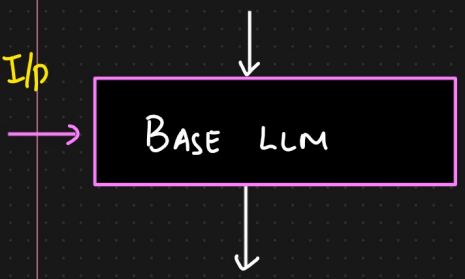
RAG Vs Fine Tuning Vs Prompt Engineering

[AI Assistant]

1) Prompt Engineering

[Act as a chef, provide me a recipe]



1) Specific Instructions
2) Structured Prompt with Clear context
3) Model Remain Unchanged

Pro's

1) No technical expertise need
2) Instant Results
3) No training cost (Free)
4) Works with any LLM.

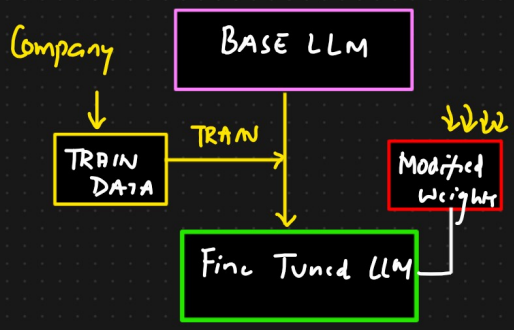
Con's

1) Limited by Models base Knowledge
2) Inconsistent Results
3) Token limit Restrict Complexity
4) Can't add New Knowledge.

Best for

1) Small Scale Application

2) Fine Tuning



1) Prepare domain specific Training data
2) Train model on data.
3) Create a Specilized version.

Pro's

1) Deeply Specialized Knowledge
2) Consistent behaviour
3) No prompt Engineering
4) Better for Specific domain

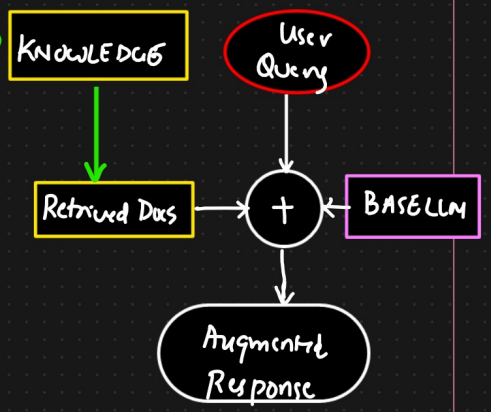
Con's

1) Expensive [GPU]
2) Require ML Expertise
3) Retraining for update

Best for

1) Specific style

3) RAG



1) Store document in Vector DB
2) Retrieve relevant docs for each Query
3) LLM generates answer from context

Pro's

1) Always up-to date info
2) No training required
3) Cost-effective
4) Accuracy is high.
5) Can private/property data.

Con's

1) Infrastructure setup
2) Retrieval Quality affects Results
3) Context Window limitation.

Best for

2) Generic purpose task
3) Quick prototyping.

② High volume
③ When accuracy is critical