RAG vs Finetuning vs Both

Supervised Finetuning (SFT) has fast become a popular method to customise and adapt foundation models for specific objectives. There has been a growing debate in the applied AI community around the application of fine-tuning or RAG to accomplish tasks.

RAG & SFT should considered as complementary, rather than competing, techniques.

RAG enhances the non-parametric memory of a foundation model without changing the parameters

SFT changes the parameters of a foundation model and therefore impacting the parametric memory

If the requirement dictates changes to the parametric memory and an increase in the non-parametric memory, then RAG and SFT can be used in conjunction

RAG Features

Connect to dynamic external data sources ✓

Reduce hallucinations ✓

Increase transparency (in terms of source of information) ✓

Works well only with very large foundation models ✗

Does not impact the style, tone, vocabulary of the foundation model ✗

SFT Features

Change the style, vocabulary, tone of the foundation model ✓

Can reduce model size ✓

Useful for deep domain expertise ✓

May not address the problem of hallucinations ✗

No improvement in transparency (as black box as foundation models) ✗

Abhinav Kimothi

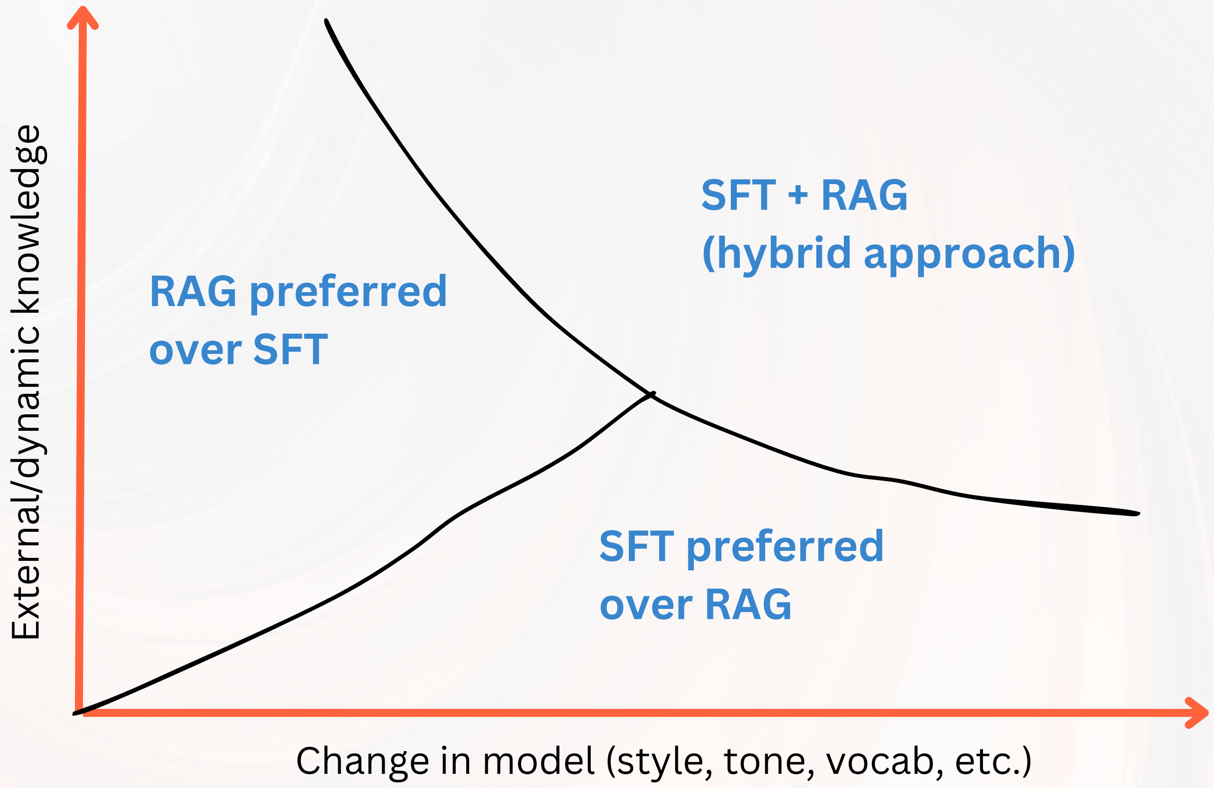
Important Use Case Considerations

Do you require usage of dynamic external data?

RAG preferred over SFT

Do you require changing the writing style, tonality, vocabulary of the model?

SFT preferred over RAG


External/dynamic knowledge
Change in model (style, tone, vocab, etc.)
RAG preferred over SFT
SFT + RAG (hybrid approach)
SFT preferred over RAG

RAG should be implemented (with or without SFT) if the use case requires

- Access to an external data source, especially, if the data is dynamic
- Resolving Hallucinations
- Transparency in terms of the source of information

For SFT, you'll need to have access to labelled training data

Other Considerations

Latency

RAG pipelines require an additional step of searching and retrieving context which introduces an inherent latency in the system

Scalability

RAG pipelines are modular and therefore can be scaled relatively easily when compared to SFT. SFT will require retraining the model with each additional data source

Cost

Both RAG and SFT warrant upfront investment. Training cost for SFT can vary depending on the technique and the choice of foundation model. Setting up the knowledge base and integration can be costly for RAG

Expertise

Creating RAG pipelines has become moderately simple with frameworks like LangChain and LlamaIndex. Fine-tuning on the other hand requires deep understanding of the techniques and creation of training data