



AI POWERED DOCUMENT SOLUTIONS FOR ENTERPRISES

Infosys Agentic Foundry, built on Infosys Topaz and accelerated by NVIDIA AI Enterprise, enables a new generation of agent-driven AI solutions with continuous learning and human-in-the-loop intelligence. The first outcome of this collaboration is an Infosys AI-powered document solution, enabled by NVIDIA AI Enterprise, which is built to interpret highly complex, multi-modal documents that span thousands of pages and traditionally require deep domain expertise. Rather than a static RAG pipeline, this is a coordinated system of specialized agents that ingest, interpret, retrieve, verify, and evaluate—delivering accurate, grounded, and actionable responses at enterprise scale. As part of the end-to-end solution, oil and gas documents were tested and inferred, which yielded higher accuracy and domain context response.

Contents

Abstract

1. Introduction	01
2. Solution approach	02
2.1. Common challenges and pain points	02
2.2. Infosys Agentic Foundry: The orchestration backbone	03
2.3. NVIDIA AI Enterprise: AI software for accelerated agent development	03
2.4. Architecture at a glance (agent-first)	04
2.5. Oil and gas reports extraction summary	05
2.6. Metrics evaluation summary	06
2.7. Production-deployable models used	07
2.8. Infosys Agentic Foundry tools	08
3. AI-powered document solution	09
3.1. Built for enterprise	09
3.2. Ensures data security	09
3.3. Unlocking business value from legacy documents	09
3.4. Lower TCO – NVIDIA infrastructure	09
3.5. Generalization across industries	09
3.6. Extensibility	10
3.7. Implementation	10
Conclusion	11
Authors	12
References	14

1. Introduction

The AI-powered document solution is powerful, domain-agnostic, and designed to automate the ingestion, interpretation, and extraction of insights from complex, multi-format documents. It leverages advanced AI techniques including multimodal understanding, context-aware extraction, and reranking strategies to

handle diverse data types such as text, images, plots, and tables.

Most systems stop at retrieval-augmented generation. We go further with agentic design patterns that make the solution reliable, auditable, and adaptable.



- **Planner-executor orchestration:** An orchestrator agent decomposes the user intent, selects tools dynamically (embedding, OCR, reranking, LLM/VLM), and manages multi-step reasoning.
- **Specialist agents:** Dedicated agents for ingestion and structure extraction, retrieval, answer generation, verification, and evaluation, each with a clear role and policy guardrails.
- **Self-verification and critique:** Evidence consistency checks, reranking, and evaluator agents ensure responses are grounded before delivery.
- **Human-in-the-Loop (HITL):** Expert review cycles and feedback capture reinforce reliability and accelerate iterative improvement.
- **Enterprise observability:** End-to-end traces, metrics, and governance controls provide auditability, compliance, and performance tuning.

2. Solution Approach

2.1. Common challenges and pain points

Enterprises contain structured, semi-structured, and unstructured documents of different formats. These have content with text, images, and charts of different formats. Extraction of the formation along with domain context poses significant challenges.

The oil and gas industry faces significant challenges in managing and interpreting highly complex, multi-format data. Reports often come as annotated PDFs spanning 100 to 1,500 pages, with over 300 documents exceeding 100 GB in total size. These documents include diverse data types such as text, images (well lithology, wellbore schematics, and well completion diagrams), and plots or graphs like drilling logs, test logs, telemetry data, and compressed well log images. The complexity is further compounded by metadata variations, multi-page tables, and handwritten notes.

Accuracy drops when dealing with such heterogeneous data, making automation critical. Streamlining data ingestion and extraction requires advanced techniques like multiple embeddings, reranking strategies, and context-aware processing to maintain relevance during interactions. Integration with external sources for continuous synchronization of standard reports and knowledge bases is essential. Verification remains a major pain point, as validating large document sets is time-consuming and demands domain expertise. Additionally, organizations need a standardized UI for repository visibility, automated email reporting on accuracy and ingestion metrics, and performance-oriented systems that can scale for large user bases. Addressing these challenges is key to improving efficiency and reducing manual effort in data-intensive domains like oil and gas.

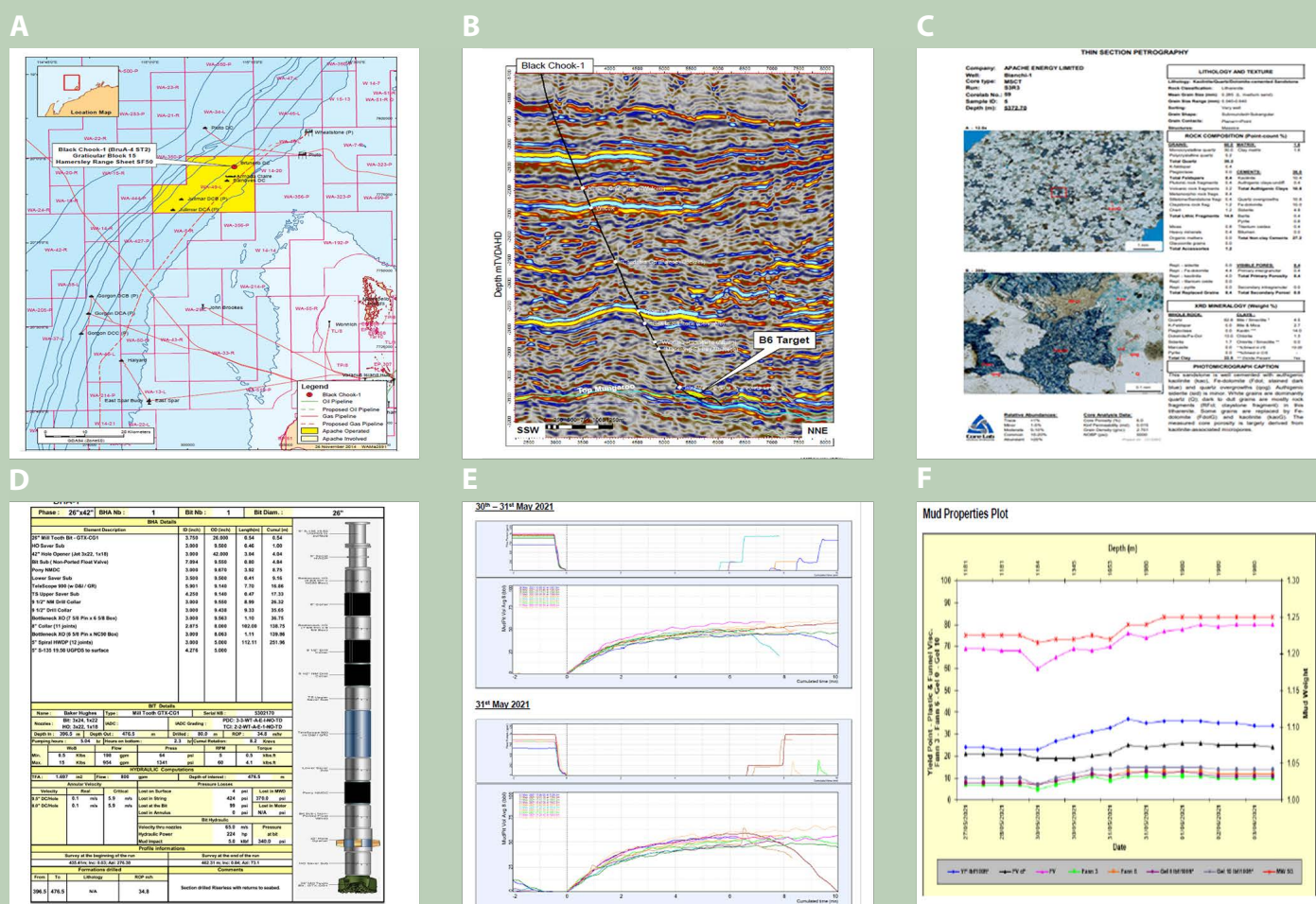
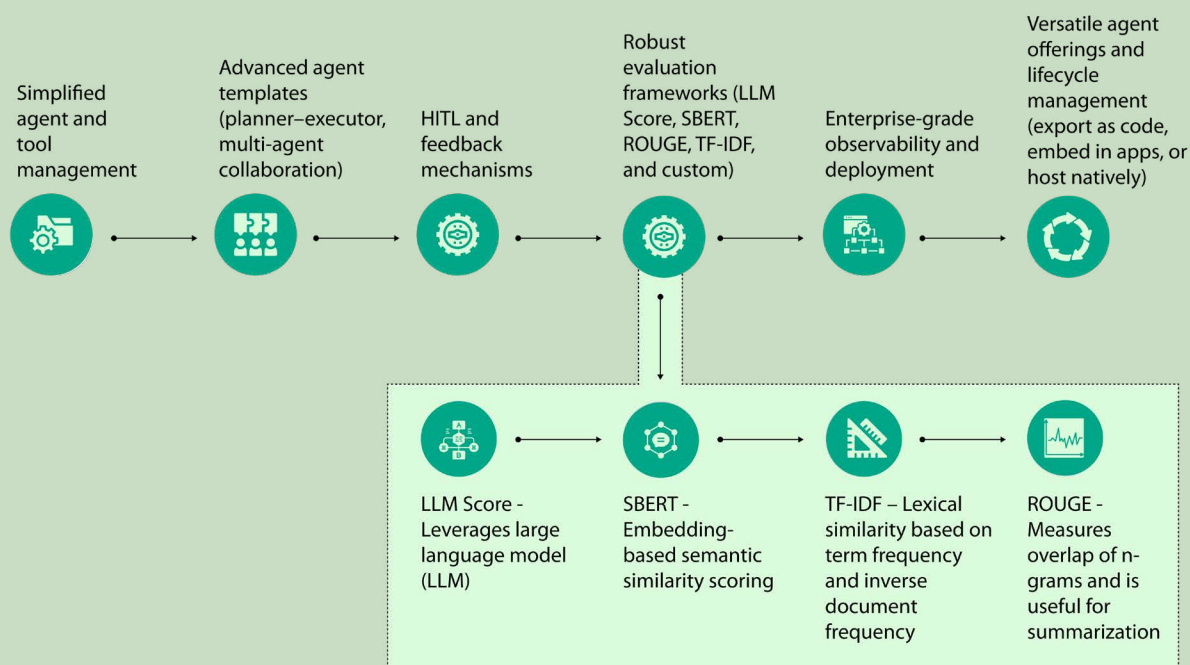


Figure 1: Sample images of oil and gas drilling documents containing images with charts of different formats - (a) to (f)

2.2. Infosys Agentic Foundry: The orchestration backbone

Infosys Agentic Foundry is a robust, open platform to build, configure, and deploy customizable AI agents with minimal coding and enterprise-grade governance.

Below are the key capabilities:



2.3. NVIDIA AI Enterprise: AI software for accelerated agent development

NVIDIA AI Enterprise delivers , production-grade software for AI agent development in accelerated environments on-prem, in the cloud, hybrid or at the edge.

Highlights leveraged in this solution:

Multimodal PDF extraction (text, tables, charts, and graphics/ infographics)

Optional VLMs for visual reasoning and image captioning

GPU-accelerated index creation and search

Leverage of embedded RAG design patterns like CRAG, self-RAG, and adaptive RAG



Custom metadata and multi-collection searchability

Reranking to improve precision and relevance

Document summarization and multi-session support

2.4. Architecture at a glance (agent-first)

- **User Query** - Sent as a JSON request through Infosys Agentic Foundry with domain/RAG parameters and policy guardrails
- **Planner/Orchestrator Agent**- Interprets intent, creates a plan, and routes to specialist agents
- **Ingestion and Structuring Agent**- Uses NVIDIA NeMo™ Retriever NIM to extract text/tables/graphics; normalizes metadata and collections
- **Retrieval Agent** - Computes embeddings, runs vector + metadata search; applies reranking for precision
- **Domain LLM Agent** - Generates a precise, contextual response with citations and structured outputs
- **Verifier Agent** - Cross-checks evidence and consistency; evaluates against evaluator policies
- **Evaluator Agents** - Score via LLM Score, semantic SBERT similarity, TF-IDF cosine, ROUGE-1/ROUGE-L, HumanEval, and more
- **HITL Review (optional)** - Expert review thresholds; feedback loops update policies/prompts
- **Observability and Metrics Store** - All actions traced; metrics persisted for audits, tuning, and release gating
- **Final Delivery** - Grounded, verifiable answer with confidence signals and citations

Figure 2 provides an integration between Infosys Agentic Foundry and NVIDIA RAG Enterprise.

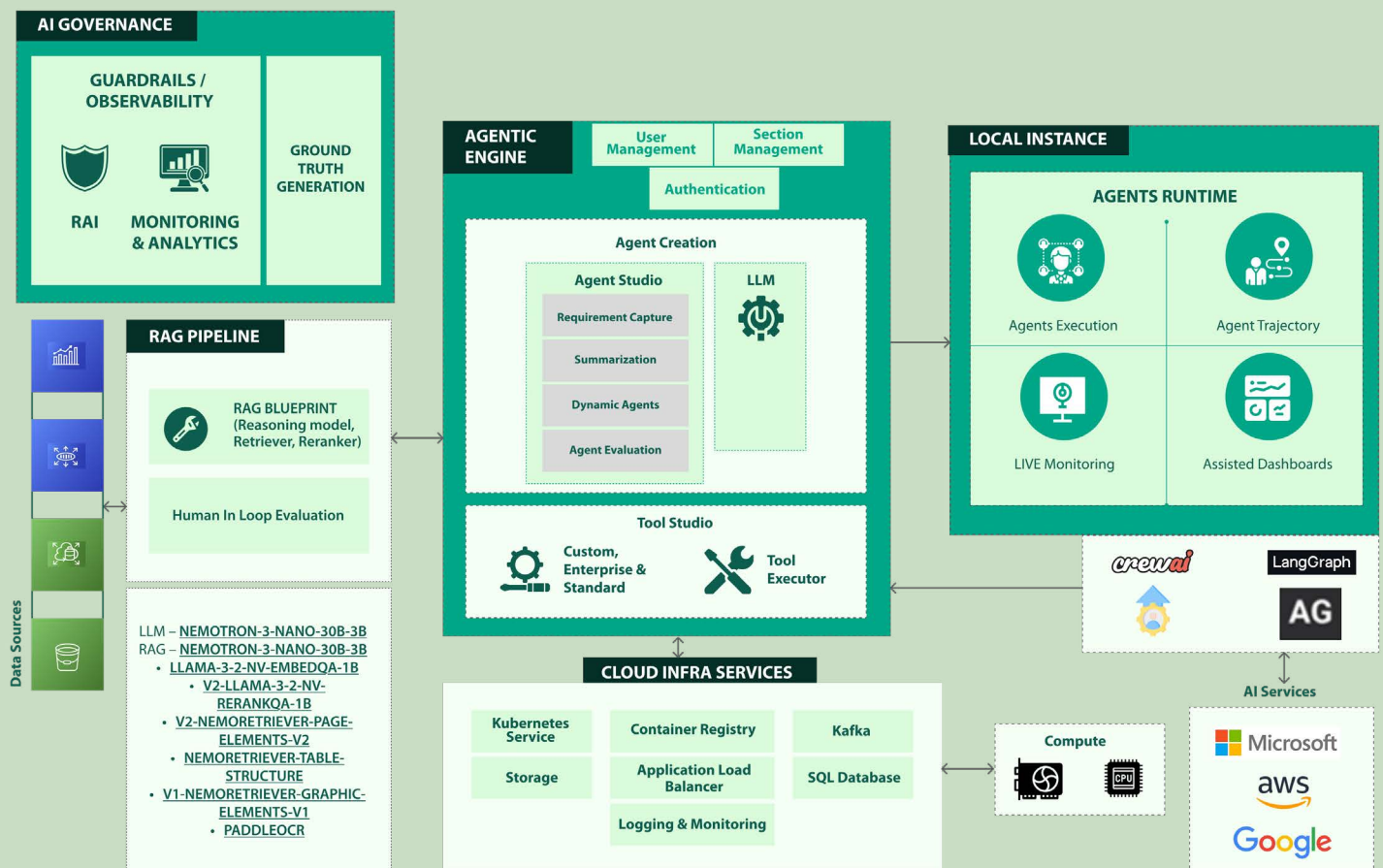


Figure 2: Infosys Agentic Foundry integration with NVIDIA RAG Enterprise

2.5. Oil and gas reports extraction summary

The oil and gas sector deals with highly technical, multi-format documents such as end-of-well reports, drilling summaries, and wireline logs, which combine text, plots, tables, schematics, and handwritten notes, making them difficult for machines to parse and time-consuming for experts to interpret. To address this, a solution was deployed using the NVIDIA RAG blueprint hosted on a local instance, with domain-specific documents ingested. Infosys Agentic Foundry orchestrates local NIM microservices—including embedding, reranking, and LLM tools—through an orchestrator agent that decomposes queries, retrieves and reranks relevant context, invokes LLM/VLM models, and performs verification and evaluation before delivering responses. Domain-specific parameters were tuned for context,

and all metrics and traces were captured within the Foundry for monitoring and optimization. The outcome was highly promising: context-rich answers for complex queries with an observed accuracy of 80–85%, an LLM score of 86.4 (indicating strong core answers, grounded evidence, and clear reasoning), and expert validation confirming contextual relevance and domain alignment.

Results show the presence of contextual information along with domain experience. Agents configured against LLM provide a quicker, efficient response for the queries executed. Figure 3 shows the response to the queries executed.

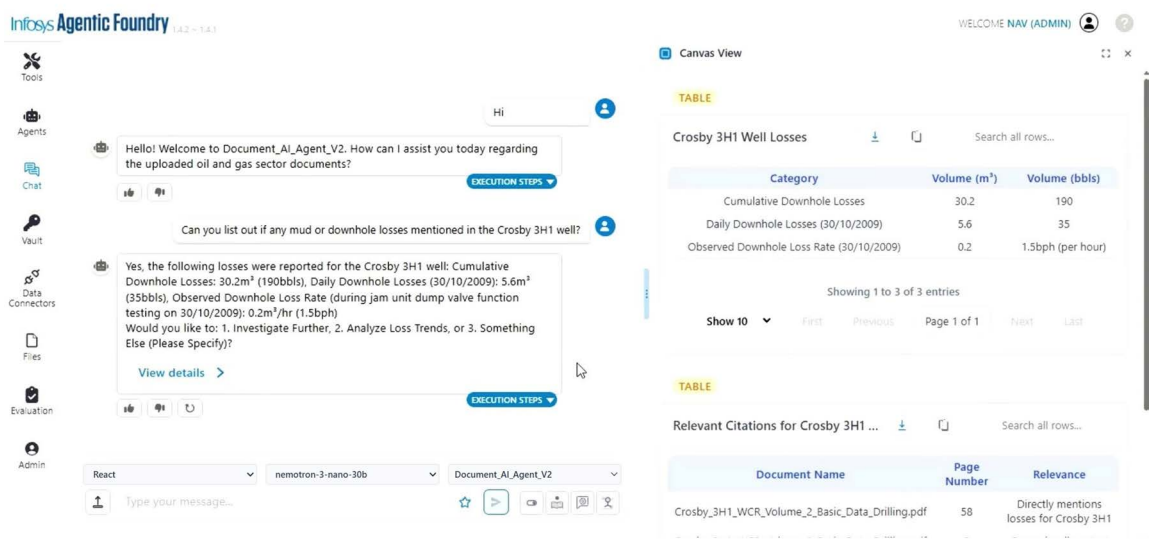


Figure 3 (a) Can you list out if any mud or downhole losses mentioned in the Crosby 3H1 well?

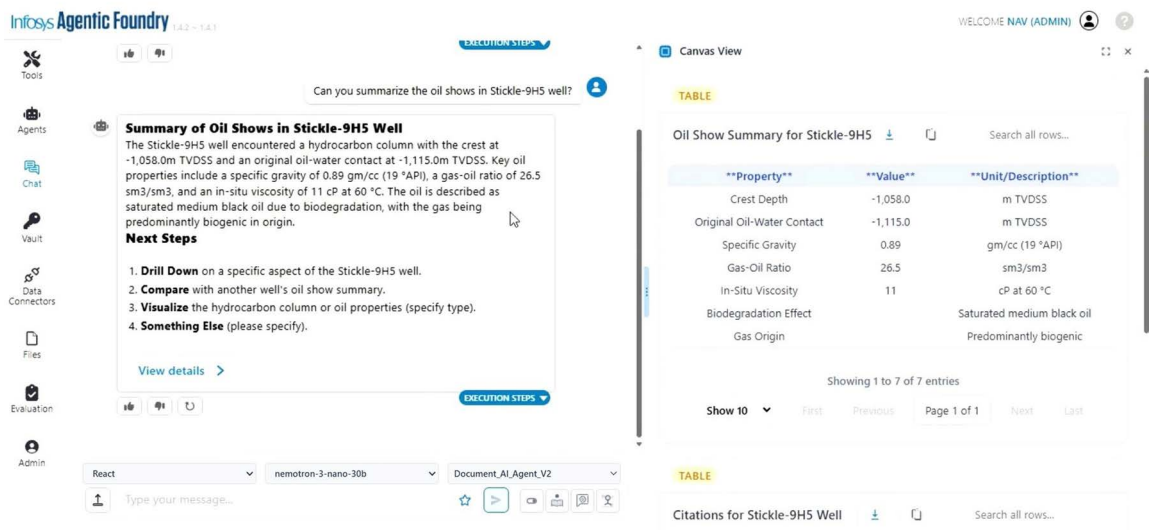


Figure 3 (b) Can you summarize the oil shows in Stickle-9H5 well?

The screenshot displays the Infosys Agentic Foundry interface. On the left, a sidebar contains icons for Tools, Agents, Chat, Vault, Data Connectors, Files, Evaluation, and Admin. The main chat area shows a conversation with an AI agent named 'Document_AI_Agent_V2'. The user asks, 'Can you provide the casing details for BruA-4 ST2? (Table)'. The agent responds with a table of casing details and lists three documents used for the response. The right pane, titled 'Canvas View', shows the generated table.

Response to User Query

Answer
Here is the BruA-4 ST2 casing details table based on the provided context:

Citations
Below are the citations supporting the answer, formatted for clarity:

- Document:** Black Chook-1_WCR.pdf
 - Page Number:** 11
 - Relevant Content Snippet:** Details on BruA-4 ST2 drilling and casing, including the decision to case and suspend the well.
- Document:** Black Chook-1_WCR.pdf
 - Page Number:** 11 (Continued from previous)
 - Relevant Content Snippet:** Confirmation of BruA-4 ST2's successful drilling in two runs and its casing.
- Document:** Black Chook-1_WCR.pdf
 - Page Number:** 10
 - Relevant Content Snippet:** Detailed well index sheet showing hole and casing specifics for BruA-4 and its sidetracks, including ST2.

Canvas View

TABLE

BruA-4 ST2 Casing Details

Hole Size (mm)	Interval mTVD/AHD (mMDRT)	Casing Size (mm)	Depth mTVD/AHD (mMDRT)	Casing Specifications	LOT/FIT
445	-373.7 to -1842.5 (396.0 to 2004.0)	340	-1838.4 (1996.0)	X56 LPWHH	LOT = 1.78 SG EMW

Showing 1 to 1 of 1 entries

Show 10 First Previous Page 1 of 1 Next Last

Figure 3 (c) Can you provide the casing details for BruA-4 ST2? (Table)

The screenshot displays the Infosys Agentic Foundry interface. On the left, a sidebar contains icons for Tools, Agents, Chat, Vault, Data Connectors, Files, Evaluation, and Admin. The main chat area shows a conversation with an AI agent named 'Document_AI_Agent_V2'. The user asks, 'Can you share the RT elevation of Ravensworth well?'. The agent responds with the RT elevation and lists three documents used for the response. The right pane, titled 'Canvas View', shows the generated table.

Response to User Query

Answer
The RT (Rotary Table) elevation of Ravensworth well is **22.25 m** above MSL (Mean Sea Level).
For further assistance or new queries, please respond with one of the following:

- Confirmation:** Is the provided RT elevation satisfactory?
- Further Assistance:** Specify additional information needed (e.g., TD depth, reservoir characteristics).
- New Query:** Enter your next question.

[View details](#)

Canvas View

TABLE

Relevant Citation for RT Elevation

Document Name	Page Number	Relevant Content Snippet
Ravensworth-SH3L1_L2 Final WCR_Rev_01.pdf	8	...Elevation RT-MSL: **22.25 m**... Well TD 3,710.0m MDRT 1,214.51m TVD...

Showing 1 to 1 of 1 entries

Show 10 First Previous Page 1 of 1 Next Last

Figure 3 (d): Can you share the RT elevation of Ravensworth well?

2.6. Metrics evaluation summary

The solution employs a comprehensive evaluation framework to ensure accuracy and reliability. Key metrics include the LLM Score (86.4), which measures core answer quality, grounding, and reasoning clarity, and Observed Accuracy, which reflects end-to-end correctness at ~80–85%. Additional measures such as SBERT Similarity and TF-IDF Cosine Similarity assess semantic alignment and contextual depth, while

ROUGE scores emphasize coverage, structure, and explainability. An agentic twist enhances robustness: evaluator agents use these scores as decision gates, automatically escalating low-confidence answers to HITL review or triggering adaptive re-planning—such as expanding retrieval scope, adjusting reranker thresholds, or invoking domain-tuned SLMs.

2.7. Production-deployable models used

The solution leverages NVIDIA's advanced AI stack to deliver high-performance document intelligence. At its core, the nemotron-3-nano model powers large language understanding. Specialized components like llama-3_2-nv-embedqa-1b-v2 and llama-3_2-nv-rerankqa-1b-v2 ensure accurate embeddings and

reranking for complex queries. Additionally, NeMo Retriever modules handle page elements, tables, and graphics, complemented by PaddleOCR for robust text extraction from images, making the stack ideal for multimodal document processing.



2.8. Infosys Agentic Foundry tools

1. PRESENTATION LAYER

**UI-based Interaction Hub**

Real-time chat, plan, tool verifiers, agent/tool creation, feedback management, evaluation, telemetry, etc.

**Export Agents**

Export single or many agents as independent solutions and to be used as API from external systems.

**Admin Dashboard**

Manages exporting agents, feedback cycles, model evaluations, and system-wide controls.

2. APPLICATION LAYER (BACKEND)

**API Gateway (FastAPI)**

Manages RESTful APIs, authentication, authorization, and session state.

**Orchestration Engine**

Coordinates all incoming requests and triggers the appropriate agent workflows.

3. CORE LOGIC LAYER (AGENTIC ENGINE)

**Agent Orchestrator**

Manages agent lifecycle, loads configurations, and connects context.

**ReAct Agent Engine**

Executes the single-agent Reason → Act → Observe loop for direct tasks.

**Multi-Agent Engine**

Runs the collaborative Planner → Executor → Critic workflow for complex tasks.

**Meta-Agent Engine**

Acts as a supervisor to delegate tasks to other agents and synthesize results.

**Tool Executor**

Dynamically loads, validates, and executes Python-based tools and functions.

**LLM Service Integration**

Abstracts connections to multiple LLMs like GPT-4, Gemini, and others.

4. DATA & PERSISTENCE LAYER

**PostgreSQL Database**

Central repository for agent/tool configs, user data, feedback, and metrics.

**File System Storage**

Stores user-uploaded assets like PDFs, documents, and SQLite databases.

**Agent Memory – Postgres and Graph**

Manages both short-term conversational context and long-term persistent knowledge.

5. OBSERVABILITY & EVALUATION

**OpenTelemetry**

Instruments the application to collect structured logs, metrics, and traces.

**Arize Phoenix**

Provides specialized observability for LLM interactions, tracing, and debugging.

**Grafana & Elasticsearch**

Visualizes operational data through real-time dashboards and log analytics.

**LLM-as-a-Judge and Ground Truth**

Automated engine for evaluating agent performance and response quality.

6. INFRASTRUCTURE & DEPLOYMENT

**Infrastructure & Deployment**

Deployable on GPU enabled Windows/ Linux, on-premise servers, or cloud platforms like Azure, AWS, GCP.

**Service Management**

Uses NSM (Windows) or systemd (Linux) to ensure services run continuously.

**Networking**

Configurable firewalls, ports, and proxy settings for secure enterprise deployment.

Figure 4: Infosys Agentic Foundry logical architecture

The Infosys Agentic Foundry is built on a robust and scalable technology stack designed for enterprise-grade AI solutions. Figure 4 provides its component view. It leverages Python as the core programming language, with LangChain and LangGraph enabling advanced agent orchestration and workflow management.

FastAPI provides high-performance microservices and API layers, while PostgreSQL ensures reliable and scalable data storage. The user experience is powered by React, delivering an intuitive and responsive interface for seamless interaction with AI-driven applications.

3. AI-powered document solution

3.1. Built for enterprise

The solution is designed with enterprise-grade capabilities to ensure security, scalability, and compliance. It offers flexible deployment options—on-premises, within a VPC, or in hybrid environments—while preserving data locality. Comprehensive observability and governance features provide full traces, guardrails, and audit trails across agents and tools, ensuring transparency and control. Its modular and extensible architecture allows organizations to swap models or tools, add specialist agents, and tailor prompts for domain-specific needs. Cost and performance are optimized through adaptive routing, using SLMs for routine tasks and LLMs for complex reasoning. Additionally, compliance is reinforced with HITL checkpoints, policy controls, and transparent evaluation logs, making the platform robust, secure, and enterprise-ready.

3.2. Ensures data security

NVIDIA AI Enterprise stack provides secured governance for model training and inference. NVIDIA [8.1]NIM comprises the best of both worlds, with the ease of use of a managed service, but the control, security, and flexibility of an open-source approach[9.1][10.1]. NVIDIA NIM for LLMs supports deployment on private cloud or on-prem infrastructure[11.1] with full control over data governance

3.3. Unlocking business value from legacy documents

Enterprises often hold terabytes of historical documents, including contracts, manuals, and compliance records, that remain largely untapped. By leveraging NVIDIA RAG

Blueprint and Infosys Agentic Foundry, organizations can build custom LLMs fine-tuned on proprietary data, unlocking valuable insights from legacy content. This transformation converts static archives into searchable, actionable intelligence, enabling better decision-making, improving productivity, and ensuring compliance across multiple business processes.

3.4. Lower TCO – NVIDIA infrastructure

NVIDIA's accelerated computing optimizes GPU utilization, reducing training and inference time, which significantly lowers infrastructure costs. By leveraging pre-trained models that are customized for different needs and fine-tuning capabilities through NVIDIA NeMo and Infosys accelerators, enterprises can minimize development effort and expenses.

3.5. Generalization across industries

The agentic AI pattern demonstrated in the oil and gas domain generalizes seamlessly across multiple industries. In financial services, it can process KYC/AML dossiers, risk memos, and regulatory filings; in manufacturing, it can interpret equipment manuals, quality logs, and sensor charts. Healthcare and life sciences benefit from their ability to handle clinical study reports and imaging notes while ensuring compliance, and the public sector and legal domains can leverage it for case bundles, legislative archives, and discovery processes. In communications, media, and technology (CMT), it enables streamlined content rights management, intelligent customer support through RAG-based retrieval, and summarization of network deployment plans. In the consumer, retail, and logistics (CRL) industry, it supports contract risk analysis, litigation



document summarization, and policy compliance validation. Further, the services, utilities, resources, and energy (SURE) space deals with large volumes of regulatory documents, audit reports, operational logs, and sustainability disclosures. Wherever data is multimodal, long-form, and compliance-sensitive, agentic orchestration consistently outperforms static pipelines, delivering context-rich, accurate, and auditable outcomes.

3.6. Extensibility

The solution enabled by NVIDIA accelerated computing can be extended with domain-specific SLMs being built to cater to the demands. To extend its capabilities, enterprises can also fine-tune models using NVIDIA NeMo and leverage the NeMo Agent Toolkit. Fine-tuning with NeMo allows domain-specific adaptation, improving accuracy for specialized terminology and complex document structures in industries like healthcare, legal, and utilities. The NeMo Agent Toolkit enables building intelligent, multi-step workflows that seamlessly integrate retrieval, reasoning, and generation. By adopting this approach, organizations benefit from higher precision in document understanding, reduced manual intervention, and faster compliance checks, while achieving lower total cost of ownership (TCO).

3.7. Implementation

To implement this solution, begin by standing up NVIDIA NIM microservices for embeddings, reranking, LLM inference, OCR, and structure extraction. Next, deploy the Infosys Agentic Foundry and select pre-built document AI templates for planner/executor, ingestion, retrieval, verification, evaluation, and HITL workflows. Connect your corpora with metadata policies and enable multi-collection search, then configure evaluation gates such as LLM Score, SBERT, ROUGE, and TF-IDF, along with HITL thresholds. Start with a pilot on representative documents and progressively specialize using SLM fine-tuning for deep domain tasks.

Conclusion

The AI-powered document solution enables automated ingestion, embedding, and verification, significantly reducing human effort, streamlining the process, and improving accuracy. It supports integration with external sources to enrich knowledge bases and ensures continuous synchronization of standard reports. The system also facilitates repository visibility through standardized UI components and generates regular email reports on ingestion metrics and performance. With proven success in the oil and gas domain, the solution demonstrates its scalability and adaptability across industries where document complexity and data diversity are major bottlenecks. This solution not only enhances operational efficiency but also empowers domain experts to focus on higher-value tasks.

Authors



Keerthi Prasad is a Senior Technology Architect at Infosys. He is part of Topaz CoE and works with Hyperscalers where he works closely on customers enterprise requirements and building solutions on these platforms. He helps with the design, development and deployment of the solutions and also supports customers in Generative AI journeys.



Prakruthi B Gowda is a Deep Learning Solution Architect at NVIDIA involved in building Generative AI and conversational AI systems across industries. She specializes in LLM-driven, agentic AI architectures, speech AI, and GPU-accelerated solutions designed for scalable, production-grade deployment. Her work focuses on delivering measurable ROI by translating latest AI research breakthroughs into real-world enterprise impact.



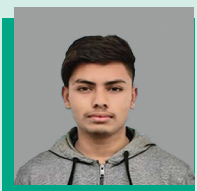
Navneet Sharma is a Digital Specialist Engineer at Infosys. He is part of the Infosys Topaz CoE and works on building generative AI solutions across new offerings.



Kiran Kori is a Lead Consultant at Infosys. He is part of the Infosys Topaz NVIDIA CoE. He works with industry leading chipmakers and OEMs to design and deliver AI and GenAI solutions that create measurable business impact. Partnering with solution architects and developers from strategy through execution, helps accelerate early technology adoption and translates innovation into tangible results.



Paulo M Prasad is a Principal Consultant at Infosys. He is part of the Infosys Topaz CoE. He focuses on driving end to end implementation of a Agentic AI powered use cases across multiple business functions. His work involves identifying opportunities for automation and intelligence, designing scalable solutions, and collaborating with cross functional teams to deliver impactful outcomes.



Amitansh Johri is a Digital Specialist Engineer at Infosys. He is part of the Infosys Topaz CoE and works on building custom agentic AI solutions

References

Infosys Agentic Foundry (Open Source): GitHub – [Infosys/Infosys-Agentic-Foundry](https://github.com/Infosys/Infosys-Agentic-Foundry)

NVIDIA Enterprise RAG Blueprint: [Build an Enterprise RAG Pipeline Blueprint by NVIDIA | NVIDIA NIM](#)

Infosys Topaz is an AI-first set of services, solutions and platforms using generative AI technologies. It amplifies the potential of humans, enterprises, and communities to create value. With 12,000+ AI assets, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a responsible by design approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale and build connected ecosystems. Connect with us at infosystopaz@infosys.com

For more information, contact askus@infosys.com



© 2026 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names, and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording, or otherwise, without the prior permission of Infosys Limited and/or any named intellectual property rights holders under this document.

[Infosys.com](https://www.infosys.com) | NYSE: INFY

Stay Connected

