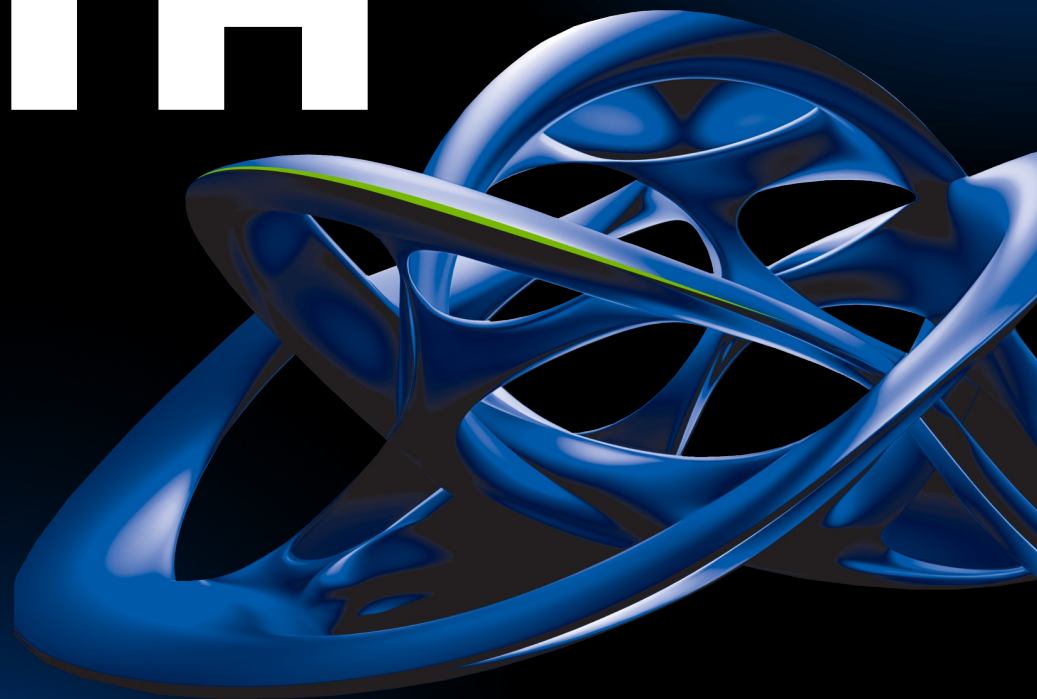


MAKING LLMs & RAG WORK WITH EASE

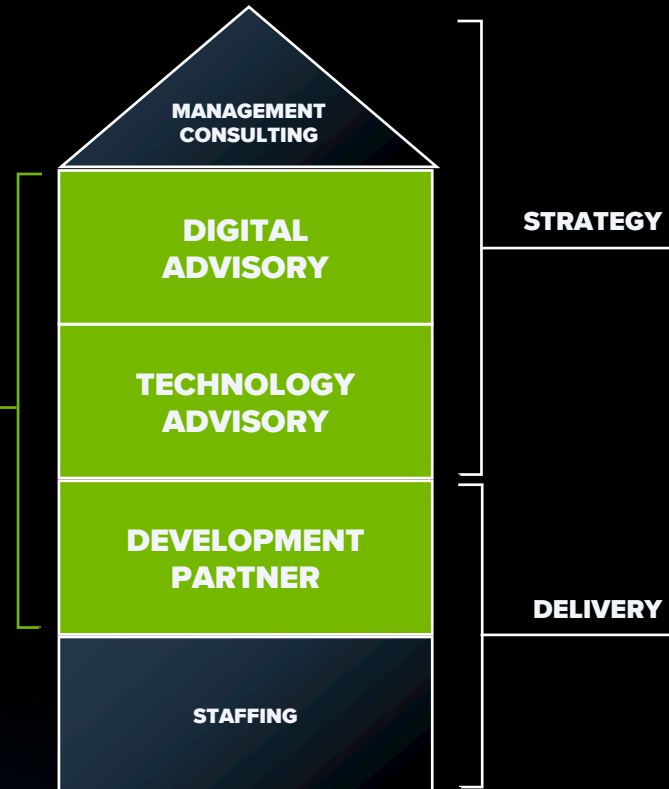


SOFTSERVE AT A **GLANCE**

**WE ARE ADVISORS
AND PROVIDERS WHO
OPERATE AT THE
CUTTING EDGE
OF TECHNOLOGY**

SOFTSERVE

We are also a **lean advisory** with iterative practical results rooted in **executable excellence**.



SPEAKER

IURII MILOVANOV

AVP, AI & Data Science
SoftServe Inc.



11,000+

Associates
worldwide

30 YEARS

Across multiple
industries

GLOBAL

61 offices,
16 countries

20,000+

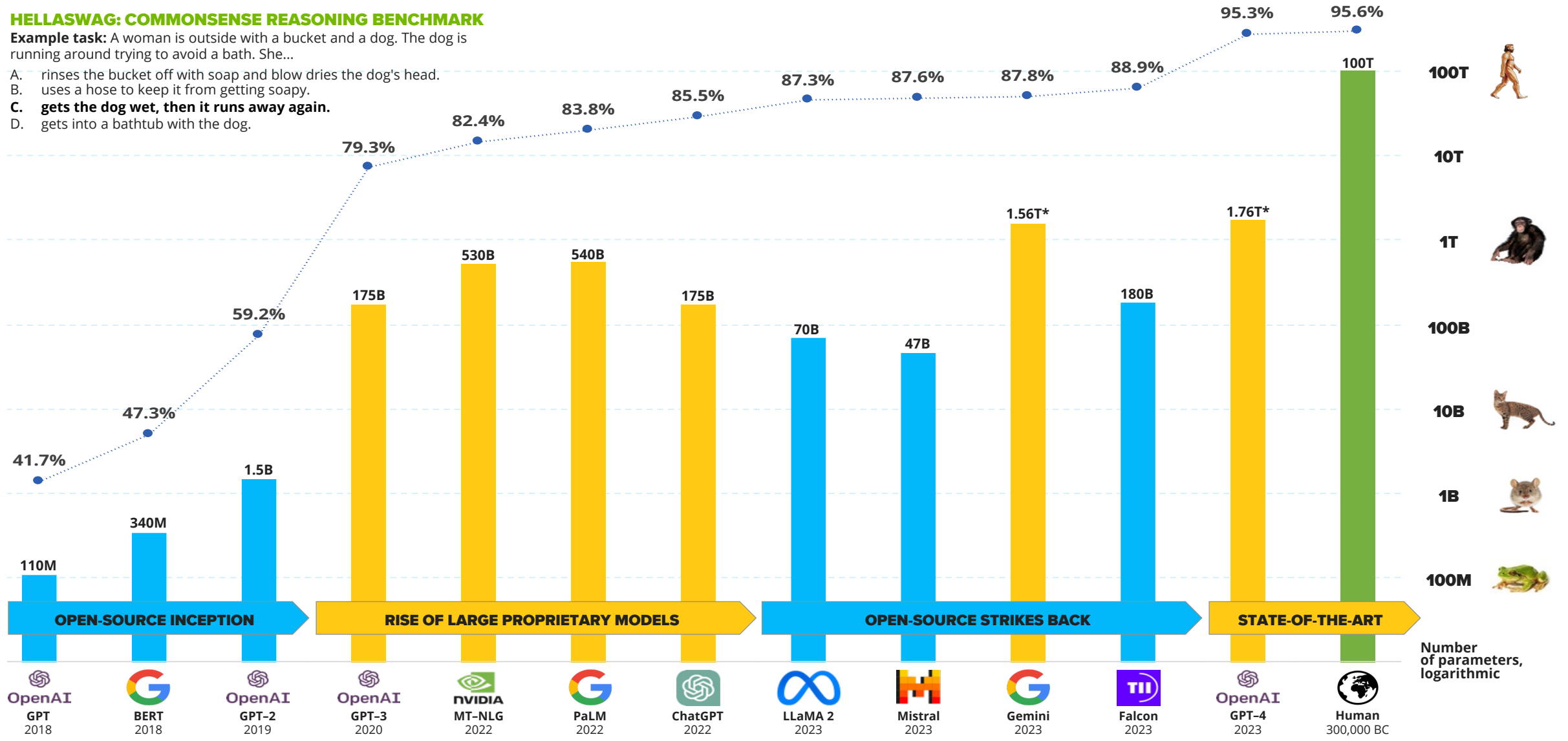
Complex projects
delivered

THE MARKET MOMENTUM WITH GENERATIVE AI

HELLASWAG: COMMONSENSE REASONING BENCHMARK

Example task: A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

- A. rinses the bucket off with soap and blow dries the dog's head.
- B. uses a hose to keep it from getting soapy.
- C. gets the dog wet, then it runs away again.
- D. gets into a bathtub with the dog.



* Model size is speculative and based on unofficial sources.

GENERATIVE AI LANDSCAPE

IMAGE

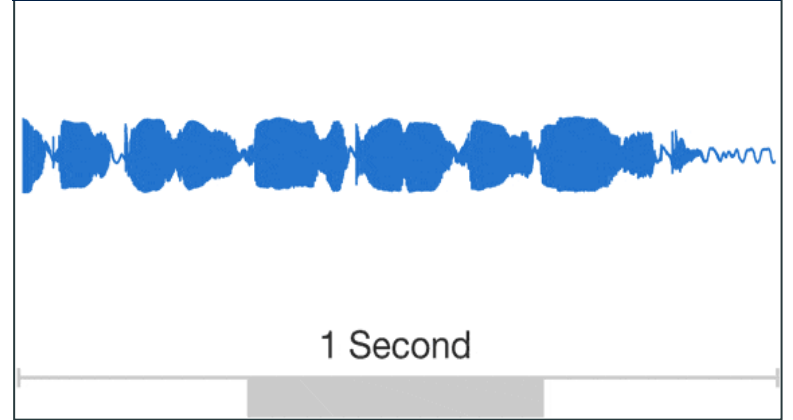


"A sci-fi panda in a yellow and blue costume eats softserve"

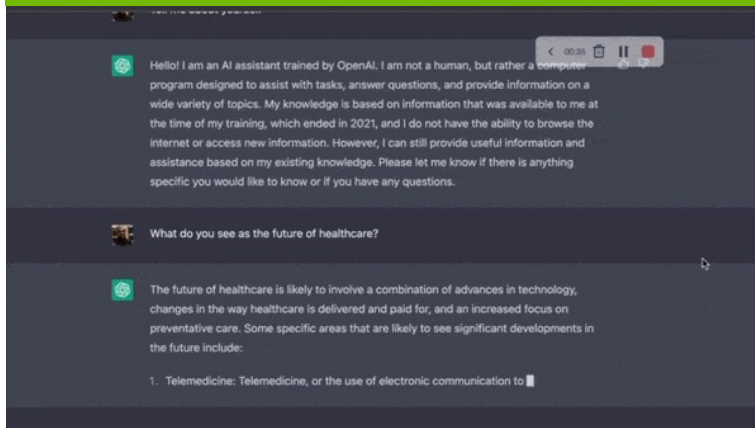
VIDEO



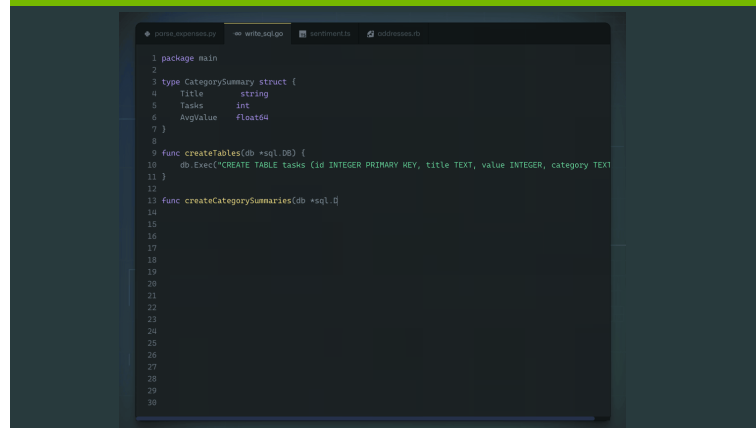
AUDIO



NATURAL LANGUAGE




SOURCE CODE





FOCUS AREA - LARGE LANGUAGE MODELS (LLMs)

UNLOCKING BUSINESS POTENTIAL OPPORTUNITIES AND CROSS-INDUSTRY GENERATIVE AI USE CASES

ASK QUESTIONS AGAINST KNOWLEDGE


 **QUESTION ANSWERING**
Enterprise search, regulatory compliance, medical discovery, troubleshooting, FAQs


 **SUMMARIZATION**
Market research, financial and legal analysis, patient history, incident reporting


 **KNOWLEDGE GRAPHS**
Inventory management, regulatory compliance, medical coding, operational excellence


 **SIMILARITY SEARCH**
Product recommendations, patient matching, investment opportunity discovery, competitor analysis

DERIVE INSIGHTS FROM KNOWLEDGE


 **REASONING**
Churn prediction, fraud detection, diagnosis assistance, root cause analysis


 **CLASSIFICATION**
Customer segmentation, transaction categorization, patient triage, defect detection


 **TOPIC RECOGNITION**
Market trends, customer sentiment, public health, emerging technologies

 **KEY-VALUE EXTRACTION**
Claims processing, KYC data collection, EHR management, order processing

GENERATE NEW DATA BASED ON KNOWLEDGE

 **CONVERSATION**
Customer support, financial advisor, telemedicine, operations assistant

 **TEXT GENERATION**
Personalized marketing, patient education, financial reports, technical documentation

 **CODE GENERATION**
Coding assistance, language conversion, API integration, test case generation

 **LANGUAGE TRANSLATION**
Multilingual support, medical research translation, global compliance

UNLOCKING BUSINESS POTENTIAL OPPORTUNITIES AND CROSS-INDUSTRY GENERATIVE AI USE CASES

BUSINESS FUNCTIONS



SEARCH



KNOWLEDGE DISCOVERY (RESEARCH)



ANALYTICS



DECISION-MAKING



DOCUMENT PROCESSING



COMMUNICATION

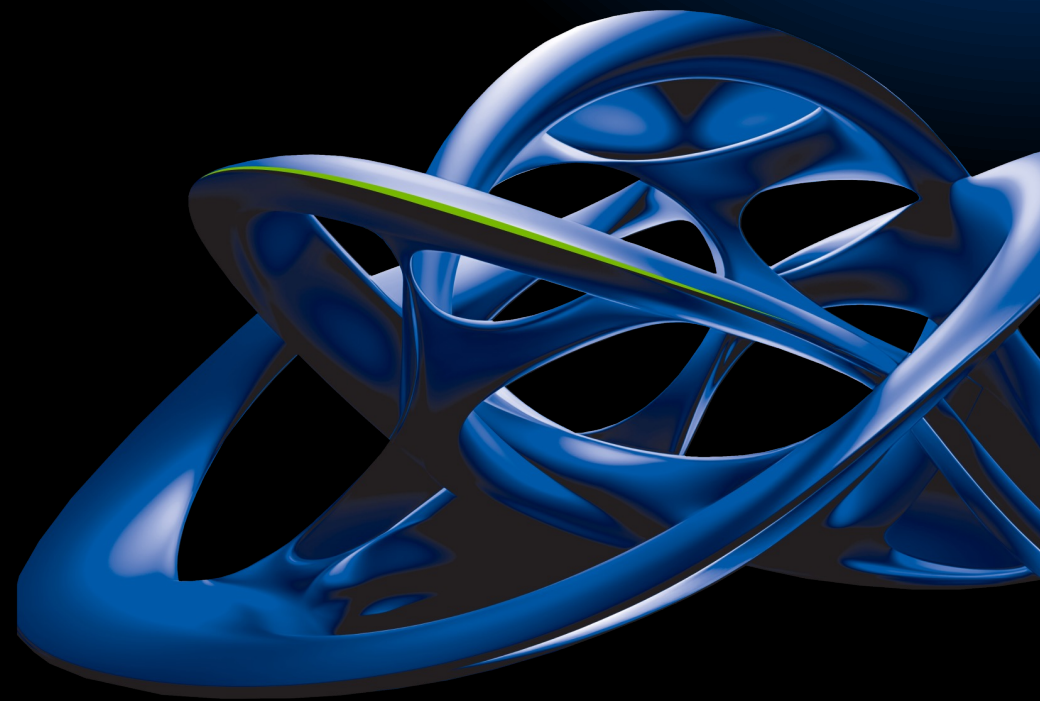


CONTENT WRITING



PROGRAMMING

WHY RAG?



GENERATIVE AI USE CASE ARCHETYPES

These archetypes simplify the understanding of Gen AI's functional outcomes and guide swift alignment with solution strategies.



DATA INSIGHT

Use cases following the **Data Insight** archetype help users quickly find and surface relevant information from large, complex, and diverse data repositories. This accelerates time-to-insight, allowing individuals and organizations to make more informed decisions faster than ever before.

Common challenges:

High data volumes, data source integration, data quality, hallucinations



VIRTUAL AGENT

Use cases adopting the **Virtual Agent** archetype enhance customer and employee experiences by integrating intelligent virtual assistants into interactions, either directly or via human augmentation. This facilitates a shift from traditional intent-based chatbots to AI agents capable of comprehending the conversation's context and delivering relevant data-driven responses.

Common challenges:

Low-latency, UX design, guardrails, privacy and security



CONTENT CREATION

Use cases following the **Content Creation** archetype automate and accelerate all forms of content creation, including code, text, images, audio, video, presentations, and documents. This improves content creation speed and quality, reducing the time and effort required for developers and creators to produce and customize content for specific audiences or tasks.

Common challenges:

Multimodality, training data, compute infrastructure, quality assurance

GENERATIVE AI USE CASE ARCHETYPES

DATA INSIGHT

Streamlining how employees access and utilize corporate knowledge.

Enabling customers to quickly find the information they need to make a purchase decision.

Relieving analysts from routine tasks and providing managers with instant data-driven responses, surpassing traditional corporate BI tools.

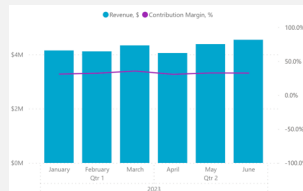
Example:

What are the P&L dynamics of products X and Y combined in APAC over the last 6 months?

Certainly, over the last six months, the combined P&L dynamics of products X and Y in the APAC region have shown some interesting trends.

Overall, both products have demonstrated a positive growth trend in terms of profitability. Product X has shown a steady increase in profits month over month, contributing significantly to the overall P&L.

Meanwhile, product Y, although starting from a lower base, has shown a steeper growth curve, suggesting a growing market acceptance.



Please feel free to ask if you have any specific questions or if there are other metrics you'd like to explore.

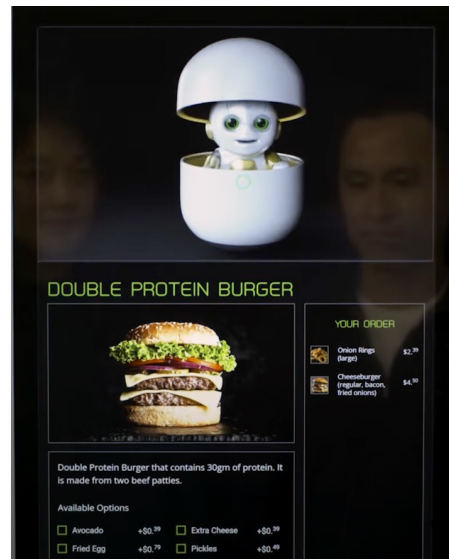
VIRTUAL AGENT

Enhancing online customer support with more intuitive virtual assistants.

Improving employee productivity with AI-powered virtual assistants.

Innovating customer service with interactive digital avatar kiosks.

Example:



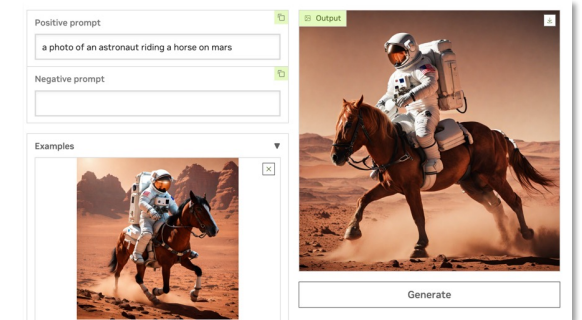
CONTENT CREATION

Boosting software developers' efficiency by auto-generating code from natural language prompts.

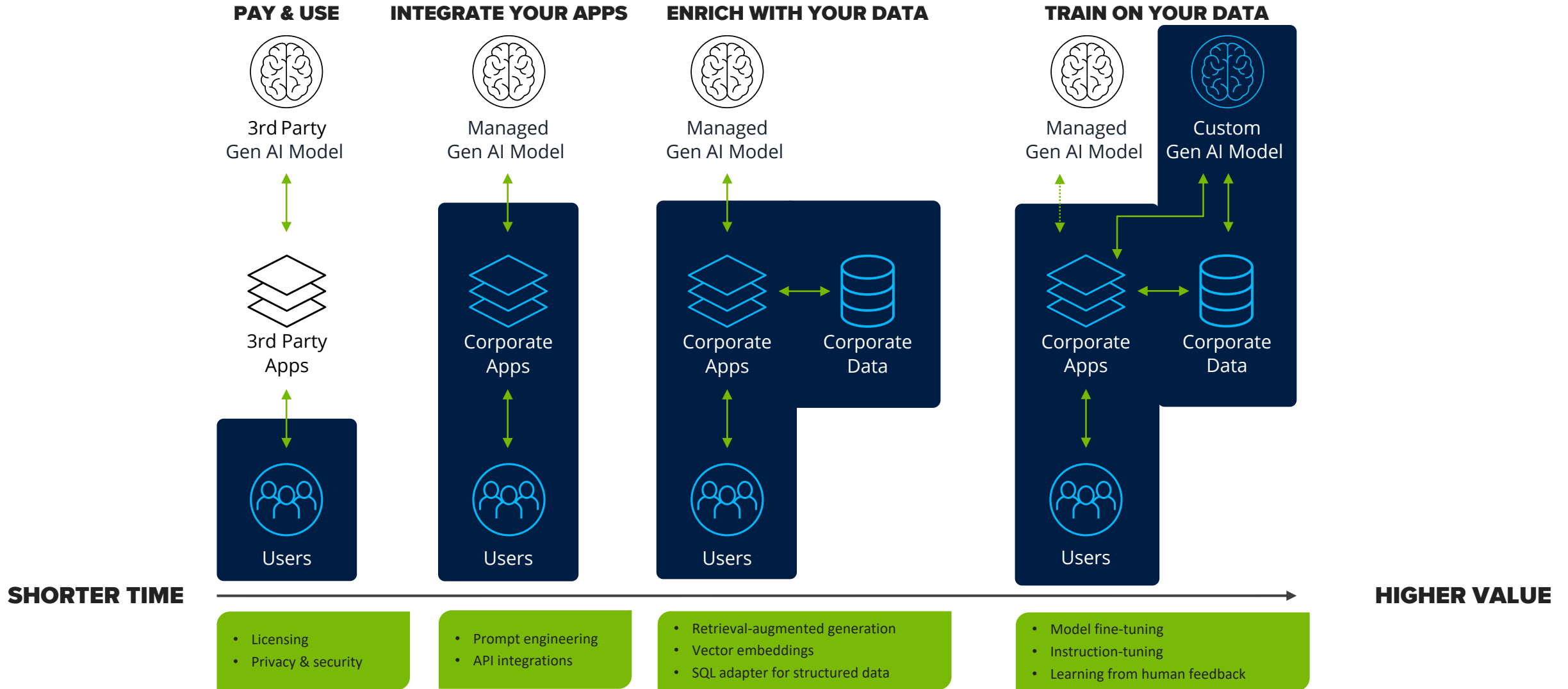
Accelerating content creation for marketing and advertising campaigns.

Creating personalized product look and description through a deep understanding of customer preferences.

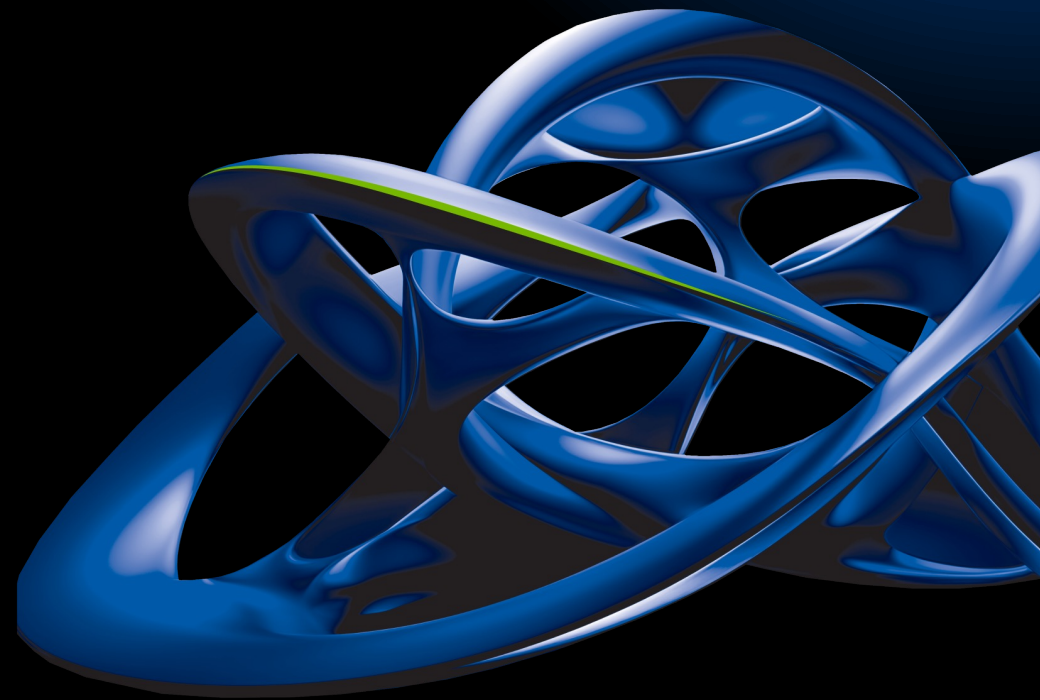
Example:



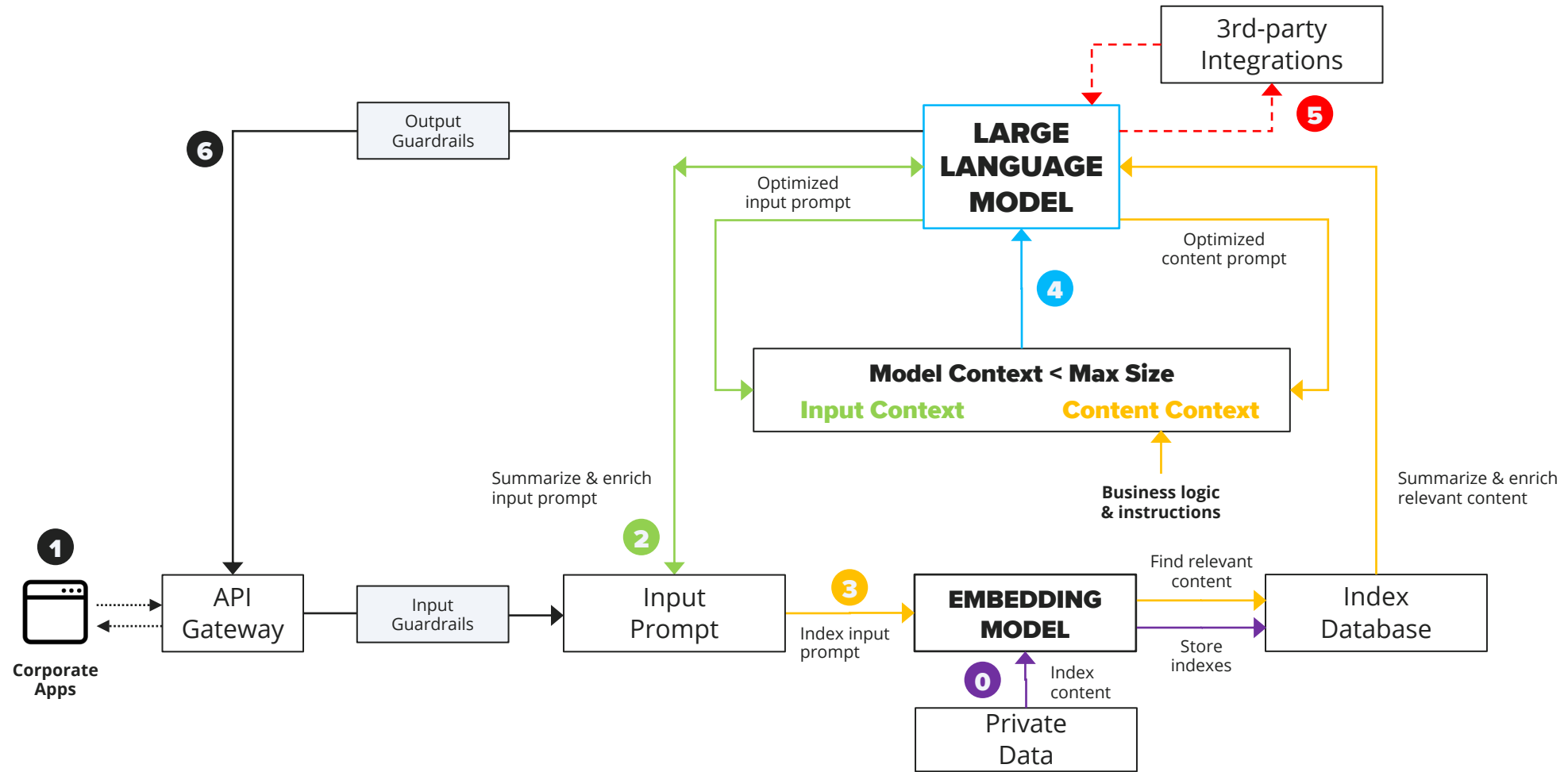
GENERATIVE AI ADOPTION PATTERNS



HOW TO RAG WITH **NVIDIA**?

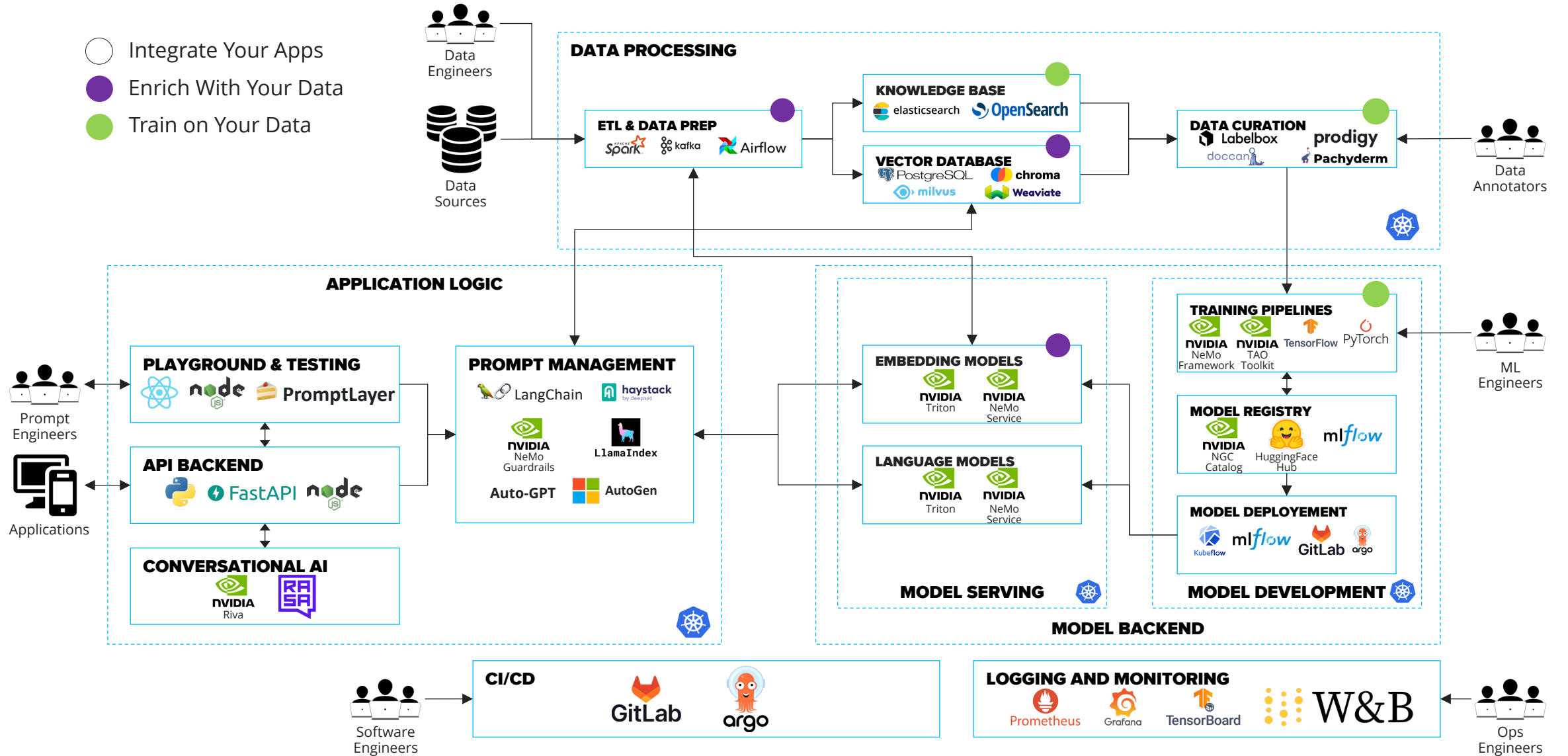


DESIGN PATTERN: RETRIEVAL AUGMENTED GENERATION (RAG)



GENERATIVE AI REFERENCE ARCHITECTURE

- Integrate Your Apps
- Enrich With Your Data
- Train on Your Data



INTEGRATING GEN AI WITH JIRA AND CONFLUENCE FOR ENHANCED KNOWLEDGE MANAGEMENT



BUSINESS CHALLENGE

The fragmentation of knowledge across Jira and Confluence platforms presented a significant challenge in accessing and utilizing information efficiently. This separation hindered the ability to perform cross-platform searches and complicated the process of finding relevant data, affecting productivity and decision-making processes within organizations.



SOLUTION

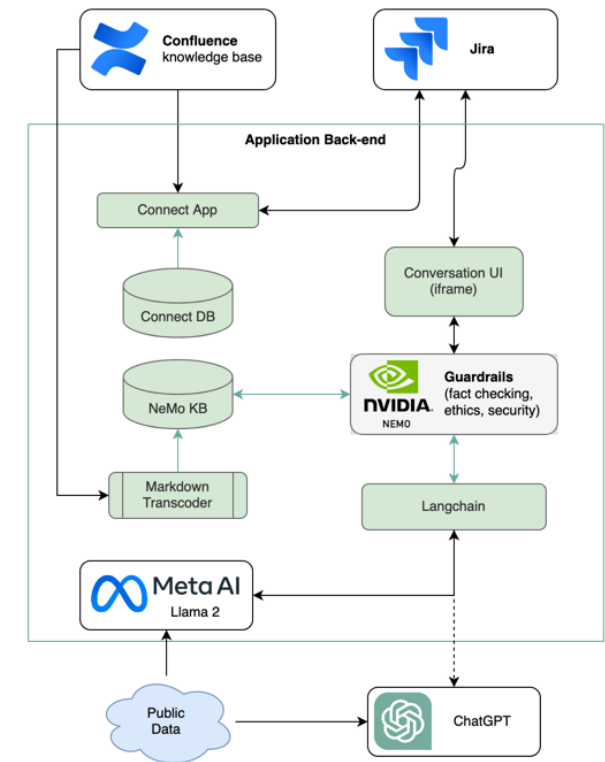
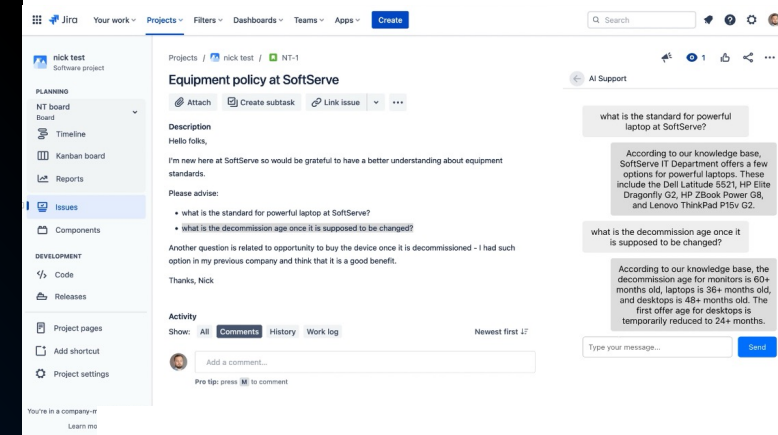
SoftServe, leveraging its status as an Atlassian Silver Solution partner, developed a proof of concept (PoC) integrating Generative AI (Gen AI) with Jira and Confluence knowledge bases. This innovative solution aimed to streamline the search process across these platforms, enhancing user experience and operational efficiency. Key features of the solution included:

- Custom integration of Gen AI services with Confluence and Jira Cloud, utilizing large language models (LLMs) for improved security and data privacy.
- A conversational UI embedded into the Jira interface, facilitating efficient knowledge base utilization and ticket processing.
- Implementation of data privacy measures, hallucination detection, and internal knowledge base synchronization to ensure accuracy and security.

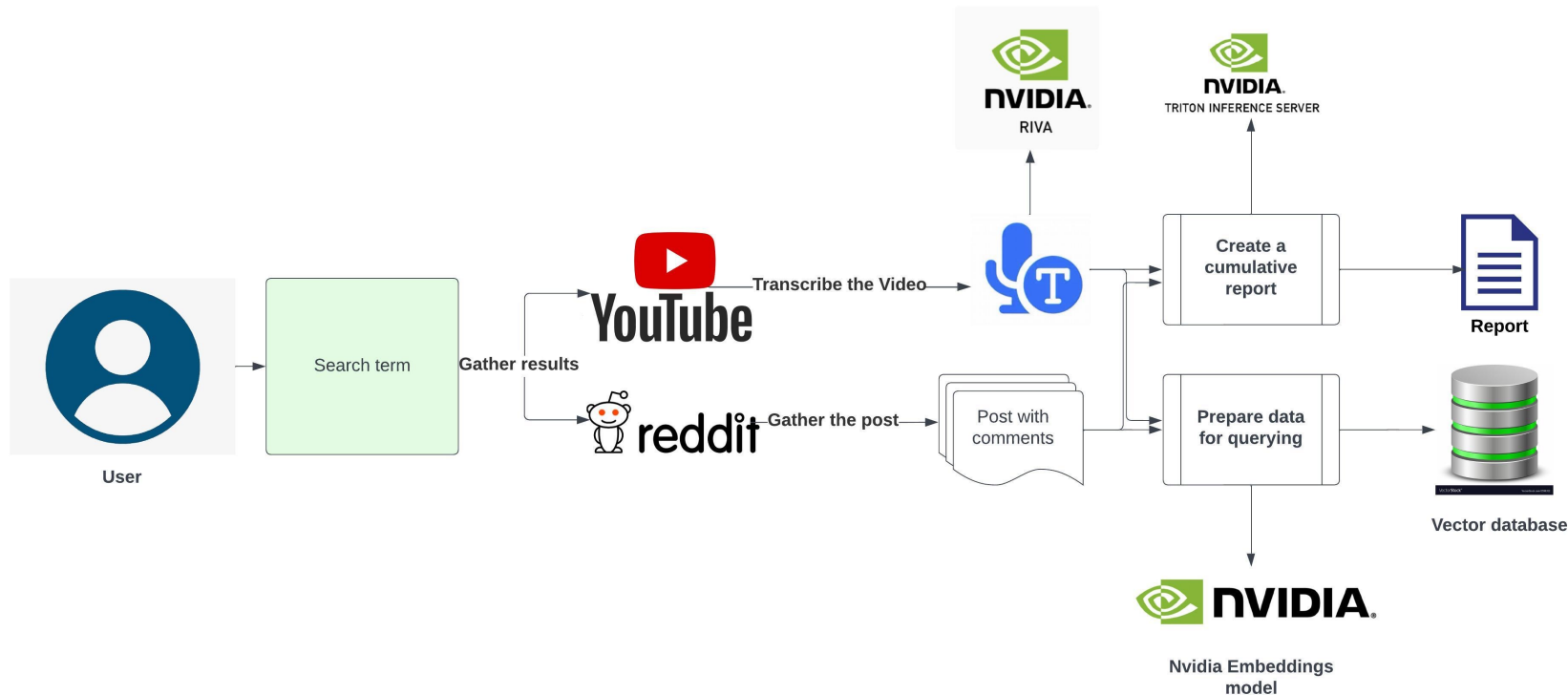


IMPACT

- **Enhanced Productivity:** The solution significantly improved the productivity of IT support teams by optimizing search processes and providing quick hints for ticket resolution, leading to faster customer service.
- **Improved Data Security and Privacy:** By employing encryption, user authentication, and hallucination detection, the solution ensured the confidentiality and integrity of data across platforms.
- **Streamlined Knowledge Management:** Automated synchronization and integration with Gen AI models facilitated seamless access to up-to-date information, overcoming the challenge of disparate knowledge repositories.



USER FEEDBACK ANALYZER USING NEMO RAG



- Solution is suitable for analyzing user feedback regarding a service or a product
- Collects data from publicly available sources such as YouTube or Reddit
- Utilizes RIVA for video transcriptions, NVIDIA Retrieval QA Embedding for text embeddings and Llama-2 as LLM
- Collected data is processed by a LLM to generate the desired reports



FOR softserve
THE
FUTURE