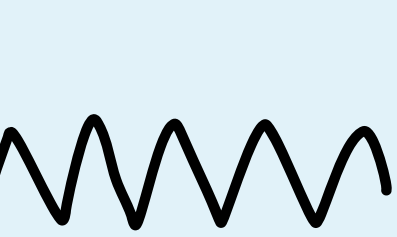
MASTERING LLM PRESENTS

COFFEE BREAK CONCEPTS

How to select the right LLM Model for use case?



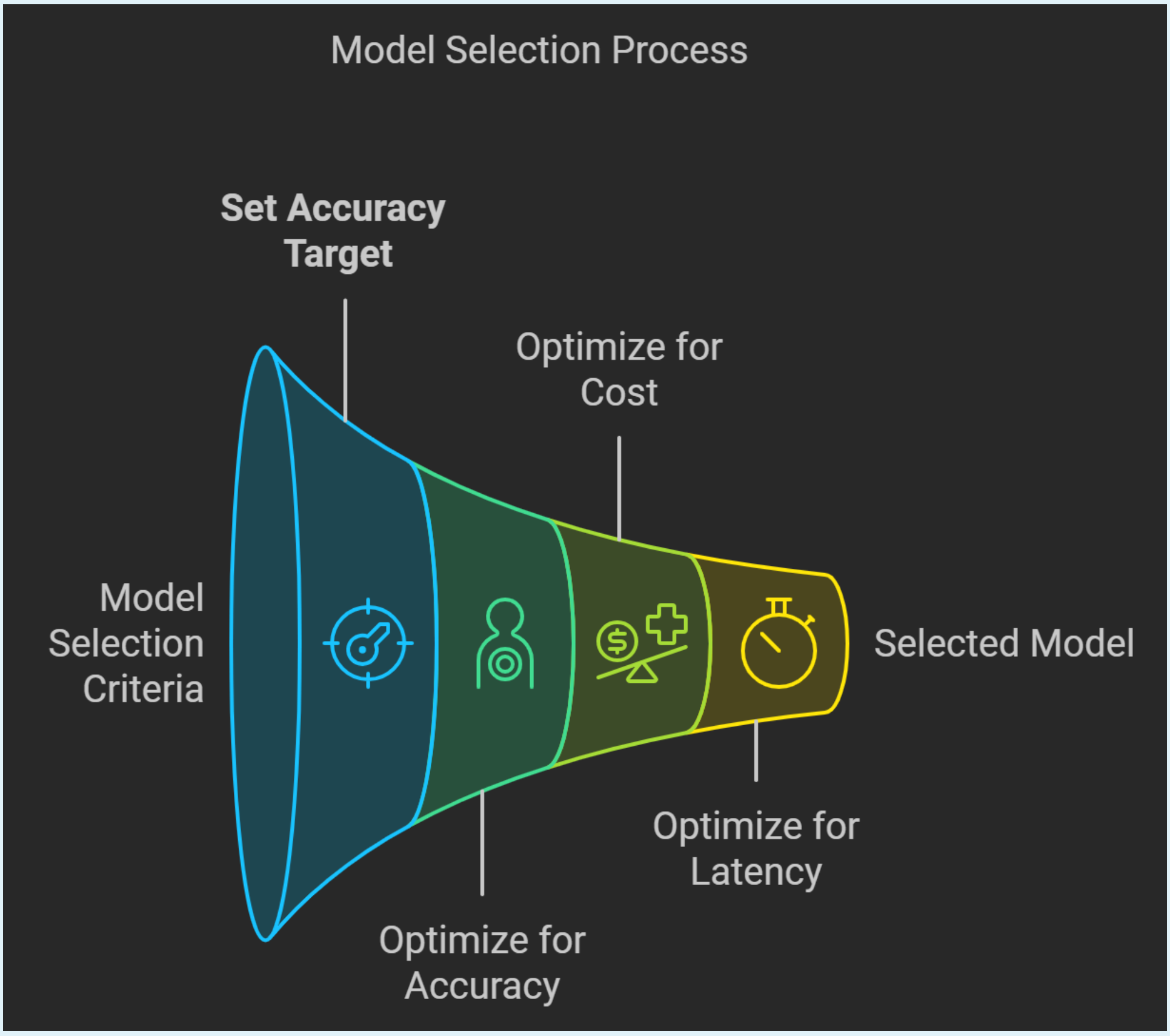
Follow us on

Credit to open AI team

Core Principles

The principles for model selection are simple:

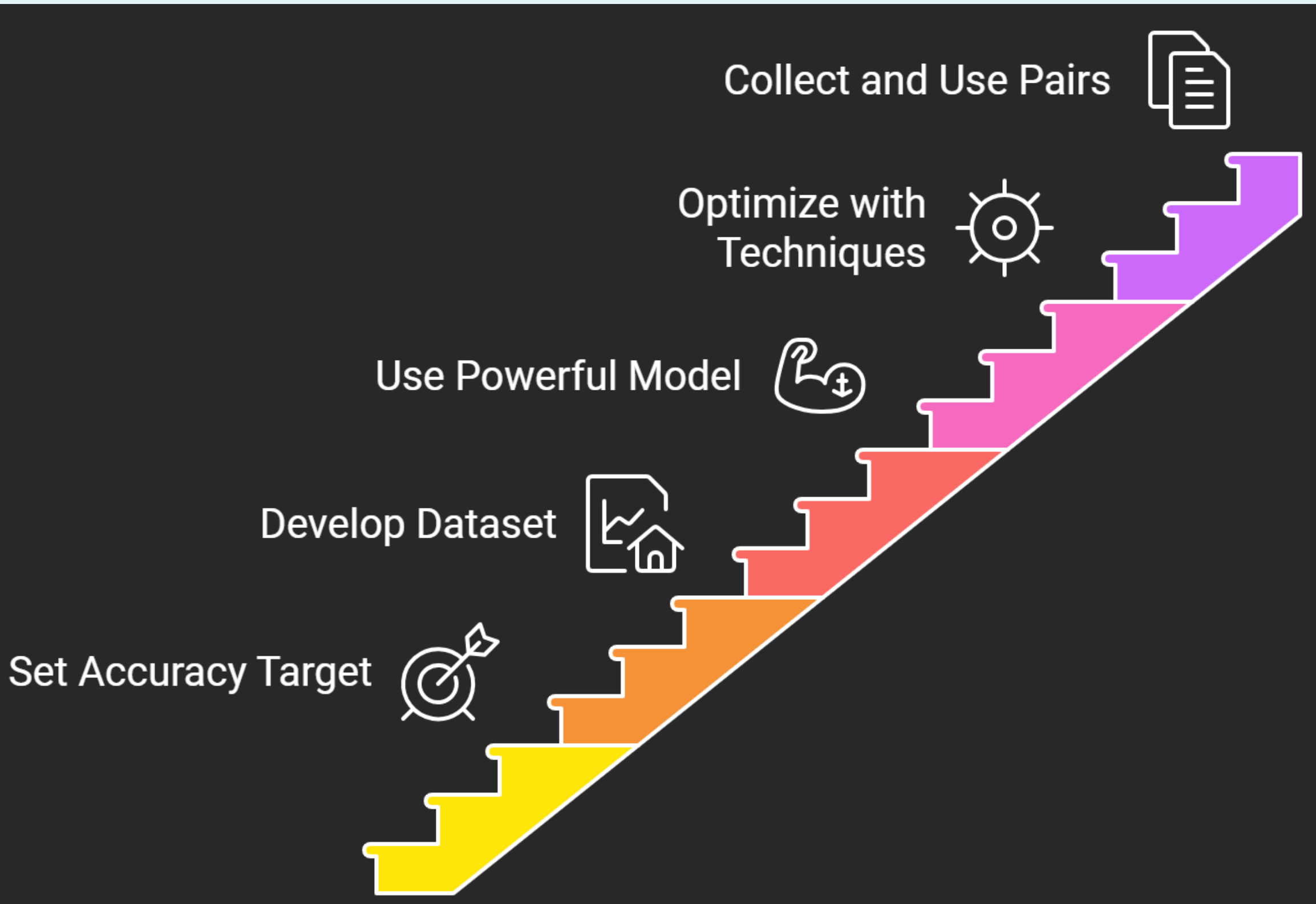
1. **Optimize for accuracy first**: Optimize for accuracy until you hit your accuracy target.
2. **Optimize for cost and latency second**: Then aim to maintain accuracy with the cheapest, fastest model possible.



Focus on Accuracy First

1. **Set a Clear Accuracy Goal**: Define what accuracy is "good enough" for your use case. Example: 90% of customer service calls triaged correctly on the first interaction
2. **Develop an Evaluation Dataset**: Create a dataset to measure the model's performance. Example: Capture 100 interaction examples, including user requests, model triage, correct triage, and accuracy
3. **Use the Most Powerful Model**: Start with the most capable model to achieve your accuracy targets. Log responses for future use.
4. **Optimize for Accuracy**
    a. Use retrieval-augmented generation
    b. Fine-tune for consistency and behavior
5. **Collect Data for Future Use**: Gather prompt and completion pairs for evaluations, few-shot learning, or fine-tuning. This practice, known as prompt baking, helps produce high-quality examples for future use.

Steps to Achieve Model Accuracy

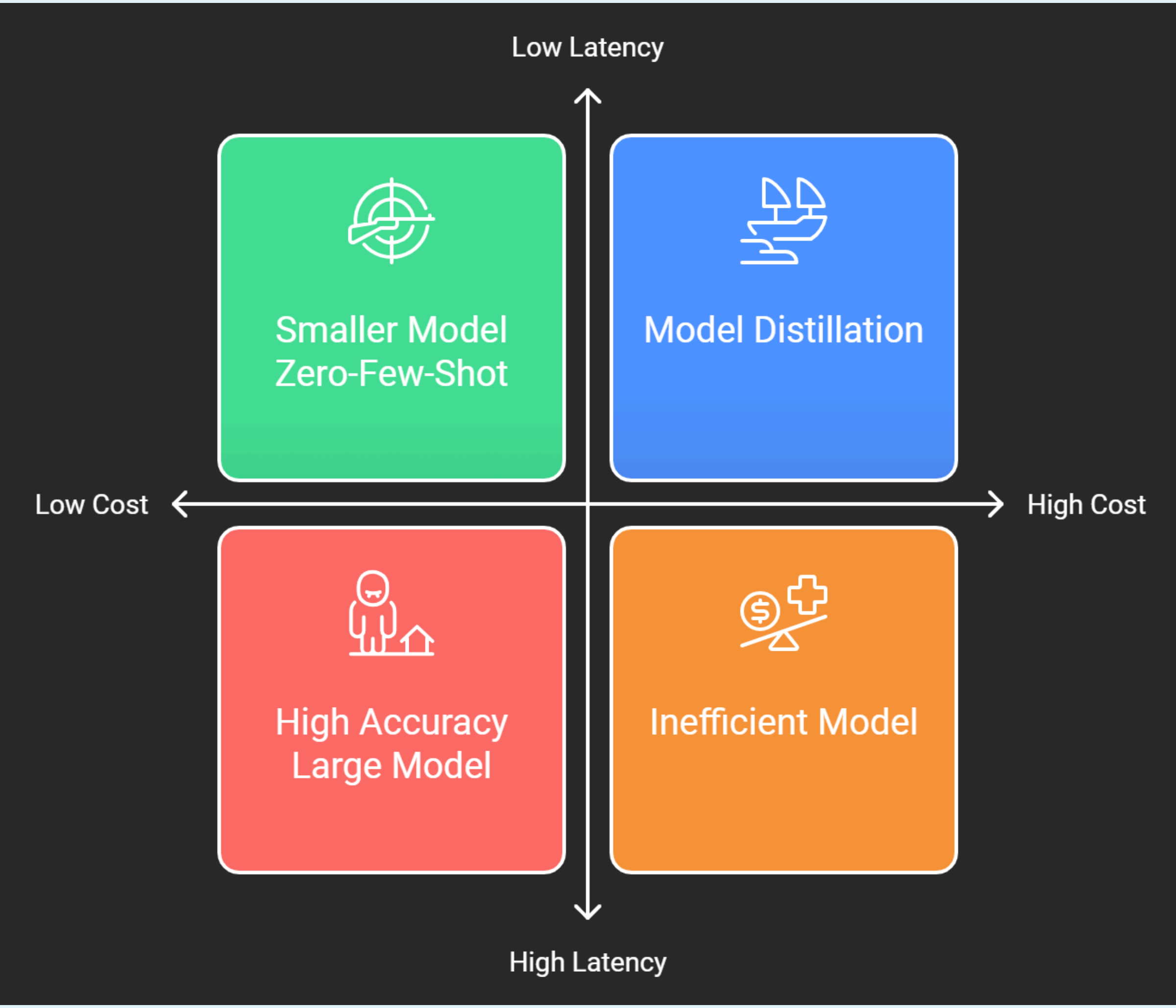


Optimize cost and latency

Cost and latency are secondary because if the model can’t hit your accuracy target then these concerns are moot. However, once you’ve got a model that works for your use case, you can take one of two approaches:

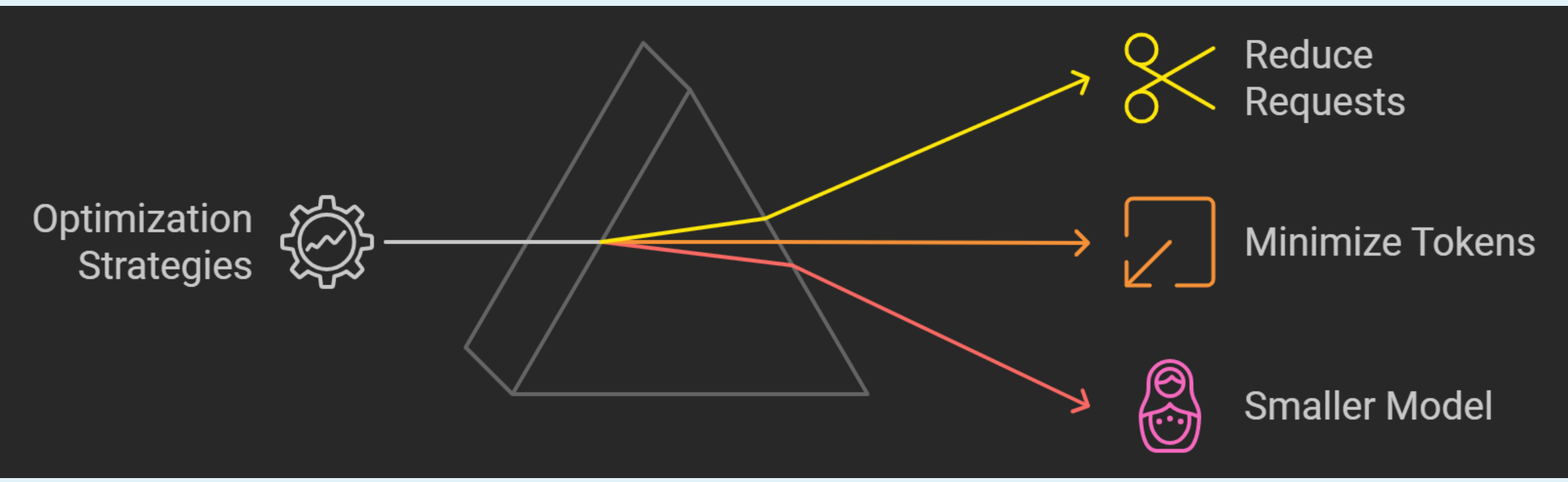
- **Compare with a smaller model zero- or few-shot:** Swap out the model for a smaller, cheaper one and test whether it maintains accuracy at the lower cost and latency point.
- **Model distillation**: Fine-tune a smaller model using the data gathered during accuracy optimization.
- **Cost and latency are typically interconnected**; reducing tokens and requests generally leads to faster processing.

Cost vs Latency



Main Stratergy

- **Reduce requests**: Limit the number of necessary requests to complete tasks.
- **Minimize tokens**: Lower the number of input tokens and optimize for shorter model outputs.
- **Select a smaller model**: Use models that balance reduced costs and latency with maintained accuracy.



Practical example from open AI

To demonstrate these principles, they have developed a fake news classifier with the following target metrics:

- **Accuracy**: Achieve 90% correct classification
- **Cost**: Spend less than $5 per 1,000 articles
- **Latency**: Maintain processing time under 2 seconds per article

ID	METHOD	ACCURACY	ACCURACY TARGET	COST	COST TARGET	AVG. LATENCY	LATENCY TARGET
1	gpt-4o zero-shot	84.5%		$1.72		< 1s	
2	gpt-4o few-shot (n=5)	91.5%	✓	$11.92		< 1s	✓
3	gpt-4o-mini fine-tuned w/ 1000 examples	91.5%	✓	$0.21	✓	< 1s	✓

They ran three experiments to reach the goal:

1. **Zero-shot**: Used GPT-4o with a basic prompt for 1,000 records, but missed the accuracy target.
2. **Few-shot learning**: Included 5 few-shot examples, meeting the accuracy target but exceeding cost due to more prompt tokens.
3. **Fine-tuned model**: Fine-tuned GPT-4o-mini with 1,000 labeled examples, meeting all targets with similar latency and accuracy but significantly lower costs.

Summary

- Optimize for accuracy first & followed by Optimization for cost and latency.
- This process is important - you often can’t jump right to fine-tuning because you don’t know whether fine-tuning is the right tool for the optimization you need, or you don’t have enough labeled examples.
- Use a large accurate model to achieve your accuracy targets, and curate a good training set - then go for a smaller, more efficient model with fine-tuning.

www.masteringllm.com

LLM Interview Course

Want to Prepare yourself for an LLM Interview?

- ✓ 100+ Questions spanning 14 categories with Real Case Studies
- ✓ Curated 100+ assessments for each category
- ✓ Well-researched real-world interview questions based on FAANG & Fortune 500 companies
- ✓ Focus on Visual learning
- ✓ Certification

Coupon Code - LLM50

Coupon is valid till 30th Sep 2024

www.masteringllm.com

AgenticRAG with LlamaIndex 

Want to learn why AgenticRAG is future of RAG?

- Master **RAG fundamentals** through practical case studies
- Understand how to overcome **limitations of RA**G
- Introduction to **AgenticRAG** & techniques like **Routing Agents, Query planning agents, Structure planning agents, and React agents with human in loop**.
- **5** real-time **case studies with code walkthroughs**