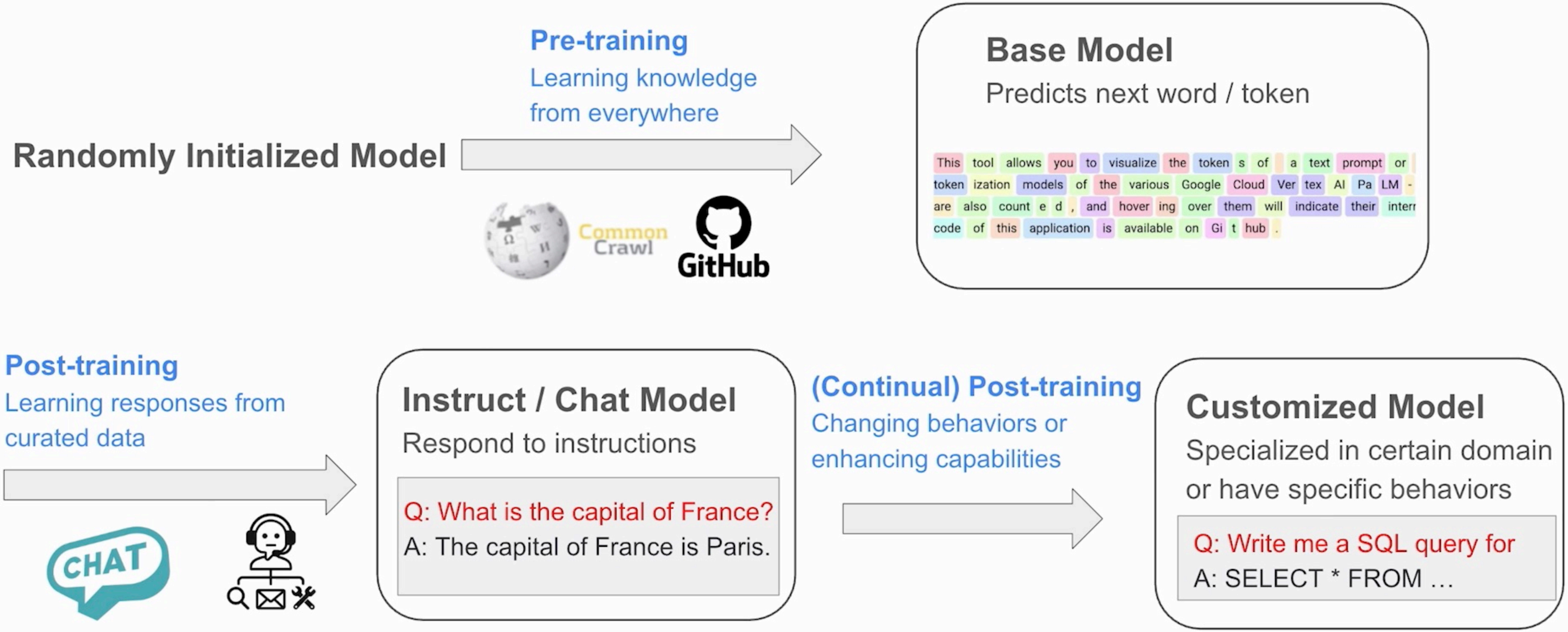
What is Post-training?



Pre-training
Learning knowledge from everywhere

Randomly Initialized Model

Base Model
Predicts next word / token

This tool allows you to visualize the token s of a text prompt or
token ization models of the various Google Cloud Ver tex AI Pa LM -
are also count e d , and hover ing over them will indicate their interr
code of this application is available on Gi t hub .

Post-training
Learning responses from curated data

Instruct / Chat Model
Respond to instructions

Q: What is the capital of France?
A: The capital of France is Paris.

(Continual) Post-training
Changing behaviors or enhancing capabilities

Customized Model
Specialized in certain domain or have specific behaviors

Q: Write me a SQL query for
A: SELECT * FROM …

Methods Used During LLM Training

Pre-Training

(Unsupervised Learning)

Unlabeled Text Corpus

Common Crawl

GitHub

>>2T tokens ⇩

"I like cats"

$$\min_{\pi} -\log \pi(\mathrm{I}) - \log \pi(\mathrm{like} \mid \mathrm{I})$$
$$- \log \pi(\mathrm{cats} \mid \mathrm{I\ like})$$

Post-training Method 1: Supervised Fine-tuning (SFT)

(Supervised / Imitation Learning)

Labeled Prompt-Response Pairs

Prompt: Explain LLM to me

Response: LLM is …

~1K-1B tokens ⇩

$$\min_{\pi} -\log \pi(\mathrm{Response} \mid \mathrm{Prompt})$$

Methods Used During LLM Training

Post-training Method 2: Direct Preference Optimization (DPO)

Prompt + Good and Bad Responses

Prompt: Explain LLM to me
Good Response: LLM is …
Bad Response: Sorry …

~1K-1B tokens ⇩

$$\min_{\pi} - \log \sigma \left(\beta \left(\log \frac{\pi(\text{Good R} \mid \text{Prompt})}{\pi_{\text{ref}}(\text{Good R} \mid \text{Prompt})} - \log \frac{\pi(\text{Bad R} \mid \text{Prompt})}{\pi_{\text{ref}}(\text{Bad R} \mid \text{Prompt})} \right) \right)$$

Post-training Method 3: Online Reinforcement Learning

Prompt + Reward Function

Prompt: Explain LLM to me
Response: LLM is …
Reward: 1.9

~1K-10M prompts ⇩

$$\max_{\pi} \text{Reward}(\text{Prompt}, \text{Response}(\pi))$$

Post-training Requires Getting 3 Elements Right

Data & algorithm co-design

- SFT
- DPO
- Reinforce / RLOO
- GRPO
- PPO
- …

Reliable and efficient library

- Huggingface TRL
- OpenRLHF
- veRL
- Nemo RL

Appropriate evaluation suite

(An Incomplete List of) Popular LLM Evals

Human Preferences for chat	**Chatbot Arena**
LLM as a judge for chat	Alpaca Eval MT Bench **Arena Hard V1 / V2**
Static Benchmarks for Instruct LLM	**LivecodeBench** **AIME 2024 / 2025** GPQA MMLU Pro IFEval
Function Calling & Agent	BFCL V2 / V3 NexusBench V1 / V2 **TauBench** **ToolSandbox**

It's easy to improve any one of the benchmarks.

It's much harder to improve **without degrading other domains**.

Do you really need post-training?

Use Cases	Methods	Characteristics
Follow a few instructions (do not discuss XXX)	Prompting	Simple yet brittle: models may not always follow all instructions
Query real-time database or knowledgebase	Retrieval- Augmented Generation (RAG) or Search	Adapt to rapidly-changing knowledgebase
Create a medical LLM / Cybersecurity LLM	Continual Pre-training + Post-training	Inject large-scale domain knowledge (>1B tokens) not seen during pre-training
Follow 20+ instructions tightly; Improve targeted capabilities ("Create a strong SQL / function calling / reasoning model")	Post-training	Reliably change model behavior & improve targeted capabilities; May degrade other capabilities if not done right