IJECS Open Access https://ijecs.in
International Journal of Engineering and Computer Science
Volume 15, Issue 07, July 2026, Page No. 28669-28677
ISSN: 2319-7242 DOI: 10.18535/ijecs.v14i07.5564

From Numbers to Narratives: A Generative Artificial Intelligence Framework for Automated Risk Report Production in IT Program Portfolio Governance

Nkem Daniel Obaloje[1], Ezeokechukwu Chiemere Victor[1]

[1]*Department of Computer Science, School of Computing and Engineering Sciences Babcock University, Ilishan-Remo, Ogun State, Nigeria*

Abstract:
Every IT program portfolio generates a continuous stream of quantitative performance data. What it rarely generates, and what governance institutions consistently demand, is a coherent written account of what those numbers mean, what risks they signal, and what program boards should do about them. Human risk analysts currently bridge this gap, translating earned value metrics and delay risk flags into governance narratives manually, at considerable cost, with variable quality, and at a pace that often lags the reporting cycle. This paper introduces PRISM, the Predictive Risk Intelligence with Synthesized Management narratives framework, a generative artificial intelligence pipeline that automates the production of professional, governance-ready risk narrative reports by combining XGBoost delay risk classification, TreeSHAP feature attribution, and a fine-tuned large language model trained on a curated corpus of 2,340 authentic program risk reports drawn from federal IT investment governance archives. The framework accepts a standard program performance data vector as input and produces a complete, contextually accurate, and institutionally appropriate governance risk narrative as output, indistinguishable in professional quality from reports produced by experienced human risk analysts. Evaluation on a held-out test set of 412 investment records demonstrates that PRISM achieves a BLEU-4 score of 0.743, a BERTScore F1 of 0.891, and a practitioner Turing evaluation pass rate of 74.6 percent, meaning that in nearly three out of four cases, certified program management practitioners rated PRISM-generated narratives as of equal or superior quality to human-authored equivalents without being able to identify their machine origin. A governance decision quality study with 56 active program board members confirms that decisions made using PRISM narratives are 31.4 percent more likely to target the correct primary risk driver than decisions made using raw performance dashboards, and that PRISM narratives reduce mean risk report production time from 4.2 hours to 11 minutes per investment without loss of governance utility. These findings establish generative AI narrative synthesis as a technically feasible, practically superior, and institutionally deployable alternative to manual risk report production in IT program portfolio governance contexts.

Keywords: Generative Artificial Intelligence, Large Language Models, Program Risk Reporting, XGBoost, SHAP, Natural Language Generation, IT Portfolio Governance, Earned Value Management..

1. Introduction

There is a persistent and costly gap at the center of IT program portfolio governance. On one side sits a growing volume of quantitative performance intelligence: earned value metrics, schedule variance computations, machine learning generated delay risk probabilities, and feature attribution scores that identify, with increasing precision, which investments are heading toward delivery failure and why. On the other side sits the governance board: a group of senior officials, procurement officers, and program sponsors who need to understand that intelligence, act on it within the constraints of institutional accountability frameworks, and communicate their decisions in professional written form. Between these two sides stands a human risk analyst, translating numbers into narratives, and the quality of that translation determines whether the intelligence actually reaches the people who need it [1], [2].

The analyst translation process is expensive, slow, and inconsistent. A typical program office managing twenty to thirty concurrent IT investments produces between forty and sixty individual risk assessment narratives per reporting cycle, each requiring a qualified analyst to interpret performance data, cross-reference prior period reports, apply professional judgment about severity and priority, and draft a governance-appropriate written account [3]. The cost of this

labor is substantial. The time required frequently exceeds the reporting cycle itself, meaning that narratives describing performance data from month one reach governance boards at month two, when the performance situation has already evolved [4]. And the quality varies systematically with the analyst's experience, workload, and familiarity with the specific investment being assessed, producing governance reports whose reliability is inversely correlated with portfolio size [5].

Natural language generation and, more recently, large language models have demonstrated the capacity to produce professional-quality written text across a wide range of domains, including legal document drafting, clinical report generation, financial commentary, and scientific abstract writing [6], [7], [8]. The application of these capabilities to program risk reporting, however, raises specific technical challenges that generic language model deployment cannot address. A governance risk narrative is not free prose. It must be factually grounded in the specific quantitative performance data of the investment being assessed. It must reflect the correct causal interpretation of that data, identifying which performance factors are driving the risk rather than merely noting that risk exists. It must employ the institutional vocabulary of program governance correctly, using terms like schedule performance index, earned value, and baseline deviation in ways that are technically precise and contextually appropriate. And it must communicate an appropriate recommendation for governance action that is proportionate to the severity of the identified risk [9], [10].

Satisfying these requirements demands more than a general-purpose language model prompted with raw performance numbers. It demands a pipeline in which the quantitative intelligence layer and the language generation layer are systematically integrated, with the language model's generation conditioned on structured, attribution-weighted feature information rather than on uninterpreted numerical inputs. This integration is the technical contribution of the PRISM framework introduced in this paper.

PRISM combines three components into a unified narrative synthesis pipeline. The first is an XGBoost delay risk classifier that generates a calibrated risk probability score for each investment from a structured program performance feature vector. The second is a TreeSHAP attribution layer that decomposes each risk prediction into per-feature contribution scores, identifying which performance dimensions are driving the risk flag and to what degree. The third is a fine-tuned large language model that receives the investment context, the risk score, and the ranked feature attributions as structured inputs, and generates a complete, governance-ready risk narrative as output. The language model is fine-tuned on a corpus of 2,340 authentic program risk reports drawn from federal IT investment governance archives, giving it the domain vocabulary, the narrative conventions, and the institutional register of professional program risk communication.

The paper makes four original contributions. First, it introduces the PRISM architecture as the first formally specified pipeline for generative AI risk narrative synthesis grounded in XAI attribution conditioning. Second, it demonstrates through empirical evaluation that PRISM-generated narratives achieve practitioner-evaluated quality comparable to human-authored equivalents on Turing evaluation criteria. Third, it provides the first evidence that governance decisions made using automated AI-generated narratives are significantly superior in intervention targeting precision to decisions made using raw performance dashboards. Fourth, it establishes an evaluation framework, combining BLEU scoring, BERTScore semantic similarity, and practitioner Turing evaluation, that future research on AI-generated governance communication can adopt as a methodological standard.

Section II reviews related literature. Section III presents the theoretical framework. Section IV describes the PRISM architecture. Section V presents the experimental evaluation. Section VI reports the governance decision quality study. Section VII discusses findings and implications, and Section VIII concludes.

2. Related Work

A.Natural Language Generation and Large Language Models in Professional Domains

Natural language generation, the automated production of grammatically and semantically coherent written text from structured data inputs, has a research history extending several decades. Early template-based systems produced serviceable but formulaic outputs that practitioners found easy to identify as machine-generated [11]. The emergence of neural sequence-to-sequence architectures and, more recently, transformer-based large language models has fundamentally changed what automated text generation can achieve. GPT-4 class models have demonstrated the ability to produce coherent multi-paragraph professional text across domains including law, medicine, and finance, with human evaluators frequently unable to distinguish AI-generated outputs from those of trained professionals [7], [8].

In professional reporting contexts, several specific deployments have been validated empirically. Chang et al. [12] fine-tuned a transformer model on a corpus of clinical discharge summaries and demonstrated that the resulting system produced clinical narratives rated by physicians as clinically adequate in 81.3 percent of cases, reducing documentation time by 68 percent without compromising accuracy. Zhou and Liang [13] applied a retrieval-augmented generation architecture to financial risk commentary production and showed that the system generated earnings analysis reports that institutional investors rated as comparably informative to analyst reports from mid-tier research firms. These results establish the technical feasibility of professional narrative generation in high-stakes institutional contexts, though neither study addressed the specific conditioned attribution problem that program risk reporting requires.

B. XAI for Project and Program Risk Communication

The application of explainable AI to project risk prediction has established that XGBoost with TreeSHAP produces both the highest predictive accuracy and the most reliable feature attribution scores among algorithms evaluated on structured program performance datasets [14], [15]. Kobayashi and Alam [16] developed a multi-module XAI framework for project risk management and confirmed that SHAP-based attribution outputs increased program managers' willingness to act on risk predictions, but documented that the technical format of SHAP outputs, specifically bar charts with numerical values, required significant analytical literacy to interpret correctly. Santos [17] evaluated explainable machine learning for project management

control and found that managers with lower quantitative literacy derived significantly less governance value from SHAP visualizations than those with stronger analytical backgrounds, establishing that the format in which attribution information is communicated, not merely its presence, determines its governance utility.

These findings motivate the translation layer that PRISM addresses. If SHAP attribution values are to function as genuine governance intelligence rather than technical artefacts visible only to specialists, they must be rendered in natural language that communicates their governance significance without requiring the reader to interpret numerical scores. Prior work has addressed this through rule-based translation schemas that map feature values to fixed narrative templates [18]. PRISM advances beyond this approach by using a fine-tuned language model to generate contextually varied, investment-specific narratives that adapt their content to the particular combination of risk factors present in each case, rather than selecting from a predetermined template library.

C. AI-Generated Text Evaluation Methodology

The evaluation of machine-generated text quality is a methodologically active area with competing frameworks. BLEU, the Bilingual Evaluation Understudy metric, measures n-gram overlap between generated and reference texts and has been widely applied in natural language generation evaluation despite known limitations in capturing semantic equivalence [19]. BERTScore addresses BLEU's limitation by computing token-level similarity using contextual BERT embeddings, capturing semantic equivalence even when surface vocabulary differs between generated and reference texts [20]. Human Turing evaluation, in which practitioners judge whether a text was produced by a human or a machine, provides an ecologically valid assessment of whether machine-generated outputs meet the professional standards of the domain but is subject to rater variability and fatigue effects [7].

Guo et al. [21] conducted a systematic comparison of automated and human evaluation metrics for professional text generation and found that BERTScore correlated most strongly with expert human quality ratings in technical domains, with BLEU providing a useful but lower-fidelity complement. The PRISM evaluation protocol employs all three methods, using BLEU and BERTScore to establish automated quality benchmarks and Turing evaluation to assess practical professional equivalence, consistent with the multi-method evaluation standard recommended in recent language model assessment literature [22].

D. Risk Communication in Program Governance

The governance literature has long recognized that the quality of risk communication, not merely the accuracy of risk identification, determines whether risk information produces appropriate organizational response [23]. Deficient risk communication produces one of two failure modes: underreaction, in which decision-makers discount risk information that is technically accurate but poorly communicated; and overreaction, in which imprecise communication produces disproportionate responses to manageable risks [24]. Program risk reports in IT governance contexts must therefore satisfy multiple simultaneous communication requirements: technical accuracy with respect to performance data, proportionality in the framing of risk severity, specificity regarding the drivers of identified risk, and actionability in the recommendation of governance response [3], [5].

Existing research on risk communication quality in program governance is primarily qualitative, offering framework guidance rather than empirical measurement [25]. Kutsch and Haass [26] reviewed the literature on AI transformation in project management and identified automated risk communication as among the highest-value and least-developed applications of artificial intelligence to project governance, noting the absence of empirical studies measuring the governance decision quality impact of AI-generated risk communications. PRISM and the evaluation methodology presented in this paper directly address this gap.

3. Theoretical Framework

A. Information Processing Theory and Communication Channel Fitness

Information Processing Theory holds that the utility of information for decision-making depends not only on the accuracy of that information but on the match between the information's format and the cognitive processing requirements of the decision task [27]. Quantitative performance dashboards presenting raw metric values impose high analytical processing demands on governance practitioners, who must perform the translation from numbers to narrative significance themselves before they can form a judgment about risk severity or governance response. Written narrative risk reports reduce this processing demand by performing the translation step prior to communication, presenting conclusions and their rationale rather than the raw data from which conclusions must be derived. The theory predicts, and the governance decision quality study in Section VI tests, that practitioners receiving translated narrative outputs will make higher-quality governance decisions than those receiving equivalent information in untranslated quantitative form.

B. Decision Support Systems Theory and the Last-Mile Problem

Decision Support Systems Theory frames the value of analytical tools in organizational settings as contingent on the degree to which those tools produce outputs that practitioners can act on within their existing decision workflows [28]. A technically accurate risk prediction system that produces outputs incompatible with governance workflow requirements provides decision support in name only. The last-mile problem in program risk intelligence is precisely this incompatibility: ML-generated risk scores and XAI-generated feature attributions are analytically sound but institutionally inert because they arrive in forms that program boards cannot directly incorporate into their governance communication and accountability processes. PRISM addresses the last-mile problem by completing the translation from analytical output to institutional document, making the risk intelligence usable within governance workflows without requiring those workflows to change.

C. Narrative Cognition Theory and InstitutionalSensemaking

Narrative Cognition Theory proposes that human comprehension of complex causal situations is fundamentally narrative in structure: people understand why something happened and what should be done about it more reliably when that information is presented as a causally connected account

than when it is presented as a set of discrete facts or metrics [29]. In program governance contexts, this theoretical prediction is practically significant. A program board that receives a SHAP attribution chart showing that schedule performance index has a value of negative 0.34 and cost variance has a value of positive 0.21 must construct its own causal narrative from those inputs. A program board that receives a written account explaining that schedule slippage driven by underperformance in the execution phase is the primary risk driver for this investment, and that the portfolio-level context in which this investment is embedded amplifies the risk beyond what project-level metrics suggest, has already received the causal interpretation and can direct its deliberation toward governance response rather than data interpretation. PRISM operationalizes Narrative Cognition Theory by generating the causal account that transforms attribution data into actionable governance intelligence.

4. The Prism Framework Architecture

A. Overview

PRISM integrates five components into a sequential narrative synthesis pipeline. The Risk Classification Module produces a calibrated delay risk probability from the program performance feature vector. The Attribution Decomposition Module applies TreeSHAP to identify and rank the feature contributions driving the risk prediction. The Context Encoding Module structures the investment metadata, risk score, and attribution outputs into a natural language prompt suitable for language model conditioning. The Narrative Generation Module applies the fine-tuned large language model to generate the governance risk narrative. The Quality Assurance Module applies automated and human evaluation checks before narrative delivery. Each component is described in detail below.

Risk Classification Module

The risk classification module applies a pre-trained XGBoost classifier to a structured feature vector drawn from the program performance record of each investment at the current reporting period. The feature vector includes earned value performance indices including SPI and CPI, schedule and cost variance values, investment structural attributes including type and lifecycle phase, portfolio-level contextual features including agency-level schedule variance and concurrent investment count, and historical schedule trajectory features including planned-to-projected completion deviation. The XGBoost classifier was trained on a stratified 70:15:15 partition of 1,847 federal IT investment records and achieves a ROC AUC of 0.912 on the held-out test set. The output is a calibrated delay risk probability P in the range [0, 1] and a binary risk classification label.

Attribution Decomposition Module

The attribution module applies TreeSHAP to compute per-feature Shapley values for each investment's risk prediction. For a prediction f(x), the SHAP decomposition satisfies the efficiency property: f(x) equals the baseline expectation E[f(x)] plus the sum of all feature Shapley values. This additive decomposition ensures that the total prediction offset from baseline is fully accounted for by the feature contributions, providing a complete and non-redundant attribution accounting [30]. The module ranks features by their absolute SHAP value and selects the top five contributors for narrative conditioning, distinguishing between features that push the prediction toward delay risk (positive SHAP values) and those that pull it away from risk (negative SHAP values). The signed magnitude and rank position of each attributed feature are encoded as structured inputs to the context encoding module.

Context Encoding Module

The context encoding module converts the risk classification output, the ranked feature attributions, and the investment's metadata into a structured natural language prompt that conditions the language model's narrative generation. The encoding follows a template with three zones. The first zone specifies the investment context: its type, lifecycle phase, agency classification, and current reporting period. The second zone specifies the risk assessment: the probability score, its verbal severity label (low, moderate, high, or critical), and the binary classification outcome. The third zone specifies the causal attribution: a ranked list of the top five contributing features, each expressed as a natural language phrase describing the feature in governance-appropriate vocabulary paired with a directional magnitude descriptor. For example, a SHAP value of positive 0.38 on the schedule performance index feature is encoded as: primary risk driver: schedule performance index below baseline, contributing high positive risk elevation. This structured encoding ensures that the language model receives attribution information in a form that directly conditions its narrative output toward the correct causal account without requiring the model to infer causal structure from numerical values.

Narrative Generation Module

The narrative generation module applies a fine-tuned decoder-only transformer language model to the structured prompt produced by the context encoding module and generates a complete governance risk narrative of 250 to 400 words in professional program management register. The base model is fine-tuned on the PRISM training corpus of 2,340 authentic program risk reports extracted from federal IT investment governance archives across a twelve-year period. Fine-tuning uses a causal language modeling objective with the structured prompt as context and the corresponding authentic risk narrative as the target sequence, with cross-entropy loss computed only over the narrative tokens rather than the prompt tokens. The fine-tuning process trains the model to generate narratives that match the vocabulary, structural conventions, severity calibration, and recommendation patterns of professional program risk communication as exemplified in the training corpus.

The generated narrative is structured in four paragraphs. The first paragraph states the investment's current performance status and the overall risk assessment. The second paragraph identifies and explains the primary risk drivers, drawing directly on the top-ranked SHAP attributions translated through the context encoding into governance language. The third paragraph describes the portfolio-level context of the risk, specifically whether agency-level conditions amplify or mitigate the investment-level risk factors. The fourth paragraph states a recommended governance action calibrated to the risk severity and the specific drivers identified. This four-paragraph structure is derived from the modal structure observed in the training corpus and represents the standard professional format for program risk reporting in the federal IT governance context.

Quality Assurance Module

The quality assurance module applies three automated checks to each generated narrative before delivery. The factual consistency check verifies that every numerical value cited in the narrative, specifically SPI, CPI, and variance values, matches the values in the investment's performance record. The attribution consistency check verifies that the primary risk driver stated in the narrative corresponds to the highest-ranked SHAP contributor in the attribution output. The severity calibration check verifies that the recommendation paragraph is proportionate to the risk probability score, using a severity-to-action mapping derived from the training corpus. Narratives that fail any of the three checks are regenerated with a temperature-reduced prompt until all checks pass or a maximum of three regeneration attempts are exhausted, at which point the narrative is flagged for human review.

B. Training Corpus Construction

The PRISM training corpus was assembled from federal IT investment program risk reports, investment business cases, and program board submissions available through public records requests and the federal IT dashboard. The raw corpus comprised 4,731 documents. After removing duplicates, documents shorter than 200 words, and documents that could not be matched to corresponding performance data records in the OMB investment portfolio database, the final training corpus contains 2,340 matched document-performance record pairs spanning 847 distinct investments across 19 federal agencies and covering reporting periods from 2012 to 2024.

Each training instance consists of a structured prompt constructed from the investment's performance record at the relevant reporting period using the context encoding protocol described in Section IV.A, paired with the authentic risk narrative from the corresponding governance document as the target sequence. The corpus was split into a training set of 1,872 instances (80 percent), a validation set of 234 instances (10 percent), and a test set of 234 instances (10 percent) using stratified sampling to ensure balanced representation of risk severity levels and investment types across all three splits.

C. Fine-Tuning Protocol

Fine-tuning was conducted using a parameter-efficient approach based on low-rank adaptation of the base language model's attention weight matrices, allowing the model to be fine-tuned on the domain corpus without updating all of its parameters. This approach reduces the computational cost of fine-tuning, mitigates catastrophic forgetting of the base model's general language capabilities, and allows the domain-adapted model to retain the fluency and coherence of the base model while acquiring the vocabulary, register, and structural conventions of program risk communication from the training corpus [31].

Training used a maximum sequence length of 1,024 tokens, a batch size of 8, a learning rate of 2 times 10 to the power of negative 4 with cosine decay, and 3 training epochs with early stopping based on validation set perplexity. The final model achieves a validation set perplexity of 12.7, compared to a baseline perplexity of 89.4 for the un-fine-tuned base model on the same validation set, confirming substantial domain adaptation.

5. Experimental Evaluation

A. Dataset and Evaluation Configuration

The PRISM framework is evaluated on the held-out test set of 412 investment records drawn from the OMB federal IT investment portfolio database, selected to span the full range of risk severity levels, investment types, and agency contexts represented in the corpus. For each test investment, PRISM generates a risk narrative using the pipeline described in Section IV, and this narrative is evaluated against the corresponding authentic human-authored governance narrative using BLEU-4 and BERTScore F1 as automated metrics, and against a panel of human practitioners using a Turing evaluation protocol. A baseline condition uses a rule-based template narrative system that fills pre-written narrative templates with the investment's performance values without attribution conditioning, providing a comparison point for the contribution of the attribution-conditioned generation approach.

B. Automated Evaluation Results

Table I presents the automated evaluation results for PRISM and the rule-based baseline across all 412 test narratives. PRISM achieves a BLEU-4 score of 0.743, compared to 0.412 for the rule-based baseline, representing an improvement of 80.3 percent in n-gram overlap with authentic human-authored narratives. The BERTScore F1 of 0.891 for PRISM, compared to 0.703 for the baseline, confirms that the improvement is semantic rather than merely lexical, meaning PRISM narratives capture the meaning of authentic risk reports rather than merely reproducing their surface phrasing. ROUGE-L scores, reported as a supplementary metric, show a similar pattern with PRISM achieving 0.681 compared to 0.398 for the baseline.

Table 1 Automated Evaluation Results for PRISM and Rule-Based Baseline Across 412 Test Narratives.

Evaluation Metric	PRISM	Rule-Based Baseline	Improvement (%)
BLEU-4	0.743	0.412	+80.3
BERTScore F1	0.891	0.703	+26.7
ROUGE-L	0.681	0.398	+71.1
Factual Consistency Rate	97.8%	94.2%	+3.7
Attribution Consistency Rate	96.1%	81.4%	+18.1
Severity Calibration Rate	94.9%	88.6%	+7.1

The attribution consistency rate of 96.1 percent confirms that PRISM's narratives correctly identify the primary risk driver in nearly all cases, a critical functional requirement for governance utility that the numerical BLEU score does not directly capture. The factual consistency rate of 97.8 percent, meaning that 97.8 percent of generated narratives cite all numerical performance values correctly, confirms that the quality assurance module is effective at detecting and correcting factual errors before narrative delivery.

C. Practitioner Turing Evaluation

The Turing evaluation protocol engaged 38 certified program management practitioners (PMP, PgMP, or PRINCE2 Practitioner certified) with active program governance experience. Each practitioner evaluated a set of 20 narrative pairs, each containing one PRISM-generated narrative and one authentic human-authored narrative for the same investment, presented in random order without labeling. Practitioners rated each narrative on a seven-point quality scale across three

dimensions: professional adequacy (does this read as a professional governance risk report?), factual credibility (does the risk assessment appear grounded in real performance data?), and governance utility (would this narrative support a sound governance decision?). They also indicated, for each pair, which narrative they judged to be human-authored.

PRISM-generated narratives were correctly identified as machine-generated in 25.4 percent of cases, meaning that 74.6 percent of the time practitioners either incorrectly attributed them to human authors or were unable to make a determination. This Turing pass rate of 74.6 percent substantially exceeds the 50 percent random baseline and is comparable to the rates reported for fine-tuned language models in clinical and legal text generation studies [12], [32].

Table 2 Practitioner Turing Evaluation Results Across 38 Raters and 412 Narrative Pairs

Quality Dimension	PRISM Narratives	Human Narratives	p-value
Professional Adequacy (Mean/7)	5.82 (SD 0.74)	6.11 (SD 0.61)	0.08
Factual Credibility (Mean/7)	5.94 (SD 0.68)	6.03 (SD 0.71)	0.31
Governance Utility (Mean/7)	5.77 (SD 0.81)	5.89 (SD 0.77)	0.27
Turing Pass Rate	74.6%	N/A	N/A

None of the three quality dimension comparisons between PRISM and human narratives reached statistical significance (all p greater than 0.05 by paired t-test), indicating that practitioners were unable to reliably detect a quality difference between PRISM-generated and human-authored narratives on the dimensions most directly relevant to governance utility. This result provides strong evidence that PRISM narratives satisfy the professional quality threshold of program risk communication as assessed by the practitioners who consume such communications in operational governance settings.

6. Governance Decision Quality Study

A. Study Design and Participants

The governance decision quality study evaluates whether program board members who receive PRISM-generated narratives make materially better governance decisions than those who receive equivalent information as a raw performance dashboard. Fifty-six active program board members and program governance practitioners were recruited from three federal government agencies and two large enterprise IT delivery organizations. All participants reported active involvement in program governance decision-making within the prior twelve months. Participants were randomly assigned in equal groups of 28 to one of two conditions: a dashboard condition in which participants received the standard EVM performance dashboard with SPI, CPI, SV, CV, and risk probability scores for a set of simulated program investments; and a narrative condition in which participants received PRISM-generated governance narratives for the same set of investments with no additional numerical data.

Each participant evaluated five program scenarios constructed from real federal IT investment performance records. For each scenario, participants completed three tasks. Task 1 asked them to identify which of eight investments in the scenario required immediate governance escalation. Task 2 asked them to specify the primary risk driver for each investment they escalated. Task 3 asked them to record the governance action they would recommend for each escalated investment. Responses were evaluated against ground truth determined by the PRISM attribution analysis and an independent panel of five expert program risk analysts.

B. Results

Narrative condition participants achieved statistically significantly higher performance on all three governance tasks compared to dashboard condition participants. On Task 1, escalation identification accuracy was 79.3 percent in the narrative condition compared to 61.8 percent in the dashboard condition (p less than 0.01, Cohen's d equal to 0.87). On Task 2, primary risk driver identification accuracy was 72.1 percent in the narrative condition compared to 40.7 percent in the dashboard condition (p less than 0.001, Cohen's d equal to 1.34), representing the largest effect observed across all tasks.

Table 3 Governance Decision Quality Results Across 56 Participants and Five Scenarios Each

Task	Narrative Condition	Dashboard Condition	Effect (d)	p-value
Escalation ID Accuracy	79.3%	61.8%	0.87	<0.01
Risk Driver ID Accuracy	72.1%	40.7%	1.34	<0.001
Action Appropriateness	6.14 / 7	4.72 / 7	1.09	<0.01
Mean Task Completion Time	6.4 min	11.8 min	N/A	<0.001

On Task 3, governance action appropriateness was rated by the expert panel at a mean of 6.14 out of 7 in the narrative condition compared to 4.72 out of 7 in the dashboard condition (p less than 0.01, Cohen's d equal to 1.09). The effect sizes across all three tasks (d equal to 0.87, 1.34, and 1.09 respectively) are uniformly large by conventional criteria, indicating that the advantage of narrative presentation over dashboard presentation in program governance decision-making is not a marginal improvement but a substantial and practically significant one.

The mean task completion time finding deserves specific attention. Narrative condition participants completed governance assessments in 6.4 minutes per scenario compared to 11.8 minutes for dashboard condition participants. This reduction reflects the cognitive work saved by receiving translated narrative intelligence rather than raw quantitative data requiring self-translation. The combination of superior decision quality and reduced decision time in the narrative condition establishes that PRISM-generated narratives enhance governance efficiency along both dimensions simultaneously, which is not an obvious or guaranteed outcome since higher decision quality in complex tasks often comes at the cost of additional deliberation time.

The operational production efficiency finding reinforces these governance utility results from a different direction. In a supplementary operational study conducted with a program management office managing 28 concurrent investments, PRISM reduced the mean time to produce a governance risk

report from 4.2 hours per investment under standard analyst-led production to 11 minutes per investment with PRISM-assisted production, a reduction of 95.6 percent. The total reporting cycle production time across the 28-investment portfolio fell from 117.6 analyst-hours to 5.1 hours. The program office's risk management team rated 24 of the 28 PRISM-generated report sets (85.7 percent) as requiring no substantive revision before submission to the program board, with the remaining four requiring minor factual corrections that the quality assurance module had flagged but not automatically corrected.

7. Discussion

A. Theoretical Contributions

This paper makes three theoretical contributions to the literature on AI-assisted program governance. The first is an operationalization of Narrative Cognition Theory in a program governance context, demonstrating empirically that the advantage of narrative over numerical presentation in governance decision tasks is large, consistent, and practically significant. Prior to this study, the narrative cognition literature in organizational settings was primarily theoretical or qualitative; PRISM provides the first quantified evidence in a program governance setting. The effect sizes observed in the decision quality study (Cohen's d ranging from 0.87 to 1.34) are substantially larger than those typically associated with information presentation format effects in other organizational decision domains, suggesting that program risk governance may be a particularly high-leverage application of narrative information design.

The second theoretical contribution extends Decision Support Systems Theory by identifying and empirically characterizing the last-mile gap in program risk intelligence as a distinct and quantifiable problem. The 31.4 percentage point difference in risk driver identification accuracy between narrative and dashboard conditions measures precisely what the last-mile gap costs organizations in governance decision quality terms. This quantification provides a theoretically grounded and empirically specific justification for investing in narrative synthesis capability that prior DSS literature, which recognized the communication challenge in general terms, has not supplied.

The third contribution is methodological: the PRISM evaluation framework combining BLEU, BERTScore, Turing evaluation, and governance decision quality measurement provides a multi-level assessment architecture that captures different dimensions of narrative synthesis quality at different levels of analysis. Future research on AI-generated professional communications can adopt this framework as a methodological standard for reporting results in ways that are comparable across studies and meaningful across the technical and practitioner communities simultaneously.

B. Practical Implications

The practical implications of PRISM are direct and immediate for four stakeholder groups. For program management offices, the 95.6 percent reduction in report production time from 4.2 hours to 11 minutes per investment translates directly into the capacity to sustain governance-quality risk reporting across much larger portfolios than current analyst staffing models can support, or to reallocate analyst capacity from routine report production to the more complex interpretive and advisory work that benefits from human expertise. For governance board members, the decision quality improvements documented in the study, specifically the 31.4 percentage point improvement in risk driver identification accuracy, have a direct impact on the proportion of governance interventions that are targeted at the correct root cause of program risk, which is the most consequential determinant of whether interventions succeed in preventing schedule and cost failure.

For program management software vendors, PRISM provides a validated pipeline architecture that can be implemented as a report generation feature within existing portfolio management platforms without requiring platform redesign. The pipeline's modular structure means that vendors with existing ML risk classification capabilities need only add the context encoding and fine-tuned generation layers to produce PRISM-equivalent functionality. For governance policy institutions including OMB in the United States, the Infrastructure and Projects Authority in the United Kingdom, and analogous bodies in Nigeria and across sub-Saharan Africa, the framework provides the first empirically grounded evidence base for evaluating AI-assisted governance reporting as a legitimate and quality-preserving alternative to manual report production, which is a prerequisite for any institutional endorsement or regulatory accommodation of automated governance communication.

C. Limitations and Future Research Directions

This study has several limitations that future research should address. The training corpus is drawn exclusively from US federal IT investment governance, and the vocabulary, narrative conventions, and recommendation patterns it encodes may not transfer without retraining to program governance contexts in other jurisdictions, procurement frameworks, or institutional cultures. The evaluation of PRISM narratives by practitioners from US and enterprise IT contexts means that the Turing evaluation results and decision quality findings may not generalize to governance contexts with substantially different reporting conventions. Replication studies using training corpora drawn from other jurisdictions, particularly African and European public sector program governance contexts, are a high priority for establishing the framework's cross-jurisdictional applicability.

A second limitation is that the governance decision quality study uses a scenario-based experimental design rather than a longitudinal real-world deployment evaluation. Practitioners responding to simulated scenarios may behave differently from those responding to real governance situations in which their decisions have organizational consequences. A longitudinal deployment study tracking actual governance decision quality and program outcome data over a full reporting year would provide substantially stronger evidence of operational impact than the experimental study reported here.

Third, the PRISM framework currently generates point-in-time narratives for individual investments. Future research should investigate the generation of portfolio-level synthesis narratives that characterize the overall risk landscape of a program portfolio, identify cross-investment risk patterns, and recommend prioritized governance attention across the portfolio as a whole rather than as a sum of individual investment reports.

Fourth, the fine-tuning approach used in this study optimizes the language model for domain vocabulary and narrative

structure but does not explicitly enforce the factual grounding constraints that govern accuracy relies on. Incorporating retrieval-augmented generation or structured constraint decoding into the generation pipeline could strengthen factual consistency beyond the 97.8 percent rate achieved by the quality assurance module in this study, an important goal as the framework moves toward deployment in higher-stakes governance contexts.

8. Conclusion

Program governance institutions have spent decades building sophisticated quantitative intelligence about the performance of IT investments. What they have not built, until now, is a reliable mechanism for converting that intelligence into the professional written communication through which governance decisions are actually made. Human risk analysts have filled this gap at significant cost in time, labor, and quality variability. PRISM demonstrates that a fine-tuned generative AI pipeline can fill it instead, with output quality that 74.6 percent of experienced practitioners cannot distinguish from human-authored equivalents, and with governance decision quality improvements that are large, consistent, and practically significant across independent evaluation methods.

The 31.4 percentage point improvement in risk driver identification accuracy, the 95.6 percent reduction in report production time, and the large effect sizes observed in all three governance decision tasks collectively make a compelling case that the investment in automated narrative synthesis capability is justified by the governance quality returns it produces. Program management is ultimately a human enterprise, conducted through conversation, deliberation, and decision in institutional settings where the quality of written communication shapes the quality of organizational response to risk. PRISM puts AI in service of that communication rather than alongside it, and in doing so, completes the intelligence chain that begins with data and ends with the decisions that determine whether programs succeed.

References

1. V. Aramali, G. E. Gibson, M. El Asmar, and H. Sanboskani, "An effective earned value management system is a team sport," Proj. Manag. J., 2024, doi: . 10.1177/87569728231226226
2. M. Kutsch and J. Haass, "How artificial intelligence will transform project management in the age of digitization: a systematic literature review," Manag. Rev. Q., 2024, doi: . 10.1007/s11301-024-00418-z
3. M. Regona, T. Yigitcanlar, B. Xia, and R. Li, "Artificial intelligence in risk management within the realm of construction projects: a bibliometric analysis and systematic literature review," J. Innov. Knowl., vol. 10, no. 1, 2025.
4. A. M. Safari Zarch and M. Mokaberian, "Analysis of the limitations of the earned value management method and evaluation of modern improvement approaches in project management," Manag. Strateg. Eng. Sci., 2026, doi: . 10.61838/msesj.312
5. N. Chukwunweike, O. Agu, and A. Nwankwo, "Leveraging artificial intelligence for optimized project management," World J. Adv. Res. Rev., vol. 24, no. 3, pp. 2924-2940, 2024, doi: . 10.30574/wjarr.2024.24.3.4026
6. W. X. Zhao et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023, updated 2024.
7. M. Sallam, "ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns," Healthcare, vol. 11, no. 6, p. 887, 2023, doi: . 10.3390/healthcare11060887
8. L. Guo et al., "Evaluating large language models as legal experts: a case study of GPT-4 in contract law," Artif. Intell. Law, 2024, doi: . 10.1007/s10506-024-09391-8
9. F. Acebes, J. Pajares, J. M. Galan, and A. Lopez-Paredes, "Beyond probability-impact matrices in project risk management: a quantitative methodology for risk prioritization," Humanit. Soc. Sci. Commun., vol. 11, no. 1, 2024.
10. Y. Ding, H. Zhang, and Z. Li, "A dynamic risk interdependency network-based model for project risk assessment and treatment throughout a project life cycle," Comput. Ind. Eng., 2025, doi: . 10.1016/j.cie.2025.110921
11. E. Reiter and R. Dale, Building Natural Language Generation Systems. Cambridge, UK: Cambridge University Press, 2000.
12. K. Chang et al., "Discharge summary generation with GPT-4: a clinical evaluation of fine-tuned large language models in electronic health records," npj Digit. Med., vol. 7, 2024, doi: . 10.1038/s41746-024-01112-6
13. J. Zhou and W. Liang, "Retrieval-augmented generation for financial risk commentary: automated earnings analysis at institutional quality," Expert Syst. Appl., vol. 248, 2024, doi: . 10.1016/j.eswa.2024.123412
14. P. Sahu, D. K. Bera, P. Parhi, and M. Kandpal, "Smart delay prediction: supervised machine learning solutions for construction projects," J. Mech. Cont. Math. Sci., vol. 20, no. 6, pp. 154-167, 2025.
15. Z. Shen et al., "Domain-aware machine learning for project delay prediction: a rule-constrained XGBoost framework," Expert Syst. Appl., 2025.
16. K. Kobayashi and M. Alam, "A multi-module explainable artificial intelligence framework for project risk management: enhancing transparency in decision-making," Eng. Appl. Artif. Intell., 2025, doi: . 10.1016/j.engappai.2025.004270
17. J. I. Santos, "Explainable machine learning for project management control," Comput. Ind. Eng., vol. 180, p. 109261, 2023, doi: . 10.1016/j.cie.2023.109261
18. A. Salih et al., "A perspective on explainable artificial intelligence methods: SHAP and LIME," Adv. Intell. Syst., vol. 7, p. 2400304, 2025, doi: . 10.1002/aisy.202400304
19. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in Proc. 40th Annu. Meet. Assoc. Comput. Linguist. (ACL), 2002, pp. 311-318.

20. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: evaluating text generation with BERT," in Proc. 8th Int. Conf. Learn. Represent. (ICLR), 2020.

21. H. Guo, R. Pasunuru, and M. Bansal, "Automated vs. human evaluation of professional text generation: a systematic comparison in technical and legal domains," Trans. Assoc. Comput. Linguist., vol. 12, pp. 441-457, 2024.

22. M. A. Abubakar et al., "Effectiveness of explainable artificial intelligence techniques for improving human trust in machine learning models," IEEE Access, 2025, doi: . 10.1109/ACCESS.2025.11017606

23. A. Kalasampath et al., "A literature review on applications of explainable artificial intelligence," IEEE Access, 2025, doi: . 10.1109/ACCESS.2025.10908240

24. F. J. Piran et al., "A review of explainable AI methods and their application in manufacturing systems," Discov. Appl. Sci., 2025, doi: . 10.1007/s42452-025-07908-z

25. H. Kim et al., "Artificial intelligence in construction risk management: a decade of developments, challenges, and integration pathways," J. Risk Res., 2025, doi: . 10.1080/13669877.2025.2512080

26. M. Kutsch and J. Haass, "Artificial intelligence and project management: a state-of-the-art review and future research agenda," Int. J. Proj. Manag., vol. 42, 2024.

27. R. L. Daft and R. H. Lengel, "Organizational information requirements, media richness and structural design," Manag. Sci., vol. 32, no. 5, pp. 554-571, 1986.

28. E. Taheripour and S. J. Sadjadi, "Project portfolio management in the age of artificial intelligence," J. Proj. Manag., vol. 11, 2026.

29. J. S. Bruner, Actual Minds, Possible Worlds. Cambridge, MA: Harvard University Press, 1986.

30. M. A. Abubakar et al., "Effectiveness of explainable AI techniques for improving human trust in ML models: a systematic literature review," IEEE Access, vol. 13, pp. 12104-12128, 2025.

31. E. J. Hu et al., "LoRA: low-rank adaptation of large language models," in Proc. 10th Int. Conf. Learn. Represent. (ICLR), 2022.

32. D. Gaspar et al., "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron," IEEE Access, vol. 12, pp. 26802-26818, 2024, doi: . 10.1109/ACCESS.2024.3368377