



# Build Your AI Companion: Crafting Custom Chatbots with Amazon Bedrock





## AGENDA

### PART 01

#### Intro

What is Generative AI?

### PART 02

#### Amazon Bedrock?

Building a GenAI Chatbot

### PART 03

#### Best Practices & Tips

Conclusion and Q&A



# INTRODUCTION

## Théo Lebrun



lebruntheo



Falydoor



@Falydoor



<https://theolebrun.com/>



**Data Engineer and Technical Manager**  
at Ippon Technologies

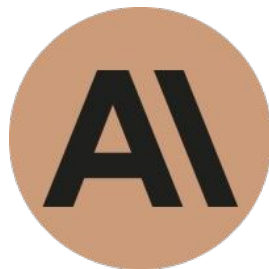


# What is Generative AI?

- **Generate Human-like Text**
  - Based on an input (or prompt)
  - Can be enhanced with a context (RAG)
- **Tailored**
  - Model provider
  - Model type
  - Temperature
- **Use Cases**
  - Customer support
  - Content creation
  - Virtual assistants
  - And more...



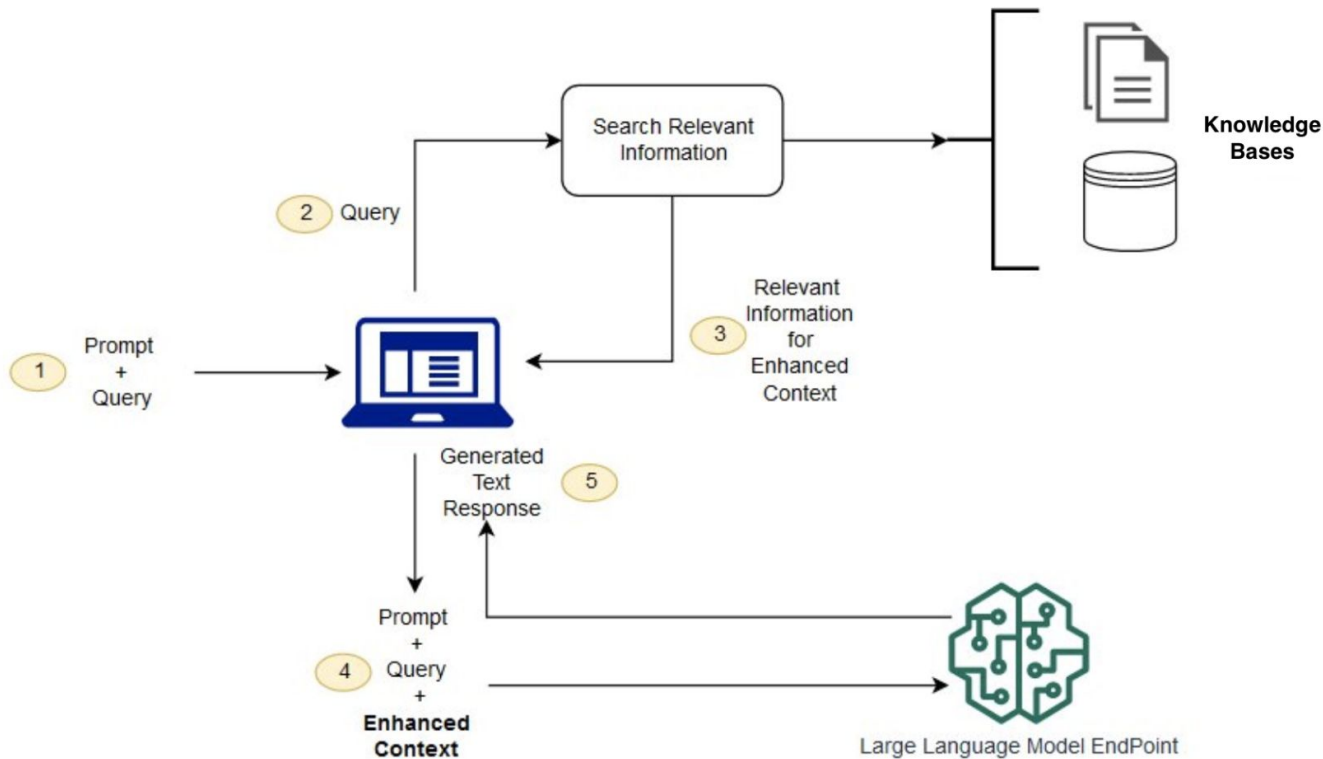
ChatGPT



Amazon Nova



# Retrieval-Augmented Generation (RAG)



# Amazon Bedrock

- **Fully Managed Serverless Service**
  - High-performing foundation models (Llama, Claude, Nova, Mistral, etc.)
  - Interactive playgrounds
  - Bring your own custom models
- **Knowledge Bases**
  - S3, web crawler
  - Third party (Salesforces, Confluence, etc)
- **Builder Tools**
  - Prompt management (versioning, caching, etc.)
  - AI Agents
- **Data Protection and Privacy**
  - Encrypted in transit and at rest
  - Not shared with model providers
  - Monitoring and logging capabilities



Amazon Bedrock



# Developing Your Chatbot

- **Picking a LLM Framework**

- **LLM and Embeddings**

- Vectorstore
- Model

- **Bot Platform**

- Gradio
- Slack, Teams, etc

- **DevOps**

- Deployment
- Scaling



**LangChain**







- **GenAI Slack Chatbot**
- **Knowledge Bases**
  - Resumes
  - Case studies
  - Internal docs (handbook, healthcare benefits, etc)



**Théo LEBRUN** 18:45

Who has Snowflake experience?



**IpponAssistant** APP 18:45

Based on the context documents provided, the following individuals at Ippon have experience with Snowflake:

- Stephen [REDACTED] - Has SnowPro Core Certification and experience architecting and implementing Snowflake data warehouses for clients.
- Christopher [REDACTED] - Has Snowflake Professional Architect certification and experience leading Snowflake implementations for clients.
- Pooja [REDACTED] - Developed Python jobs to replicate data from various sources into Snowflake using Airflow and has experience maintaining and contributing to data architecture in Snowflake.
- Nicolas [REDACTED] - Has experience managing Snowflake implementations and migrations as part of his previous role as Business Manager at Ippon.

I did not see any other mentions of individuals having specific Snowflake experience in the context documents provided. Let me know if you need any additional details about the Snowflake experience of these individuals. (edited)





# Knowledge Bases

- **Sources**

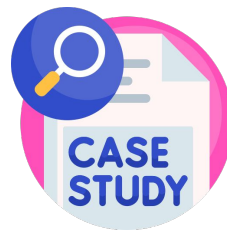
- File based
- Database
- Third party

- **Data Requirements**

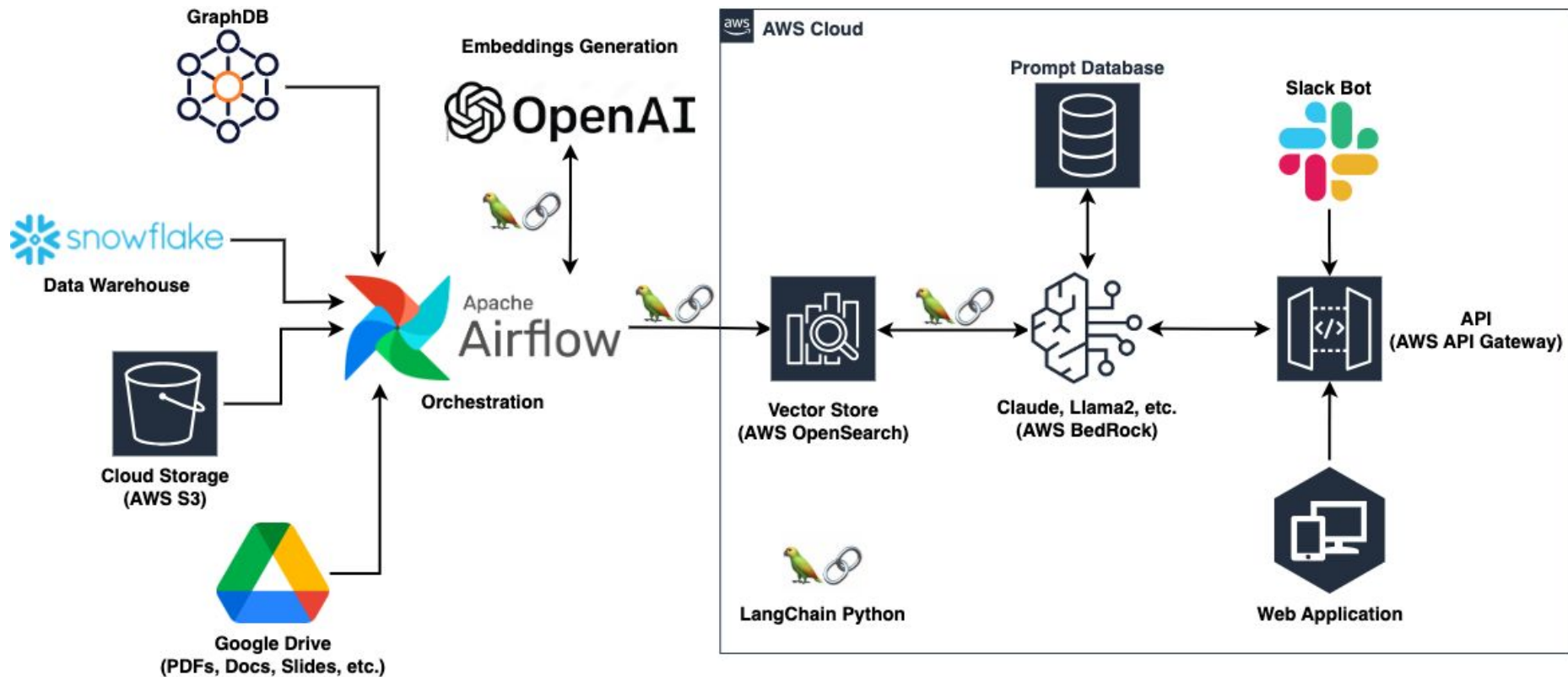
- Collect, clean and standardize
- Update frequency

- **Document Integration**

- Type of documents (doc, pdf, txt, etc)
- Chunking strategies



# Architecture



# Best Practices

- **Security Considerations**
  - User's data not shared
  - Amazon Bedrock Guardrails
- **Ongoing Maintenance**
  - Monitoring
  - Feedback system
- **AI Agents for Amazon Bedrock**
  - Automate repetitive tasks
  - Improve generated response
  - Reduce LLM hallucinations



# Tips and Tricks

- **Conversational Retrieval**

- Timestamp, who is speaking, etc
- History, converse API
- Include source documents
- Multi-index and prompt (reduce LLM hallucination)

- **Performance Optimization**

- Streaming
- Max tokens
- Model type (SLMs vs LLMs)



# Bedrock Pricing

- **Pricing Plans**

- On Demand (pay for what you use)
- Batch (50% cheaper)
- Provisioned Throughput (charged by the hour)

- **SLMs vs LLMs**

Meta models	1,000 input tokens	1,000 output tokens	1,000 input tokens (batch)	1,000 output tokens (batch)
Llama 3.1 Instruct (8B)	\$0.00022	\$0.00022	\$0.00011	\$0.00011
Llama 3.1 Instruct (70B)	\$0.00072	\$0.00072	\$0.00036	\$0.00036
Llama 3.1 Instruct (405B)	\$0.0024	\$0.0024	\$0.0012	\$0.0012



# Useful Links

- <https://aws.amazon.com/bedrock/>
- <https://www.langchain.com/>
- <https://blog.ippon.tech/building-a-document-powered-chatbot-with-langchain-amazon-bedrock-and-rag>
- <https://github.com/Falydoor/talks>



# Thank You for Your Time!

I would love to chat more with you about building a successful Gen AI Chatbot using Amazon Bedrock or any other technology.

You can email me at [tlebrun@ipponusa.com](mailto:tlebrun@ipponusa.com) or message me on 





# Q & A

