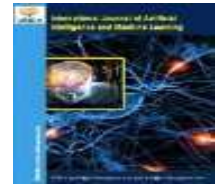




SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Design and Development of an Intelligent Governance and Compliance Framework for Prompt Engineering and AI Output Validation in Regulated ERP Environments Using Generative AI

Vinay Gautam¹, Dr. Atul Agrawal²

¹Dept. of Computer Applications, Oriental University, Indore, Indore, M.P., India, Mail ID: gettovinay@gmail.com

²Dept. of Computer Applications, Oriental University, Indore, Indore, M.P., India, Mail ID: atulagrawal@orientaluniversity.in

ABSTRACT

In enterprise resource planning (ERP) environments, generative AI is becoming more and more part and parcel of the work, and while AI prompts are useful, the risks of being triggered, sensitive to transactions, and plagued by unreliable outputs can compromise accountability and operational control. In this study, this was a proof-of-concept, script-based governance framework that encompasses the operationalization of AI-output validation (hallucination-risk assessment), prompt-injection detection, and ERP transaction-risk classification. Three secondary datasets were combined into 5,000 synthetic ERP-GenAI scenarios, which were then fed into specific machine-learning modules and a deterministic decision engine that yielded Approved, Human Review, or Rejected results with structured audit reasons. Source label and pre-defined governance rules were used to generate reference decisions. It attained the following results: exact agreement rate of 76.96% with these decisions, unsafe detection rate of 90.89% and unsafe approval rate of 9.11%. This produced 1,377 Approved, 1,386 Human Review and 2,237 Rejected outcomes. Error analysis revealed that the largest number of unsafe approvals were due to missed prompt attacks (76.72%), followed by missed hallucinations and ERP-risk errors. The results prove multi-layer governance to be achievable, incorporating transaction context, prompt integrity, output risk, escalation logic and auditability. The results, however, show a synthetic integration benchmark, and are NOT a confirmation of regulatory conformity or production readiness. Better detection, ERPS based verification and independent risk labelling is needed before operational deployment.

Keywords: generative artificial intelligence, ERP governance, prompt-injection detection, hallucination-risk assessment, human review, auditability

1. INTRODUCTION

Generative AI has moved beyond early experimentation to become a part of organizations, transforming knowledge work, decision support, process execution, and human-machine interaction. For enterprises, LLMs can help employees to understand transactions, write reports, look up information on procedures, and aid in operational decisions. However, these are not without risk of unreliable generation, timely tampering, lack of transparency of accountability, and inappropriate automation. Generative AI can contribute to productivity and augment human work in human resource management, and it can also pose problems in terms of bias, transparency, governance, and managerial responsibility [1]. These issues are more impactful when AI generated content is fed into integrated enterprise systems than it is when it is only a tool for conversations. These risks are exacerbated in an ERP-driven environment as recommendations can affect both financially significant transactions and process controls in addition to records that are regulated.

Generative AI is distinct from traditional predictive analytics in that it generates an output that is dependent on the context as well as on the natural-language command, establishing a dynamic relationship between user intent, model behavior, and downstream action. It has the ability to synthesize, transform and generate information that extend enterprise applications and make verification and control more complex [2]. There are significant challenges in enterprise resource planning (ERP), where finance, procurement, logistics, inventory, human resources, and compliance-sensitive records are tied to each other via common process architectures. An unsupported output can affect a purchase order, payment recommendation, audit response or management report, and a malicious or poorly defined prompt can cause policy circumvention or unauthorized disclosure, or incorrect guidance. Here, governance means that the conditions of the transactions need to be assessed jointly, integrity is achieved in a timely way, output is reliable, escalation requirements are not ignored, and decisions are traceable.

Systematic study of the current LLM challenges on the topics of transparency, robustness, reproducibility, human evaluation and societal impact have been requested by current research [3]. However, enterprise governance is still in a disjointed state – spread over different technical and organizational fields. Prompt-

engineering research is concerned with demonstrating and providing instructions and templates which can generate useful model behavior, while security research is concerned with prompt injection, instruction manipulation and safeguard circumvention. Unsupported or inconsistent generation is studied under hallucination, and transaction anomalies, process conformance and operational risk are studied under ERP analytics. These streams are used to give a basic level of foundation and are typically treated separately as a control. This means that when AI responds to a prompt, an organization can identify it as suspicious without evaluating the prompt itself when it was called out, or know that a prompt is unsafe but not be able to assess the reliability of the output for the specific business context.

Prompt engineering goes beyond the optimizing of responses in regulated environments. It also relates to escalation, misuse prevention, authorization and provenance. The intention of the present study is not to optimize prompts, manage prompt lifecycle, or improve the quality of responses to prompts, but to identify prompts that are injection-oriented and/or policy-circumventing. Likewise, AI-output validation is implemented as text risk for hallucination-related output. It does not carry with it any direct evidence of the facts relative to the ERP records, supporting documents, transaction values or organization policies. These differences represent the scope of implementation and ensure that prompt governance or output validation isn't understood in a way that is broader than the computational parts can handle.

The main research problem is that there is a lack of an integrated and auditable decision mechanism that is able to integrate the risk from ERP with prompt-level risk and AI-output risk prior to the acceptance of generated content in an enterprise workflow. Existing auditing approaches acknowledge the need to evaluate their LLM at governance, model, and application levels [4], however, there has been relatively little empirical research linking these levels to transaction-level ERP decisions. While a model may work well in general testing, it may give a less than safe response in a procurement or financial situation. On the other hand, a regular business can be accompanied by an instant that is meant to disable protections. In the absence of a single pipeline, organizations may experience disparate thresholds, duplicate controls, poor logic for escalation, and limited rationale for why certain actions by AI were accepted, rejected or referred for review.

This is a challenge of responsibility, as well. Key elements of effective AI accountability include clear responsibilities, accountability systems, and linkages between technical outcomes and organizational decision makers [5]. The final risk score is not enough for a governance framework to be ERP oriented. Component-level risk signals and applicable control logic, human review requirements and reasons for dispositioning should be captured in the decision record. The distinction between trustworthy AI and acceptable risk indicates that confidence in a system does not determine whether a specific use or residual risk should be permitted [6]. In turn, governance needs to be implemented as context-specific decision control in order to be understood not as an abstract model property.

Structured oversight is strengthened by the regulatory and assurance viewpoints. From deployment to documentation, monitoring, and corrective actions, the emphasis on responsible AI use moves beyond pre-market testing to the entire use lifecycle within conformity assessment and post-market monitoring frameworks [7]. The case studies in health care implementations also reveal a need for validation pathways, integration with workflows, human oversight, and the need to manage reliability and safety risks as they are explicitly mentioned [8]. While the contexts of ERP environments vary, they share the need for decisions to be taken through traceability and escalation to be controlled, with outputs from artificial intelligence accepted based on evidence. These requirements are used to facilitate a layered architecture, coordinating specialized risk components by using transparent rules, instead of a single opaque classifier.

As far as the authors know, few studies have been performed so far to combine empirically ERP transaction context, prompt-injection risk, AI-output hallucination risk, governance disposition and structured audit reasons in one single computational pipeline. The present study fills this gap by providing a proof-of-concept, script-based framework that utilizes separate machine learning (or rule-based) components that evaluate ERP transaction risk, prompt risk, and output risk, followed by a deterministic governance component that determines Approved, Human Review or Rejected outcomes. Human Review is not just an informal check, but is represented as a governable state. It is a design that aligns with the belief that the agency between human and AI is spread out in the technical outputs, organizational procedures, and decision authority [9].

The study is conceptually limited to governance decisions related to generative-AI enabled ERP scenarios and methodologically limited to secondary datasets, derived ERP labels through rules, supervised classifiers, synthetic scenarios for ERP integration, built reference decisions and structured audit records. The regulated ERP context is control oriented, referring to the demands for and characteristics of enterprise processes in a regulated context, not the evaluation in a "live" regulated organization. The study cannot offer a legal certification, production readiness, a cryptographic audit of immutability, factual support for the results on the operation ERP databases, and universal generalization for all industries, jurisdictions, platforms, and models.

In these confines, this study has three contributions to make. In theory, it posits AI governance as a multi-signal decision-making problem that depends on the context, such as the conditions of the transaction, prompt integrity, risk of the output, and the possibility of accountability escalation. In terms of methodology, it shows how to combine multiple and different datasets and classifiers into an end-to-end experimental framework using safety measures as well as standard predictive ones. In reality it gives an understandable method to

document component forecasts, final outcomes, human audits and structured reasons for audits. To this end, the study has three goals: First, to identify use cases for the development of computational components for classification of transaction risks in ERP systems, second, to develop governance components for prompt-injection detection for prompt-engineering governance, and third, to develop use cases for the assessment of the behavior of the components and the framework using constructed reference decisions, safety-oriented metrics, error decomposition, ablation analysis, and audit-record completeness.

2. LITERATURE REVIEW

The study of enterprise governance with regard to generative artificial intelligence can be divided into four related topics: prompt design, output reliability, machine-learning safety, and institutional oversight. The importance of instructions, demonstrations, templates, and contextual cues for promoting model behavior has been established from a variety of prompt scholarship studies [10]. Much of this literature, however, has focused on the quality of the task performance and quality of the response, rather than enterprise control requirements like authorization, timely provenance, injection resistance, misuse prevention, and escalation. In an ERP environment, prompts can seek access to potentially sensitive process information, try to circumvent organizational security measures, or prompt actions that go against the authority of the organization. Prompt engineering thus offers a valuable technical underpinning but is not prompt governance in itself. For regulated environments, the governance process should also consider the authorization, traceability, consistency and appropriateness of an instruction for downstream use.

A second stream looks at hallucination, inconsistency and output faithfulness. The evaluation and the behavior of the model are still highly variable, and reviews of natural-language generation indicate that the problem of hallucination can be caused by data limitations, model behavior, inference procedures, and the evaluation design [11]. There are methods to control the generation of data-to-text at the word level that can increase the faithfulness at the local level, but these methods require task-specific supervision and do not ensure that the output is faithful to a set of external business rules [12]. While natural-language-inference (NLI) methods offer a scalable way to find inconsistencies, there may be cases of numerical discrepancies, approval breaches, unsupported transaction references, and ERP policy violations that are not detected by NLI [13]. It has also been noted that there can be conceptual critiques that the outputs may be informative, but without truth-directed processes [14]. These studies do not necessarily create a definitive answer to whether generated claims are correct in comparison to authoritative ERP records, ERP approval rules, transaction values, or supporting documents, but they do provide support for AI-output validation including hallucination-risk and consistency assessment.

The third is the theme of risk-based regulation, human oversight and machine-learning safety. Fragmentation of obligations by application context and by the degree of risk, but not enough on how to combine transaction, prompt and output signals into a single decision mechanism [15]. There is a collaborative dimension of reciprocal human-machine learning, and the focus is not on disposition, formal governance or on audit traceability, but primarily on the collaborative performance [16]. The concept of machine-learning safety taxonomies goes beyond accuracy to robustness, monitoring, uncertainty and failure management [17]. Socio-technical safety research also relates to vulnerabilities and coordination issues in the organization, but tends to be more structural in nature than focusing on executable controls to address ERP-specific prompt and output issues [18].

Previous ERP governance studies have been centered around transaction controls, anomaly detection, process conformance, segregation of duties, authorization and auditability. Usually these mechanisms are intended for enterprise records that are structured and have specific business processes, and not natural language prompts and generative outputs. To resolve this governance issue, a proven logic for ERP-control must be linked with timely threats, output in-less reliability, escalation needs, and human decision-making powers.

The institutional governance scholarship extends the analysis to consider legitimacy, accountability, coordination and enforcement. The research literature on global governance of AI reveals tensions between innovation and regulatory control, and between the regulatory control and institutional capacity, but is too abstract to provide any clear sense of how transactions-level can be managed more effectively [19]. However, recent efforts on generative-AI governance highlight the importance of distributed accountability, adaptive regulation, and lifecycle oversight of generative-AI systems, but lack empirical evidence of how to connect disparate technical risk signals into operating decision pipelines focused on ERP systems [20].

Thus, there are no models or methods lacking in the literature to prompt, evaluate hallucinations, analyse risks, control ERP, or govern institutions. There are, however, no separate models or methods lacking in the literature to prompt, assess hallucinations, analyse risks, control ERP, or govern institutions. The more specific gap is how the ERP transaction context, prompt-injection risk, AI-output validation operationalized by the hallucination-risk assessment, governance disposition, and structured audit reasons are not integrated in a single computational framework. This is addressed by the current research work that proposes a proof-of-concept, multi-layer pipeline, which comprises of three specialized pipeline components and a deterministic decision engine. Its added value is in integration, clear escalation, component error attribution, and the ablatability assessment, not in assuring regulatory compliance, production readiness, or fact checking for live

ERP systems.

3. METHODOLOGY

3.1 Research Design and System Development Strategy

The study was based on the design-science approach and computational experiments to build and test a proof-of-concept governance artifact for generative-AI-enabled ERP settings. This method was appropriate for the goal of creating and evaluating an implementable system as opposed to estimating an organizational relationship that naturally occurs. The framework consisted of three layers of risk assessment, a deterministic decision making engine and an audit record generator. The synthetic ERP-prompt-output was used as an experimental unit, which is fed into the pipeline. The analytical domain included enterprise workflow-related ERP transactions, process events, prompts, and outputs generated. It was an evaluation of controlled pipeline behavior, not production performance or regulatory conformity, since relying on secondary data and creating scenarios.

3.2 Data Sources and Dataset Integration

There were three secondary data sets. In this paper, the ERP component uses around 1.46 million events from the procurement-process events of BPI Challenge 2019 [21]. The AI-output validation was carried out using valid and hallucinated responses from a set of labelled responses provided by HaluEval [22] for hallucination-risk assessment. For the prompt component, 662 examples were taken from the Prompt Injections dataset by deepset [23] with 399 safe prompts and 263 injection prompts. Data were programmatically loaded, inspected, harmonized and processed.

The datasets were created for other purposes and were not obtained as transaction-prompt-response pairs. To evaluate end-to-end, 5,000 synthetic ERP-GenAI scenarios were created, with one context, prompt, and output from ERP for each scenario. A total of 1,250 scenarios were obtained for each ERP-risk class using stratified sampling. Some prompt/out instances may repeat due to the vast difference in size of the datasets. The benchmark did not reflect typical workflows or operational risk occurrences.

3.3 Data Preparation, Feature Engineering, and Target Construction

Categorical ERP variables were normalized, missing values were explicitly represented, transaction amounts were discretized and process-stage indicators were created from activity fields. The targets for ERP were Low, Medium, High or Critical and were established according to preset rules for transactions and process control, and not based on the actual compliance results.

Prompt text was encoded with term-frequency-inverse-document-frequency features, with the labels on the different texts used to indicate whether they were safe or injection-oriented prompts. Text was represented using word-level and character-level TF-IDF vectors, along with engineered linguistic features, with labels indicating the source of the text, valid or hallucinated. Table 1 provides an overview of operational definitions used for the constructs of the main study.

Table 1. Operational definitions of study constructs

Construct	Operational definition
ERP risk	Rule-derived process or transaction exposure categorized as Low, Medium, High, or Critical
Prompt-engineering governance risk	Safe or injection-oriented prompt classification
AI-output validation risk	Valid or hallucinated output classification
Reference decision	Governance outcome derived from source labels and predefined decision rules
Unsafe approval	Reference-unsafe scenario assigned Approved

The framework separated ERP risk, prompt-engineering governance risk, AI-output validation risk, reference decisions, and unsafe approvals into distinct operational entities as seen in Table 1. These constructs indicated different aspects of risk and should not be viewed as synonymous measures of formal regulatory violations.

3.4 Component-Level Risk Models

To compare the performance of the variants of the Random Forest along with Extra Trees on the multiclass prediction using transaction and process features in the ERP layer. Logistic Regression and Linear Support Vector Machine (SVM) models were evaluated for sparse text by the prompt layer. The output layer tested classical classifiers with the lexical features and engineered features to estimate hallucination-related risk. Where applicable, stratifications for training and test partitions were applied.

The candidate configurations were evaluated with both aggregate and class-specific metrics, and the selected ones were considered as proof-of-concept configurations. The analysis was not used to determine the global optimum. Special attention focused on recall of injection, recall of hallucination, and Critical ERP-risk recall,

due to inability to identify false negatives for downstream decisions.

3.5 Integrated Governance Decision Framework

Governance-risk categories were assigned to the component predictions using pre-defined, deterministic mappings. The classes Low, Medium, High and Critical remained the same for ERP predictions. Safe prompts were assigned to the category of Low risk and Injection prompts were assigned to the category of Critical risk. The outputs that were valid were mapped to the category Low risk while Hallucinated outputs were mapped to the High risk. These mappings were not learned from organizational decision records, but were preset. A hierarchical rule engine then was used to assign one of three outcomes: Approved, Human Review, or Rejected. Human Review conditions took precedence over Rejection conditions which took precedence over approval. A hierarchy of decisions was created and summarized in Table 2.

Table 2. Governance decision rules

Priority	Component condition	Final decision
1	Prompt classified as Injection and mapped to Critical risk	Rejected
2	Output classified as Hallucinated and mapped to High risk	Rejected
3	ERP risk classified as High or Critical	Human Review
4	No rejection or review condition identified	Approved

Table 2 reveals that prompt-injection / hallucination-risk conditions resulted in rejection, while High / Critical ERP risk resulted in Human Review. Approval was only granted if a rejection or review condition was not identified. The engine did not learn rule priorities from organizational decisions, but used already given priorities. There were unique identifiers, component predictions, mapped risks, final disposition, human review indicator, applied rule, and audit reason in each record. The behavior in the event of model failure or missing evidence was not assessed separately.

3.6 Evaluation and Statistical Analysis

The performance of the components was evaluated based on accuracy, precision, recall, F1-score, macro F1, balanced accuracy, confusion matrices. Prompt-injection, hallucinated-output and Critical ERP-risk detection were measured by class-specific recall.

The precision of end-to-end performance was measured by exact-agreement, unsafe approval, unsafe detection, reference-Rejected interception, safe approval and Human Review exact-match. Exact agreement was 3,848/5,000 (76.96%). An unsupported scenario was created with a decision of Human Review or Rejected. Unsafe approval was 335/3,677 (9.11%), whereas unsafe detection was 3,342/3,677 (90.89%). Reference-Rejected interception was 2,014/2,318 (86.89%) for Reference-Rejected scenarios that were Reviewed as Rejected or Human Review. Safe approval was 1,042/1,323 (78.76%), and Human Review exact-match precision was 1,089/1,386 (78.57%).

Since Reference-Rejected interception was done with a smaller subset, it could not be interchangeable with unsafe detection. The estimates of uncertainty intervals were derived via bootstrap resampling and unsafe approvals were traced back to missing risks in ERP, attacks (when they did occur) or hallucinated outputs. These measures were for agreement with labels, and for pre-defined logic, not for operational regulatory compliance.

3.7 Validation, Ablation, and Auditability Assessment

The ablation experiments were conducted one by one, and compared the results of exact agreement and unsafe approval with the full structure, a single-model baseline, and an oracle upper bound with ground-truth labels. The oracle did not exist in practice and had no prediction error.

The extent to which identifiers are unique, fields are complete, the explanation is covered, rules are covered and consistency between the predicted risk and disposition were measured for auditability. These measures were structured audit recording completeness and not immutability, tamper resistance, access control, or admissibility in court. An instrument for further research was developed with the help of the expert, but it was not used as an empirical instrument. Informed consent was not applicable because only secondary data (public or de-identified) were used, and no participants were recruited.

4. Results

4.1 Dataset and Scenario Characteristics

The pipeline combined ERP context, prompt risk, and AI-output validation through hallucination-risk assessment. The ERP source contained approximately 1.46 million procurement-process events. The prompt dataset comprised 662 labelled instances, including 399 safe prompts and 263 injection examples, while the output dataset contained labelled valid and hallucinated responses. For the integrated evaluation, 5,000 synthetic ERP-GenAI scenarios were constructed, with 1,250 scenarios assigned to each ERP-risk category. The

composition of the source datasets and their representation in the integrated benchmark is summarized in Table 3.

Table 3. Dataset composition and synthetic integration benchmark

Component	Source size	Integrated representation
ERP context	Approximately 1,463,017 events	5,000 scenarios
Prompt assessment	662 prompts	5,000 prompts
Output validation	Labelled response corpus	5,000 outputs
ERP-risk strata	Four classes	1,250 per class

Table 3 shows that ERP contexts, prompts, and outputs were computationally aligned within a common 5,000-scenario benchmark. The integrated sample contained 3,027 safe prompts, 1,973 injection prompts, 3,665 valid outputs, and 1,335 hallucinated outputs. These scenarios were synthetic analytical composites rather than naturally observed ERP-prompt-response sequences.

4.2 ERP Risk-Assessment Performance

The ERP-risk classifier achieved a macro F1-score of 0.6582 and an overall accuracy of 74.20%. All 1,250 Low-risk cases were classified correctly, as were 1,225 Medium-risk and 1,231 High-risk cases. By contrast, only four of the 1,250 Critical cases were correctly identified; the remaining Critical cases were predominantly assigned to the High category, with smaller numbers classified as Low or Medium. The class-specific distribution of predictions is presented in Figure 1.

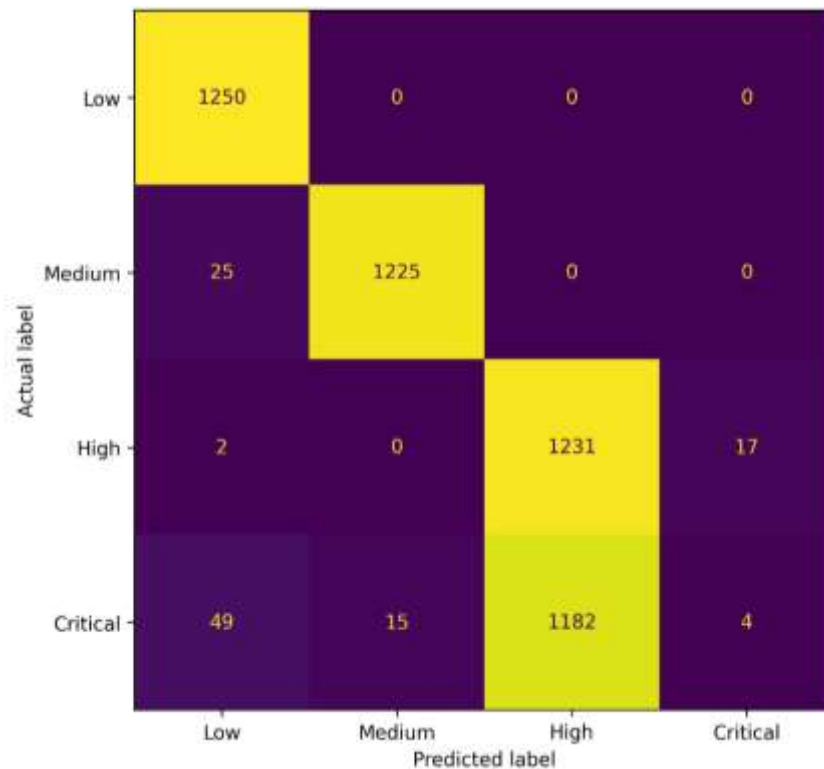


Figure 1. Confusion matrix of ERP-risk classifications on the synthetic integration benchmark

Figure 1 shows that classification error was concentrated in the Critical-risk class. Most Critical cases were predicted as High, indicating that the classifier often preserved an elevated severity signal but did not reliably distinguish the most severe category.

Critical recall was 0.32%, compared with 100.00% for Low, 98.00% for Medium, and 98.48% for High. Overall, 3,710 predictions were correct, and only 21 instances were assigned to the Critical class. These results indicate strong discrimination for the lower three classes but very limited sensitivity to Critical-risk cases.

4.3 Prompt-Engineering Governance Performance

The prompt-engineering governance was evaluated based on the binary prompt-injection classification. The accuracy and macro F1-score of the classifier were 76.34% and 0.7042, respectively. 100% of the 3,027 safe prompts were correctly identified. Of the 1,973 injection prompts, 790 were detected, and 1,183 were deemed to be safe. The resulting true/false prediction distribution is as presented in Figure 2.

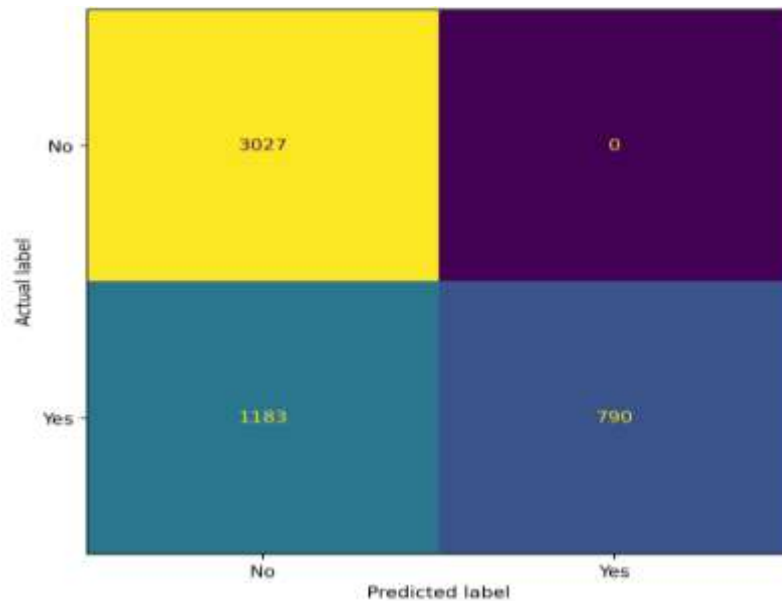


Figure 2. Confusion matrix of prompt-injection classifications on the synthetic integration benchmark

Figure 2 illustrates that the classifier identified a significant amount of injection prompts but failed to classify over half of them. Meanwhile, no safe prompt was wrongly classified as an injection, yielding a very skewed error pattern.

The results were 100.00% for precision injection and 40.04% for injection recall. The result was a pattern without any false positives and a 59.96% false negative rate. The results thus show that the classifier was very conservative in assigning the label injection, and that a lot of the classification errors were due to undetected attacks.

4.4 AI-Output Validation Performance

AI-output validation was operationalized as hallucination-risk classification. On the synthetic integration benchmark, the classifier achieved 82.94% accuracy and a macro F1-score of 0.8041. Of the 3,665 valid outputs, 2,972 were correctly classified and 693 were labelled hallucinated. Of the 1,335 hallucinated outputs, 1,175 were detected and 160 were missed. The confusion structure for this benchmark is shown in Figure 3.

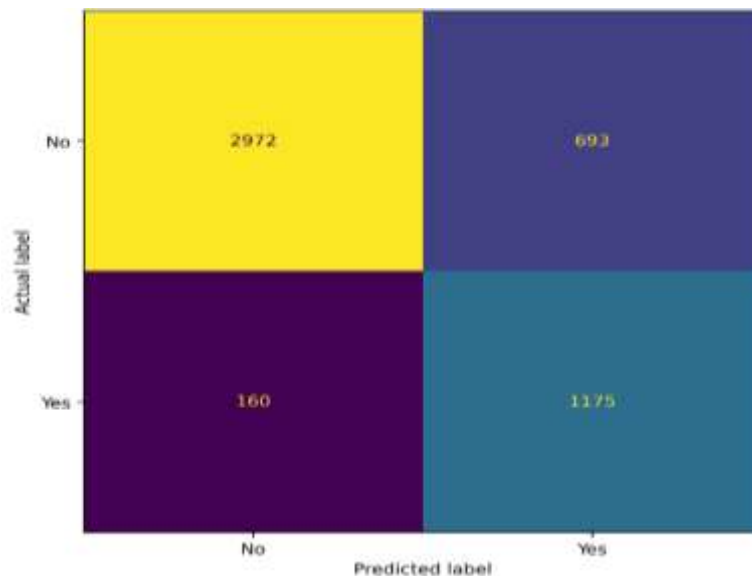


Figure 3. Confusion matrix of AI-output hallucination-risk classifications on the synthetic integration benchmark

Figure 3 shows that the classifier detected most hallucinated outputs while also generating a noticeable number of false-positive hallucination predictions. The resulting pattern indicates stronger sensitivity to hallucinated content than specificity to valid outputs.

Hallucination recall was 88.01%, specificity was 81.09%, and hallucination precision was 62.90%. A separate holdout evaluation produced 72.28% accuracy, 63.72% balanced accuracy, a macro F1-score of 0.6081, and

50.31% hallucination recall. Because the benchmark and holdout evaluations differed in composition and conditions, their estimates are not interchangeable. The integrated result was therefore benchmark-specific and did not demonstrate factual verification against ERP records or business documents.

4.5 Integrated Governance Decision Outcomes

The governance engine classified 1,377 scenarios as Approved, 1,386 as Human Review and 2,237 as Rejected. Built-in reference decisions were created from source labels and governance rules that have been defined and these outcomes were compared with the built-in reference decisions. There were 3,848 agreements and 1,152 disagreements, with 76.96% of the time being spent on agreement. This measure reflected the consistency with the experimental governance logic as opposed to agreement with independently observed organizational decisions.

The total number of reference unsafe scenarios is 3,677, the number of which were incorrectly marked Approved is 335 giving an unsafe approval percentage of 9.11%. The remaining 3,342 scenarios were deemed unsafe, either because they were deemed unsafe by the humans that have been given scenarios, or because they were rejected. The rest of the scenarios were deemed unsafe by the humans that have been given scenarios (3,342) or as Rejected (0 cases), giving an unsafe detection rate of 90.89%. Among those scenarios where the constructed reference decision was Rejected (2,318), 2,014 (86.89%) were given a Human Review or Rejected decision. This measure was not unsafe detection since only reference-Rejected scenarios were taken into account.

Of the 1,323 scenarios, 1,042 were assigned to Approved and correctly (78.76%) assigned to Approved. The 1,386 scenarios that were assigned to Human Review resulted in 1,089 scenarios that matched the Human Review state that was constructed, which gives a Human Review exact-match precision of 78.57%. The main governance arrangements for the end-to-end are outlined in Table 4.

Table 4. End-to-end governance performance

Measure	Result
Exact agreement with constructed reference decisions	76.96%
Unsafe approval rate	9.11%
Unsafe detection rate	90.89%
Reference-Rejected interception rate	86.89%
Safe approval rate	78.76%
Human Review exact-match precision	78.57%

Table 4 shows that the framework intercepted most reference-unsafe scenarios, although 9.11% were still approved. The bootstrap confidence interval for the unsafe approval rate was 8.24%–10.04%. These measures reflected agreement with constructed labels and predefined governance logic rather than regulatory certification or operational compliance performance.

4.6 Error Analysis and Safety Failures

Error decomposition identified 335 unsafe approvals. Missed prompt attacks accounted for 257 cases, or 76.72% of the total. Missed hallucinated outputs contributed 47 cases, or 14.03%, while missed High or Critical ERP risks accounted for 31 cases, or 9.25%. The distribution of unsafe approvals by error source is presented in Table 5.

Table 5. Sources of unsafe approvals

Error source	Number	Percentage
Missed prompt attacks	257	76.72%
Missed hallucinated outputs	47	14.03%
Missed High or Critical ERP risks	31	9.25%
Total	335	100.00%

Table 5 shows that prompt-level false negatives were the dominant source of unsafe approvals, accounting for more than three-quarters of all such cases. Output- and ERP-related errors contributed smaller but still material proportions.

Decision disagreements included both under-escalation and over-escalation. Under-escalation occurred when scenarios requiring Human Review or Rejection were assigned Approved, whereas over-escalation occurred when approvable scenarios were assigned Human Review or Rejected. Only under-escalation to Approved contributed to the unsafe approval rate.

4.7 Ablation, Baseline, and Auditability Results

The omission of any governance level resulted in a decrease in end-to-end performance. If the ERP layer is not

used, the agreement rate drops to 55.18%, while the number of unsafe approvals rises to 46.80%. The agreement in the removed prompt assessment was 69.58%, and the unsafe approval rate was 14.60% when removing output assessment, the agreement was 68.70%, and the unsafe approval rate was 22.63%. The overall system performance had 76.96% exact agreement, 90.89% unsafe detection rate and 9.11% unsafe approval rate. The comparison between the performance of the full framework, ablated variants and baseline configurations is provided as Figure 4.

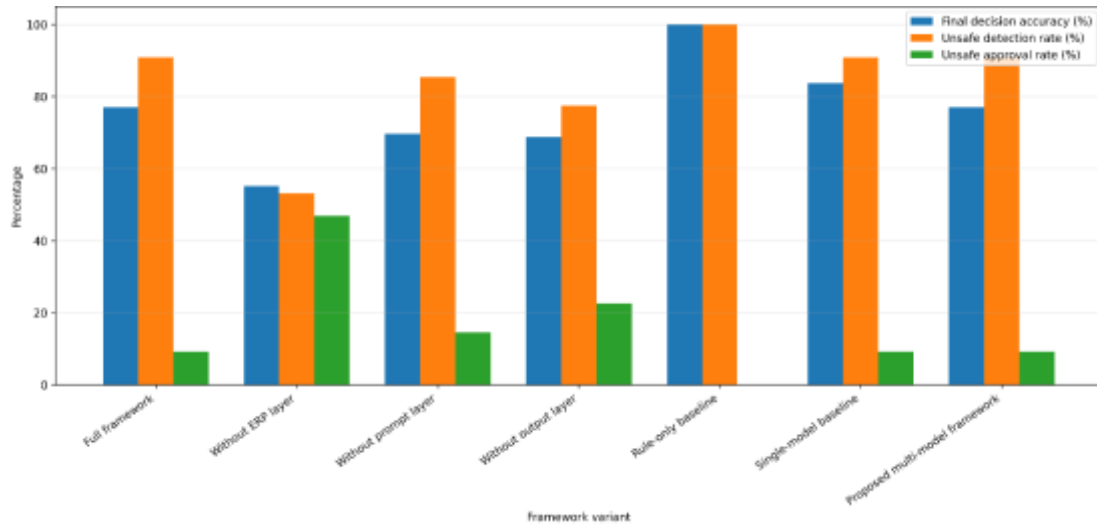


Figure 4. Comparison of governance-layer ablations and baseline configurations

The results of removing the ERP layer are shown in figure 4, and is the largest drop in safety performance. The prompt and output layers also gave complementary governance signals, with removing them leading to more unsafe approval rates than the whole system.

The single-model baseline had a higher rate of exact agreement (83.74%) than the entire framework, and a nearly identical (9.22%) unsafe approval rate compared to the framework (9.11%). A statistical analysis was not performed to determine if this difference was significant. The upper bound of the oracles with respect to the ground-truth component labels yielded 100.00% exact agreement and 0.00% unsafe approvals, while also having 0.00% unsafe detections. It was a theoretical upper bound as it used known ground truth component labels instead of predicted component outputs.

A total of 5,000 unique records were found with all required fields, explanations, and decision rules filled in for the auditability assessment. This 100% result indicates that structured audit-record completeness was achieved in the benchmark that was generated, but did not prove immutability, tamper resistance, access control or legal admissibility. There was no validation by experts of the results.

5. Discussion

The results show that the interaction between the transaction context, the integrity of the prompt, the risk of the output, and the escalation logic of the transaction are the factors that influence the performance of governance in generative-AI-enabled ERP environments, and not any one classifier. The framework reproduced most of the constructed reference decisions, but maintained an unacceptable residual rate of unsafe approval, indicating that there is not only a need for a rate of agreement to be the sole assurance criterion, but a need for it to be a very high percentage. Partly, end-to-end agreement was assessed by reproducing the decision logic of the experimented sources instead of by matching with any independently made judgments of the organization, due to the fact that they drew their references from the source labels and pre-defined governance rules. However, the unsafe approval results themselves are still relevant to the operation as even one or two of these false approvals might allow for unsupported, manipulated or contextually inappropriate outputs. This pattern helps to make the governance, model, and application-level behavior auditable separately, but operationally integrated [4].

A primary source of residual risk was found to be prompt level false negatives, according to the error distribution. In a more conservative manner, the prompt classifier detected no false-positive injection alerts, but did lead to significant under-detection of attacks. This is an issue because a control may only block a copy of the legitimate prompt, but might not be able to block a copy of an adversarial prompt if it is not detected. This extends accountability research in this area by demonstrating that there is also a need to be able to trace responsibility for an unsafe disposition back to the upstream component that influenced the decision path, that is, to the decision maker that made the prediction [5].

The dedicated holdout evaluation showed lower performance for the AI-output validation layer (or hallucination-risk assessment). This is a discrepancy, suggesting sensitivity to experimental conditions and/or

to the composition of a dataset. The integrated result should thus not be seen as a stable factually verified result across ERP contexts. The component identified as textual hallucination risk did not check the generated claims against authoritative transactions, numerical transactions, supporting documents, and organizational policies. As such, it is based on output-risk screening and not full factual validation.

The ERP-risk failed with a different failure pattern. A high rating (most Critical cases) meant that the model often failed to differentiate between the most critical case and maintained an elevated risk classification. Many of the downstream exposures were contained because in most cases, High and Critical predictions led to escalation. Weak Critical recall limits, however, prioritization, urgency assignment and allocation of review resources. This discovery aligns with the notion that human-AI agency is distributed throughout the technical forecast, organizational process, and decision authority, specifically when Human Review serves as an institutionalized checking mechanism as opposed to an informal corrective mechanism [9].

The ablation results additionally show that the three layers provided complementary governance signals. The layer of the ERP removal resulted in the highest level of unsafe approvals and omitting either the prompt or output layer also had a negative impact on safety performance. The single-model baseline resulted in higher exact agreement with the complete framework, and a similar (but slightly higher) number of unsafe approvals. This small difference was not statistically significant with any inferential comparison made. This finding, however, suggests that predictive compactness and the quality of governance are not one and the same. It also is congruent with the research that has demonstrated that output reliability is not merely expediency in surface fluency or the ability to perform generic classifications [11]. In theory, the study expands the definition of enterprise AI governance beyond a single-signal perspective to one that includes both risk context and risk accumulations, and considers risk as a function of the interactions of specialized controls. On the practical level, the framework offers a clear means to identify failure triggers, elevate unclear instances and keep structured audit causes while maintaining the traceability of individual risk signals.

There are some limitations in the interpretation and generalization. ERP labels were created based on pre-defined rules, not upon actual compliance results. The scenarios were naturalistic, but based on naturalistic classes and computed sequences rather than natural ERP-prompt-response sequences. There may have been a number of source instances that were repeated and/or prompt templates that reinforced pattern recognition. In the output component, linguistic hallucination risk and not evidence-based correctness was evaluated. Constructed reference decisions were based on source labels and pre-defined governance logic and audit completeness was based on fields populated, not cryptographic immutability, access control, tamper resistance or legal admissibility. But validation by experts was not completed. The restrictions are based on already noted machine-learning safety issues of distribution shift, monitoring and failure containment [17]. The framework should therefore not be taken as evidence of conformity to any specific regulation, standard or legal requirement but rather as compliance oriented.

Case level temporal partitions, independent validated ERP-risk labels, unseen prompt-attack families, multilingual and indirect injections, and semantically coherent prompt-response pairs, based on authoritative transactions should be utilized in future work. Retrieval-based evidence, numerical reconciliation, policy checks, provenance, and claim level checks are some of the elements that should be included in AI-output validation. Future research should investigate reviewer behavior, escalation criteria, drift in the models, industry applicability, and longitudinal safety outcomes in real-world ERP use. These extensions would take the framework beyond controlled proof of concept evaluation and lifecycle governance of technical controls and adaptive institutional control [20].

6. Conclusion

In this study, a proof-of-concept, script-based, intelligent governance and compliance framework was designed and tested, which combines enterprise resource planning (ERP) transaction-risk classification, prompt-injection detection, AI-output validation by hallucination-risk assessment, deterministic decision control and structured generation of audit records in a single pipeline. The results demonstrate that layered governance can detect most of the reference defined unsafe scenarios, and it can generate transparent Approved, Human Review, and Rejected outcomes. However, some residual risk persisted for missed prompt attacks, the lack of discrimination between a High and Critical ERP case and the output performance characteristic of the dataset. Theoretically, enterprise AI governance is not a single, standalone classification problem, but rather one that relies on multiple signals and is tied to the specific context. In practice, it provides a traceable format for representing the components' failures, progressing to uncertain cases and recording rationales for decisions. The study has been conducted in an experimental setting, not a production context, to illustrate how ERP can be synthesized and validated through a training process based on rule-generated ERP, in addition to factual validation of ERP in real systems. In the lead up to operational deployment, organisations should increase their adversarial prompt detection, independently validate ERP-risk labels, add output checking with evidence, ramp up their escalation thresholds, and ensure that their audit controls are secure and tamper-proof. Human Review should be formalized by putting it under a person's responsibility, formally recording overrides, service-level requirements and feedback mechanisms. As for future work, it is recommended to utilize real ERP transactions, temporal case splits, multilingual attacks, authoritative grounding and independent expert and

field validation. Model drift, model transferability, reviewers' behavior, and longitudinal safety should also be evaluated in such studies.

References

1. P. Budhwar, S. Chowdhury, G. Wood, H. Aguinis, G. J. Bamber, J. R. Beltran, et al., "Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT," *Human Resource Management Journal*, vol. 33, no. 3, pp. 606–659, 2023.
2. L. Banh and G. Strobel, "Generative artificial intelligence," *Electronic Markets*, vol. 33, no. 1, Art. no. 63, 2023.
3. E. A. van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "ChatGPT: Five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.
4. J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: A three-layered approach," *AI and Ethics*, vol. 4, no. 4, pp. 1085–1115, 2024.
5. C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: What it is and how it works," *AI & Society*, vol. 39, no. 4, pp. 1871–1882, 2024.
6. J. Laux, S. Wachter, and B. Mittelstadt, "Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk," *Regulation & Governance*, vol. 18, no. 1, pp. 3–32, 2024.
7. J. Mökander, M. Axente, F. Casolari, and L. Floridi, "Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation," *Minds and Machines*, vol. 32, no. 2, pp. 241–268, 2022.
8. M. M. Raza, K. P. Venkatesh, and J. C. Kvedar, "Generative AI and large language models in health care: Pathways to implementation," *NPJ Digital Medicine*, vol. 7, no. 1, Art. no. 62, 2024.
9. S. Krakowski, "Human-AI agency in the age of generative AI," *Information and Organization*, vol. 35, no. 1, Art. no. 100560, 2025.
10. P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, Art. no. 195, 2023, doi: 10.1145/3560815.
11. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
12. C. Rebuffel, M. Roberti, L. Soulier, G. Scouteeten, R. Cancelliere, and P. Gallinari, "Controlling hallucinations at word level in data-to-text generation," *Data Mining and Knowledge Discovery*, vol. 36, no. 1, pp. 318–354, 2022.
13. P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 163–177, 2022.
14. M. T. Hicks, J. Humphries, and J. Slater, "ChatGPT is bullshit," *Ethics and Information Technology*, vol. 26, no. 2, pp. 1–10, 2024.
15. T. D. Krafft, K. A. Zweig, and P. D. König, "How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications," *Regulation & Governance*, vol. 16, no. 1, pp. 119–136, 2022.
16. D. Te'eni, I. Yahav, A. Zagalsky, D. Schwartz, G. Silverman, D. Cohen, et al., "Reciprocal human-machine learning: A theory and an instantiation for the case of message classification," *Management Science*, vol. 72, no. 1, pp. 167–192, 2026.
17. S. Mohseni, H. Wang, C. Xiao, Z. Yu, Z. Wang, and J. Yadawa, "Taxonomy of machine learning safety: A survey and primer," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, 2022.
18. A. Gazos, J. Kahn, I. Kusche, C. Büscher, and M. Götz, "Organising AI for safety: Identifying structural vulnerabilities to guide the design of AI-enhanced socio-technical systems," *Safety Science*, vol. 184, Art. no. 106731, 2025.
19. J. Tallberg, E. Erman, M. Furendal, J. Geith, M. Klamberg, and M. Lundgren, "The global governance of artificial intelligence: Next steps for empirical and normative research," *International Studies Review*, vol. 25, no. 3, Art. no. viad040, 2023.
20. A. Taeihagh, "Governance of generative AI," *Policy and Society*, vol. 44, no. 1, pp. 1–22, 2025.
21. B. F. van Dongen, "BPI Challenge 2019," 4TU.ResearchData, Eindhoven University of Technology, Jan. 31, 2019. [Data set]. doi: 10.4121/uuid:d06aff4b-79f0-45e6-8ec8-e19730c248f1.