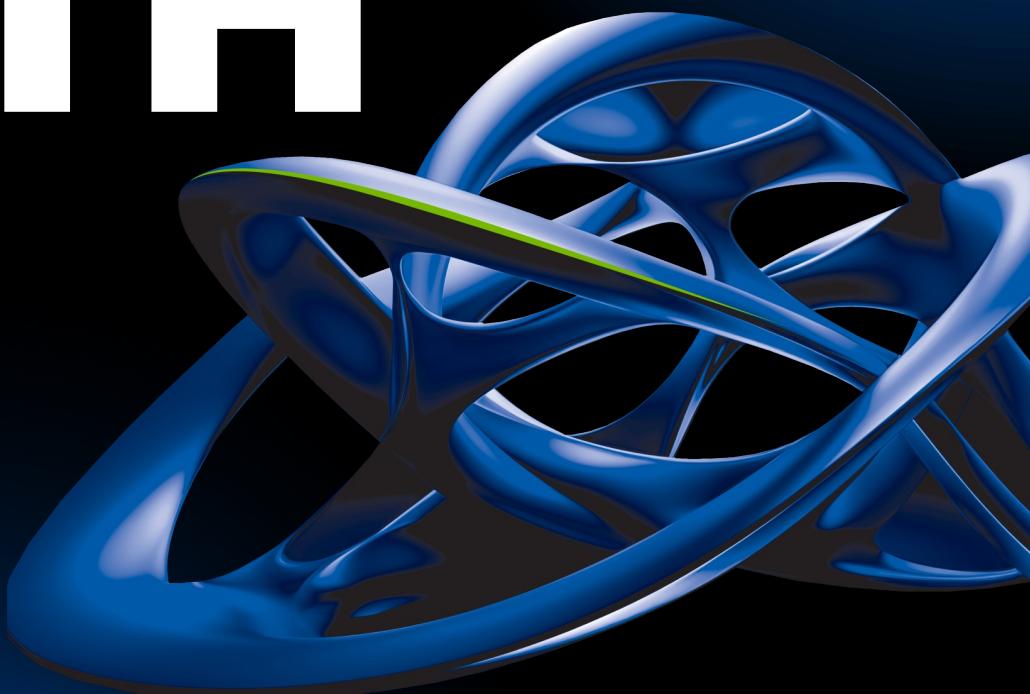


# MAKING LLMs & RAG WORK WITH EASE



March, 2024

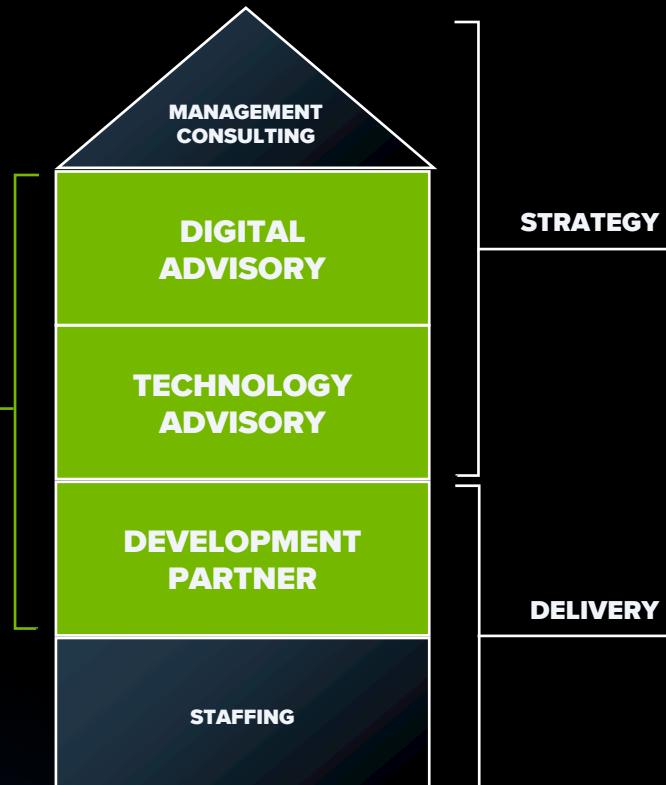
 NVIDIA. softserve

# 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**.



**11,000+**

**30 YEARS**

**GLOBAL**

**20,000+**

Associates worldwide

Across multiple industries

61 offices,  
16 countries

Complex projects delivered

**SPEAKER**

# **IURII MILOVANOV**

AVP, AI & Data Science  
SoftServe Inc.

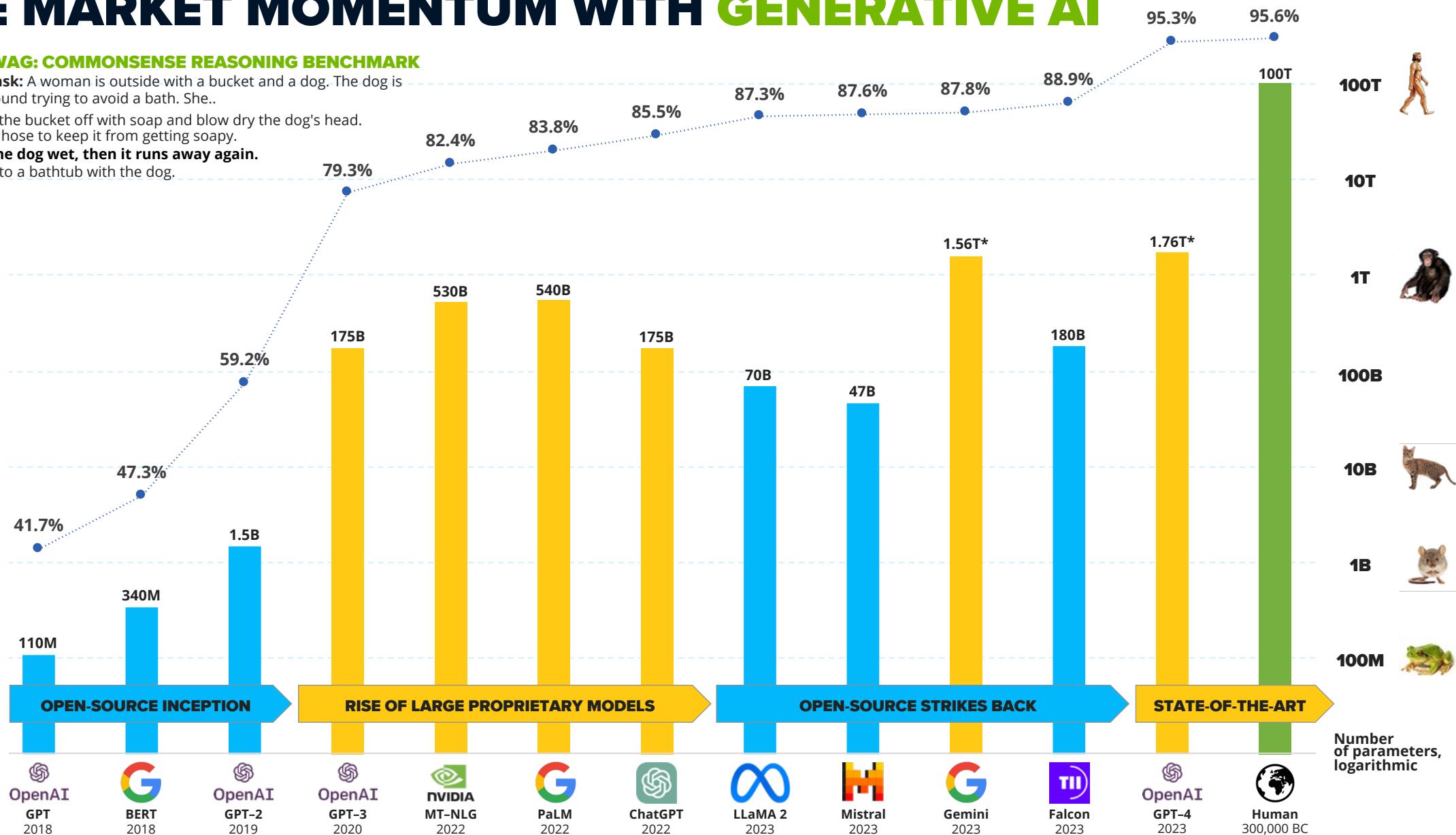


# THE MARKET MOMENTUM WITH GENERATIVE AI

## 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 dry 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.

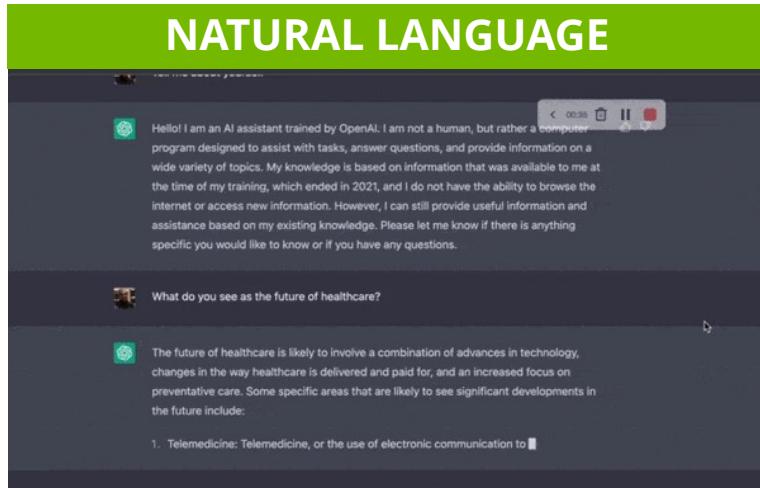
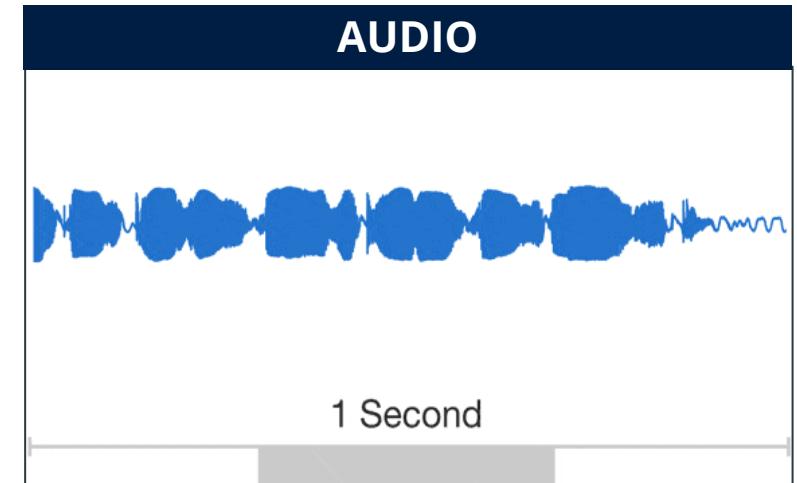
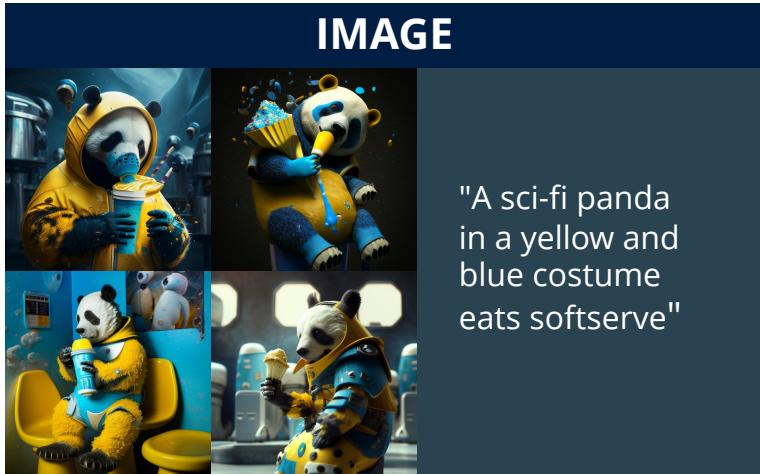


\* Model size is speculative and based on unofficial sources.



softserve

# GENERATIVE AI LANDSCAPE



### SOURCE CODE

```
1 package main
2
3 type CategorySummary struct {
4     Title      string
5     Tasks      int
6     AvgValue   float64
7 }
8
9 func createTables(db *sql.DB) {
10    db.Exec(`CREATE TABLE tasks (id INTEGER PRIMARY KEY, title TEXT, value INTEGER, category TEXT)
11    `)
12 }
13 func createCategorySummaries(db *sql.DB) {
14 }
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30 }
```

## 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

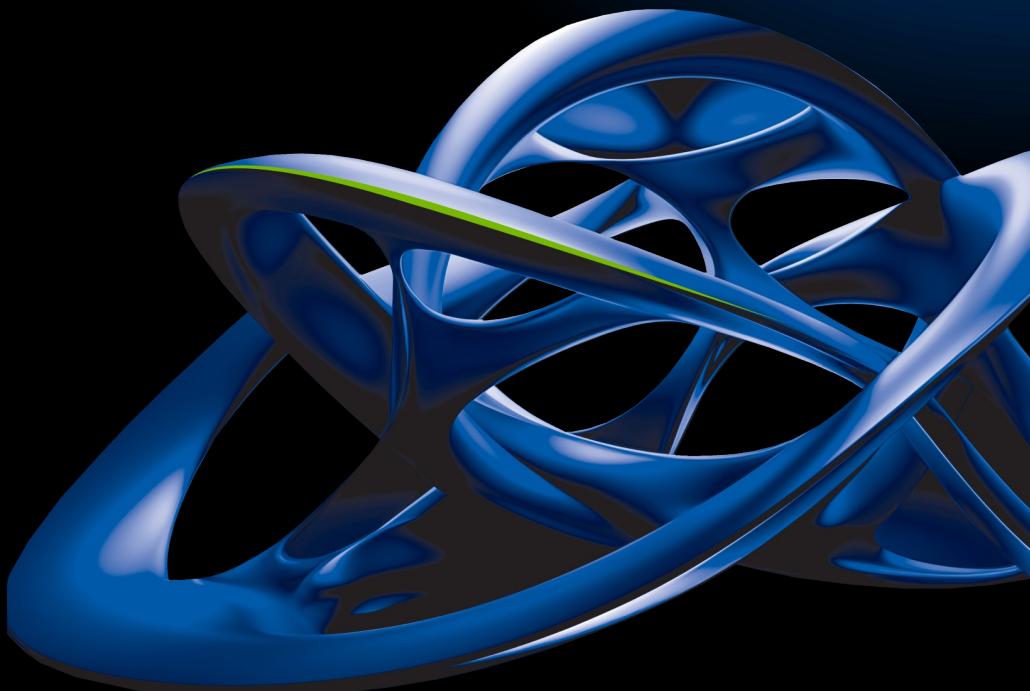


CONTENT WRITING



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

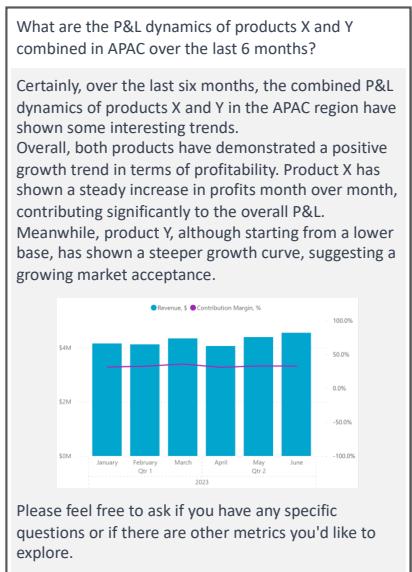
## DATA INSIGHT

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:



## 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:



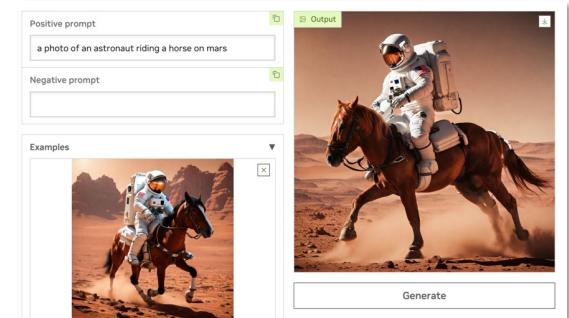
## CONTENT CREATION

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

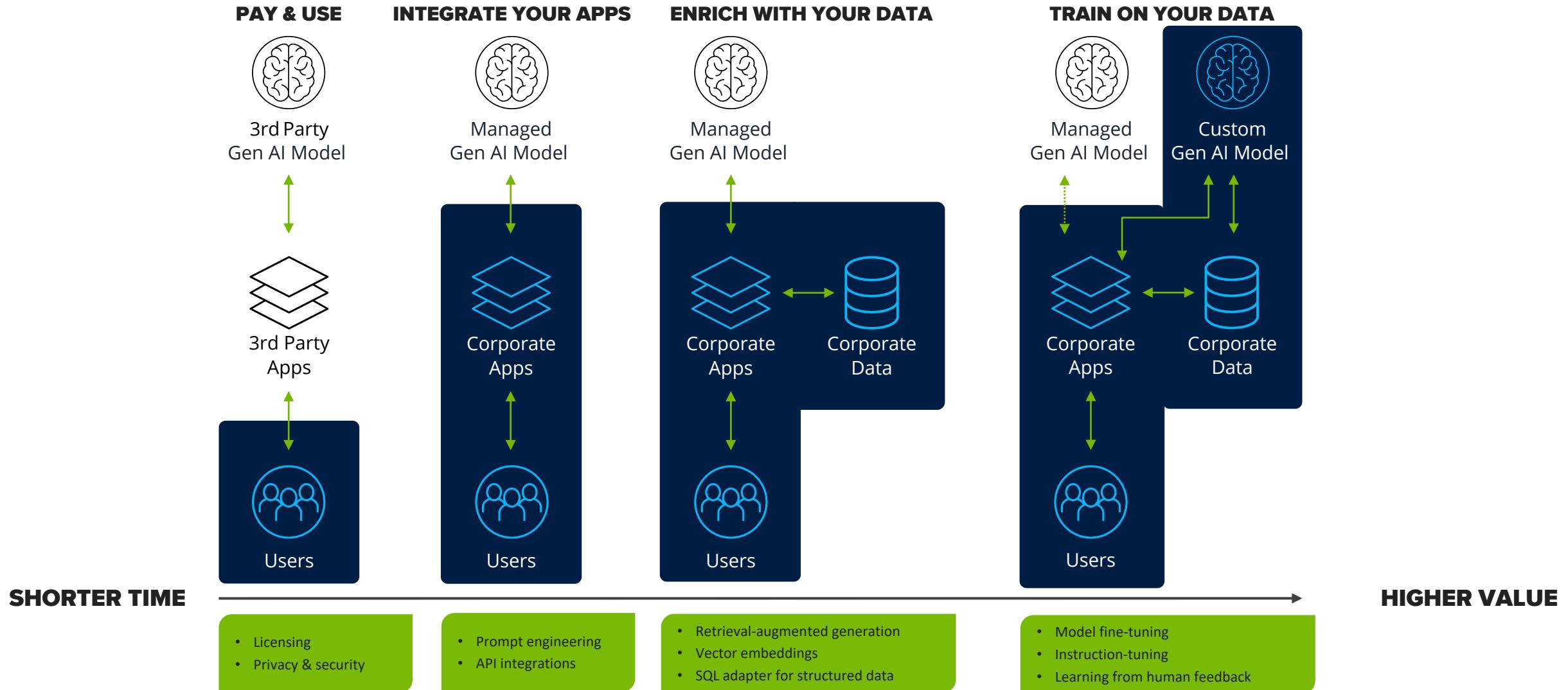
Accelerating content creation for marketing and advertising campaigns.

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

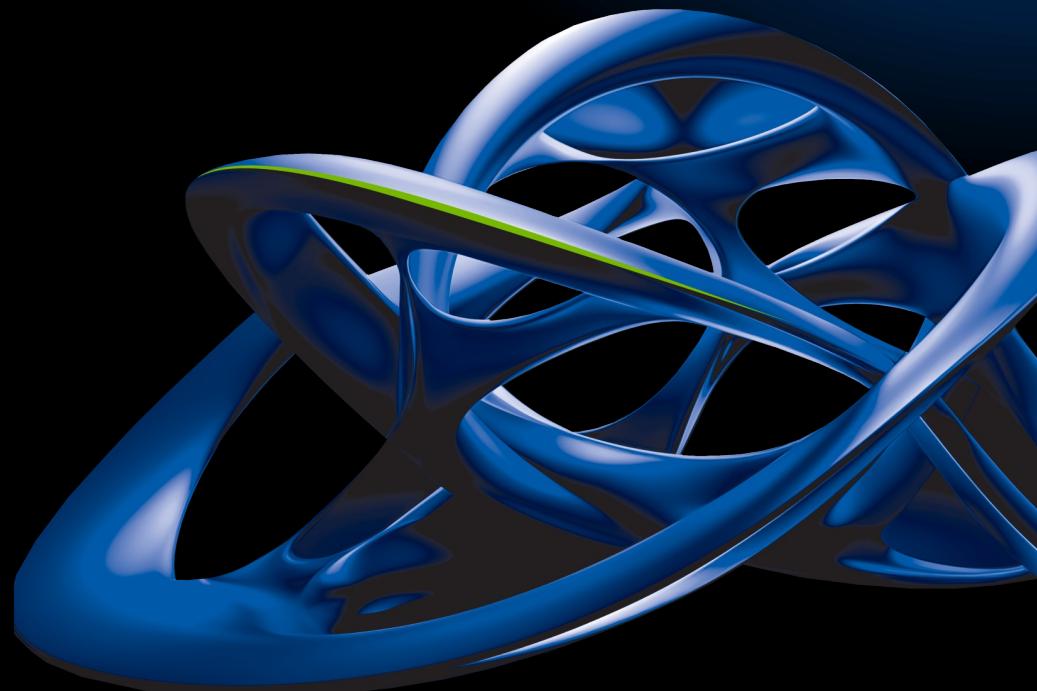
Example:



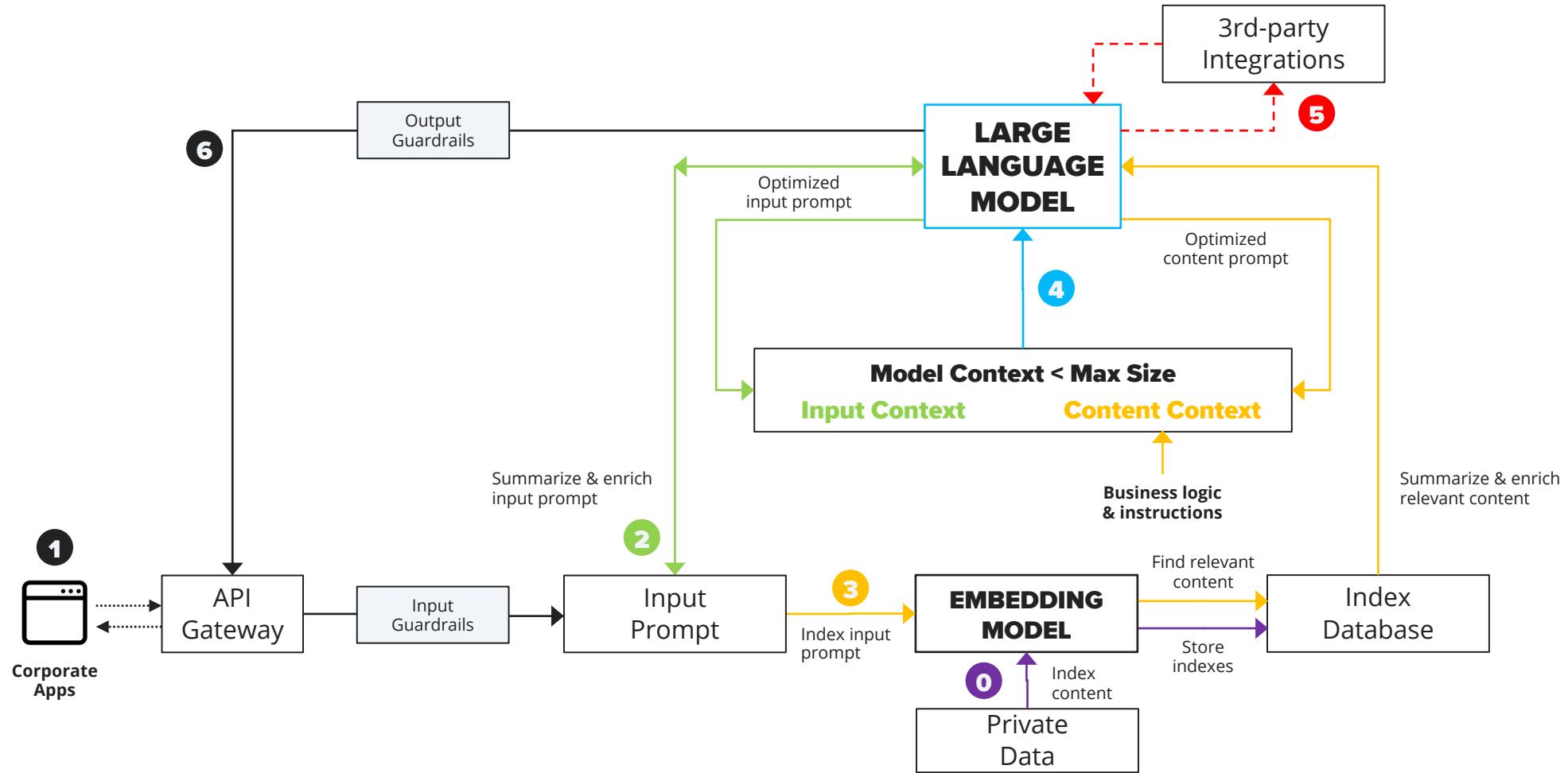
# GENERATIVE AI ADOPTION PATTERNS



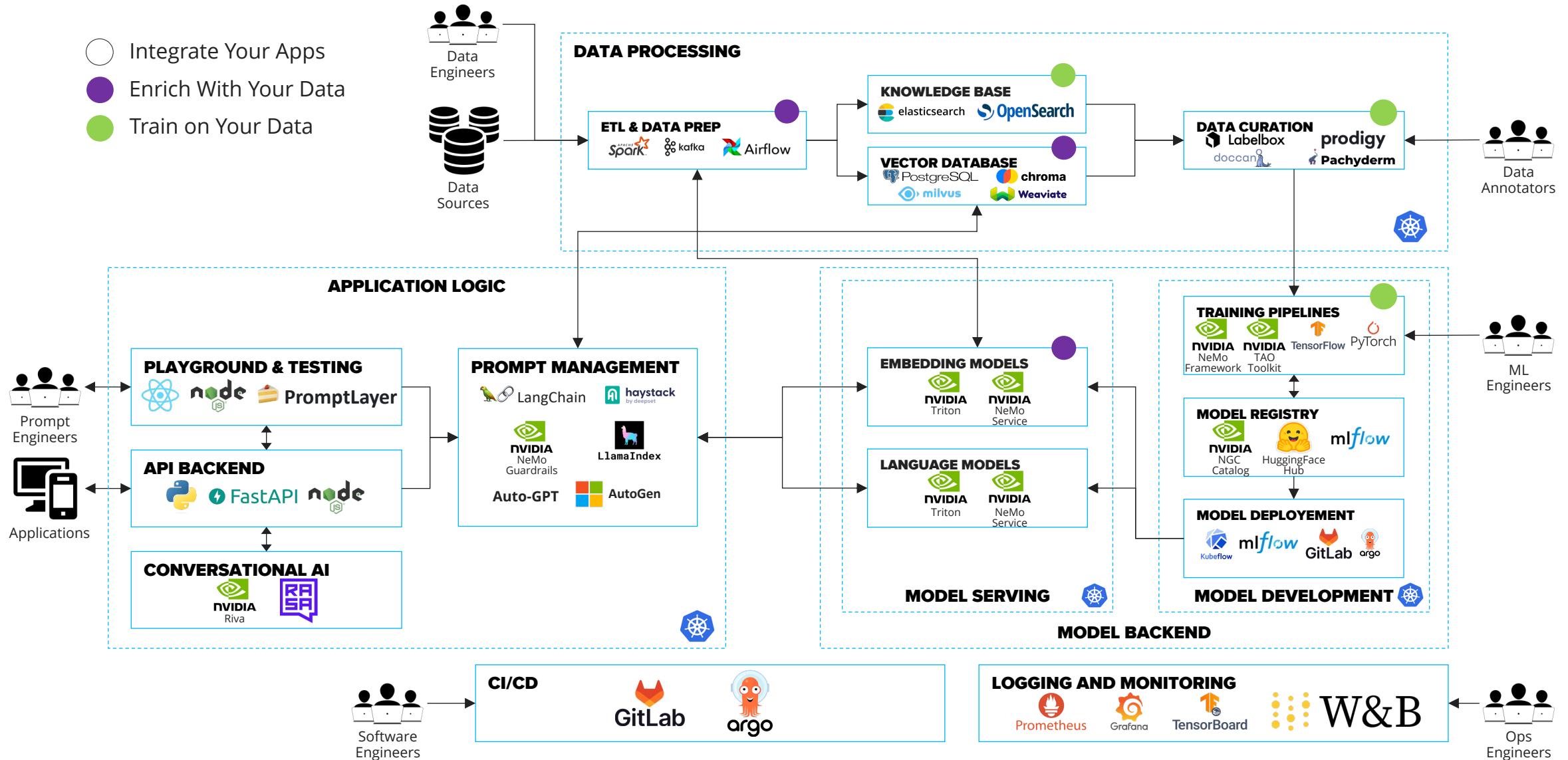
# HOW TO RAG WITH NVIDIA?



# DESIGN PATTERN: RETRIEVAL AUGMENTED GENERATION (RAG)



# GENERATIVE AI REFERENCE ARCHITECTURE



# 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:

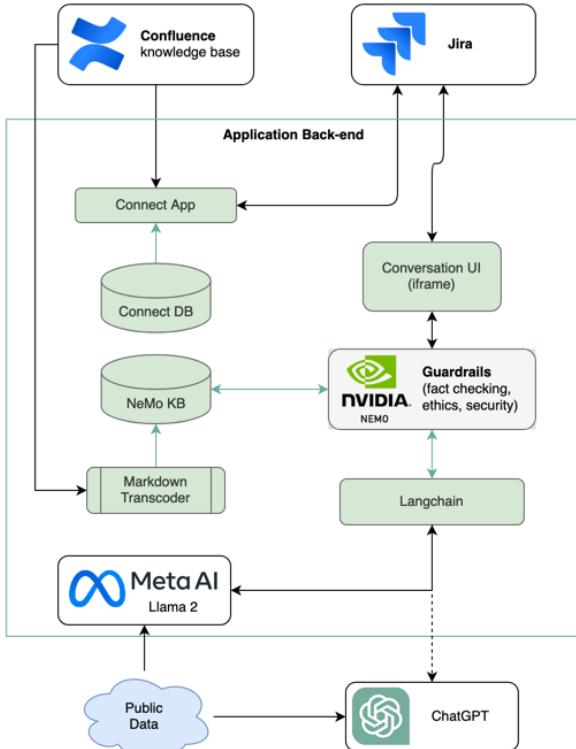
- Custom integration of Gen AI services with Confluence and Jira Cloud, utilizing large language models (LLMs) for improved security and data privacy.
- A conversational UI embedded into the Jira interface, facilitating efficient knowledge base utilization and ticket processing.
- Implementation of data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

The screenshot shows a Jira ticket titled "nick test / nick test / NT-1 Equipment policy at SoftServe". The ticket contains a message from an AI support bot asking about laptop standards. The interface includes navigation menus like "PLANNING", "DEVELOPMENT", and "ISSUES", and a sidebar with "Project pages" and "Add shortcut".



# ASSISTANT FOR CONFLUENCE: APPROACH

The screenshot shows a Jira issue page for a project named 'nick test'. The issue is titled 'Equipment policy at SoftServe'. The description section contains a message from a user asking about equipment standards and decommissioning age. The activity section shows a comment from 'Nick Turskyi' asking for advice. The right sidebar displays the issue's details, including assignee (Unassigned), reporter (Nick Turskyi), priority (Medium), and AWS Service Catalog (Open Requested Product). The sidebar also includes a link to 'Open AI Support'.

nick test Software project

PLANNING

- NT board Board
- Timeline
- Kanban board
- Reports

Issues

DEVELOPMENT

- Code
- Releases

Project pages

Add shortcut

Project settings

You're in a company-managed project

Learn more

Standard IT Workplace Equipment X Manage apps - Confluence X [NT-1] Equipment policy at Soft... X + https://nturskyitest.atlassian.net/browse/NT-1 4 13 1 Search ...

nick test / NT-1

## Equipment policy at SoftServe

Attach Create subtask Link issue ...

Description

Hello folks,

I'm new here at SoftServe so would be grateful to have a better understanding about equipment standards.

Please advise:

- what is the standard for powerful laptop at SoftServe?
- what is the decommission age once it is supposed to be changed?

Another question is related to opportunity to buy the device once it is decommissioned - I had such option in my previous company and think that it is a good benefit.

Thanks, Nick

Activity

Show: All Comments History Work log Newest first ↴

Add a comment... Pro tip: press M to comment

Backlog Actions

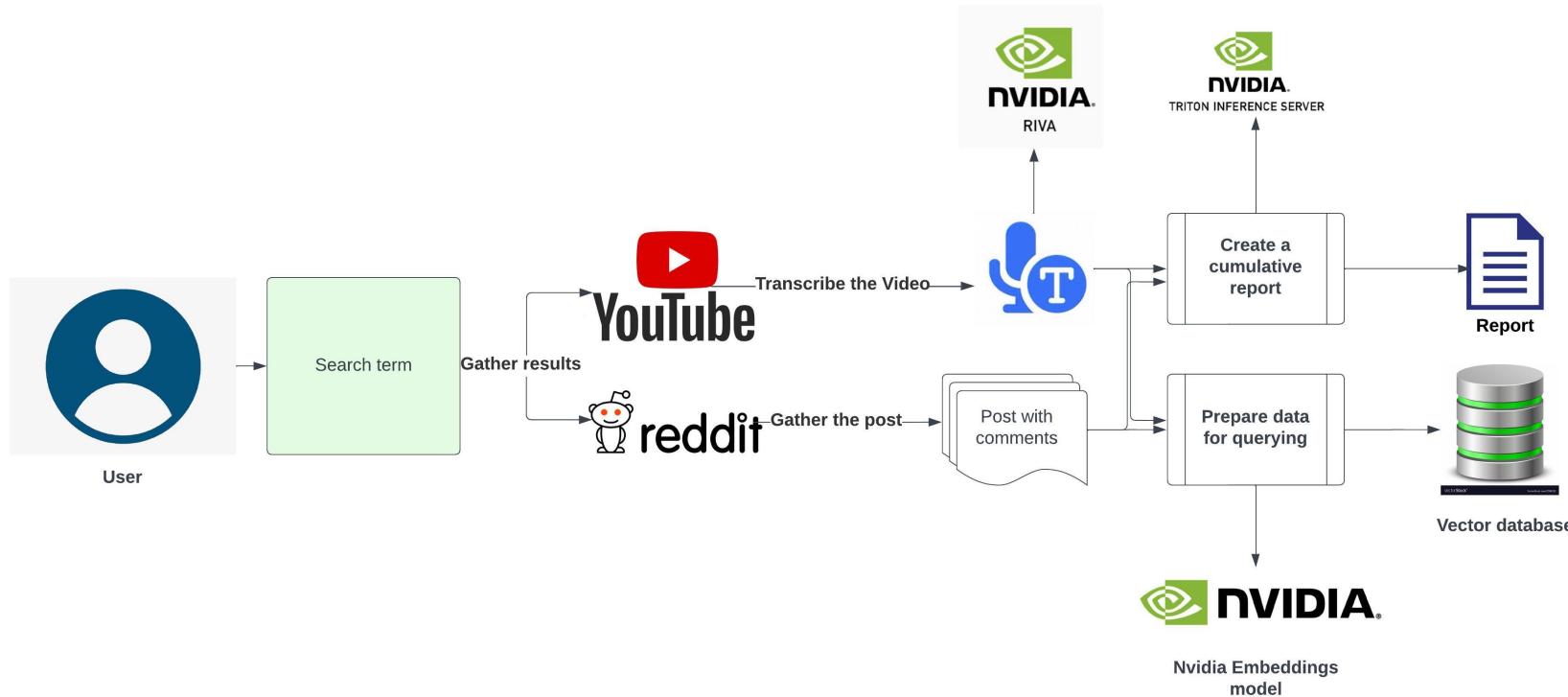
Details

Assignee Unassigned Assign to me Reporter Nick Turskyi Labels None Priority Medium AWS Service Catalog aws Open Requested Product AI Support Open AI Support

More fields Story Points, Original estimate, Time tracking...

Created December 8, 2021 at 1:14 PM Updated 6 minutes ago Configure

# 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

# USER FEEDBACK ANALYZER: DEMO

The image shows a screenshot of a user interface for a "User Feedback Analyzer: DEMO". The interface is divided into two main sections.

**Left Panel:** This panel contains three dropdown menus:

- YouTube data
- Reddit data
- Summaries

**Right Panel:** This panel has a title "Nvidia RAG demo" and a list of five items, each with a dropdown arrow:

- Collect YouTube data
- Collect Reddit data
- Create report
- Create collection in database
- Query the data

In the bottom right corner of the interface, there is a small cursor icon pointing towards the bottom right.

# Q&A

**FEELING LUCKY?**

**SIGN UP TO WIN  
A NINTENDO SWITCH**



**FOR  
THE  
FUTURE**

softserve