HYBRID INFERENCE QUERY FRAMEWORK FOR RAG-DRIVEN TRUTH-PROMPT PAIRS

Wilhelm Henri R. Alegrado
Jhon Vincent A. Gupo
Alberto Y. Zaldivar Jr.

Operation Excellence – Global Operations
Western Digital Philippines, 109 Technology Ave., SEPZ, Laguna Technopark, Binan, Laguna, Philippines 4024
wilhelmhenri.alegrado@wdc.com, jhon.vincent.gupo@wdc.com, Alberto.Zaldivar@wdc.com

ABSTRACT

Manufacturing operations across Western Digital's various sites generate a large amount of technical documentation, operational reports and other miscellaneous knowledge that traditionally exist not only in siloed environments but also in diverse formats. However, by utilizing LLMs (Large Language Models) with RAG (Retrieval Augmented Generation), organizations now have the capability to create a unified knowledge repository from these previously fragmented knowledge sources to support quick and grounded decision-making in day-to-day manufacturing operations. While LLMs have advanced far in understanding and generating human-like responses, the risk of hallucination remains even with off-the-shelf RAG implementations and thus poses significant risks in manufacturing environments where precision is paramount.

In this study, the authors improved upon Amazon Claude's existing RAG framework by introducing a hybrid inference query framework which combines existing semantic query techniques with keyword-based query techniques to provide a more comprehensive context base for the LLM's response. This framework leverages AWS knowledge base architecture as a PostgreSQL database to execute complex SQL queries. To measure the effectiveness of the hybrid framework, the authors measured and compared the misalignment rate (whether an LLM response aligns with the user ground truth) between the current RAG framework and the new hybrid RAG framework.

The results of the study show that utilizing hybrid inference query reduced misalignment rate from 91.18% to 17.65% across 68 user-submitted truth-prompt pairs. While this represents a substantial improvement, the limited number of truth-prompt pairs does not capture the overall variety of questions submitted to LLM. As the knowledge base expands and the user base grows, continuous optimization of the hybrid inference query framework is needed to maintain and improve overall response quality.

1. 0 INTRODUCTION

Western Digital's global manufacturing network, including HDD plants in Thailand collectively produces over 100,000 enterprise-grade drives per day, underscoring both its scale and the need for seamless coordination. Such geographic dispersion enhances supply-chain resilience and market responsiveness but also fragments critical operational knowledge across disparate systems and formats, creating data silos that jeopardize real-time analytics. Engineering designs, quality-assurance records, and maintenance logs often reside in isolated PLM databases, ERP modules, or ad-hoc spreadsheets, impeding unified access and elongating decision cycles. As manufacturing precision and complexity continue to rise, there is an urgent need for a decision-support mechanism that can rapidly consolidate, verify, and interpret heterogeneous data sources without risking ungrounded or hallucinated insights.

1.1. Background & Motivation

1.1.1, Overview of Western Digital's distributed manufacturing footprint

Western Digital operates multiple manufacturing sites worldwide, including HDD production plants in Thailand, wafer fabs in the U.S., and facilities in Japan, China, Malaysia, and Philippines. These sites leverage Industry 4.0 technologies such as connected sensors, digital twins, and real-time analytics to optimize throughput and quality across a geographically dispersed network. In 2022, Western Digital reported that its facilities produced over 100,000 enterprise-grade drives per day in Thailand alone, underscoring both the scale and the criticality of tightly coordinated operations.

1.1.2, Challenges from siloed, heterogeneous documentation and formats

Large manufacturing organizations frequently contend with content silos where engineering drawings, test reports, and maintenance logs reside in disparate repositories. Each using different schemas and file formats. This heterogeneity

hampers cross-site traceability; for example, a single part's design revisions may be documented in PLM systems, while validation data are stored in a separate quality-management database, making unified retrieval laborious. Moreover, unstandardized metadata and evolving documentation practices over decades introduce schema mismatches and interoperability gaps, delaying root-cause analysis when anomalies surface.

1.2. Problem Statement

1.2.1. Need for rapid, accurate decision support in high-precision manufacturing

High-precision manufacturing demands decision cycles measured in seconds or minutes, whether adjusting spindle speeds for micrometer-level tolerances or re-sequencing test slots to avert bottlenecks. Digital twins and real-time process visualization architectures have shown that embedding computational intelligence directly into operational workflows can reduce defect rates and improve throughput—but only if the underlying data queries and model inferences are both fast and trustworthy.

1.2.2. Limitations of vanilla LLM and off-the-shelf RAG pipelines (hallucination risk)

Standard LLM deployments, when fed only their training-data distributions, frequently produce confident yet incorrect outputs, a phenomenon known as hallucination which is intolerable in precision manufacturing contexts where errors can cost millions in scrap or downtime. Out-of-the-box RAG systems mitigate some hallucinations by retrieving external documents, but they do not inherently verify the credibility or timeliness of those sources; misaligned or outdated retrievals can still lead to misleading model responses.

Furthermore, vanilla RAG pipelines typically lack a tightly coupled verification layer to cross-check model-generated assertions against authoritative operational data, leaving a residual risk of ungrounded inference.

2. 0 REVIEW OF RELATED WORK

Large language models (LLMs) such as GPT-4 and its variants have recently demonstrated the capacity to transform manufacturing operations by automating the extraction and summarization of maintenance logs and standard operating procedures (SOPs), with emerging studies showing that RAG-powered interfaces reduce operator search times by over 40 % while maintaining high answer accuracy. Akos Nagy et al.[1]. In industrial settings, frameworks integrating LLMs like GPT and Claude-Opus have been deployed to parse unstructured maintenance records and technical manuals, converting them into structured action plans for preventive maintenance and troubleshooting[2]. Real-time decision-support systems further combine sensor data, historical fault logs, and procedural documentation into a unified conversational interface, enabling rapid root-cause analysis and maintenance scheduling. A notable case study on smart factory operations by Manjurul Islam et al.[3] illustrates how LLM agents can not only retrieve SOP steps but also recommend optimized task sequences for assembly line reconfiguration, thereby improving throughput by 15 % on average.

To ground LLM outputs in factual data and mitigate the well-documented risk of hallucinations, Shailja Gupta et al.[4] study shows that Retrieval-Augmented Generation (RAG) architectures have been adopted, wherein an initial retrieval step fetches relevant documents from a domain-specific corpus before conditioning the LLM's generation on those documents. Early RAG implementations, inspired by the work of Lewis et al. and subsequent surveys, utilize a dual-encoder model to produce dense embeddings for both queries and documents, enabling semantic retrieval via approximate nearest-neighbor search in vector stores such as FAISS[5]. State-of-the-art RAG pipelines often incorporate a re-ranking component, typically a transformer-based cross-encoder that refines the initial retrieval results by evaluating the semantic alignment between the query and each candidate passage, significantly boosting downstream answer accuracy[6].

Despite these advances, hallucinations where the LLM fabricates details do not present in the retrieved documents remain a critical challenge for safety-critical manufacturing applications. A recent comprehensive survey highlights that hallucination rates in RAG systems can exceed 20 % when retrieval quality is suboptimal, underscoring the need for robust evaluation metrics beyond standard IR or generation scores[7]. To address this, Shane Connelly[8] share methods such as open-source Hallucination Evaluation Models (HEM) have been proposed, assigning quantitative "hallucination scores" to generated outputs and enabling continuous monitoring and fine-tuning of the retrieval-generation pipeline. Additional mitigation strategies presented by Lei Huang et al.[7] include multi-pass retrieval where the query is reformulated and reissued to capture diverse context—and automated fact-checking modules that cross-validate LLM responses against a secondary document store.

A key determinant of RAG performance lies in the choice of retrieval technique. Sparse keyword search using BM25 excels at rapid matching of exact terms and scales efficiently to corpora of millions of documents with minimal hardware requirements[10]. However, it suffers from mismatch vocabulary when queries use synonyms or paraphrases are not present in the stored text. Dense semantic retrieval, by

contrast, encodes both queries and documents into a shared vector space where semantic similarity is measured via cosine distance, thereby capturing conceptual relationships but at the expense of higher computing costs and potential retrieval of contextually irrelevant passages[11]. Hybrid retrieval techniques leverage the complementary strengths of both approaches: one common pattern prunes a large candidate set with BM25, then re-ranks the reduced set using dense embeddings, achieving superior recall without sacrificing precision.

Looking forward, continued research is focusing on self-adaptive retrieval strategies that dynamically adjust the sparse-dense balance based on query characteristics, as well as the incorporation of real-time sensor streams into RAG corpora for true conversational Digital Twin experiences[12]. There is also growing interest in multi-modal RAG combining text, CAD drawings, and sensor imagery to support cross-format troubleshooting in smart factories[1].

Overall, by uniting the precision of keyword search, the breadth of semantic retrieval, and the generative power of LLMs, hybrid RAG systems are poised to deliver reliable, explainable, and high-performance decision support across the full spectrum of manufacturing operations.

3.0 METHODOLOGY

3.1 Dataset Preparation

For this study, a knowledge base is first established using 9 document sources ranging from weekly product development reports to organizational policy documents in PowerPoint and PDF format. Each document is ingested to the knowledge base by converting each page/slide to an image to be interpreted as text by an LLM agent. The text-converted information is finally stored as vector embeddings in the knowledge base.

In parallel, document owners supply "truth–prompt" pairs (a representative user query plus the expected ground-truth answer) for evaluation. These pairs form the test set on which we compare the baseline semantic-search RAG pipeline against our hybrid inference query framework.

3.2 LLM Context Retrieval Components

Before going to the overall architecture, an understanding of the individual retrieval components used for both the base line and hybrid inference frameworks must first be established.

3.2.1 LLM Context Retrieval Components

Semantic search uses vector embeddings generated from document text and user queries. These embeddings capture contextual meaning, enabling the system to retrieve relevant passages even when the user's prompt doesn't share exact keywords with the source content. A nearest-neighbor search is performed in the embedding space to identify the most similar aligned documents. Semantic similarity is usually expressed in terms of distance. In this study, a lower value indicates higher similarity.

Keyword search, by contrast, relies on direct textual matches between the query and the knowledge base. Using SQL full-text search and pattern-matching (e.g., LIKE), it is effective for retrieving documents containing specific terms, acronyms, or structured phrases that may be underemphasized in semantic space. In this study, keyword search relevance is measured by the number of matched keywords from a list of relevant keywords per user prompt.

The hybrid inference query framework combines both semantic search and keyword search to give the LLM a broader and more accurate set of context documents to generate grounded and reliable responses.

```
SELECT
id,
chunks,
similarity (embeddings comparison),
keyword_count (no. of matched keywords per row)

ORDER BY keyword_count DESC, similarity ASC
```

Fig. 1. Hybrid Inference Query SQL Template. Both semantic search and keyword search

Fig. 1 shows the SQL query template for the hybrid inference query. In this query statement, similarity is calculated per row from the user prompt embeddings and the stored vector embeddings constituting the semantic search portion of the framework. Meanwhile, keyword count counts the number of matched keywords from a list of relevant keywords. The results are sorted by keyword count in descending order first follow by similarity in ascending order. The query thus returns the most relevant rows prioritizing the greatest number of matched keywords first followed by the most similarity (lower distance value).

Finally, a reranking step is performed to further filter rows based on lexical relevance and semantic alignment. First, each retrieved row is assigned a keyword-based relevance score—using TF-IDF computed weights to highlight terms

that are both frequent in the chunk and uniquely informative across the corpus. The results are filtered to retain only the highest 20 % rows based on this score, focusing on the most lexically pertinent passages. Lastly, the filtered results from the previous step are further truncated to retain only the lowest 20% rows by similarity value, focusing on rows that match the overall query the most in meaning. This two-step process ensures that only rows with both high keyword importance and strong conceptual alignment are passed to the LLM.

3.3 RAG Architecture Comparison

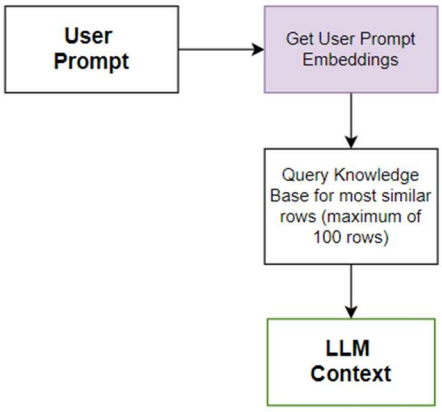

Fig. 2. Baseline RAG Semantic Search Framework. In a standard semantic search, vector embeddings are generated from user prompt and the most similar rows (rows with the least Euclidean distance are returned). The maximum number of rows that can be returned is 100 (as stated in the AWS documentation).

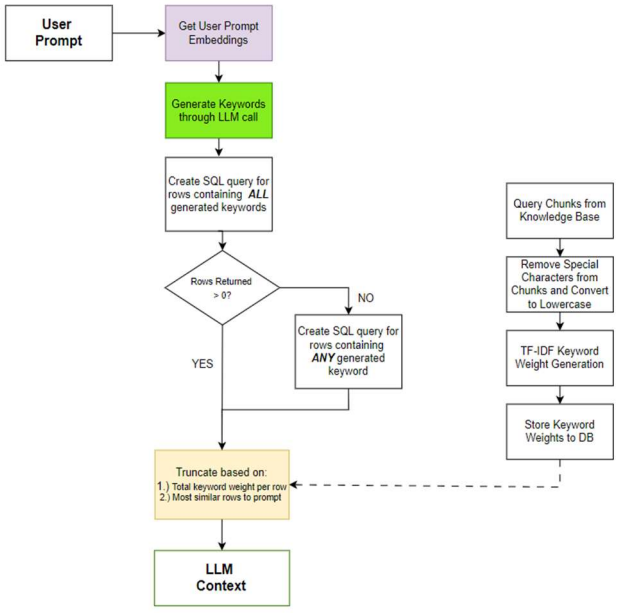

Fig. 3. Hybrid Inference Query Framework. Relevant keywords in addition to embeddings are generated from a given user prompt. A custom query is then generated incorporating both embeddings similarity and keyword matching to be executed against the knowledge base. Finally, a truncation step is added which uses calculated keyword weights and embedding similarity to further maximize the relevance of the retrieved context.

Figs. 2-3 show the overall framework for a baseline RAG Semantic Search Framework and the hybrid inference query framework used in this study respectively. The hybrid inference query framework builds upon the baseline framework by augmenting embedding similarity search with a keyword matching search and reranking filter.

3.4 Framework Evaluation

To evaluate the quality of each RAG framework, the study uses misalignment rate as an evaluation metric. In particular, misalignment rate refers to the number of misaligned responses over the total number of LLM responses. Each response is evaluated by a separate LLM agent into 4 categories based on the amount of matching information between the LLM response and the ground truth.

Table 1. Response Evaluation Categories

Response Category	Definition
Misaligned	Does not match the expected answer or contains inconsistencies
Subset	Partially correct but misses key information
Exact	Fully matches or captures the intended meaning of the expected answer
Superset	Correct but includes extra information

Table 1 shows the response evaluation categories. An exact and superset response is preferrable over a subset response and especially a misaligned response.

3.5 Evaluation Truth-Prompt Pairs

Prompt: What is the build production status?

Truth: Awaiting confirmation from testing team

Prompt: What is the required action if damage is found during machine inspection?

Truth: Operator to immediately inform technician and supervisor for further action.

Prompt: As a tool, what x-ray diffraction reveals about materials?

Truth: Phase, crystal structure and crystallinity

Fig. 4. Truth-Prompt Pairs for Evaluation. Truth prompt pairs are submitted by users along with documents for ingestion to the knowledge base. These pairs are used to evaluate if the LLM response aligns with the user's submitted ground truth.

Figure 4 shows sample pairs across 68 user-submitted truth-prompt pairs. These pairs were submitted along with the documents for ingestion to the knowledge base and serve to evaluate if the resulting LLM response from both the baseline semantic search framework and hybrid inference query search framework aligns with the user's submitted ground truth or not. These questions were sourced across various types of documents such as operating manuals, technical documentation and development updates across different departments in the organization.

4.0 RESULTS AND DISCUSSION

Table 2. Response Evaluation Comparison

Response Score	Baseline Semantic Search	Hybrid Inference Query Search
Misaligned	**91.18%**	**17.65%**
Subset	0%	10.29%
Exact	1.47%	11.76%
Superset	7.35%	60.29%

Table 2 shows a performance comparison between the baseline semantic search framework and hybrid inference query search framework. Using only semantic search, the baseline RAG framework performs poorly with a misalignment rate of 91.18% in retrieving the correct context over the defined knowledge base. The high number of misaligned responses can be broken down into two cases. First, the returned context does not provide any relevant information at all to answer the question. Second, the returned context contains similar information and the LLM misleadingly uses this to answer the question but ultimately does not come from the correct document source leading to a misaligned response. Using semantic search on its own means that the framework is highly dependent on the quality of the context contained inside the user's question. If the context is too vague or generic, the context retrieved may not be fully relevant to the user's expected answer.

On the other hand, the hybrid inference query search with keyword search incorporated shows a substantial reduction in the misalignment rate down to 17.65%. In addition, an increase in the percentage of Subset, Exact and Superset can be observed as previously misaligned responses have been recategorized under the hybrid inference query framework.

The improvement can be attributed to the utilization of keyword search as it alleviates the sensitivity to user question quality. As long as unique keywords are found in the user question, the hybrid inference query framework considerably boosts the context retrieval power of any LLM-RAG framework.

For the remaining misaligned responses, the questions used in the misaligned responses have either too few unique keywords or keywords that are too common in the knowledge base and thus gives too broad of a context for the LLM to be useful in answering the user's question. This observation shows there is still further room for improvement in the hybrid inference query framework.

5.0 CONCLUSION

In this study, the authors demonstrated that a hybrid inference query framework combining both standard RAG semantic search with keyword search together with a reranking filter was able to substantially reduce the misalignment rate from 91.18% to 17.65%. The improvement is attributed to the advantages of keyword search being able to alleviate the weaknesses of the standard semantic search.

A lower misalignment rate provides users with more grounded and confident answers that can be used to drive decision-making in manufacturing operations. It also instills more confidence in the reliability of an LLM-RAG pipeline, leading to a potential increase in the user base.

However, the evaluation on 68 truth–prompt pairs represent a limited sample and may not capture the full diversity of potential manufacturing queries. As the knowledge base expands and new query types emerge, ongoing assessment and optimization of the hybrid framework will be essential to sustain and further improve response quality.

6.0 RECOMMENDATIONS

As the study was conducted with a small truth-prompt pair sample size on a moderately-sized knowledge base, monitoring the performance of the hybrid inference query framework should be the focus of any future continuations. Response quality, through misalignment rate, may vary as more types of documents are added to the knowledge base and more diverse questions are submitted by users.

In response to possible fluctuations in response quality, more advanced retrieval techniques may be needed to maintain or improve response quality.

Finally, the study focuses on optimizing the data retrieval in an LLM-RAG pipeline, but response quality can also be improved through prompt engineering and reinforcement learning. Clarity and reasoning in prompts greatly enhance response quality by tailoring the LLM response to the user's provided context and logic. Incorporating reinforcement learning enables the framework to dynamically personalize responses by learning user preferences and behavior over time.

7.0 ACKNOWLEDGMENT

The authors would like to thank Ma'am Janoah Delos Santos, Sir Albert Zaldivar, Sir Chee Kheng Lim and all the team members who worked on Western Digital's GenAI solution. Their support and passion for innovation and digital transformation has continually provided the authors with the freedom and confidence to explore and deliver on this study.

8.0 REFERENCES

1. A. Nagy, Y. Spyridis, and V. Argyriou, "Cross-Format Retrieval-Augmented Generation in XR with LLMs for Context-Aware Maintenance Assistance," *arXiv.org*, Feb. 21, 2025. https://arxiv.org/abs/2502.15604

2. Y. F. Zhao, E. Niforatos, T. Custis, Y. Lu, and J. Luo, "Large language models in design and manufacturing," *Journal of Computing and Information Science in Engineering*, pp. 1–6, Dec. 2024, doi: 10.1115/1.4067319.

3. M. M. Manjurul Islam, ***Springer Series in Advanced Manufacturing***. Large Language Models (LLMs) for Smart Manufacturing and Industry X.0, *March 2025*, pp 97-119.

4. *S. Gupta, R. Ranjan, and S. N. Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): evolution, current landscape and future directions," arXiv (Cornell University), Oct. 2024, doi: 10.48550/arxiv.2410.12837.*

5. *N. N. Doan, A. Härmä, R. Celebi, and V. Gottardo, "A Hybrid Retrieval Approach for Advancing Retrieval-Augmented Generation Systems," ACL Anthology, Oct. 01, 2024. https://aclanthology.org/2024.icnlsp-1.41/*

6. Adnan Masood, Re-Ranking Mechanisms in Retrieval-Augmented Generation Pipelines. *April 2025.*

7. *L. Huang et al., "A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2311.05232.*

8. Shane Connelly. Measuring Hallucinations in RAG Systems. *November 2023*.

9. Akash Desai. Hybrid Search: Combining BM25 and Semantic Search for Better Results with Langchain. *December 2023.*

10. Aaron Tay. Boolean vs Keyword/Lexical search vs Semantic, keeping things straight. *November 2023.*

11. Djordje Grozdic, Oleg Smirnov. Transforming business process automation with retrieval-augmented generation and LLMs. *October 2023*

9.0 ABOUT THE AUTHORS

Wilhelm Henri R. Alegrado is a Sr. Data Scientist at Western Digital Storage Technologies for Operation Excellence. He holds a master's degree in data science at Asian Institute of Management. He has been with the company for 13 years with experience from previous departments with an Engineering and Development background.

Jhon Vincent A. Gupo is an Associate Data Scientist at Western Digital Storage Technologies for Operation Excellence. He holds a bachelor's degree in computer science at Laguna State Polytechnic University – Los Baños.

Alberto Zaldivar is a Technology Lead at Western Digital, with a total of 20 years of experience. He holds a Bachelor of Science in Computer Engineering from Adamson University.