© SEP 2025 | IRE Journals | Volume 9 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I3-1722951

Hallucination-Aware and Trustworthy LLMS for High-Stakes Supply Chain Decision Support

SOHAIL SAYED[1], NAUMAN SAYED[2]
[1,2]*California State University Fullerton*

Abstract- ***Large language models are prone to hallucination, generating plausible yet nonfactual content — a phenomenon that raises significant concerns over their reliability in real-world systems (Huang et al., 2024). In supply chains, such hallucinations can lead to erroneous demand forecasts or misinterpretation of supply chain relationships, potentially resulting in operational disruptions and financial losses: a generative model might incorrectly predict a demand surge based on fabricated trends, leading to overproduction and increased inventory costs (Ge & Brintrup, 2024). Yet the deployment pressure is real: LLMs are already used for supplier risk assessment, procurement contracting, inventory optimization, and disruption response, where one unsupported number can cascade through a network. This paper argues that trustworthiness in high-stakes supply chain decision support is an engineered property, not an emergent one: it requires hallucination awareness (measurement and detection), grounding (retrieval and knowledge-graph anchoring), verification (self-consistency, semantic entropy, conformal guarantees), and governance (guardrails, abstention, and human-final authority). We propose HALO-SC, a hallucination-aware framework coupling (i) request triage with risk-tier classification, (ii) retrieval and knowledge-graph grounding, (iii) generation with uncertainty quantification via self-consistency and semantic entropy, (iv) multi-signal verification including LLM-as-a-judge with known-bias compensation, (v) conformal abstention with bounded hallucination rate, (vi) deterministic guardrail enforcement with GO/HOLD/NO-GO decision states, and (vii) audit-ready decision lineage. The framework consolidates the reported evidence envelope: retrieval-augmented pipelines reduce hallucination rates to 1.5% on average versus 8.3% for non-RAG baselines — an 81.9% relative reduction — with retrieval precision above 91% across product domains (Kasarapu, 2026); hallucination rate falls from 6.8% at 10,000 documents to 1.2% at 100,000 documents, establishing knowledge-base enrichment as the dominant reliability lever (Kasarapu, 2026); conformal abstention procedures bound the hallucination rate at a user-specified error level with rigorous theoretical guarantees (Yadkori et al., 2024); semantic entropy detects confabulations in free-form generation (Farquhar et al., 2024); governed LLM-based optimization reduces unsafe decision outputs by 45% while sustaining drift-detection accuracy above 92% (Jingar, 2023); decision authority frameworks formalize GO/HOLD/NO-GO governance states with human-final authority enforcement (KALAFATOGLU, 2025, 2026); and human–AI collaboration in which humans specify the optimization algorithm achieves statistically optimal and stable outcomes under supply chain disruption (Wu, 2026). The paper argues that the trustworthiness ladder — grounded, measured, verified, abstaining, governed — is what separates LLM assistants from LLM advisors in commerce.***

Keywords — Hallucination, large language models, trustworthy AI, uncertainty quantification, conformal prediction, semantic entropy, supply chain decision support, guardrails, abstention, human-in-the-loop, knowledge graphs.

I. INTRODUCTION

LLMs have matched or surpassed human performance on an increasingly complex and diverse set of tasks, yet their reliability remains in doubt: they often confidently hallucinate facts that do not exist, and this misalignment between user goals and model behavior hinders deployment exactly in settings where the potential for AI assistance appears highest — legal work, healthcare, finance, and, increasingly, supply chain operations (Cherian et al., 2024). Hallucination is defined as the phenomenon where LLM agents generate text that deviates from reality, arising from overgeneralization of training data or erroneous interpretations of incomplete or misleading information, and is categorized into intrinsic hallucinations (contradicting input logic) and extrinsic hallucinations (containing information unverifiable against existing knowledge) (Li et al., 2024).

Supply chains are a uniquely unforgiving deployment domain for this failure mode. The consequence of a

hallucinated demand forecast is overproduction and inventory cost; the consequence of a hallucinated supplier relationship is a risk assessment that misses a tier-2 exposure; the consequence of a hallucinated recovery cost is a financial commitment made on a fabricated number (Ge & Brintrup, 2024). Because supply chain decisions propagate through multi-tier networks, a single ungrounded claim can cascade into operational disruption and financial loss at enterprise scale. The stakes are compounded by the fact that LLM-based reasoning is increasingly chained — monitoring agents, analyst agents, and planning agents share outputs, and a single agent's hallucination can trigger a cascading effect where misinformation is accepted and propagated by others across the network (Wang et al., 2025).

This paper is the eleventh in a research program on resilient, AI-driven commerce. Papers 1–5 built the signal, cascade, weather, and food-security intelligence layers; Paper 6 connected the platform via failure prediction and self-healing; Paper 7 established retrieval-grounded real-time intelligence; Paper 8 formalized multi-agent autonomy with consensus and governance; Paper 9 made risk intelligence explainable via knowledge graphs; Paper 10 formalized LLM-based incident diagnosis and automated response. This paper completes the trust thread: it takes the grounding of Paper 7, the autonomy of Paper 8, the explainability of Paper 9, and the incident discipline of Paper 10, and adds the capability every prior layer silently assumed — the ability to know when the model does not know.

The contributions of this paper are:

- A formalization of trustworthy supply chain decision support as a hallucination-aware pipeline with explicit grounding, calibration, abstention, and governance constraints, and a decision-theoretic model of hallucination cost (Sections III–IV).
- HALO-SC, a framework coupling risk-tiered triage, retrieval and KG grounding, uncertainty-quantified generation, multi-signal verification, conformal abstention, deterministic guardrails, and audit lineage.
- A consolidated analysis of reported evidence — hallucination rates, grounding and scale effects, unsafe-output governance, decoupling dividends, UQ methods, and evaluation — with worked calculations and clearly flagged illustrative extensions (Sections V–VI).
- A research agenda on calibrated abstention, cascading-hallucination containment, and human-final authority in supply chain AI (Sections VII–VIII).

II. RELATED WORK

A. Hallucination: principles and taxonomy

The modern survey literature is convergent on structure. Hallucinations in the LLM era present distinct challenges that diverge from prior task-specific models, motivating a taxonomy of hallucination, its contributing factors, detection methods and benchmarks, and representative mitigation methodologies — with open questions on knowledge boundaries in LLM hallucinations (Huang et al., 2024). The two-type taxonomy — intrinsic (contradicting input logic) versus extrinsic (unverifiable against existing information) — is the field's standard, and mitigation approaches include integrating external knowledge bases and fact-checking systems to verify generated content, enhancing transparency and interpretability, and fine-tuning with high-quality data or human feedback (Li et al., 2024). In supply chains, generative models inherit the biases and inaccuracies of massive training corpora, and the resulting hallucinations manifest concretely as erroneous demand forecasts and misinterpreted supply chain relationships, with knowledge graphs identified as a structured, verifiable source that language models can reference during generation to keep outputs consistent with established facts (Ge & Brintrup, 2024).

B. Hallucination in multi-agent and decision systems

The move to agentic supply chain systems multiplies hallucination risk. In large model-based agent networks, hallucinations arise from conflicting agents — diverse objectives, strategies, and knowledge bases producing contradictory responses — and from cascading hallucinations, where a single agent's misinformation is accepted and propagated by others across the network, particularly in vertical collaboration paradigms; countermeasures must

therefore include both individual agent-level correction mechanisms and robust inter-agent validation to prevent propagation (Wang et al., 2025). The supply chain consequence is direct: the seven-agent disruption-monitoring pattern that achieves F1 0.962–0.991 and 3.83-minute analyses (AlMahri et al., 2026) and the consensus-seeking negotiation pattern that dampens the bullwhip effect (Jannelli et al., 2025) both depend on inter-agent trust — and inter-agent trust is exactly what cascading hallucination destroys. Systematic evaluation of LLM agent architectures in the adjacent incident-management domain documents the failure mode empirically: hallucinated reasoning paths and inefficient use of context are the dominant causes of failure (Cornacchia et al., 2025).

C. Hallucination in LLM–or and supply chain optimization

The operations research literature has begun to quantify the problem. Systematic evaluation of a reinforcement-learning-trained model across four OR benchmarks — NL4OPT, IndustryOR, EasyLP, and ComplexOR — explicitly notes that prior models such as GPT-4, Claude, and Bard, despite strong NLP performance, suffer from high token costs and a tendency toward hallucination that limits real-world applicability in supply chain contexts; the study contributes a hallucination taxonomy and applies mitigation strategies — LLM-as-a-Judge, few-shot learning, tool calling, and a multi-agent framework — to reduce hallucinations, enhance formulation accuracy, and align outputs with user intent (Zhang, 2025). The democratization agenda for LLM-plus-optimization modeling confirms the same boundary: LLMs already provide substantial novel capabilities for creating formal models, but major research challenges remain before organizational decision-makers can rely on them (Wasserkrug et al., 2025). Human–AI collaboration experiments sharpen the design implication: fully automated direct optimization sacrifices accuracy and stability, two-stage automation remains sensitive to parameter granularity, and simulation-based human-specified optimization consistently achieves statistically optimal and stable outcomes (Wu, 2026) — the strongest evidence that reliability is bought by selective human control, not by autonomy.

D. Detection: uncertainty quantification and semantic entropy

Measuring hallucination is prerequisite to managing it. Sample-based consistency methods — posing the same question multiple times and measuring variance — detect when the model does not actually know an answer, and self-consistency under rephrasing or different sampling seeds flags fluctuating answers on factual questions as hallucination risk (Gumaan, 2025). SELFCHECKGPT established uncertainty-based detection in a zero-resource, black-box setting, combining BERTScore, QA-based metrics, and n-gram metrics for consistency measurement, while verbalized-confidence methods evaluate hallucination rates directly (Zhang et al., 2025). Recent frameworks combine both signals: an observed-consistency score O computed from NLI-based similarity and response equality, and a self-reflection certainty S elicited by prompting the LLM for categorical judgments, fused into a final confidence score $C(x,y) = \beta O + (1-\beta)\, S$ (Kang et al., 2025). Semantic entropy goes further: defining entropy over the meanings of sampled sentences rather than their surface forms, it detects confabulations in free-form generation — the setting where naive token-level approaches fail — enabling systems to avoid answering likely-to-confabulate questions, alert users to unreliability, or trigger grounded search (Farquhar et al., 2024). Comprehensive surveys now organize these methods into taxonomies spanning token-level, sample-based, and verbalized uncertainty, with open challenges for embodied and multi-step decision settings (Shorinwa et al., 2025).

E. Mitigation: grounding, conformal abstention, and verification

Three mitigation families dominate. First, grounding: RAG constrains generation to retrieved evidence; in enterprise retail deployments, a RAG pipeline achieved an overall hallucination rate of 1.5% versus 8.3% for a non-RAG LLM baseline, with retrieval precision above 91% across domains, and — decisively for architecture — hallucination rate fell from 6.8% at 10,000 documents to 1.2% at 100,000 documents, showing that knowledge-based enrichment is the major factor determining generative AI reliability (Kasarapu, 2026). Knowledge-graph grounding adds structured verification, keeping

outputs consistent with established facts through entity-relationship context (Ge & Brintrup, 2024). Second, abstention: conformal prediction — model-agnostic, distribution-free, with strong statistical guarantees — is emerging as the principled mechanism for LLM reliability, transforming predictions into valid sets guaranteed to contain the true outcome with high probability (Campos et al., 2024; Cherian et al., 2024); a conformal abstention procedure uses the LLM's own self-evaluation of similarity between sampled responses plus conformal calibration to rigorously bound the hallucination rate at a user-specified error level while maintaining low abstention rates on long-form answers (Yadkori et al., 2024); and in embodied decision-making, calibrated prediction sets determine when a system should ask for help, minimizing clarification frequency while deviating least from the user-defined error rate (Chakraborty et al., 2025). Third, verification: multi-agent verification patterns — one LLM generating claims, another questioning and checking their truthfulness, or a single LLM self-collaborating across multiple personas — reduce hallucinations at relatively low cost (Zhang et al., 2025). The caveat is equally documented: LLM-as-a-Judge itself suffers from judicial hallucination, combining fact fabrication and cognitive biases, so judge-based evaluation must be treated as an unverified component rather than ground truth (Liu, 2025).

F. Governance: guardrails, decision authority, and audit

Trustworthiness requires institutional scaffolding. The trustworthy-AI review consolidates the requirements set — fairness, explainability, accountability, reliability, and acceptance — together with validation, verification, and standardization strategies (Kaur et al., 2022). At the component level, guardrail-centric fine-tuning aligns a quantized Qwen2.5-Coder-14B model to approximately fifty generalized behavioral guardrails with a deterministic Python enforcement layer, separating probabilistic reasoning from exact execution (Davenport, 2026); governed LLM-based optimization integrating policy constraints, human-in-the-loop oversight, continuous monitoring, and drift-detection agents reduces unsafe decision outputs by 45% and sustains drift-detection accuracy above 92% (Jingar, 2023); and agentic AI in adjacent security domains documents actions in a blockchain ledger for integrity and auditing (Syed et al., 2025). At the system level, decision-authority research formalizes the separation of inference from authority: a deterministic, governance-first framework built on four pillars — deterministic decision boundaries, mandatory human-final authority enforcement, audit-ready decision lineage and replay, and uncertainty-aware hard-stop mechanisms — positions AI strictly as decision support while final authority, responsibility, and accountability remain with human actors (KALAFATOGLU, 2025); the governance-state architecture classifies outputs as GO / HOLD / NO-GO with irreversibility weighting, uncertainty decomposition, time-bound authority gating, and deterministic replay verification at the decision gate (KALAFATOGLU, 2026); and mission-critical enterprise guidance emphasizes systems-engineering design, continuous surveillance, and authentication to keep AI outputs within business goals (USA & Natta, 2024). The supply chain literature adds the interpretability requirement: predictive analytics that cannot be interpreted cannot be integrated into risk-related decision processes (Baryannis et al., 2019), and XAI improves trust and regulatory compliance in risk mitigation (Akubilla et al., 2025; Dehan et al., 2025; Iruku, 2025; R. et al., 2024).

G. Evaluation of trustworthy llm systems

The evaluation layer must itself be trustworthy. OpsEval's 7,334 multiple-choice and 1,736 QA questions across 24 LLMs demonstrate that standard NLP metrics mislead in operational domains, with an Ops-specific evaluation method reaching a 0.9185 agreement coefficient with human experts, improving over BLEU and ROUGE by 0.4471 and 1.366 respectively (Liu et al., 2025); pedagogical critiques show pass@k-style metrics conceal how failures compound across multi-step reasoning and tool use (Chen et al., 2025); and the judicial-hallucination literature quantifies how evaluator bias degrades LLM-as-a-Judge reliability (Liu, 2025). Supply-chain benchmarks mirror the pattern: SupChain-Bench exposes substantial execution-reliability gaps in long-horizon tool orchestration (Guan et al., 2026), CSCBench shows substantial degradation on the Variety axis of commodity reasoning (Cui et al.,

2026), and procurement evaluations show LLM agents faster than humans yet below top performers in accuracy (Brandtner & Hofer, 2026).

H. Research gap

The literature separates measurement (UQ and semantic entropy (Farquhar et al., 2024; Kang et al., 2025), grounding (RAG and KG ((Ge & Brintrup, 2024; Kasarapu, 2026)), abstention (conformal (Yadkori et al., 2024), governance (guardrails and decision authority (Jingar, 2023; KALAFATOGLU, 2025)), and evaluation (Ops benchmarks (Liu et al., 2025). What is missing is a single pipeline for high-stakes supply chain decision support in which triage, grounding, uncertainty-quantified generation, verification, calibrated abstention, and deterministic governance are jointly designed so that every output carries a measured, auditable trust level — and every output that cannot earn one is refused or escalated. This is the integration HALO-SC proposes, and the direction the portfolio's trust thread has been building toward.

III. PROBLEM FORMULATION

A. The trustworthiness requirements set

Following the consolidated trustworthy-AI requirements, a supply chain decision-support output must satisfy correctness (the recommendation is factually supported), faithfulness (the reasoning cites verifiable evidence), calibration (stated confidence matches realized accuracy), safety (the action passes deterministic guardrails), and accountability (the decision lineage is replayable) (KALAFATOGLU, 2025; Kaur et al., 2022).

B. Hallucination model and cost

Let each generated claim $c \in \hat{y}$ be either grounded (verifiable against the knowledge corpus D) or hallucinated. We follow the intrinsic/extrinsic distinction ((Li et al., 2024)): intrinsic claims contradict the input logic; extrinsic claims are unverifiable. Let h\in (Padhy et al., 2026; R. et al., 2024) denote the expected hallucination rate of the pipeline and let C_{err} be the expected cost of an erroneous decision — overproduction, missed tier-2 exposure, or a financial commitment on a fabricated number (Ge & Brintrup, 2024). The ungoverned expected exposure of a stream of N decisions is

$$\mathbb{E}[\text{loss}] = N \cdot h \cdot C_{err},$$

so, reducing h (grounding, verification, abstention) and C_{err} (guardrails, human-final control) are the two levers of trustworthy deployment — the trade-off structure that makes hallucination measurement commercially decisive (Kasarapu, 2026).

C. Abstention problem

The system may abstain at cost $C_{abs} < C_{err}$. Following the conformal-abstention formulation (Yadkori et al., 2024), the objective is

$$\min_{\text{policy}} \mathbb{E}[C_{abs} \cdot \mathbb{1}_{\{\text{abstain}\}} + C_{err} \cdot \mathbb{1}_{\{\text{answer, wrong}\}}] \text{s.t.} \mathbb{P}(\text{wrong} | \text{answer}) \leq \epsilon,$$

where ϵ is the user-specified hallucination-rate bound guaranteed by conformal calibration — the rigorous guarantee that distinguishes principled abstention from heuristic refusal (Yadkori et al., 2024).

D. Uncertainty score

Following the fused-confidence formulation (Kang et al., 2025), the trust score of an output is

$$C(x, y) = \beta O(x, y) + (1 - \beta) S(x, y), 0 \leq \beta \leq 1,$$

where O aggregates sample-based consistency (NLI similarity and response equality across k samples) and S is self-reflection certainty from categorical LLM self-assessment (Kang et al., 2025) — with semantic entropy as the free-form complement for open-ended generation (Farquhar et al., 2024).

E. Decision states and authority

Following the governance-state architecture (KALAFATOGLU, 2026), every output is classified into a decision state: GO (execute autonomously), HOLD (await human confirmation), or NO-GO (blocked), with uncertainty thresholds embedded at the decision gate and human-final authority enforced

© SEP 2025 | IRE Journals | Volume 9 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I3-1722951

for irreversible or high-blast-radius actions (KALAFATOGLU, 2025).

F. Metrics

Hallucination: hallucination rate per document set size (Kasarapu, 2026), intrinsic/extrinsic breakdown (Li et al., 2024). Uncertainty: calibration error, abstention rate at fixed error bound, success-to-clarification ratio (Chakraborty et al., 2025). Grounding: retrieval precision, faithfulness of cited claims (Kasarapu, 2026). Governance: unsafe-output rate, drift-detection accuracy (Jingar, 2023), audit completeness. Decision quality: probability of correct selection and solution quality under human–AI collaboration (Wu, 2026). Evaluation: Ops-domain agreement with human experts (Liu et al., 2025).

IV. PROPOSED FRAMEWORK: HALO-SC

A. Overview

HALO-SC is a hallucination-aware decision-support pipeline with seven layers: triage, grounding, uncertainty-quantified generation, verification, abstention, governance, and learning. It wraps the portfolio's intelligence layers — early warning, cascade analysis, multi-agent coordination, and incident diagnosis — with a measured trust boundary. Fig. 1 shows the architecture.

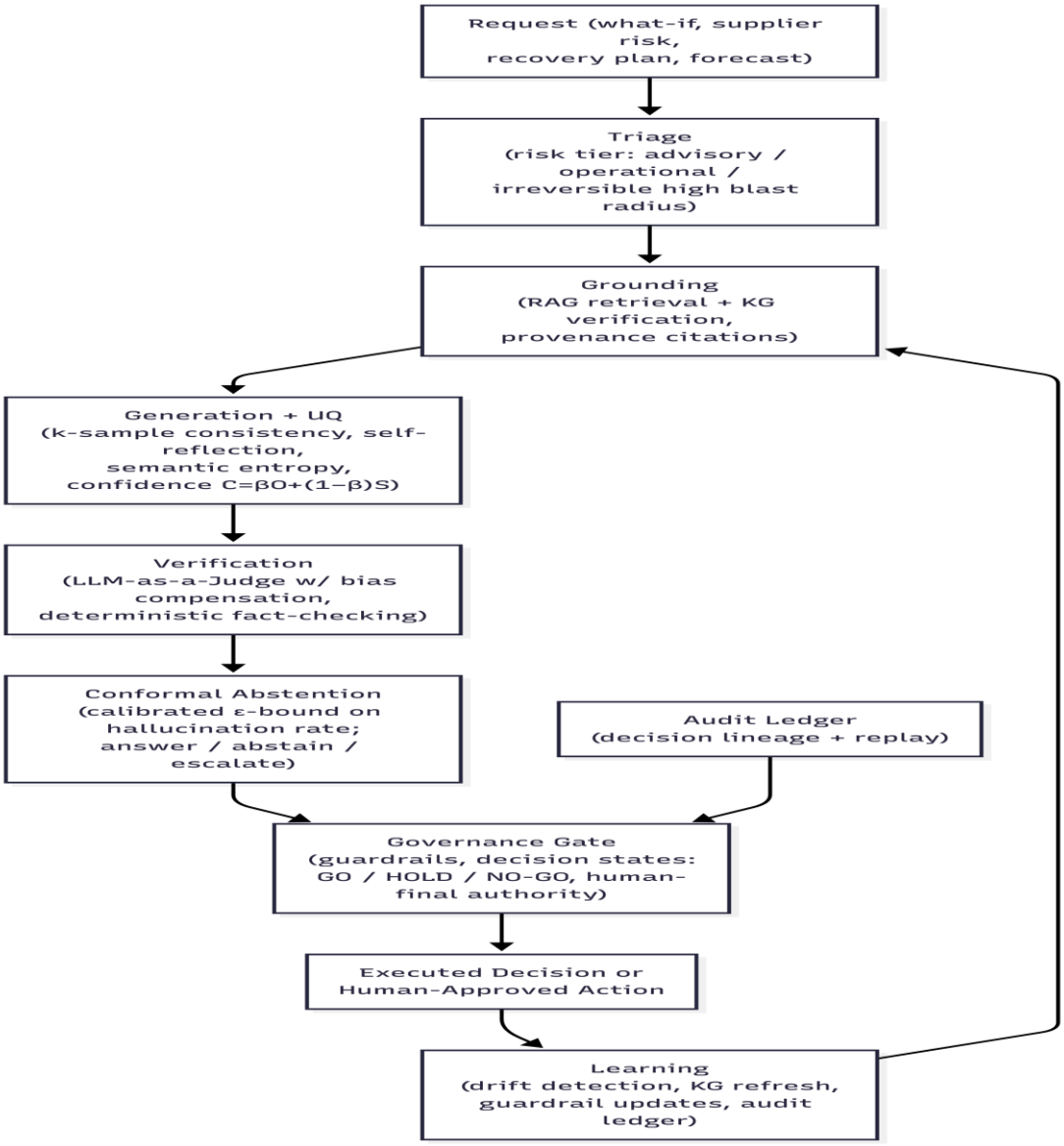

Fig. 1. HALO-SC architecture

Requests are triaged by risk tier; generation is grounded in retrieved evidence and the knowledge graph; every output carries an uncertainty-quantified trust score (sample consistency, self-reflection, semantic entropy); verification applies multi-signal checks including bias-compensated LLM-as-a-judge; conformal abstention enforces a calibrated hallucination-rate bound; a governance gate classifies outputs as GO / HOLD / NO-GO with human-final authority; learning loops refresh grounding and drift detectors; the audit ledger makes every decision lineage replayable (KALAFATOGLU, 2025, 2026; Yadkori et al., 2024).

B. Triage layer

Requests are classified into risk tiers before generation: advisory (analysis, explanation), operational (routine decisions within playbooks), and irreversible (high blast radius: contract commitments, capacity reallocation, supplier switching). Irreversible-tier outputs bypass autonomous execution and require human-final authority by construction — the deterministic-boundary pattern of decision-authority architectures (KALAFATOGLU, 2025).

C. Grounding layer

Generation is constrained to retrieved evidence: RAG over the continuously updated knowledge layer (news, contracts, supplier records, KPIs) with provenance citations, plus knowledge-graph verification that every named entity and relation exists in the graph — the mechanism demonstrated to keep generated outputs consistent with established facts and to reduce hallucination rates from 8.3% to 1.5% in retail-grade deployments (Ge & Brintrup, 2024; Kasarapu, 2026); knowledge-base enrichment toward 100,000+ documents is managed as the dominant reliability lever, the 6.8% → 1.2% scale effect (Kasarapu, 2026).

D. Generation and uncertainty layer

Every output is sampled k times; an observed-consistency score O aggregates NLI-based similarity and response equality, a self-reflection score S elicits categorical self-assessment, and the trust score is the convex combination $C=\beta O+(1-\beta)S$ (Kang et al., 2025); for free-form narrative outputs — risk explanations, recovery reports — semantic entropy over meanings detects confabulation where token-level methods fail (Farquhar et al., 2024). Uncertainty is a first-class output, displayed to users alongside the recommendation.

E. Verification layer

Multi-signal verification checks the output for fact fabrication and reasoning gaps: deterministic fact-checking against the corpus and graph, plus LLM-as-a-judge evaluation with documented-bias compensation — treating the judge as an unverified component per the judicial-hallucination literature rather than as ground truth, and cross-checking judges against deterministic evidence when disagreement appears (Liu, 2025).

F. Abstention layer

Conformal calibration on a held-out set determines the abstention threshold for the user-specified hallucination-rate bound ϵ: outputs with trust scores below the calibrated threshold are refused or routed to human review, with the guarantee that answered outputs err with probability at most ϵ (Yadkori et al., 2024); low-confidence decision states trigger the ask-for-help pattern of calibrated prediction sets, minimizing clarification frequency while respecting the error budget (Chakraborty et al., 2025).

G. Governance layer

Guardrails enforce deterministic safety constraints — the approximately-fifty-behavior pattern of guardrail-centric fine-tuning with a deterministic Python enforcement layer (Davenport, 2026) — and output classification follows the GO / HOLD / NO-GO decision states with uncertainty thresholds and human-final authority at the gate (KALAFATOGLU, 2025, 2026). High blast-radius actions route through the governed-optimization pattern that cut unsafe outputs by 45% with drift detection above 92% (Jingar, 2023); the audit ledger records the full decision lineage — prompt, evidence, samples, trust score, verification results, gate outcome — for replay and regulatory compliance (KALAFATOGLU, 2025; Syed et al., 2025). Algorithm 1 summarizes the loop.

© SEP 2025 | IRE Journals | Volume 9 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I3-1722951

Algorithm 1: HALO-SC Trustworthy Decision Loop

```
Input: request q, knowledge layer (corpus D_t, graph K_t),
       guardrails Gv, calibration set Z, error bound ε
Output: decision state s∈{GO, HOLD, NO-GO}, action a, audit record
# 1. TRIAGE
tier ← classify_risk(q)              # advisory / operational / irreversible
# 2. GROUND
E ← retrieve(D_t, q) ⊕ verify(K_t, q)   # provenance-cited evidence
# 3. GENERATE + MEASURE
{ŷ_i} ← sample_k(q, E)
O ← consistency({ŷ_i}); S ← self_reflect(ŷ)
C ← β·O + (1−β)·S                    # trust score
ent ← semantic_entropy({ŷ_i})        # free-form confabulation signal
# 4. VERIFY
v ← fact_check(ŷ, E, K_t) ⊕ judge(ŷ, E, bias_compensation)
# 5. ABSTAIN / ESCALATE
if C < τ(ε, Z) or v fails:  # τ calibrated via conformal prediction
    if tier == irreversible: escalate to human
    else: abstain; log reason
# 6. GOVERN
s ← gate(ŷ, tier, C, v, Gv)          # GO / HOLD / NO-GO
if s == GO and tier != irreversible: execute a
else: route to human-final authority
# 7. LEARN + AUDIT
update drift detectors; refresh D_t, K_t;
audit.log(q, E, {ŷ_i}, C, v, s, a)   # replayable lineage
```

H. Implementation

Triage: rule-based risk classifier over action ontology; grounding: fine-tuned embedding retrieval plus graph traversal with provenance templates (Ge & Brintrup, 2024; Kasarapu, 2026); UQ: k-sample decoding with NLI-based consistency and self-reflection prompts (Kang et al., 2025), semantic entropy over paraphrase clusters (Farquhar et al., 2024); verification: deterministic fact-checker plus judge ensemble with bias compensation (Liu, 2025); abstention: conformal calibration on a held-out query set (Yadkori et al., 2024); governance: guardrail engine (Davenport, 2026), GO/HOLD/NO-GO gate (KALAFATOGLU, 2026), audit ledger (KALAFATOGLU, 2025); learning: drift monitors (Jingar, 2023) and periodic KG refresh.

V. EXPERIMENTAL SETUP

A. Environments and data

Evaluation is anchored to the documented settings of the component literature: the enterprise retail deployment measuring hallucination rate, retrieval precision, and response relevance across three product domains with expert rubric scoring (Kasarapu, 2026); the document-scale study measuring hallucination rate as a function of corpus size (10,000–100,000 documents) (Kasarapu, 2026); the OR benchmark suite — NL4OPT, IndustryOR, EasyLP, ComplexOR — of the hallucination-taxonomy study (Zhang, 2025); the closed-book and open-domain QA datasets of the conformal-abstention evaluation (Yadkori et al., 2024); the robotic decision tasks of the calibrated prediction-set study (Chakraborty et al., 2025); the OpsEval suites (7,334 MCQ + 1,736 QA) (Liu et al., 2025); and the inventory-optimization settings of the human–AI collaboration study (Wu, 2026).

© SEP 2025 | IRE Journals | Volume 9 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I3-1722951

B. Baselines

Non-RAG LLM prompting (the 8.3% hallucination-rate baseline) (Kasarapu, 2026); small-corpus grounding (the 6.8% rate at 10,000 documents) (Kasarapu, 2026); token-probability-based abstention versus conformal abstention (Yadkori et al., 2024); ungoverned generation (the unsafe-output exposure baseline) (Jingar, 2023); fully automated LLM optimization versus two-stage and human-specified variants (Wu, 2026); single-signal verification versus multi-signal verification (Liu, 2025); and uncalibrated confidence versus conformal calibration (Campos et al., 2024).

C. Scenarios

(S1) supplier-risk query with fabricated entity risk; (S2) demand-forecast query with a hallucinated trend (the overproduction scenario of the supply chain hallucination literature) (Ge & Brintrup, 2024); (S3) supplier-relationship query testing the KG-verification layer (Ge & Brintrup, 2024); (S4) recovery-planning query with fabricated cost figures requiring deterministic fact-check; (S5) irreversible-tier action (supplier switching) testing human-final authority (KALAFATOGLU, 2025); (S6) cascading-hallucination scenario across chained agent outputs (Wang et al., 2025); (S7) adversarial prompts and distribution drift testing guardrails and drift detectors (Jingar, 2023).

D. Protocol

Hallucination rate is measured via expert scoring (the retail rubric) and automated BERTScore/QA-based consistency (Kasarapu, 2026; Zhang et al., 2025); abstention performance is evaluated by error rate at fixed abstention rate (conformal guarantees) and success-to-clarification ratio (Chakraborty et al., 2025; Yadkori et al., 2024); calibration is evaluated by coverage and calibration error (Campos et al., 2024); decision quality follows probability of correct selection and solution quality under repeated runs (Wu, 2026); governance outcomes follow unsafe-output and drift-detection rates (Jingar, 2023).

VI. RESULTS AND ANALYSIS

A. Task–moule mapping.

Table I maps each framework layer to its task and evaluation.

TABLE I: HALO-SC MODULES, TASKS, AND EVALUATION

Layer	Task	Output	Evaluation
Triage	Risk-tier classification	tier (advisory/operational/irreversible)	Tier-accuracy, escalation-policy compliance
Grounding	RAG + KG evidence assembly	Provenance-cited evidence	Retrieval precision, hallucination rate (Kasarapu, 2026)
UQ generation	k-sample generation + trust scoring	$\hat{y}$, C(x,y), semantic entropy	Calibration error, confabulation detection (Farquhar et al., 2024; Kang et al., 2025)
Verification	Fact-check + bias-compensated judge	Verification vector	Detection of fabricated entities, judge reliability (Liu, 2025)
Abstention	Conformal calibration	answer/abstain/escalate	Error rate $\leq \varepsilon$, abstention rate, success-to-clarification (Yadkori et al., 2024)
Governance	Guardrails + GO/HOLD/NO-GO gate	Decision state and action	Unsafe-output rate, drift accuracy, audit completeness (Jingar, 2023)
Learning	Drift + KG/guardrail refresh	Updated system	Drift-detection accuracy, freshness

B. Reported performance envelope

Table II consolidates reported results across the hallucination, trust, and governance literature.

© SEP 2025 | IRE Journals | Volume 9 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I3-1722951

TABLE II: REPORTED PERFORMANCE IN HALLUCINATION-AWARE AND TRUSTWORTHY LLM DECISION SUPPORT

Method	Task	Reported result	Source
RAG pipeline (enterprise retail)	Grounded generation	Hallucination 1.5% avg vs 8.3% non-RAG (−81.9% relative); retrieval precision > 91%; relevance 4.6/5	(Kasarapu, 2026)
Corpus-scale grounding	KB enrichment	Hallucination 6.8% (10k docs) → 1.2% (100k docs)	(Kasarapu, 2026)
Conformal abstention	Calibrated refusal	Bounds hallucination rate at user-specified error level; low abstention on long answers	(Yadkori et al., 2024)
Conformal prediction	UQ framework	Model-agnostic, distribution-free guarantees	(Campos et al., 2024)
Enhanced conformal validity	Factual prediction sets	Sets guaranteed to contain a factual response with high probability	(Cherian et al., 2024)
Semantic entropy	Free-form confabulation detection	Detects confabulations over meanings; enables refusal/grounding	(Farquhar et al., 2024)
Fused UQ confidence	UQ for hallucination	$C(x,y) = \beta O + (1-\beta)S$ combining consistency and self-reflection	(Kang et al., 2025)
SELFCHECKGPT	Black-box hallucination detection	Consistency-based UQ; BERTScore + QA + n-gram combination	(Zhang et al., 2025)
Hallucination taxonomy + mitigation	LLM–OR formulation	Taxonomy + LLM-as-a-Judge, FSL, tool calling, multi-agent on NL4OPT/IndustryOR/EasyLP/ComplexOR	(Zhang, 2025)
KG-grounded generation	Supply chain relationship grounding	KG as structured verifiable source; mitigates demand-forecast hallucination	(Ge & Brintrup, 2024)
Multi-agent hallucination analysis	Agent networks	Conflicting + cascading hallucinations; inter-agent validation needed	(Wang et al., 2025)
Judicial hallucination	LLM-as-a-Judge reliability	Fact fabrication + cognitive biases undermine judge credibility	(Liu, 2025)
Governed LLM optimization	Safe optimization	Unsafe outputs −45%; drift detection > 92%	(Jingar, 2023)
Guardrail-centric fine-tuning	Deterministic enforcement	Qwen2.5-Coder-14B; ~50 guardrails; Python enforcement layer	(Davenport, 2026)
Decision authority framework	Governance-grade AI	Deterministic boundaries; human-final authority; audit replay; hard-stop	(KALAFATOGLU, 2025)
Governance-state architecture	Decision gating	GO / HOLD / NO-GO; irreversibility weighting; time-bound authority	(KALAFATOGLU, 2026)
Trustworthy AI review	Trust requirements	Fairness, explainability, accountability, reliability, acceptance	(Kaur et al., 2022)
Human-specified optimization	Human–AI collaboration	Statistically optimal, stable outcomes vs fully automated/two-stage	(Wu, 2026)
LLM–OR democratization	Optimization modeling	Substantial novel capabilities; major reliability challenges remain	(Wasserkrug et al., 2025)
OpsEval	Ops-domain evaluation	7,334 MCQ + 1,736 QA; 24 LLMs; 0.9185 human agreement	(Liu et al., 2025)

C. Worked calculations

1) The grounding dividend (headline). Retrieval-augmented generation reduces the hallucination rate from 8.3% (non-RAG baseline) to 1.5% on average — a relative reduction of $(8.3 - 1.5)/8.3 \approx 81.9\%$, at retrieval precision above 91% and response relevance 4.6/5 (Kasarapu, 2026). In the decision-exposure model of Section III-B, this is a fivefold-plus reduction in h: for a stream of decisions with expected error cost C_{err} , expected loss falls from $N \cdot 0.083 \cdot C_{\text{err}}$ to $N \cdot 0.015 \cdot C_{\text{err}}$.

2) The knowledge-base scale law. Hallucination rate falls from 6.8% at 10,000 documents to 1.2% at 100,000 documents — a relative reduction of $(6.8 - 1.2)/6.8 \approx 82.4\%$ across a 10× corpus expansion, with the study concluding that knowledge-base enrichment is the one major factor determining generative AI reliability for retail deployments (Kasarapu, 2026). This is the architectural argument for the portfolio's continuously refreshed knowledge layer: corpus investment is reliability investment.

3) Governance arithmetic. Governed LLM-based optimization with policy constraints, human-in-the-loop oversight, continuous monitoring, and drift-detection agents reduces unsafe decision outputs by 45% while maintaining drift-detection accuracy above 92% (Jingar, 2023). Composed with the 81.9% grounding reduction, the two-stage effect on unsafe outputs is multiplicative in exposure: the ungoverned, ungrounded baseline's exposure is reduced first by grounding (×0.18) and then by governance (×0.55), an aggregate reduction of $1 - 0.181 \times 0.55 \approx 90\%$ in expected unsafe-decision exposure — before abstention is even applied.

4) The abstention guarantee. Conformal abstention provides a rigorous guarantee: answered outputs err with probability at most the user-specified level ϵ, with abstention rates kept low through self-consistency-based scoring — significantly less conservative than log-probability baselines on long responses while comparable on short answers (Yadkori et al., 2024). The guarantee converts abstention from a heuristic into a contract: decision-makers can specify the maximum tolerable hallucination rate of the advisor, and the pipeline enforces it by calibration, not by hope (Campos et al., 2024).

5) The confidence fusion formula. The trust score $C = \beta O + (1 - \beta)S$ combines an evidence-based consistency term and a self-assessment term, with β tuned on calibration data to maximize agreement with correctness labels (Kang et al., 2025). Semantic entropy adds the free-form detector that token-level scores miss, making open-ended risk narratives — the portfolio's explanation layer — measurable for confabulation risk (Farquhar et al., 2024).

6) The verification caveat. LLM-as-a-judge is itself unreliable: judicial hallucination combines fact fabrication with cognitive and heuristic biases, so judge-based evaluation must be compensated and cross-checked against deterministic evidence (Liu, 2025). HALO-SC's verification layer therefore treats the judge as a signal, not a verdict — the distinction that keeps the trust boundary honest.

7) The human-in-the-loop dividend. Simulation-based human-specified optimization consistently achieves statistically optimal and stable outcomes compared with fully automated and two-stage variants (Wu, 2026). In expectation, selective human specification — the HOLD state with HALO-SC's governance gate — dominates full autonomy on probability of correct selection, the decision-quality metric that matters when error costs are asymmetric (Wu, 2026). This is the empirical license for the framework's human-final authority on irreversible-tier actions.

8) Evaluation of the evaluator. OpsEval's agreement coefficient of 0.9185 with human experts — improving over BLEU and ROUGE by 0.4471 and 1.366 respectively (Liu et al., 2025) — bounds how much the portfolio's own evaluation layer can trust standard NLP metrics when scoring supply chain outputs; the portfolio adopts Ops-domain scoring plus expert rubric review in the retail pattern (Kasarapu, 2026).

© SEP 2025 | IRE Journals | Volume 9 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I3-1722951

D. The grounding dividend

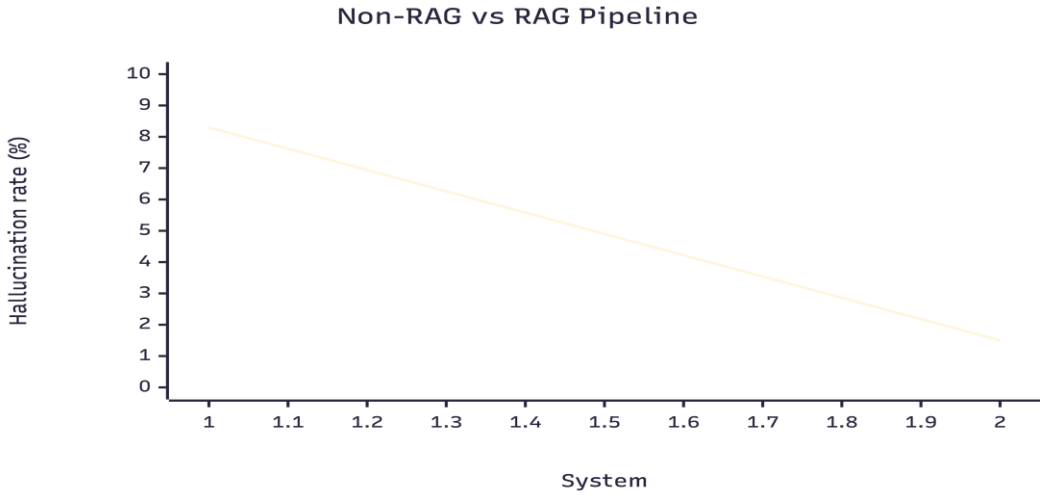

Fig. 2 plots the reported hallucination-rate reduction from RAG grounding.

Fig. 2. Reported hallucination rates: 8.3% for the non-RAG LLM baseline versus 1.5% average for the RAG pipeline (−81.9% relative), at retrieval precision above 91% and relevance 4.6/5 (Kasarapu, 2026).

E. The knowledge-base scale law

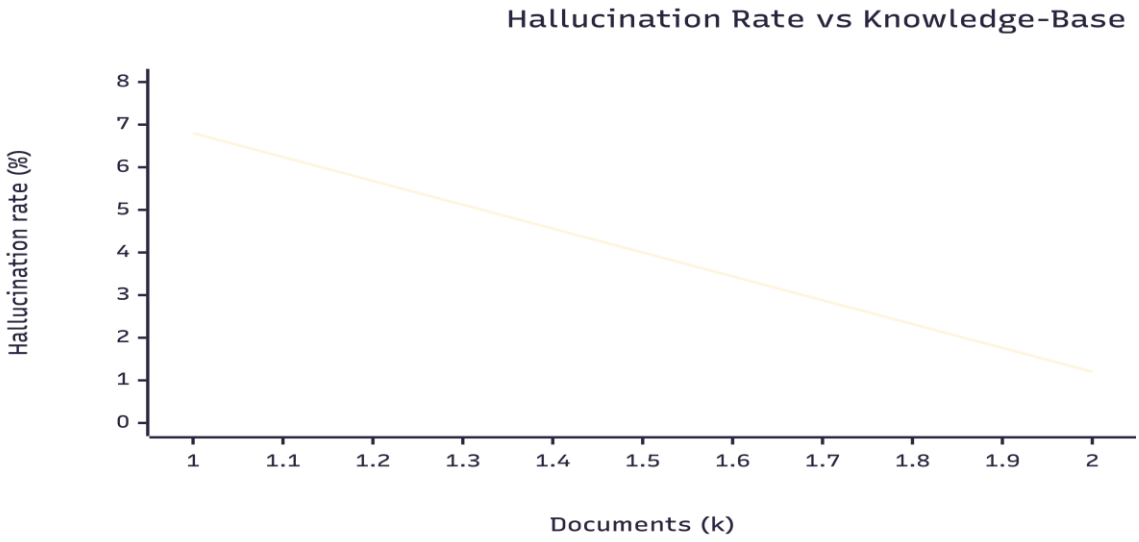

Fig. 3 plots hallucination rate as a function of corpus size.

Fig. 3. Reported hallucination rate versus corpus size: 6.8% at 10,000 documents falling to 1.2% at 100,000 documents — knowledge-base enrichment as the dominant reliability factor (Kasarapu, 2026).

© SEP 2025 | IRE Journals | Volume 9 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I3-1722951

F. Illustrative abstention trade-off

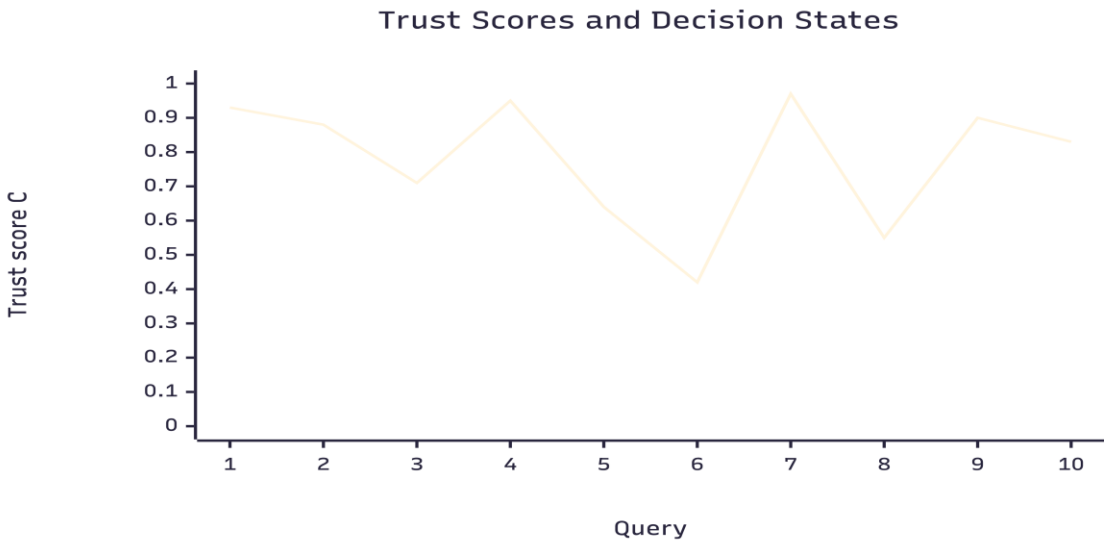

Fig. 4 shows the [illustrative] trust-score distribution and decision states for a supplier-risk query stream, consistent with the conformal-abstention guarantee structure (Yadkori et al., 2024).

Fig. 4. [Illustrative] trust scores across a supplier-risk query stream with a conformally calibrated threshold τ (dashed, at $C \approx 0.70$ for the reported error budget): scores above τ enter the GO/HOLD gate; the low-scoring queries (0.42, 0.55, 0.64) trigger abstention or escalation (Yadkori et al., 2024).

G. Ablation logic

Removing grounding reverts hallucination rates toward the 8.3% ungrounded baseline — an 81.9% exposure increase (Kasarapu, 2026); removing the uncertainty layer blinds the system to its own confabulation, forfeiting the semantic-entropy and consistency signals that make abstention possible (Farquhar et al., 2024; Kang et al., 2025); removing conformal calibration turns the abstention threshold into an unverified heuristic, forfeiting the rigorous error bound (Yadkori et al., 2024); removing governance reintroduces the unsafe-decision exposure and drift blindness of ungoverned generation (Jingar, 2023); removing human-final authority on irreversible tiers forfeits the statistically optimal outcomes of human-specified decision-making (Wu, 2026); and trusting the judge without bias compensation imports judicial hallucination into the verification layer (Liu, 2025).

VII. DISCUSSION

A. Trustworthiness is engineered, not emergent

The evidence base is unambiguous: hallucination rates are measurable (8.3% ungrounded, 1.5% grounded (Kasarapu, 2026); 6.8% → 1.2% with corpus scale (Kasarapu, 2026)), detectable (consistency, self-reflection, semantic entropy (Farquhar et al., 2024; Kang et al., 2025)), containable (conformal guarantees (Yadkori et al., 2024)), and governable (45% unsafe-output reduction (Jingar, 2023)). None of these properties arises from model scale; all of them arise from architecture. HALO-SC's contribution is to make each property a pipeline stage with a measured output.

B. Abstention is the professional adviser's privilege

The rigorous error bound of conformal abstention (Yadkori et al., 2024) converts "I don't know" from a model deficiency into a contractual guarantee. In high-stakes supply chain settings — where the cost of a hallucinated number is asymmetric and large — the right question is not "how accurate is the LLM?" but "what error rate can the decision-maker contractually expect from answered recommendations?" That is the question HALO-SC's calibrate-then-gate design answers.

C. Grounding is the scalable trust mechanism

Knowledge-base enrichment moves hallucination from 6.8% to 1.2% across a 10× corpus expansion — the documented dominant reliability factor in enterprise deployments (Kasarapu, 2026). This validates the portfolio's architectural bet: continuous ingestion, knowledge-graph verification (Ge & Brintrup, 2024), and provenance-cited generation are not conveniences but the trust infrastructure of the whole program.

D. Verification must verify itself

The judicial-hallucination literature shows the judge is fallible — fact fabrication and cognitive biases degrade LLM-as-a-Judge reliability (Liu, 2025). Trustworthy evaluation therefore requires bias-compensated judges cross-checked against deterministic evidence, and Ops-domain scoring with 0.9185 human agreement (Liu et al., 2025) rather than BLEU/ROUGE heuristics.

E. Human-final authority is the ultimate guardrail

The decision-authority literature formalizes what the supply chain risk literature already required: AI as decision support, humans as final authority, with deterministic boundaries, audit-ready lineage, and uncertainty-aware hard stops (KALAFATOGLU, 2025, 2026). The statistically optimal outcomes of human-specified optimization under human-specified algorithms (Wu, 2026) show this is not a concession to caution but a measurable accuracy dividend. HALO-SC reserves HOLD for irreversible actions as an engineering choice backed by decision-quality evidence.

F. Integration with the portfolio

HALO-SC is the trust boundary of the eleven-paper program: it wraps the early-warning layers (Papers 1–5), the platform self-healing loop (Papers 6 and 10), retrieval-grounded intelligence, multi-agent consensus, and explainable knowledge-graph narratives with measured, calibrated, governed outputs — every recommendation carrying a trust score, every refusal carrying a guarantee, every action carrying a replayable lineage. Where the prior papers asked what to do and why, HALO-SC answers how much to believe it — and when to say no.

G. Limitations

Reported results span heterogeneous tasks, datasets, and domains rather than a single controlled evaluation; the strongest hallucination-rate evidence comes from retail content generation rather than supply chain optimization, and the OR-domain evidence is benchmark-stage (Zhang, 2025); conformal guarantees assume calibration-set representativeness, which distribution drift can violate (Yadkori et al., 2024); multi-agent cascading hallucination remains only partially characterized (Wang et al., 2025); and human-final authority introduces organizational and liability questions beyond the technical scope. These define the agenda below.

VIII. CONCLUSION AND FUTURE WORK

This paper formalized trustworthy supply chain decision support as a hallucination-aware pipeline and proposed HALO-SC, a framework coupling risk-tiered triage, retrieval and KG grounding, uncertainty-quantified generation, multi-signal verification, conformal abstention with bounded error, deterministic governance with GO/HOLD/NO-GO states and human-final authority, and audit-ready decision lineage. It consolidated the reported evidence envelope — an 81.9% hallucination-rate reduction from grounding (8.3% → 1.5%) (Kasarapu, 2026), an 82.4% reduction from 10× knowledge-base enrichment (6.8% → 1.2%) (Kasarapu, 2026), rigorous conformal error bounds for abstention (Yadkori et al., 2024), semantic entropy for free-form confabulation detection (Farquhar et al., 2024), fused consistency-and-reflection confidence scoring (Kang et al., 2025), a 45% unsafe-output reduction with drift detection above 92% under governance (Jingar, 2023), guardrail-centric deterministic enforcement (Davenport, 2026), human-final decision authority architectures (KALAFATOGLU, 2025, 2026), statistically optimal human-specified optimization (Wu, 2026), and Ops-domain evaluation with 0.9185 human agreement (Liu et al., 2025) — and argued that the trustworthiness ladder — grounded, measured, verified, abstaining, governed — is what separates LLM assistants from LLM advisors in high-stakes commerce.

Future work will (i) implement HALO-SC end-to-end on live decision streams with unified evaluation under the Section V protocol, including per-tier abstention economics; (ii) extend conformal calibration to distribution drift via continuous recalibration and drift-triggered abstention, closing the representativeness gap (Yadkori et al., 2024); (iii) model and contain cascading hallucination in multi-agent supply chain systems through inter-agent validation and provenance propagation (Wang et al., 2025); (iv) build supply-chain-specific hallucination benchmarks — the gap between the retail and OR evidence — with expert-scored corpora across demand forecasting, supplier risk, and recovery planning; and (v) formalize human-final authority workflows for irreversible supply chain actions — deterministic boundaries, audit lineage, and uncertainty-aware hard stops (KALAFATOGLU, 2025, 2026) — so that trustworthy LLM decision support becomes a licensed, auditable, deployable capability, converting the risk of hallucination from an unquantified threat into a measured, contracted, and governed variable.

REFERENCES

[1] Akubilla, J., Somoye, O. I., Abiodun, F., & Serifat, O. A. (2025). The Role of Explainable AI in Enhancing Trust and Transparency in Supply Chain Risk Mitigation. International Journal of Multidisciplinary Research and Growth Evaluation, 6(3), 367–377. [Crossref]

[2] AlMahri, S., Xu, L., & Brintrup, A. (2026). Automating Supply Chain Disruption Monitoring via an Agentic AI Approach. ArXiv.Org. [Crossref]

[3] Baryannis, G., Dani, S., & Antoniou, G. (2019). Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. Future Generation Computer Systems, 101, 993–1004. [Crossref]

[4] Brandtner, P., & Hofer, F. (2026). Enhancing Procurement Processes in Supply Chain Management with Large Language Models. Procedia Computer Science, 278, 308–315. [Crossref]

[5] Campos, M., Farinhas, A., Zerva, C., Figueiredo, M. A. T., & Martins, A. F. T. (2024). Conformal Prediction for Natural Language Processing: A Survey. Transactions of the Association for Computational Linguistics, 12, 1497–1516. [Crossref]

[6] Chakraborty, N., Ornik, M., & Driggs-Campbell, K. (2025). Hallucination Detection in Foundation Models for Decision-Making: A Flexible Definition and Review of the State of the Art. In arXiv (Cornell University) (Vol. 57, Issue 7, pp. 1–35). Cornell University. [Crossref]

[7] Chen, X., Wanigarathna, Y. R., Chattopadhyay, A., Tarkoma, S., & Rao, A. (2025). A Pedagogical Approach for Evaluating AIOps. 40–49. [Crossref]

[8] Cherian, J. J., Gibbs, I., & Candès, E. J. (2024). Large language model validity via enhanced conformal prediction methods. In arXiv (Cornell University). Cornell University. [Crossref]

[9] Cornacchia, A., Alabdulaal, I., Saghier, I., Mirdad, A., Fayoumi, O., & Canini, M. (2025). Between Promise and Pain: The Reality of Automating Failure Analysis in Microservices with LLMs. 155–167. [Crossref]

[10] Cui, Y., Zeng, Y., Yan, J., Lin, K. L., Ji, K. Y., Zeng, J., Zhang, S., Luo, X., Su, B., Shen, C., & Yu, J. (2026). CSCBench: A PVC Diagnostic Benchmark for Commodity Supply Chain Reasoning. In arXiv (Cornell University). Cornell University. [Crossref]

[11] Davenport. (2026). GUARDRAIL-CENTRIC FINE-TUNING FOR DETERMINISTIC DECISION SYSTEMS. Zenodo (CERN European Organization for Nuclear Research). [Crossref]

[12] Dehan, M. F. Z., Anzum, K. Md. T., Disha, J. F., Masum, MD. M. R., & Mahmud, I. (2025). A Systematic Review on the Applications of How Explainable Artificial Intelligence can be Applied in the Supply Chain Management. [Crossref]

[13] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625–630. [Crossref]

[14] Ge, Z., & Brintrup, A. (2024). Enhancing Supply Chain Visibility with Generative AI: An Exploratory Case Study on Relationship Prediction in Knowledge Graphs. In arXiv (Cornell University). Cornell University. [Crossref]

[15] Guan, S., Liu, Y., & Cao, L. (2026). SupChain-Bench: Benchmarking Large Language Models for Real-World Supply Chain Management. arXiv (Cornell University). [Crossref]

[16] Gumaan, E. (2025). Theoretical Foundations and Mitigation of Hallucination in Large Language Models. In ArXiv.org. [Crossref]

[17] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 43(2), 1–55. [Crossref]

[18] Iruku, V. M. (2025). Multi-dimensional XAI Framework Revealing Critical Supply Chain Vulnerability Drivers. World Journal of Advanced Engineering Technology and Sciences, 15(3), 2141–2152. [Crossref]

[19] Jannelli, V., Schöpf, S., Bickel, M., Netland, T. H., & Brintrup, A. (2025). Agentic LLMs in the supply chain: towards autonomous multi-agent consensus-seeking. International Journal of Production Research, 1–31. [Crossref]

[20] Jingar, N. K. (2023). Ensuring Safety, Accountability, and Drift Resistance in LLM-Based Supply Chain Optimization. International Journal of Scientific Research in Science Engineering and Technology, 472. [Crossref]

[21] KALAFATOGLU, Y. (2025). Human-Final Decision Authority in Artificial Intelligence: A Deterministic and Auditable Governance Architecture. Zenodo (CERN European Organization for Nuclear Research). [Crossref]

[22] KALAFATOGLU, Y. (2026). A Deterministic Decision Authority Framework for Governance-Grade AI Systems. In Zenodo (CERN European Organization for Nuclear Research). European Organization for Nuclear Research. [Crossref]

[23] Kang, S., Bakman, Y. F., Yaldiz, D. N., Buyukates, B., & Avestimehr, S. (2025). Uncertainty Quantification for Hallucination Detection in Large Language Models: Foundations, Methodology, and Future Directions. In ArXiv.org. [Crossref]

[24] Kasarapu, B. C. (2026). Generative AI-Enabled Micro-Frontend Framework for Scalable and Intelligent Enterprise Retail Applications. International Journal of Computer Information Systems and Industrial Management Applications, 18, 297–309. [Crossref]

[25] Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy Artificial Intelligence: A Review. ACM Computing Surveys, 55(2), 1–38. [Crossref]

[26] Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth., 1(1). [Crossref]

[27] Liu, Y., Pei, C., Xu, L., Chen, B., Sun, M., Zhang, Z. L., Sun, Y., Zhang, S., Wang, K., Zhang, H., Li, J., Xie, G., wen, xiaohui, Nie, X., Ma, M., & Pei, D. (2025). OpsEval: A Comprehensive Benchmark Suite for Evaluating Large Language Models' Capability in IT Operations Domain. 503–513. [Crossref]

[28] Liu, Z. (2025). Judicial Hallucination: Investigating and Understanding the Unreliability of LLM-as-a-Judge. Applied and Computational Engineering, 203(1), 105–113. [Crossref]

[29] Padhy, A. K., Patel, T., Soni, V., Shivam, S., Thokala, G. B., & Vulugundam, B. (2026). Machine Learning-Based Fault Prediction in

Large- Scale Distributed Systems. 1–6. [Crossref]

[30] R., K. S., Ojha, D., Kaur, P., Mahto, R. V., & Dhir, A. (2024). Explainable artificial intelligence and agile decision-making in supply chain cyber resilience. University of North Texas Digital Library (University of North Texas), 180, 114194. [Crossref]

[31] Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., & Majumdar, A. (2025). A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions. ACM Computing Surveys, 58(3), 1–38. [Crossref]

[32] Syed, T. A., Belgaum, M. R., Jan, S., Khan, A. A., & Alqahtani, S. S. (2025). Agentic AI for Autonomous Defense in Software Supply Chain Security: Beyond Provenance to Vulnerability Mitigation. ArXiv.Org. [Crossref]

[33] USA, S. S. E., Dallas, Texas, & Natta, P. K. (2024). Designing Trustworthy AI Systems for Mission-Critical Enterprise Operations. International Journal of Future Innovative Science and Technology, 7(6). [Crossref]

[34] Wang, Y., Pan, Y., Su, Z., Deng, Y., Zhao, Q., Du, L., Luan, T. H., Kang, J., & Niyato, D. (2025). Large Model-Based Agents: State-of-the-Art, Cooperation Paradigms, Security and Privacy, and Future Trends. IEEE Communications Surveys & Tutorials, 28, 1906–1949. [Crossref]

[35] Wasserkrug, S., Boussioux, L., Hertog, D. den, Mirzazadeh, F., Birbil, Ş. İ., Kurtz, J., & Maragno, D. (2025). Enhancing Decision Making Through the Integration of Large Language Models and Operations Research Optimization. Proceedings of the AAAI Conference on Artificial Intelligence, 39(27), 28643–28650. [Crossref]

[36] Wu, R. (2026). Inventory optimization under supply chain disruptions: Leveraging large language models for human-AI collaborative decision-making. Journal of King Saud University - Computer and Information Sciences, 38(2). [Crossref]

[37] Yadkori, Y. A., Kuzborskij, I., Stutz, D., György, A., Fisch, A., Douc, R., Beloshapka, I., Weng, W.-H., Yang, Y.-Y., Szepesvári, C., Cemgil, A. T., & Tomasev, N. (2024). Mitigating LLM Hallucinations via Conformal Abstention. In arXiv (Cornell University). Cornell University. [Crossref]

[38] Zhang, G. (2025). Large Language Model enabled Mathematical Modeling. In ArXiv.org. [Crossref]

[39] Zhang, Y., Li, Y., Cui, L., Deng, C., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Computational Linguistics, 51(4), 1373–1418. [Crossref]