ChatGPT

LLM Training Phases

Large language models (LLMs) are typically trained in three sequential phases – **pre-training**, **supervised fine-tuning (SFT)**, and **reinforcement learning from human feedback (RLHF)** – each with different data, objectives, and compute requirements. In pre-training, a transformer-based model (e.g. GPT, LLaMA) learns general language patterns by predicting the next token on massive unlabeled corpora (CommonCrawl, books, Wikipedia, code, etc.) [1] [2]. This stage trains **all parameters from scratch** and produces a *foundation model* with broad knowledge. The model is then fine-tuned on task-specific or instruction-following data in a supervised manner (SFT), adjusting weights on much smaller labeled datasets (e.g. human-written question-answer pairs, instruction–output examples) to create an *instruction-tuned* model [3]. Finally, RLHF uses human feedback to further refine the model: a **reward model** is trained on pairwise human rankings of outputs, and the LLM's policy is updated (e.g. via PPO) to maximize this reward, yielding an *aligned* model that better matches user preferences and safety constraints [4] [5].

- **Pre-Training:** Uses massive, unlabelled text corpora (filtered Common Crawl, books, etc.) and self-supervised objectives (next-token prediction) [1] [2]. Models learn general language structure; **all parameters** are trained from random initialization via gradient descent. The goal is broad fluency and knowledge acquisition (no specific task). State-of-the-art pre-training runs on extremely large compute clusters for **weeks or months**. For example, GPT-3 (175B) was trained on ~300 billion tokens [1] and required ~355 GPU-years (V100) of compute (~$4.6M) [6]. GPT-4 reportedly used ~13 trillion tokens [7] (on perhaps 1.8T parameters), costing over $100 million [8]. Meta's Llama 2 (70B) was pre-trained on ~2 trillion tokens [2]. Smaller open models (e.g. EleutherAI's models) use proportionally fewer tokens and GPU cost. The output is a **foundational model** (e.g. GPT, PaLM, LLaMA), which may be proprietary (OpenAI's GPT) or open (Meta's LLaMA, Mosaic's MPT).

- **Supervised Fine-Tuning (SFT):** Starts from the pre-trained weights and trains on curated *labelled* data. Common setups include "instruction tuning" on human-written Q&A pairs or dialog (StackExchange, user instructions, crowdsourced dialogues), or domain-specific corpora. Only a fraction of weights may be updated (though often all are), typically with smaller learning rates. This stage focuses on teaching the model to perform desired tasks (e.g. answer questions, follow instructions) rather than raw language modeling. It uses supervised objectives (cross-entropy on input–output pairs). Data scale is much smaller: often **tens of thousands to millions of examples**. For instance, Stanford's Alpaca fine-tuned LLaMA-7B on 52K instruction pairs (cost ≈$600 on 8×A100 GPUs for 3 hours) [3]. Large-scale instruct models (e.g. InstructGPT, ChatGPT's GPT-3.5) use a few hundred thousand high-quality examples (e.g. crowdworker demonstrations). Fine-tuning typically takes **hours to days** on modern GPUs and costs from hundreds to a few thousand dollars (depending on model size). The output is an *instruction-following model* (e.g. InstructGPT, LLaMA-2-Chat, Claude-instruct), which is often released for applications.

- **Reinforcement Learning from Human Feedback (RLHF):** Further refines the SFT model using human preference data. In practice, this means collecting a dataset of **comparative feedback**: for a variety of prompts, human raters rank or score multiple model outputs. A separate *reward model* is trained (usually by fine-tuning a copy of the LLM) to predict these preferences. Then the original

model is updated with a policy gradient (PPO) to maximize the reward model's output [4] [5]. Only a small amount of data is needed (e.g. 10^5–10^6 ranking comparisons). OpenAI's InstructGPT collected ~50K prompts with 4–9 responses each (yielding 0.3–1.8M labeled pairs) [5]. Anthropic's HH-RLHF used on the order of 170K–318K comparisons. RLHF training (the PPO step) typically takes days on GPUs (plus time for data labeling). The cost includes human labor for annotations (often the largest expense) plus compute (~similar to SFT). The result is an *aligned* model (e.g. ChatGPT-3.5/4, Claude) that is more helpful, truthful, and safe. Empirically, RLHF substantially improves user alignment over SFT alone [4] [5].

- **Industry Context:** Major AI labs follow this general pattern. OpenAI's GPT series were trained this way (GPT-3 foundation; InstructGPT/ChatGPT SFT; GPT-3.5/4 with RLHF) [1] [4]. Anthropic's Claude models use similar steps (sometimes called Constitutional AI) with large custom RLHF datasets. Meta's LLaMA models (open-source) were pre-trained on public data [2] and then instruction-tuned (LLaMA-2-Chat) using fine-tuning and alignment, though Meta has not detailed its RLHF process. Companies differ in scale (parameters and data) but share methods: e.g. Mistral and Falcon used pre-training then fine-tuning; DeepMind's Sparrow/Chinchilla applied RLHF. Open models like LLaMA or MPT allow the community to perform their own fine-tuning, while closed models (GPT, Claude) keep foundation weights proprietary.

The following table summarizes key aspects of each phase:

Aspect	**Pre-Training**	**Supervised Fine-Tuning**	**RLHF**
Data/Tools	Massive raw text corpora (CommonCrawl, books, Wikipedia, code); transformer architecture; self-supervised learning [1] [2]. Frameworks: PyTorch/ TPU, distributed training.	Curated labelled datasets: instruction–response pairs, dialogue transcripts, examples; supervised cross-entropy loss. Uses SFT frameworks (e.g. HuggingFace, DeepSpeed) on pretrained model.	Human preference data: prompts + ranked model outputs; train a reward model (fine-tune LM) and apply RL (PPO). Requires human labeling tools and RL libraries.
Parameters	All model weights trained from scratch (starting random). Large parameter counts (10^7–10^12+). Eg. GPT-3 (175B), GPT-4 (~1.8T); LLaMA-2 (7B–70B) [7] [2].	Fine-tuning starts from pretrained weights; typically update most or all weights (low LR). Parameter count same as base model.	Policy starts from SFT model; weights adjusted via policy gradient (often only policy heads) to maximize reward.
Scope/Goal	Learn general language modeling: syntax, semantics, world knowledge. Broad, domain-agnostic.	Teach model to follow human instructions or perform tasks (summarization, Q&A, coding, etc.). Introduce task structure (dialogue style, formats).	Align model with human values/ preferences: make outputs helpful, safe, truthful. Optimize subjective criteria not captured by SFT.

Aspect	**Pre-Training**	**Supervised Fine-Tuning**	**RLHF**
Tasks/ Performance	Next-word/token prediction. The result is a **foundation model** (general-purpose).	Instruction following, task performance. E.g. answer queries, follow step-by-step prompts. Creates an **instruction-tuned model** (e.g. "InstructGPT", "ChatGPT", "LLaMA-2-Chat").	Preference fulfillment. Improves on metrics of quality/safety beyond likelihood (e.g. human preference rate, toxicity, truthfulness). Produces an **aligned model** (e.g. ChatGPT-3.5/4, Claude).
Training Data Size	Very large: often hundreds of billions to trillions of tokens. *(Examples:)* GPT-3 used ~300B tokens [1]; GPT-4 ~13T [7]; LLaMA-2 ~2T [2]. Smaller foundation models use less.	Moderate: from 10^4 to 10^7 examples (millions of tokens). *Example:* Alpaca used 52K instruction pairs [3]. Commercial instr. models may use few hundred thousand examples.	Small: ~10^5–10^6 examples of (prompt, pair of outputs) comparisons. E.g. InstructGPT labeled ~50K prompts (~0.3–1.8M preference pairs) [5]; Anthropic ~170K–318K comparisons.
Compute Time	Very long: weeks to months on large GPU clusters. E.g. GPT-3 required ~355 GPU-years (V100) [6]. State-of-art training uses thousands of GPUs (GPT-4 reportedly ~3 months on 8K H100s [9]).	Shorter: hours to days. E.g. fine-tuning Alpaca-7B took ~3 hours on 8×A100 [3]. Large-scale SFT (like InstructGPT) may take days/weeks on clusters.	Similar to SFT: days of GPU time for PPO runs (often on millions of tokens). Plus human annotation time (weeks to collect feedback).
Compute Cost	Very high: ranges from ~$10^5 up to ~$10^8. *Examples:* MosaicML reports training a GPT-3-quality (30B) model for < $0.5M [10]. GPT-3 (175B) was ~$4.6M [6]. GPT-4 cost >$100M [8] [11]. Newer models (Gemini Ultra, etc.) may be ~$200M [11].	Moderate: from hundreds to tens of thousands of dollars. Small-scale SFT (e.g. <1B models) can be done for ~$100–1,000. Alpaca (7B) cost <$600 [3]. Large SFT on 100B+ models could be ~$10k–$100k depending on scale.	Additional annotation costs: (thousands of human-hours * pay rate) can be millions for big ops. Compute for fine-tuning is modest relative to pre-training (typically <$100k even for large models).
Resulting Model	Foundation model (e.g. GPT-n, BERT-family, LLaMA-base). Broadly knowledgeable but unaligned; used as base for all downstream tuning. Closed or open depending on developer.	Instruction-tuned model (e.g. InstructGPT/text-davinci, Claude-instruct, LLaMA-2-Chat). Capable of following tasks/ instructions. Often still requires safety alignment.	Fully aligned model for deployment (e.g. ChatGPT 3.5/4, Claude). Designed for user interaction. Examples: *closed* ChatGPT/GPT-4, *open* LLaMA-2-Chat.

Illustration: Reinforcement Learning from Human Feedback (RLHF) aligns a supervised model to human preferences [4] [5].

In summary, pre-training builds a **foundation model** on vast unlabeled text (requiring billions of tokens and huge compute), SFT specializes this model to instructions or tasks (using moderate-scale labeled data), and RLHF aligns it to human feedback (using preference data). OpenAI, Anthropic, Meta, and others all follow these stages: e.g. OpenAI's workflow produced GPT (pretrained), then InstructGPT/ChatGPT (SFT+RLHF) [1] [4]. Anthropic's Claude uses constitutional RLHF on top of a pretrained model [5]. Meta's LLaMA (open) was pretrained on public data [2] and fine-tuned into LLaMA-2-Chat for instructions.

Industry Data: As a rough scale, training a medium-sized LLM (~10–30B params) at "GPT-3 quality" might cost on the order of $10^5–10^6 [10], while cutting-edge models (GPT-4, Gemini Ultra) cost on the order of $10^8 [11]. Fine-tuning costs are orders of magnitude lower (e.g. hundreds of dollars for smaller models [3]). RLHF adds human-labeling overhead but further improves model alignment [4] [5]. These training phases yield the progression from a **foundation model** to an **instruction-following** model to a fully **aligned assistant** – powering today's chatbots and AI services.

Sources: Industry reports and blog posts provide these estimates (OpenAI GPT-3 paper, Meta model cards, MosaicML blog, OpenAI alignment blog, etc.) [1] [4] [5] [11] [2], corroborating the token counts, costs, and procedures outlined above.

---

[1] [6] OpenAI's GPT-3 Language Model: A Technical Overview
https://lambda.ai/blog/demystifying-gpt-3

[2] meta-llama/Llama-2-70b · Hugging Face
https://huggingface.co/meta-llama/Llama-2-70b

[3] Stanford CRFM
https://crfm.stanford.edu/2023/03/13/alpaca.html

[4] Aligning language models to follow instructions | OpenAI
https://openai.com/index/instruction-following/

[5] RLHF: Reinforcement Learning from Human Feedback
https://huyenchip.com/2023/05/02/rlhf.html

[7] [8] GPT-4 - Wikipedia
https://en.wikipedia.org/wiki/GPT-4

[9] [11] What is the cost of training large language models?
https://www.cudocompute.com/blog/what-is-the-cost-of-training-large-language-models

[10] Mosaic LLMs: GPT-3 quality for
https://www.databricks.com/blog/gpt-3-quality-for-500k