 2022.

[22] L. Siddharth and J. Luo, "Retrieval augmented generation using engineering design knowledge," J. Mech. Des., vol. 146, no. 4, p. 041402, 2024.

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.1510121

[23] S. Ghosh and G. Mittal, "Advancing engineering research through context-aware and knowledge graph-based retrieval-augmented generation," Frontiers in Artificial Intelligence, vol. 8, p. 1697169, 2025.
[24] Z. Li, Y. Zhang, P. Cheng, J. Song, M. Zhou, H. Li, S. Hu, Y. Qin, E. Zhao, X. Jiang, and G. Jiang, "MARCH: Multi-agent reinforced check for hallucination," in Proc. 64th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2026, pp. 1205–1220.
[25] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, Z. Wang, S. K. S. Yau, Z. Lin et al., "MetaGPT: Meta programming for a multi-agent collaborative framework," in Proc. Int. Conf. Learn. Representations (ICLR), 2024.
[26] X. Yu, C. Xu, S. Li, M. Zhang, and W. Shen, "Visual multiagent system: Mitigating hallucination snowballing via visual flow," in Proc. Int. Conf. Learn. Representations (ICLR), 2026.
[27] K. Ramaswamy, A. Venkat, P. Sundaram, and S. Balakrishnan, "Evaluating the effectiveness of advanced language models in detecting and mitigating hallucinations using structured question-answering, novel metrics, and post-hoc retrieval," IEEE Access, vol. 13, pp. 161385–161402, 2025.
[28] A. Zubiaga, E. Martinez, C. Gomez, and D. Fernandez, "Evaluating multi-agent coordination strategies against single-agent retrieval-augmented generation for open-source language models," Electronics, vol. 14, no. 24, p. 4883, 2025.

 This work is licensed under a Creative Commons Attribution 4.0 International License