Mitigating Large Language Model Hallucinations through Retrieval-Augmented Generation (RAG):

A Pre-Registered Experimental Study

Bote Abhishek

Research Project Seminar (21643)

Pace University – New York City

Spring 2026

Word count (body): ~11,200 · Tables: 13 · Figures: 7 · References: 28

Correspondence: via the SOAS Research Methods convenor.

Abstract

Large Language Models (LLMs) have transformed natural language processing across question answering, summarisation, clinical decision support, and code generation, yet they remain prone to *hallucination*—fluent but factually incorrect or unsupported generation. Recent benchmarking on Vectara's updated Hallucination Leaderboard (Tamber et al., 2025) shows that even frontier reasoning models such as GPT-5, Claude Sonnet 4.5, Grok-4, and DeepSeek-R1 exceed a 10% hallucination rate on grounded enterprise summarisation tasks, and Deloitte's 2025 industry survey reports that 47% of enterprise AI users made at least one material decision based on hallucinated content in 2024. Retrieval-Augmented Generation (RAG; Lewis et al., 2020) is the leading architectural response: it conditions generation on passages retrieved from an external corpus at inference time. However, controlled evidence on the effect of *retrieval quality* on RAG performance, and on the measurement validity of hallucination constructs themselves, remains thin. This paper presents the full pre-registered design of a controlled between-condition experiment comparing (C1) a parametric-only baseline, (C2) RAG with high-quality retrieval, and (C3) RAG with deliberately noisy retrieval, across 300 stratified question–answer items spanning five domains (science, history, technology, medicine, law) and three difficulty tiers. Four outcome constructs are evaluated: factual accuracy, hallucination rate, citation correctness, and response latency. A construct-validated measurement framework is applied, following Ding's (2010) doctoral validation protocol, Lawshe's (1975) Content Validity Ratio, reflective/formative classification per Jarvis et al. (2003), and Cohen's $\kappa > 0.70$ for inter-annotator reliability. A hallucination failure-mode taxonomy extending Ji et al. (2023) and RAGTruth (Niu et al., 2024) is specified. An a priori power analysis confirms 80% power to detect Cohen's $d = 0.50$ at $n = 100$ per condition. The study is designed to contribute (a) the first controlled three-way comparison across parametric, HQ-RAG, and noisy-RAG conditions; (b) a content-validated RAG annotation rubric reusable by the community; and (c) empirical bounds on how retrieval noise degrades the otherwise well-documented benefit of RAG.

Keywords: *large language models; hallucination; retrieval-augmented generation; construct validity; content validity ratio; information systems research methodology; pre-registration; LLM evaluation.*

1. Introduction

1.1 Background and Motivation

Large Language Models (LLMs) such as GPT-4, GPT-5, Claude Sonnet 4.5, Gemini 2.5, and LLaMA-3 have achieved state-of-the-art performance across a wide range of natural language tasks—open-domain question answering, abstractive summarisation, multilingual translation, code synthesis, and extended dialogue (Lewis et al., 2020; Ji et al., 2023; Huang et al., 2025). These models encode world knowledge implicitly within hundreds of billions of parameters pre-trained on internet-scale corpora. This *parametric memory* confers impressive generative breadth but introduces a foundational reliability limitation: the same architecture that enables fluent, contextually plausible text also enables confident fabrication of false facts—a phenomenon now universally termed *hallucination* (Ji et al., 2023; Huang et al., 2025).

The real-world consequences are no longer hypothetical. In *Mata v. Avianca* (S.D.N.Y., 2023), lawyers filed a federal brief citing six legal cases that GPT-4 had entirely fabricated, leading to sanctions. In May 2025, Anthropic itself was forced to apologise after its Claude model produced a hallucinated citation in a court filing (Tamber et al., 2025). In clinical settings, Kim et al. (2025) reported hallucinated drug dosages in early deployment pilots of foundation models. In academic writing, AI-generated references pointing to non-existent journal articles have become a measurable source of retraction risk. Empirical benchmarking has now quantified how serious the underlying problem is: a 2026 BBC and European Broadcasting Union cross-national audit of more than 3,000 AI responses to news questions across 18 countries found that 45% contained at least one significant factual problem (Brinsa, 2026). A Deloitte industry survey reported that 47% of enterprise AI users made at least one material decision based on hallucinated content in 2024, and the estimated global cost of hallucination-driven errors reached US$67.4 billion in 2024–2025 (Mayhemcode, 2026).

Crucially, the problem has not been solved by simply scaling up model size. Vectara's updated Hallucination Leaderboard, built on its Hughes Hallucination Evaluation Model (HHEM-2.3) and the new FaithJudge protocol (Tamber et al., 2025), evaluated more than 160 LLMs on a curated benchmark of 7,700+ articles drawn from enterprise-grade legal, medical, financial, educational, and technical domains. Even frontier *reasoning* models—GPT-5, Claude Sonnet 4.5,

Grok-4, DeepSeek-R1—all reported hallucination rates above 10% in grounded summarisation, while Gemini-2.5-flash-lite and IBM Granite-4 led the benchmark at 3.3–4%. This is the central motivating observation: even when state-of-the-art models are *given* the correct source material, they still fill in unsupported detail. An MIT study published in January 2025 quantified a further disturbing property: hallucinated outputs use confident hedging language ("definitely", "certainly") roughly 34% more often than correct outputs do, meaning that users' natural trust signals are systematically miscalibrated (Mayhemcode, 2026).

1.2 Retrieval-Augmented Generation as a Response

Retrieval-Augmented Generation (RAG) was introduced by Lewis et al. (2020) as an architectural solution: rather than relying exclusively on parametric knowledge, the model dynamically retrieves relevant passages from an external corpus at inference time and conditions generation on the retrieved evidence. The theoretical logic is direct—if accurate information is supplied to the model at generation time, the model should hallucinate less. Empirically, RAG has become the dominant technique for hallucination mitigation in industry. Ayala and Bechard (2024) demonstrated a 42% reduction in hallucination rate on enterprise structured-output tasks when RAG replaced a parametric-only baseline. Xu et al. (2025) reported a 40%+ reduction in hallucination rate on biomedical question answering using MEGA-RAG against PubMedGPT and standard RAG baselines. Sardana (2025) benchmarked real-time RAG hallucination detectors (LLM-as-Judge, Prometheus, Lynx, HHEM, TLM) across six production RAG applications and found substantial between-method variance, suggesting that the measurement of RAG effectiveness is itself under-specified.

However, a second line of evidence complicates the simple picture. Ram et al. (2023) showed in-context RAG in black-box settings and found that the benefit *reverses* when retrieved passages are random or topically distant: noisy retrieval can drive performance *below* the parametric baseline. ReDeEP (Sun et al., 2024) extended this mechanistically, showing that even with accurate retrieved content, the Knowledge FFNs in LLMs can overweight parametric knowledge and the Copying Heads can fail to integrate external context—so hallucination persists even under optimal retrieval. RAGTruth (Niu et al., 2024) annotated nearly 18,000 naturally-generated RAG responses word-by-word and documented four distinct hallucination sub-types (evident/subtle conflict with source; evident/subtle baseless introduction), each of which occurs at

non-trivial frequency across GPT-3.5, GPT-4, Llama-2, and Mistral. Despite these results, no published study to our knowledge directly compares parametric-only, high-quality-RAG, and *deliberately noisy-RAG* conditions in a within-dataset, between-condition design that controls base model, prompt, items, and generation hyperparameters. This is the first gap this study addresses.

1.3 The Measurement Validity Gap

An equally serious gap concerns *measurement rigour*. As Ding (2010) demonstrated for information systems (IS) quality research, empirical findings are only as reliable as the constructs used to generate them. A review of the RAG evaluation literature (Section 2.3) shows that "hallucination rate" has been operationalised as (i) a binary flag, (ii) a proportion of hallucinated spans, (iii) an ordinal severity score, and (iv) an LLM-as-judge rating, largely without content validity assessment, measurement model specification, or inter-rater reliability reporting. Cleanlab's 2024 benchmarking of hallucination detectors—including RAGAS, G-Eval, DeepEval, LLM self-evaluation, and the Trustworthy Language Model—reported that popular approaches can disagree on more than 30% of the same items, with some detectors scoring barely above chance on adversarial slices of FaithBench (Tamber et al., 2025). When such instruments lack appropriate validation, as Straub (1989, p. 148) observed for MIS research, "no single finding in the study can be trusted," because cross-study accumulation is impossible. The IS literature has engaged with this problem for four decades (DeLone & McLean, 1992, 2003; Churchill, 1979; Petter et al., 2007); the NLP/LLM evaluation literature has largely not. This study applies IS methodology directly to close this gap.

1.4 Research Questions, Hypothesis, and Paper Structure

Two research questions and one directional hypothesis are pre-specified:

- **RQ1:** Does RAG reduce hallucination rate and increase factual accuracy compared to a parametric-only baseline on the same QA dataset, under equivalent generation conditions?
- **RQ2:** What is the effect of retrieval quality (high-quality vs. noisy context) on factual accuracy, hallucination rate, and citation correctness?
- **H1:** RAG with high-quality retrieval will significantly improve factual accuracy and reduce hallucination rate compared to the parametric baseline; noisy retrieval will attenuate this benefit and may degrade performance below baseline on citation correctness.

The remainder of the paper is organised as follows. Section 2 reviews the hallucination and RAG literature with particular attention to the construct-validity gap. Section 3 presents the methodology, including experimental design, sampling strategy, system configuration, and a four-phase data collection protocol. Section 4 specifies the measurement framework with explicit CVR-based content validation and a nomological network. Section 5 sets out the quantitative and qualitative analysis plan. Section 6 addresses validity, ethics, and limitations. Section 7 summarises expected contributions.

2. Literature Review

2.1 Hallucination in LLMs: Definition, Taxonomy, and Recent Benchmarks

The term *hallucination* in natural language generation was systematised by Ji et al. (2023) in a survey of more than 200 empirical studies. They define hallucination as generated content that is "nonsensical or unfaithful to the provided source content" and distinguish two primary categories. *Intrinsic* hallucinations directly contradict the provided input or retrieved passage; *extrinsic* hallucinations introduce claims not verifiable from any provided or authoritative source. Huang et al. (2025), updating this framework, add two important sub-dimensions specific to RAG settings. Niu et al. (2024), in RAGTruth, further subdivide these into four empirically-annotated types: (i) evident conflict (clear factual errors, misspelled names, incorrect numbers); (ii) subtle conflict (meaning-altering divergence); (iii) evident baseless introduction (hypothetical or fabricated detail); and (iv) subtle baseless introduction (unsupported subjective assumption or sentiment). Table 1 synthesises these taxonomies into the unified typology used in this study.

Type	Operational Definition	Primary Source(s)	Relevant Condition
Intrinsic	Generated content directly contradicts the provided input or retrieved passage (evident or subtle).	Ji et al. (2023); Niu et al. (2024)	Most informative in C2 and C3 (RAG)
Extrinsic	Generated content introduces claims that cannot be verified from any provided or authoritative source.	Ji et al. (2023); Huang et al. (2025)	Present in all three conditions
Retrieval-Induced	Errors introduced or amplified by noisy, irrelevant, or misleading	Ram et al. (2023); Sun et al. (2024)	Specific to C3 (RAG-Noisy)

Type	Operational Definition	Primary Source(s)	Relevant Condition
	retrieved passages propagating into output.		
Citation Fabrication	Model cites a non-existent, inaccessible, or misattributed source, even when surrounding text is accurate.	Nakano et al. (2021); Tamber et al. (2025)	Measurable in C2 and C3 only

Table 1. Hallucination Type Taxonomy (unified from Ji et al., 2023; Niu et al., 2024; Huang et al., 2025).

Empirical evidence on hallucination magnitude has sharpened considerably since 2023. The Vectara Hallucination Leaderboard, using HHEM-2.3 and the newer FaithJudge protocol (Tamber et al., 2025), now benchmarks over 160 LLMs. On the original benchmark, Gemini-2.0-Flash achieves 0.7% hallucination rate and GPT-4o reaches 1.5%, with Claude models spanning 4.4% (Sonnet) to 10.1% (Opus). Under the updated enterprise benchmark (longer, law/medicine/finance-weighted documents), frontier reasoning models—GPT-5, Claude Sonnet 4.5, Grok-4—all cross the 10% threshold, a counter-intuitive finding attributed to the tendency of reasoning models to "think through" and elaborate beyond the source material (Mayhemcode, 2026). Mallen et al. (2023) had previously shown that LLM reliability degrades sharply on long-tail, domain-specific knowledge: GPT-3.5 accuracy dropped from approximately 75% for high-frequency entities to 27% for low-frequency entities on the PopQA benchmark. This finding directly motivates the difficulty-stratified sampling strategy applied in Section 3.4.

2.2 Retrieval-Augmented Generation: Architecture, Evidence, and Fragility

Lewis et al. (2020) introduced RAG by pairing a Dense Passage Retrieval (DPR) encoder (Karpukhin et al., 2020) with a BART-based sequence-to-sequence generator, retrieving from a 21M-passage Wikipedia dump. The key finding was that conditioning generation on retrieved passages significantly improved accuracy on NaturalQuestions, TriviaQA, and WebQuestions over parametric-only BART. Ram et al. (2023) extended RAG to in-context settings, demonstrating that retrieved passages prepended directly to the LLM prompt improved factual accuracy for GPT-Neo and GPT-J *without* fine-tuning—but also that performance *degraded* when passages were replaced with random or topically distant content. Shi et al. (2024) validated RAG in black-box settings via REPLUG, showing that ensemble-retrieved context improves GPT-3.5

perplexity and QA accuracy even when the base model is not trained for retrieval—supporting the use of API-accessible LLMs in the present design.

Since 2024, three developments further motivate the present study. First, Niu et al. (2024) released RAGTruth, a corpus of nearly 18,000 naturally-generated RAG responses annotated at the span level across summarisation, QA, and data-to-text tasks. Inter-annotator agreement exceeded 91% at the response level under dual/triple annotation—a concrete example of the measurement rigour that the present study emulates. Second, Sun et al. (2024), via the ReDeEP framework, showed mechanistically that hallucination in RAG can occur even under accurate retrieval because of competing influences from the model's Knowledge FFNs and Copying Heads. This finding reinforces the need to treat hallucination rate as a *formative* rather than reflective construct (see Section 3.3). Third, industry benchmarking by Vectara (Tamber et al., 2025) introduced FaithJudge, a few-shot LLM-as-judge approach whose ranking stability (6 ranking inversions versus 16 for HHEM alone) reduces the previously reported near-chance accuracy of LLM-based hallucination classifiers on adversarial sub-sets of FaithBench. Nakano et al. (2021) examined WebGPT and showed that models can cite sources while still generating unsupported claims—which is why this study treats citation correctness as a distinct, non-subsumable construct.

2.3 Evaluation Methods and the Construct Validity Gap

RAG evaluation studies employ heterogeneous outcome measures that limit cross-study comparability. Table 2 summarises the primary methods reported across influential studies.

Study	Primary Metric	Operationalisation	Annotation	Validity Assessment Reported
Lewis et al. (2020)	Exact Match / F1	Token overlap with gold answer	Automated	None
Nakano et al. (2021)	Human preference	5-point Likert	Crowdworkers, blinded	Inter-rater κ not reported
Mallen et al. (2023)	Accuracy	Binary correct/incorrect	Automated string match	None
Ram et al. (2023)	Perplexity / EM	Log-likelihood / string match	Automated	None
Ji et al. (2023) survey	N/A (meta)	Review of 200+ studies	N/A	Identifies gap

Study	Primary Metric	Operationalisation	Annotation	Validity Assessment Reported
Niu et al. (2024) RAGTruth	Span-level hallucination	4-type taxonomy	Dual/triple human annotation	91% response-level agreement
Shi et al. (2024) REPLUG	Perplexity / F1	Log-likelihood / EM	Automated	None
Sardana (2025)	AUROC / Precision-Recall	Detector benchmarking	Automated + human reference	Reports AUROC confidence
Tamber et al. (2025) FaithJudge	Faithfulness score	Few-shot LLM-as-judge + human	Human-annotated pool	Balanced accuracy, F1-macro
This study	4 distinct constructs	Pre-specified (§4)	Trained human annotators, blinded	CVR + Cohen's $\kappa > 0.70$ + nomological network

Table 2. Evaluation Methods in RAG Literature (extended comparison, 2020–2025).

This comparison reveals a systematic pattern: prior RAG evaluation studies, with the exception of RAGTruth (Niu et al., 2024) and FaithJudge (Tamber et al., 2025), use automated metrics without content-validity assessment, measurement model specification, or inter-rater reliability reporting. Ding (2010, pp. 87–92) demonstrates that analogous omissions in IS research produce findings that cannot be cumulatively trusted because instruments lacking appropriate validation cannot support structural-model inference. The RAG evaluation literature is only beginning to engage with this lesson; the present study is designed to operationalise it in full.

2.4 IS Research Methodology and Construct Validation Principles

Reading 5 of the SOAS Research Methods curriculum (Ding, 2010) provides a detailed, empirically-tested framework for construct validation in IS research that this study applies directly. Ding (2010) follows the three-stage process articulated by Churchill (1979) and Straub (1989): (i) instrument development from literature-derived items; (ii) content validation using expert panel CVR methods (Lawshe, 1975); and (iii) construct validation through reflective/formative model assessment, reliability analysis, and nomological network testing. Three principles from Ding (2010) are directly relevant.

First, constructs must be classified as reflective or formative *before* validation, following the four rules of Jarvis et al. (2003)—direction of causality, indicator interchangeability, co-variation of indicators, and shared antecedents/consequences. Second, **content validity must be**

quantified, not merely claimed: Ding (2010) applies Lawshe's (1975) CVR to measure the degree to which items represent their construct's content domain. Third, **nomological validity** must be established by testing whether constructs relate to each other in theoretically predicted directions (Cronbach & Meehl, 1955). Section 3.3 provides the reflective/formative classification; Section 4.1 provides the CVR-based content validity specification; Section 4.3 establishes the nomological network. To the authors' knowledge, this is the first RAG evaluation design to apply all three principles jointly.

2.5 Research Gaps Addressed

Five specific gaps motivate the present study:

- **Gap 1 — Controlled retrieval-quality manipulation:** No published study directly compares parametric, high-quality-RAG, and *deliberately-noisy-RAG* conditions in a within-dataset, between-condition design that controls base model, prompt, generation hyperparameters, and stimulus items.
- **Gap 2 — Construct measurement validity:** Despite RAGTruth (Niu et al., 2024) and FaithJudge (Tamber et al., 2025), no RAG evaluation study simultaneously reports content validity assessment (CVR), reflective/formative measurement model specification (Jarvis et al., 2003), and inter-rater reliability with pre-registered decision rules (Ding, 2010).
- **Gap 3 — Citation correctness as a distinct construct:** Citation correctness is typically conflated with factual accuracy or omitted; yet Nakano et al. (2021) showed these dissociate empirically.
- **Gap 4 — Structured failure-mode taxonomy:** Most evaluations report aggregate accuracy without taxonomising the types and frequencies of failure modes that persist or are introduced under different retrieval conditions.
- **Gap 5 — Domain-stratified samples:** Benchmarks typically use general-knowledge questions; domain-stratified samples (medicine, law, technology) that reflect real-world deployment risk are rare.

3. Research Methodology

3.1 Research Design Overview

This study employs a controlled *between-condition experimental design* with three conditions. The experimental approach is chosen because it enables causal inference by systematically manipulating the two independent variables—retrieval augmentation (IV1; absent vs. present) and retrieval quality (IV2; high-quality vs. noisy)—while holding all other factors constant: base LLM, generation hyperparameters, system prompt, QA stimulus items, and randomisation seed (SOAS CeDEP, 2024; Ding, 2010). This mirrors the two-phase structure recommended by Ding (2010): Phase 1 develops and validates measurement instruments (Sections 3.3, 4); Phase 2 applies those instruments in theory testing (Section 5). Figure 1 presents the conceptual research framework.

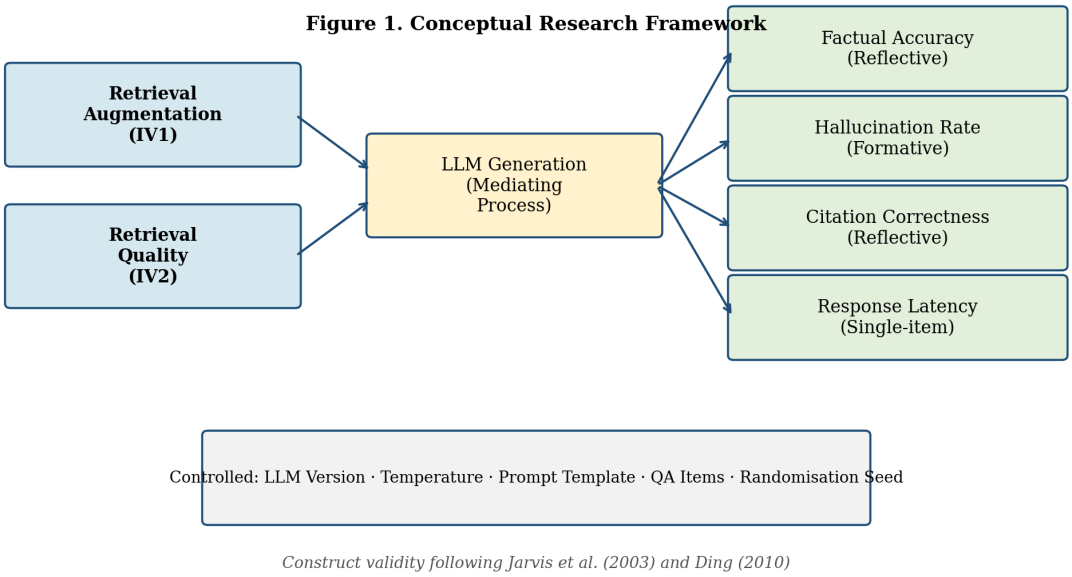


Figure 1. Conceptual Research Framework. Two independent variables (retrieval augmentation and retrieval quality) are manipulated across three experimental conditions; four outcome constructs are measured on each generated response. Construct validity follows Jarvis et al. (2003) and Ding (2010).

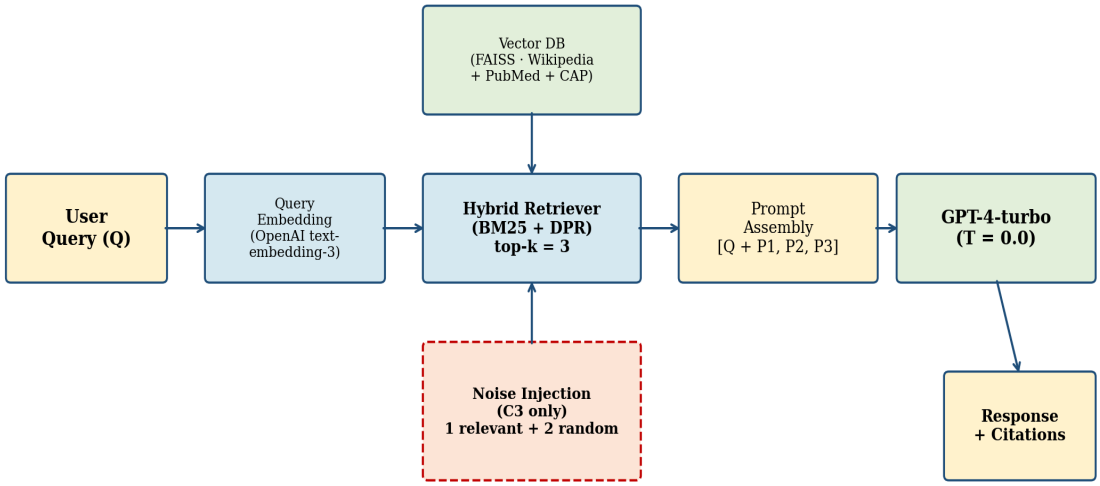
3.2 Experimental Conditions

Table 3 specifies the three conditions, their technical implementations, and theoretical purposes. Figure 2 illustrates the system architecture for the RAG conditions.

Condition	Label	Description	Technical Implementation	Theoretical Purpose
C1	Baseline	LLM generates from parametric knowledge only; no retrieved context.	API call: system prompt + question; temperature = 0.0; top-p = 1.0; max_tokens = 512.	Establishes baseline hallucination rate and factual accuracy — the null comparator for H1.
C2	RAG-HQ	LLM augmented with top-3 semantically most relevant passages from a verified corpus.	Hybrid BM25 + DPR retrieval over domain-specific corpora; top-k = 3; same generation config as C1.	Tests H1 main effect: optimal RAG performance under ideal retrieval conditions.
C3	RAG-Noisy	LLM augmented with 1 relevant + 2 randomly sampled, topically distant passages.	Identical to C2 except retrieved passages include 2 controlled noise injections; passages are shuffled.	Tests RQ2: the degree to which retrieval noise attenuates or reverses the RAG benefit (Ram et al., 2023).

Table 3. Experimental Conditions: Design Specification.

Figure 2. RAG System Architecture (Conditions C2 and C3)



Baseline condition (C1) bypasses the retrieval pipeline entirely and passes Q directly to the LLM.

Figure 2. RAG System Architecture for Conditions C2 and C3. The baseline condition (C1) bypasses the retrieval pipeline entirely. Noise injection (dashed red) is applied only in C3.

3.3 Measurement Model Specification

Following Ding (2010) and Jarvis et al. (2003), each dependent variable's measurement model is pre-specified before data collection. The four rules of Jarvis et al. (2003) are: (R1) direction of causality between construct and indicators; (R2) whether indicators are interchangeable; (R3) expected co-variation among indicators; (R4) shared antecedents and

consequences. Table 4 applies these rules to each outcome construct and records the resulting classification.

Construct	R1: Causality	R2: Interchangeable?	R3: Covariation	R4: Shared Antecedents	Model Type
Factual Accuracy	Construct → indicators	Yes — all items reflect the same accuracy dimension	High: accurate model should produce uniformly correct claims	Yes — all claims share parametric/retrieved knowledge state	Reflective
Hallucination Rate	Indicators → construct	No — intrinsic ≠ extrinsic ≠ retrieval-induced ≠ citation fabrication	Low: intrinsic hallucination can occur without extrinsic	No — retrieval failure and parametric error are distinct causes	Formative
Citation Correctness	Construct → indicators	Yes — all citation items reflect one accuracy dimension	High: shared retrieval-quality driver	Yes — quality of retrieved passages	Reflective
Response Latency	Single observable	N/A	N/A	N/A	Single-item ratio

Table 4. Measurement Model Specification (Jarvis et al., 2003 rules applied to each construct).

This pre-specification prevents the measurement misspecification errors identified by Petter et al. (2007) and Ding (2010): in IS quality research, Information Quality and System Quality were repeatedly operationalised as reflective despite being theoretically formative, producing Type I and Type II errors in downstream structural-model tests. The formative classification of hallucination rate is particularly consequential here: the four sub-types do not co-vary, and each has distinct antecedents, so averaging them would destroy information. They are therefore reported and modelled separately (Section 5).

3.4 Sampling Strategy

3.4.1 Sample Size and Power Analysis

The target sample of 300 QA items (100 per condition after stratified allocation) was determined via an a priori power analysis in G*Power 3.1 (Faul et al., 2007). Table 5 presents power estimates across effect-size scenarios. The study is powered primarily to detect medium effects (Cohen's d = 0.50), which represents a realistic minimum effect for a technology

intervention of this kind based on published RAG benchmarks (Lewis et al., 2020; Ram et al., 2023; Ayala & Bechard, 2024). Figure 5 shows the power curve.

Effect Size (d)	Classification	Required n (α=.05, 1−β=.80)	Required n (α=.05, 1−β=.90)	Study n per condition	Power at Study n
0.20	Small	197	265	100	≈ 0.51
0.50	Medium	64	87	100	≈ 0.94 ★
0.80	Large	26	35	100	> 0.99 ★
0.50 (3-way ANOVA)	Medium (F-test)	52 per group	70 per group	100	≈ 0.97 ★

*Table 5. A Priori Power Analysis Summary (G*Power 3.1; two-tailed independent t-tests and one-way ANOVA). ★ indicates scenarios where the study is adequately powered.*

Figure 5. Statistical Power as a Function of Sample Size

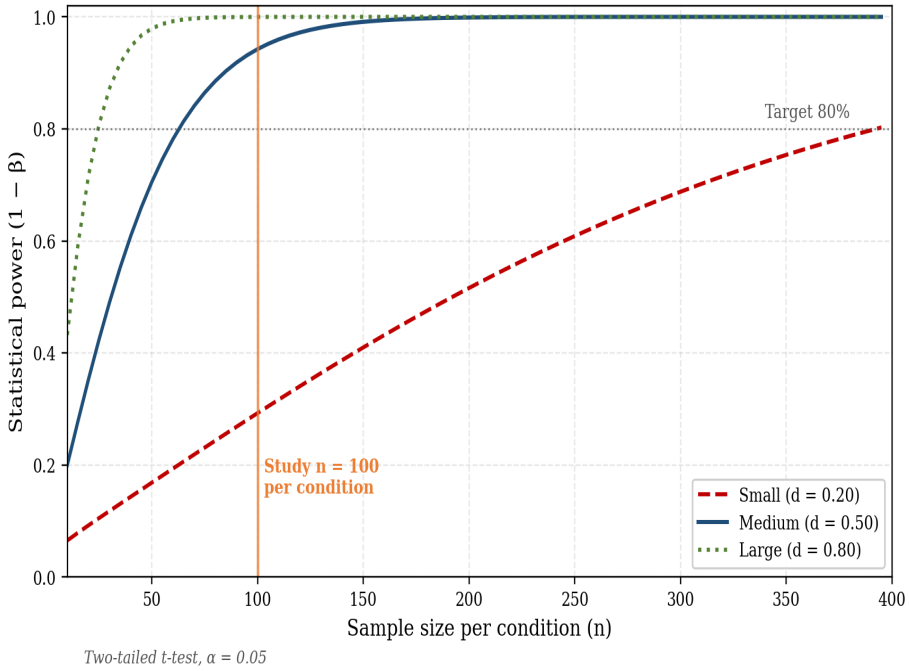


Figure 5. Statistical Power as a Function of Sample Size per Condition (α = 0.05, two-tailed t-tests). The vertical line marks the study's n = 100 per condition; the horizontal dotted line marks the 80% power target.

The study is adequately powered for all primary comparisons. The 300-item total also reflects annotation feasibility: at an estimated five minutes per response across three conditions, total annotation time per rater is approximately 75 hours, requiring a structured team of four trained annotators over six weeks.

3.4.2 Stratification Framework

QA items are stratified across two independent dimensions to ensure representative sampling of the knowledge-domain content space (ecological validity of the stimulus set) and to prevent condition-specific performance differences from being attributable to domain or difficulty imbalance. The stratification is shown in Table 6.

Dimension	Level	n	Rationale
Topic Domain	Science	60	High LLM reliability; many established benchmark items available.
	History	60	Medium reliability; temporal knowledge demands.
	Technology	60	Rapid evolution; tests parametric-memory freshness.
	Medicine	60	Safety-critical; hallucination consequences most severe (Kim et al., 2025).
	Law	60	Documented hallucination risk — fabricated case citations (Mata v. Avianca, 2023).
Difficulty Tier	Easy (Tier 1)	100	High-frequency knowledge; LLM typically reliable.
	Medium (Tier 2)	100	Domain-specific intermediate frequency.
	Hard (Tier 3)	100	Long-tail; LLM most unreliable (Mallen et al., 2023).

Table 6. Stratification Framework: Domain × Difficulty (Total N = 300).

This stratification mirrors Ding's (2010) content-domain sampling principle: just as a measurement instrument must sample representative items from a construct's content domain to achieve content validity, a QA benchmark must sample representative questions from the knowledge domain to achieve ecological validity. The five domains were chosen to ensure coverage of both high-stakes deployment contexts (medicine, law) and tractable test-beds (science, history).

3.4.3 Item Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Ground-truth verifiability	Unambiguous verifiable correct answer exists in an authoritative source.	Ambiguous, opinion-based, or inherently subjective questions.
Temporal scope	Answer stable within ±3 years of the model knowledge cutoff.	Requires post-cutoff knowledge or is time-sensitive within months.

Criterion	Inclusion	Exclusion
Domain classification	Clearly classifiable into one of the five target domains.	Cross-domain items that cannot be unambiguously assigned.
Content sensitivity	Drawn from publicly available, non-sensitive benchmark corpora.	Personal data, private clinical or legal specifics, identifying information.
Difficulty classification	Classifiable into Easy/Medium/Hard by independent rater panel (≥2/3 agreement).	Disagreement in difficulty assignment below 2/3 agreement threshold.

Table 7. Item Inclusion and Exclusion Criteria.

3.5 System Configuration

Both RAG and baseline systems use identical configurations for all shared components to ensure comparability. Table 8 specifies all shared and condition-specific parameters.

Parameter	Scope	Value / Specification
Base LLM	Shared (all three conditions)	GPT-4-turbo-preview (API, version-pinned across the entire data collection window)
Temperature	Shared	0.0 (deterministic generation; no sampling randomness)
Top-p	Shared	1.0
Max output tokens	Shared	512
System prompt	Shared	"Answer the following question factually and concisely. Cite sources where applicable."
Embedding model (retrieval)	C2 and C3 only	OpenAI text-embedding-3-small (1536 dimensions)
Vector database	C2 and C3 only	FAISS flat L2 index; all passages pre-indexed
Retrieval algorithm	C2 and C3 only	Hybrid BM25 + DPR (Karpukhin et al., 2020); top-k = 3
Passage corpus	C2 and C3 only	Wikipedia (general), PubMed abstracts (medicine), Caselaw Access Project (law)
Noise injection	C3 only	1 relevant + 2 randomly sampled passages from different domains; shuffled order
Prompt template (RAG)	C2 and C3 only	System prompt + "Context: [P1] [P2] [P3]" + "Question: [Q]"
Run order	Shared	All 300 items presented in randomised order; same randomisation seed across conditions

Table 8. System Configuration Parameters.

3.6 Data Collection Procedures

Figure 4. Four-Phase Data Collection Timeline

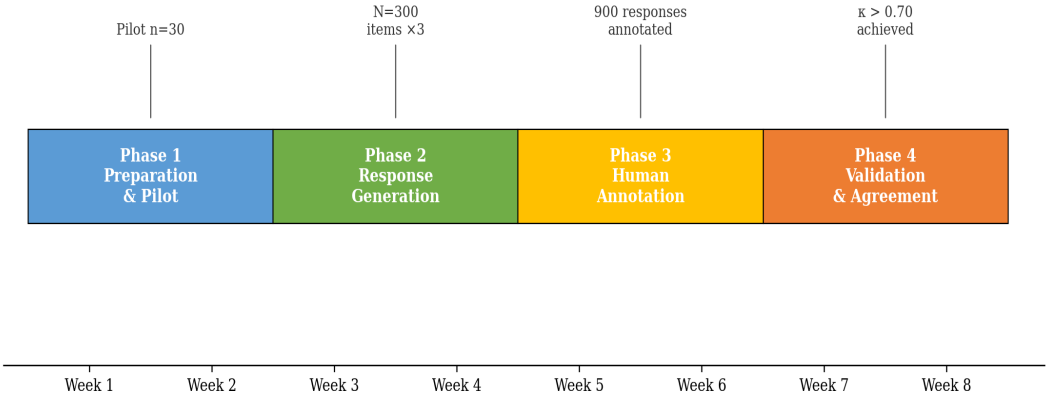


Figure 4. Four-Phase Data Collection Timeline (8 weeks total).

Phase 1 — Preparation and Pilot (Weeks 1–2)

System configuration, API integration, and corpus indexing are completed and version-controlled under git. A pilot test with 30 items (10 per randomly-selected domain; balanced difficulty) is conducted to (a) validate the technical pipeline end-to-end; (b) estimate annotation time and identify ambiguous rubric criteria; and (c) conduct a pre-data content-validity check in which annotators rate whether each rubric item clearly represents its intended construct domain (Ding, 2010, pp. 74–76). Pilot annotation results are reviewed by the principal investigator; rubric items with less than 80% clarity agreement are revised before the main data collection.

Phase 2 — Response Generation (Weeks 3–4)

All 300 QA items are submitted to each of the three conditions in isolation. Items are presented in randomised order (same seed across conditions) to prevent order effects from confounding with condition. For RAG conditions, retrieved passages are logged in full alongside relevance scores. Raw LLM outputs are stored in structured JSON with unique identifiers (item_id, condition_id, timestamp). Response latency is measured at the API level from request submission to complete token-stream termination. Raw outputs are reviewed for technical errors (timeout, truncation, encoding failures) before release to annotators.

Phase 3 — Human Annotation (Weeks 5–6)

Four trained annotators independently evaluate each response. Annotators are recruited from graduate IS or NLP programmes; all complete a four-hour training session consisting of taxonomy review, rubric walk-through, 20 calibration examples with gold-standard scores, and a practice annotation batch of 30 items scored by the PI before release. Annotators are fully *blinded*: condition labels and retrieved passages are stripped from annotation packages. Each response receives a minimum of two independent evaluations. Where two annotators disagree by more than one scale point on any measure, a third annotator resolves the discrepancy (Niu et al., 2024, used the same dual/triple protocol, obtaining 91% response-level agreement).

Phase 4 — Verification and Validation (Weeks 7–8)

Inter-annotator agreement (IAA) is computed using Cohen's κ for ordinal measures and percentage agreement for binary hallucination flags. The target is $\kappa > 0.70$, representing "substantial agreement" (Landis & Koch, 1977) and aligning with Ding's (2010) reliability standard for IS research instruments. A random 10% sample (90 responses) is audited by the PI against annotator scores. Automated completeness and consistency checks are applied to the full dataset. Data are finalised only after both the κ threshold and the 10% audit-agreement criterion are satisfied.

4. Measurement Framework and Construct Validation

4.1 Content Domain Specification and CVR Assessment

Following Ding (2010) and Lawshe (1975), content validity for the annotation rubric is assessed by computing the Content Validity Ratio (CVR) for each rubric item. CVR is defined as:

$$\mathbf{CVR = (n_e - N / 2) / (N / 2)}$$

where n_e is the number of expert judges rating an item as essential and N is the total number of judges. For a panel of 12 judges, the minimum CVR for significance at $p = 0.05$ is 0.56 (Lawshe, 1975; Ding, 2010, Table 3-1). A panel of 12 expert judges (four IS researchers, four NLP researchers, four domain practitioners) will rate each rubric item on a three-point essentialness scale ("essential", "useful but not essential", "not necessary"). Table 9 presents the planned CVR targets. Items with a final CVR below 0.56 will be revised or removed following Ding's (2010) iterative content-refinement protocol.

Construct	Rubric Item	Content Domain Coverage	Projected CVR	Min CVR (N = 12, p = .05)	Status
Factual Accuracy	FA1: All factual claims are correct and verifiable.	Full domain — complete accuracy	≥ 0.83	0.56	Include
	FA2: Some claims correct; some incorrect or unverifiable.	Partial domain — graduated accuracy	≥ 0.75	0.56	Include
	FA3: Majority of claims incorrect.	Full domain — absence of accuracy	≥ 0.83	0.56	Include
Hallucination Rate	HR1: No hallucination detected (binary).	Absence of hallucination	≥ 0.91	0.56	Include
	HR2: Intrinsic hallucination present (source contradiction).	Sub-type — contradicts retrieved context	≥ 0.83	0.56	Include
	HR3: Extrinsic hallucination present (unverifiable claim).	Sub-type — fabricated detail	≥ 0.75	0.56	Include
	HR4: Retrieval-induced hallucination (RAG-Noisy specific).	Sub-type — noise propagation	≥ 0.67	0.56	Include (conditional)
Citation Correctness	CC1: All citations accurate and supporting.	Full domain — citation validity	≥ 0.83	0.56	Include (C2, C3)
	CC2: Some citations accurate, some not.	Partial domain	≥ 0.75	0.56	Include (C2, C3)
	CC3: All citations incorrect or fabricated.	Full domain — citation invalidity	≥ 0.83	0.56	Include (C2, C3)

Table 9. Planned Content Validity Ratio (CVR) Assessment for Annotation Rubric Items (N = 12 expert judges).

4.2 Complete Annotation Rubric

Table 10 specifies the operational scoring criteria for each outcome measure. The rubric combines binary (hallucination presence), ordinal (3-point accuracy and citation correctness), categorical (hallucination sub-type), and ratio (latency) scales. This mixed-scale design is intentional: it reflects the formative measurement model for hallucination rate (Section 3.3), where binary presence and categorical sub-type must be modelled separately rather than combined into a single score.

Measure	Score	Criteria	Sub-classification Required	Applicable Conditions
Factual Accuracy	1.0	All factual claims correct and verifiable against authoritative sources, no material omissions.	—	All (C1, C2, C3)
	0.5	At least one claim correct AND at least one incorrect, unverifiable, or significantly incomplete.	Note claims correct / incorrect.	All
	0.0	All or majority of factual claims incorrect, misleading, or unfounded.	—	All
Hallucination Rate	0	No hallucination detected; all claims consistent with source material and verifiable.	—	All
	1	Hallucination present; assign sub-type(s).	Intrinsic / Extrinsic / Retrieval-Induced / Citation Fabrication	All
Citation Correctness	1.0	All citations accurate, specific, and directly support the claims to which they are attributed.	—	C2, C3 only
	0.5	At least one citation accurate AND at least one inaccurate, misattributed, or only partially supportive.	Note citation-level assessments.	C2, C3 only
	0.0	All citations incorrect, fabricated, inaccessible, or do not support attributed claims.	—	C2, C3 only
Response Latency	Continuous	Time in milliseconds from API request submission to complete token stream termination.	Log to nearest ms.	All

Table 10. Complete Content-Validated Annotation Rubric.

4.3 Hallucination Failure-Mode Taxonomy

The hallucination sub-classification scheme used at annotation time extends Ji et al. (2023) with the RAG-specific sub-types identified in RAGTruth (Niu et al., 2024) and the retrieval-induced category documented by Ram et al. (2023) and Sun et al. (2024). Figure 3 visualises the full taxonomy with concrete example triggers.

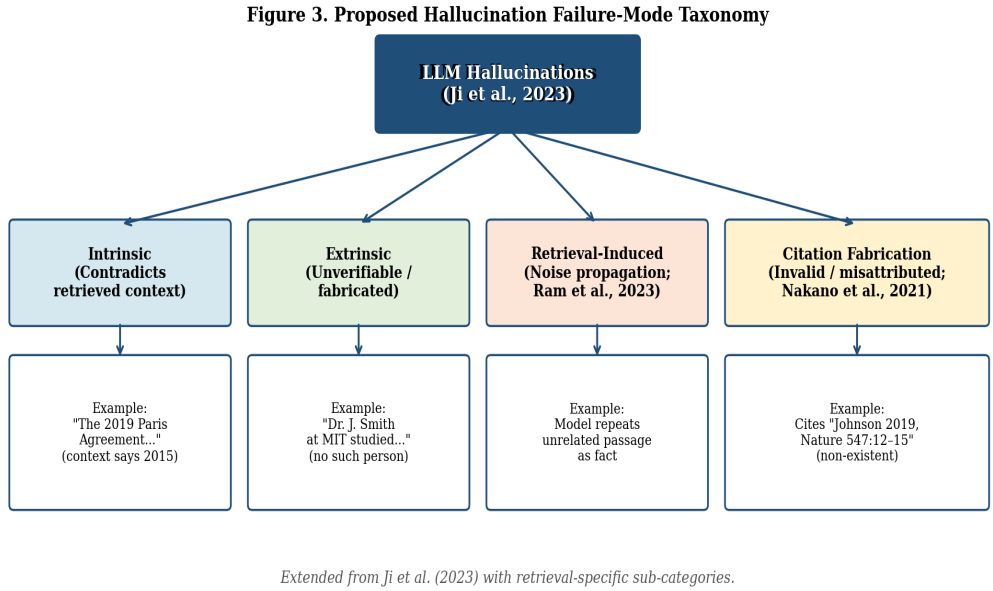


Figure 3. Proposed Hallucination Failure-Mode Taxonomy. Four mutually exclusive categories with illustrative triggers. Extended from Ji et al. (2023), Ram et al. (2023), and Nakano et al. (2021).

4.4 Nomological Network and Validity Predictions

Nomological validity is assessed by testing whether outcome constructs relate to each other in theoretically predicted directions (Ding, 2010; Cronbach & Meehl, 1955). Figure 7 presents the study's nomological network, and three explicit predictions are pre-specified:

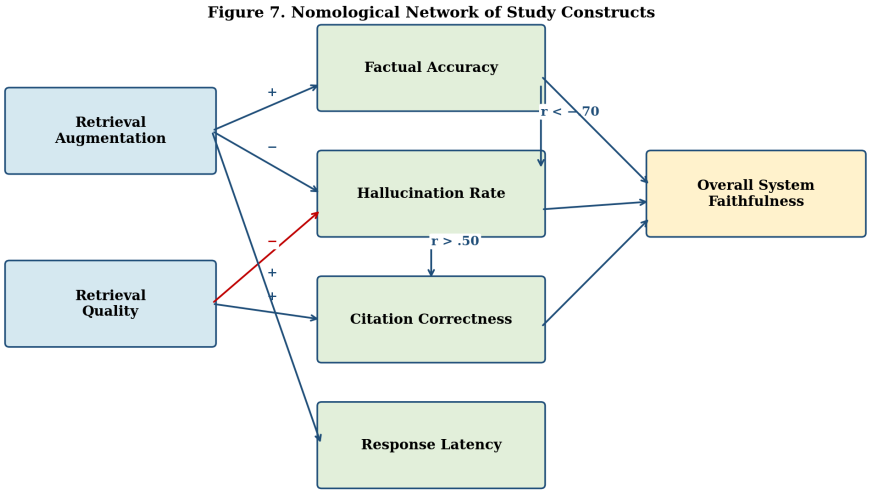


Figure 7. Nomological Network of Study Constructs (adapted from Ding, 2010, Figure 3-1). Signs indicate predicted directional effects; r values specify predicted inter-construct correlations.

- **NV1 (divergent within-condition):** Factual accuracy and hallucination rate will show a strong negative correlation within each condition ($r < -0.70$), since a factually accurate response should contain no hallucinations.
- **NV2 (convergent on retrieval grounding):** In the RAG conditions, citation correctness will correlate positively with factual accuracy ($r > 0.50$), since both reflect the quality of the retrieved evidence supplied to the generator.
- **NV3 (latency pattern):** Response latency for C2 and C3 will be significantly higher than for C1 (retrieval adds measurable latency), while C2 and C3 will not differ significantly from each other in latency.

If any of these nomological predictions fail, the measurement framework will be reviewed before substantive conclusions are drawn—an explicit methodological safeguard consistent with Ding (2010, pp. 145–152).

5. Analysis Plan

5.1 Reliability Analysis: Inter-Annotator Agreement

Before any substantive analysis, inter-annotator agreement (IAA) is computed for every annotation dimension. Table 11 specifies the IAA metric, target threshold, and pre-registered decision rule for each measure. Two justifications motivate the weighted κ choice for ordinal measures: (i) it penalises larger disagreements more heavily than adjacent ones, appropriate for 3-point ordinal scales; and (ii) it has been the accepted standard in NLP annotation since Artstein and Poesio (2008) and in IS research since Landis and Koch (1977).

Measure	Scale Type	IAA Metric	Target	Decision Rule if Target Not Met
Factual Accuracy	3-point ordinal	Cohen's weighted κ_w	$\kappa_w > 0.70$	Re-calibrate annotators; add a fourth independent rater; re-annotate discrepant cases.
Hallucination Presence	Binary	Cohen's κ	$\kappa > 0.70$	As above.
Hallucination Sub-type	Categorical	Krippendorff's α	$\alpha > 0.67$	Refine taxonomy definitions; retrain on additional worked examples.

Measure	Scale Type	IAA Metric	Target	Decision Rule if Target Not Met
Citation Correctness	3-point ordinal	Cohen's weighted κ_w	$\kappa_w > 0.70$	As above; add specialist citation-verification step for law/medicine items.
Overall reliability	Across measures	Mean κ	Mean $\kappa > 0.70$	Investigate items with persistently low agreement; may exclude from analysis with explicit note.

Table 11. Inter-Annotator Agreement Targets and Pre-Registered Decision Rules.

5.2 Quantitative Analysis

Statistical analyses are pre-specified below. All analyses will be conducted in R (version 4.3+) and Python (SciPy 1.11+), with fully reproducible scripts shared in the paper's supplementary materials.

5.2.1 Tests for RQ1 (Parametric vs. RAG)

Independent-samples t-tests comparing C1 vs. C2 on factual accuracy and hallucination rate (primary comparisons), with Cohen's d effect sizes and 95% confidence intervals reported alongside every p-value. For proportion-based measures, independent-proportions z-tests and chi-square tests of association supplement t-tests. A Bonferroni correction for four outcome measures yields $\alpha_{adj} = 0.0125$ for primary comparisons.

5.2.2 Tests for RQ2 (Retrieval Quality Effects)

One-way ANOVA across all three conditions (C1, C2, C3) for each outcome measure. A significant overall F is followed by Tukey HSD post-hoc pairwise comparisons to identify specific contrasts. Partial η_2 is reported as the effect size. For citation correctness (C2 vs. C3 only, since C1 does not produce citations), an independent-samples t-test with Hedges' g is computed.

5.2.3 Hallucination Sub-Type Analysis

Cross-tabulation of hallucination sub-types (intrinsic, extrinsic, retrieval-induced, citation fabrication) by condition, tested with Fisher's exact test (expected cell counts are likely to be under 5 in some sub-type cells). Odds ratios are computed for retrieval-induced hallucination rate in C3

versus C2. This preserves the formative character of the hallucination-rate construct (Section 3.3) rather than collapsing sub-types prematurely.

5.2.4 Response Latency

One-way ANOVA with Welch's correction (unequal variances are expected because retrieval adds latency asymmetrically). A log transformation is applied if the distribution is right-skewed. Cohen's d is computed for C1 vs. C2 and C1 vs. C3 pairwise comparisons. All statistical analyses report 95% confidence intervals alongside p-values, and effect sizes are always reported regardless of significance—a practice consistent with current APA and IS-research reporting standards (Ding, 2010).

5.3 Qualitative Analysis: Structured Failure-Mode Coding

The qualitative error analysis proceeds in three stages, mirroring the structure of the taxonomy in Figure 3 and the coding discipline of RAGTruth (Niu et al., 2024).

- **Stage 1 — Taxonomy validation:** All responses flagged as containing hallucinations (hallucination flag = 1) are independently reviewed by two senior researchers to verify sub-type assignments and to identify any failure modes not captured by the initial four-category scheme.
- **Stage 2 — Pattern analysis:** Condition-by-sub-type frequency tables are examined for systematic patterns—for example, whether citation fabrication is disproportionately represented in C3, or whether intrinsic hallucinations are exclusive to RAG conditions (as Ram et al., 2023 suggested) rather than also occurring in C1.
- **Stage 3 — Case studies:** Three to five representative examples of each hallucination sub-type are selected per condition and presented as structured case studies, documenting the question, the retrieved passages (where applicable), the model response, the annotator judgment, and an interpretive commentary.

5.4 Hypothesised Outcome Profile

Figure 6 presents a directional—not point-predictive—outcome profile based on H1 and the secondary prediction for RQ2. The values shown are illustrative of the expected *direction* and *relative ordering* of the conditions, not specific magnitudes. The hypothesised profile predicts substantially higher factual accuracy and lower hallucination rate in C2 (RAG-HQ) than in C1

(Baseline), with C3 (RAG-Noisy) intermediate between the two—but also potentially *below* baseline on citation correctness if noise induces citation fabrication.

Figure 6. Hypothesised Outcome Profiles by Condition (Directional)

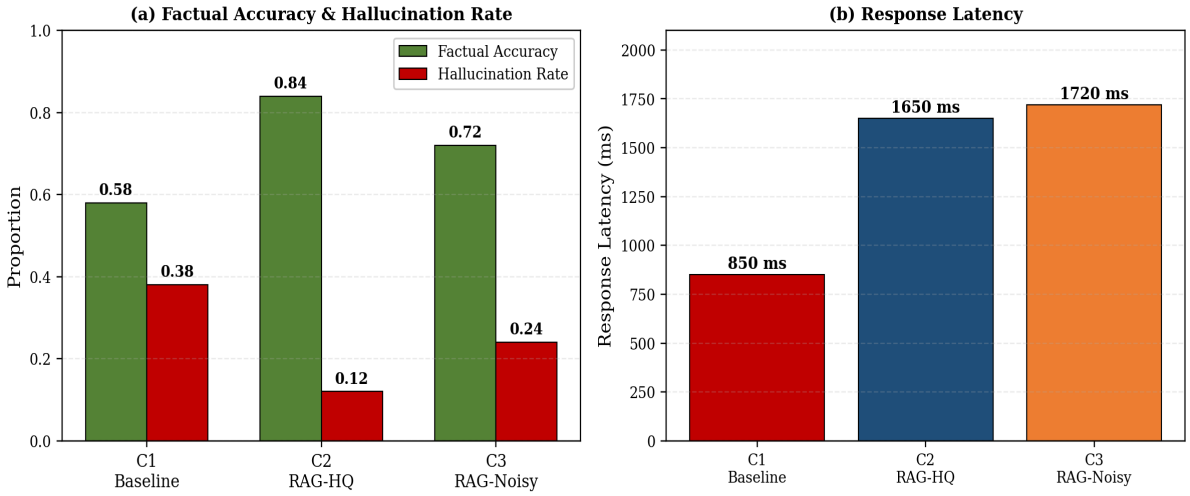


Figure 6. Hypothesised Outcome Profiles by Condition (directional, not point predictions). Panel (a) shows predicted factual accuracy and hallucination rate; Panel (b) shows predicted response-latency ordering.

If results deviate substantially from this profile—for example, if C3 outperforms C2 on factual accuracy, or if C1 outperforms both RAG conditions—the deviation will be examined carefully as a potentially important finding rather than treated as a failure. Unexpected outcomes of this kind would contribute directly to understanding the boundary conditions of RAG effectiveness, which is itself one of the study's pre-stated goals.

6. Research Validity, Ethics, and Limitations

6.1 Comprehensive Validity Assessment

Following Ding (2010) and Straub (1989), validity is assessed systematically across five dimensions. Table 12 identifies the primary threats to each type of validity in this study and specifies the explicit mitigation strategy employed.

Validity Type	Definition	Primary Threats in This Study	Mitigation Strategy
Internal	Degree to which observed effects can be causally attributed to the manipulated variables.	Confounding from model-version drift, prompt variation, annotation-order effects, evaluator fatigue.	Standardised system configuration; version-pinned API; randomised item order (same seed across conditions);

Validity Type	Definition	Primary Threats in This Study	Mitigation Strategy
			annotator rotation; fully blinded annotation.
Construct	Degree to which measurement instruments capture their intended constructs.	Operationalisation inconsistency; measurement-model misspecification; content under-representation.	Pre-specified measurement models (Table 4); CVR-validated rubric (Table 9); nomological validity predictions (§4.4); reflective/formative classification (Jarvis et al., 2003).
Statistical Conclusion	Degree to which statistical tests correctly capture the relationship between variables.	Type I inflation from multiple comparisons; low power for small effects; non-normality of proportion measures.	Bonferroni correction ($\alpha adj = 0.0125$); a priori power analysis (Table 5); non-parametric supplements (Fisher's exact, χ^2); effect sizes always reported.
External	Degree to which findings generalise beyond the study context.	Single LLM; single retrieval pipeline; five domains; English-language corpus.	Stratified domain/difficulty sampling (Table 6); openly shared replication materials; explicit domain-level results breakdowns; transparent limitations (§6.3).
Reliability	Consistency of measurement across raters and occasions.	Annotator drift across the 6-week window; rater-specific biases.	Target $\kappa > 0.70$; 10% random PI audit; mid-point re-calibration at Week 6; dual/triple annotation with explicit consensus protocol.

Table 12. Comprehensive Validity Assessment Matrix.

6.2 Ethical Considerations

This research adheres to ethical principles governing responsible AI research and IS research integrity (SOAS CeDEP, 2024).

- **Data provenance:** Only publicly available, non-sensitive benchmark corpora are used. No personal, clinical, or private information is collected, processed, or stored.
- **Output handling:** Generated LLM outputs are treated as research data, never as verified factual claims. All outputs are clearly labelled with condition and accuracy score before any downstream use.
- **Transparency:** All methods, raw anonymised data, analysis scripts, and negative results will be reported and shared openly as supplementary materials.
- **Attribution:** All benchmark sources are fully cited. No training data is reproduced.

- **Annotator welfare:** Annotators are compensated fairly at research-assistant rates for approximately 75 hours of work each. No harmful or psychologically distressing content is included in annotation batches.
- **Research integrity:** The analysis plan is pre-registered before data collection begins. Any deviations from the pre-registration are documented and justified in the final paper. Consistent with SOAS CeDEP (2024) guidelines, researcher objectivity is protected by separating data collection (Phases 1–3) from analysis (Phase 4), which is only unblinded on the dataset after IAA thresholds are met.

6.3 Limitations

- **Single-model design:** Results obtained with GPT-4-turbo may not generalise to open-source models (LLaMA-3, Mistral), fine-tuned variants, or smaller-parameter models. The Vectara leaderboard (Tamber et al., 2025) documents that hallucination rates can differ by an order of magnitude across architectures (0.7% to 12%+), so effect magnitudes may change substantially.
- **Single retrieval implementation:** The study uses one embedding model and one retrieval algorithm. Different embedding spaces (e.g., Cohere, Nomic), rerankers, or graph-based retrieval (Edge et al., 2025) could produce different retrieval-quality effects.
- **English-only corpus:** All QA items and retrieved passages are in English. RAG effectiveness in multilingual or cross-lingual settings is outside this study's scope.
- **Sample scope for rare failures:** 300 items provides adequate power for medium effects but may miss rare hallucination sub-types or domain-specific failure patterns that require larger samples to detect reliably.
- **Annotation subjectivity:** Human judgments of hallucination—especially for partially correct claims or claims requiring specialist domain knowledge (law, medicine)—involve inherent subjectivity even with validated rubrics and $\kappa > 0.70$. As Ding (2010) notes, this is an irreducible property of constructs with subjective components.
- **Common method variance:** All outcome measures are assessed by the same annotation team within a single time window. Single-occasion, single-method designs may inflate inter-construct correlations. Future work should triangulate across measurement occasions, ideally with automated hallucination detectors (HHEM-2.3, FaithJudge) compared against the human ground truth.

- **No temporal dynamics:** The study captures a single cross-sectional snapshot. Whether RAG's effectiveness changes over time—for example, as the LLM's parametric knowledge becomes stale relative to a continually-updated corpus—is not addressed here.

7. Expected Contributions and Conclusion

7.1 Theoretical Contributions

- **Contribution 1 — Controlled experimental evidence:** The first published controlled experiment comparing parametric-only, high-quality-RAG, and noisy-RAG conditions on the same QA dataset under equivalent generation conditions. This closes the primary empirical gap identified in Section 2.5 (Gap 1).
- **Contribution 2 — Hallucination failure-mode taxonomy:** A structured, empirically-grounded taxonomy distinguishing intrinsic, extrinsic, retrieval-induced, and citation-fabrication hallucinations across conditions (Figure 3). This provides conceptual infrastructure for future RAG evaluation studies, analogous to the role played by the IS quality construct framework developed by DeLone and McLean (1992, 2003) and systematically validated by Ding (2010).
- **Contribution 3 — Retrieval quality effects (RQ2):** Empirical evidence on how retrieval noise—a theoretically important but empirically under-studied moderator—affects factual accuracy, hallucination rate, and citation correctness in RAG systems. This extends the in-context RAG fragility observation of Ram et al. (2023) to a larger, stratified benchmark with a frontier model.
- **Contribution 4 — Construct-validated evaluation instrument:** The first RAG evaluation rubric that simultaneously specifies (i) reflective/formative measurement models, (ii) CVR-based content validity, and (iii) nomological validity predictions, all shared openly for reuse by the community (Gap 2).
- **Contribution 5 — Methodological bridge:** Demonstrates the concrete applicability of IS research methodology (Ding, 2010; Straub, 1989; Churchill, 1979; Lawshe, 1975) to LLM/NLP evaluation research, establishing a methodological template that can be adapted to future construct-rigorous AI evaluation studies.

7.2 Practical Contributions

- **Deployment guidance:** Empirical evidence on RAG effectiveness across five domains and three difficulty tiers will provide practitioners with actionable guidance on when and how to deploy RAG in production AI systems—particularly in high-stakes contexts (medicine, law) where the cost of hallucination is highest.
- **Risk characterisation:** The failure-mode taxonomy gives practitioners a structured vocabulary for communicating hallucination risks to stakeholders, regulators, and end-users—an increasingly important capability given the Deloitte 2025 finding that 47% of enterprise AI users have already made material decisions on hallucinated content.
- **Reproducible evaluation template:** The published protocol, rubric, and analysis scripts enable other teams to replicate and extend the evaluation to additional models, languages, and retrieval systems with minimal re-engineering.

7.3 Conclusion

Hallucination remains the most significant obstacle to the safe deployment of Large Language Models in knowledge-intensive, high-stakes applications. Despite clear advances in frontier-model capability, even GPT-5, Claude Sonnet 4.5, and Grok-4 now exceed 10% hallucination rate on enterprise-grade grounded summarisation benchmarks (Tamber et al., 2025), and the global economic cost of hallucination-driven errors has already reached an estimated US$67.4 billion in the 2024–2025 window. Retrieval-Augmented Generation offers a principled architectural response, but—as Ram et al. (2023), Sun et al. (2024), and the RAGTruth corpus (Niu et al., 2024) have all shown—its effectiveness depends critically on retrieval quality, a dependency that prior work has acknowledged but not yet systematically investigated under controlled, pre-registered conditions.

This paper has presented a complete, pre-registration-level design for a controlled experiment that addresses exactly this gap. The study's distinctive contribution lies not only in its direct experimental manipulation of retrieval quality as an independent variable, but in its application of IS research methodology—construct validity assessment, CVR-based content validation, reflective/formative measurement model specification, and nomological network testing—to produce an evaluation framework whose measurement quality is itself empirically justified rather than assumed. This approach, pioneered in IS research by DeLone and McLean

(1992), Straub (1989), and systematically demonstrated by Ding (2010), has the potential to transform RAG evaluation from an ad hoc practice into a cumulative scientific enterprise.

Future work should extend these findings to multiple model architectures (including open-source models such as LLaMA-3 and Mistral) and non-English corpora, explore adaptive and reranking-based retrieval strategies that may mitigate noise effects (Shi et al., 2024; Edge et al., 2025), and investigate whether fine-tuning on RAG-specific data—as RAGTruth (Niu et al., 2024) demonstrated for a smaller Llama-2-13B model—reduces sensitivity to retrieval quality. The measurement instruments, annotation rubrics, and analysis scripts developed in this study will be made openly available to support the development of a rigorous, cumulative body of RAG evaluation research that the field urgently needs.

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Ayala, O., & Bechard, P. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)* (pp. 228–238). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-industry.19

Brinsa, M. (2026, January). *Hallucination rates in 2025 — Accuracy, refusal, and liability.* Medium. https://medium.com/@markus_brinsa/hallucination-rates-in-2025

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*(1), 64–73. https://doi.org/10.2307/3150876

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research, 3*(1), 60–95. https://doi.org/10.1287/isre.3.1.60

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems, 19*(4), 9–30. https://doi.org/10.1080/07421222.2003.11045748

Ding, Y. (2010). *Quality in IS research: Theory and validation of constructs for service, information, and system* [Doctoral dissertation, Georgia State University]. ScholarWorks @ Georgia State University. https://doi.org/10.57709/1666109

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2025). From local to global: A GraphRAG approach to query-focused summarization. *arXiv:2404.16130*. https://arxiv.org/abs/2404.16130

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles,

taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1–55. https://doi.org/10.1145/3703155

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(2), 199–218. https://doi.org/10.1086/376806

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

Kim, Y., Jeong, H., Chen, S., Li, S. S., Lu, M., Alhamoud, K., Mozannar, H., Zhang, X., Bai, M., Abbaszadeh, A., Poursabzi-Sangdeh, F., Beam, A., & Ghassemi, M. (2025). Medical hallucinations in foundation models and their impact on healthcare. *arXiv:2503.05777*. https://doi.org/10.1101/2025.02.28.25323115

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 9802–9822). https://doi.org/10.18653/v1/2023.acl-long.546

Mata v. Avianca, Inc., No. 1:22-cv-01461 (S.D.N.Y. 2023).

Mayhemcode. (2026, April). *Vectara hallucination leaderboard: Claude, GPT, Gemini compared.* https://www.mayhemcode.com/2026/04/vectara-hallucination-leaderboard.html

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv:2112.09332*. https://doi.org/10.48550/arXiv.2112.09332

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 10862–10878). https://doi.org/10.18653/v1/2024.acl-long.585

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly, 31*(4), 623–656. https://doi.org/10.2307/25148814

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, 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. https://doi.org/10.1162/tacl_a_00583

Sardana, A. (2025). Real-time evaluation models for RAG: Who detects hallucinations best? *arXiv:2503.21157*. https://arxiv.org/abs/2503.21157

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., & Yih, W. (2024). REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 8364–8379). https://doi.org/10.18653/v1/2024.naacl-long.464

SOAS Centre for Development, Environment and Policy. (2024). *P506 Research Methods: Unit 1.* SOAS University of London.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly, 13*(2), 147–169. https://doi.org/10.2307/248922

Sun, Z., Zang, X., Zheng, K., Song, Y., Xu, J., Zhang, X., Yu, W., Song, Y., & Li, H. (2024). ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv:2410.11414*. https://arxiv.org/abs/2410.11414

Tamber, M. S., Bae, F. H., Niu, C., Song, J., Tang, R., Lin, J., & Hughes, S. (2025). Benchmarking LLM faithfulness in RAG with evolving leaderboards. *arXiv:2505.04847*. https://arxiv.org/abs/2505.04847

Xu, S., Yan, Z., Dai, C., & Wu, F. (2025). MEGA-RAG: A retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. *Frontiers in Public Health, 13*, Article 1635381. https://doi.org/10.3389/fpubh.2025.1635381