www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

TOWARDS TRUSTWORTHY LLMS FOR PROJECT MANAGEMENT: BENCHMARKING A VECTOR HYBRID AND GRAPH RAG ARCHITECTURES

SUBMITTED: February 2026
PUBLISHED: August 2026
EDITOR: Bimal Kumar
DOI: 10.36680/j.itcon.2026.042

Eduardo Navarro-Bringas, Assistant Professor in Construction Management
Institute of Sustainable Built Environment (ISBE), Heriot-Watt University
https://orcid.org/0000-0003-0308-3860
E.Navarro_Bringas@hw.ac.uk

SUMMARY: *Retrieval Augmented Generation (RAG) is a practical response, yet there is limited comparative evidence on how different RAG architectures trade off efficiency and answer quality for major project portfolio data. This article benchmarks two RAG pipelines using UK Government Major Projects Portfolio (GMPP) datasets (from 2020-2021 to 2023-2024) and a fixed query set, holding the foundation LLM model (i.e. GPT 4.0.), prompt template, and generation settings constant to isolate the effects of the two retrieval architectures. A vector-based hybrid pipeline (metadata filtering, dense retrieval, BM25 re-ranker) is compared with a labelled property graph-based pipeline using graph schema-aware retrieval. Results, evaluated through RAGAS metrics and a supplementary blind expert evaluation, indicate task-dependent trade-offs between the two architectures. While the vector-based RAG pipeline achieves lower query time latency and lower token consumption per query, the graph RAG pipeline improves answer quality based on RAGAS dimensions linked to trustworthiness. Particularly, graph RAG outperforms in terms of faithfulness (0.81 vs 0.78), context recall (0.87 vs 0.83) and answer relevancy (0.74 vs 0.69), suggesting better grounding in retrieved evidence and more complete coverage of the information required to provide relevant answers. Context precision is marginally higher for the vector pipeline (0.86 vs 0.85), which might be linked with the limited scoped document retrieval settings, while answer correctness performs comparably across both architectures. Expert evaluation also rated graph-based responses higher for accuracy, completeness and trustworthiness across the query subset, although preferences varied by query type. Overall, the findings highlight that architecture choice should be driven by expected query type, information risk, and the relative importance of efficiency, provenance and decision-support reliability in project and portfolio management.*

KEYWORDS*: LLMs, Retrieval Augmented Generation (RAG), vector index, knowledge graphs, major projects.*

REFERENCE*: Navarro-Bringas, E. (2026). Towards trustworthy LLMs for project management: Benchmarking a vector hybrid and graph RAG architectures. Journal of Information Technology in Construction (ITcon), 31, 990-1014. https://doi.org/10.36680/j.itcon.2026.042*

1. INTRODUCTION

Large Language Models (LLMs) have rapidly advanced the state of the art in natural language processing (NLP), enabling fluent text generation, summarisation, and question answering that are increasingly being adopted across professional domains (Brown et al., 2020; Bubeck et al., 2023). Within construction and project-based industries, early studies and practitioner-facing reviews suggest that LLMs can support a wide range of lifecycle activities (e.g., briefing, information retrieval, reporting, and decision support), yet also emphasise that current applications remain constrained by domain specificity, governance requirements, and the structure of project information (Saka et al., 2024; Adebayo et al., 2025; Samsami, 2024; Marzouk & Enaba, 2019; Pan & Zhang, 2021). These opportunities are particularly salient in major projects, where teams routinely operate across heterogeneous datasets (e.g. narrative reports, spreadsheets, and document repositories) and where timely, explainable access to reliable information is central to delivery performance (NAO, 2025; Too and Weaver, 2014).

Despite their capabilities, multiple studies have documented persistent limitations when LLMs are applied to knowledge-intensive and decision-critical tasks. Key issues include reliance on static training corpora, constrained access to domain-specific or proprietary information, and hallucinated outputs that may be presented with high linguistic confidence (Maynez et al., 2020; Huang et al., 2023). These limitations are amplified in project and portfolio management contexts, where the "ground truth" often sits inside organisational systems rather than the public web, and where errors can translate into material cost, schedule, or governance impacts. In this regard, Flyvbjerg (2025) argues that contemporary AI systems may produce persuasive but unreliable answers in management-relevant settings, framing this failure mode as "*artificial ignorance*". While some recent deployments incorporate tool use and retrieval actions at query time (e.g., calling external resources), access to authoritative project data is frequently restricted by organisational boundaries, information security, and data governance constraints (Schick et al., 2023).

The project management and construction management literature similarly indicates strong interest in AI-enabled decision support, but continued challenges related to trust, transparency, explainability, and practical integration with real organisational data. Reviews of AI in project management and construction project management report increasing application breadth (e.g., cost, schedule, risk, and quality), alongside recurring barriers such as fragmented datasets, limited interpretability, and difficulties operationalising models across changing project contexts (Adebayo et al., 2025; Pan & Zhang, 2021; Bento et al., 2022; Love et al., 2023; Taboada et al., 2023; Marzouk and Enaba, 2019). From an information-management perspective, construction and major projects remain document- and communication-intensive, making retrieval, traceability, and context reconstruction persistent challenges, which are issues that pre-date LLMs and were already prominent in earlier research on electronic document management and project information workflows (Björk, 2003).

In this regard, Retrieval Augmented Generation (RAG) has emerged as a dominant architectural response to these limitations by coupling an LLM with an explicit retrieval mechanism, enabling responses to be generated conditionally on retrieved evidence rather than solely relay on data used on the training of the FM. Lewis et al. (2020) formalised RAG for knowledge-intensive NLP tasks, demonstrating that RAG can improve factuality, answer relevance, and robustness relative to foundation model baselines. Subsequent work in dense retrieval and retrieval and generation systems reinforced that retrieving authoritative context improves LLM performance when factual grounding is needed (Karpukhin et al., 2020; Izacard & Grave, 2021). In construction-related settings, recent research increasingly explores RAG approaches to support information access and reasoning over construction management systems, reflecting the sector's reliance on large volumes of unstructured and semi-structured documentation (Wu et al., 2025).

The RAG architecture and design research space has expanded in recent years, covering distinct RAG architectures, such as naïve, advanced, and modular configurations, and recent surveys highlight that evaluation practices remain inconsistent and that architecture selection is often under-justified relative to task characteristics and operational constraints (Gao et al., 2023). In parallel, graph-based RAG (often referred to as GraphRAG) has attracted increasing attention as graph indexes can explicitly encode entities, relationships, and heterogeneous structures that are difficult to preserve in vector similarity-based retrieval (Han et al., 2026). Yet, despite rapid growth in proposed variants, there remains limited empirical benchmarking that contrasts graph-based and vector-based RAG architectures under consistent conditions and on datasets that resemble real portfolio and project reporting, and with evaluation designs that link automated RAG metrics to domain-expert judgement.

This paper addresses that gap through a comparative evaluation of two RAG pipelines designed for knowledge-intensive project and portfolio queries: (1) a vector-based RAG pipeline using dense retrieval, with metadata and lexical re-ranking, and (2) a graph-based RAG pipeline operating over a property graph representation of the same underlying dataset. The study is grounded in the UK Government Major Projects Portfolio (GMPP) transparency datasets collected by the National Infrastructure and Service Transformation Authority (NISTA), which provides departmental-level, semi-structured portfolio reporting across multiple years, combining quantitative performance indicators with qualitative narrative fields. The analysis draws on the detailed departmental releases available up to the 2023-24 reporting cycle (with 2024 data published in January 2025) and is complemented by the associated annual report materials (NISTA, 2025).

The aim of the research is to benchmark how alternative RAG architectures trade off query-time cost and performance against retrieval effectiveness and answer quality for project-management queries over a semi-structured portfolio data. Specifically, the study evaluates both pipelines using a standardised benchmark query set and compares outcomes using retrieval effectiveness metrics and automated RAG quality measures, and a supplementary blind expert evaluation of selected responses, enabling analysis both at aggregate level and by query type. The intended contribution is a replicable experimental methodology and an evidence-based characterisation of architecture trade-offs, contrasting the automated evaluation with expert judgement, to inform the design of RAG-enabled decision support systems for project and portfolio management in construction and adjacent domains.

2. LITERATURE REVIEW

2.1 Retrieval Augmented Generation (RAG)

In recent years, LLMs have showcased strong capabilities in NLP and response generation across a wide range of domains, driving rapid adoption in both research and applied settings (Brown et al., 2020; Bubeck et al., 2023). However, multiple studies have documented persistent limitations when such models are applied to knowledge-intensive and decision-critical tasks, including reliance on static training corpora, constrained access to domain-specific or proprietary information, and the generation of hallucinated outputs presented with high linguistic confidence (Maynez et al., 2020; Ji et al., 2023). While recent LLM deployments increasingly incorporate tools or agentic retrieval mechanisms, allowing models to interact with external resources at query-time, access to authoritative project data often remains restricted by organisational and data governance constraints, and stored in private repositories. These limitations are particularly relevant in PM contexts, where project professionals need to question heterogeneous and evolving datasets, and where inaccuracies can have material financial and operational consequences (Flyvbjerg, 2014; Locatelli et al., 2023).

RAG addresses these limitations by coupling large language models with explicit information retrieval mechanisms, enabling responses to be generated based on externally retrieved evidence rather than relying solely on parametric model knowledge. In a typical RAG pipeline, a user query is first used to retrieve relevant documents or records from an external knowledge base, after which the retrieved context is provided to the LLM to condition response generation. Lewis et al. (2020) formalised this paradigm for knowledge-intensive NLP tasks, demonstrating that RAG models outperform FM-only baselines in terms of factual accuracy, robustness, and answer relevance. Subsequent studies have reinforced these findings, showing that retrieval augmentation improves performance across a range of tasks where precise factual grounding is required (Guu et al., 2020; Izacard & Grave, 2021; Karpukhin et al., 2020). More recent evaluations comparing foundation model performance with RAG variants further indicate improved performance, with reported improvements of approximately 15 - 40% in factual accuracy, answer relevance, and faithfulness for knowledge-intensive queries, alongside significant reductions in hallucination rates (OpenAI, 2023; Es et al., 2024). A RAG paradigm, which relies in curated data and information embeddings, is particularly aligned with PM activities, which typically involve synthesising organisational knowledge, historical project data, and performance records (Vanhoucke et al., 2016), often stored privately, rather than relying on general open-access knowledge.

Beyond response accuracy and relevance, RAG has been shown to reduce hallucinations and improve trustworthiness in applied settings. Shuster et al. (2021) provide empirical evidence that retrieval augmentation significantly reduces hallucinated responses in conversational agents by constraining generation to retrieved

evidence from the curated knowledge base. Similar performance gains have been reported across distinct knowledge-intensive domains, including scientific question answering (Izacard & Grave, 2021), legal document analysis and question answering (Chalkidis et al., 2022), or enterprise or organisational knowledge management systems (Thorne & Vlachos, 2021). In more project specific industries, like construction, RAG applications have also shown promising applications for knowledge-intensive tasks such as contract management and administration (Zheng et al., 2025), or the management of construction information utilised during the project delivery phase (Wu et al., 2025).

Finally, RAG introduces advantages regarding provenance, transparency, and system maintainability. By decoupling the knowledge base and knowledge storage from the LLM itself, RAG enables inspection of retrieved sources, facilitates updating the underlying knowledge base without retraining the model, and supports traceability between outputs and original embedded data (Lewis et al., 2020). These characteristics are particularly aligned with the requirements of PM and portfolio analysis, where accountability, auditability, and contextual interpretation of information is key (NAO, 2025). As a result, RAG has become a dominant paradigm for deploying LLMs in knowledge-intensive and decision-support applications, forming the foundation for more advanced retrieval architectures explored in subsequent sections of this paper.

2.2 Vector-based RAG

The most implemented form of RAG relies on vector-based retrieval, in which textual content is embedded into a continuous semantic space and retrieved based on similarity to a query embedding (Lewis et al., 2020). In its most simple form, where only a semantic vector query is embedded, this type of RAG is known as naïve RAG. In naïve RAG pipelines, documents (i.e., unstructured or semi-structured) are typically segmented into text chunks, which are encoded using dense embedding models, and subsequently stored in a vector database. At inference time, a user query is embedded into the same vector space, and the top-k most similar chunks are retrieved using a similarity metric such as cosine similarity. The retrieved top-k responses are then provided to the LLM as contextual grounding for response generation (Lewis et al., 2020; Reimers & Gurevych, 2019).

Vector-based RAG systems have proven effective for document-centric and text-heavy tasks, where semantic similarity is a strong proxy for relevance and filtering. Dense retrieval methods, such as dense passage retrieval, outperform sparse lexical approaches in question answering, document retrieval, and summarisation tasks by capturing semantic relationships in the data that go beyond exact term overlap (Karpukhin et al., 2020). This capability underpins the adoption of vector databases in contemporary RAG systems, enabling nearest-neighbour search over a large document corpus (Johnson et al., 2019: Malkov & Yashunin, 2020).

In practical deployments, vector-based RAG pipelines have commonly incorporated re-ranking mechanisms to improve and fine-tune retrieval precision. A common strategy is to combine dense vector retrieval with sparse lexical methods such as BM25, either as a hybrid retrieval architecture or as a lightweight re-ranker applied to the retrieved top-k candidates through the sparse lexical method (e.g. cosine similarity) (Robertson & Zaragoza, 2009). BM25 remains widely used due to its interpretability, robustness, and strong baseline performance in document retrieval (Yu et al., 2025). More advanced pipelines may employ neural re-rankers, which model fine-grained query/document interactions and usually improvements in ranking quality at the expense of increased computational resources (Nogueira & Cho, 2019).

Despite their effectiveness for document retrieval, vector-based RAG approaches are limited when applied to datasets characterised by complex structures, temporal evolution of data, or relational dependencies, as dense retrieval primarily optimises for semantic similarity rather than preserving entity identity or relational coherence (Karpukhin et al., 2020; Wang et al., 2022). Consequently, retrieved passages may be topically and semantically relevant yet drawn from the wrong sources, time/dates, or organisational context. This limitation becomes particularly evident when queries require distinguishing between entities with similar names, for instance tracking changes in project status across reporting periods, or filtering information by organisational unit (Wang et al., 2022). While as shown in research naïve and vector-based RAG usually outperform foundation LLMs operating without RAG, these constraints have motivated the investigation of more structured and advanced retrieval architectures.

2.3 Modular and Graph-based RAG

In recent years, research on RAG architectures has moved beyond naïve, single-step RAG pipelines, even more advanced architectures with re-ranking features, towards modular RAG architectures, in which indexing, retrieval and output generation are decomposed into explicitly defined stages that allow for selecting appropriate methods at each. These architectures introduce intermediate components such as query reformulation, multi-stage retrieval, filtering, aggregation, or evidence re-ranking/selection, which reflect a recognition that more complex information needs cannot always be adequately addressed through a vector similarity metrics (Gao et al., 2023).

A comprehensive survey by Gao et al. (2023) formalises this evolution by proposing a taxonomy of RAG architectures, distinguishing between naïve, modular, iterative, and structured variants. While modular RAG approaches improve controllability and reasoning capability, the survey also highlights practical drawbacks, including increased system complexity, orchestration overhead, computational cost, and challenges in maintaining reliability across interacting components. These limitations have motivated interest in architectures that introduce structure, adaptability and control without excessive procedural complexity (Asai et al., 2023).

Within this design space, graph-based RAG has emerged as a prominent form of structured RAG, in which retrieval is grounded in an explicit knowledge graph representing entities, relationships, and attributes. Rather than relying on multi-stage procedural reasoning, graph-based RAG constrains retrieval through graph construction and traversal, preserving entity identity and relational context throughout the retrieval process. Han et al. (2026) provide a detailed decomposition of Graph-RAG pipelines, explicitly outlining stages for entity recognition, entity linking, graph construction, and graph-guided retrieval. Their work clarifies how Graph-RAG operationalises structure as an alternative to fully modular or agent-based reasoning, while maintaining transparency and control over retrieval behaviour.

Recent industry-led research further demonstrates the practical viability of graph-based retrieval in large-scale systems. The industry example (e.g., Salesforce chatbot) graph-architecture from Hilel et al. (2025) reports that entity-centric graph representations, combined with lightweight vector similarity, improve retrieval precision and contextual coherence in enterprise-scale deployments, particularly for queries requiring disambiguation across entities and organisational contexts. These studies position Graph-RAG as a structured yet operationally feasible subset of modular RAG, suitable for real-world decision-support settings, and consider the solution an architecture that can be adapted to different knowledge-intensive domains.

The construction and maintenance of knowledge graphs also possess well-acknowledged challenges in the knowledge engineering and ontology development community, including schema design effort, data integration complexity, and the need for ongoing curation and maintenance as source data evolve over time (Ehrlinger & Wöß, 2016; Hogan et al., 2021). Graph embeddings and learned representations further introduce trade-offs between expressiveness and interpretability, often requiring iterative refinement to align with domain semantics (Nickel et al., 2016). Consequently, graph-based RAG is not universally optimal, but represents a deliberate architectural choice where the benefits of more explicit and knowledge driven structures outweigh the associated costs with developing and maintaining graphs.

In project and portfolio management contexts, queries often require distinguishing between similar entities, tracking changes across project reporting periods, and aggregating information by organisational responsibility (NAO, 2025), the trade-offs of developing a graph could be justified. For such settings, graph-based RAG provides a structured and operationally feasible alternative to vector-based retrieval, while avoiding the complexity of fully modular or multi-agent RAG pipelines. This study develops a graph-based RAG architecture aligned with these principles and evaluates its performance relative to a vector-based RAG in the following sections.

3. METHODOLOGY

This study follows a Design Science Research approach, in which an IT artefact is developed to address a class of problems and its utility is evaluated through systematic empirical assessment (Hevner et al., 2004, Peffers et al., 2007). The artefact in this study comprises two alternative pipeline implementations of RAG to develop a UK's Government Major Project Portfolio (GMPP) question retrieval and answering system: (i) a vector-based RAG pipeline and (ii) a graph-based RAG pipeline. To assess the impact of architectural choice, we adopt a comparative

experimental benchmarking design, in which both pipelines are evaluated under identical conditions. This includes a common source dataset, standardised query set, same LLM foundational model with identical settings, including prompts and instructions (Sanderson and Zobel, 2005, Clough and Sanderson, 2013). The evaluation of the pipelines was also undertaken using a common set of performance and retrieval quality evaluation metrics known as Retrieval Augmented Generation Assessment (RAGAS) (Es et al., 2024) as well as a supplementary blind human evaluation of a sample of generated responses (van der Lee et al., 2019). This combination of artefact-building and controlled comparative evaluation aligns with established guidance on DSR evaluation and information retrieval system assessment, enabling transparent comparison of retrieval quality, answer quality, and computational performance across pipelines (Peffers et al., 2007; Prat et al., 2015).

3.1 Dataset preparation

3.1.1 Dataset and scope

The empirical evaluation presented in this study is based on publicly available data published by NISTA, which reports annually on the Government Major Projects Portfolio (GMPP) (NISTA, 2025). The portfolio comprises major public-sector projects spanning infrastructure, construction, information technology, and service delivery programmes across the different UK government agencies and ministries. While a proportion of the projects relate to the infrastructure and the built environment, the inclusion of non-construction initiatives positions the dataset as representative of broader project and portfolio management context.

The data used in this study integrates GMPP datasets from financial years 2020-21 to 2023-24 (both inclusive), selected to capture the most recent period with consistent multi-year coverage as well as expanded reporting fields such as project benefits data which are relevant to RAG question answering.

3.1.2 Data pre-processing

The NISTA, previously known as the Infrastructure and Projects Authority (IPA), data on major projects is released annually in a comma separated value (csv) format by each governmental department/ministry, which in some cases possesses variations in schema, attribute availability, and naming conventions across reporting years (NISTA, 2025). To enable consistent longitudinal analysis, all datasets were normalised to a common schema through a structured preprocessing workflow. This process involved harmonising column names, aligning categorical values, and standardising data types, while preserving the financial data of numerical cells as well as semantic content in the unstructured narrative fields.

The empirical analysis uses the detailed departmental GMPP datasets available up to the 2023-24 reporting cycle; at the time of data capture (accessed 16 July 2025), the subsequent 2024-25 cycle was only available as an annual report and did not provide equivalent departmental-level tabular releases suitable for pipeline ingestion and benchmarking (NISTA, 2025). Such changes are common due to departmental restructuring, following government and ministerial restructuring, as well as project and programme reclassifications. Rather than collapsing these variations into a single static representation, the consolidated dataset retains project-year records, with the aim to support queries that require checks across multiple project financial years and entity-specific reasoning.

The resulting consolidated dataset is stored in .csv format and serves as the common input for both RAG pipelines evaluated in this study. The dataset, after data pre-processing, is available in an open repository (see data availability statement).

3.2 Pipeline development and evaluation

To enable a controlled comparison of RAG architectures, two pipelines were developed and evaluated in parallel. Both pipelines were designed to answer natural language queries over the same underlying dataset, use the same large language model for response generation, and follow the same evaluation procedure. The pipelines differ only in their retrieval and knowledge representation architecture, allowing the impact of architectural choice to be isolated.

Both pipelines follow a common high-level workflow: (i) data ingestion and processing, (ii) knowledge representation and indexing, (iii) query-time retrieval, and (iv) response generation using a foundation LLM.

3.2.1 Pipeline 1: Vector-based RAG

Pipeline 1 implements a vector-based hybrid RAG architecture representative of common document-centric deployments, combining metadata filtering with dense semantic retrieval (Gao et al., 2023). The consolidated GMPP csv dataset is ingested via an orchestration layer with LlamaIndex©, which performs record-to-text transformation, chunking of project-year entries, and attachment of structured metadata fields (e.g., reporting year, department, confidence rating) used for filtering at retrieval. Each chunk is embedded using a pre-trained embedding model (known as "*text-embedding-3-small*") and stored in a Pinecone© vector index together with the metadata.

At query time, the user query is embedded into the same vector space and matched against the indexed corpus using cosine similarity to retrieve an initial candidate set. To improve retrieval precision, a BM25 lexical re-ranking stage is applied to the retrieved candidates, prioritising passages that also exhibit strong term-level alignment with the query (Robertson & Zaragoza, 2009). The top-ranked chunks are then assembled into a context window and passed to the LLM for response generation using the fixed prompt template applied across both pipelines. The overview of the system architecture is shown in Figure 1.

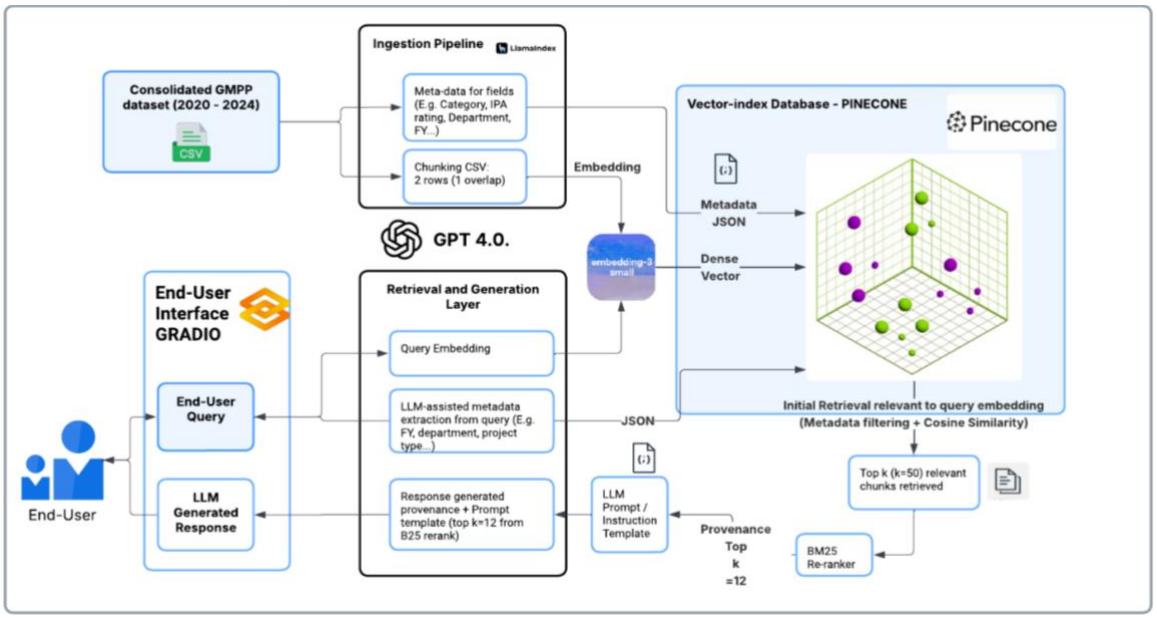

Figure 1: Vector-based RAG: Hybrid architecture with metadata filtering, dense retrieval and BM25 re-ranker.

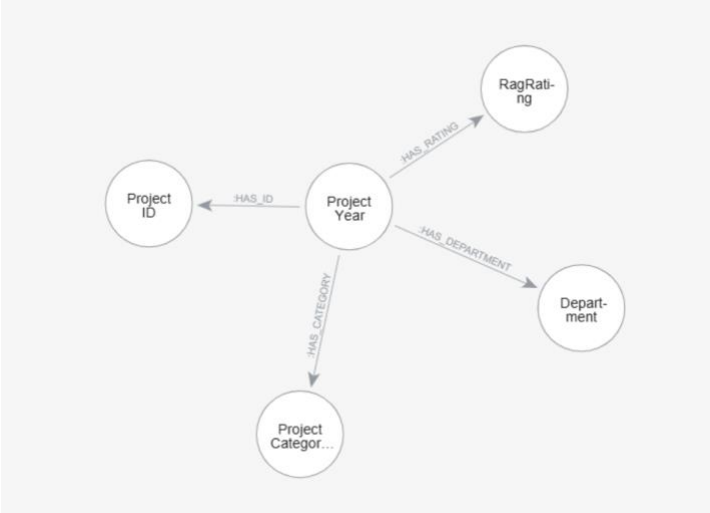

Figure 2: Labels and nodes reference schema in Neo4j ©.

3.2.2 Piepeline 2: Graph RAG

Pipeline 2 implements a graph-based RAG architecture grounded in an explicit Neo4j© labelled property-graph representation of the same GMPP dataset. Ingestion is performed via a Python ETL process that maps the consolidated CSV to a predefined graph schema (Figure 2) (i.e. node labels, relationship types, and property mappings), loading entities such as projects, departments, and reporting periods as nodes, with relationships capturing continuity and associations across organisational units and time. Narrative fields and quantitative indicators are stored as node properties and/or linked nodes according to this schema. Unlike Pipeline 1, graph construction does not rely on chunk embeddings or a vector index; retrieval is supported by structured graph representation rather than semantic similarity search.

At query time, the LLM operates in a schema-aware mode to interpret the user query and translate it into Cypher retrieval instructions constrained by the graph schema. The resulting Cypher query is executed against Neo4j© to retrieve a task-appropriate subgraph (entities, properties, and relationships), which is then serialised into a textual context window. This retrieved context is provided to the same LLM using the same fixed prompt template as vector-based pipeline 1 to generate the final response, ensuring comparability of the generation stage across both architectures. The overview of the graphRAG architecture is presented in Figure 3.

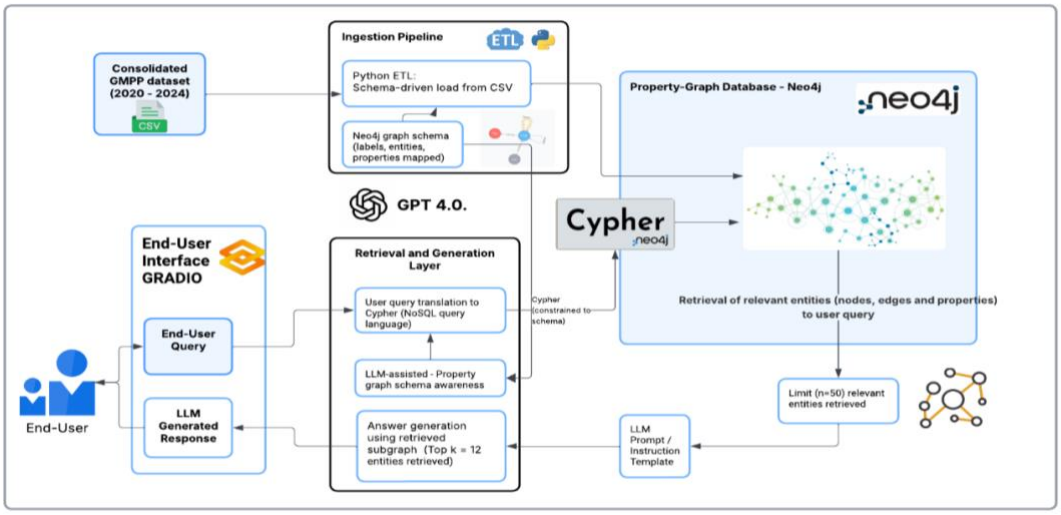

Figure 3: Graph RAG labelled property graph and LLM schema-aware retrieval architecture.

3.3 Retrieval settings and query set

3.3.1 Prompts at response generation

Both pipelines employ the same foundation LLM, GPT 4.0., within the pipeline for response generation. Additionally, the study utilised an identical prompt template and generation parameters to ensure consistency and comparable response generation. This includes maintaining consistent temperature, maximum token limits, k retrieval settings, and system instructions across both experimental pipelines. Holding LLM settings constant ensures that observed differences in response quality, faithfulness, and efficiency can be attributed to the retrieval architecture, rather than variation in generation behaviour from the use of different foundational LLM (Rau et al., 2024; Zhao et al., 2024). The used of consistent settings and prompts is aligned with the purpose of evaluating the performance of two distinct RAG architectures in the context of project and programme management.

Similarly, query handling and response formatting were standardised across pipelines. Where retrieval-specific parameters differed (e.g. vector similarity versus graph traversal), these differences are intrinsic to the pipeline architecture and not related to LLM configuration.

3.3.2 Standardised 20 query set

Evaluation of RAG systems typically relies on a fixed set of benchmark queries, allowing consistent comparison of alternative retrieval architectures under identical information needs. This approach is well established in

traditional information retrieval research, where systems are evaluated using shared query sets and test collections to assess retrieval effectiveness and robustness (Clough & Sanderson, 2013). More recent RAG evaluation studies similarly adopt curated query sets to compare retrieval strategies and end-to-end answer quality (Lewis et al., 2020; Es et al., 2024).

In this study, a benchmark set of 20 different natural language queries was developed to reflect realistic information needs encountered in project and portfolio management and related to the information stored in the GMPP dataset. Moreover, the queries were designed to span a range of retrieval and reasoning challenges, including narrative-heavy questions, project-specific queries, portfolio-level aggregation, temporal comparisons across reporting periods, and queries requiring identification or ranking based on quantitative attributes (e.g. cost, confidence ratings, or delivery status). The use of a diverse query set aligns with prior RAG evaluation studies, which emphasise depth and interpretability of results over large-scale benchmarking when evaluating applied or domain-specific systems (Lewis et al., 2020; Es et al., 2024), such as this specific GMPP dataset.

Ground-truth/reference answers were developed to provide a consistent basis for the AI generated output evaluation. For each query, the lead researcher prepared an initial reference answer by manually inspecting the relevant fields in the consolidated GMPP dataset, including project identifiers, department, reporting year, delivery confidence, cost and schedule fields, and narrative summaries where applicable. The answer was written using only information available in the underlying dataset. This follows the logic of test-collection-based information retrieval evaluation, where fixed information needs are paired with reference judgements or answers to enable comparable system assessment (Clough & Sanderson, 2013). To improve robustness, two additional researchers reviewed the ground-truth answers to check whether key records or data fields had been omitted and whether the answers were grounded in the original database columns. Any omissions or ambiguities were resolved by rechecking the source dataset before RAGAS scoring.

The benchmark queries, query rationale, typology of expected response, generated responses from each pipeline, and ground-truth answers used for the automated evaluation are available in an open repository (see Data Availability Statement). A subset of ten queries was subsequently selected from this benchmark set for supplementary blind human evaluation, with two queries selected from each query type (Available in Appendix A).

3.4 Analysis metrics

To enable systematic comparison between the two RAG pipelines, evaluation metrics were defined across three dimensions: computational performance, retrieval effectiveness, answer quality, and human judgement of response quality. These metrics capture differences in efficiency/cost of the RAG system, retrieval behaviour, end-to-end response quality, as well as expert perceptions of response quality attributable to each pipeline's architecture.

3.4.1 Performance metrics

Computational performance was evaluated using token consumption, which serves as a proxy for both cost and efficiency in large language model–based systems. A token refers to a discrete unit of text processed by a language model, typically corresponding to character fragments rather than full words (each token equating around 4 characters processed by the LLM), and represents the basic unit of computation for modern transformer-based models like GPT 4.0. (OpenAI, 2023). Foundational LLMs pricing models are token dependent, and as such token use is used to estimate model performance in terms of token usage (and cost per query).

For each query, token usage was recorded as the following components:

- Prompt tokens are those tokens consumed by the system prompt, user query, and retrieved context
- Completion tokens are those tokens generated by the language model in the final response.
- Embedding tokens are those tokens consumed during embedding operations (e.g. when establishing the vector index from the dataset).

$$Total\ LLM\ tokens = Prompt\ tokens + Completion\ tokens + Embedding\ tokens \quad (1)$$

Total token usage (1) was then computed as the sum of these components for each pipeline on responding the 20 queries. In any case, token accounting differs between pipelines. In the vector-based RAG pipeline 1, embedding tokens include an initial one-off cost associated with embedding the dataset into the vector database/index, in addition to query-time embeddings. In contrast, the graph-based RAG pipeline 2 did not require embedding for the graph construction; embedding tokens were incurred only at query time where required for retrieval or generation. Token usage was aggregated per query to enable direct comparison of operational efficiency across pipelines. This metric is particularly relevant for applied deployment scenarios, where differences in token consumption directly affect system cost and scalability.

3.4.2 Retrieval and answer quality metrics: RAGAS

End-to-end answer quality was evaluated using RAGAS, an automated evaluation framework specifically designed for RAG systems (Es et al., 2024). Unlike traditional NLP evaluation metrics that focus solely on answer similarity, RAGAS decomposes RAG performance into interpretable components that assess both the quality of retrieved context and the faithfulness and relevance of generated answers. All RAGAS metrics were computed using GPT-4o-mini as a consistent external evaluator model across all pipelines, following best practice to minimise self-evaluation bias and ensure comparable scoring behaviour (Es et al., 2024).

RAGAS evaluates the interaction between four core elements: the user query, the retrieved context for the response, the LLM-generated response, and the reference human response (known as the ground-truth answer). By analysing these relationships, the framework enables systematic assessment of whether a response is supported by retrieved evidence, whether the retrieved context is appropriate and sufficient, and whether the final answer addresses the original information need. Evaluation was performed using standard, conservative decoding settings prioritising stable and deterministic behaviour. This decomposition is particularly well suited to comparative evaluation of RAG architectures, as it allows retrieval and generation behaviour to be examined jointly and across a range of tests rather than in isolation. Figure 4 showcases how the different RAGAS metrics relate and evaluate the relation among the human ground-truth response, the query, the context retrieved (i.e. vector chunks and/or graph nodes), and the LLM's response.

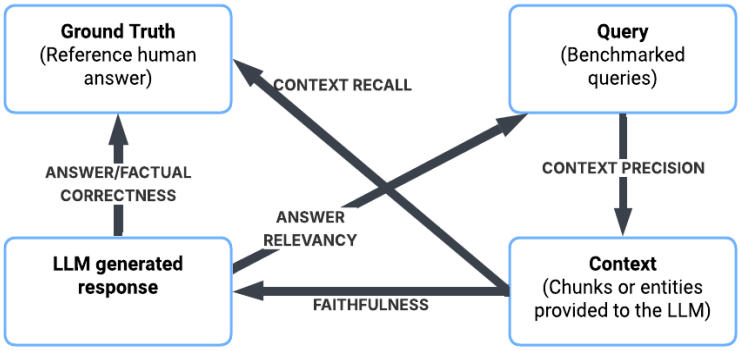

Figure 4: Overview of RAGAS metrics in relation to system constituents (Es et al., 2024).

Faithfulness (2)

$$Faithfulness = \frac{|\{c \in \mathcal{C}(a): c \text{ is supported by } C\}|}{|\mathcal{C}(a)|} \quad (2)$$

Faithfulness measures the extent to which the LLM's generated answer is grounded in the retrieved context. The response a is first decomposed into a set of atomic claims $\mathcal{C}(a)$ (i.e. factual statements). Each claim $c \in \mathcal{C}(a)$ is then checked against the retrieved context C to determine whether it is *supported* (i.e., can be inferred from, or is explicitly evidenced by, the retrieved passages). The metric is computed as the proportion of supported claims relative to the total number of claims in the answer. A faithfulness score close to 1 indicates that the response is fully evidence-grounded in C, whereas lower scores indicate that some content is unsupported and may reflect hallucination or over-generation beyond the retrieved evidence.

Context Precision (3)

$$Context\ Precision\ @K = \frac{1}{K}\sum_{k=1}^{k} v_k \tag{3}$$

(where $v_k = 1$if the k-th retrieved context item is relevant to the query (q), else 0.)

Context precision evaluates precision of the retrieval, this is, how much of the retrieved context is relevant to the query. Given the top-K retrieved context items (chunks, nodes, or records) for a query q, each item is assigned a relevance label (e.g. $v_k = 1$ if relevant, 0 otherwise). Precision@K is the fraction of the top-K items that are relevant. High precision indicates that the retriever returns on-topic evidence with minimal noise, improving the likelihood that the LLM receives a useful context for answer generation.

Context Recall (4)

$$Context\ Recall = \frac{|\{u \in \mathcal{U}(gt): u\ is\ covered\ by\ C\}|}{|\mathcal{U}(gt)|} \tag{4}$$

Context recall assesses retrieval completeness from a generation-oriented perspective, this is, whether the retrieved context contains the information needed to answer the query. The ground-truth (reference answer) gt is decomposed into information units $\mathcal{U}(gt)$ (e.g. facts or claims). Each unit is checked for coverage within the retrieved context C. The metric reports the fraction of reference units that are covered by the retrieved C. High recall suggests that the retriever surfaced most of the evidence needed to reproduce the reference answer, whereas low recall indicates that key information was not retrieved even if the generated answer appears plausible, which often refers to hallucinations.

Answer/Factual Correctness (5)

$$Answer\ Correctness = \frac{|\{c \in \mathcal{C}(a): c\ is\ correct\ gt\}|}{|\mathcal{C}(a)|} \tag{5}$$

Answer correctness, also known as factual correctness, measures alignment between the generated response a and the ground-truth reference answer gt. The answer is decomposed into claims $\mathcal{C}(a)$, and each claim is judged for correctness with respect to gt. The score is computed as the proportion of claims in $a$that are correct relative to the total claims. Scores close to 1 indicate strong factual agreement with the reference, while lower scores indicate incorrect, misleading, or unsupported statements when compared against gt.

Answer Relevancy (6)

$$Answer\ Relevancy = sim(\phi(a), q) \tag{6}$$

Answer relevancy evaluates whether the response addresses the user's query, independent of stylistic similarity to a reference answer. It reflects how focused the generated content is on the information need expressed in q. Operationally, relevancy can be estimated via semantic similarity between the query and a representation of the answer content (e.g., an embedding or LLM-derived representation $\phi(a)$). Higher values indicate that the response content remains on-topic and responsive to q, whereas lower values indicate the answer contains tangential content or generic filler that does not satisfy the query intent.

3.4.3 Human evaluation of automated generated text

To complement the automated RAGAS evaluation, a supplementary human evaluation was undertaken to assess how PM domain experts judged the quality and practical usefulness of the generated responses. Human evaluation is widely used in natural language generation (NLG) research (van der Lee et al., 2019; Howcroft et al., 2020). A purposive sample of 8 PM professionals and academics was invited to participate. Participants were selected because of their expertise and familiarity with PM, project controls, construction management, infrastructure delivery, or portfolio-level decision-making. The use of subject matter expert (SME) evaluators was considered important as human judgements of generated text vary depending on evaluator expertise, task familiarity, and criteria applied (Belz & Reiter, 2006; van der Lee et al., 2019). The purpose of the sample was therefore to provide expert judgement on the practical adequacy of RAG-generated responses in a project and portfolio management context. Given the exploratory purpose of the exercise, the expert panel was considered appropriate for

triangulating automated evaluation results and identifying whether observed metric differences are meaningful from a practitioner perspective.

The human evaluation used a subset of 10 queries selected from the original 20-query benchmark set. This reduced the assessment burden on participants while preserving coverage across the five query types used in the automated analysis: LIST, TOPN, COM, SUM, and FIND. Two queries were selected from each type to ensure that the review included both direct retrieval tasks and more complex tasks involving project comparisons, narrative syntheses, and absence-detection. Limiting queries is also intended to reduce participant fatigue, which can affect the reliability of human evaluation judgements (Howcroft et al., 2020). The evaluation followed a blind paired-comparison design. Participants were shown the full query, the two generated responses, as well as relevant provenance information based on projects from the original dataset. Responses were labelled as Response A and Response B, and participants were not told which architecture had generated each response. The order of the two responses was randomised across the questionnaire to reduce order/position bias. This design reflects recommended practice in human evaluation of generated text, where response presentation and ordering should be controlled to avoid systematic bias in evaluator judgement (van der Lee et al., 2019; Clark et al., 2021).

Each response was rated using five criteria: accuracy, relevance, completeness, trustworthiness, and clarity (Table 1). These criteria were selected to connect established concerns in human evaluation of generated text with the specific requirements of RAG-based decision support, including factual correctness, responsiveness to the query, sufficiency of information, confidence in use, and usability of the response. The use of explicitly defined criteria follows recommendations that human evaluation should avoid single quality judgements and instead provide multiple defined criteria to evaluate each response (van der Lee et al., 2019). Each criterion was rated on a five-point Likert scale, where 1 represented “very poor” and 5 represented “excellent”.

Table 1: Criteria used in the questionnaire (van der Lee et al., 2019; Howcroft et al., 2020).

Criterion	Definition provided in the survey	Justification
Accuracy	The response appears factually correct and does not contain misleading or unsupported statements.	Captures factual adequacy. A concern when comparing generated outputs with automated evaluation metrics.
Relevance	The response directly addresses the query and remains focused on requested information.	Captures whether the response satisfies the user’s information need.
Completeness	The response provides sufficient information to answer the query without important omissions.	Captures coverage and adequacy for task completion.
Trustworthiness	The response gives confidence that it could support preliminary project or portfolio analysis.	Captures practitioner confidence and perceived reliability in a practical context.
Clarity	The response is understandable, structured, and usable by a PM professional.	Captures communicative usability while remaining distinct from factual quality.

Finally, results of the human evaluation were analysed descriptively. Mean scores were calculated for each criterion and pipeline, and overall preference counts were summarised across the full 10-query subset and by query type. Given the small purposive sample and exploratory design, inferential statistical testing was not undertaken. The results were instead used as a triangulation layer to assess whether expert judgement aligned with, qualified, or contradicted the automated RAGAS results. This reporting approach is consistent with best-practice recommendations that human evaluation studies should clearly state their purpose, scale, aggregation procedure, and limits of interpretation, particularly where the evaluation is exploratory rather than confirmatory (van der Lee et al., 2019; Howcroft et al., 2020).

4. RESULTS

4.1 Performance metrics

4.1.1 Latency per query

Figure 5 presents the distribution of query latency for both pipelines, while Figure 6 decomposes average token consumption per query into prompt, completion, and embedding components. These figures illustrate the computational trade-offs introduced by alternative RAG architectures.

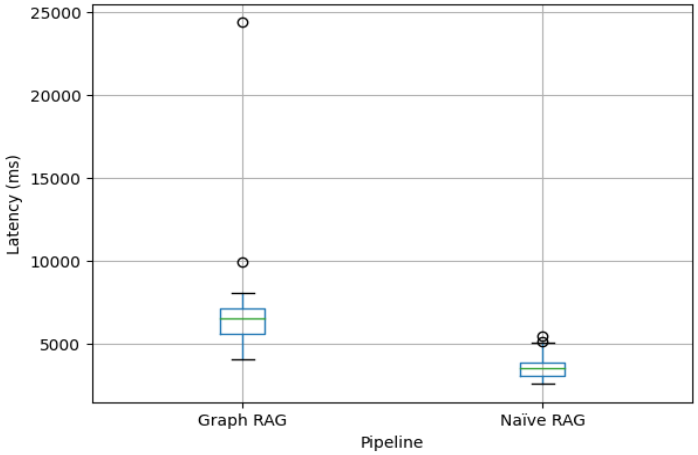

Figure 5: Boxplot of the response latency (ms) for each pipeline.

The vector-based RAG pipeline exhibits consistently lower query latency, with a mean latency of approximately 3.6 seconds and a median of 3.5 seconds per query. Latency variation is limited, with a maximum observed latency of 5.4 seconds, indicating stable and predictable response times across the benchmark query set. This behaviour reflects the relatively simple retrieval workflow, in which dense vector search and lightweight BM25 re-ranking are followed by response generation.

In contrast, the graph-based RAG pipeline incurs substantially higher latency, with a mean latency of approximately 7.3 seconds and a median of 6.5 seconds per query. While many queries fall within a comparable range to the vector RAG pipeline, the latency distribution exhibits a pronounced right tail, with a maximum latency exceeding 24 seconds. These outliers correspond to queries that require more complex graph traversal, aggregation, or multi-step reasoning, such as list-based or comparative queries spanning multiple projects or reporting periods. As a result, latency in the graphRAG pipeline is sensitive to query complexity.

4.1.2 Token usage per query

Figure 6 decomposes average token usage per query into prompt tokens, completion tokens, and embedding tokens using the token accounting recorded during the experiments. In the vector-based RAG pipeline, the embedding token component is constant across queries, indicating that it represents an amortised estimate of the one-off index construction embedding cost, distributed evenly across the 20 evaluation queries. Under this accounting approach, embedding tokens therefore capture a fixed per-query share of index-build cost in addition to query execution.

For the vector-based RAG pipeline, average prompt token usage is approximately 3,354 tokens per query, while completion tokens averaged 176 tokens, indicating that response generation contributes a relatively small share of total token consumption. The dominant component is embedding tokens, averaging approximately 18,197 tokens per query, reflecting the embedding-intensive nature of vector database indexing when index construction costs are amortised across a limited number of queries. As these is a one-off cost, associated with index embedding (once the vector index is developed this can be reused for each subsequent query), at a scale of 100 or 1,000 queries (low volume for a retrieval system), the embedding cost per query would be substantially reduced, vs the 20 queries tested in this paper.

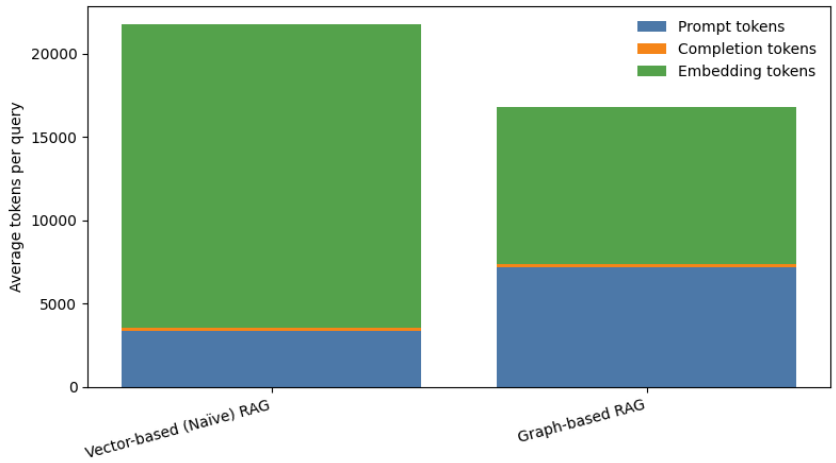

Figure 6: Average token usage breakdown for each pipeline.

For the graph RAG pipeline, average prompt token usage is higher at approximately 7,154 tokens per query, consistent with the inclusion of structured retrieval instructions and graph-derived context in the prompt. Completion tokens remain low at approximately 231 tokens per query, comparable to the vector-based pipeline. Embedding tokens are substantially lower, averaging approximately 9,376 tokens per query, and vary by query, reflecting that graph construction does not rely on embedding-based indexing in the same manner; instead, embedding costs arise primarily at query time where required for retrieval or scoring.

Overall, the token breakdown highlights different cost profiles across the two architectures. The vector-based pipeline exhibits an embedding-dominated cost structure driven by the vector-index construction, whereas the graph-based pipeline shifts a greater share of cost towards prompt construction and query-time processing. This distinction is important when interpreting token consumption as a proxy for operational cost, as amortised index-build costs are sensitive to query volume expected and the needs to refresh the vector-index in deployment scenarios.

4.2 Answer quality and retrieval effectiveness (RAGAS)

Figure 7 presents the average RAGAS scores across the full benchmark query set for the vector-based RAG and graph-based RAG pipelines. Overall, the results indicate that the two architectures exhibit distinct strengths across different dimensions of answer quality, rather than a uniform performance advantage.

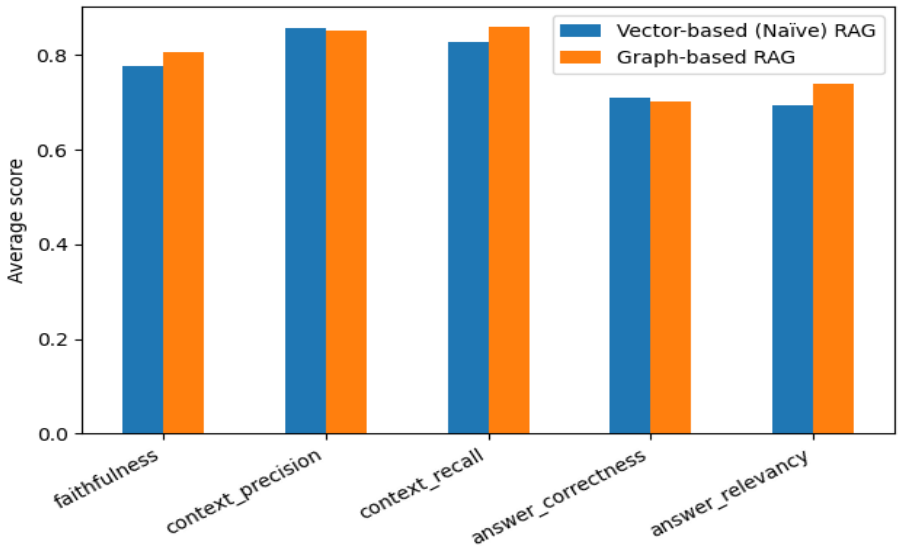

Figure 7: Average RAGAS scores for the benchmarked query set for each pipeline.

The graph-based RAG pipeline achieves higher average scores for faithfulness (≈0.81 vs ≈0.78), context recall (≈0.87 vs ≈0.83), and answer relevancy (≈0.74 vs ≈0.69). These improvements indicate that responses generated using graph-based retrieval are more consistently grounded in the retrieved context, capture a greater proportion of the information required to answer the query, and align more closely with the intent of the user query. In contrast, the vector-based pipeline achieves marginally higher context precision (≈0.86 vs ≈0.85), suggesting that dense vector retrieval combined with BM25 re-ranking remains effective at filtering irrelevant information. Average answer correctness is comparable across pipelines (≈0.71 for vector-based vs ≈0.70 for graph-based), indicating that both architectures can produce factually aligned responses when sufficient context is retrieved.

Taken together, these average results across the 20-query set indicate that graph-based RAG improves coverage and grounding of retrieved information, while vector-based RAG maintains slightly higher precision in scoped retrieval scenarios. To provide more detail, the RAGAS results across pipelines were also compared by query type.

4.2.1 RAGAS performance per query type

To examine how these differences manifest across distinct project-management information needs, Figure 8(a–e) report RAGAS performance disaggregated by query type, grouping queries into LIST type of queries (i.e. queries to retrieve and list relevant projects based on department/project types filters), TOPN (i.e. queries prompting for a rank of projects, for example, identified projects cost overrun or by lower/higher cost variances), COM (queries seeking comparisons across projects or financial years), SUM (queries seeking a narrative summary of projects), and FIND (queries aiming to find and identify key terms within project narratives) query types. Each sub-figure corresponds to one RAGAS metric.

For LIST queries, both pipelines exhibit consistently high performance across all metrics (Figure 8(a)), with faithfulness and answer correctness exceeding 0.9 in both cases. This indicates that enumeration-based tasks involving well-defined entities are well supported by both retrieval architectures. For TOPN queries, which require ranked or ordered responses, graph-based RAG demonstrates a clear advantage in faithfulness (≈0.95 vs ≈0.67; Figure 8(a)) and comparable or slightly higher answer relevancy (Figure 8(e)). Although vector-based RAG achieves slightly higher context recall for this query type (Figure 8(c)), this does not correspond to higher faithfulness, indicating that greater retrieval volume alone does not necessarily improve grounded generation for ranking tasks.

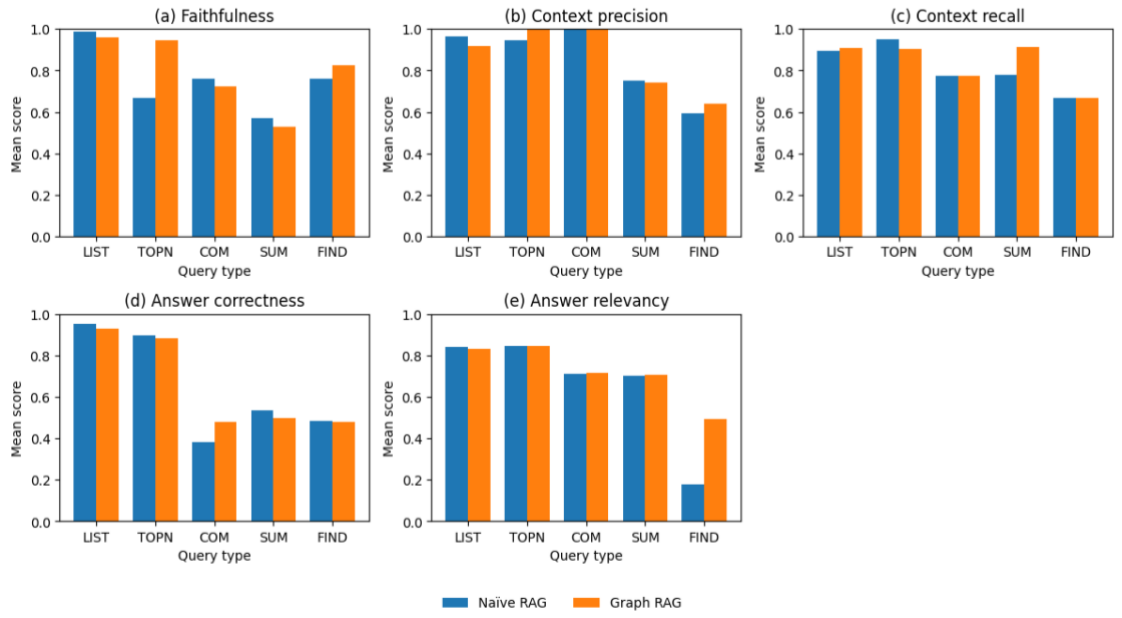

Figure 8 (a-e): RAGAS results per pipeline and query type (averaged).

For the queries were narratives play a key part, starting with narrative comparison (COM) queries, graph-based RAG records higher answer correctness (≈0.48 vs ≈0.39; Figure 8(d)), while faithfulness remains similar across pipelines (Figure 8(a)). This reflects differences in how each architecture supports comparative reasoning across

multiple projects or attributes. For those focused on more general narrative summary (SUM) queries of multiple projects, both pipelines exhibit lower scores across all RAGAS metrics (Figure 8(a–e)), reflecting the greater difficulty of narrative synthesis tasks. Differences between architectures are relatively small, with vector-based RAG showing marginally higher faithfulness and correctness, while graph-based RAG achieves higher context recall.

Finally, those queries requesting to provide summaries for specific terms within the narratives (e.g. “supplier insolvency” or “industrial action”), i.e. FIND queries, graph-based RAG substantially outperforms vector-based RAG in answer relevancy (≈0.49 vs ≈0.18; Figure 8(e)) and also achieves higher faithfulness (Figure 8(a)). In this specific case, the ground-truth answer indicated that no projects in the dataset contained information on supplier insolvency. While both pipelines acknowledged this absence, the prompting constraints required a multi-sentence response. As a result, vector-based RAG produced additional, tangential content to satisfy response length requirements, whereas graph-based RAG more consistently aligned its response with the absence of relevant data, leading to higher relevancy and faithfulness scores under the RAGAS evaluation.

4.3 Expert evaluation results

The supplementary blind expert evaluation was completed by nine PM professionals and academics. The panel included five senior project or programme managers, two PM academics, one public-sector project delivery professional, and one programme management consultant/advisor. Participants were experienced in their PM career; one reported 6-10 years of relevant experience, four reported 11-20 years, and the final four reported 21-30 years.

Across the 10-query subset included in the questionnaire, the graph RAG pipeline received higher mean expert scores across all five human evaluation criteria (Figure 9). The largest differences were observed for completeness, accuracy, and trustworthiness. Graph RAG achieved an overall mean score of 4.10 across all criteria, compared with 3.42 for the vector-based RAG pipeline. This suggests that, when judged against the reference answer and query intent, experts perceived the graph-based responses as more complete, accurate, and suitable for preliminary project or portfolio analysis. The difference for clarity was smaller, indicating that the advantage of graph RAG was not simply a matter of response presentation or readability.

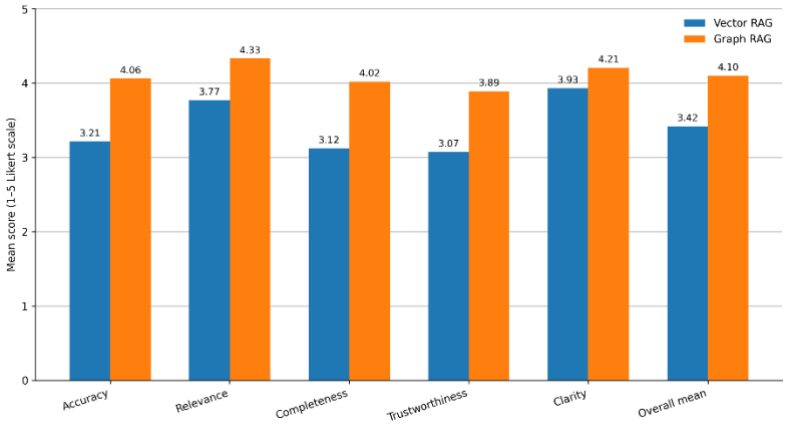

Figure 9: Mean expert evaluation scores per criterion.

Finally, the overall preference results also favoured the graph-based RAG pipeline, but the pattern varied by query type. Across 90 paired comparisons, experts preferred graph RAG in 51 cases, vector RAG in 23 cases, and indicated no clear preference in 16 cases. For the selected LIST and TOPN queries, GraphRAG was preferred in all 36 paired expert comparisons. This indicates a strong preference for structured listing and ranking tasks in the selected human-evaluation. In contrast, preferences were more mixed for COM and SUM queries. For FIND queries, vector RAG was preferred more frequently, while a high number of no-clear-preference judgements was also observed. This is notable because the RAGAS results indicated stronger GraphRAG performance for FIND queries, suggesting that expert judgement was sensitive to aspects of response usefulness that were not fully captured by the automated metrics.

The query-type results show that expert judgement did not indicate universal superiority of one architecture over the other. Instead, human evaluation reinforced the task-dependent nature of the architecture trade-off. Graph RAG was strongly preferred where the task required structured retrieval, ranking, and accurate identification of relevant project records. In these sub-set, respondents ranked the trustworthiness and accuracy of vector RAG poorly (as responses failed to capture the ranks and all projects accurately). The graph RAG capability to extract numerical data proved stronger. However, for narrative comparison, summary, and term-finding tasks, expert preferences were less consistent. This suggests that, while graph-based retrieval improved perceived accuracy, trustworthiness and completeness in many cases, the practical advantage of the architecture depends on the nature of the information need and the way retrieved evidence is transformed into a final answer. Moreover, the human evaluation indicates that graph-based RAG provides stronger performance on quality dimensions associated with answer grounding and completeness. However, the expert results also highlight sharper variations by query type, particularly for FIND queries, where vector RAG responses were often judged as equally or more useful.

Table 2: Expert preference by query type (pairwise comparison results).

Query type (number of pairwise comparissons)	Vector RAG preferred	Graph RAG preferred	No clear preference
Listing (LIST) (n=18)	0	18	0
Ranking (TOPN) (n=18)	0	18	0
Cross-year comparisson (COM) (n=18)	7	5	6
Narrative summaries (SUM) (n=18)	6	9	3
Searches in narratives (FIND) (n=18)	10	1	7
Total (90)	23	51	16

4.4 Performance vs. RAGAS compounded metric

Figure 10(a-b) compares computational performance and answer quality for the vector-based and graph-based RAG pipelines. Figure 10(a) relates query-time LLM token usage to answer quality, while Figure 10(b) contrasts latency against the same quality measure. Answer quality is represented using a composite RAGAS score, defined as the mean of faithfulness, context recall, and answer relevancy. These metrics represent answer grounding, coverage, and alignment with user intent.

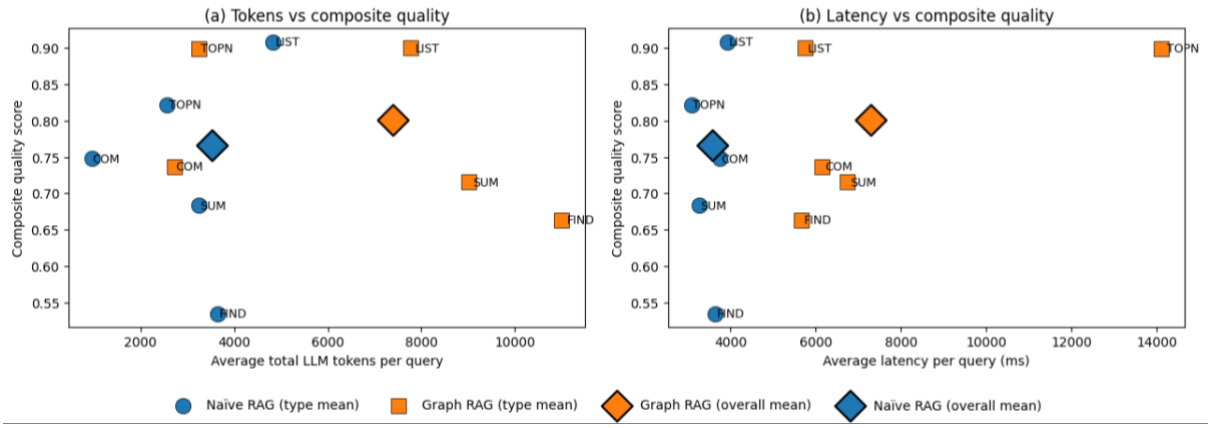

Figure 10 (a-b): Performance-quality trade-off by query type and pipeline.

Token values in Figure 10(a) reflects query-time LLM consumption only. For the vector-based pipeline, embedding and index-construction costs are excluded, and only tokens associated with prompt construction and response generation are included. For the graph-based pipeline, token usage is taken directly from logged LLM calls during query execution. This provides a like-for-like comparison of runtime computational cost.

Across the full query set, the graph-based pipeline achieves a higher average composite RAGAS score, but with substantially higher token usage and latency. On average, graph-based RAG incurs approximately double the query-time token consumption and latency of the vector-based approach, corresponding to a modest but consistent improvement in answer quality. Disaggregation by query type indicates that this trade-off is query dependent. For simpler list (LIST) and ranking (TOPN) queries, vector-based RAG achieves comparable quality at substantially lower cost. For more complex term-related (FIND) and multi project summary (SUM) queries, graph-based RAG attains higher quality scores, although with increased computational overhead. When considered together with the expert evaluation, graph-based RAG can provide higher perceived answer quality, but at increased query-time cost and with advantages that vary by query type.

5. DISCUSSION

5.1 Pipeline, performance, expert evaluation and retrieval trade-offs

The comparative evaluation of vector-based and graph-based RAG pipelines reinforces findings from recent RAG research were retrieval augmentation architectures shape and improve the reliability of LLM outputs for question-answering systems for knowledge-intensive tasks (Lewis et al., 2020; Gao et al., 2023). However, the results of this study do not support a simple interpretation in which one architecture is uniformly superior. Instead, the RAGAS scores, computational performance measures, and expert evaluations indicate more conditional trade-offs. In general terms, graph-based RAG improves perceived and measured answer quality, while vector-based RAG remains more efficient and performs strongly for lower-complexity tasks.

The vector-based RAG pipeline demonstrates the continuing value of comparatively lightweight retrieval architectures. It achieved substantially lower latency and lower query-time overhead, confirming that dense retrieval combined with metadata filtering and lightweight re-ranking (*i.e*. BM25) provides an efficient baseline for project and portfolio question answering (Karpukhin et al., 2020). These lower costs can be particularly important for high-volume or routine information access, where responsiveness or operating cost may be more important than marginal improvements in answer quality. However, the expert evaluation nuanced the interpretation of vector-RAG performance for structured listing (LIST) and ranking (TOPN) tasks. For these queries, experts assigned very low trustworthiness and accuracy scores where a project was omitted from the list or where an incorrect project was hallucinated. Metrics such as RAGAS register such errors as partial factual inaccuracies within a broader response, whereas expert evaluators may treat them as more consequential because they affect the actionability and reliability of the answer.

The graph-based RAG pipeline achieved higher average RAGAS scores for faithfulness, context recall, and answer relevancy, and was also rated more highly by experts, particularly for accuracy, completeness, and trustworthiness. The expert preferences were unanimous across the listing and ranking queries, which strengthens the interpretation that explicit modelling (i.e. knowledge graph) of entities, relationships, and reporting periods can improve both measured and perceived response quality. These findings are consistent with recent studies, which show that graph-based retrieval can improve grounding and contextual coherence for relational queries (Han et al., 2026; Hilel et al., 2025). In this study, the benefit was particularly visible where the answer depended on identifying the correct project records (e.g. LIST or TOPN queries), preserving temporal context, or distinguishing between similar entities. By limiting retrieval to accurate subgraphs, graph-based RAG can reduce risks of retrieving semantically related but contextually inappropriate evidence, a potential limitation of purely vector-based RAG (Gao et al., 2023).

These quality gains are accompanied by additional implementation and operational costs. Even when index-construction embedding costs are excluded, graph-based RAG exhibits higher query token usage and latency, reflecting the overhead of schema-aware retrieval, graph traversal, and processing of structured sub-graph context when answering the query. Beyond this runtime costs/latency, graph-based RAG also requires additional design effort to define, populate, validate, and maintain the graph schema as source data evolve, particularly in contrast to embedding-based vector retrieval, where the indexing process is comparatively easier to automate. Similar trade-offs between robustness, controllability, and efficiency have been reported in modular and structured RAG architectures (Han et al., 2026; Gao et al., 2023). From a systems design perspective, these results therefore indicate that graph-based RAG is most appropriate where answer reliability, entity coherence, and completeness outweigh latency, cost, and maintenance constraints.

As such, numerical differences between architectures should not be interpreted as definitive. Average RAGAS differences across the dataset are modest, and several query categories, particularly narrative focused queries, showed overlapping or mixed performance. The choice of architecture within a PM use-case might be more dependent on the decision context in which the system is used. In low-risk exploratory search or routine portfolio monitoring, a small gain in faithfulness or completeness may not justify additional latency, cost, and the graph development and maintenance costs. In contrast, in governance-sensitive settings, where answers may inform assurance or cross-project learning, improvements in accuracy, completeness, and trustworthiness may be meaningful and justify the additional cost. Moreover, the expert evaluation also shows why automated metrics, such as RAGAS, should be interpreted as one layer of evidence rather than as a definitive measure of practical usefulness. While expert scores broadly favoured graph-based RAG across the selected query subset, the preference results varied by query type. Within narrative queries (*e.g.* FIND queries) expert evaluations differed from the automated RAGAS results, with experts more often preferring vector-RAG responses or indicating no clear preference. This suggests that domain evaluators may value concise absence-detection, more cautious wording, or practical interpretability in ways that are not fully captured by RAGAS metrics. For PM decision support, this distinction is important as an output can score well on automated grounding metrics while still being considered less useful by a practitioner if it is verbose, indirect, or insufficiently clear about the limits of the grounding evidence.

Overall, the findings support a design-contingent rather than architecture-focused interpretation of RAG performance. The question is not whether graph-RAG is generally "*better*" than vector-based RAG, but whether the additional structure, cost, and maintenance burden of graph retrieval is justified. This interpretation aligns with work on adaptive RAG, where retrieval is treated as configurable pathway rather than a fixed component, and where routing mechanisms can select retrieval strategies according to the query, task, or data source (Asai et al., 2023). For routine information retrieval, vector-based RAG remains an efficient and practical baseline. For tasks requiring reliable project identification, ranking, temporal comparison, or relational coherence, graph-based RAG offers clearer advantages. This is also consistent with recent comparative work showing that vector and graph-based RAG exhibit distinct strengths across task types, and that integration strategies may combine the advantages of both approaches (Han et al., 2026). In PM contexts, the implication is therefore not to replace a RAG architecture with another, but to design retrieval workflows that can route queries to the least complex architecture capable of producing a sufficiently reliable answer.

5.2 RAG for project management practice: applicability and implications

Beyond technical performance, the results have implications for the application of RAG-based systems in PM practice. Prior research on artificial intelligence in PM and construction management has largely focused on prediction, optimisation, and automation, including schedule forecasting, risk prediction, and resource allocation (Marzouk & Enaba, 2019; Pan & Zhang, 2021). Comparatively less attention has been given to AI systems that support knowledge-centric querying and sensemaking across heterogeneous project portfolios.

PM is increasingly recognised as a knowledge-intensive activity, involving the interpretation of narrative reports, performance histories, and contextual information across organisational and temporal boundaries (Flyvbjerg, 2014; Locatelli et al., 2023). The present findings demonstrate how RAG architectures can augment this form of work by enabling natural-language questioning of longitudinal, semi-structured project datasets, such as those maintained by public-sector infrastructure bodies. The expert evaluation reinforced this as participants did not assess responses only as technically correct, but as potentially usable answers for preliminary project or portfolio analysis. This highlights the importance of evaluating RAG systems in domain-context using a range of criteria such as accuracy, completeness, trustworthiness and clarity, rather than relying solely on automated retrieval metrics (van der Lee et al., 2019; Howcroft et al., 2020).

The results suggest that vector-based RAG is well suited to routine PM information needs, including broad filtering or exploratory retrieval of projects by status, budget category, or delivery confidence, where responsiveness and efficiency are prioritised. These use cases align with operational PM tasks such as portfolio monitoring and reporting, which dominate existing digital PM tools (Too & Weaver, 2014; Pan & Zhang, 2021). Conversely, graph-based RAG appears particularly relevant for higher-order PM tasks that require contextual reasoning across projects, organisations, and reporting periods. Examples include comparing delivery confidence across departments, tracing changes in project narratives over time, or identifying structurally related risks across a

portfolio. Such tasks align with calls in the PM literature for improved systems to support strategic oversight and cross-project learning, particularly for major and megaprojects (Flyvbjerg, 2014; Locatelli et al., 2023).

From a governance perspective, the provenance and transparency afforded by graph-based retrieval are also significant. Public-sector PM increasingly operates under audit, accountability, and assurance requirements, where the ability to trace analytical outputs back to authoritative project records is essential (NAO, 2025). This is where the methodological design of the study is also practically relevant. The use of reference answers and expert review showed that response quality cannot be judged only by fluency or apparent plausibility of the outputs; it must be assessed against underlying project evidence. In practice, RAG systems intended for PM decision support should therefore expose the basis of generated outputs, including records, reporting periods and assumptions used to produce the response. This is particularly important in contexts where a minor omission, hallucinated output, or unsupported summary could influence project decision-making. In graph-based RAG, the retrieved subgraphs can provide a visually explicit representation of project entities and relationships used to construct the response, supporting traceability and auditability of the generated outputs (Kumar & Teo, 2021).

Nevertheless, practical deployment challenges remain. Graph-based RAG requires upfront investment in schema design, data curation, and ongoing maintenance of structured representations, echoing challenges previously identified in the knowledge graph literature (Hogan et al., 2021; Barrasa & Webber, 2023). As such, adoption should be aligned with organisational data maturity and intended use cases, rather than pursued as a universal solution.

6. CONCLUSION

This paper set out to benchmark how two commonly deployed RAG architectures, (i) a vector-based hybrid pipeline and (ii) a graph-based pipeline grounded on a labelled property graph, trade-off query efficiency against answer quality when responding to knowledge intensive queries over a multi-year, semi-structured major project portfolio dataset. Using a controlled comparative design with a common dataset, fixed query set, identical response-generation settings, automated RAGAS evaluation, and a supplementary blind expert evaluation, the study isolated how retrieval architecture influences efficiency, answer quality, and perceived usefulness in PM contexts.

Across the query benchmark, the findings show that neither architecture is universally superior. Vector-based RAG provides a lower-latency and lower-overhead baseline for routine or exploratory information retrieval, particularly where responsiveness and scalability are prioritised. Graph-based RAG achieved stronger performance on several grounding and coverage-oriented dimensions, including faithfulness and context recall, and was rated more highly by experts for accuracy, completeness and trustworthiness across the selected query subset. However, these advantages were task-dependent rather than uniform. Expert preferences varied by query type, and the results reinforce that automated RAG metrics should be interpreted alongside domain-informed judgement rather than treated as definitive measures of practical usefulness.

The main implication is that system architecture choice should be treated as a design decision, not a ranking exercise. In general project and portfolio management, vector-based RAG may be sufficient for broad filtering, exploratory search and high-volume reporting tasks. Graph-based RAG is more justified where queries require reliable project identification, ranking, temporal comparison, relational coherence, or higher traceability to authoritative datasets. This is particularly relevant in governance-sensitive PM environments, where omitted projects, incorrect inclusions, unsupported summaries, or weak provenance can affect decision-making.

Several limitations should be acknowledged. The evaluation is based on a relatively small, curated query set designed to reflect project and portfolio information needs. Broader benchmarking with a larger and more diverse query set would improve generalisability across additional project control tasks and activities. Second, while RAGAS provides an efficient and interpretable approach to decomposing RAG quality into retrieval- and generation-linked metrics, automated scoring remains dependent on the behaviour of the underlying evaluator model and may inherit associated biases or fail to capture domain-specific notions of correctness or usefulness. In this study, a consistent external evaluator model was used across all pipelines to support fair comparison. Additionally, to mitigate shortcomings of automated evaluations, the results were triangulated with PM professional blind evaluation of a subset of responses. The expert evaluator sample size (n=9) is relatively small limiting the generalisability of the expert evaluation results. Finally, the study holds the LLM foundation model

and prompting constant to isolate retrieval effects. While this is necessary for a fair comparison, results are conditional on the chosen LLM and prompt design and may shift with other models, tool-augmented settings, or alternative context and prompt formatting. Moreover, the graph pipeline depends on a predefined schema and ETL pipeline. Schema design choices, maintenance costs, and organisational data maturity can strongly influence the feasibility and performance of knowledge graph approaches in real-world deployment contexts.

Looking ahead, the results of this paper provide three promising directions for RAG research in construction and project-based industries. First, future studies should develop shared benchmark suites for project and portfolio queries that reflect temporal evolution, departmental restructuring, and mixed narrative and quantitative fields, enabling reproducible comparison across architectures and datasets. Second, there is scope for adaptive and hybrid RAG strategies that route queries to vector retrieval, graph traversal, or combined retrieval paths based on query intent, data structure, complexity and risk. Potentially, such architectures could optimise cost for simple queries while retaining the graph strengths for relational and temporal reasoning needed in more complex queries. Third, governance-oriented evaluation should become more explicit. Beyond accuracy, research should measure provenance quality, auditability, and failure modes under incomplete or missing data. This is particularly important in light of Flyvbjerg's (2025) critique of "*artificial ignorance*", where fluent AI outputs may hide weak evidential grounding or unreliable reasoning in PM contexts. RAG offers a credible response by constraining response generation through retrieved evidence, but this study findings show that RAG alone is not sufficient. Architecture choice, domain-grounded evaluation, expert judgement, information governance and traceable evidence all shape whether RAG systems can support project decision-making and reduce risk, or merely repackage incomplete information in more fluent forms.

DATA AVAILABILITY STATEMENT

The benchmark query set, processed GMPP dataset, and full evaluation outputs (i.e. tokens, latency, retrieval metrics, and RAGAS scores) are available via the Open Science Framework repository

[Available via OSF (anonymised/view-only link for peer review): <https://osf.io/b5gqw/overview?view_only=7a9bcd5b8c194e9c9f6bd50cb437980c>]

REFERENCES

Adebayo, Y., Udoh, P., Kamudyariwa, X. B., & Osobajo, O. A. (2025). Artificial intelligence in construction project management: A structured literature review of its evolution in application and future trends. *Digital*, *5*(3), 26. https://doi.org/10.3390/digital5030026

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8

Barrasa, J., & Webber, J. (2023). *Building Knowledge Graphs: A Practitioner's Guide*. O'Reilly Media.

Belz, A., & Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 313–320. Association for Computational Linguistics. https://aclanthology.org/E06-1040/

Bento, S., Pereira, L., Gonçalves, R., Dias, Á., & da Costa, R. L. (2022). Artificial intelligence in project management: Systematic literature review. *International Journal of Technology Intelligence and Planning*, *13*(2), 143–163. https://doi.org/10.1504/IJTIP.2022.126841

Björk, B.-C. (2003). Electronic document management in construction—research issues and results. *Journal of Information Technology in Construction (ITcon)*, *8*, 105–117. https://www.itcon.org/paper/2003/9

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.5555/3495724.3495883

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4310–4330. https://doi.org/10.18653/v1/2022.acl-long.297

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

Clough, P., & Sanderson, M. (2013). Evaluating the performance of information retrieval systems using test collections. *Information Research*, *18*(2), Paper 582. http://www.informationr.net/ir/18-2/paper582.html

Ehrlinger, L., & Wöß, W. (2016). Towards a definition of knowledge graphs. *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems, CEUR Workshop Proceedings*, *1695*, 13–16. https://ceur-ws.org/Vol-1695/paper4.pdf

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

Flyvbjerg, B. (2014). What you should know about megaprojects and why: An overview. *Project Management Journal*, *45*(2), 6–19. https://doi.org/10.1002/pmj.21409

Flyvbjerg, B. (2025). AI as artificial ignorance. *Project Leadership and Society*, *6*, 100208. https://doi.org/10.1016/j.plas.2025.100208

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning, PMLR 119*, 3929–3938. https://doi.org/10.5555/3524938.3525306

Han, H., Ma, L., Wang, Y., Shomer, H., Lei, Y., Qi, Z., Guo, K., Hua, Z., Long, B., Liu, H., Aggarwal, C. C., & Tang, J. (2026). RAG vs. GraphRAG: A systematic evaluation and key insights. *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 8966–8977. https://doi.org/10.1145/3770855.3817575

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Hilel, M., Karmim, Y., de Bodinat, J., Sarehane, R., & Gillon, A. (2025). Building specialized software-assistant chatbot with graph-based retrieval-augmented generation. arXiv:2511.05297. https://doi.org/10.48550/arXiv.2511.05297

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Labra Gayo, J. E., Kirrane, S., Neumaier, S., Polleres, A., Navigli, R., Ngonga Ngomo, A.-C., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, *54*(4), Article 71, 1–37. https://doi.org/10.1145/3447772

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., & Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs

evaluation sheets and standardised definitions. *Proceedings of the 13th International Conference on Natural Language Generation*, 169–182. https://doi.org/10.18653/v1/2020.inlg-1.23

Huang, S., Mamidanna, S., Jangam, S., Zhou, Y., & Gilpin, L. H. (2023). Can large language models explain themselves? A study of LLM-generated self-explanations. arXiv:2310.11207. https://doi.org/10.48550/arXiv.2310.11207

Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open-domain question answering. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880. https://doi.org/10.18653/v1/2021.eacl-main.74

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248, 1–38. https://doi.org/10.1145/3571730

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Kumar, V., & Teo, E. A. L. (2021). Exploring the application of property graph model in visualizing COBie data. *Journal of Facilities Management*, *19*(4), 500–526. https://doi.org/10.1108/JFM-09-2020-0066

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://doi.org/10.5555/3495724.3496517

Locatelli, G., Ika, L., Drouin, N., Müller, R., Huemann, M., Söderlund, J., Geraldi, J., & Clegg, S. (2023). A manifesto for project management research. *European Management Review*, *20*(1), 3–17. https://doi.org/10.1111/emre.12568

Love, P. E. D., Fang, W., Matthews, J., Porter, S., Luo, H., & Ding, L. (2023). Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. *Advanced Engineering Informatics*, *57*, 102024. https://doi.org/10.1016/j.aei.2023.102024

Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

Marzouk, M., & Enaba, M. (2019). Text analytics to analyze and monitor construction project contract and correspondence. *Automation in Construction*, *98*, 265–274. https://doi.org/10.1016/j.autcon.2018.11.018

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

National Audit Office (NAO). (2025). *Governance and decision-making on mega-projects*. https://www.nao.org.uk/insights/governance-and-decision-making-on-mega-projects/

National Infrastructure and Service Transformation Authority (NISTA). (2025). *Major projects data*. GOV.UK. https://www.gov.uk/government/collections/major-projects-data (accessed 16 July 2025).

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*(1), 11–33. https://doi.org/10.1109/JPROC.2015.2483592

Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv:1901.04085. https://doi.org/10.48550/arXiv.1901.04085

OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, *122*, 103517. https://doi.org/10.1016/j.autcon.2020.103517

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems*, *32*(3), 229–267. https://doi.org/10.1080/07421222.2015.1099390

Rau, D., Déjean, H., Chirkova, N., Formal, T., Wang, S., Clinchant, S., & Nikoulina, V. (2024). BERGEN: A benchmarking library for retrieval-augmented generation. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7640–7663. https://doi.org/10.18653/v1/2024.findings-emnlp.449

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Saka, A., Taiwo, R., Saka, N., Salami, B. A., Ajayi, S., Akande, K., & Kazemi, H. (2024). GPT models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment*, *17*, 100300. https://doi.org/10.1016/j.dibe.2023.100300

Samsami, R. (2024). Optimizing the utilization of generative artificial intelligence (AI) in the AEC industry: ChatGPT prompt engineering and design. *CivilEng*, *5*(4), 971–1010. https://doi.org/10.3390/civileng5040049

Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 162–169. https://doi.org/10.1145/1076034.1076064

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, *36*, 68539–68551. https://doi.org/10.52202/075280-2997

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Taboada, I., Daneshpajouh, A., Toledo, N., & de Vass, T. (2023). Artificial intelligence enabled project management: A systematic literature review. *Applied Sciences*, *13*(8), 5014. https://doi.org/10.3390/app13085014

Thorne, J., & Vlachos, A. (2021). Evidence-based factual error correction. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3298–3309. https://doi.org/10.18653/v1/2021.acl-long.256

Too, E. G., & Weaver, P. (2014). The management of project management: A conceptual framework for project governance. *International Journal of Project Management*, *32*(8), 1382–1394. https://doi.org/10.1016/j.ijproman.2013.07.006

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368. https://doi.org/10.18653/v1/W19-8643

Vanhoucke, M., Coelho, J., & Batselier, J. (2016). An overview of project data for integrated project management and control. *The Journal of Modern Project Management*, *3*(2), 6–21. https://journalmodernpm.com/manuscript/index.php/jmpm/article/view/218

Wang, Y., Hou, Y., Wang, H., Miao, Z., Wu, S., Chen, Q., Xia, Y., Chi, C., Zhao, G., Liu, Z., Xie, X., Sun, H., Deng, W., Zhang, Q., & Yang, M. (2022). A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, *35*, 25600–25614. https://doi.org/10.52202/068431-1856

Wu, C., Ding, W., Jin, Q., Jiang, J., Jiang, R., Xiao, Q., Liao, L., & Li, X. (2025). Retrieval augmented generation-driven information retrieval and question answering in construction management. *Advanced Engineering Informatics*, *65*, 103158. https://doi.org/10.1016/j.aei.2025.103158

Yu, Y., Kim, S., Lee, W., & Koo, B. (2025). Evaluating ChatGPT on Korea's BIM Expertise Exam and improving its performance through RAG. *Journal of Computational Design and Engineering*, *12*(4), 94–120. https://doi.org/10.1093/jcde/qwaf033

Zhao, S., Huang, Y., Song, J., Wang, Z., Wan, C., & Ma, L. (2024). Towards understanding retrieval accuracy and prompt quality in RAG systems. arXiv:2411.19463. https://doi.org/10.48550/arXiv.2411.19463

Zheng, C., Wong, S., Su, X., Tang, Y., Nawaz, A., & Kassem, M. (2025). Automating construction contract review using knowledge graph-enhanced large language models. *Automation in Construction*, *175*, 106179. https://doi.org/10.1016/j.autcon.2025.106179

APPENDIX A: SUB-SET OF 10 QUERIES USED FOR EXPERT EVALUATION*

**The questionnaire for the participants presented the RAG generated responses blindly as Response A / Response B (randomised). This appendix unblinds the corresponding Vector-RAG and Graph-RAG answers. The table is presented for readability while retaining responses. The full set of 20 queries is accessible via OSF:* <https://osf.io/b5gqw/overview?view_only=7a9bcd5b8c194e9c9f6bd50cb437980c>

Table A1: 10 query sub-set (2 of each type) used for the expert evaluation

Query type	Query	Ground-truth / Reference answer	Vector-RAG answer	Graph-RAG answer
LIST	List 2022/23 projects with RAG of Red, sorted by whole-life cost.	1. Future Combat Air System | £12556m 2. Emergency Services Mobile Communications Programme (ESMCP) | £11263m 3. Housing Infrastructure Fund | £4334m 4. Core Production Capability | £3772m 5. PHE Science Hub | £2656m 6. Frontline Digitisation | £1982m 7. T Level Programme | £1754m 8. NS&I Rainbow | £1158m 9. Nature for Climate Fund | £928m 10. Replacement Analytical Project | £712m 11. GP IT Futures Programme | £448m 12. Unity Programme | £214m	Financial Year 2022/23 projects with a RAG: Red Sorted by whole-life cost: Rail Passenger Services | £14480m Future Combat Air System | £12556m Emergency Services Mobile Communications Programme (ESMCP) | £11263m Housing Infrastructure Fund | £4334m Core Production Capability | £3772m New Hospital Programme | £3690m Manston Transformation Programme | £2713m PHE Science Hub | £2656m Frontline Digitisation | £1982m Collective Training Transformation Programme | £1981m T Level Programme | £1754m New Plan for Immigration | £1574m	Financial Year 2022/23 projects with a RAG: Red Sorted by Whole-Life Cost: Future Combat Air System | £12556m Emergency Services Mobile Communications Programme (ESMCP) | £11263m Housing Infrastructure Fund | £4334m Core Production Capability | £3772m PHE Science Hub | £2656m Frontline Digitisation | £1982m T Level Programme | £1754m NS&I Rainbow | £1158m Nature for Climate Fund | £928m Replacement Analytical Project | £712m GP IT Futures Programme | £448m Unity Programme | £214m
LIST	List 2023/24 DESNZ projects with the highest whole-life cost (top 10).	1. Geological Disposal Facility Programme (GDF) | £20300m 2. Smart Metering Implementation Programme | £20177m 3. Social Housing Decarbonisation Fund | £4558m 4. Public Sector Decarbonisation Scheme | £2565m 5. Hydrogen Allocation Round 1 (HAR1) | £2366m 6. Sellafield Product and Residue Store Retreatment	Top 10 in 2023/24 projects in the Department of Energy Security and Net Zero (DESNZ) sorted by Whole-Life cost: Geological Disposal Facility Programme (GDF) | £20300m Smart Metering Implementation Programme | £20177m Energy Bills Support Scheme | £12455m Social Housing Decarbonisation Fund | £4558m Public Sector Decarbonisation Scheme (PSDS) | £2565m Hydrogen Allocation Round 1 (HAR1) | £2366m	Top 10 in 2023/24 projects in the Department of Energy Security and Net Zero (DESNZ) sorted by Whole-Life cost: Geological Disposal Facility Programme (GDF) | £20300m Smart Metering Implementation Programme | £20177m Social Housing Decarbonisation Fund | £4558m Public Sector Decarbonisation Scheme (PSDS) | £2565m Hydrogen Allocation Round 1 (HAR1) | £2366m Sellafield Product and Residue Store Retreatment Plant | £1433m

		Plant | £1433m 7. Home Upgrade Grant and Local Authority Delivery (HULA) schemes | £1300m 8. SIXEP Continuity Plant | £1076m 9. Replacement Analytical Project | £735m 10. Box Encapsulation Plant Product Store (BEPPS2) | £727m	Sellafield Product and Residue Store Retreatment Plant | £1433m Home Upgrade Grant and Local Authority Delivery (HULA) schemes | £1307m SIXEP Continuity Plant | £1076m Replacement Analytical Project | £735m	Home Upgrade Grant and Local Authority Delivery (HULA) schemes | £1307m SIXEP Continuity Plant | £1076m Replacement Analytical Project | £735m BEPPS2 | £727m
TOPN	Top 5 cost variances in 2023/24 within Department for Transport.	1. Midland Main Line Programme | +733% variance 2. Transpennine Route Upgrade | +8% variance 3. A428 Black Cat to Caxton Gibbet | +5% variance 4. A66 Northern Trans-Pennine | +0% variance 5. HS2 Phase 1 | +0% variance 6. Lower Thames Crossing | +0% variance 7. Midlands Rail Hub | +0% variance	Top 5 yearly variances (descending) of projects in the Department for Transport (DFT) in 2023/24: Midland Main Line Programme | 733% Transpennine Route Upgrade | 8% A428 Black Cat to Caxton Gibbet | 5%	Top 5 yearly variances of projects in the Department for Transport (DFT) in 2023/24: Midland Main Line Programme | 733% Transpennine Route Upgrade | 8% A428 Black Cat to Caxton Gibbet | 5% A66 Northern Trans-Pennine | 0% HS2 Phase 1 | 0%
TOPN	Top 7 underspends in 2022/23 within DHSC.	1. ESR Transformation Programme | -47% variance 2. Integrated Single Financial Environment | -22% variance	Top 7 underspends in the Department of Health and Social Care (DHSC) for 2022/23: Digital Transformation of Screening_old | -75% New Hospital Programme | -53% ESR Transformation Programme | -47% Integrated Single Financial Environment | -22% Federated Data Platform | -17% Shared Care Records | -13%	Top 7 underspends in the Department of Health and Social Care (DHSC) for 2022/23: ESR Transformation Programme | -47% Integrated Single Financial Environment | -22%
COM	Provide a narrative comparison for the East West rail project between 2021/22 and 2023/24.	The East West Rail programme includes two projects in the MPA dataset, configuration stage 1 and connection stages 2&3. In 2021/22, CS1 was rated Amber, with progress on milestones such as Bletchley Flyover and Winslow Station, but risks remained around	Between the financial years 2021/22 and 2023/24, the East West Rail project has seen notable developments and changes in its delivery confidence and scheduling. In 2021/22, the project, particularly East West Rail Connection Stage 1, maintained an Amber rating in the Infrastructure Project Authority's Delivery Confidence Assessment, indicating steady progress with	In the financial year 2021-22, the East West Rail Configuration State 1 project was rated Amber, with a whole life cost of £1353m and a budget variance of -20%. The project was on schedule to finish by September 1, 2025, and made significant progress on infrastructure milestones, although operational readiness remained a focus. In contrast, the East West Rail Connection Stage 2 & 3 project was

		integration with HS2, and the year showed a budget variance above 5%. By 2023/24, EWR CS1 continued with an amber RAG, infrastructure works were largely complete, and risks centred on rolling stock delivery, industrial action, and driver availability. For EWR connection stages 2&3, the 2021/22 narrative described early concept and design work with significant uncertainty over scope and costs, reflected in an Amber RAG and estimates in the £4–7bn range. In 2023/24, these stages were re-baselined with statutory consultations planned and enabling works funded, though whole-life cost estimates remained uncertain at £3,970m within a £4–7bn range, and the established end date of 2030-10-09 was similarly uncertain.	infrastructure works like the Bletchley Flyover and a projected completion by September 2025. By 2023/24, the Delivery Confidence Assessment for Connection Stages 2 and 3 improved from Red to Amber, reflecting increased governmental commitment and strategic planning for economic development around the stations. The project's end date for these stages is now set for October 2030, with ongoing efforts to refine delivery options and secure funding. The shift in confidence and scheduling highlights a more structured approach to addressing challenges and ensuring the project's benefits are realized. However, there remains a degree of uncertainty in the overall project schedule, pending further work and public engagement.	rated Red during the same period, indicating greater concerns regarding its delivery confidence. This project was still in the concept and design stage, with a budget variance of -15%, reflecting challenges in refining delivery strategies and securing funding. By the financial year 2023-24, the East West Rail Configuration State 1 project maintained its Amber rating, indicating continued progress but with ongoing risks, particularly around rolling stock and operational readiness. The project remained on track for its completion date in 2025, with a forecasted budget variance of -6%.
COM	Provide a narrative comparison for Crossrail Programme between 2021/22 and 2022/23.	In 2021/22, Crossrail moved from final testing into readiness. The Delivery Confidence Assessment improved from Amber/Red to Amber as mass-evacuation exercises were completed, reliability updates were installed and “shadow running” began. The central section (Paddington–Abbey Wood) was confirmed to open on 24 May 2022, with full end-to-end services still targeted for May 2023. Costs were broadly stable at around £18.8bn with a small uplift driven by station works (notably Bond Street), and a funding shortfall remained	In the financial year 2021/22, the Crossrail Programme saw a decrease in the Infrastructure Project Authority's Delivery Confidence Assessment rating from Amber/red to Amber, indicating an improvement in project delivery confidence. The central section of the line, from Paddington to Abbey Wood, was announced to open on 24 May 2022, with successful completion of large-scale evacuation exercises and system updates to enhance reliability. Despite these advancements, there remained a funding shortfall above the additional £2.97 billion agreed in previous years, which the department aimed to address through negotiations. By 2022/23, the project maintained its scheduled end date of 14 May 2023, with full end-to-end services targeted for May 2023. The	In the financial year 2021-22, the Crossrail Programme was rated Amber, with a Whole Life Cost (WLC) of £18,814m and a variance of -21% compared to the baseline. The project was scheduled to finish on 14 May 2023, with the central section opening confirmed for 24 May 2022. However, there were concerns regarding service reliability that could affect the overall delivery timeline. In contrast, for the financial year 2022-23, the project improved to a Green rating, indicating enhanced confidence in delivery. The WLC slightly increased to £18,815m, but the Anticipated Final Crossrail Direct Costs decreased, reflecting progress towards completion. The successful opening of the Elizabeth line and the introduction of additional services contributed to this positive shift.

		to be resolved through the Transport for London settlement. By 2022/23, the focus shifted to operations and close-out. The rating rose to Green, the central section opened as planned, through-running began in November 2022, Sunday services were introduced, frequency increased, and Bond Street opened in October 2022. A long-term funding settlement with TfL in August 2022 underwrote completion. Whole-life cost stayed essentially flat (c. £18.8bn) while direct costs fell as contracts completed and risks closed. Some performance risk persisted on Network Rail's western route, but improvement plans were under way. Full end-to-end services at 24 trains per hour (peak) were due in May 2023. Overall, the programme moved from late-stage testing and funding uncertainty to running the railway, with higher confidence and costs broadly stable	Baseline Whole Life Cost of the project increased slightly from £18,812.10 million to £18,814.30 million, primarily due to an increase in the Anticipated Final Crossrail Direct Costs.	
SUM	Summarise narrative drivers for Red ratings in 2021/22 for MoD.	In 2021/22, nine major Ministry of Defence programmes were rated Red, reflecting persistent delivery challenges. The Armoured Cavalry Programme faced multiple technical issues with Ajax platforms, including noise and vibration problems, with design modifications under trial and risk to its 2025 completion. Infrastructure programmes such as Clyde and Core Production Capability experienced schedule	In the 2021/22 financial year, several Ministry of Defence (MoD) projects received Red ratings due to various challenges. The MODnet Evolve project faced a Red rating primarily because of the need to pause and replan End User Services, which led to an unsuccessful procurement of Electronic Records Management. The Future Combat Air System was rated Red due to ongoing dependencies from other MOD areas and a shortage of resources impacting skills and capabilities. The Core Production Capability project encountered significant challenges in meeting the required core delivery	In the financial year 2021-22, several projects within the Ministry of Defence (MOD) were rated Red, indicating significant challenges. The Lightning Programme faced slow Force Growth, which hindered the delivery of a second Front-Line Squadron, compounded by affordability issues related to necessary infrastructure. The Armoured Cavalry 2025 project encountered technical difficulties with the Ajax platforms, particularly concerning noise and vibration, impacting the overall schedule. The Sea Venom Programme, extracted from the Complex Weapons Portfolio,

		pressure linked to wider programme dependencies and the criticality of sustaining deterrent milestones. Crowsnest continued to underperform against Initial Operating Capability requirements, with whole-life costs increasing due to a VAT re-assessment. Meanwhile other portfolios including Future Cruise Anti-Ship Weapon, Sea Venom and MODNet Evolve were red-rated for governance, resourcing and re-planning issues, with several details redacted. The Lightning Programme slipped further, with Force Growth delays and affordability issues for supporting infrastructure leading to a revised out-of-service date of 2048 and additional whole-life cost growth from inflationary pressures and revised cost estimates. The New Style of Information Technology (Base) project narrative indicated the rating should have been Amber, highlighting a possible error in reporting. Overall the Red ratings reflected technical challenges, re-baselining, supply chain pressures and affordability constraints.	date for the Dreadnought, crucial for sustaining the Continuous at Sea Deterrent. Additionally, the Clyde Infrastructure Programme faced external challenges such as market shortages of experienced personnel and supply chain issues, despite being well-managed locally.	was also rated Red, with no baseline date available for comparison, indicating governance and assurance challenges. MODnet Evolve's rating was affected by the need to pause and replan End User Services, leading to procurement issues. The Future Cruise Anti Ship Weapon Programme was similarly assessed as Red due to the installation of revised governance arrangements.
SUM	Summarise schedule-related causes of delay in 2023/24 for Infrastructure and Construction projects.	In 2023/24, schedule delays across infrastructure and construction programmes most often came from planning and consent timelines, loss of access to worksites (including industrial action), supply-chain and contractor capacity, policy and assurance dependencies, and difficult	Several infrastructure and construction projects in 2023/24 experienced schedule-related delays due to various factors. The Housing Infrastructure Fund project extended its end date from March 2026 to March 2028 following a revised business case in 2023. The Lower Thames Crossing project saw its completion date pushed from March 2032 to April 2034, primarily due to the rephasing of construction as	In the 2023-24 financial year, several infrastructure and construction projects experienced schedule-related delays due to various factors. The New Hospital Programme, rated Red, faced significant challenges, resulting in a forecast variance of -53%, indicating severe issues in meeting timelines. Similarly, Project Gigabit reported a substantial variance of -65%, attributed to delays in contract start dates and changes in eligible premises,

		site conditions or first-of-a-kind projects. Cultural estate moves slipped where the V&A Museum's completion pushed the Blythe House programme, with contingency added to protect the sale date. The Youth Investment Fund reprofiled delivery as few schemes were ready to commence works, planning decisions took longer, benfeciary capability varied and due diligence extended pre-construction. Rail programme milestones slipped despite overall end dates holding: the East Coast Digital Programme lost engineering access and faced supply-chain constraints, while the East Coast Main Line pushed timetable and resilience works into 2025. Decarbonisation and telecoms funds reported policy-design and approval delays (e.g. Rapid Charging Fund), early-market delivery issues and withdrawals (e.g. Net Zero Hydrogen Fund), and contract-start delays with uneven regional delivery (e.g. Project Gigabit). Across prison-estate projects, planning permission delays, challenging ground conditions and first-generation designs caused delays.	confirmed by a Written Ministerial Statement. The River Thames Scheme faced increased forecast durations for consenting and construction activities, leading to later start and readiness dates. Additionally, the A303 Amesbury to Berwick Down project could not provide a definitive schedule due to ongoing legal proceedings. The Submarine Waterfront Infrastructure Future project maintained its schedule, but timely completion remains at risk due to the pace of construction.	which impacted the overall schedule. The Science Capability in Animal Health Programme also encountered delays, with a variance of -52%, as it worked to finalize its scope and prepare for construction. The Natural History Museum Unlocked project maintained its Amber rating, with no deviations from the planned schedule, but it is still set to complete in 2031, reflecting the long-term nature of such initiatives. The East Coast Digital Programme and the Housing Infrastructure Fund both reported minor variances, indicating some level of schedule stability.
FIND	Summarise challenges faced by projects affected by 'industrial action' in 2022/23.	In 2022/23, industrial action affected rail programmes in different ways. For the East Coast Digital Programme, strike periods reduced engineering access, delaying intermediate	The Rail Passenger Services project faced significant challenges due to ongoing industrial action, which impacted rail services and led to disputes with trade unions. Despite these issues, progress was made with a pay deal secured with Network Rail RMT members, resulting in the	The challenges faced by projects affected by industrial action in the financial year 2022-23 are evident across several initiatives. The Rail Passenger Services project encountered delays due to industrial relations issues, which impacted the overall delivery confidence and

		milestones and raising the risk of additional cost even though the overall end date held. Rail Passenger Services reported continuing disruption to operations from industrial action, alongside wider budget and resourcing pressures; a pay deal with Network Rail staff removed some immediate strike dates, but disputes persisted. For the Rail Transformation Programme, the funding settlement left a tight cost envelope for workforce reforms, increasing the likelihood of industrial action and forcing rescoping and re-baselining to keep reforms on track. Overall, the impact ranged from lost worksite access and milestone slippage to operational disruption and heightened strike risk tied to reform and affordability.	calling off of strike dates in March and April. However, the project still contends with budgetary and resourcing challenges, as well as complexities in delineating the Passenger Services Contract schedule. The Delivery Confidence Assessment rating improved from Red to Amber by the end of the 2022/23 financial year, indicating some stabilization in the operating market. The project's end date was extended from October 2022 to April 2024, reflecting these ongoing challenges. Other projects did not report being affected by industrial action, or the information was not provided.	extended the project's timeline. Similarly, the Rail Transformation Programme also faced challenges, although specific details regarding industrial action were not explicitly mentioned. The Mechanised Infantry Programme reported increased risks due to global supply chain issues, which may have been exacerbated by industrial action in related sectors. In contrast, the New Style of Information Technology Deployed project faced delivery challenges in both Land and Maritime environments, indicating that operational disruptions could be linked to broader industrial action contexts. The Home Upgrade Grant (HUG) Phase 1 project experienced procurement delays and supply chain problems, which could also be influenced by industrial action.
FIND	Find 2022/23 projects with ‘supplier insolvency’ in narratives.	No cases of supplier insolvency were identified in 2022/23. I checked project narratives for explicit insolvency language including various related terms (e.g., entered administration/liquidation, contractor failure/closurem, administrators appointed, cva, ceased trading...). Several records mention supplier or general supply-chain issues, but none indicate insolvency events.	The provided CSV chunks do not contain any projects from the 2022/23 period with narratives explicitly mentioning 'supplier insolvency.' Several projects, such as the Met Office Supercomputing 2020+ Programme and the Mine Hunting Capability, mention issues related to supplier performance and delays, but not insolvency. The Debt Transformation project also notes concerns regarding supplier capacity, but again, insolvency is not specified. If further details are needed, additional data sources or records may be required to identify projects with 'supplier insolvency' in their narratives.	Insufficient data. No records explicitly mention 'supplier insolvency' in the narratives for the 2022/23 projects. The available project narratives focus on various challenges such as procurement delays, supply chain issues, and budget variances, but do not specifically address supplier insolvency. The projects listed include a range of initiatives from energy efficiency upgrades to flood risk management, with varying degrees of delivery confidence and financial performance. Notably, the Automotive Transformation Fund has a significant budget variance of -96%, indicating potential financial challenges, but again, there is no mention of supplier insolvency.