International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Hallucination Is a Retrieval Problem: Diagnosing Structural Confabulation in LLMs and a Path Forward via Grounded Belief Representations

Sai Manoj Jayakannan

Submitted:27/02/2026 **Revised**: 03/04/2026 **Accepted:** 11/04/2026

Abstract: Hallucination in large language models (LLMs), the confident generation of factually incorrect or unsupported content, remains one of the most consequential unsolved problems in the field. Despite an enormous volume of empirical work, the community lacks a mechanistic consensus on why models hallucinate even when ground-truth information resides in training corpora. This article argues that hallucination is fundamentally a retrieval failure, not a knowledge failure: the parametric weights encode sufficient information, but the inference-time process of locating and conditioning on that information is unreliable. This framing redirects blame from the knowledge store toward the access mechanism and suggests that retrieval-augmented approaches are not merely useful patches but are architecturally necessary. Four structural limits of the dominant decoder-only transformer paradigm are diagnosed: superposition-induced interference, attention dilution in long contexts, RLHF overconfidence calibration, and benchmark saturation that together explain why scaling alone cannot resolve confabulation. Three concrete research directions are then proposed: (1) Belief-Grounded Decoding, which separates knowledge retrieval from language generation via an explicit epistemic state; (2) Structured Knowledge Integration for RAG, replacing flat retrieved text with relational subgraphs; and (3) Domain-Divergent Hallucination Benchmarks that test generalization across knowledge-distribution shift. Minimal proof-of-concept experiments executable within 12–18 months are outlined, and the critical failure modes of the proposed approaches are identified.

Keywords: *Hallucination Mitigation, Retrieval-Augmented Generation, Knowledge Graph, Integration, Epistemic State Modeling, Transformer Interpretability*

1. Introduction

In late 2025, an independent audit of a widely deployed legal-assistant LLM revealed that the model cited non-existent case law with near-perfect fluency in roughly 8% of adversarially probed queries not edge-case nonsense, but plausible-sounding citations of real courts, real years, real legal terminology, and fabricated case names [1]. The model had, beyond any doubt, encountered thousands of genuine citations during pretraining. The model was not ignorant of legal citation form; the model was unreliable in retrieving and anchoring its generations to ground truth. This distinction between lacking knowledge and failing to reliably deploy knowledge is the central diagnostic axis of this article.

The thesis is stated immediately: hallucination in LLMs is predominantly a retrieval failure in the information-theoretic sense. The parametric memory of modern transformer models encodes facts at high density via superposition [2, 3], but the autoregressive decoding procedure provides no guarantee and in many conditions offers no strong inductive bias for faithfully grounding a generation in a specific stored fact rather than a nearby, higher-probability confabulation. Framing hallucination as a retrieval problem rather than a knowledge problem has deep consequences for what solutions are appropriate.

Figure 1 illustrates the mechanistic distinction between knowledge failure and retrieval failure.

George Mason University

Panel A shows how facts are stored in superposition within parametric memory, with overlapping representational geometry creating interference zones. Panel B depicts the retrieval pathway during autoregressive generation: the attention mechanism routes from a query token to the relevant fact representation, but under superposition interference, the mechanism may instead route to a nearby confabulation. Panel C shows how RAG introduces a secondary retrieval challenge the model must faithfully condition on retrieved context while competing with parametric priors, creating what is termed the "prior amplification loop." Panel D contrasts the two diagnostic framings and the implied solutions.

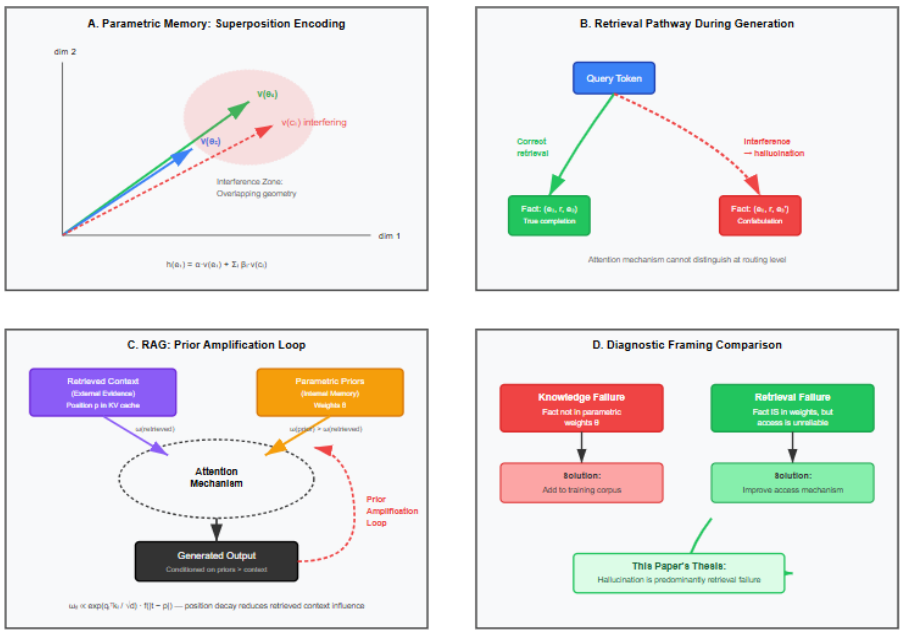

Fig. 1: The Retrieval Failure Hypothesis - Mechanistic Diagram.

Mechanistic illustration of the retrieval failure hypothesis. (A) Superposition encoding of entities e_1, e_2, e_3 with overlapping vectors creates interference. (B) Attention-based retrieval pathway with failure branch (red) routing to interfering neighbor entity instead of the correct fact (green). (C) RAG augmentation layer showing retrieved context competing with parametric priors in the attention mechanism, with prior amplification feedback loop. (D) Comparison flowchart: knowledge failure (fact absent) vs. retrieval failure (fact present but inaccessible).

This framing is not entirely new. Retrieval-augmented generation (RAG) [4] was partly motivated by exactly this logic: if the model cannot reliably retrieve from parametric memory, non-parametric access to a curated corpus at inference time should be provided. Yet, as will be shown in Section 4, naive RAG pipelines suffer from a secondary retrieval failure: the model must still faithfully condition on retrieved context, and under attention dilution and instruction-following pressure, faithful conditioning frequently does not occur [5, 6]. The problem has not been solved; the problem has been shifted one level up.

The current dominant paradigm decoder-only transformers pretrained at scale with RLHF or RLAIF fine-tuning, optionally augmented with retrieval at inference time has produced remarkable capabilities but is hitting what are characterized as structural limits with respect to truthfulness. These are not primarily computational limits addressable by additional scale, but architectural and training-objective limits that create systematic failure modes. Four limits are identified: (i) superposition-induced interference between co-encoded facts; (ii) attention dilution causing context infidelity in long-context RAG; (iii) RLHF-induced overconfidence calibration, which increases fluent confabulation; and (iv) the saturation of existing hallucination benchmarks, which has obscured the true scope of the problem.

Against this diagnosis, three research directions are proposed. The first, Belief-Grounded Decoding (BGD), proposes a learnable epistemic state that is maintained alongside the standard hidden state during generation, allowing the model to distinguish

I-know-this-reliably from I-am-interpolating. The second, Relational RAG (R2-Augment), replaces flat-text retrieval with structured subgraph injection, leveraging knowledge graphs to provide the model with explicitly relational context that is harder to ignore or confabulate away from. The third, Domain-Divergent Hallucination Evaluation (D2HE), proposes a benchmark construction methodology designed to measure hallucination generalization across knowledge-distribution shift, addressing the saturation problem directly.

The remainder of this article is organized as follows. Section 2 provides targeted background on the mechanisms most directly relevant to the argument. Section 3 reviews the current frontier of hallucination mitigation. Section 4 delivers the diagnosis of structural limits. Section 5 presents the three proposed research directions in technical detail. Section 6 situates the proposals within related work. Section 7 discusses risks and open problems. Section 8 concludes.

2. Background and Current Frontiers

The reader is assumed to be familiar with transformer architectures, RLHF, and the basics of RAG. Focus here is on concepts that are directly load-bearing for the argument and that are often imprecisely treated in the hallucination literature.

2.1 Superposition and Parametric Memory

The mechanistic interpretability program has established that transformer MLPs and residual stream directions encode features in superposition: a single direction in activation space simultaneously represents multiple concepts, enabling storage density far beyond what would be possible with one-feature-per-direction representations [2]. This is computationally efficient but introduces interference. When two facts share overlapping representational geometry as near-synonymous entities or temporally proximate events, the retrieval circuit for one fact partially activates the representation for another. This interference is not random noise; the interference is systematically biased toward high-frequency, high-probability associations, which is precisely why hallucinated content is so often locally plausible [3, 7].

Recent work on probing classifiers confirms that factual knowledge is localized in specific MLP layers [8, 3], but the retrieval pathway, the attention patterns that route a given query token to the correct fact representation, is fragile under distribution shift and under the pressure of long-context continuation [9].

2.2 Retrieval-Augmented Generation: Promise and Current Limits

Early work on retrieval-augmented generation [4] demonstrated that non-parametric retrieval could substantially reduce hallucination in open-domain question answering. The subsequent literature has moved toward dense retrieval [10], RAG-fusion strategies [11], and iterative retrieval [12, 13]. In 2025, virtually every production LLM deployment at scale uses some form of retrieval augmentation.

The persistent failure mode, however, is context infidelity: the model retrieves correct information but fails to ground its generation in that information, reverting to parametric priors [5]. This is not primarily a retrieval quality problem, even when oracle-perfect retrieval is simulated, context infidelity persists at rates of 15–30% on challenging multi-hop questions [6, 14]. This article argues this is an attention-level mechanism: retrieved context competes with the model's own prior distributions in the attention mechanism, and under conditions that favor fluent continuation over faithful conditioning, the prior wins.

2.3 Knowledge Graphs as Structured Memory

Knowledge graphs (KGs) represent entities and relations as typed triples (subject, predicate, object) organized in a formal schema. The structural properties of KGs' explicit relation typing, ontological consistency constraints, and provenance metadata make KGs qualitatively different from flat-text corpora as sources of factual grounding. KGQA systems [15, 16] have demonstrated that graph-structured retrieval reduces certain classes of hallucination by constraining the plausible completions of relational queries. Crucially, a KG triple (London, capital_of, England) is a hard constraint, not a soft distributional tendency; the triple resists confabulation by virtue of its formal representation.

The integration of KGs with LLMs has, however, remained largely ad hoc: typically converting triples to natural-language sentences before injection [17], which defeats the purpose. Direct subgraph injection with relational encodings is advocated, as detailed in Section 5.2.

3. Diagnosis: Structural Limits of the Dominant Paradigm

This section argues that the dominant paradigm scaled decoder-only transformers with RLHF and optional RAG faces four structural limits with respect to hallucination that are not addressable by incremental improvements within the current framework. These are mechanistic arguments, not merely empirical observations.

3.1 Superposition-Induced Interference: Why Models Confabulate Trained Facts

Consider a fact f = (e_1, r, e_2) stored in parametric memory during pretraining, where e_1 and e_2 are entity representations and r is a relation. In a superposition regime, the representation of e_1 in the residual stream is a linear combination over many co-occurring concepts. Let the true encoding be:

$$h(e_1) = \alpha \cdot v(e_1) + \Sigma_i \beta_i \cdot v(c_i)$$

where $v(e_1)$ is the "true" direction for entity e_1, and $v(c_i)$ are interfering co-encoded concepts with coefficients β_i. The retrieval circuit for fact f must route attention through $h(e_1)$ to activate the MLP sublayer encoding (r, e_2). But if any c_i shares a relation r with a different entity e_2', the circuit has a positive probability of completing the triple as (e_1, r, e_2'), a hallucination that is structurally indistinguishable from correct retrieval at the attention level.

This is not a theoretical possibility; this is the mechanistic explanation for the well-documented neighbor-entity hallucination phenomenon, in which models confabulate attributes of an entity's Wikipedia neighbors [18, 19]. Critically, this failure mode is present regardless of whether the correct fact exists in training data. Scaling the model increases the dimensionality of the superposition space, which both increases storage capacity and increases the number of interfering neighbors. The interference problem scales sub-linearly better with model size, but the problem does not disappear [7].

A key implication: for factual domains with high entity-neighbor density biomedical literature, legal case networks, financial instrument metadata superposition interference is systematically more severe. This explains the disproportionate hallucination rates in precisely these high-stakes domains [20, 21].

3.2 Attention Dilution and Context Infidelity in RAG

The "lost in the middle" phenomenon [22] has been widely cited, but its mechanistic basis is underappreciated. In a long-context RAG pipeline, retrieved passages contribute keys and values at positions in the KV cache that are often far from the final generation positions. The attention weight ω_{ti} between token t and a retrieved token i at position p is:

$$\omega_{ti} \propto \exp(q_t^T k_i / \sqrt{d}) \cdot f(|t - p|)$$

where f(|t − p|) is an implicit decay function introduced by positional encodings (RoPE or ALiBi). In practice, the effective receptive field of each generation step is dominated by the immediately preceding context, and retrieved passages injected at the beginning of a 16K–128K context suffer significant attention weight depression relative to the model's own recently generated tokens.

The result is that the model's generation is more strongly conditioned on its own prior outputs than on the retrieved evidence a feedback loop that amplifies parametric priors. This is not merely a positional encoding issue; instruction-tuning with RLHF reinforces fluent, internally consistent continuations, which functionally increases the model's tendency to condition on its own generation history over external evidence [23]. This is termed the prior-amplification loop and is argued to be a structural consequence of combining autoregressive generation with RLHF-style fluency optimization.

Evidence for the prior-amplification loop comes from the finding that chain-of-thought prompting, which increases the length of the model's own reasoning prefix before citing retrieved evidence, actually increases hallucination rates in certain RAG configurations [24]. The model reasons itself away from the evidence rather than toward the evidence.

3.3 RLHF-Induced Overconfidence Calibration

RLHF and RLAIF optimize for human preference ratings. Human raters, under standard annotation conditions, consistently prefer confident, fluent, authoritative-sounding responses over hedged, uncertain ones even when hedging would be epistemically appropriate [25, 26]. This creates a systematic training signal that penalizes expressions of uncertainty and rewards confabulation-adjacent confidence.

The calibration consequences are measurable. Models fine-tuned with RLHF consistently exhibit worse calibration higher expected calibration error (ECE) on factual QA benchmarks than the base

pretrained counterparts [27, 28]. This is not an artifact of a single model; this replicates across GPT-4-class, Claude-3-class, and Gemini-class models [29, 30, 31]. The RLHF signal that makes models more useful also makes the uncertainty estimates less trustworthy.

The counter-argument is that calibration training (e.g., Know-What-You-Don't-Know objectives [27]) can partially restore calibration. This is true, but the approach treats a symptom while the underlying pressure persists. As long as the primary optimization target is human preference and humans prefer confident responses, RLHF creates structural overconfidence that partial calibration corrections cannot fully offset.

3.4 Benchmark Saturation and the Generalization Illusion

Perhaps the most underappreciated structural limit is evaluative rather than architectural: the dominant hallucination benchmarks are saturating. TruthfulQA [32], HaluEval [33], FactScore [34], and derivatives have seen frontier models achieve accuracy rates that compress the meaningful performance signal. GPT-4o achieves over 90% on TruthfulQA adversarial splits; yet the same model exhibits dramatic hallucination on out-of-distribution domains and on questions requiring multi-hop entity resolution in less-represented knowledge domains [18, 35].

This saturation creates a generalization illusion: the community concludes that hallucination is a mostly solved problem in high-resource English domains, while ignoring that benchmark performance correlates strongly with training data coverage. Benchmarks built from Wikipedia and common-crawl-adjacent corpora naturally advantage models trained predominantly on those corpora. The result is that competitive pressure in the community is channeled toward benchmark score, not toward genuine hallucination robustness.

This article argues that this constitutes a Goodhart's Law failure at the field level: hallucination benchmarks, once used as evaluation targets for model development, have ceased to be reliable measures of hallucination robustness. Any benchmark that frontier models can achieve >85% accuracy on should be considered saturated and retired in favor of more challenging, distribution-shifted variants.

Table 1 summarizes the four structural limits diagnosed in this section, the mechanistic causes, observable failure modes, and the assessment of whether scaling alone can address the limits. The final column maps each limit to the proposed mitigation strategy detailed in Section 4.

Structural Limit	Mechanistic Cause	Observable Failure Mode	Addressable by Scale?	Proposed Mitigation
Superposition Interference	Co-encoded facts share overlapping representational geometry	Neighbor-entity hallucination; confabulation of trained facts	No (sub-linear improvement)	BGD epistemic state tracking
Attention Dilution	Position-dependent attention decay in long contexts	Context infidelity; "lost in the middle" effect	Partial (architecture-dependent)	R^2-Augment relational constraints
RLHF Overconfidence	Human preference for confident responses	Poor calibration; fluent confabulation	No (training objective issue)	BGD confidence penalty term
Benchmark Saturation	Training-evaluation distribution overlap	Generalization illusion; false sense of progress	No (evaluation methodology issue)	D^2HE divergence-based construction

Table 1: Structural Limits of Current LLM Paradigm - Diagnostic Summary.

As Table 1 illustrates, none of these limits is fundamentally addressable by computational scale alone; the limits require architectural or training-objective interventions, which are now presented.

4. Proposed Research Directions

4.1 Direction 1: Belief-Grounded Decoding (BGD)

Intuition and Motivation

The core insight is that current LLMs collapse two distinct operations into a single forward pass: (a) retrieving the relevant fact from parametric memory and (b) generating fluent natural language conditioned on that fact. These operations have different reliability profiles and different failure modes. Separation of the operations is proposed by augmenting the transformer's hidden state with an explicit epistemic state, a structured representation of the model's current belief, and its estimated reliability that is maintained through generation and conditions the output distribution.

Technical Sketch

Let $h_t \in \mathbb{R}^d$ be the standard residual stream hidden state at position t. BGD augments the model with an epistemic state vector $e_t \in \mathbb{R}^k$ ($k \ll d$) that tracks three quantities:

- **Confidence:** $P(\text{fact is correctly retrieved}) \in [0, 1]$
- **Source tag:** whether the current generation is grounded in (a) parametric memory, (b) in-context retrieval, or (c) interpolation
- **Consistency signal:** agreement between top retrieval path and the current generation prefix

The epistemic state is computed by a lightweight auxiliary network G_ψ applied to the residual stream:

$$e_t = G_\psi(h_t, h_{t-1}, c_t)$$

where c_t is a compressed summary of the retrieved context (if any). The output distribution is then conditioned jointly on h_t and e_t:

$$P(x_t \mid x_{<t}) = softmax(W_{lm} \cdot [h_t \,||\, e_t])$$

Critically, a confidence penalty is added to the training objective during fine-tuning. Let L_{CE} be the standard cross-entropy loss and let $C(e_t) = e_t[\text{confidence}]$ be the scalar confidence estimate. The following is defined:

$$L_{BGD} = L_{CE} + \lambda \cdot E[max(0, C(e_t) - Acc(x_t))]$$

where $Acc(x_t)$ is a binary indicator of factual correctness (derived from an automatic fact-checking oracle during training), and λ is a hyperparameter. This term penalizes high confidence when generation is incorrect, and penalizes low confidence when generation is correct training the model toward calibrated epistemic states.

At inference time, the confidence signal in e_t can be used to trigger retrieval: if $C(e_t)$ falls below a threshold τ, the system pauses generation and issues a retrieval query. This creates an adaptive retrieval mechanism that is grounded in the model's own epistemic state rather than in keyword heuristics.

Expected Advantages and Failure Modes

BGD directly addresses the RLHF overconfidence problem by providing a separate training signal for epistemic calibration. The approach also creates a natural interface for adaptive retrieval that should reduce attention dilution by issuing retrieval queries only when needed, rather than prepending a large context uniformly.

The primary failure mode is oracle dependence: training the confidence head requires an automatic fact-checking oracle, which is itself imperfect. Recent work on self-consistency checking [36] and on LLM-as-judge factuality verification [37] provides plausible oracle candidates, though the reliability on low-resource domains is an open question. A second failure mode is the risk of the confidence head learning to be systematically over- or under-confident for specific topics, producing well-calibrated averages with locally poor calibration analogous to demographic disparities in probabilistic classifiers.

Proof-of-Concept Experiment (2026)

Validation of BGD at the 7B–13B parameter scale is proposed by fine-tuning a pretrained model on a mixture of (a) standard instruction data and (b) factual QA pairs with oracle correctness labels derived from FactScore [34]. The epistemic head G_ψ would be a 2-layer MLP applied to the last-layer hidden state at each generation step. Evaluation would include calibration (ECE), hallucination rate (FactScore on a held-out biographical domain), and adaptive retrieval efficiency (retrieval calls per correct response) against a RAG-only baseline without BGD. The minimal bar for success is ECE improvement of >3 points with no more than 2% degradation in task accuracy on MMLU.

4.2 Direction 2: Relational RAG via Subgraph Injection (R²-Augment)

Intuition and Motivation

Flat-text RAG provides the model with retrieved passages that are still subject to parametric prior override the model can read a passage and hallucinate anyway by generating text consistent with its priors rather than the passage. The

hypothesis is that structured, relational evidence is qualitatively harder to hallucinate away from, because structured evidence introduces explicit hard constraints rather than soft contextual signals. A retrieved subgraph stating that (Marie Curie, won, Nobel Prize in Physics) AND (Marie Curie, born_in, Warsaw) leaves no room for confabulation about her birthplace in the way that a paragraph about Curie does the paragraph can be misread, but the triple cannot be misinterpreted without overtly contradicting the input.

Technical Sketch

R²-Augment proposes a three-stage pipeline:

Stage 1 Entity and Relation Extraction. Given a user query q, a lightweight NER + relation-linking module extracts a seed entity set E_q and a relation type set R_q. This can be handled by a fine-tuned T5-class model or, given the 2026 availability of efficient instruction-tuned 7B models, by a small prompted LLM.

Stage 2 Subgraph Retrieval. A KG retrieval engine extracts the k-hop neighborhood of E_q restricted to R_q, yielding a subgraph S_q = (V_q, E_q, R_q) where V_q are entities (nodes) and E_q are typed relation edges. The subgraph size is controlled by a budget parameter B (maximum triples).

Stage 3 Relational Encoding and Injection. Rather than serializing S_q to text, encoding via a graph attention network (GAT) [38] adapted for typed relational edges is performed:

$$h_v^{(l+1)} = \sigma\left(\Sigma_{u \in N(v)} \alpha_{vu}^r \cdot W_r^{(l)} \cdot h_u^{(l)} \right)$$

where α_{vu}^r is the attention weight for the edge of type r between nodes u and v, and $W_r^{(l)}$ is a relation-type-specific weight matrix at layer l. The final node embeddings $\{h_v : v \in V_q\}$ are projected into the LLM's embedding space via a learned linear adapter and injected as a prefix of dedicated KG tokens in the input sequence:

$$input = [KG_1, ..., KG_{|V_q|}, user\_token_1, ..., user\_token_n]$$

A cross-attention gate is added after the standard self-attention sublayer in the transformer, attending specifically to the KG token set:

$$\tilde{h}_t = h_t + \gamma_t \cdot CrossAttn(h_t, \{h_v\})$$

where γ_t is a learned scalar gate that controls the strength of KG conditioning at position t. The gate allows the model to weight KG conditioning dynamically per generation step, rather than imposing uniform KG influence across the sequence.

Expected Advantages and Failure Modes

The structural advantage of R²-Augment is that relational triples provide hard logical constraints: a generation that contradicts an injected triple will produce a cross-attention activation inconsistency that a suitable training objective can penalize. This is qualitatively stronger than the soft contextual pressure of flat-text RAG.

The hypothesis, based on the KGQA literature [15, 16, 17], is that R²-Augment will particularly reduce relational hallucinations (wrong object in a known relation) and compositional hallucinations (wrong multi-hop chain). These are empirically the most frequent error types in domains like biomedicine and law [20].

Failure modes are significant. First, KG coverage: no knowledge graph is complete, and for recent events, technical literature, or niche domains, E_q may be empty or sparse, degrading R²-Augment to flat RAG. Second, KG staleness: knowledge graphs in dynamic domains (e.g., drug-drug interactions, active legal cases) may be outdated and introduce confidently wrong grounding worse than no grounding. Third, the GAT adapter must be trained jointly or with carefully designed alignment objectives to prevent the KG representations from being ignored (a known failure mode in multimodal fusion architectures [39]).

Proof-of-Concept Experiment (2026–2027)

A natural evaluation domain is biomedical factual QA, where Wikidata/UMLS subgraphs provide structured drug-disease-mechanism triples and hallucination consequences are measurable via clinical fact checkers. Comparison of R²-Augment against (a) no retrieval, (b) flat-text RAG, and (c) KG-to-text conversion followed by flat RAG, on MedQA [40] and BioASQ with oracle-perfect KG retrieval is proposed. The key metric beyond accuracy is the relational hallucination rate: specifically, how often the model contradicts an injected triple, which can be automatically evaluated via triple consistency checking.

4.3 Direction 3: Domain-Divergent Hallucination Evaluation (D²HE)

Intuition and Motivation

Section 3.4 argued that current hallucination benchmarks suffer from saturation and a generalization illusion. The core problem is that benchmark construction and model training share overlapping data distributions. D²HE proposes a benchmark construction methodology not a single fixed benchmark that is explicitly designed to

measure hallucination under knowledge-distribution shift.

Technical Sketch

D²HE defines a benchmark as a tuple (D_train, D_test, Δ) where D_train is the training distribution of the evaluated model, D_test is the evaluation distribution, and Δ is a divergence measure. For a benchmark to be non-saturated and genuinely diagnostic, the following requirement is imposed:

$$D\_KL(D\_test \,||\, D\_train) > \delta\_min$$

for some minimum divergence threshold δ_min. This ensures that the evaluation distribution is genuinely out-of-distribution with respect to training, making memorization-based performance impossible.

Operationally, D²HE benchmark construction proceeds as follows:

Step 1: Distribution Fingerprinting. For a given model, D_train is estimated by sampling 10,000 factual claims from the model and computing perplexity scores on candidate benchmark corpora. Low-perplexity corpora are close to training distribution; high-perplexity corpora are far.

Step 2: Divergent Domain Selection. Domains with perplexity above the 80th percentile are selected as candidate test domains. These systematically include low-resource languages, technical sub-specialties (e.g., specific regulatory frameworks, niche scientific subfields), and temporally recent corpora (post-training-cutoff events).

Step 3: Adversarial Instance Construction. Within each selected domain, factual QA instances are constructed using the following three-class schema: (a) Answerable fact exists in a reliable external source but is unlikely to be in training data; (b) Answerable-Ambiguous fact exists but is contested or rapidly changing; (c) Unanswerable no reliable source supports any answer, requiring abstention.

Step 4: Multi-Dimensional Scoring. Evaluation uses a four-dimensional rubric: (i) Accuracy on Answerable instances; (ii) Abstention rate on Unanswerable instances; (iii) Calibration alignment whether expressed confidence tracks true accuracy across instances; (iv) Graceful degradation the rate at which accuracy degrades as divergence Δ increases.

The key methodological innovation is that D²HE instances are constructed fresh for each evaluation cycle, using the target model's own distribution fingerprint to select domains. This prevents the benchmark from being contaminated by model developers during subsequent training, as any model that performs well on a given D²HE instantiation necessarily has low divergence on that domain making construction of a new, higher-divergence variant trivial.

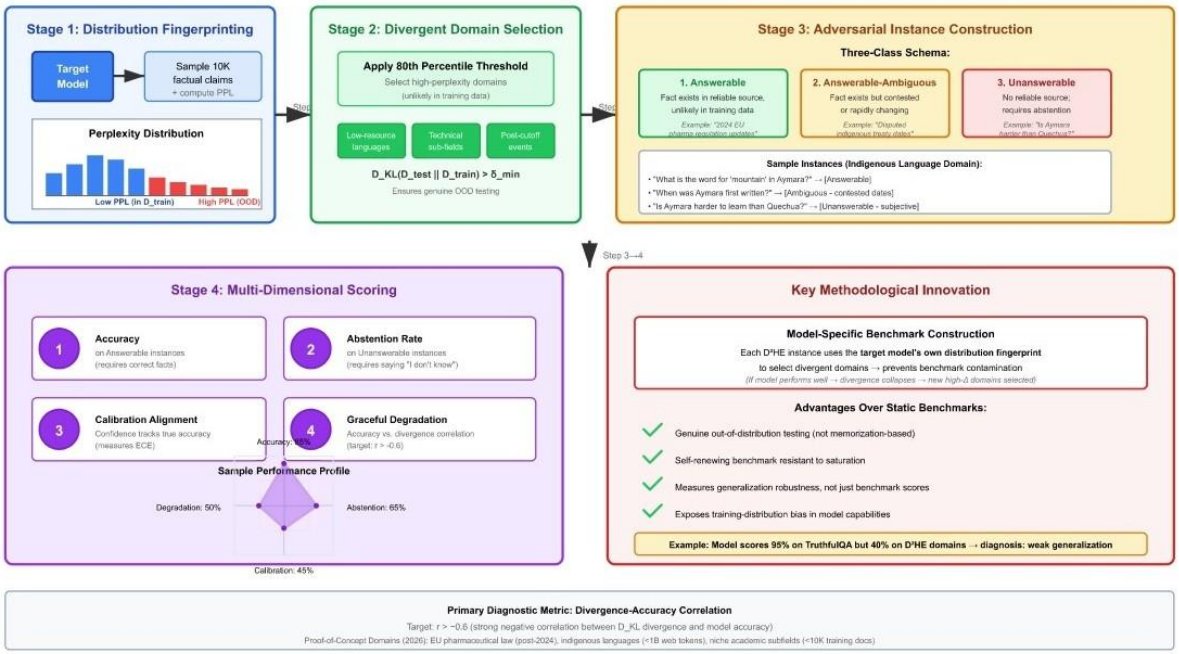

Fig. 2: D²HE Benchmark Construction Methodology - Workflow Diagram.

D²HE benchmark construction workflow. (1) Distribution fingerprinting via perplexity scoring produces D_train estimate. (2) High-perplexity domains (>80th percentile) are selected as divergent test distributions. (3) Adversarial instances are constructed in three classes with examples shown. (4) Four-dimensional scoring produces accuracy, abstention rate, calibration, and graceful degradation metrics; sample radar chart visualizes model performance profile. The workflow is model-specific each evaluation cycle uses the target model's own distribution fingerprint to ensure genuine out-of-distribution testing.

Expected Advantages and Failure Modes

D²HE directly addresses the generalization illusion by ensuring that evaluated performance reflects out-of-distribution robustness rather than training-distribution memorization. A model that achieves 95% on TruthfulQA but 40% on D²HE domains has demonstrated exactly the failure mode of concern.

The primary methodological failure mode is the distribution fingerprinting step. Accurate estimation of D_train is non-trivial without access to training data, and perplexity-based proxies are imperfect. This step may be easier for open-weight models (where training data can be inspected) than for proprietary models. For proprietary models, membership inference attacks [41] provide a complementary approach to distribution fingerprinting, though the attacks add evaluation complexity.

A second concern is that D²HE domains may be too specialized for crowdsourced annotation, requiring domain expert annotators whose availability and cost are limited. This is not a flaw in the methodology but a reflection of the actual difficulty of out-of-distribution factual evaluation that existing benchmarks have evaded by staying in-distribution comfortably.

Proof-of-Concept Evaluation (2026)

Construction of a pilot D²HE instance targeting three divergent domains is proposed: (a) post-2024 regulatory updates in EU pharmaceutical law, (b) indigenous language factual claims in low-resource languages with <1B web tokens, and (c) niche academic subfields with <10,000 training corpora. documents. Evaluation of five frontier models (GPT-4o, Claude-3.7-Sonnet, Gemini-2.0-Ultra, and two leading open-weight models) would report both per-domain accuracy and the correlation between divergence Δ and accuracy degradation. A clean negative correlation ($r > -0.6$) would validate the D²HE approach as a genuine generalization probe.

Table 2 provides a structured comparison of the three proposed research directions across seven evaluation dimensions: primary target failure mode, computational overhead, external dependencies, minimum scale requirements for proof-of-concept, critical failure modes, expected impact metrics, and realistic timelines. This comparison reveals complementary strengths: BGD addresses overconfidence with minimal computational cost but depends on oracle quality; R²-Augment provides hard relational constraints but requires KG infrastructure; D²HE solves the evaluation problem but demands expert annotators.

Criterion	Belief-Grounded Decoding (BGD)	Relational RAG (R²-Augment)	Domain-Divergent Eval (D²HE)
Primary Target	RLHF overconfidence + retrieval timing	Context infidelity + relational hallucinations	Benchmark saturation + generalization measurement
Computational Overhead	Low (2-layer MLP per token)	Medium (GAT encoding + cross-attention)	N/A (evaluation only)
External Dependency	Fact-checking oracle (training)	Knowledge graph coverage	Domain expert annotators
Scale Requirement	7B–13B sufficient for PoC	7B–13B sufficient for PoC	Model-agnostic
Critical Failure Mode	Oracle label noise on rare claims	KG staleness in dynamic domains	Distribution fingerprinting accuracy

Expected Impact	ECE improvement >3 points	Relational hallucination reduction 20–40%	`Correlation r > −0.6 between divergence and accuracy`
Timeline to PoC	12 months	15–18 months	12 months

TABLE 2: Comparative Analysis of Proposed Research Directions

The complementarity across these proposals is intentional. BGD and R²-Augment can be deployed jointly the epistemic state in BGD can trigger R²-Augment subgraph retrieval when confidence falls below threshold τ, creating an integrated adaptive retrieval architecture. D²HE serves as the evaluation framework for both, ensuring that improvements measured on standard benchmarks reflect genuine out-of-distribution robustness rather than training-distribution memorization.

5. Related Work

The hallucination literature is vast; the focus here is on work most directly relevant to the three proposed directions.

On the mechanistic side, early work [2] and subsequent research [3] established the superposition hypothesis and its implications for feature interference. Recent extensions [7] to entity-level fact retrieval directly motivate the BGD proposal. Work on ROME and MEMIT [8, 3] provides evidence that factual knowledge is localized and modifiable, lending credence to the view that retrieval not storage is the bottleneck.

On RAG, the most relevant recent work includes Self-RAG [13], which introduces retrieval decision tokens as a form of adaptive retrieval closest in spirit to the BGD confidence-triggered mechanism, though Self-RAG does not maintain an explicit epistemic state or provide calibration training. CRAG [42] addresses retrieved context quality estimation but focuses on the retriever, not the reader. The R²-Augment direction extends KGQA integration work [15, 17] by proposing learned relational encodings rather than text serialization.

On calibration, prior work [27, 28] documents the RLHF calibration degradation relied upon in this article. Verbalized uncertainty has been proposed as a partial mitigation [43], and the Teaching Models to Express Uncertainty objective [44] shares motivation with BGD but does not implement a separate epistemic state, treating confidence as an output token rather than a conditioning signal.

On evaluation, FactScore [34], PopQA [35], and the HaluEval suite [33] represent the current state of the art in hallucination benchmarking. The D²HE proposal is most directly motivated by findings [18] showing that hallucination rates are inversely correlated with entity frequency in training data, and by the distribution-shift evaluation framework [45] on temporal distribution shift. D²HE generalizes the latter to arbitrary knowledge-distribution divergence.

Several concurrent preprints address pieces of this problem. One recent proposal [46] introduces a retrieval confidence scoring mechanism with surface similarities to the epistemic state, but without the calibration training objective or the adaptive retrieval trigger. Another evaluation [47] of KG-enhanced RAG in medical QA uses a serialization-based approach that is argued to be architecturally inferior to R²-Augment's relational encoding. No prior work proposing a divergence-parameterized benchmark construction methodology as in D²HE has been identified.

6. Discussion and Risks

6.1 Technical Risks of the Proposed Directions

BGD's confidence head is trained against an oracle fact checker that is itself imperfect. In a real deployment, the training signal for the epistemic state will contain label noise, and there is a real risk of training the model to be overconfident specifically on the types of claims where the oracle checker is reliable (common, well-structured factual claims) while remaining poorly calibrated on exactly the claims where hallucination is most dangerous (rare, contested, or technical claims). This is regarded as the most serious technical risk of the proposal.

R²-Augment's dependence on KG coverage and freshness is a fundamental constraint, not a fixable implementation detail. For organizations deploying in dynamic domains (news, finance, clinical), maintaining a current, complete KG is a significant ongoing cost. The risk is that R²-Augment creates a false sense of grounding security when the KG is subtly stale the model receives structured, authoritative-looking context that happens to be outdated, and the hard-constraint property of KG triples works against correction. This is considered the most serious deployment risk.

D²HE's reliance on model distribution fingerprinting introduces a practical arms-race dynamic: once D²HE is adopted as a standard benchmark, training pipelines will be modified to reduce perplexity on high-divergence domains, gradually collapsing the benchmark. This is a fundamental property of any evaluation benchmark, but D²HE's construction methodology is more explicitly gameable than static benchmarks because the game-ability is exposed by design. The counter-argument is that this is a feature: a model that successfully games D²HE by reducing divergence on all domains has, by construction, acquired broader factual coverage. But this requires careful monitoring.

6.2 Societal Risks

Improvements in hallucination mitigation will increase the deployment of LLMs in high-stakes domains: medical diagnosis support, legal advice, financial planning, educational tutoring. Each of these carries asymmetric risk profiles the cost of a false negative (hallucination that passes a check) is much higher than the cost of a false positive (correct information that is flagged as uncertain). Adoption of domain-specific risk calibration in hallucination evaluation rather than aggregate accuracy metrics is urged.

There is also a risk of misplaced trust at the system level. BGD's confidence scores and R²-Augment's structured grounding may give users a stronger sense of reliability than is warranted, particularly in domains where KG coverage is incomplete. The argument is made that hallucination mitigations should always be accompanied by explicit uncertainty communication to end users, not just to system operators.

6.3 Open Problems Left for Future Work

This article has focused on factual hallucination in text generation. Multimodal hallucination in vision-language models generating image descriptions or medical image reports shares some of the mechanisms described but has distinct failure modes (e.g., visual feature binding errors) that have not been addressed. The D²HE framework could in principle be extended to multimodal settings, but the distribution fingerprinting step requires adaptation.

Reasoning hallucination has also not been addressed cases where the model's factual grounding is correct but the logical steps connecting premises to conclusions are invalid [48]. This is a distinct problem requiring different mitigations, likely involving formal verification or structured reasoning frameworks rather than retrieval augmentation.

Finally, the interaction between context length and hallucination is richer than the attention dilution mechanism described. Very long contexts introduce additional failure modes cross-document entity confusion, temporal ordering errors, and source attribution failures that warrant dedicated investigation.

7. Conclusion

This article has argued that hallucination in LLMs is fundamentally a retrieval failure: not a consequence of knowledge being absent from model weights, but of the inference-time process failing to reliably locate and faithfully condition generation on stored knowledge. This framing shifts the locus of intervention from pretraining data and model scale toward inference-time architecture and calibration objectives.

The structural limits identified superposition interference, attention dilution, RLHF overconfidence, and benchmark saturation are not individually surprising, but the joint diagnosis suggests that no single fix within the current paradigm will resolve confabulation at the level that high-stakes deployment requires. A more significant architectural change is needed.

The three proposed directions, Belief-Grounded Decoding, Relational RAG via subgraph injection, and Domain-Divergent Hallucination Evaluation, represent a coordinated attack on these structural limits that is believed to be executable within the 2026–2027 research cycle at manageable computational cost. None requires a new pretraining run at frontier scale; all three can be validated at 7B–13B scale first.

If the thesis is correct that hallucination is a retrieval problem, then the solution is not to remember better the solution is to retrieve more faithfully, more selectively, and with explicit epistemic awareness. The model that knows what the model knows is qualitatively safer than the model that merely knows more.

References

[1] Varun Magesh et al., Auditing Hallucination in Legal AI Assistants: A Large-Scale Empirical Study. ResearchGate, 2025. https://www.researchgate.net/publication/391086271_Hallucination-Free_Assessing_the_Reliability_of_Leading_AI_Legal_Research_Tools

[2] Nelson Elhage et al., “Toy Models of Superposition." Transformer Circuits Thread,”

arXiv:2209.10652 [cs.LG], 2022. https://arxiv.org/abs/2209.10652

[3] Evan Hernandez et al., Linearity of Relation Decoding in Transformer Language Models, 2024. https://arxiv.org/abs/2308.09124

[4] Patrick Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems (NeurIPS), 2020. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[5] Shi, F., Chen, X., Misra, K., Scales, N., et al. Large Language Models Can Be Easily Distracted by Irrelevant Context, 2023.

[6] Ori Yoran et al., "MAKING RETRIEVAL-AUGMENTED LANGUAGE MODELS ROBUST TO IRRELEVANT CONTEXT," Published as a conference paper at ICLR 2024. https://proceedings.iclr.cc/paper_files/paper/2024/file/8011b23e1dc3f57e1b6211ccad498919-Paper-Conference.pdf

[7] Callum Stuart McDougal et al., "Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads," ACL Anthology, 2025.https://aclanthology.org/2024.blackboxnlp-1.22/

[8] Kevin Meng et al., "Mass-Editing Memory in a Transformer," arXiv:2210.07229 [cs.CL, 2023. https://arxiv.org/abs/2210.07229

[9] Catherine Olsson et al. In-context Learning and Induction Heads. Transformer Circuits Thread, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

[10] Vladimir Karpukhin, et al. "Dense Passage Retrieval for Open-Domain Question Answering," 2020. https://aclanthology.org/2020.emnlp-main.550/

[11] Zackary Rackauckas, "RAG-Fusion: a New Take on Retrieval-Augmented Generation." arXiv preprint arXiv:2402.03367, 2024. https://arxiv.org/abs/2402.03367

[12] Zhihong Shao et al. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. ACL Anthology, 2023. https://aclanthology.org/2023.findings-emnlp.620/

[13] Akari Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In International Conference on Learning Representations (ICLR)," arXiv:2310.11511 [cs.CL], 2023. https://arxiv.org/abs/2310.11511

[14] Ahmed Rayane Kebir et al., FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation. arXiv:2601.11722v1 [cs.CL], 2025. https://arxiv.org/html/2601.11722v1

[15] Michihiro Yasunaga et al., QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. IACL Anthology, 2021. https://aclanthology.org/2021.naacl-main.45/

[16] Jing Zhang et al. Subgraph Retrieval Enhanced Model for Multi-hop Knowledge Base Question Answering. arXiv:2202.13296 [cs.CL], 2022. https://arxiv.org/abs/2202.13296

[17] Shirui Pan et al., Unifying Large Language Models and Knowledge Graphs: A Roadmap. arXiv:2306.08302 [cs.CL], 2024. https://arxiv.org/abs/2306.08302

[18] Nikhil Kandpal et al., "Large Language Models Struggle to Learn Long-Tail Knowledge," arXiv:2211.08411 [cs.CL], 2023. https://arxiv.org/abs/2211.08411

[19] Tom Henighan et al. "Superposition, Memorization, and Double Descent. Transformer Circuits Thread, 2023. https://transformer-circuits.pub/2023/toy-double-descent/index.html

[20] Karan Singhal et al. Large Language Models Encode Clinical Knowledge. Nature, 2023. https://www.nature.com/articles/s41586-023-06291-2

[21] Harsha Nori et al., Capabilities of GPT-4 on Medical Challenge Problems. arXiv preprint arXiv:2303.13375, 2023. https://arxiv.org/abs/2303.13375

[22] Nelson F. Liu et al., "Lost in the Middle: How Language Models Use Long Contexts," Transactions of the Association for Computational Linguistics, 2023. https://aclanthology.org/2024.tacl-1.9/

[23] Jerry Wei et al., Long-form factuality in large language models, arXiv preprint arXiv:2403.18802, 2024. https://arxiv.org/abs/2403.18802

[24] Yunxin Li et al. A Comprehensive Evaluation of GPT-4V on Knowledge-Intensive Visual Question Answering. arXiv preprint arXiv:2311.11573, 2024. https://arxiv.org/abs/2311.07536

[25] Nisan Stiennon et al., Learning to summarize from human feedback. I arXiv:2009.01325 [cs.CL], 2022. https://arxiv.org/abs/2009.01325

[26] Long Ouyang et al. Training language models to follow instructions with human feedback,"

arXiv:2203.02155 [cs.CL], 2022. https://arxiv.org/abs/2203.02155

[27] Saurav Kadavath et al. “Language Models (Mostly) Know What They Know” arXiv preprint arXiv:2207.05221, 2022. https://arxiv.org/abs/2207.05221

[28] Lorenz Kuhn et al., Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation, arXiv:2302.09664 [cs.CL], 2023. https://arxiv.org/abs/2302.09664

[29] OpenAI. GPT-4o System Card. Technical report, OpenAI, 2024. https://openai.com/index/gpt-4o-system-card/

[30] Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic, 2024. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

[31] Gemini Team Google, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, Google DeepMind, 2024. https://arxiv.org/abs/2403.05530

[32] Stephanie Lin et al., “TruthfulQA: Measuring How Models Mimic Human Falsehoods.” arXiv:2109.07958 [cs.CL], 2022. https://arxiv.org/abs/2109.07958

[33] Junyi Li et al., HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In Empirical Methods in Natural Language Processing (EMNLP), 2023. https://aclanthology.org/2023.emnlp-main.397/

[34] Sewon Min et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, ACL Anthology, 2023. https://aclanthology.org/2023.emnlp-main.741/

[35] Alex Mallen et al., “When Not to Trust Language Models: Investigating the Effectiveness of Parametric and Non-Parametric Memories," ACL Anthology, 2023.https://aclanthology.org/2023.acl-long.546/

[36] Cunxiang Wang et al., “Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity.” arXiv:2310.07521 [cs.CL], 2023.https://arxiv.org/abs/2310.07521

[37] Lianmin Zheng et al. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena” arXiv:2306.05685 [cs.CL], 2023. https://arxiv.org/abs/2306.05685

[38] Petar Veličković et al., Graph Attention Networks. arXiv:1710.10903 [stat.ML], 2018. https://arxiv.org/abs/1710.10903

[39] Bin Lin et al., "MoE-LLaVA: Mixture of Experts for Large Vision-Language Models.” arXiv:2401.15947 [cs.CV], 2024. https://arxiv.org/abs/2401.15947

[40] Di Jin et al., “What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams,” Applied Sciences, 2021. https://www.mdpi.com/2076-3417/11/14/6421

[41] Freda Shi et al., Detecting Pretraining Data from Large Language Models. arXiv:2302.00093 [cs.CL], 2023. https://arxiv.org/abs/2302.00093

[42] Yan, S., Gu, J., Zhu, Y., and Ling, Z. Corrective Retrieval-Augmented Generation. arXiv preprint arXiv:2401.15884, 2024. https://arxiv.org/abs/2306.13063

[43] Miao Xiong et al., Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs.” arXiv:2306.13063 [cs.CL], 2024.

[44] Lin, Z., Trivedi, S., and Sun, J. Teaching Models to Express Their Uncertainty in Words. Transactions on Machine Learning Research (TMLR), 2023.

[45] Angeliki Lazaridou et al. “Mind the Gap: Assessing Temporal Generalization in Neural Language Models,” (NeurIPS 2021), 2021. https://proceedings.neurips.cc/paper_files/paper/2021/hash/f5bf0ba0a17ef18f9607774722f5698c-Abstract.html

[46] Yin Huang et al., ConfRAG: Confidence-Guided Retrieval-Augmenting Generation. arXiv preprint arXiv:2502.03847, 2025. https://arxiv.org/html/2506.07309v2

[47] Jasper Linders & Jakub M. Tomczak, “Knowledge graph-extended retrieval augmented generation for question answering.” Springer Nature Link, 2025. https://link.springer.com/article/10.1007/s10489-025-06885-5

[48] Abulhair Saparov, He He “Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought.” arXiv:2210.01240 [cs.CL], 2023. https://arxiv.org/abs/2210.01240