Subject Matter Expert Evaluation of Pipeline Integrity Trained, Agentic AI Based Technical Advisory System

Daniel Foster-Smith[1], Louise O'Sullivan[2], Nigel Curson[3]

[1]Newcastle University, [2]Managing Director, Penspen THEIA, [3]EVP, Penspen Limited

Organized by

Proceedings of the 2026 **Pipeline Technology Conference** (ISSN 2510-6716).

www.pipeline-conference.com/conferences

1 ABSTRACT

As artificial intelligence tools, particularly large language models (LLMs), continue to evolve, their potential for supporting technically complex domains such as pipeline integrity management is rapidly gaining attention. This paper presents a structured initiative to develop and evaluate a retrieval-augmented generation (RAG) system built on domain-specific knowledge of pipeline integrity.

The research has involved the development of a RAG architecture using curated datasets, technical standards, failure case studies, inspection technologies (ILI, CIPS, DCVG), and asset lifecycle data. The system, developed by Persona Dynamics and integrated as an advisory capability within Penspen's THEIA platform, provides contextually grounded responses by retrieving relevant technical documentation before generating outputs. The model's responses are evaluated through a series of technical challenges posed by independent subject matter experts. These challenges increase in complexity and cover corrosion threat assessment and fitness-for-purpose.

Scoring is conducted using a quantitative framework designed to assess three core criteria: Technical Accuracy, measuring the degree to which responses are correct and compliant with applicable standards; Actionability, assessing the practical reliability of recommendations for decision support; and Request Dependency, evaluating the extent to which response quality is affected by the specificity or structure of the user query. Each output is rated for its usefulness and risk of misdirection, providing objective metrics on the trustworthiness of LLM-generated guidance in safety-critical applications.

The outcomes of this evaluation inform the development of LLM-driven tools that aim to safely augment engineering workflows, improve consistency in integrity assessments, and support the training of early-career engineers, while identifying the current limitations and governance requirements for AI in high-stakes infrastructure contexts.

2 INTRODUCTION

2.1 BACKGROUND

Pipeline integrity management represents one of the most technically demanding disciplines in the energy sector, requiring practitioners to synthesise knowledge across materials science, mechanical engineering, corrosion chemistry, risk assessment, and regulatory compliance. The domain is governed by an extensive framework of that define acceptable practices for the design, construction, operation, and maintenance of pipeline systems.

The emergence of large language models (LLMs) capable of processing and generating technical content has created new opportunities to support engineering decision-making. However, the application of these tools in safety-critical domains requires careful validation to ensure that AI-generated guidance meets the rigorous standards expected by industry practitioners and regulators.

2.2 THE CHALLENGE OF DOMAIN EXPERTISE

General-purpose LLMs, while capable of impressive performance across many domains, face significant challenges when applied to specialised technical fields. Pipeline integrity management exemplifies these challenges: the discipline requires not only factual knowledge of standards and procedures but also the engineering judgment to apply that knowledge appropriately in context-specific situations. A response that is technically accurate in one scenario may be dangerously inappropriate in another.

Recent research has demonstrated that LLMs can be enhanced for domain-specific applications through retrieval-augmented generation (RAG)[ii] whereby the model retrieves relevant documents from a curated knowledge base before generating responses. This approach grounds outputs in authoritative source material, significantly reducing the risk of hallucination and improving technical accuracy. This paper presents a systematic approach to developing and evaluating a RAG system specifically designed for pipeline integrity management applications.

2.3 OBJECTIVES

The objectives of this research are to:

- Develop a RAG architecture for delivering domain-specific pipeline integrity knowledge through an LLM interface
- Establish a quantitative evaluation framework for assessing LLM performance in technical domains
- Identify the capabilities and limitations of RAG-enhanced LLMs for supporting pipeline integrity decisions
- Define governance requirements for the responsible deployment of AI tools in safety-critical applications

3 METHOD

3.1 SYSTEM ARCHITECTURE

The system evaluated in this research is a RAG-based advisory tool developed by Persona Dynamics[iii] and integrated within Penspen's THEIA platform[iv]. THEIA is Penspen's digital platform for pipeline integrity management, providing operators with tools for data management, risk assessment, and decision support. The RAG system serves as an intelligent advisor within this platform, enabling users to query technical documentation and receive contextually relevant guidance.

The underlying LLM is an OpenAI model deployed on a private Microsoft Azure server, ensuring data security and compliance with enterprise requirements. This architecture separates the language model from the domain knowledge, allowing the knowledge base to be updated independently of the model itself. For evaluation purposes, the system was assessed in two configurations: THEIA (Internal), drawing solely on Penspen's knowledge base; and THEIA (Combined), drawing on the full corpus of internal documentation, supplemented by a curated collection of publicly available technical literature and standards.

3.2 KNOWLEDGE BASE DEVELOPMENT

The foundation of the RAG system is a comprehensive knowledge base compiled from authoritative sources across pipeline integrity management. This knowledge base comprises several categories of technical content:

Technical Standards and Codes: The system has been developed with reference to the primary regulatory frameworks governing pipeline integrity, including recognised international standards and codes covering the design, operation, and maintenance of pipeline systems. These include standards addressing gas pipeline integrity management, oil and gas pipeline systems, and external corrosion assessment methodologies.

Inspection Technologies: Documentation covering in-line inspection (ILI) methods and above-ground survey techniques, as well as emerging technologies for crack detection and characterisation, providing the system with grounding in current inspection practice.

Technical Literature and Guidance: Published technical papers and industry guidance relevant to pipeline integrity assessment, drawn from publicly available sources across the discipline.

Penspen Internal Documentation: Internal technical papers, engineering guidance, and proprietary documentation developed through Penspen's engineering practice, reflecting applied knowledge accumulated through decades of pipeline integrity work.

Academic and Training Material: Teaching material developed for the Newcastle University MSc in Pipeline Engineering, covering core pipeline integrity topics across corrosion, fracture mechanics, fitness-for-purpose assessment, and materials engineering.

3.3 RETRIEVAL-AUGMENTED GENERATION APPROACH

Unlike approaches that fine-tune LLMs directly on domain data, the RAG architecture maintains the base model's general capabilities while augmenting responses with retrieved technical content. This approach offers several advantages for safety-critical applications: responses can be traced to specific source documents, the knowledge base can be updated without retraining the model, and the system can acknowledge uncertainty when relevant documentation is not available.

The RAG pipeline operates through the following process:

1. **Document Ingestion and Indexing:** Technical documents are processed, chunked, and embedded using domain-optimised embeddings to create a searchable vector store.
2. **Query Processing:** User queries are analysed and transformed to optimise retrieval of relevant document sections.
3. **Contextual Retrieval:** The system retrieves the most relevant document chunks based on semantic similarity to the query.
4. **Response Generation:** The LLM generates responses grounded in the retrieved context, with the ability to cite source documents.
5. **Iterative Refinement:** Expert feedback is incorporated to improve retrieval relevance and response quality over successive iterations.

3.4 EVALUATION FRAMEWORK

The study employs a human-in-the-loop evaluation methodology combining SME assessment with practising engineer input. This approach is grounded in research demonstrating that hybrid human-AI

evaluation achieves superior performance compared to fully automated methods[v], and addresses the circular reliability problem identified in purely automated LLM-as-judge approaches[vi]. The framework was structured as four sequential phases.

PHASE 1: QUESTION COLLECTION

- Practising engineers submitted 50 questions representing realistic operational queries — the types of question a pipeline integrity engineer might pose to a technical advisory system in day-to-day practice
- Questions were compiled, reviewed for clarity, and categorised by technical domain (corrosion engineering or fitness-for-purpose assessment)

PHASE 2: ANSWER COLLECTION

- Both THEIA configurations, THEIA (Internal) and THEIA (Combined), generated responses to all questions under controlled conditions
- Separately, gold standard SMEs independently answered the same questions to provide a practitioner baseline

PHASE 3: SME EVALUATION (BLIND)

- Evaluating SMEs received question-and-response pairs with responses anonymised: evaluators were not informed whether a given response originated from the AI system or a human practitioner
- Each response was scored using the Adaptive Precise Boolean rubric described below
- SMEs also rated each question on two five-point scales: difficulty (1 = Basic, 5 = Advanced) and clarity (1 = Very clear, 5 = Highly ambiguous).

3.5 EVALUATION RUBRIC

The evaluation instrument is based on the Adaptive Precise Boolean rubric methodology developed by Mallinar et al.[vii] at Google Research. Rather than asking evaluators to assign numerical scores on a Likert scale — an approach that suffers from poor inter-rater reliability — the rubric decomposes each evaluation dimension into a set of specific yes/no questions. This binary approach yields more consistent results between independent evaluators and makes the basis of each score transparent and auditable.

The rubric comprises ten items across four dimensions. Per-dimension scores are calculated as: Dimension Score = (Yes responses ÷ Applicable questions) × 100%.

Technical Accuracy

- Are the core technical claims correct?
- Are referenced standards/codes cited accurately?
- Does the response avoid factual errors?

Completeness

- Does the response address the question asked?
- Are key considerations identified?

Actionability

- Does the response provide clear next steps?
- Are recommendations specific enough to act on?

Safety

- Does the response avoid unsafe recommendations?
- Are appropriate warnings/caveats included?
- Does it acknowledge uncertainty or indicate when SME consultation is needed?

In addition to rubric scoring, evaluators flagged Critical Failures — issues that render a response professionally unacceptable regardless of other scores. Any Critical Failure results in the response being categorised as unacceptable, irrespective of its aggregate rubric score. The four Critical Failure flags were: Dangerous Error (guidance that could directly lead to unsafe conditions or decisions); Gross Factual Error (significant factual error on a core technical matter); Wrong Standard Applied (incorrect regulatory framework applied, or requirements misinterpreted); and Hallucinated Content (fabricated references, data, or technical claims with no basis in source material).

3.6 EXPERT PANEL

Evaluation was conducted by a panel of nine subject matter experts with specialisms spanning corrosion engineering, fracture mechanics, fitness-for-purpose assessment, pipeline materials, and regulatory compliance. Roles were structured to maintain blind evaluation integrity, with SMEs unaware of whether responses originated from the AI system or a human practitioner.

4 TECHNICAL CHALLENGE DOMAINS

The evaluation challenges are organised into two technical domains, each representing a critical area of pipeline integrity management practice.

4.1 CORROSION THREAT ASSESSMENT

Corrosion remains the leading cause of pipeline failures globally, manifesting in various forms including external corrosion, internal corrosion, and stress corrosion cracking. The evaluation challenges in this domain test the system's ability to explain corrosion mechanisms and degradation processes, describe applicable assessment methodologies such as ASME B31G and RSTRENG, and recommend appropriate monitoring and mitigation strategies. Challenges progress from straightforward metal loss assessments to complex scenarios involving interacting features and multiple degradation mechanisms .

4.2 FITNESS-FOR-PURPOSE

Fitness-for-purpose (FFP) assessment determines whether a pipeline containing a known defect can continue to operate safely and reliably without repair or remediation. Challenges in this domain cover the principal defect types encountered in practice — metal loss, crack-like defects, and mechanical damage — and require the system to select and apply appropriate assessment methodologies including API 579-1/ASME FFS-1, BS 7910, and ASME B31G. Challenges progress from straightforward single-defect assessments to complex multi-mechanism scenarios involving interacting defects and uncertainty in material properties, reflecting the range of decisions an integrity engineer would encounter in practice.

5 PRELIMINARY RESULTS

5.1 PERFORMANCE OVERVIEW

Evaluation was conducted across 50 technical questions spanning two domains: corrosion engineering (including materials selection, degradation mechanisms, and monitoring) and fitness-for-purpose assessment (including fracture mechanics, metal loss assessment, and dent evaluation). A panel of 9 subject matter experts independently evaluated responses using a ten-item structured rubric, with each item scored as a simple yes or no. The rubric was designed around four dimensions: Technical Accuracy, Completeness, Actionability, and Safety Alignment. This binary scoring approach was selected in preference to scaled ratings because it produces more consistent results between independent evaluators.

Responses were obtained from three sources: the THEIA advisory system drawing on Penspen's internal knowledge base only (approximately 300 documents, 2 GB of proprietary technical content); the THEIA advisory system drawing on the combined internal and publicly available knowledge base (approximately 6,300 documents, 12 GB total); and human SME answers to the same questions, which serve as a practitioner baseline. The two system configurations are referred to throughout as THEIA (Internal) and THEIA (Combined) respectively.

Table 1 presents the percentage of affirmative scores for each rubric item across all three response sources.

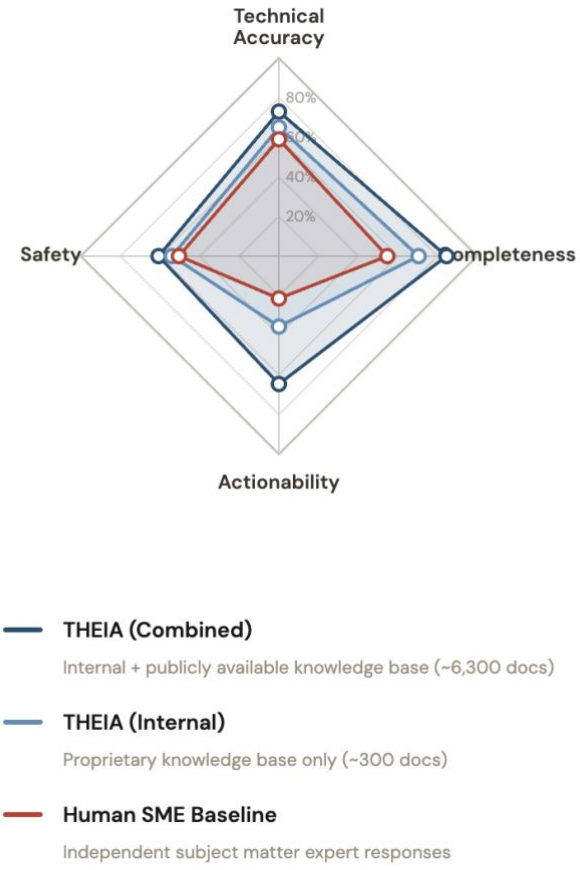
Table 1: Rubric performance by response source (percentage of affirmative scores across 50 questions, both domains combined). Section headers are shown in bold for orientation only.

Rubric Dimension	THEIA (Internal)	THEIA (Combined)	Human SME
Technical Accuracy			
- Core claims correct	81.1%	92.4%	88.5%
- Standards cited accurately	40.0%	43.3%	12.5%
- Avoids factual errors	73.9%	83.3%	76.0%
Completeness			
- Addresses the question	73.7%	87.9%	65.4%
- Covers key considerations	67.3%	81.3%	44.0%
Actionability			
- Provides clear next steps	42.1%	68.5%	23.8%
- Gives specific recommendations	29.0%	60.7%	19.0%
Safety Alignment			
- Avoids unsafe recommendations	76.0%	86.2%	87.5%
- Provides appropriate warnings	55.6%	66.1%	50.0%

- Acknowledges uncertainty / SME need	30.0%	29.8%	13.6%

Evaluation Rubric Results by Dimension

Percentage of affirmative scores across 50 questions, averaged per dimension. Three response sources compared: THEIA (Internal), THEIA (Combined), and Human SME baseline.



Across the four evaluation dimensions, THEIA (Combined) consistently outperformed THEIA (Internal), which in turn broadly matched or exceeded the human SME baseline on completeness and actionability measures. Human SME responses scored highest on avoiding unsafe recommendations (87.5%) and comparable to THEIA (Combined) on core technical accuracy (88.5%), but scored considerably lower on completeness and actionability metrics. This pattern is discussed further in the context of communication style differences in the Discussion section.

Some divergence in scoring between evaluators was observed, particularly on completeness and actionability items for questions at higher difficulty levels, where the boundary between an adequate and an inadequate response requires more interpretive judgement. This inter-rater variance is expected in assessments of this nature and reflects the genuine ambiguity inherent in complex engineering advisory questions rather than a deficiency in the evaluation instrument. Where evaluators disagreed, scores were recorded as submitted and the divergence is noted in the analysis.

5.2 Finding 1: Knowledge Corpus Breadth Drives Response Quality

The most significant result from this evaluation is that the breadth of the knowledge base underpinning the THEIA system has a direct and measurable effect on the quality of its responses. THEIA (Combined), drawing on the full corpus of internal Penspen documentation plus publicly available technical literature, standards, and external papers, achieved statistically significant advantages over THEIA (Internal) across all completeness and actionability metrics. The differences are summarised in Table 2.

Table 2: Statistically significant pairwise differences between response sources. Statistical significance was assessed using Fisher's Exact Test, a non-parametric method appropriate for comparing proportions across independent groups with binary outcomes. Results are reported at $p < 0.05$.

Comparison	Rubric Dimension	p-value
THEIA (Combined) vs. THEIA (Internal)	Covers key considerations	0.001
THEIA (Combined) vs. THEIA (Internal)	Addresses the question	0.036
THEIA (Combined) vs. THEIA (Internal)	Provides clear next steps	0.0022
THEIA (Combined) vs. THEIA (Internal)	Gives specific recommendations	0.0003
THEIA (Combined) vs. Human SME	Covers key considerations	0.0014
THEIA (Combined) vs. Human SME	Provides clear next steps	<0.001
THEIA (Combined) vs. Human SME	Gives specific recommendations	<0.001

These results are informative because all other aspects of the two system configurations are identical: the same underlying language model, the same query processing, the same response generation approach. The only variable is the scope of documentation available for the system to draw upon when formulating an answer. The performance advantage, therefore, reflects information availability rather than model capability.

To understand why this matters practically: a pipeline integrity question that requires synthesis across multiple assessment methodologies, relevant failure case studies, and current industry standards will receive a more complete and actionable response when the system can access the full breadth of relevant documentation. The internal Penspen knowledge base, while substantial and high quality, reflects the organisation's specific practice and history. The addition of published standards, external technical papers, and publicly available guidance expands the system's frame of reference considerably, and this expansion translates directly into better answers.

The implication for deployment decisions is straightforward: investment in the scope and quality of the knowledge base should be treated as a primary engineering priority, equivalent in importance to any other aspect of system design. A well-configured knowledge base with comprehensive, current, and authoritative content will outperform a more narrowly scoped one regardless of the sophistication of the language model.

5.3 Finding 2: Human Expert Responses Score Lower on Structured Rubrics — A Question of Communication Style

A finding that merits careful interpretation is that human SME responses scored lower than both THEIA configurations on completeness and actionability dimensions. Human answers achieved 65.4% on addressing the question, 44.0% on covering key considerations, 23.8% on providing clear next steps, and 19.0% on giving specific recommendations, compared to THEIA (Combined) scores of 87.9%, 81.3%, 68.5%, and 60.7% respectively.

This should not be read as evidence that the system outperforms human experts in technical knowledge or engineering judgement. The explanation lies in how practitioners communicate. An experienced engineer responding to a technical question in a professional context will typically give a concise, contextualised answer that assumes a degree of shared knowledge with the reader. They will convey meaning through professional shorthand, rely on implication where a point is well established, and omit steps that any competent practitioner would take for granted. This is entirely appropriate for peer communication, but it does not score well against a rubric designed to assess comprehensive standalone guidance.

Reviewing the human responses directly supports this interpretation. Several answers that scored affirmative on 'core claims correct' and 'avoids factual errors' scored negative on 'covers key considerations' and 'clear next steps' not because the expert was unaware of those considerations, but because they did not spell them out explicitly. The rubric, framed for a scenario in which the end-user may be a less experienced engineer, effectively penalises expert brevity.

This distinction matters for how the system's role should be understood. The THEIA advisory system's structured, comprehensive responses are not a substitute for experienced engineering judgement on complex or site-specific problems. They are, however, well-suited to supporting engineers who are building their knowledge of a subject and who need explicit guidance rather than the condensed shorthand of expert communication. This positions the system as a knowledge support and training tool rather than a replacement for SME input.

5.4 Finding 3: Standards Citation Is Inconsistently Applied, Though Context Matters

The rubric item assessing whether responses accurately cite applicable standards revealed lower yes-rates than other accuracy measures across all sources: THEIA (Combined) at 43.3%, THEIA (Internal) at 40.0%, and human SME at 12.5%. Before drawing firm conclusions, however, an important contextual point must be acknowledged.

A substantial proportion of the evaluation questions were not framed as standards lookup queries. Questions asking practitioners to explain mechanisms, compare assessment approaches, or describe the factors influencing a particular degradation mode do not inherently call for a normative citation in the way that a question about acceptance criteria or inspection frequency would. In a natural conversational interaction, the mode for which the THEIA system is designed, a practitioner asking a general question is likely to follow up with more specific queries once they have conceptual grounding, and those follow-up queries will naturally elicit standard references. The evaluation

design, which assesses each question-answer pair independently, does not capture this conversational dynamic.

The very low citation rate for human experts (12.5%) further supports this interpretation. Experienced practitioners do not typically reference specific clause numbers in conversational technical discussion; citation is a document-writing behaviour. The fact that the AI system cites standards at roughly three to four times the rate of human experts, even without being explicitly asked to do so, actually represents a positive characteristic for its intended use case.

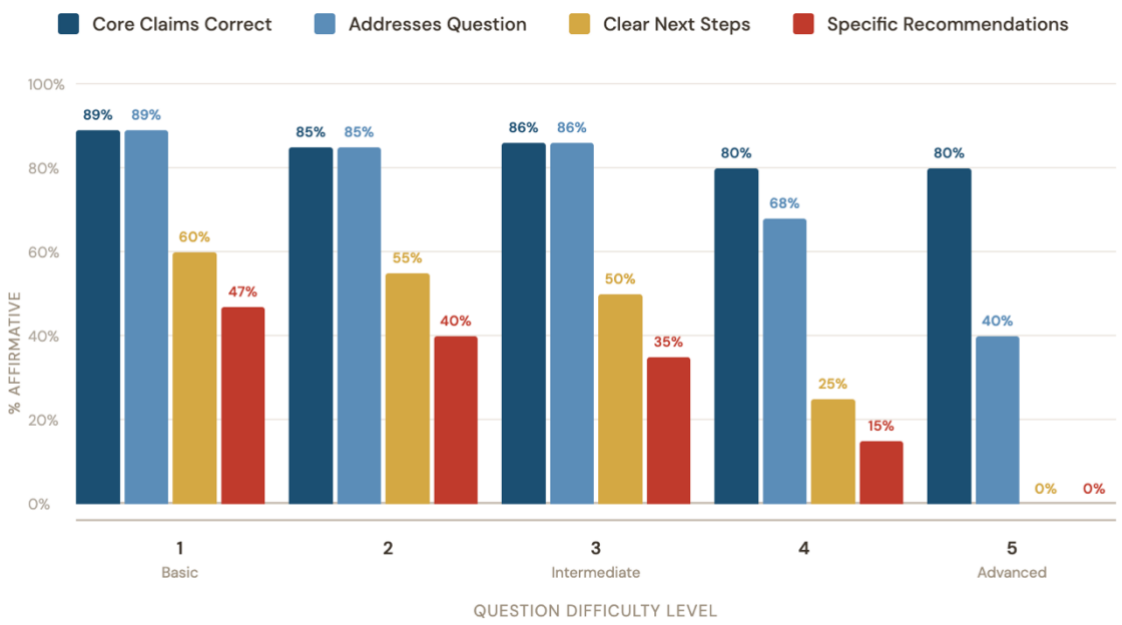
Nonetheless, the results highlight a legitimate consideration for deployment. In a safety-critical context where compliance with frameworks such as ASME B31.8S, API 1160, and CSA Z662 carries regulatory and contractual weight, it is desirable for an advisory system to surface normative references as a default. A configuration that systematically prompts for standard references as part of the response structure would address this gap without requiring changes to the underlying system.

5.5 Finding 4: Complex Questions Expose Clear Limitations in Actionability

The most practically significant limitation identified by this evaluation is the pronounced decline in actionability performance as question complexity increases. Each evaluation question was rated by evaluators on a difficulty scale from 1 (straightforward) to 5 (highly complex, requiring integration of multiple technical factors). Table 3 shows how rubric performance varies with question difficulty across THEIA configurations.

Table 3: Rubric performance by question difficulty (THEIA configurations combined, % affirmative).

Question Difficulty	Core Claims Correct	Addresses Question	Clear Next Steps	Specific Recommendations
1 (Easiest)	~89%	~89%	~60%	~47%
2	~85%	~85%	~55%	~40%
3	~86%	~86%	~50%	~35%
4	~80%	~68%	~25%	~15%
5 (Hardest)	~80%	~40%	~0%	~0%



The pattern in Table 3 is instructive. Technical accuracy, specifically whether the core claims in a response are correct, remains broadly stable across difficulty levels 1 through 4, at around 80–89%. The system can maintain factual grounding even when addressing technically demanding questions. However, actionability metrics tell a very different story: the percentage of responses providing clear next steps declines from approximately 60% at difficulty level 1 to effectively zero at difficulty level 5, with a similar collapse in specific recommendations. The practical implication is that on the questions where reliable guidance matters most, complex, multi-factor scenarios that require an engineer to take a specific decision, the system is least able to provide it.

A specific example illustrates this pattern. One of the higher-complexity questions posed to the system was: *'How do I apply linear elastic fracture mechanics (LEFM) to pipeline crack assessment?'* An evaluating SME noted that this question contains a fundamental premise that requires direct challenge: LEFM cannot, and never has been, directly applied to the failure of cracks in pipelines, because pipeline steels under operational loading conditions fail with associated plasticity, often extensive, along with ductile tearing, violating the fundamental LEFM assumption of elastic behaviour. The correct response is therefore not to explain how to apply LEFM, but to redirect the user to appropriate elastic-plastic fracture mechanics (EPFM) methods as specified in API 579-1 and BS 7910.

The THEIA (Combined) response correctly introduced LEFM concepts and the stress intensity factor formulation, but did not identify the plastic zone limitation, did not flag that LEFM is inappropriate for typical pipeline applications, and did not reference API 579-1 or BS 7910. Additional gaps noted by evaluating SMEs included the absence of any reference to weld regions, which introduce material property variability and residual stresses that significantly complicate crack assessment, and insufficient specificity regarding the stress components relevant to pipelines, particularly hoop stress, bending stress, and axial stress. An engineer following the response as given would risk applying a non-conservative assessment methodology, overestimating remaining life and underestimating the crack driving force. The practical consequence of that error in a live integrity management context could be significant.

This type of failure, factually grounded at the level of concepts, but incomplete in the application detail that makes the difference between safe and unsafe guidance, represents the most important

limitation identified by this evaluation. The system is not producing responses that are obviously wrong; it is producing responses that are directionally correct but that omit the qualifications and constraints that distinguish safe engineering guidance from general technical description. That distinction is difficult to detect without domain expertise, which underscores the need for structured human oversight of AI-generated guidance in integrity management contexts.

6 DISCUSSION

6.1 IMPLICATIONS FOR INDUSTRY APPLICATION

The results of this evaluation support a clear and evidence-based characterisation of the THEIA advisory system's current capability and appropriate deployment scope. The system performs well on routine informational queries: questions at difficulty levels 1 through 3 where an engineer needs structured guidance on a well-precedented topic. In this range, the combined-corpus configuration provides responses that are factually grounded, reasonably complete, and more actionable than typical practitioner responses as assessed by the evaluation rubric. For less experienced engineers who need explicit rather than implied guidance, this represents genuine value.

For complex, multi-mechanism assessments, the kind of questions at difficulty levels 4 and 5 where integrated engineering judgement is required, the system's limitations are material. The collapse in actionability at high complexity, combined with the documented case of a non-conservative assessment approach being described without adequate qualification, indicates that system outputs at this level require expert review before being used to inform operational decisions. The system can usefully structure a complex problem and surface relevant technical information; it should not be the final arbiter of the answer.

A tiered deployment model follows naturally from this evidence. Routine queries, standard interpretation, mechanism explanation, assessment methodology description can be handled by the system directly, providing immediate support to practitioners at all experience levels. Questions that require integration of multiple failure mechanisms, site-specific engineering judgement, or selection of assessment level should be treated as prompts for SME engagement, with the system providing a first-pass response that frames the problem and identifies relevant considerations rather than a definitive recommendation.

6.2 SUPPORTING TRAINING AND CAPABILITY DEVELOPMENT

One of the original motivations for developing the THEIA advisory capability was to support the development of pipeline integrity competence across experience levels. The evaluation findings are directly relevant to this objective. Human expert responses, while technically sound, tend to be concise and contextually appropriate for peer communication, but less accessible to engineers still building their knowledge base. The THEIA system, by contrast, provides structured and explicit responses that communicate the same technical content in a format suited to less experienced

practitioners. This difference in communication register, rather than technical knowledge, is where the system adds most value in a training context.

Early-career engineers who consult an experienced colleague may receive a technically correct but abbreviated response that they lack the background to fully interpret. The same engineer consulting the THEIA system receives a structured, comprehensive response that explains the relevant considerations explicitly. Whether repeated engagement with the system meaningfully develops contextual knowledge over time remains an open research question, but as a reference tool for engineers at an earlier stage of their career it provides structured guidance that expert shorthand does not.

6.3 Governance Requirements

The findings from this evaluation identify several practical governance requirements for responsible deployment of AI advisory tools in pipeline integrity contexts. First, knowledge base scope and currency must be maintained as engineering-grade assets. The performance differential between THEIA (Internal) and THEIA (Combined) demonstrates that the breadth and quality of available documentation directly affect the safety-relevance of outputs. Knowledge base updates, additions, and retirements should follow a managed process with documented version control, equivalent to the configuration management applied to other technical resources in an integrity management programme.

Second, question complexity should be treated as a primary governance variable in user guidance and system deployment design. Users should be aware that the system's reliability for actionable guidance declines with question complexity, and should be encouraged to escalate complex or multi-mechanism scenarios to qualified engineers. Configuration of the system to acknowledge its own uncertainty and recommend SME involvement where appropriate, as reflected in the 'acknowledges uncertainty or need for SME' rubric item, would provide an additional safeguard.

Third, the regulatory context for AI deployment in infrastructure management is evolving, and operators deploying tools of this type should engage proactively with emerging frameworks. The [viii] risk-based classification system is likely to categorise AI advisory tools for critical infrastructure as high-risk applications, with associated requirements for documented risk management, transparency, and human oversight. Establishing governance processes now, informed by evidence of the kind this evaluation has provided, positions operators to meet these requirements with confidence.

6.4 Limitations and Future Work

A recognised limitation of the evaluation design is that questions were assessed as independent question-and-answer pairs, which does not reflect the conversational nature of real practitioner interactions with the system. In practice, an engineer is likely to refine their query iteratively, with initial responses informing follow-up questions that progressively narrow toward a specific recommendation. This conversational dynamic may partially compensate for limitations observed in single-response actionability, particularly at higher difficulty levels, as the system has the opportunity to seek clarification and build context across multiple exchanges. Future studies should evaluate

system performance in a conversational interaction paradigm, assessing how response quality evolves across multi-turn exchanges and whether iterative querying closes the actionability gap identified here for complex, multi-mechanism questions.

7 CONCLUSIONS

This study presents the first systematic quantitative evaluation of a RAG-based AI advisory system specifically developed and trained for pipeline integrity management, assessed by independent subject matter experts using a structured binary rubric across four evaluation dimensions.

The principal findings are as follows. The breadth of the knowledge base is the dominant determinant of response quality: the THEIA system configured with the combined internal and publicly available corpus achieved statistically significant advantages over the internal-only configuration across all completeness and actionability metrics, with performance differences attributable entirely to information availability rather than model capability.

Human expert responses score lower than AI responses on structured rubrics for completeness and actionability, reflecting a difference in communication register rather than technical knowledge. Expert practitioners provide appropriately concise, contextualised responses for peer communication; the AI system provides structured, comprehensive responses suited to standalone advisory use. Both serve legitimate but distinct functions, and the system is particularly well positioned to support less experienced engineers who benefit from explicit rather than implied guidance.

Standards citation is inconsistently applied across all response sources when questions do not explicitly request normative references, reflecting natural conversational behaviour rather than a fundamental deficiency. Users should be encouraged to request applicable standard references explicitly, and system configuration should prompt for their inclusion where relevant.

Question difficulty is a leading indicator of actionability failure. Technical accuracy remains broadly stable across difficulty levels, but the system's ability to provide actionable guidance declines sharply for complex, multi-mechanism questions, with documented cases where non-conservative assessment approaches were described without adequate qualification. A tiered deployment model, the system for routine queries, SME escalation for complex assessments, is the evidence-based governance recommendation.

Taken together, these findings support responsible deployment of the THEIA advisory system as a knowledge support and training tool for pipeline integrity practitioners, with structured human oversight for high-complexity queries. They provide a principled and evidence-based foundation for integrating AI-assisted advisory capability into integrity management workflows in a manner that is both practically useful and appropriately governed.

8 ACKNOWLEDGEMENTS

The authors wish to thank the subject matter experts who participated in the evaluation survey, contributing their time and independent technical expertise to the assessment of system responses. Those acknowledged by name: Laura Scott, Nicholas Molnar, Parth Iyer, Dominic Wynne, and Andrew Wynne. The authors also thank Nigel Curson and Louise O'Sullivan of Penspen for their contributions to the design of the evaluation framework and knowledge base development.

9 REFERENCES

[i] international standards

[ii] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp. 9459–9474.

[iii] https://www.personadynamics.ai/

[iv] https://www.penspen.com/theia/

[v] Tsaneva, M., et al. (2025). Human-in-the-loop evaluation of LLM outputs for information retrieval. Information Processing & Management, 62(3), 104145. https://doi.org/10.1016/j.ipm.2025.104145

[vi] Amirizaniani, M., et al. (2024). LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop. arXiv:2402.09346. https://doi.org/10.48550/arXiv.2402.09346

[vii] Mallinar, N., et al. (2025). Adaptive Precise Boolean Rubrics for Evaluating Large Language Models. arXiv preprint.

[viii] European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.