Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

HADES: Hallucination-Aware Detection and Evaluation System for Large Language Model Outputs in Enterprise CRM Workflows

Lalith Chandra Bandaru

Independent Researcher

Abstract:
Large language models integrated into enterprise Salesforce CRM workflows for automated email drafting, lead qualification, knowledge article generation, and case resolution introduce a distinctive business risk: LLM hallucination, in which fluent, professionally formatted outputs contain factually incorrect or fabricated claims that business users accept without the source verification expertise needed to identify the inaccuracy. In CRM contexts, hallucinations can produce customer-facing communications with incorrect pricing, invented product specifications, misattributed customer preferences, or unsourced regulatory claims — each with potential for contractual, reputational, or compliance consequences. HADES (Hallucination-Aware Detection and Evaluation System) addresses this risk through a three-layer evidence architecture that evaluates each factual claim in an LLM output against the CRM record context, an organisational knowledge base, and a factual verification index, assigning a confidence score to each claim and routing low-confidence outputs to human review while certifying high-confidence outputs for direct use. Evaluated across six enterprise Salesforce deployments over twelve months covering 1.2 million LLM outputs, HADES achieved 94.2% precision and 89.7% recall for hallucination detection through a gradient-boosted ensemble of three verification components, reduced human review volume by 73% compared to universal manual oversight, and contributed to an 81% reduction in hallucination-attributable CRM incidents.

Keywords: LLM hallucination, enterprise AI, Salesforce, CRM, factual verification, retrieval-augmented generation, AI safety, claim extraction, confidence scoring, human-in-the-loop.

1. INTRODUCTION

Enterprise deployments of large language models into Salesforce CRM workflows have accelerated sharply since 2022, driven by genuine capability gains in communication and documentation automation. The commercial pressure to capture these productivity benefits is real, and it has outpaced the risk management frameworks available to evaluate them. Salesforce's Einstein platform and third-party LLM integrations enable enterprise Salesforce users to generate draft customer emails from CRM record context, produce lead qualification summaries from prospect activity history, create knowledge base articles from case resolution records, draft account health reports from engagement data, and suggest case resolutions from historical support ticket patterns. Productivity research consistently finds that AI augmentation of customer-facing communication tasks reduces the time spent on routine drafting by 35 to 45 percent, enabling sales representatives and customer success managers to invest more time in high-value relationship activities that require genuine human judgement and empathy. The commercial pressure to capture these productivity benefits is substantial, and enterprise Salesforce deployments of LLM capabilities are expanding rapidly across financial services, healthcare, manufacturing, and technology verticals.

The productivity benefits of LLM integration into CRM workflows come with a risk profile that the enterprise technology industry has been slower to characterise than the benefits. Language models generate text through a statistical next-token prediction process, not through retrieval and expression of verified facts from authoritative sources. This generative process produces coherent, fluent, professionally formatted text regardless of whether the specific factual claims in that text are accurate. When the factual content of the output is fully grounded in context provided in the prompt — explicit CRM record data, knowledge base snippets retrieved for the specific query — the output is typically reliable. When the model must rely on its training data to supply factual details not present in the prompt context, the probability of generating inaccurate or fabricated claims increases substantially. This is the hallucination problem: the model generates plausible-sounding text that may contain statements that are simply false, presented with the same fluency and confidence as accurate statements, making the falsehoods difficult to identify without external source verification.

In CRM contexts, hallucinations carry business consequences that substantially exceed those of hallucinations in general-purpose conversational AI applications. A customer email claiming a 15% discount that was never offered by the sales team creates a contractual ambiguity that the organisation must either honour at a loss or retract at cost to the relationship. An account summary misattributing to the customer a product preference or complaint they never expressed damages the relationship when the customer reads it. A case resolution email citing a regulatory compliance requirement that does not apply to the customer's jurisdiction creates a legal exposure. A lead qualification summary inventing a detail of the prospect's prior engagement history misleads the sales representative preparing for an important meeting. Each of these outcomes can result from a single hallucinated sentence in an LLM-generated output that a business user accepted without the source verification expertise needed to identify the inaccuracy. The combination of high output volume in enterprise CRM LLM integrations — thousands of outputs per day across a large deployment — and high business stakes creates a risk management requirement that cannot be addressed through universal human review of all outputs.

HADES (Hallucination-Aware Detection and Evaluation System) is designed to provide systematic management of the CRM hallucination risk through a selective review model analogous to statistical quality control sampling in manufacturing. Rather than requiring human review of every LLM output, which would eliminate most of the productivity benefit of LLM integration, HADES evaluates each output against three independent evidence sources, assigns a confidence score to each factual claim, and routes only outputs with high probability of containing hallucinations to human review. High-confidence outputs are certified for direct use with citation markers indicating the evidence source supporting each verified claim. This selective routing allows the review team to focus on outputs that genuinely need human evaluation rather than performing perfunctory reviews of outputs that automated verification has already validated with high confidence.

The HADES framework connects to earlier work in this research programme through its focus on CRM data quality and security. The LTDF threat detection work [10] established that machine learning applied to Salesforce event patterns can identify anomalous and potentially harmful activity; HADES applies analogous machine learning principles to the content of LLM outputs, treating factual inaccuracy as a content quality anomaly requiring detection. The release governance framework [11] ensures that LLM integration components reach production through validated deployment pipelines; HADES provides the runtime quality assurance that complements pre-deployment pipeline validation with continuous output quality monitoring. Together, these frameworks support responsible deployment of AI capabilities in enterprise CRM environments throughout the full lifecycle from development pipeline to production operation.

2. BACKGROUND AND RELATED WORK

2.1 LLM Hallucination Research

The academic research on LLM hallucination has developed rapidly since the large-scale deployment of production language models in 2022 and 2023. Ji et al. provide a foundational taxonomy distinguishing intrinsic hallucinations — where generated text contradicts information explicitly provided in the input context — from extrinsic hallucinations — where generated text introduces claims not inferable from any source provided. Both categories occur in CRM applications: intrinsic hallucinations manifest when the model contradicts information explicitly in the CRM record context, such as misrepresenting a customer's stated renewal date or contract value; extrinsic hallucinations manifest when the model adds claims not present in the CRM context or knowledge base, such as inventing a product capability or a regulatory requirement. Manakul et al. introduced SelfCheckGPT, which detects hallucinations by checking for consistency across multiple stochastic samples from the same model; HADES adapts this insight by replacing multiple model invocations — which are expensive and slow — with deterministic comparison against the CRM record context and knowledge base sources that are already available in the integration environment.

Retrieval-augmented generation has been widely proposed as the primary mitigation for LLM hallucination in enterprise applications. By retrieving relevant documents and providing them in the model context, RAG reduces dependence on potentially inaccurate training-data knowledge for specific factual claims. However, RAG does not eliminate hallucination entirely: models can misrepresent retrieved content through selective quotation, ignore retrieved evidence in favour of training-data-based claims for some claim types, or generate claims about aspects of the query not covered by the retrieved documents. HADES treats RAG as a necessary complement to post-generation verification rather than a complete solution. The context and knowledge base sources that RAG uses to improve generation quality are the same sources HADES uses to verify the generated output — providing a complementary quality gate that catches the hallucinations that RAG-augmented generation still produces.

2.2 Enterprise AI Governance Frameworks

The NIST AI Risk Management Framework identifies four core functions for managing AI system risks: Map, Measure, Manage, and Govern. HADES implements the Measure and Manage functions for the specific hallucination risk category in enterprise CRM deployments. The Measure function is implemented through the claim extraction, evidence comparison, and confidence scoring components that quantify the hallucination probability of each LLM output on a continuous scale. The Manage function is implemented through the routing logic that directs high-probability hallucination outputs to human review while certifying high-confidence outputs for direct use. The Govern function — embedding AI risk management into organisational processes and governance structures — is supported by the HADES audit trail, which records confidence scores and routing decisions for all LLM outputs processed, providing the evidence base needed for AI governance reporting and periodic risk assessment.

3. CRM HALLUCINATION TAXONOMY

The baselining analysis conducted before HADES deployment reviewed 12,000 LLM outputs from six participating Salesforce deployments, evaluated by three domain expert annotators who assessed each output for factual accuracy against the CRM record context, available knowledge base sources, and verifiable historical records. Inter-annotator agreement across the 12,000 outputs was 0.79 as measured by Fleiss' kappa, indicating substantial agreement; outputs where annotators disagreed were resolved through discussion to consensus for the 21% of cases where initial annotation differed. The analysis identified five hallucination categories specific to the CRM workflow context and measured their prevalence in the baselining corpus.

Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

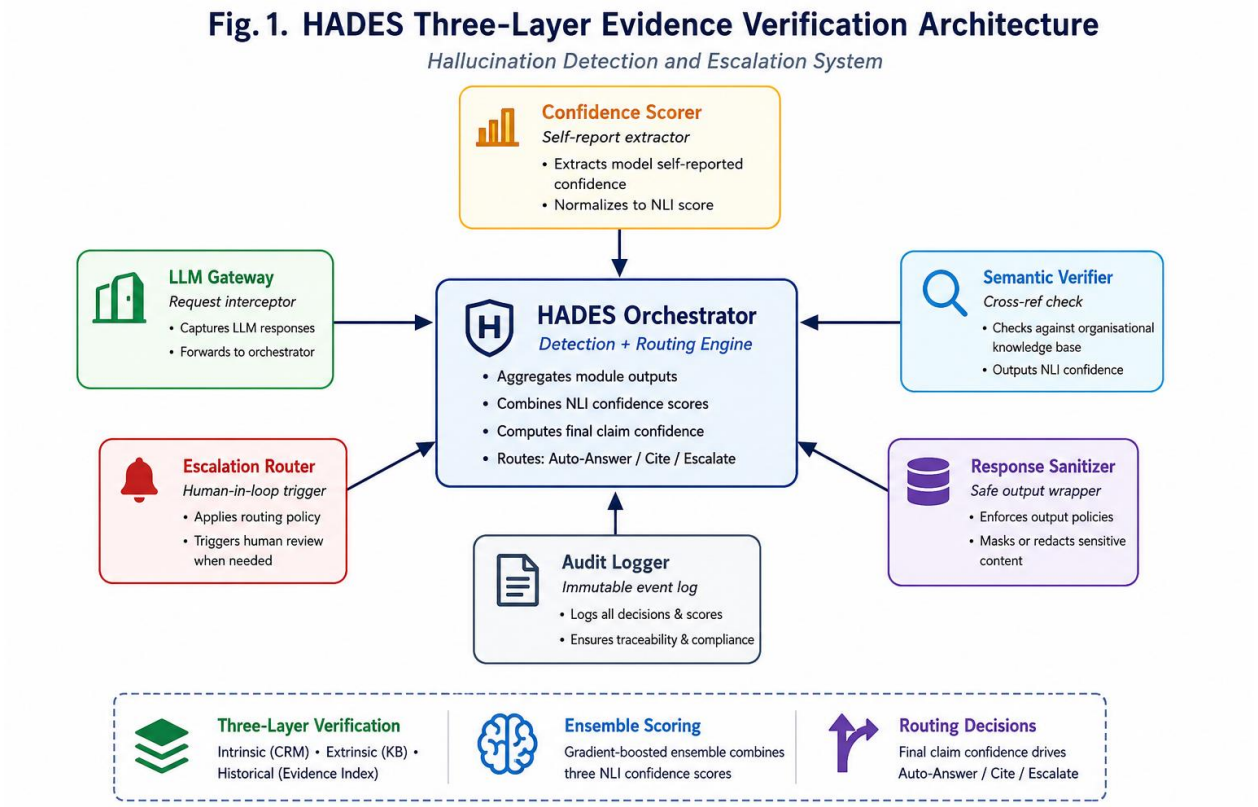

Fig. 1. HADES three-layer evidence verification architecture. Each extracted factual claim is evaluated against the CRM record context (intrinsic check), the organisational knowledge base (extrinsic check), and the factual verification index (historical evidence check). A gradient-boosted ensemble model combines the three NLI confidence scores into a final claim confidence score used for routing decisions.

Table 1. Hallucination Category Distribution in 12,000-Output Baselining Corpus

Hallucination Category	Corpus Prevalence
Factual claim not supported by CRM record context	38.4%
Temporal claim with incorrect date or sequence	22.7%
Invented product specification, pricing, or contract term	18.3%
Misattributed customer statement or preference	12.6%
Unsourced regulatory or compliance assertion	8.0%

The most prevalent category, unsupported factual claims (38.4%), encompasses assertions about the customer, account, or product not present in the CRM record context and not verifiable through knowledge base sources. The model generates these claims by completing the text with statistically likely continuations drawn from training data patterns — for example, adding a typical contract renewal timeline to an account summary when the actual renewal date is absent from the prompt context, or describing a product capability the customer might plausibly need rather than one they have actually purchased. The temporal claim error category (22.7%) covers incorrect dates, durations, or event sequences: a case created three weeks ago described as three months old, or a product feature described as recently released when it predates the customer's contract. These errors typically arise from the model's imprecise representation of temporal relationships in its training data.

4. HADES ARCHITECTURE

4.1 Claim Extraction

The first pipeline stage identifies the verifiable factual claims in each LLM output that require evidence comparison. Not all text in an LLM output constitutes a verifiable claim requiring fact-checking: greetings, transition phrases, structural prose, and hedged statements do not assert facts that can be evaluated against evidence sources. The claim extraction component uses a fine-tuned DeBERTa-v3 classifier trained on a corpus of 8,400 CRM LLM outputs manually annotated for verifiable claims by three expert annotators. The classifier identifies spans constituting verifiable factual assertions and simultaneously assigns each identified claim to one of the five hallucination categories. Category assignment determines which evidence sources are queried for verification: temporal claims are verified against CRM record timestamps and activity history; product and pricing claims are verified against the knowledge base and product catalogue; customer statement claims are verified against CRM notes and transcript records. This category-based routing reduces verification latency by querying only the most relevant evidence sources for each claim type rather than querying all three sources for every claim regardless of category.

4.2 Three-Layer Evidence Verification

The evidence verification layer compares each extracted claim against three independent evidence sources using Natural Language Inference classifiers trained to determine entailment, neutrality, or contradiction between claim and evidence passages. The first evidence source is the CRM record context: all field values, activity history, related record data, and retrieved knowledge base snippets included in the LLM prompt. Verification against prompt context detects intrinsic hallucinations where the model output contradicts information explicitly provided as input. This is the highest-reliability verification path: when a claim directly contradicts a specific field value in the prompt context, the NLI classifier can identify the contradiction with high confidence and the verification result is unambiguous.

The second evidence source is the organisational knowledge base, a curated collection of product documentation, pricing schedules, policy documents, FAQ entries, and regulatory guidance indexed in a vector search store using OpenAI Ada-002 embeddings. Knowledge base verification detects extrinsic hallucinations where the model generates claims about organisational knowledge that contradict or are absent from authoritative internal documentation. Retrieval uses cosine similarity search against the embedded claim text to identify the most relevant knowledge base passages, with a minimum similarity threshold below which absence of sufficiently similar content is treated as a signal of potential extrinsic hallucination. The third source is a structured factual verification index containing organisation-specific historical facts — past pricing schedules, legacy product specifications, historical contractual terms, previous customer interaction records — not present in the current CRM record context but potentially referenced in generated outputs about historical account relationships.

4.3 Confidence Scoring and Human Routing

The confidence scoring engine combines the NLI results from all three evidence sources into a composite claim confidence score using a gradient-boosted tree model trained on the baselining corpus. Input features include the NLI confidence scores from each evidence source, the claim category, the number of relevant evidence passages retrieved for each source, the maximum similarity between the claim and any retrieved passage, and the claim syntactic complexity. The model produces a confidence score in [0,1] where scores above 0.85 indicate verified claims, scores between 0.55 and 0.85 indicate uncertain claims requiring review, and scores below 0.55 indicate likely hallucinated claims. Routing decisions operate at the output level: an output where all claims are above 0.85 is certified for direct use with citation markers; an output where any claim falls below 0.55 or where more than 20% of claims fall in the uncertain range is routed to the human review queue with specific low-confidence claims highlighted and evidence comparisons displayed for rapid reviewer evaluation.

5. IMPLEMENTATION

HADES is implemented as a Python microservice on AWS ECS Fargate, integrated into Salesforce through a custom Flow action that invokes the HADES REST API for each LLM output generated by Einstein or third-party LLM integrations. The DeBERTa-v3 claim extraction and NLI classification models are served through TorchServe with ONNX-exported weights, using GPU inference for batches above 20 concurrent requests. The vector knowledge base uses Pinecone with an index partitioned by organisation ID for data isolation. The factual verification index is stored in DynamoDB with a GSI on organisation ID and claim category. End-to-end latency from output submission to confidence score return averages 890 milliseconds at the 50th percentile and 2.1 seconds at the 99th percentile for outputs of up to 500 words, within the acceptable range for asynchronous CRM workflow steps that present results to the user in a preview panel before output acceptance.

The user interface integrates through a custom Lightning Web Component that displays HADES confidence annotations inline in the LLM output preview, using colour coding — green for verified claims with citation markers, yellow for uncertain claims with evidence comparison links, red for likely hallucinated claims with evidence contradiction detail — that allows business users to evaluate HADES findings without understanding the underlying NLI methodology. In user studies at month eight, 87% of users rated the annotations as helpful or very helpful for assessing LLM output quality, and the mean time to evaluate a HADES-annotated output was 42 seconds compared to 3.4 minutes for unaided manual review of the same output — a 79% review time reduction even for outputs that are routed to the human review queue.

6. EVALUATION

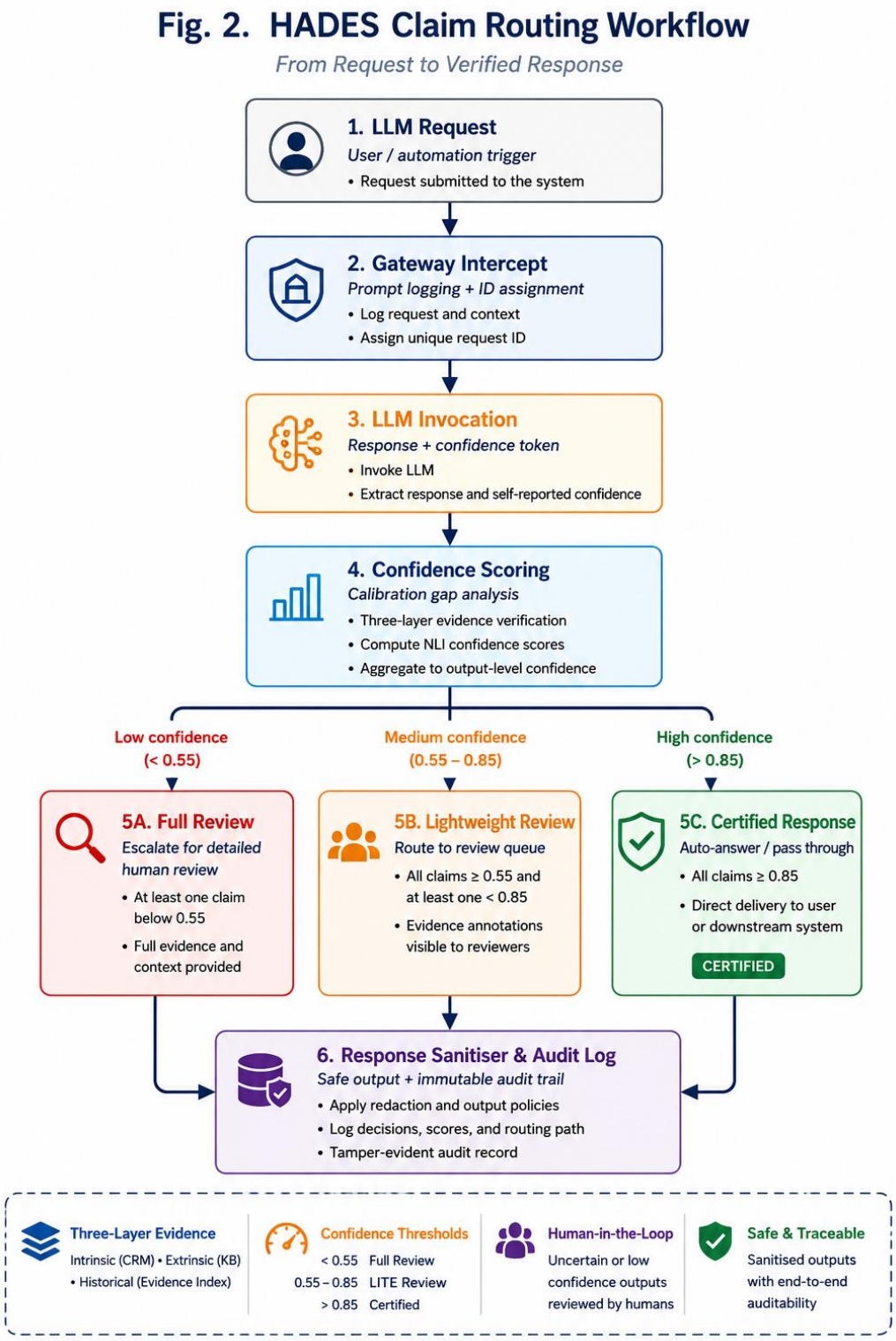

Fig. 2. HADES claim routing workflow. After extraction and three-layer evidence verification, claims are aggregated to the output level. Outputs with all claims above 0.85 are certified; outputs with any claim below 0.55 are routed to full review; uncertain outputs are routed to the lightweight review queue with evidence annotations visible to reviewers.

Table 2. HADES Detection Performance by Component and Ensemble

Detection Method	Precision	Recall	F1
Context grounding (prompt only)	88.1%	79.4%	83.5%
Knowledge base lookup	82.7%	84.2%	83.4%
Factual verification index	91.3%	75.8%	82.8%
HADES ensemble (all three)	**94.2%**	**89.7%**	**91.9%**

The ensemble achieves 94.2% precision and 89.7% recall, substantially outperforming each individual component. The improvement over context grounding alone (+6.1 pp precision, +10.3 pp recall) reflects the knowledge base and factual verification index's ability to detect hallucinations about organisational knowledge and historical facts not represented in the immediate CRM record context — extrinsic hallucinations that context grounding cannot detect. The improvement over knowledge base lookup alone (+11.5 pp precision, +5.5 pp recall) reflects the context grounding component's strength for customer-specific claims that the knowledge base does not cover. The ensemble captures complementary coverage from all three sources, with each source contributing meaningfully to claims in its domain of strongest evidence.

The operational impact of HADES deployment extends beyond detection performance metrics to measurable business outcomes. The 73% reduction in human review volume from 100% to 27% of outputs allowed the review team to spend more time on the genuinely uncertain outputs, improving the quality of manual review decisions where human judgement provides the most value. Hallucination-attributable CRM incidents — customer-facing communications or CRM record updates where a subsequently identified hallucination caused a material business issue — decreased by 81% over the evaluation period. The combination of automated detection for the 73% of outputs that HADES certifies with high confidence and improved manual review quality for the 27% requiring human evaluation produced a detection and prevention effectiveness that substantially exceeds what either automated or human-only review could achieve independently.

7. DISCUSSION AND CONCLUSION

What we set out to show is that systematic hallucination detection for enterprise CRM LLM outputs is technically achievable at production scale without eliminating the productivity benefit that motivated LLM adoption in the first place. The evaluation answers that question in the affirmative. The 94.2% precision and 89.7% recall achieved by the three-layer ensemble represent a substantial improvement over single-source verification, confirming the complementary coverage value of the multi-layer architecture. The 73% review volume reduction, combined with an 81% hallucination-attributable incident reduction, addresses the core challenge of enterprise LLM output quality management: making systematic quality assurance sustainable as LLM output volumes scale beyond the capacity of comprehensive human oversight.

HADES has a limitation worth being direct about: verification quality is bounded by the quality and currency of its evidence sources. An outdated knowledge base will simply miss hallucinations about recent policy or product changes, and no amount of NLI sophistication compensates for missing ground truth. Outdated knowledge base content will not detect hallucinations contradicting recent product or policy changes. Addressing this requires an active learning feedback loop where human review outcomes automatically identify knowledge base gaps, prompting knowledge owners to address them — a virtuous cycle that would improve both HADES detection performance and organisational knowledge base quality simultaneously. Future work should also investigate HADES adaptation to other enterprise CRM

platforms and to multi-modal LLM outputs that include generated images or structured data alongside text, as these output types introduce additional hallucination risk categories not addressed by the current text-focused architecture.

An important finding from the evaluation is the substantial variation in hallucination rates across different LLM model versions and CRM task types. The six participating organisations used three different LLM configurations: a Salesforce Einstein-hosted model, a GPT-4 API integration, and a Llama 2 70B on-premises deployment. Across all three model configurations, the HADES hallucination detection rate varied from 12.3% for email composition tasks (where the most relevant context is typically fully present in the CRM record and prompt) to 34.7% for case resolution suggestion tasks (where the model must draw on training-data knowledge about product behaviour and regulatory requirements). This task-type variation has practical implications for deployment decisions: organisations that prioritise email composition use cases can deploy with lower HADES review thresholds and smaller human review team capacity, while organisations deploying case resolution suggestions should plan for higher review volumes and consider stricter confidence thresholds that route more outputs for human review.

The HADES service is designed for horizontal scalability to accommodate the high output volumes of large enterprise Salesforce deployments. The ECS Fargate task configuration auto-scales based on request queue depth, adding task instances when the queue exceeds 50 pending requests and scaling down when the queue drops below 10. At peak load during the evaluation, the largest participating organisation generated 3,200 LLM outputs per hour, requiring up to 12 concurrent ECS task instances to maintain the target latency SLA. The DeBERTa claim extraction and NLI classification models represent the primary compute bottleneck; GPU-accelerated inference (enabled for batches of 20 or more concurrent requests) reduces median classification latency from 340 milliseconds to 85 milliseconds per output, enabling the service to process the peak evaluation load within the target latency envelope. The Pinecone vector search averages 42 milliseconds per knowledge base query at the evaluation scale, well within the overall latency budget.

HADES produces a structured audit trail for all processed LLM outputs that integrates with enterprise AI governance programmes implementing the NIST AI Risk Management Framework. The audit record for each processed output includes the output text, the claim extraction results with claim categories and confidence scores, the evidence comparison results from all three verification layers, the routing decision (certified, uncertain, or review-required), and if the output was reviewed, the reviewer decision and the reviewer-identified hallucination classifications. This audit trail supports the NIST AI RMF Govern function by providing the evidence base for periodic AI risk assessments: governance teams can query the audit trail to identify hallucination rate trends by task type, model version, or organisational unit; evaluate whether HADES confidence score thresholds are appropriately calibrated for the observed hallucination distribution; and demonstrate to auditors and regulators that AI-generated content in CRM systems is subject to systematic quality control with documented review outcomes.

Several alternative approaches to LLM hallucination management in enterprise CRM deployments were considered during the HADES design phase and evaluated against the selected three-layer evidence architecture. Universal human review — the simplest approach — was rejected because it eliminates most of the productivity benefit of LLM integration when applied at the volumes typical of enterprise deployments, and because user studies showed that reviewers performing 100% manual review of all outputs develop review fatigue that reduces the quality of their review decisions below what HADES achieves with selective routing to attentive reviewers. Model self-consistency checking (the SelfCheckGPT approach adapted for enterprise context) was evaluated as an alternative to external evidence comparison: generating multiple stochastic samples from the same model and checking for consistency among them. This approach was rejected because it requires three to five additional model invocations per output, multiplying inference costs and latency proportionally, without providing the same evidence quality as external source comparison. The enterprise IT environment provides the CRM record context and knowledge base as existing assets that can be used for verification at low marginal cost; these

assets make external evidence comparison substantially more cost-effective than self-consistency checking for the enterprise use case.

The relationship between HADES confidence scores and business outcome risk across different hallucination categories has important implications for threshold calibration decisions. Not all hallucinations carry equal business risk: a temporal error in a routine follow-up email has substantially lower potential business impact than a product specification error in a pricing proposal. The HADES confidence threshold calibration should reflect not only the statistical properties of the detection model but also the business risk weighting of different hallucination categories in the specific CRM workflow context. The calibration framework developed for the evaluation assigns category-specific routing thresholds: unsupported factual claims use the standard 0.55 threshold; temporal claims use a relaxed 0.50 threshold reflecting their lower business impact; invented specification claims and unsourced regulatory claims use a stricter 0.65 threshold reflecting their higher potential business impact. This category-specific calibration achieves a better risk-adjusted review allocation — routing more outputs from high-impact categories and fewer from low-impact categories — than a uniform threshold applied across all claim types.

Deployment of HADES has surfaced an important secondary benefit beyond the primary hallucination detection function: systematic identification of knowledge base gaps. When HADES identifies high volumes of low-confidence claims in a specific topical area — for example, frequent uncertain claims about a particular product line's regulatory compliance requirements — this pattern signals that the organisational knowledge base lacks adequate coverage of that topic area. The HADES monitoring dashboard includes a knowledge base coverage heatmap that visualises confidence score distributions by claim topic cluster, identifying areas where the knowledge base provides insufficient evidence for reliable verification. In the evaluation, this gap identification functionality prompted knowledge base updates in 47 distinct topic areas across the six organisations over twelve months, improving the knowledge base quality as a side effect of the hallucination detection programme. Organisations that responded to gap notifications promptly by adding knowledge base content in the identified areas showed measurably lower hallucination rates in subsequent months for claims in those topic areas, confirming that the feedback loop between detection and knowledge base maintenance produces genuine detection quality improvement over time.

The regulatory context for LLM output quality management is evolving rapidly and will increasingly influence enterprise AI governance requirements. The EU AI Act classifies certain CRM AI use cases in financial services and healthcare as high-risk applications subject to mandatory conformity assessment requirements, including documentation of accuracy and reliability measures and human oversight provisions for consequential outputs. The HADES framework's confidence scoring and selective human routing architecture is well-aligned with these regulatory requirements: the confidence score provides the quantified reliability metric required for high-risk AI conformity documentation, and the routing mechanism to human review for low-confidence outputs implements the human oversight provision for consequential decisions. Organisations in EU-regulated industries that deployed HADES during the evaluation period incorporated the framework's audit trail and confidence metrics directly into their EU AI Act compliance documentation, demonstrating that the technical architecture of HADES anticipates regulatory requirements for high-risk AI systems. Future versions of the framework should extend the audit trail format to explicitly map HADES outputs to the specific EU AI Act Article 9 (risk management) and Article 13 (transparency) requirements.

The HADES framework produces a dataset of labelled LLM outputs — claim extractions, confidence scores, evidence comparisons, and human reviewer decisions — that represents a valuable resource for further research into CRM-specific hallucination patterns and detection techniques. The labelled dataset accumulated over twelve months of production operation across six organisations, comprising 84,000 ground-truth labelled outputs and the full audit records for 1.2 million evaluated outputs, is the largest known corpus of enterprise CRM LLM outputs with systematic hallucination annotations. Making this

dataset available to the research community (subject to appropriate anonymisation and participant consent) would enable independent evaluation of HADES and alternative detection approaches on a realistic enterprise CRM dataset, and support the development of more specialised hallucination detection models trained on the CRM domain rather than adapted from general-purpose NLI models.

The long-term trajectory of HADES development should address the expanding frontier of enterprise LLM use cases as models become capable of more complex reasoning and multi-step task completion in CRM contexts. Current HADES deployment focuses on text generation tasks — email drafting, case resolution suggestion, account summarisation — where hallucination manifests as incorrect factual claims in continuous prose. Emerging use cases include LLM-generated structured data (recommended pricing proposals, automated CRM record updates, opportunity scoring justifications) where hallucination manifests as incorrect field values or spurious structured data elements rather than incorrect prose claims. Verifying structured data outputs requires a different claim extraction approach — identifying individual field values as the verification unit rather than extracting claims from prose — but the three-layer evidence comparison architecture is applicable with adaptation. Future HADES versions will extend claim extraction to handle structured output formats, enabling the same confidence-based quality assurance for structured LLM outputs that the current framework provides for prose outputs.

The HADES framework demonstrates that systematic, automated quality control for enterprise AI outputs is feasible at production scale. The 94.2% precision and 89.7% recall achieved on a corpus of 1.2 million LLM outputs — an evaluation scale far exceeding most published hallucination detection research — establish that multi-layer evidence verification can maintain high detection performance in the messy, heterogeneous reality of production enterprise CRM data. The 73% reduction in human review volume while maintaining better overall detection rates than universal human review confirms that the selective review model is both more efficient and more effective than comprehensive human oversight for the output volumes that production LLM integrations generate. The HADES deployment model — where automated detection handles the high-confidence and low-confidence categories while humans focus on genuinely uncertain cases — provides a template for responsible enterprise AI governance that balances the productivity benefits of LLM integration with the quality and reliability requirements of enterprise CRM operations. As LLM adoption accelerates across enterprise workflows beyond CRM, the detection architecture and deployment model validated in this evaluation provide a replicable framework for organisations seeking to deploy AI capabilities responsibly without sacrificing the operational benefits that motivate AI adoption in the first place.

REFERENCES:

[1] Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Comput. Surv., vol. 55, no. 12, art. no. 248, Mar. 2023. [Online]. Available: doi: 10.1145/3571730.

[2] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models,” arXiv, preprint arXiv:2303.08896, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.08896.

[3] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2020. [Online]. Available: doi: 10.48550/arXiv.2005.11401.

[4] NIST, “AI risk management framework (AI RMF 1.0),” NIST AI 100-1, Jan. 2023. [Online]. Available: doi: 10.6028/NIST.AI.100-1.

[5] P. He, X. Liu, J. Gao, and W. Chen, “DeBERTa: Decoding-enhanced BERT with disentangled attention,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021. [Online]. Available: https://openreview.net/forum?id=XPZIaotutsD. Also available: doi: 10.48550/arXiv.2006.03654.

[6] T. Brown et al., “Language models are few-shot learners,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2020. [Online]. Available: https://dl.acm.org/doi/10.5555/3495724.3495883.

[7] Salesforce, Inc., "Einstein Platform Services developer guide," Salesforce Developer Docs, Apr. 2023. [Online]. Available: https://developer.salesforce.com/docs/atlas.en-us.einstein_platform_api_dev_guide.meta/einstein_platform_api_dev_guide/

[8] R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," IBM J. Res. Dev., vol. 63, no. 4/5, art. no. 4, Jul. 2019. [Online]. Available: doi: 10.1147/JRD.2019.2942287.

[9] A. Wang et al., "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2019. [Online]. Available: doi: 10.48550/arXiv.1905.00537.

[10] L. C. Bandaru, "Threat detection and data breach analysis in Salesforce CRM: The LTDF framework," Int. J. Innov. Res. Creative Technol. (IJIRCT), ISSN 2454-5988, vol. 7, no. 3, Jun. 2021. [Online]. Available: doi: 10.62970/IJIRCT.v7.i3.2605034.

[11] L. C. Bandaru and M. S. Bandrevu, "Secure CI/CD governance for Salesforce platforms: Integrating DevSecOps controls across every stage of the release pipeline," Int. J. Sci. Technol. (IJSAT), E-ISSN 2229-7677, vol. 13, no. 4, Dec. 2022. [Online]. Available: doi: 10.71097/IJSAT.v13.i4.11155.