

A Review of Hallucination Mitigation Techniques in Retrieval-Augmented Generation Systems

¹Emmanuel Eniola Ajala, ²Harouna Wendpanga Yann Christian Sankara, ³Serif Oyindamola Oyesiji, ⁴Stanley Nwakamma.

¹Independent Researcher, Atlanta, Georgia; emmanuelajala22@gmail.com

²Independent Researcher, Dallas, TX, USA; christiansankara@gmail.com

³Harrisburg University of Science and Technology, Harrisburg, PA; oyesijiserif@gmail.com

⁴Independent Researcher, USA; nwakammastanley@gmail.com

Corresponding Author; emmanuelajala22@gmail.com

Abstract: Retrieval-augmented generation has become the default architecture for grounding large language models in external knowledge, and it is widely deployed precisely because it promises to reduce hallucination, the generation of fluent content unsupported by fact or source. The promise is real but partial: retrieval-augmented systems continue to hallucinate, and they do so through failure modes that differ from those of closed-book models, including retrieval of irrelevant or conflicting evidence, faithful generation from unfaithful sources, unfaithful generation from faithful sources, and fabricated attribution. This review synthesizes the literature on hallucination and its mitigation in retrieval-augmented generation systems. It first consolidates definitions and origins of hallucination in neural text generation, distinguishing factuality from faithfulness and intrinsic from extrinsic error. It then reviews the architectural foundations of retrieval augmentation, from dense retrieval and fusion-based readers through end-to-end retrieval-augmented models, and organizes documented failure points into a pipeline-aligned taxonomy spanning query formulation, retrieval, context construction, generation, and attribution. Mitigation research is examined at each stage: retrieval-side methods including query rewriting, reranking, and robustness training against irrelevant context; generation-side methods including grounded and contrastive decoding, self-verification, uncertainty estimation, and calibrated abstention; and pipeline-level methods including active and iterative retrieval, self-reflective generation, verification chains, and structured knowledge integration. The evaluation landscape is reviewed across faithfulness metrics, attribution measures, and RAG-specific benchmarks, followed by evidence from high-stakes domain deployments. The review closes with open problems in knowledge conflict resolution, long-context utilization, calibration, and evaluation validity, and with directions for future research.

1. Introduction

Large language models generate text by modeling conditional probability over sequences, and the same statistical machinery that produces their fluency also produces their most consequential defect: content that is plausible, confident, and wrong. The phenomenon, commonly termed hallucination, was documented in neural generation systems well before the current model generation, notably in abstractive summarization where generated summaries routinely asserted facts absent from or contradicted by their source documents (Maynez et al., 2020), and it has been consolidated into taxonomies that span tasks, causes, and metrics (Ji et al., 2023). With the scaling of general-purpose models into widely deployed assistants (Brown et al., 2020; Ouyang et al., 2022), hallucination ceased to be a task-specific curiosity and became the central reliability obstacle of the technology, catalogued at length in surveys dedicated to its forms in large models (Zhang et al., 2023; Huang et al., 2025). The stakes are organizational as much as technical, since enterprises now route consequential decisions through model output and must govern that dependence explicitly (Shah et al., 2025), and in clinical settings the fluency that makes assistance valuable is precisely what makes unflagged error costly (Ilodigwe & Adesemoye, 2025a).

Retrieval-augmented generation arose as the principal architectural response. Rather than relying on knowledge stored implicitly in parameters, a retrieval-augmented system fetches evidence from an external corpus at inference time and conditions generation on it, an approach formalized in end-to-end retrieval-augmented models (Lewis et al., 2020; Guu et al., 2020) and now standard in deployed question answering, enterprise search, and assistant products. The motivations are threefold. Parametric knowledge is bounded by training data and decays with time, whereas an index can be updated continuously; parametric knowledge fails disproportionately on rare, long-tail facts that models see too seldom to memorize (Kandpal et al., 2023); and parametric answers offer no provenance, whereas retrieved evidence supports citation and audit. Empirical work confirmed the core promise early: augmenting conversational generation with retrieval measurably reduced hallucinated content relative to closed-book baselines (Shuster et al., 2021). Comparable motivations drive grounding in operational decision support, where systems are expected to point at the record an answer came from rather than assert a conclusion from an opaque model (Adelanwa et al., 2024).

The promise, however, is partial, and the gap between retrieval-augmented and hallucination-free has become a research field of its own. Retrieval can fail, returning irrelevant, outdated, or contradictory passages that the generator then faithfully reproduces. Generation can fail, ignoring retrieved evidence in favor of parametric priors, blending sources incoherently, or asserting

more than the evidence supports. Attribution can fail, decorating unsupported claims with citations to documents that do not contain them. Systematic engineering studies of deployed pipelines catalogue failure points across the entire chain from ingestion to answer (Barnett et al., 2024), and benchmark studies show that models frequently mishandle exactly the hard cases retrieval is meant to solve: noisy evidence, absent evidence, and evidence that conflicts with what the model already believes (Chen et al., 2024; Longpre et al., 2021). Deployed systems add an adversarial dimension absent from benchmark settings, since the retrieval corpus and the tools a model may call are themselves attack surfaces for injection and content poisoning (Nwakamma et al., 2024b).

This review synthesizes the literature on hallucination and its mitigation specifically in retrieval-augmented settings. Its organizing premise is that mitigation is a pipeline property: hallucination in a retrieval-augmented system is not one phenomenon but a family of failures arising at identifiable stages, and the mitigation literature is best understood as a set of interventions aligned to those stages. The paper proceeds as follows. Section 2 states scope and approach. Section 3 consolidates definitions and origins of hallucination. Section 4 reviews retrieval-augmented architectures. Section 5 organizes failure modes into a pipeline-aligned taxonomy. Sections 6 through 8 review mitigation at the retrieval side, the generation side, and the pipeline level respectively. Section 9 treats evaluation, Section 10 reviews domain deployments, Sections 11 and 12 discuss open problems and future directions, and Section 13 concludes.

2. Scope and Approach of the Review

The review covers systems in which a neural language model conditions its output on evidence retrieved at inference time from an external corpus, whether the coupling is loose, as in prompt-concatenation pipelines over commodity vector stores, or tight, as in models trained jointly with their retrievers. Hallucination is treated in both of its standard senses, factuality error, in which output conflicts with world knowledge, and faithfulness error, in which output conflicts with or exceeds the provided evidence, because retrieval-augmented systems are accountable to both standards and their divergence is itself a characteristic failure setting. Work on hallucination in closed-book generation is included where it supplies definitions, mechanisms, or mitigation techniques that transfer; work on retrieval for purposes other than grounding generation, and on multimodal hallucination, is excluded.

The literature base spans natural language processing venues, where retrieval augmentation, faithfulness evaluation, and decoding-time mitigation are published; machine learning venues, where retrieval-augmented architectures and uncertainty methods appear; information retrieval venues, for the retrieval and reranking components; and an emerging software engineering and systems literature that studies deployed pipelines as engineered artifacts. Comprehensive surveys of retrieval-augmented generation were used to bound the architectural landscape (Gao et al., 2024), and hallucination surveys were used to anchor terminology (Ji et al., 2023; Huang et al., 2025). Emphasis falls on work from 2020 onward, the period in which retrieval-augmented generation became the dominant grounding paradigm, with earlier foundational work in open-domain question answering and faithfulness evaluation retained where the field builds on it directly. An adjacent applied literature on model governance, auditing, and regulated deployment was consulted where it bears on how mitigation is specified, verified, and evidenced in production rather than in evaluation (Ominyi & Anichukwueze, 2023; Aliliele et al., 2025a).

Two synthesis choices shape the presentation. First, mitigation methods are organized by the pipeline stage at which they intervene rather than by technique family, because deployed systems compose interventions across stages and the stage view exposes which combinations are complementary and which redundant. Second, throughout the review, claims about mitigation effectiveness are tied to the evaluation instruments that support them, because the field's metrics are themselves model-based and imperfect, and several apparent disagreements in the literature trace to differences in what was measured rather than in what works.

3. Hallucination in Language Models: Definitions and Origins

Terminological precision matters because mitigation targets differ by error type. The generation literature distinguishes intrinsic hallucination, output that contradicts the provided source, from extrinsic hallucination, output that is unsupported by the source but not necessarily false, a distinction developed in summarization faithfulness research (Maynez et al., 2020) and generalized across generation tasks in the survey literature (Ji et al., 2023). Orthogonally, factuality names agreement with the world, while faithfulness names agreement with given evidence. A retrieval-augmented answer can be faithful but false, when the evidence is wrong; false but supported by the model's training distribution, when parametric priors override evidence; or true but unfaithful, when the model answers correctly from parameters while citing evidence that does not support the claim. Surveys of hallucination in large models refine this space further, separating factuality errors from faithfulness errors and instruction-inconsistency, and cataloguing contributing causes across data, training, and inference (Zhang et al., 2023; Huang et al., 2025). Applied treatments of accountability for automated decision systems draw a parallel distinction between an output being wrong and an output being unexplainable, and a retrieval-augmented system is answerable on both counts (Ominyi & Anichukwueze, 2023; Annan, 2024).

The mechanistic origins relevant to retrieval-augmented settings fall into three groups. The first is knowledge representation in parameters. Language models demonstrably store and retrieve relational knowledge (Petroni et al., 2019) and can answer many questions closed-book (Roberts et al., 2020), but this parametric knowledge is frequency-sensitive, failing systematically on long-tail entities and relations (Kandpal et al., 2023; Mallen et al., 2023), and time-bound, degrading as the world moves past the training

corpus (Lazaridou et al., 2021; Dhingra et al., 2022). Retrieval exists to compensate for both properties, and hallucination risk concentrates where compensation is incomplete: rare entities retrieval also struggles to cover, and temporal questions whose answers differ between index and parameters. Because parametric coverage is a product of what the training corpus contained and how it was curated, data-side intervention on the composition of training and adaptation corpora bears directly on which facts a model can be expected to hold reliably (Oyesiji et al., 2024), and coverage thins sharply outside high-resource languages, where adaptation method rather than scale carries the burden (Oyesiji et al., 2023).

The second group concerns training objectives and behavior. Likelihood training rewards fluent continuation rather than epistemic accuracy, instruction tuning and preference optimization reward helpfulness signals that correlate imperfectly with truth (Ouyang et al., 2022), and models therefore answer confidently by default rather than abstaining. Studies of self-knowledge complicate the picture productively: models carry usable internal signal about the reliability of their own claims, with self-evaluation well calibrated in some regimes (Kadavath et al., 2022), and factual truthfulness partially recoverable from internal representations even when surface output is wrong (Li et al., 2023). This latent signal is the resource on which several generation-side mitigations of Section 7 draw. Work on interpretable models in applied prediction settings argues the same point from the opposite direction, treating recoverable internal signal as the precondition for trusting an output at all (Komi & Adamolekun, 2021).

The third group concerns the interaction between context and parameters, which is the distinctive locus of retrieval-augmented failure. When retrieved evidence conflicts with parametric belief, models exhibit systematic and non-obvious preferences, sometimes over-trusting parameters against contradicting evidence and sometimes over-trusting persuasive but wrong context, with behavior modulated by evidence coherence and popularity of the fact (Longpre et al., 2021; Xie et al., 2024). Even absent conflict, context use is imperfect: models are measurably distracted by irrelevant material inserted into prompts (Shi et al., 2023), and their attention over long contexts is positionally biased, using information at the beginning and end of the window far better than material in the middle (Liu et al., 2024). These properties convert retrieval imperfection into generation error, and they explain why mitigation cannot live at the retrieval stage alone.

4. Retrieval-Augmented Generation: Foundations and Architectures

The model side of the paradigm rests on the transformer architecture and its pre-training lineage, from bidirectional encoders through unified text-to-text formulations to openly released large models (Vaswani et al., 2017; Devlin et al., 2019; Raffel et al., 2020; Touvron et al., 2023). The architecture of retrieval augmentation assembled itself from open-domain question answering. The modern template, retrieve then read, was established by pipelines coupling term-based retrieval over Wikipedia with neural readers (Chen et al., 2017), with classical probabilistic ranking functions supplying the retrieval component (Robertson & Zaragoza, 2009). Dense retrieval replaced term matching with learned embeddings, training dual encoders so that questions and supporting passages map to nearby vectors (Karpukhin et al., 2020), with hard-negative mining sharpening the learned space (Xiong et al., 2021), and late-interaction designs recovering token-level matching precision at scale (Khattab & Zaharia, 2020). Neural reranking layered cross-encoder precision over first-stage recall (Nogueira & Cho, 2019). The vector search infrastructure underneath, GPU-scale similarity indexes and graph-based approximate nearest neighbor structures, made corpus sizes of billions of passages practical (Johnson et al., 2021; Malkov & Yashunin, 2020). Standard resources for the resulting task family include Natural Questions and TriviaQA for open-domain question answering (Kwiatkowski et al., 2019; Joshi et al., 2017), multi-hop composition benchmarks (Yang et al., 2018), evidence-based verification datasets (Thorne et al., 2018), the KILT suite unifying knowledge-intensive tasks over a shared snapshot (Petroni et al., 2021), and heterogeneous zero-shot retrieval evaluation (Thakur et al., 2021). The operational substrate beneath these components is ordinary data engineering, since corpora must be extracted, transformed, and loaded before they can be embedded or indexed (Badmus et al., 2019a), the governance of that substrate determines what the index is permitted to contain (Aliliele et al., 2025b), and the elastic serving components beneath it inherit the capacity-planning problems of any served system (Ahmed et al., 2020).

On this substrate, three coupling patterns between retrieval and generation emerged. End-to-end trained models integrate the retriever into the generative likelihood, as in retrieval-augmented generation over dense indexes (Lewis et al., 2020) and retrieval-augmented pre-training in which the retriever is learned from the language modeling signal itself (Guu et al., 2020); fusion-in-decoder architectures scale evidence integration by encoding many passages independently and fusing them in the decoder (Izacard & Grave, 2021), and retrieval-in-pre-training designs interleave retrieval throughout the model's computation over trillions of tokens (Borgeaud et al., 2022), with few-shot variants demonstrating strong knowledge-intensive performance from compact models (Izacard et al., 2023). Nearest-neighbor language models interpolate parametric token distributions with distributions induced by retrieved contexts (Khandelwal et al., 2020). Finally, and dominant in deployment, in-context augmentation treats the language model as frozen, concatenating retrieved passages into the prompt (Ram et al., 2023), optionally adapting only the retriever to the black-box model's preferences (Shi et al., 2023b). Instruction-following models trained to browse and cite demonstrated the pattern's product form, answers with quoted supporting sources (Nakano et al., 2021). Deployment constraints increasingly determine which coupling is available in practice: quantization, pruning, and distillation make capable models viable on constrained hardware (Nwakamma et al., 2025), ranking components face the same compression pressure (Oyesiji et al., 2025), and portable runtimes have made cloud-to-edge placement of the whole pipeline a live architectural choice (Ojukwu et al., 2025; Komi & Ganiu, 2023).

The deployed pipeline that this review's taxonomy addresses is the in-context pattern elaborated with engineering stages: document ingestion and chunking, embedding and indexing, query processing, first-stage retrieval, reranking, context assembly under a token budget, generation with citation instructions, and post-hoc verification. Surveys of the paradigm organize its variants into naive, advanced, and modular configurations and catalogue the optimization surface at each stage (Gao et al., 2024). Two properties of this pipeline matter for everything that follows. It is a cascade, so errors propagate and compound forward, and stages are trained or configured in isolation, so each stage's failure distribution is unknown to the others. The taxonomy of the next section is, in essence, an accounting of where that cascade leaks.

5. A Pipeline-Aligned Taxonomy of Retrieval-Augmented Hallucination

Failures enter at query formulation. User questions are underspecified, multi-intent, or phrased in vocabulary mismatched to the corpus, and compositional questions require evidence no single query surfaces, a gap quantified as the compositionality gap between what models can answer piecewise and what they answer directly (Press et al., 2023). A pipeline that retrieves on the raw question therefore begins with degraded recall, and the generator downstream answers anyway.

Failures enter at retrieval. First-stage retrieval returns passages that are topically adjacent but evidentially useless, stale relative to the question's time reference, or drawn from unreliable strata of the corpus; engineering studies of deployed systems identify missing content and missed top-ranked documents among their most common failure points (Barnett et al., 2024). Benchmarks constructed to probe these conditions show sharp degradation: when noise is injected into retrieved sets, when no passage contains the answer, and when passages contradict one another, models frequently fail to reject, abstain, or integrate correctly (Chen et al., 2024). The conflict case is the most treacherous, because model behavior under evidence-parameter disagreement is systematic but not reliable, varying with fact popularity and evidence framing (Longpre et al., 2021; Xie et al., 2024). A distinct class of retrieval failure is adversarial rather than accidental, since a corpus can be seeded with content engineered to be retrieved and believed, which makes index hygiene a security property and not only a quality one (Nwakamma et al., 2024b; Adebayo et al., 2022).

Failures enter at context construction. Chunking severs claims from their qualifying context, concatenation interleaves sources without provenance boundaries, and position effects mean that evidence placed mid-context is systematically underused (Liu et al., 2024), while irrelevant filler measurably distracts generation (Shi et al., 2023). The result is a context window that contains the answer yet does not deliver it. Provenance is lost at this stage as much as anywhere, and enterprise practice for classifying material by sensitivity and tracing it to its system of record supplies exactly the metadata that grounded generation and citation later require (Aliliele et al., 2024a; Aneke et al., 2023).

Failures enter at generation. The generator may ignore evidence in favor of parametric priors, producing answers ungrounded in the supplied passages; may over-abstract, asserting stronger claims than the evidence licenses, the extrinsic pattern documented since summarization (Maynez et al., 2020); may blend incompatible sources into a fluent composite true of neither; or may complete gaps with invented specifics, names, numbers, dates, that inherit the surface style of the retrieved material. Because retrieval-augmented prompts lengthen contexts, generation failures interact with the long-context weaknesses above, and because instruction-tuned models are rewarded for answering, the default failure is confident fabrication rather than refusal.

Failures enter at attribution. Citation-generation studies show that models cite documents that do not support the attributed claim, omit citations for claims that need them, and cite partially, covering some conjuncts of a compound claim while leaving others unsupported (Gao et al., 2023). Attribution failure is the most corrosive category for trust, because it counterfeits exactly the verifiability that motivates retrieval augmentation, and formal frameworks for attribution were developed precisely to make such claims testable (Rashkin et al., 2023). Where output reaches populations who cannot readily verify it, unearned confidence compounds, which is the concern that trust-centered designs for public-facing generated communication are built around (Odejide, 2024).

The taxonomy earns its keep through diagnosis: the visible symptom, a wrong or unsupported answer, is compatible with any of these entry points, and mitigation misallocated across stages is a recurring engineering failure. The next three sections review the mitigation literature stage by stage.

6. Retrieval-Side Mitigation

Query-stage methods attack recall before generation can inherit its absence. Query rewriting trains or prompts models to reformulate user questions into retrieval-effective forms, and trainable rewriters optimized against downstream answer quality improve end-to-end performance over retrieving on raw questions (Ma et al., 2023). Decompositional prompting addresses compositionality directly, eliciting sub-questions whose individual answers assemble into the final one and narrowing the measured compositionality gap (Press et al., 2023), while interleaved formulations alternate reasoning steps with retrieval so that each step's queries are conditioned on what reasoning has established so far (Trivedi et al., 2023). Agentic formulations generalize the pattern, treating retrieval as an action within a reasoning loop that can search, read, and reconsider (Yao et al., 2023). Query understanding

over noisy, informal, and multilingual sources is itself a mature applied problem, and the normalization and reformulation strategies developed there transfer directly to retrieval-augmented pipelines operating on comparable corpora (Atakpa et al., 2023).

Ranking-stage methods raise the evidential quality of what enters the context. Cross-encoder reranking remains the workhorse for precision at the top of the ranking (Nogueira & Cho, 2019), and retriever adaptation methods tune the retrieval distribution against the frozen generator's actual utility, treating the language model's likelihood of the correct answer as supervision for what to retrieve (Shi et al., 2023b). Heterogeneous benchmark evaluation cautions against overfitting any single retrieval configuration, documenting large cross-domain variance in dense retriever effectiveness (Thakur et al., 2021), which for deployed systems translates into a recommendation of hybrid first stages and mandatory in-domain evaluation. In enterprise settings the candidate pool is additionally constrained by entitlement, so retrieval must respect identity and access boundaries before relevance is considered at all, and a permission-aware first stage is a correctness requirement rather than a deployment detail (Mbonu et al., 2020; Badmus et al., 2024).

Robustness methods accept that retrieval will be imperfect and harden generation against its characteristic imperfections. Training with deliberately mixed relevant and irrelevant evidence teaches models to use the former and ignore the latter, substantially improving robustness to retrieval noise relative to models trained only on clean evidence (Yoran et al., 2024), and the distraction results that motivate such training also support a cheap inference-time counterpart, instructing models explicitly that supplied context may be irrelevant, which recovers part of the degradation (Shi et al., 2023). Context construction practice contributes the remaining retrieval-side gains: provenance-preserving formatting that keeps source boundaries and metadata legible to the generator, deduplication and conflict surfacing so that disagreement among sources is presented as disagreement rather than silently concatenated, and evidence ordering informed by positional bias so that the strongest passages occupy the positions models actually use (Liu et al., 2024). Systematic failure-point studies close the loop operationally, recommending instrumentation at each stage so that missing-content and ranking failures are observed rather than inferred from bad answers (Barnett et al., 2024). Where source systems are heterogeneous, much of the realized gain comes from integration and data-quality discipline upstream of retrieval rather than from the retriever itself (Ilodigwe & Adesemoye, 2021).

7. Generation-Side Mitigation

Decoding-time methods modify how the model produces tokens without retraining it. Contrastive layer decoding exploits the observation that factual knowledge localizes differentially across transformer layers, amplifying the signal of mature layers against premature ones and improving factuality on knowledge-intensive benchmarks (Chuang et al., 2024). Representation-level intervention goes one step deeper, identifying attention heads whose activations track truthfulness and shifting activations along truth-correlated directions at inference time (Li et al., 2023). Factuality-oriented training complements decoding: continued training regimes designed around factual completion improve the factual reliability of open-ended generation (Lee et al., 2022), and truthfulness benchmarks that penalize imitation of common misconceptions supply the evaluation pressure such methods answer to (Lin et al., 2022). Explainability work in applied deployment settings converges on the same premise, that usable reliability depends on exposing the basis of an output and not only its conclusion (Sanni et al., 2026).

Self-assessment methods convert the model's latent self-knowledge into usable reliability signals. Direct self-evaluation, asking the model for the probability that its answer is true, is informative and partially calibrated in large models (Kadavath et al., 2022), though the general calibration literature cautions that modern networks are systematically overconfident and require post-hoc correction (Guo et al., 2017). Sampling-based consistency provides a black-box alternative: hallucinated content proves unstable under resampling, so agreement across stochastic generations serves as a factuality signal without model internals (Manakul et al., 2023). Semantic formulations sharpen this idea by clustering samples by meaning before measuring dispersion, yielding semantic entropy, an uncertainty estimate that detects confabulated answers substantially better than token-level measures and has been validated across question answering domains (Kuhn et al., 2023; Farquhar et al., 2024). In retrieval-augmented pipelines these signals gate behavior: high-uncertainty answers can be suppressed, hedged, routed to further retrieval, or escalated to human review, converting hallucination from an undetected output into a managed event. Treating predictive uncertainty as an explicit modeling target rather than a residual is standard practice in applied forecasting, and the framing carries over directly (Wedraogo & Sanni, 2024).

Verification and revision methods have the model check and repair its own output. Chain-of-verification prompts the generator to draft an answer, plan verification questions against it, answer them independently, and revise, reducing hallucination across tasks without external tools (Dhuliawala et al., 2024). Self-refinement generalizes draft-feedback-revise into a task-agnostic loop (Madaan et al., 2023), and tool-augmented variants ground the critique step in external evidence, having the model verify its claims against search results or other tools and correct accordingly (Gou et al., 2024), with pipeline instantiations wrapping black-box models in retrieval-verify-repair loops driven by automated feedback (Peng et al., 2023). Reasoning-style prompting interacts with all of the above: eliciting intermediate reasoning improves multi-step accuracy (Wei et al., 2022), but reasoning traces are themselves generated text subject to fabrication, which is why the verification methods deliberately re-derive claims rather than trusting the trace. Where residual risk is unacceptable, the loop is closed by a person rather than a model, and the human-in-the-loop

literature supplies the design vocabulary for where that intervention sits, what it costs, and how generated drafts should be presented for review rather than for acceptance (Ladapo et al., 2022; Ladapo et al., 2024).

Abstention closes the generation-side repertoire. Adaptive retrieval policies decide when evidence is needed at all, retrieving for long-tail questions while answering popular ones parametrically, using popularity and confidence signals to route (Mallen et al., 2023), and the same signals support declining to answer when neither parameters nor evidence suffice. Calibrated abstention is arguably the definitive mitigation, since a system that reliably knows when it does not know converts residual hallucination risk into explicit uncertainty, and its dependence on the calibration and uncertainty methods above makes those methods infrastructural rather than optional.

8. Pipeline-Level and Structural Mitigation

Active and iterative retrieval restructures the pipeline from retrieve-once to retrieve-as-needed. Forward-looking active retrieval monitors generation confidence token by token, triggering fresh retrieval when the model is about to generate low-confidence content and regenerating the affected span with the new evidence (Jiang et al., 2023). Interleaved reasoning-retrieval alternates the two operations so that multi-hop questions accumulate their evidence chain stepwise (Trivedi et al., 2023), and agentic loops make retrieval one action among several in a deliberative policy (Yao et al., 2023). These designs attack the taxonomy's query and retrieval failures jointly, because each retrieval is conditioned on an explicit, evolving information need rather than on the original question alone. Multi-agent formulations extend the same logic to division of labor among specialized components acting under partial information, whether those components are reasoning roles or coordinating physical agents (Akanbi et al., 2025).

Self-reflective architectures train the generator to manage its own grounding. Self-reflective retrieval-augmented generation trains models to emit control tokens deciding when to retrieve, whether each retrieved passage is relevant, whether the generated segment is supported by it, and whether the answer is useful, yielding a single model that retrieves on demand and critiques its own evidence use, with reported gains in factuality and citation accuracy over conventional pipelines (Asai et al., 2024). The design internalizes at training time the checks that verification chains perform at inference time, and the two approaches are complementary in deployment, structural self-critique for routine grounding and explicit verification for high-stakes claims. Comparable division-of-labor designs appear in regulated document review, where separate agents draft, check, and adjudicate rather than a single model performing all three roles (Ominyi et al., 2025a).

Structured knowledge integration substitutes or supplements free-text corpora with knowledge graphs, whose triples carry provenance and compositional structure that text chunks lack. Roadmaps for unifying language models with knowledge graphs organize the design space, graphs as retrieval targets, as constraints on generation, and as verification references against which generated claims are checked (Pan et al., 2024). For hallucination specifically, the graph's contribution is decidability: a claim mapped to a triple is checkable in a way that a claim floating in prose is not, at the cost of coverage and of the mapping step itself. Where the requirement is tamper-evidence rather than compositional structure, append-only and cryptographically verifiable records serve the same auditability goal for the evidence trail itself (Anichukwueze et al., 2021; Oshoba et al., 2020; Sanni et al., 2024), and where corpora cannot be centralized across organizational boundaries, federated designs preserve the evidence relationship without pooling the underlying data (Sanni et al., 2023).

Finally, systems-level practice treats the pipeline as an engineered artifact with observable stages. The failure-point literature recommends per-stage instrumentation, retrieval hit rates, reranker agreement, citation coverage, verification pass rates, so that hallucination regressions localize to a stage rather than surfacing only as end-task complaints (Barnett et al., 2024), and automated evaluation frameworks purpose-built for retrieval-augmented systems supply the measurement layer such practice requires (Es et al., 2024; Saad-Falcon et al., 2024). The consistent lesson across pipeline-level work is compositionality: no single intervention removes hallucination, but staged combinations, robust retrieval, grounded generation, uncertainty gating, and verification, reduce it multiplicatively because their failure modes are imperfectly correlated. Process mining over pipeline execution logs offers a complementary view, reconstructing what a system actually did on a given request rather than what its configuration says it should do (Sanni et al., 2025), while dashboarding practice determines whether those signals are acted on or merely collected (Ogbole et al., 2021), and governance frameworks for AI-assisted workflows assign ownership for the residual risk that instrumentation exposes (Badmus et al., 2025).

9. Evaluation: Metrics, Benchmarks, and Their Limits

Faithfulness measurement matured in summarization before migrating to retrieval-augmented generation. Entailment-style classifiers score whether source documents support generated claims (Kryscinski et al., 2020), with natural language inference reformulations improving robustness through sentence-level decomposition (Laban et al., 2022); question-based metrics test whether questions generated from the output are answered identically by the source (Durmus et al., 2020; Wang et al., 2020); and meta-evaluations of summarization metrics document both progress and the persistent gap between automatic scores and human judgments (Fabbri et al., 2021), with standardized comparison surfaces consolidating faithfulness evaluation across tasks (Honovich et al.,

2022). Fine-grained factuality scoring decomposes long-form output into atomic claims and verifies each against a reference corpus, yielding precision-style factuality scores that expose error rates invisible to answer-level metrics (Min et al., 2023).

Attribution has its own instruments. Formal frameworks define when output is attributable to identified sources and how annotators should judge it (Rashkin et al., 2023), and citation-quality benchmarks operationalize the definition for generated text with citations, measuring citation recall, whether supported claims carry citations, and citation precision, whether cited passages actually support their claims, alongside fluency and correctness (Gao et al., 2023). RAG-specific evaluation frameworks assemble these dimensions into pipeline dashboards: reference-light suites score faithfulness, answer relevance, and context relevance jointly (Es et al., 2024), and trained-judge frameworks calibrate lightweight evaluators on synthetic data with statistical correction against small human-labeled sets (Saad-Falcon et al., 2024). Behavioral benchmarks complete the toolkit by probing the failure conditions of Section 5 directly, noise robustness, negative rejection when evidence is absent, information integration, and conflict handling (Chen et al., 2024), while truthfulness suites measure resistance to imitative falsehood (Lin et al., 2022) and knowledge-intensive task collections standardize provenance-aware comparison (Petroni et al., 2021). Judge-model evaluation has been surveyed in its own right, with attention to the biases judges import, the prompt sensitivity of their verdicts, and the release-gating decisions that come to rest on them (Nwakamma et al., 2024a).

Three validity cautions qualify this apparatus. Most faithfulness and attribution metrics are themselves model judgments, inheriting the biases and blind spots of judge models, which is why frameworks emphasize calibration against human labels (Saad-Falcon et al., 2024) and why metric-only gains warrant skepticism. Benchmarks age: as evaluation sets circulate, contamination and optimization pressure erode their discriminative value, and static snapshots drift from the live corpora deployed systems retrieve over (Petroni et al., 2021). And aggregate scores obscure the asymmetric costs of error types, a system that hallucinates rarely but confidently in the negative-rejection condition may dominate benchmarks while being operationally worse than a more abstinent competitor. Reporting practice that separates error types, discloses judge models, and includes abstention behavior is the field's available correction. Two adjacent lines sharpen the caution. Annotator and judge disagreement is not noise to be averaged away but signal about the underlying task, and treating it as noise inflates apparent reliability (Ladapo et al., 2026). And continuous validation practice from security engineering, in which controls are re-tested against live conditions rather than certified once, is the operational analogue of the living benchmarks the field lacks (Dosunmu & Ogundele, 2024a; Dosunmu & Ogundele, 2025b), with acceptance testing against stated user requirements remaining the last check that measured gains correspond to usable ones (Eyetsemitan et al., 2023b).

10. Domain Deployments and Case Evidence

High-stakes domains supply the clearest evidence of both the necessity and the insufficiency of retrieval augmentation. In medicine, large models encode substantial clinical knowledge and approach expert-level performance on medical question answering under careful evaluation rubrics, while the same evaluations document residual error modes, incorrect retrieval-free recall, incomplete answers, and unsupported reasoning, that motivate grounding and abstention in clinical use (Singhal et al., 2023). The clinical setting sharpens every requirement of this review: evidence must be current, provenance must be auditable, and confident error is costlier than refusal, which elevates uncertainty gating and verification from options to obligations. The surrounding literature on clinical deployment reaches the same conclusion from the operational side, documenting both the gains available from grounded decision support (Adelanwa et al., 2024; Afrihyia et al., 2023) and the governance, readiness, and accountability conditions without which those gains do not materialize (Afrihyia et al., 2024; Afrihyia et al., 2025; Ildigwe & Adesemoye, 2025a). Trust in patient-facing systems is contingent on privacy posture and regulatory alignment as much as on accuracy (Akintola et al., 2025; Ekechi et al., 2026), and retrieval over clinical records inherits every protection obligation attached to those records (Nnaji & Akinlolu, 2024a). Population-level applications add a further requirement, since risk stratification that is confidently wrong misallocates scarce response capacity (Oparah et al., 2021).

In law, systematic measurement found hallucination pervasive when general models answer legal questions, with error rates that vary by task and jurisdiction and that concentrate exactly where practitioners are least able to detect them, in plausible misstatements of authority (Dahl et al., 2024). Legal deployments consequently exhibit the full mitigation stack of this review, retrieval over curated primary sources, citation-first generation, and attribution checking, and they illustrate the taxonomy's attribution stratum vividly, since a fabricated citation to a real reporter is more dangerous than an obvious fabrication. The regulatory literature around these deployments is now substantial: enterprise use of large language models raises privilege and confidentiality questions distinct from accuracy (Ominyi et al., 2025c), automated decisions carry anti-discrimination exposure (Annan, 2025a), authorship and ownership of generated text remain unsettled (Annan, 2023), and organizations face concrete readiness obligations under emerging statutory regimes (Ominyi et al., 2024b).

Enterprise assistants and search products constitute the largest deployment class, and the engineering literature drawn from them grounds the taxonomy of Section 5: missing content in corpora, chunking that severs context, ranking misses, and specification failures between stages recur across case studies (Barnett et al., 2024), while product-scale browsing assistants demonstrated early that citation-forward generation is feasible and preferred by users when the plumbing supports it (Nakano et al., 2021). Across domains, deployment evidence converges on three practices: curation of the retrieval corpus dominates downstream cleverness,

because no generation-side method rescues an index that lacks or misstates the answer; citation with verification is the trust interface, because users and auditors act on provenance, not on model confidence; and abstention pathways are load-bearing, because every deployed system faces queries its corpus cannot answer and its behavior there defines its failure reputation. Governance frameworks written for this deployment class make ownership of these practices explicit rather than leaving them to implementation judgment (Ominyi & Anichukwueze, 2024).

Financial services present a third pattern, in which retrieval-augmented systems are deployed less to answer open questions than to assemble evidence for a human adjudicator. Fraud and financial-crime work has long paired statistical detection with case narratives that must be traceable to source records (Atakpa & Abetoh, 2022; Okoruwa et al., 2025), and compliance monitoring at transaction scale imposes the same evidence-linking requirement (Ominyi et al., 2024a). The audit literature supplies the governing constraint: a system whose output enters a controlled process must itself be auditable, which places model risk management and assurance of AI-enabled financial systems on the critical path rather than at the margin (Fobellah, 2025b; Ominyi & Anichukwueze, 2023), while the human factors shaping how analysts weigh machine-generated evidence belong to the risk picture rather than sitting outside it (Fobellah, 2025a). Analytics platforms built for this setting increasingly assume these properties rather than adding them afterward (Walawalkar et al., 2026).

Education and public communication show the same failure modes at their most diffuse. Adaptive and assistive learning systems place generated explanation in front of learners least equipped to detect error, which makes grounding and calibrated hedging pedagogical requirements rather than technical niceties (Onwubuya et al., 2026; Fawehinmi et al., 2026; Joseph & Uzundu, 2024a). Public safety communication is the sharper case, since alerting content is consumed under time pressure, often in translation, by populations with uneven access to verification, and the design literature there treats trust, accessibility, and equity as first-order system requirements rather than downstream concerns (Odejide, 2023; Odejide, 2025). Data governance determines what such systems may retrieve and about whom, and is therefore upstream of every mitigation discussed above (Orise & Niniola, 2025).

11. Challenges and Open Problems

Knowledge conflict remains the least solved failure condition. Systems need principled policies for evidence-parameter disagreement, evidence-evidence disagreement, and staleness, policies that current models approximate with unstable heuristics sensitive to popularity and framing (Longpre et al., 2021; Xie et al., 2024). Progress requires both behavioral specification, what the correct action is when sources disagree, and mechanisms that implement it, and the temporal dimension makes the problem permanent, since indexes and parameters will always date at different rates (Dhingra et al., 2022; Lazaridou et al., 2021).

Long-context utilization is the second structural obstacle. Context windows have grown faster than the reliability of attention over them, and positional underuse of mid-context evidence converts retrieval success into generation failure at exactly the scale deployments now operate (Liu et al., 2024), while distraction by irrelevant material penalizes the recall-oriented retrieval that long windows invite (Shi et al., 2023). Whether the resolution is architectural, training-based, or a retrieval discipline that curates less and orders better is open, and evaluation must catch up, since answer-level metrics do not localize positional failure.

Calibration and abstention at deployment grain constitute the third cluster. Semantic uncertainty and self-assessment work under controlled conditions (Farquhar et al., 2024; Kadavath et al., 2022), but thresholds drift across domains, models, and prompt formats, and the field lacks standards for reporting operating characteristics of abstention, the false-refusal versus hallucination trade-off, that deployments actually tune. Relatedly, verification methods inherit a regress, the verifier is a model with its own error distribution, and characterizing when verification helps versus merely launders confidence remains open (Dhuliawala et al., 2024; Gou et al., 2024). The organizational counterpart is equally unsettled, since who owns an abstention threshold, and against what standard it is defended when it is wrong, are governance questions that current practice answers inconsistently (Ominyi et al., 2026b).

Evaluation validity is the fourth. Judge-model bias, benchmark contamination, and the divergence between static benchmarks and live corpora jointly threaten the field's ability to measure its own progress (Saad-Falcon et al., 2024; Petroni et al., 2021), and attribution evaluation at the granularity users need, claim by claim, citation by citation, remains expensive enough that it is sampled rather than routine (Gao et al., 2023; Min et al., 2023). Finally, the systems dimension is understudied relative to its deployment weight: chunking, index freshness, corpus governance, and stage instrumentation determine much of realized hallucination behavior, yet occupy a small fraction of the literature (Barnett et al., 2024). Fairness adds a dimension that aggregate accuracy conceals, since a system can improve on average while becoming systematically less reliable for some populations than others, a pattern documented wherever predictive systems are evaluated only in aggregate (Komi, 2023).

12. Future Directions

Four directions appear most promising. The first is native grounding: architectures and training regimes in which evidence use is a learned, supervised behavior rather than a prompt-time convention, extending the self-reflective line in which retrieval decisions, relevance judgments, and support judgments are model outputs with their own training signal (Asai et al., 2024), and integrating retrieval deeply enough that attribution is generated with, not after, the claim (Borgeaud et al., 2022; Izacard et al., 2023).

The second is uncertainty as a first-class output. Semantic uncertainty estimation validated at scale (Farquhar et al., 2024), self-assessment (Kadavath et al., 2022), and consistency signals (Manakul et al., 2023) point toward systems that emit calibrated support levels per claim, enabling downstream policies, hedging, verification, escalation, abstention, to be tuned against explicit operating characteristics rather than fixed prompts.

The third is conflict-aware evidence reasoning: models and pipelines that detect disagreement among sources and between sources and parameters, represent it explicitly, and resolve it by recency, authority, and provenance policies specified by the deployer, a direction for which knowledge-graph integration supplies structure (Pan et al., 2024) and behavioral benchmarks supply measurement (Chen et al., 2024; Xie et al., 2024). Resolution policy is ultimately institutional rather than technical, which argues for frameworks that connect legal, quantitative, and operational judgment instead of treating conflict resolution as a modeling problem alone (Ominyi et al., 2026a).

The fourth is evaluation infrastructure worthy of the deployment stakes: living benchmarks coupled to evolving corpora, judge calibration as standard practice (Saad-Falcon et al., 2024), claim-level attribution auditing cheap enough to run continuously (Min et al., 2023; Gao et al., 2023), and reporting norms that disclose abstention behavior and error-type decomposition alongside headline accuracy. The trajectory of neighboring subfields suggests that such infrastructure, more than any single mitigation, converts a fast-moving literature into reliable engineering practice.

13. Conclusion

Retrieval-augmented generation reframed the hallucination problem without dissolving it. Grounding generation in retrieved evidence removes the largest single cause of fabricated content, the closed-book reliance on parametric memory for facts it stores poorly or not at all, and the empirical record confirms substantial reductions in hallucination when retrieval is added. What remains is a different and more tractable problem: a cascade of identifiable failure modes, in query formulation, retrieval, context construction, generation, and attribution, each with a maturing mitigation literature and each observable with purpose-built evaluation.

The synthesis offered here supports three conclusions. First, mitigation is compositional: robust retrieval, grounded and self-critical generation, uncertainty gating, and verification address imperfectly correlated failures, and deployed reliability comes from their staged combination rather than from any single technique. Second, the model's relationship to its evidence, not the evidence's availability, is now the frontier: conflict handling, long-context utilization, and faithful attribution are where current systems fail most consequentially. Third, measurement is the binding constraint on progress: the field's metrics are model judgments requiring calibration, its benchmarks age quickly, and its most important behaviors, abstention and attribution, are the least standardized in reporting.

The practical outlook is nonetheless positive. The combination of native grounding, calibrated uncertainty, conflict-aware reasoning, and continuous evaluation defines an engineering agenda in which hallucination becomes a managed, measured error budget rather than an unpredictable defect, and the pace of the literature reviewed here suggests that agenda is being executed. Systems built on these principles will not be free of error, but they will be honest about it, and for the knowledge-intensive applications that motivated retrieval augmentation in the first place, calibrated honesty is the property on which everything else depends.

References

- Adebayo, A., Adegbite, M. P., & Ahmed, M. O. (2022). Adversarial machine learning in critical infrastructure: A conceptual framework for threat modeling AI enabled OT systems. *World Journal of Innovation and Modern Technology*, 6(1), 184-234.
- Adelanwa, A., Basnet, A., & Anene, U. N. (2024). Advanced AI based decision support systems for healthcare operations and resource planning. *Gyanshauryam, International Scientific Refereed Research Journal*, 7(1), 244-276.
- Afrihiya, E., Akinse, S. G., & Ojukwu, P. U. (2023). Development and implementation of an AI-driven early detection system for non-communicable diseases. *International Journal of Advanced Multidisciplinary Research and Studies*, 3(6), 2680-2691.
- Afrihiya, E., Akinse, S. G., & Ojukwu, P. U. (2024). Comparative governance of AI-driven healthcare management: Executive oversight, regulatory structures, and accountability in the United States and developing countries. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(6), 3087-3102.
- Afrihiya, E., Akinse, S. G., & Ojukwu, P. U. (2025). Organizational readiness for generative AI integration in healthcare operations: Comparative management capabilities between the U.S. and low- and middle-income countries. *Iconic Research and Engineering Journals*, 8(10), 1673-1697.
- Ahmed, K. S., Odejebi, O. D., & Oshoba, T. O. (2020). Predictive model for cloud resource scaling using machine learning techniques. *Journal of Frontiers in Multidisciplinary Research*, 1(1), 173-183.
- Akanbi, O., Ganiu, O. S., & Sunday, E. A. (2025). A multi-agent AI framework for swarm intelligence in autonomous mobile robot (AMR) fleet coordination. *International Journal of Scientific Research in Humanities and Social Sciences*, 2(1), 81-109.
- Akintola, A. S., Fawehinmi, Y. O., Aiyenitaju, O., Chinemerem, B., & Emmanuel, O. (2025). Digital transformation in telehealth: A systematic review of user trust, privacy, and regulatory governance in AI-powered remote monitoring systems. *Journal of Scientific Research and Reports*, 31(11), 163-182.

- Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2024a). A conceptual framework for enterprise data sensitivity classification and regulatory traceability mechanisms. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(6), 3103-3124.
- Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025a). A review of AI assisted continuous auditing systems in technology risk and cybersecurity oversight. *Gyanshauryam, International Scientific Refereed Research Journal*, 8(4), 210-250.
- Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025b). Advances in data lakehouse governance architectures for enterprise data loss prevention and compliance assurance. *Shodhshauryam, International Scientific Refereed Research Journal*, 8(4), 171-213.
- Aneke, O. B., Sydney, E., & Adesemoye, A. C. (2023). Data integrity and governance in regulated life sciences operations: A review of failures, controls, and remediation strategies. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(10), 538-568.
- Anichukwueze, C. C., Osuji, V. C., & Oguntegbe, E. E. (2021). Blockchain-based architectures for tamper-proof regulatory recordkeeping and real-time audit readiness. *International Journal of Multidisciplinary Research and Growth Evaluation*, 2(6), 485-504.
- Annan, A. O. (2023). Intellectual property ownership and authorship in AI-generated works. *Gyanshauryam, International Scientific Refereed Research Journal*, 6(3), 425-450.
- Annan, A. O. (2024). Algorithmic accountability and trade secret protection in artificial intelligence. *Shodhshauryam, International Scientific Refereed Research Journal*, 7(5), 315-347.
- Annan, A. O. (2025a). Automated decision-making and anti-discrimination compliance under U.S. law. *International Journal of Advanced Multidisciplinary Research and Studies*, 5(2), 2522-2540.
- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of the International Conference on Learning Representations*.
- Atakpa, M. I., & Abetoh, N. F. (2022). A systematic review of machine learning advances in financial fraud detection for banking systems. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 8(2), 771-801.
- Atakpa, M. I., Abolaji, T. O., & Akeju, B. (2023). A review of natural language processing for social media-based public health surveillance and analytics. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(6), 1030-1060.
- Badmus, O., Dosunmu, A. A., & Anunagba, C. O. (2025). A governance framework for AI-assisted CRM workflows: Agentforce deployment, risk management, and organizational readiness in enterprise Salesforce environments. *International Journal of Scientific Research in Humanities and Social Sciences*, 2(1), 59-80.
- Badmus, O., Dosunmu, A. A., & Ozowara, D. E. (2019a). A conceptual model for ETL design and data integration in Salesforce-centric enterprise architectures. *Iconic Research and Engineering Journals*, 3(5).
- Badmus, O., Jooda, D., & Anunagba, C. O. (2024). A privacy-by-design framework for role-based security architecture in Salesforce environments handling sensitive personal data. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(6), 3244-3254.
- Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. *Proceedings of the IEEE/ACM International Conference on AI Engineering*, 194-199.
- Borgeaud, S., Mensch, A., Hoffmann, J., et al. (2022). Improving language models by retrieving from trillions of tokens. *Proceedings of the International Conference on Machine Learning*, 2206-2240.
- Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1870-1879.
- Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 17754-17762.
- Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., & He, P. (2024). DoLa: Decoding by contrasting layers improves factuality in large language models. *Proceedings of the International Conference on Learning Representations*.
- Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64-93.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 4171-4186.
- Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., & Cohen, W. W. (2022). Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10, 257-273.
- Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-verification reduces hallucination in large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, 3563-3578.
- Dosunmu, A. A., & Ogundele, P. O. (2024a). Breach and attack simulation frameworks for continuous validation of enterprise security controls. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 10(3), 1100-1119.
- Dosunmu, A. A., & Ogundele, P. O. (2025b). Cyber defense performance measurement frameworks for executive and board level governance. *Computer Science and IT Research Journal*, 6(11), 895-913.
- Durmus, E., He, H., & Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 5055-5070.

- Ekechi, N. V., Ozowara, D. E., & Anunagba, C. O. (2026). Conceptual framework for AI governance, data privacy compliance, and financial sustainability in digital health. *Computer Science and IT Research Journal*, 7(4), 275-299.
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.
- Eyetsemitan, R. A., Ambali, K. B., Oyeleye, A. O., & Fadayomi, O. (2023b). User acceptance testing in small business technology deployment: A structured validation framework for lean operational environments. *International Journal of Multidisciplinary Research and Growth Evaluation*, 4(6), 1512-1531.
- Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9, 391–409.
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), 625–630.
- Fawehinmi, Y. O., Okereke, G. M. T., Williams, P., Chijioke-Churuba, J. N., & Asare-Badu, B. (2026). AI-human collaboration in education and psychology: Personalised learning, language acquisition and student support systems. *Journal of Political Science and International Relationship*, 3(1), 54-64.
- Fobellah, A. N. (2025a). Cognitive biases in financial decision-making: Implications for audit and risk management in large corporations: A conceptual review. *International Journal of Science and Research Archive*, 16(2), 17-22.
- Fobellah, A. N. (2025b). Navigating digital transformation: Auditing artificial intelligence-powered financial systems: A conceptual review. *International Journal of Science and Research Archive*, 16(2), 23-28.
- Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 6465–6488.
- Gao, Y., Xiong, Y., Gao, X., et al. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: Large language models can self-correct with tool-interactive critiquing. *Proceedings of the International Conference on Learning Representations*.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the International Conference on Machine Learning*, 1321–1330.
- Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). Retrieval augmented language model pre-training. *Proceedings of the International Conference on Machine Learning*, 3929–3938.
- Honovich, O., Aharoni, R., Herzig, J., et al. (2022). TRUE: Re-evaluating factual consistency evaluation. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 3905–3920.
- Huang, L., Yu, W., Ma, W., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 42.
- Ilodigwe, L., & Adesemoye, A. C. (2021). The data backbone of health system transformation: A governance and architecture framework for interoperability, data quality, and advanced analytics at national scale. *International Journal of Health and Pharmaceutical Research*, 6(2), 52-76.
- Ilodigwe, L., & Adesemoye, A. C. (2025a). Advances, risks, and implementation challenges of artificial intelligence as a force multiplier for clinical decision making and health system efficiency. *International Journal of Medical Evaluation and Physical Report*, 9(7), 165-185.
- Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 874–880.
- Izacard, G., Lewis, P., Lomeli, M., et al. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251), 1–43.
- Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 248.
- Jiang, Z., Xu, F. F., Gao, L., et al. (2023). Active retrieval augmented generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 7969–7992.
- Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547.
- Joseph, O. B., & Uzundu, N. C. (2024a). Integrating AI and machine learning in STEM education: Challenges and opportunities. *Computer Science & IT Research Journal*, 5(8), 1732-1750.
- Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1601–1611.
- Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. *Proceedings of the International Conference on Machine Learning*, 15696–15707.
- Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 6769–6781.
- Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2020). Generalization through memorization: Nearest neighbor language models. *Proceedings of the International Conference on Learning Representations*.
- Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48.

- Komi, N. M. (2023). When accurate forecasts are not fair forecasts: Deep learning load prediction and equity across districts. *Shodhshauryam, International Scientific Refereed Research Journal*, 6(6), 476-546.
- Komi, N. M., & Adamolekun, A. (2021). Interpretable machine learning for early failure prediction in distributed renewable energy assets. *International Journal of Multidisciplinary Research and Growth Evaluation*, 2(6), 1015-1038.
- Komi, N. M., & Ganiu, O. S. (2023). Edge intelligence for resilient microgrid control: Advances, energy sovereignty, and open challenges. *Gyanshauryam, International Scientific Refereed Research Journal*, 6(3), 525-586.
- Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 9332–9346.
- Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Proceedings of the International Conference on Learning Representations*.
- Kwiatkowski, T., Palomaki, J., Redfield, O., et al. (2019). Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7, 453–466.
- Laban, P., Schnabel, T., Bennett, P. N., & Hearst, M. A. (2022). SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10, 163–177.
- Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2022). Human-in-the-loop machine learning: A state of the art. *Journal of Frontiers in Multidisciplinary Research*, 3(1), 656-669.
- Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2024). Keeping humans in the loop: Human-centered automated annotation with generative AI. *International Journal of Multidisciplinary Futuristic Development*, 5(1), 81-95.
- Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2026). On the disagreement problem in human-in-the-loop federated machine learning. *International Journal of Multidisciplinary Research and Growth Evaluation*, 7(3), 178-192.
- Lazaridou, A., Kuncoro, A., Gribovskaya, E., et al. (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34, 29348–29363.
- Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P., Shoenybi, M., & Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35, 34586–34599.
- Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 41451–41530.
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 3214–3252.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173.
- Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-based knowledge conflicts in question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 7052–7063.
- Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting for retrieval-augmented large language models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 5303–5315.
- Madaan, A., Tandon, N., Gupta, P., et al. (2023). Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 46534–46594.
- Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4), 824–836.
- Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 9802–9822.
- Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 9004–9017.
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1906–1919.
- Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Uzoka, E. (2020). A review of identity and access management integration strategies in hybrid and multi cloud environments. *International Journal of Multidisciplinary Research and Growth Evaluation*, 1(5), 795-810.
- Min, S., Krishna, K., Lyu, X., et al. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 12076–12100.
- Nakano, R., Hilton, J., Balaji, S., et al. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- Nnaji, N., & Akinlolu, V. S. (2024a). Enhancing clinical record protection through structured health data security frameworks in regulated healthcare environments. *International Journal of Scientific Research in Humanities and Social Sciences*.

- Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv preprint arXiv:1901.04085.
- Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2024a). LLM-as-a-judge for automated model governance: A review of methods, biases, and release-gating practices. *World Journal of Innovation and Modern Technology*, 8(6), 185-216.
- Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2024b). Securing agentic AI enterprise workflows against prompt injection, tool poisoning, memory manipulation, and excessive agency threats. *World Journal of Innovation and Modern Technology*, 8(6), 217-248.
- Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2025). On-device multimodal small language models for privacy-preserving edge intelligence through quantization, pruning, distillation, and energy-efficient inference. *World Journal of Innovation and Modern Technology*, 9(12), 271-302.
- Odejide, G. S. (2023). Artificial intelligence in multilingual and accessible emergency alerting: A scoping review of evidence, applications, and gaps in the United States. *International Journal of Scientific Research in Science, Engineering and Technology*, 10(6), 499-512.
- Odejide, G. S. (2024). A trust-centered conceptual model for AI-driven multilingual emergency communication. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 10(6), 2781-2811.
- Odejide, G. S. (2025). Toward inclusive emergency alerting: A conceptual framework for accessibility and equity in AI-driven public safety systems. *Iconic Research and Engineering Journals*, 9(5).
- Ogbole, J. I., Okoruwa, P. O., Babatope, O. M., & Oyewole, T. (2021). Developing an integrated data visualization model for continuous business performance monitoring and optimization. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 7(5), 441-467.
- Ojukwu, J., Oyesiji, S. O., & Nwakamma, S. (2025). WebAssembly component model and WASI for portable cloud-to-edge software: Language interoperability, security, performance, and container comparisons. *International Journal of Computer Science and Mathematical Theory*, 11(12), 232-263.
- Okoruwa, P. O., Fadayomi, O., Bello, A. D., Elebe, O., Hammed, N. I., & Omoegun, G. O. (2025). An artificial intelligence-driven financial crime investigation framework for analyst decision support. *International Journal of Multidisciplinary Research and Growth Evaluation*, 6(6).
- Ominyi, M., & Anichukwueze, C. C. (2023). A review of algorithmic accountability and model risk management approaches in financial services. *Journal of Accounting and Financial Management*, 9(12), 218-238.
- Ominyi, M., & Anichukwueze, C. C. (2024). Conceptualizing responsible AI governance structures for automated decision systems in regulated industries. *International Journal of Social Sciences and Management Research*, 10(11), 403-430.
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2024a). Developing a conceptual framework for AI-assisted compliance monitoring and anomaly detection in high-volume transaction environments. *World Journal of Innovation and Modern Technology*, 8(6), 204-228.
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2024b). A systematic review of organizational readiness for the EU Artificial Intelligence Act in multinational enterprises. *World Journal of Innovation and Modern Technology*, 8(6), 185-203.
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2025a). A conceptual framework for multi-agent AI quality control in the review of regulated documents. *International Journal of Social Sciences and Management Research*, 11(8), 534-557.
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2025c). Reviewing legal privilege and confidentiality challenges in enterprise deployment of large language models. *Journal of Law and Global Policy*, 10(3), 150-176.
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2026a). A conceptual model for integrating legal, quantitative, and operational disciplines in next-generation regulatory risk leadership. *International Journal of Social Sciences and Management Research*, 12(7), 524-550.
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2026b). Conceptualizing enterprise AI governance functions as institutional standards for auditable and defensible risk management. *Journal of Accounting and Financial Management*, 12(7), 326-356.
- Onwubuya, L. I., Fawehinmi, Y. O., Sunday, D., Igwenagu, M. O., Adeniran, A. O., & Agyapong, E. A. (2026). Adaptive learning systems powered by artificial intelligence for STEM education improvement. *Asian Journal of Education and Social Studies*, 52(6), 553-566.
- Oparah, O. S., Ezeh, F. E., Olatunji, G. I., & Ajayi, O. O. (2021). AI-based risk stratification framework for large-scale public health emergency preparedness and response planning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 7(1), 332-366.
- Orise, C., & Niniola, S. (2025). Data governance in digital health learning systems: A systematic review of implementation frameworks and best practices. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(5), 520-538.
- Oshoba, T. O., Hammed, N. I., & Odejobi, O. D. (2020). Blockchain-enabled compliance and audit trail model for cloud configuration management. *Journal of Frontiers in Multidisciplinary Research*, 1(1), 193-201.
- Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2023). Parameter-efficient multilingual adaptation of on-device language models: A survey. *International Journal of Engineering and Modern Technology*, 9(3), 287-320.
- Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2024). A conceptual framework for context-aware data curation in fine-tuning pipelines. *International Journal of Engineering and Modern Technology*, 10(11), 197-228.
- Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2025). Compressing recommender and ranking models for edge deployment: A survey. *International Journal of Engineering and Modern Technology*, 11(12), 205-237.
- Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), 3580–3599.

- Peng, B., Galley, M., He, P., et al. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:2302.12813.
- Petroni, F., Piktus, A., Fan, A., et al. (2021). KILT: A benchmark for knowledge intensive language tasks. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2523–2544.
- Petroni, F., Rocktäschel, T., Riedel, S., et al. (2019). Language models as knowledge bases? Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2463–2473.
- Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., & Lewis, M. (2023). Measuring and narrowing the compositionality gap in language models. Findings of the Association for Computational Linguistics: EMNLP 2023, 5687–5711.
- Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.
- Ram, O., Levine, Y., Dalmedigos, I., Muhlga, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11, 1316–1331.
- Rashkin, H., Nikolaev, V., Lamm, M., et al. (2023). Measuring attribution in natural language generation models. Computational Linguistics, 49(4), 777–840.
- Roberts, A., Raffel, C., & Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? Proceedings of the Conference on Empirical Methods in Natural Language Processing, 5418–5426.
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389.
- Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 338–354.
- Sanni, J. O., Iwuanyanwu, U. A., & Essien, M. A. (2025). Problem-oriented process mining for auditable marketing automation lifecycle control. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 1933-1947.
- Sanni, J. O., Iwuanyanwu, U. A., & Essien, M. A. (2026). Designing explainable AI based marketing automation architectures for healthcare and financial applications. World Scientific News, 213, 119-148.
- Sanni, J. O., Iwuanyanwu, U. A., Essien, M. A., & Attah, A. (2023). Lifecycle-aware marketing automation using federated learning for secure cross-organizational data management. Gyanshauryam, International Scientific Refereed Research Journal, 6(6), 337-364.
- Sanni, J. O., Iwuanyanwu, U. A., Essien, M. A., & Wedraogo, L. (2024). Integrating blockchain-enabled smart contracts for transparent and verifiable marketing workflows. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2891-2906.
- Shah, R., Oziri, S. T., & Seyi-Lande, O. B. (2025). A framework for leveraging artificial intelligence in strategic business decision-making. Gulf Journal of Advance Business Research, 3(11), 1517-1558.
- Shi, F., Chen, X., Misra, K., et al. (2023). Large language models can be easily distracted by irrelevant context. Proceedings of the International Conference on Machine Learning, 31210–31227.
- Shi, W., Min, S., Yasunaga, M., et al. (2023b). REPLUG: Retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652.
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. Findings of the Association for Computational Linguistics: EMNLP 2021, 3784–3803.
- Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.
- Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. Proceedings of the NeurIPS Datasets and Benchmarks Track.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 809–819.
- Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2023). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 10014–10037.
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
- Walawalkar, G., Oduleye, T. E., Adesuyi, M. O., & Kalu, A. (2026). Next-generation financial analytics frameworks for AI-enabled enterprises. International Journal of Advanced Multidisciplinary Research and Studies, 6(1), 1779-1791.
- Wang, A., Cho, K., & Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 5008–5020.
- Wedraogo, L., & Sanni, J. O. (2024). Machine learning models addressing uncertainty in cross-channel campaign performance forecasting accuracy. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2875-2890.
- 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.
- Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2024). Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. Proceedings of the International Conference on Learning Representations.

- Xiong, L., Xiong, C., Li, Y., et al. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. Proceedings of the International Conference on Learning Representations.
- Yang, Z., Qi, P., Zhang, S., et al. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2369–2380.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. Proceedings of the International Conference on Learning Representations.
- Yoran, O., Wolfson, T., Ram, O., & Berant, J. (2024). Making retrieval-augmented language models robust to irrelevant context. Proceedings of the International Conference on Learning Representations.
- Zhang, Y., Li, Y., Cui, L., et al. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.