DATA & AI

FINE-TUNING VS RAG

How to choose the right approach?

REPLY
CLUSTER

MAXIMIZE IMPACT AND EFFICIENCY

FINE-TUNING A LANGUAGE MODEL

Customizing a pre-trained model on a specific dataset to improve performance for a particular task or domain.

RETRIEVAL-AUGMENTED GENERATION (RAG)

Combining a retrieval system with a generative model to generate answers by fetching relevant information from a large dataset or knowledge base.

BUSINESS NEEDS

Real-time analysis of video sources

ADAPTABILITY AND SCALABILITY

Businesses need AI systems that quickly adapt to new data and able to handle varying data volumes, complexities, and growth over time.

ACCURACY

An effective chatbot should prioritize both efficiency and accuracy. It should quickly detect questions while maintaining a high level of precision in the answers.

COST EFFICIENCY

Managing costs is crucial, particularly regarding development, maintenance, and scaling.

CUSTOMIZATION

Tailored Functionality configuring algorithms and models to perform specific tasks, addressing unique requirements and scenarios for specialized applications.

KEY FEATURES

FINE-TUNING	RAG
High customization	Dynamic information retrieval
Requires domain-specific data	Leverages existing knowledge bases
Higher cost	Lower cost
Better for static environments	Better for dynamic environments

SUMMARY

	Customization	Opex Cost	Scalability	Performance	Adaptability
FINE-TUNING	High, domain-specific	High (training)	Limited by data and model size	High for specific tasks	Moderate (needs retraining)
RAG	Medium, dependent on retrieval quality	Lower	Highly scalable with large knowledge base	Flexible, but performance varies	High (dynamic retrieval from updated data)

DATA & AI

REPLY
CLUSTER

Contact us!

Fabio Solazzo
f.solazzo@reply.it

Davide Pietrasanta
d.pietrasanta@reply.it

WWW.CLUSTERREPLY.IT