Fine Tuning Foundational Models: With and Without RAG

This presentation explores the differences and best practices for fine-tuning foundational models, with and without retrieval augmented generation (RAG), providing insights and practical examples.

by The XYZ Company

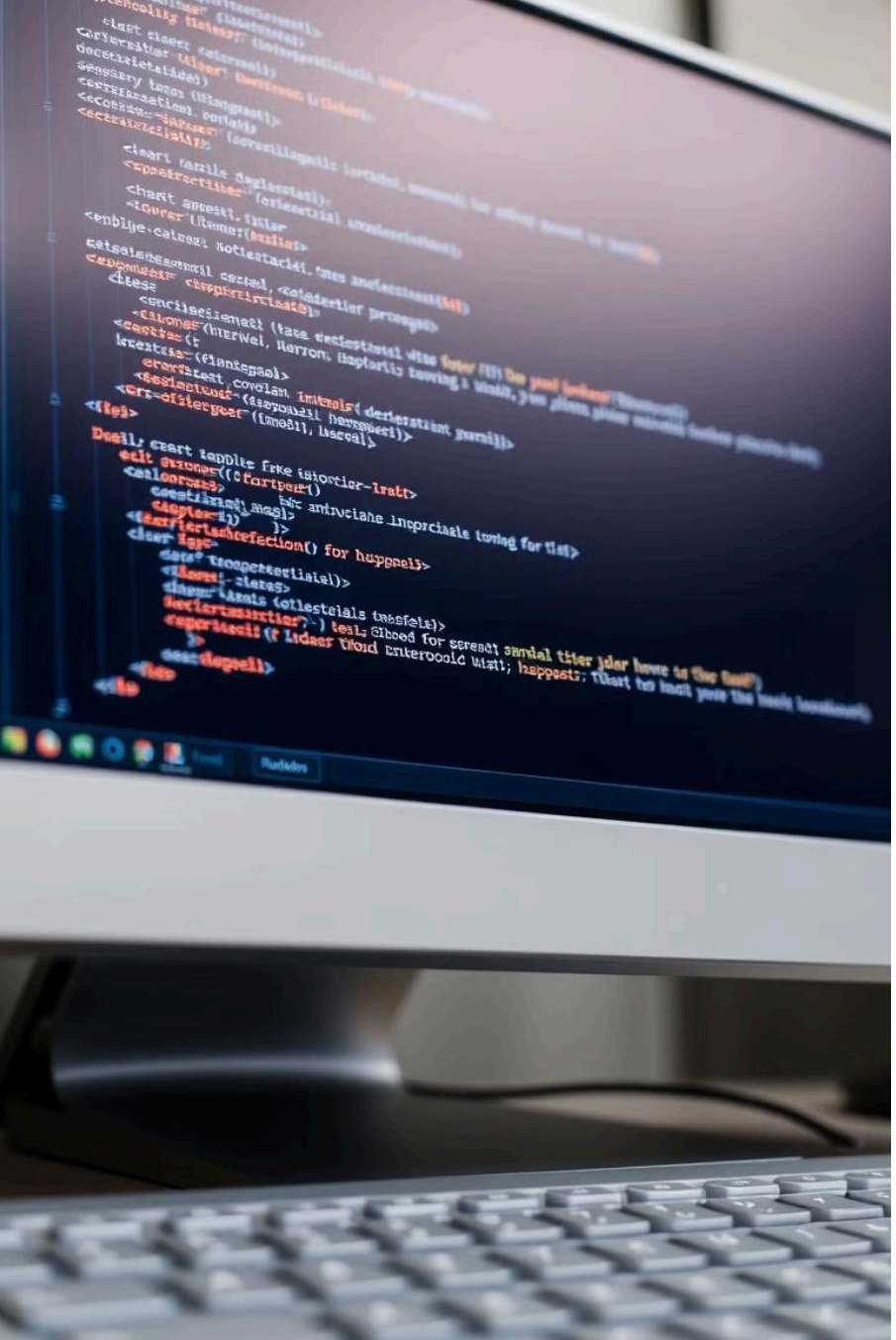
Introduction to Foundational Models

Definition

Foundational models are large language models (LLMs) trained on vast datasets of text and code. They possess a wide range of capabilities, including text generation, translation, and code completion.

Examples

Examples of foundational models include GPT-3, LaMDA, and PaLM. These models are often used as a starting point for building specific AI applications.



Fine Tuning: Tailoring the Model

Advantages

Fine-tuning allows for customization to specific domains and tasks. It can improve performance and accuracy in targeted scenarios.

Limitations

Fine-tuning requires a significant amount of labeled data, which can be expensive and time-consuming to acquire. It can also lead to overfitting, where the model performs well on training data but poorly on new data.

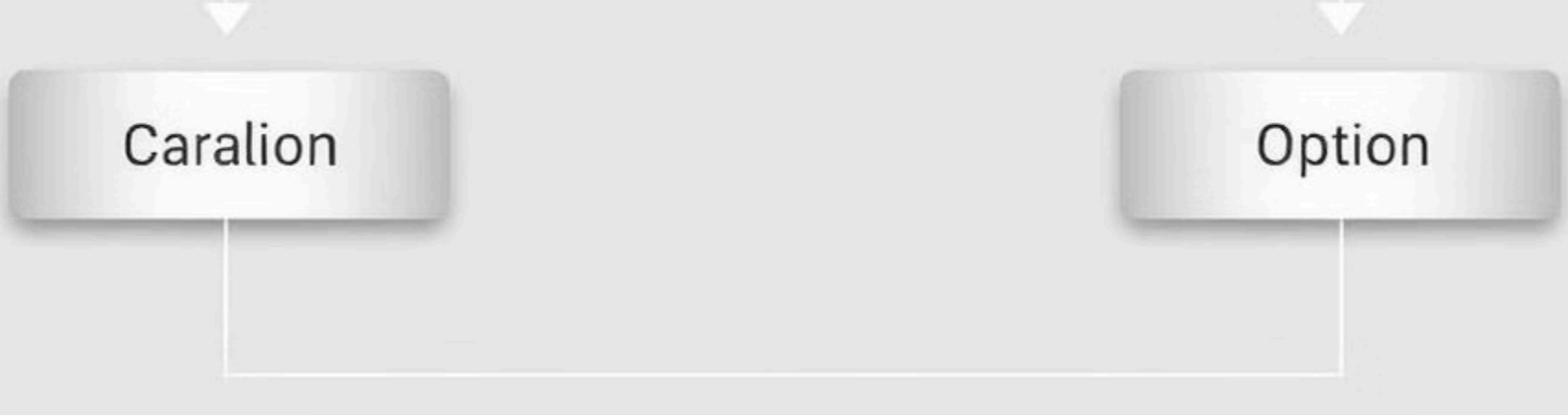
Retrieval Augmented Generation (RAG)

Retrieval

RAG combines retrieval and generation. It uses a knowledge base or external data sources to retrieve relevant information and integrate it into the generated text.

Generation

The retrieved information is then used by a language model to generate more informed and accurate responses.



Fine Tuning vs RAG: When to Use Each

1

Fine Tuning

Use fine-tuning when you have a large dataset of labeled data specific to your task and domain.

2

RAG

Use RAG when you have a vast external knowledge base or need to access information dynamically without re-training the model.

Architectural Considerations

Fine Tuning

The model architecture is typically based on the foundational model, with additional layers added for domain-specific learning.

RAG

The architecture involves a retrieval component, a language model, and a knowledge base or external data source. The retrieval component identifies relevant information, which is then passed to the language model for generation.

Practical Examples and Use Cases

1

Chatbots

Fine-tuning or RAG can be used to create chatbots that are more knowledgeable and engaging.

2

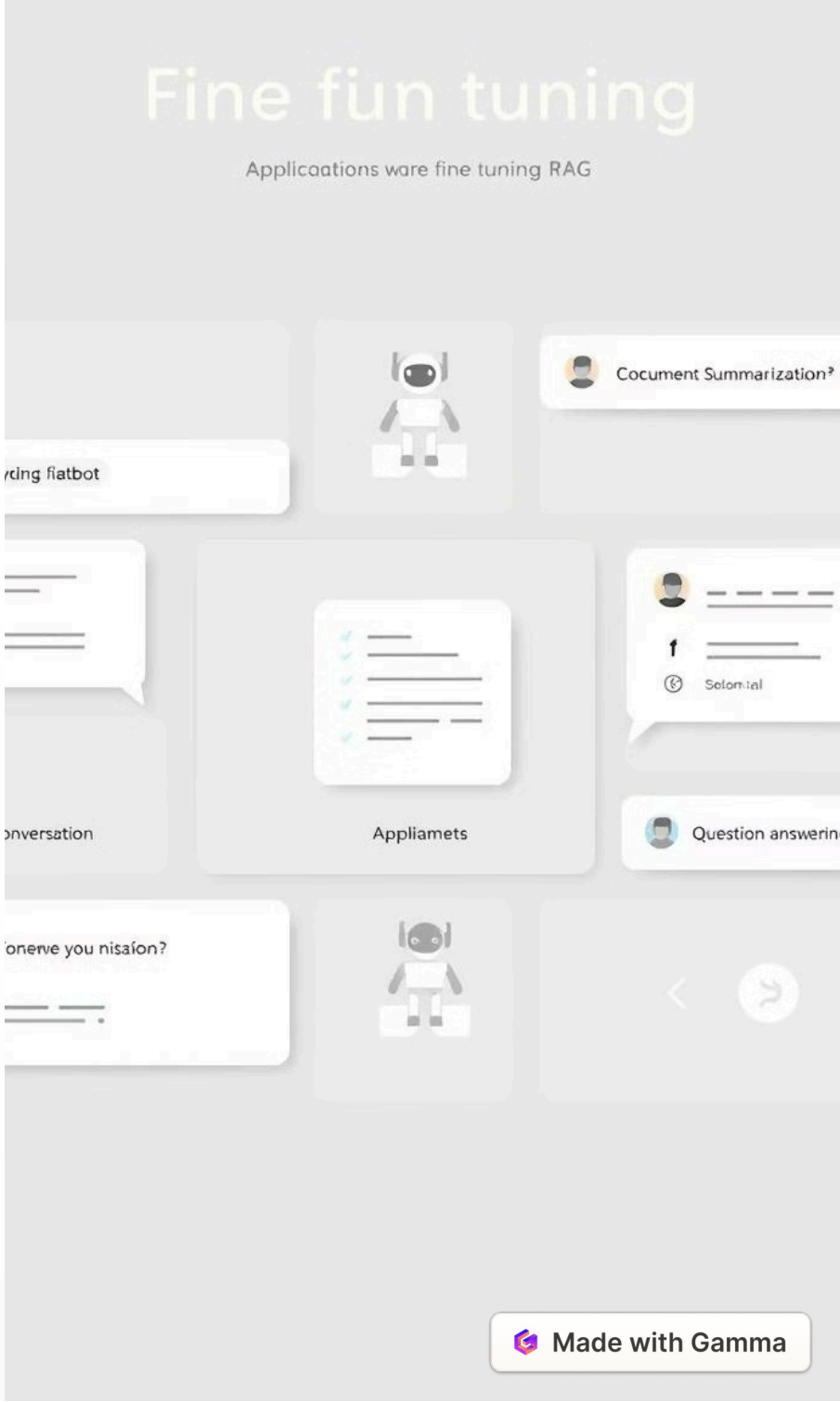
Document Summarization

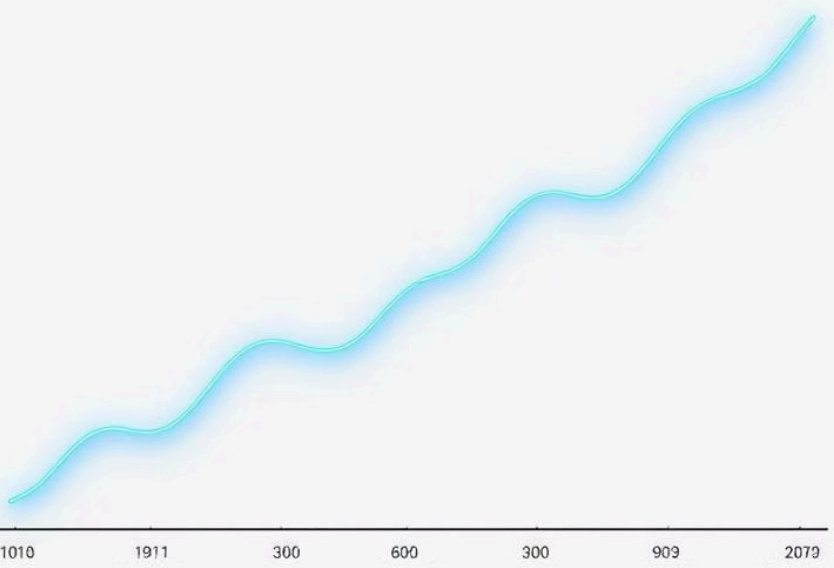
RAG can summarize lengthy documents by retrieving and highlighting key information.

3

Question Answering

RAG can answer questions by retrieving relevant information from a knowledge base and generating a response.





Conclusion and Key Takeaways

Both fine-tuning and RAG are valuable techniques for enhancing foundational models. Fine-tuning is suitable for domain-specific tasks with labeled data, while RAG excels in scenarios requiring dynamic information retrieval. Selecting the right approach depends on the specific requirements of your application.