FINE-TUNING ⇝ Create a specialized model

- Reduces hallucinations
- Customizes the model to a specific case

Prompt Engineering vs. Finetuning

	Prompting	Finetuning
Pros	• No data to get started • Smaller upfront cost • No technical knowledge needed • Connect data through retrieval (RAG)	• Nearly unlimited data fits • Learn new information • Correct incorrect information • Less cost afterwards if smaller model • Use RAG too
Cons	• Much less data fits • Forgets data • Hallucinations • RAG misses, or gets incorrect data	• More high-quality data • Upfront compute cost • Needs some technical knowledge, esp. data
	Generic, side projects, prototypes	Domain-specific, enterprise, production usage, ...privacy!

Benefits of finetuning your own LLM.

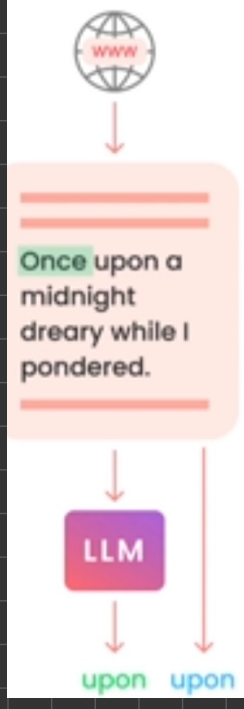
PERFORMANCE
- Stop hallucinations
- Increase consistency
- Reduce unwanted information

PRIVACY
- on-prem or VPC (Virtual private cloud)
  - Application hosted locally
  - On site ⇝ within a physical data center
- Prevent leakage
- No breaches

RELIABILITY
- control uptime
- lower latency
- moderation

COST
- lower cost per request
- increased transparency
- greater control

Pretraining



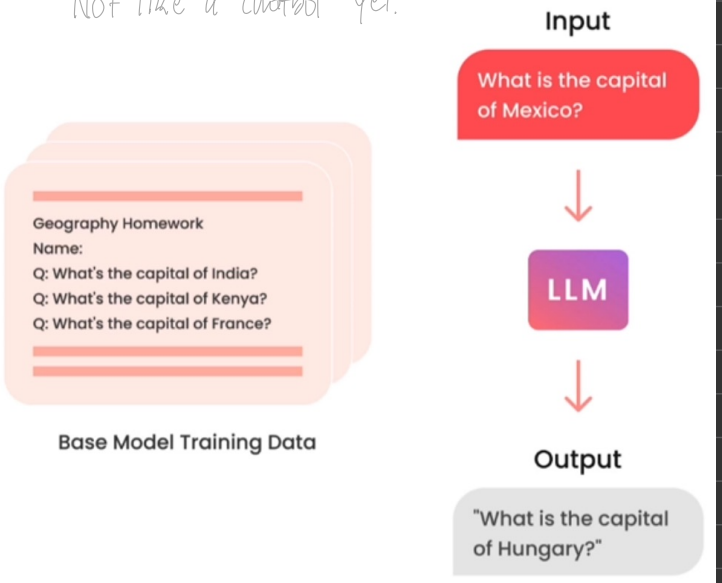
- Model at the start:
  - Zero knowledge about the world
  - Can't form English words
- Next token prediction
- Giant corpus of text data
- Often scraped from the internet: "unlabeled"
- Self-supervised learning
- After Training
  - Learns language
  - Learns knowledge

What is "data scraped from the internet"?

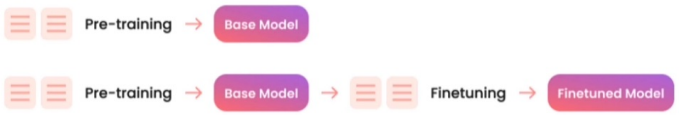
- Often not publicized how to pretrain
- Open-source pretraining data: "The Pile"
- Expensive & time-consuming to train

Limitations of pretrained base models

Not like a chatbot yet.



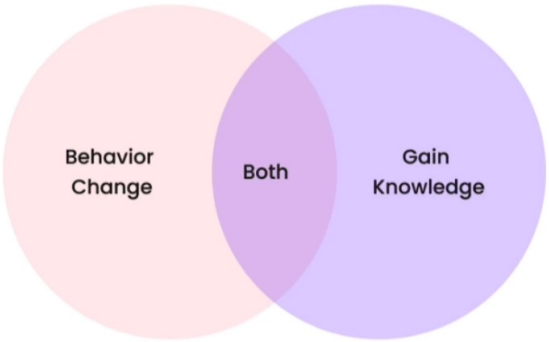
Finetuning after pretraining



- Finetuning usually refers to training further
  - Can also be self-supervised unlabeled data
  - Can be "labeled" data you curated
  - Much less data needed
  - Tool in your toolbox
- Finetuning for generative tasks is not well-defined:
  - Updates entire model, not just part of it
  - Same training objective: next token prediction
  - More advanced ways reduce how much to update (more later!)

What is finetuning doing for you?

- Behavior change
  - Learning to respond more consistently
  - Learning to focus, e.g. moderation
  - Teasing out capability, e.g. better at conversation
- Gain knowledge
  - Increasing knowledge of new specific concepts
  - Correcting old incorrect information



Tasks to finetune

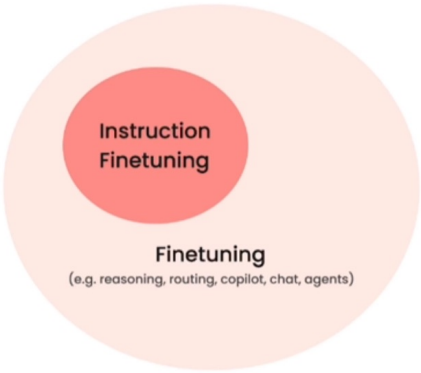
- Just text-in, text-out:
  - Extraction: text in, less text out
    - "Reading"
    - Keywords, topics, routing, agents (planning, reasoning, self-critic, tool use), etc.
  - Expansion: text in, more text out
    - "Writing"
    - Chat, write emails, write code
- Task clarity is key indicator of success
- Clarity means knowing what's bad vs. good vs. better

⇝ FINE-TUNING STEP

1. Identify task(s) by prompt-engineering a large LLM
2. Find tasks that you see an LLM doing ~ok at
3. Pick one task
4. Get ~1000 inputs and outputs for the task
5. Finetune on this data.

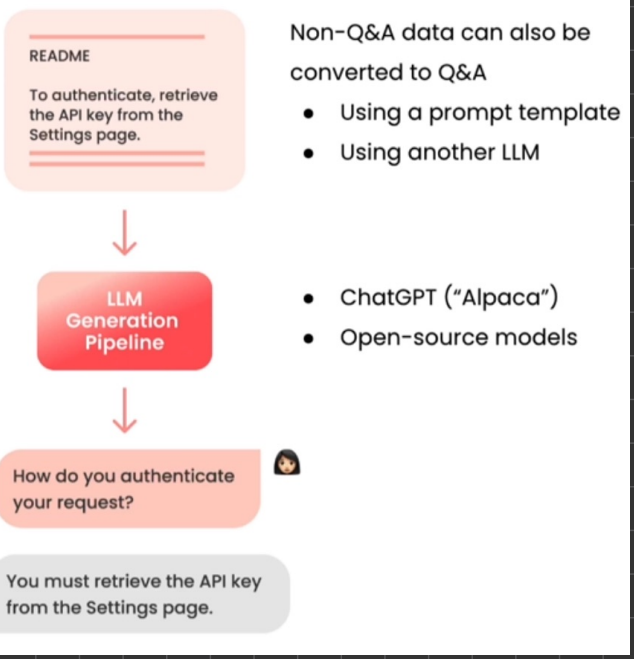
INSTRUCTION FINETUNING.

What is instruction finetuning?



- AKA "instruction-tuned" or "instruction-following" LLMs
- Teaches model to behave more like a chatbot
- Better user interface for model interaction
  - Turned GPT-3 into ChatGPT
  - Increase AI adoption, from thousands of researchers to millions of people

LLM Data Generation



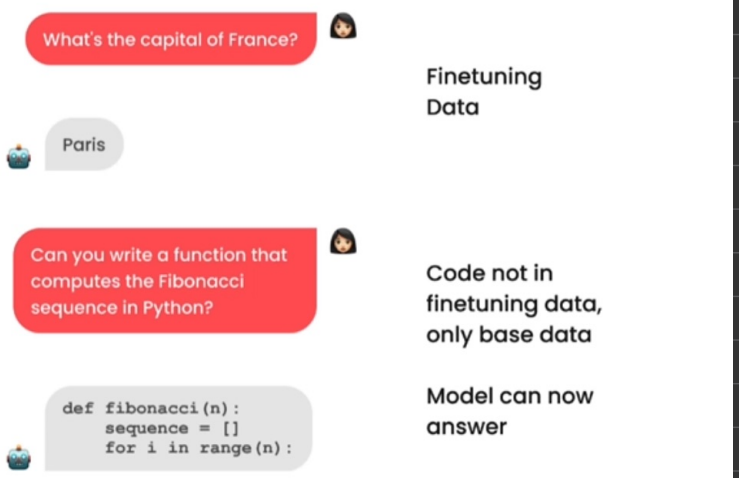
Non-Q&A data can also be converted to Q&A

- Using a prompt template
- Using another LLM

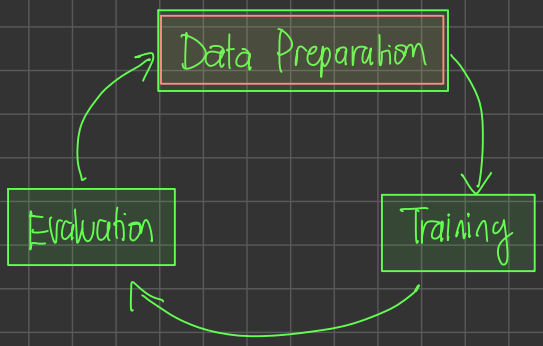
- ChatGPT ("Alpaca")
- Open-source models

Instruction Finetuning Generalization

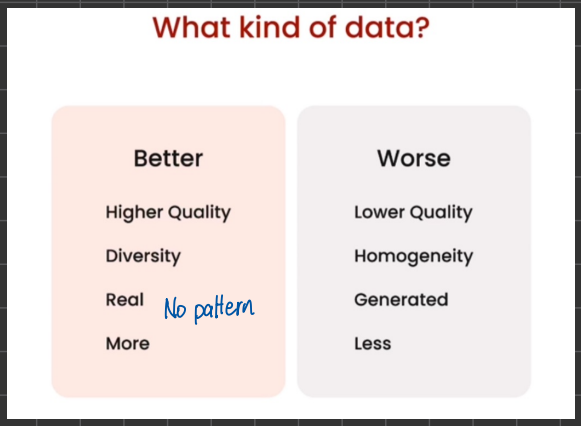
- Can access model's pre-existing knowledge
- Generalize following instructions to other data, not in finetuning dataset



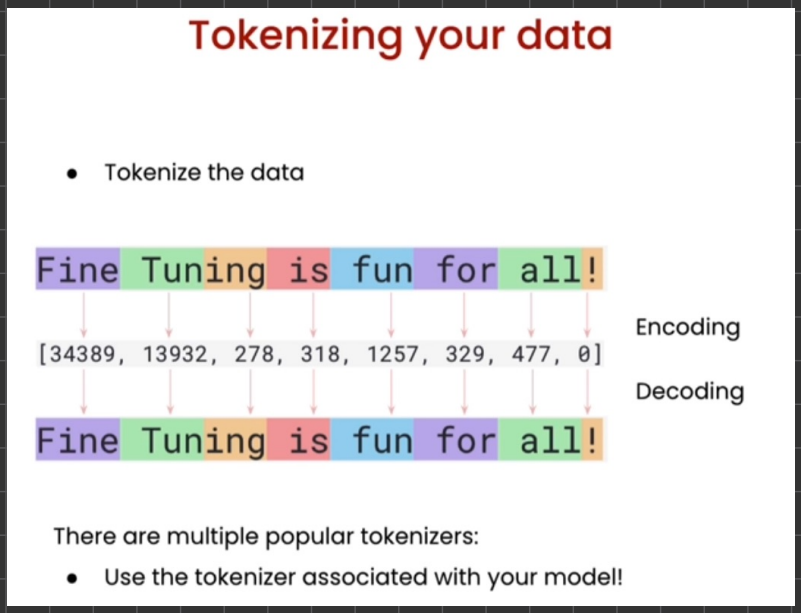
OVERVIEW



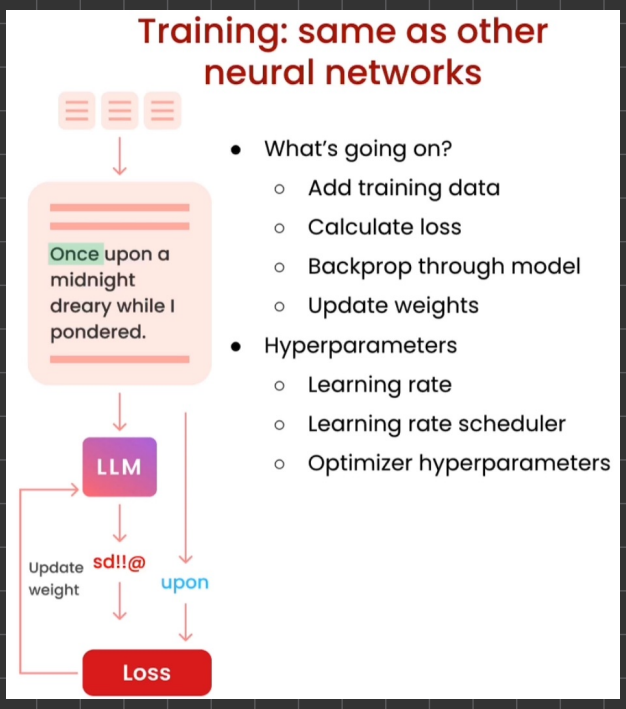
DATA PREPARATION



1. Collect instruction-response pairs
2. Concatenate pairs ⇝ add prompt template
3. Tokenize: Pad, Truncate ⇝ size of the model
4. Split into train/set



TRAINING PROCESS



EVALUATION

Evaluating generative models is notoriously difficult

Human Evaluation

Test Suites

Elo Rankings

- Human expert evaluation is most reliable
- Good test data is crucial
  - High-quality
  - Accurate
  - Generalized
  - Not seen in training data
- Elo comparisons also popular

LLM Benchmarks: Suite of Evaluation Methods

Common LLM benchmarks:

- ARC is a set of grade-school questions.
- HellaSwag is a test of common sense.
- MMLU is a multitask metric covering elementary math, US history, computer science, law, and more.
- TruthfulQA measures a model's propensity to reproduce falsehoods commonly found online.

Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
LLama-2	67.3	67.3	87.3	69.8	44.9
FreeWilly2	71.4	71.1	86.4	68.2	59.4
FreeWilly1	68.7	68.2	85.9	64.8	55.8

Error Analysis

- Understand base model behavior before finetuning
- Categorize errors: iterate on data to fix these problems in data space

Category	Example with Problem	Example Fixed
Misspelling	"Your kidney is healthy, but your lever is sick. Go get your lever checked."	"Your kidney is healthy, but your liver is sick."
Too long	"Diabetes is less likely when you eat a healthy diet, because eating a healthy diet makes diabetes less likely, making..."	"Diabetes is less likely when you eat a healthy diet."
Repetitive	"Medical LLMs can save healthcare workers time and money and time and money and time and money."	"Medical LLMs can save healthcare workers time and money."