



The evolving role of high quality-data in AI


Vahan Petrosyan
CEO & Co-founder


Leo Lindén
PMM




| databricks




 Round 1 | Prompt
What happens when I eat grape seeds?


 Round 1 | Completion
The black seeds in regular grape are matu that would sprout into a grape plant if you the soil. They are safe to eat and will pass body, as they are insoluble fiber

 Round 2 | Prompt
Is there a maximum amount our body can


Comments (4)


Assigned AI Trainer

 Rita Troskin
Linguist

 Jeremy Astor
Linguist

Comments (4)

 Rita Troskin
Hey @jeremy I find this sentence to not be entirely correct grammatically, could you check?
09.04.2024, 08:15

 Jeremy Astor
Yup, you're right! I'll flag this as an error in our database.. 🙏
09.04.2024, 08:15

1. Intro

- 1.1 The three pillars of AI
- 1.2 Data as the scaling limit
- 1.2 Diminishing data pool

2. Data and LLM Training

- 2.1 Pre-training
- 2.2 Continued Pre-Training
- 2.3 Fine-Tuning
- 2.4 Preference Training
- 2.5 Synthetic Data
- 2.6 The modern training pipeline

3. The future of data in AI

- 3.1 Trends
- 3.2 Agentic Systems
- 3.3 Multimodal

S

Home × Item_0001

Prompt

What is love?

Submit

Completion

What is love? Baby, don't hurt me
Baby, don't hurt me no more
Oh, I don't know, what can I do?
What else can I say? It's up to you

Rating*

★★★★☆

Select Completion 1

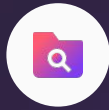
Prompt

The three pillars of AI development



Algorithm (AI Labs)

Algorithms are the core of LLMs, determining how models learn, process, and generate language. The transformer architecture is critical to their success.



Data (Data Foundries)

High-quality data is crucial for LLMs. Custom datasets from foundries like SuperAnnotate ensure models are trained on relevant and diverse data.



Compute (Hardware)

LLMs need massive computing power. GPUs and AI-specific hardware, like those from NVIDIA, make it possible to train these complex models efficiently.



The Data Bottleneck



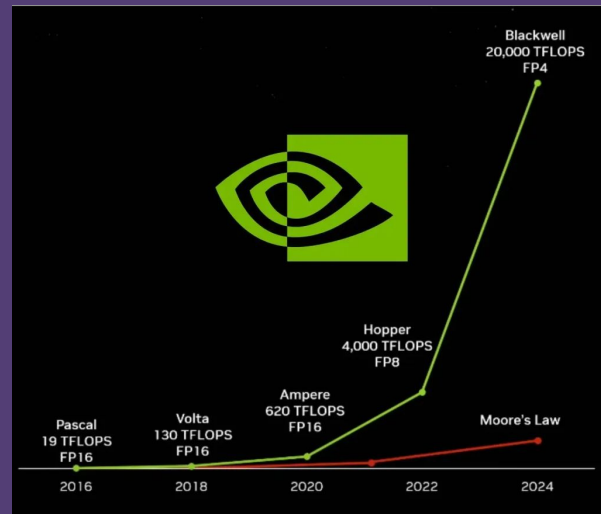
The Data Bottleneck

Compute used to be the limit.

For decades, Moore's Law predicted that computing power would double every two years. This enabled the rise of more sophisticated algorithms and models but also limited development speed.

Data access is now a limit.

Large Language Models and Foundation Models require vast amounts of high-quality, domain-specific data for training. However, the growth in high-quality data availability is not keeping pace with the scale of models.



The Data Bottleneck

Increasing difficulty

As models become more and more advanced, they need more and more complicated data to improve performance.

Highly Skilled

Creating datasets for machine learning models is moving from low-skill labor to, in some cases, requiring PhD level of understanding of the topics.

Why a \$14 Billion Startup Is Now Hiring PhD's to Train AI From Their Living Rooms



[Link](#)



Data and LLM Training



Building and deploying AI products require four things



Data

Existing datasets need to be annotated, or LLM specific datasets need to be created for fine-tuning



Training

A system to efficiently manage the training process and the necessary resources.



Deployment

A platform that can host and manage the model governance and run inference



Evaluation

Tooling to do in depth evaluations with domain experts or red teaming of models.



Pre-Training



Language understanding

- Before pre-training initial output is random and not meaningful.
- Training adjusts the model's parameters to produce the most likely token given an input sequence.
- The model is trained on a large amount of text data
- Trillions of **tokens**
- **Quality** still matters even in this phase:
[Textbooks are all you need](#)

Input sequence:

It is so hot outside; it would be great to cool down by eating an ...

Output before pre-training:

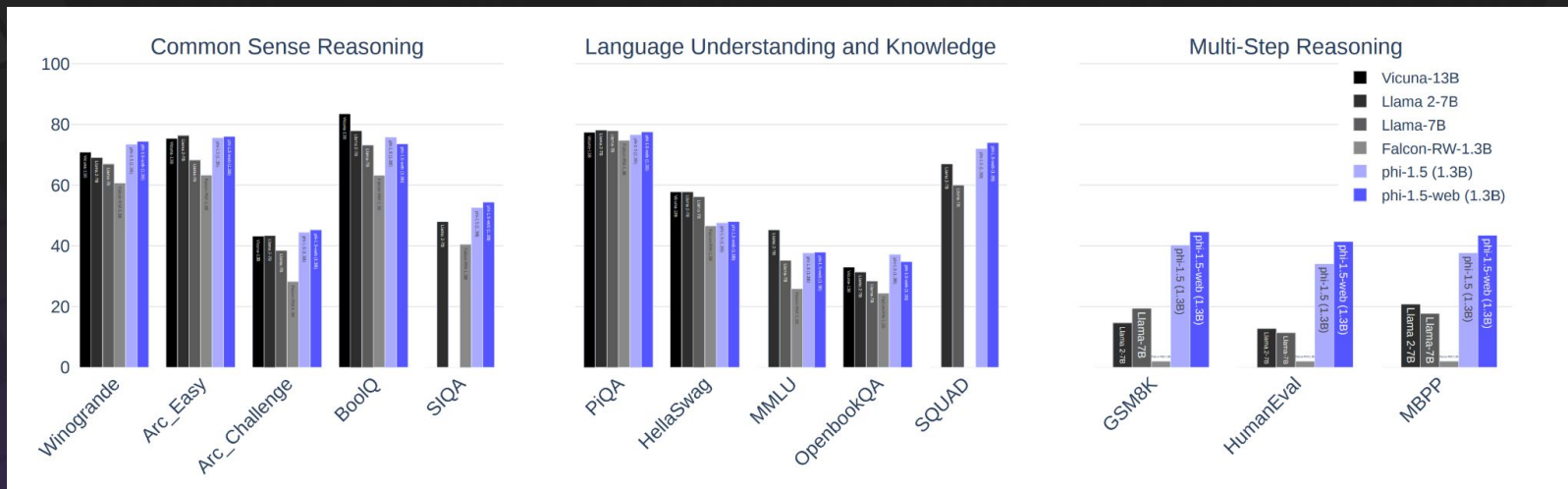
x!saöosjd

Output after pre-training:

ice cream

PRE-TRAINING

Textbooks are all you need



We build phi-1.5, a 1.3 billion parameter model trained on a dataset of 30 billion tokens, which achieves common sense reasoning benchmark results comparable to models ten times its size that **were trained on datasets more than ten times larger**



<https://arxiv.org/pdf/2309.05463>

Less available pre-training data

Licensing deals

Companies are making significant licensing deals to access premium content from news agencies, Reddit, books, and more. (Reddit alone is rumored to cost \$60M annually.)

Transcription

Transcription from video content is another method being explored to expand available text data for training.

Synthetic data

New techniques like generating synthetic tokens (e.g., Phi3 paper) are emerging as an alternative to real-world data limitations.



Continued Pre-Training



Domain Knowledge

- After pre-training the model can “understand” general language
- Continued pre-training on a specific domain can improve performance on that domain
- The model is trained on a large amount of domain specific text data
- Billions of tokens

Input Sequence:

The party seeking damages must demonstrate that there was a breach of ...

**Output before Continued Pre-training:
the party's feelings**

**Output after Continued Pre-training (Legal Domain):
contract and that the breach caused quantifiable harm**

Pre-Training vs Continued-Pre-Training

Pre-Training

Learning Everything

Learn language modelling and everything completely from scratch

Compute intensive

Requires trillions of tokens and thousands of GPUs over weeks or months

Unsupervised

Uses unsupervised methods

Continued-Pre-Training

Leveraging base

Basically transfer learning on a base model to acquire new domain knowledge

Less compute intensive

Require billions of tokens and hundreds of GPUs over days

Unsupervised

Uses unsupervised methods



Supervised Fine Tuning



Task Specialization

What is SFT: Model trained on labeled datasets for specific tasks.

How it Works: Use small, task-specific datasets for supervised learning.

Challenges: Availability of labeled data, overfitting.

Input Sequence: "What are the key elements required to establish negligence in a legal case?"

Output before Supervised Fine-Tuning:

- A. Duty of care, breach, causation, damages
- B. Intent, action, result
- C. Agreement, breach, remedy.

What are the common defenses against negligence claims?"

Output after Supervised Fine-Tuning:

The key elements required to establish negligence in a legal case are duty of care, breach of duty, causation, and damages. The plaintiff must prove that the defendant had a duty of ...

Fine-Tuning vs Continued-Pre-Training

Fine-Tuning

Tasks

Focuses on task-specific execution (e.g., summarization, tool use) not found in the data

Formatted Data

Relies on structured, curated datasets

Supervised

Uses Supervised learning methods

Continued-Pre-Training

Knowledge

Ideal for teaching domain-specific jargon and knowledge

Unstructured

Ideal for teaching domain-specific jargon and knowledge

Unsupervised

Uses unsupervised methods



FINE-TUNING

Data Quality



SFT (Mix)

SFT (Annotation)

Choice of vendor

Differences between annotation providers can have massive impact on model performance

Less is more

10s of thousands high quality better than millions low quality.

Reaches human level quickly

In training LLaMA2 the researchers noted that the model quickly reached the same level as many human annotators



Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, D., Batra, S., Bhargava, A., Bhosale, S., et al. (2023). **LLaMA 2: Open Foundation and Fine-Tuned Chat Models**. *arXiv:2307.09288*. Available at: <https://arxiv.org/abs/2307.09288>

Task Specialization

Full Fine-Tuning

- Updates all model weights based on task-specific data.
- Requires significant memory and compute resources.
- Risk of "catastrophic forgetting" where model loses prior knowledge.

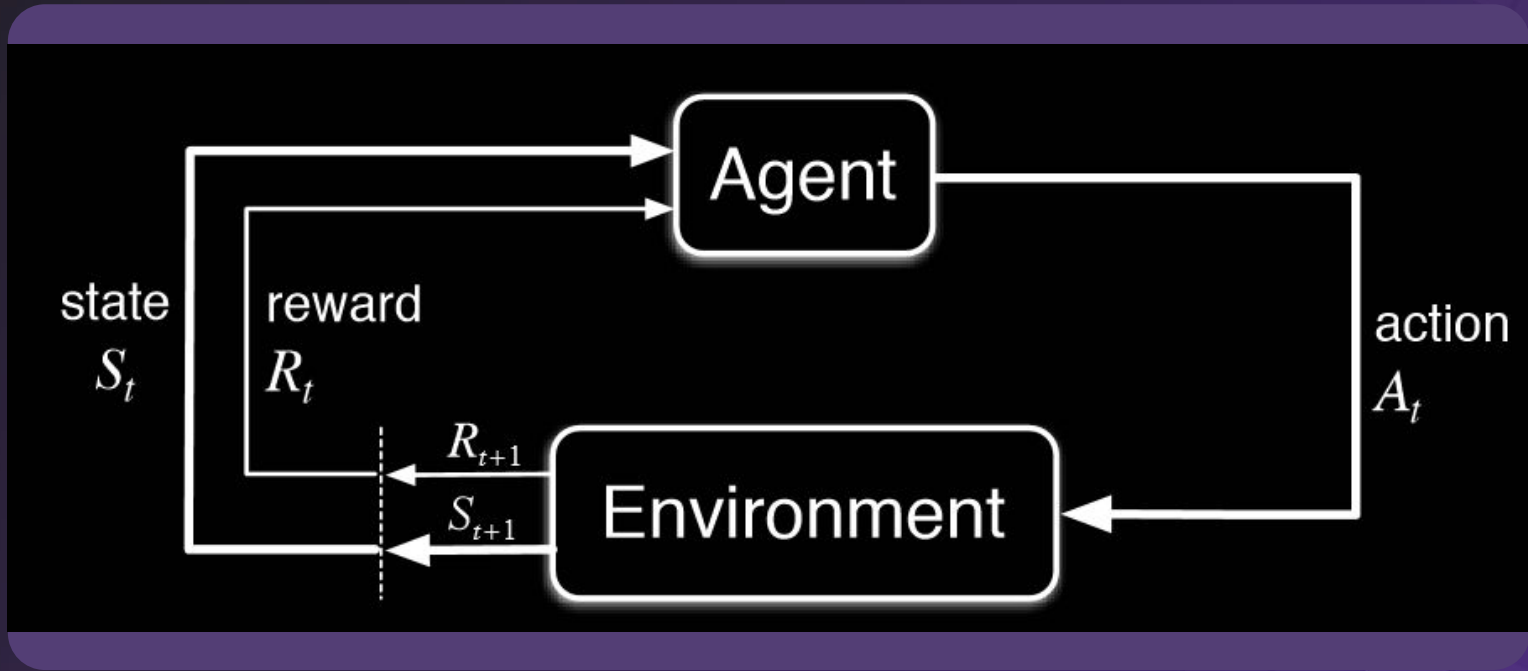
Parameter-Efficient Fine-Tuning (PEFT)

- Only updates a subset of parameters, freezing the rest.
- Lower memory requirements (e.g., LoRA reduces trainable parameters by up to 10,000x).
- Helps retain previous knowledge while adapting to new tasks. (Mix of data might be good too)

Reinforcement Learning



In a Nutshell



Human preference

Aligns a model with human preference

- The model generates 2 or more answers to each prompt
- A human or AI rater decides which answer is the best
- A reward model is trained on the preference data.
- Reward model used to defined a loss function for the main model

Prompt

What is Neurips, and give your answer in a few words

Generate

Completion 1

NeurIPS (Conference on Neural Information Processing Systems) is an annual research conference that focuses on interdisciplinary research in machine learning and artificial intelligence.

Rate the answer



Completion 2

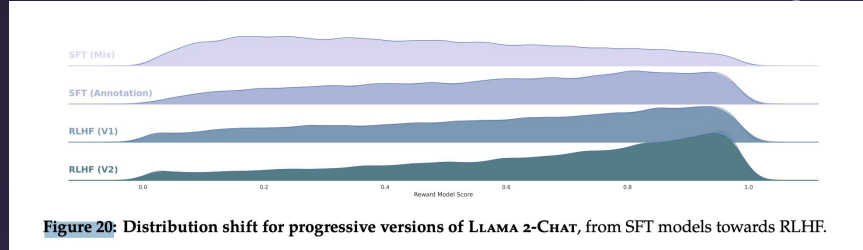
NeurIPS (Conference on Neural Information Processing Systems) is a leading annual conference in the field of machine learning and artificial intelligence.

Rate the answer



REINFORCEMENT LEARNING

Preference vs Fine-Tuning



Human Preference

Easier

Hard to write a mozart quality concerto but easy to say which out of two are the best

Scalable

Ranking of responses is much faster than creating one

Reinforcement Learning

Uses various reinforcement learning methods

Supervised Fine-Tuning

Tasks

Focuses on task-specific execution (e.g., summarization, tool use) not found in the data

Formatted Data

Relies on structured, curated datasets

Supervised

Uses Supervised learning methods



Traditional Method

Preference Dataset

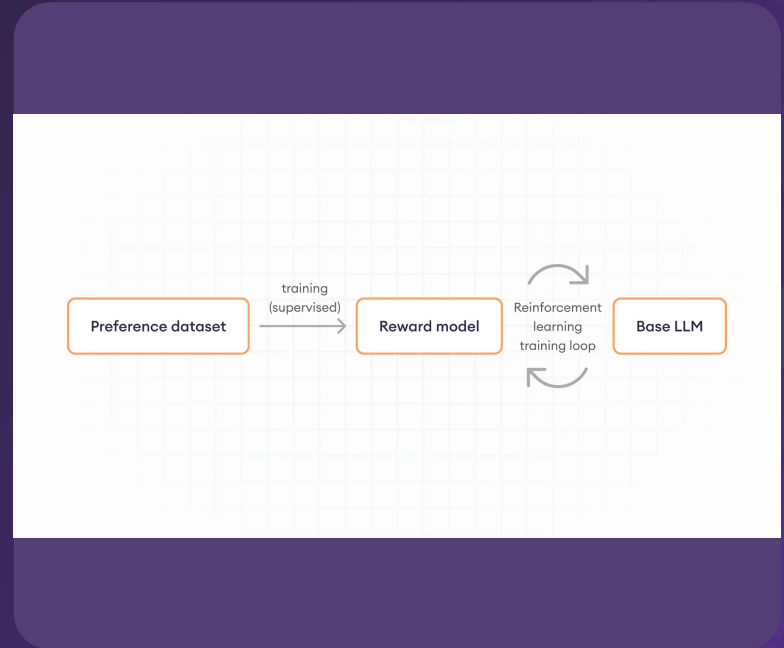
A dataset with human preference for different answers is collected

Reward Model

The dataset is used to train a reward model that rates responses

Alignment

This reward model is used to train the LLM



Data Collection

Preference Dataset

A dataset with human preference for different answers is collected.

Ranking

Ordering multiple responses from better to worse

Rating

Rating responses on a scale, or according to a separate evaluation matrix,

Rewriting

If none of the responses are good you might want to rewrite one of them



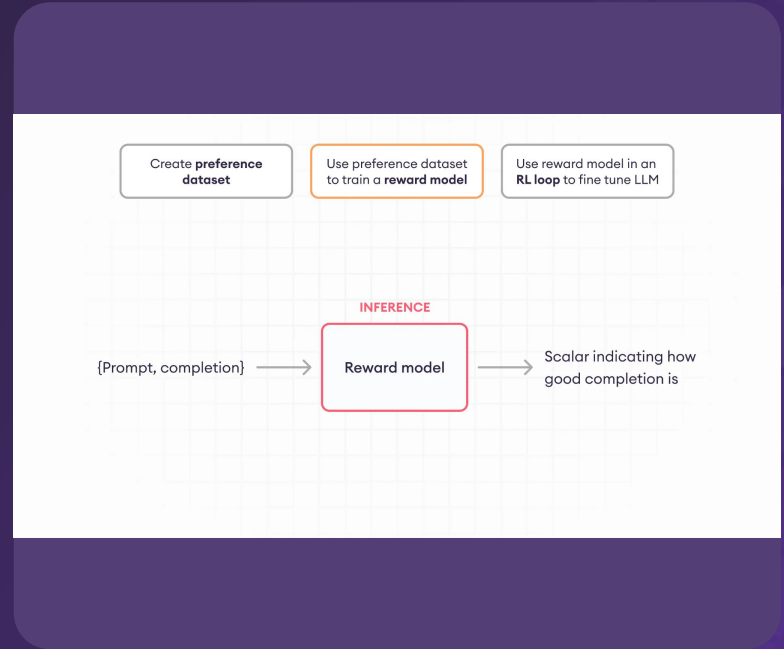
Human Preference → Reward Model

Reward Model

The dataset is used to train a reward model that rates responses

Model type

The base for this model is sometimes the same as the LLM but with the final layer replaced with a digit output instead of language



REINFORCEMENT LEARNING

Human Preference → Reward Model

Multiple Reward Models

By having multiple reward models focused on different tasks trade-offs between different goals can be avoided

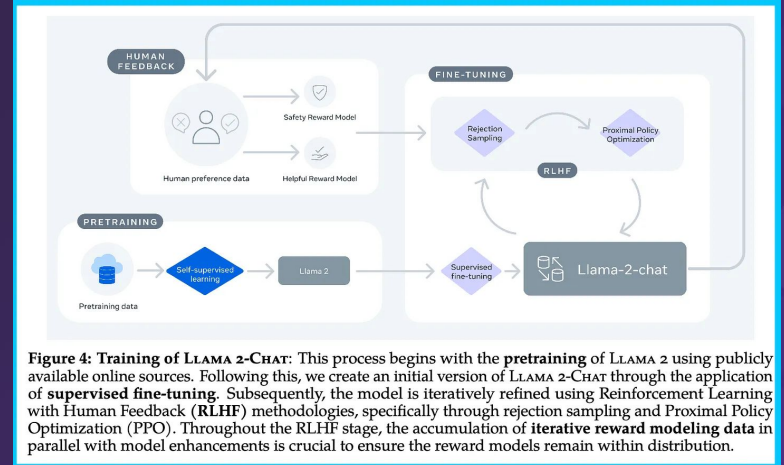


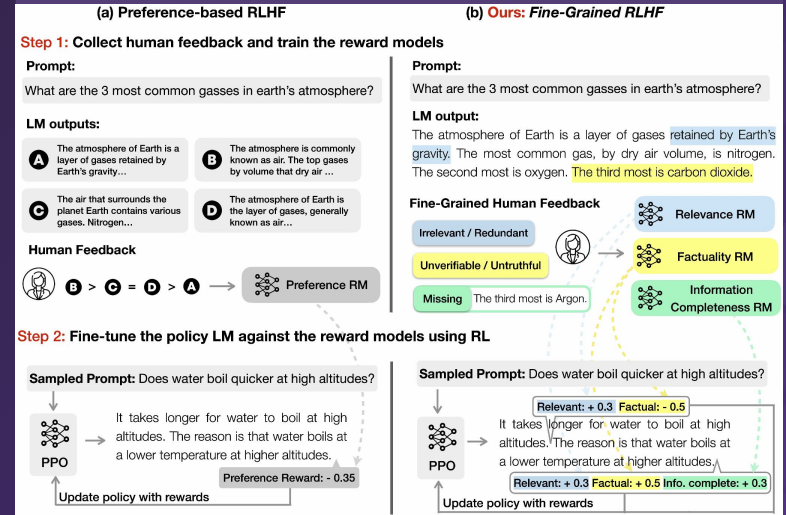
Figure 4: Training of LLAMA 2-CHAT: This process begins with the **pretraining** of LLAMA 2 using publicly available online sources. Following this, we create an initial version of LLAMA 2-CHAT through the application of **supervised fine-tuning**. Subsequently, the model is iteratively refined using Reinforcement Learning with Human Feedback (RLHF) methodologies, specifically through rejection sampling and Proximal Policy Optimization (PPO). Throughout the RLHF stage, the accumulation of **iterative reward modeling data** in parallel with model enhancements is crucial to ensure the reward models remain within distribution.

REINFORCEMENT LEARNING

Human Preference → Reward Model

Fine-Grained Reward Models

By having multiple reward models and detailed tagging on a token level, human annotations can be of higher quality, and subjectivity in ranking can be avoided



Reward model → Better LLM

Training

The final reward model is used to train the language model using PPO or similar algorithms.

Improvement

Generally shows a clear improvement over just SFT

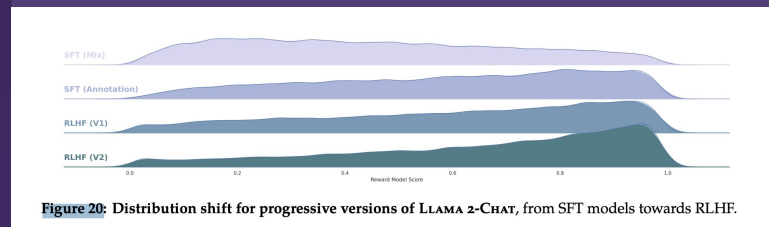
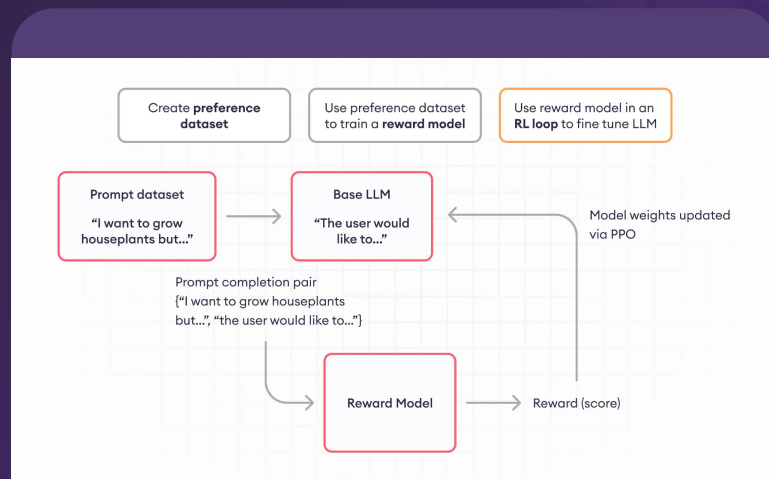
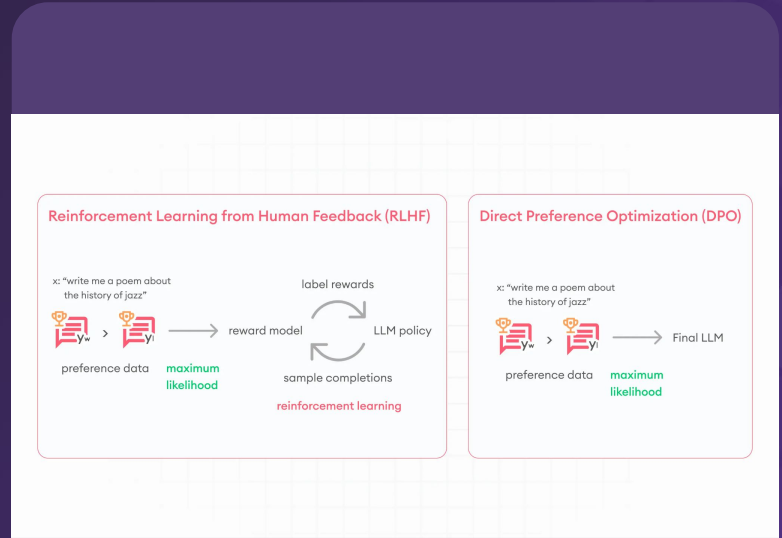


Figure 20: Distribution shift for progressive versions of LLAMA 2-CHAT, from SFT models towards RLHF.

DPO vs Reward Model(PPO)

Different Approaches

Direct Preference Optimization (DPO) is a training approach that integrates preference data directly into the learning process, eliminating the need for an intermediate reward model.



Synthetic Data





Quicker

Building Synthetic datasets is substantially faster than fully human but there are some traps



More competent models

As models get better and better they are producing better output than most human annotators



Model Collapse

There has been research indicating that training on synthetic data should lead to model collapse.



Works (Sometimes)

A lot of models now use synthetic data and it seems to work just fine.



Why it should not work

Model Collapse:

Recursively training AI models on data generated by earlier versions leads to loss of information and performance degradation.

Loss of Distribution Tails:

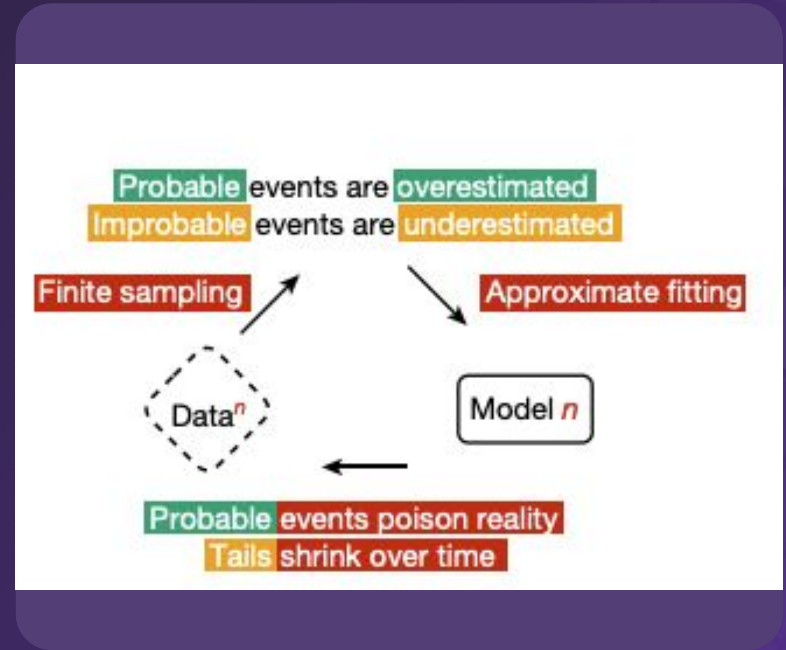
Over time, models forget rare events, causing the original data distribution to shrink and collapse.

Generative Models Affected:

When trained on model-generated data, large language models (LLMs), VAEs, and GMMs all exhibit this degenerative effect.

Importance of Human Data:

To prevent collapse, access to accurate, human-generated data is essential for maintaining model accuracy over generations.



Our evaluation suggests a ‘first mover advantage’ when it comes to training models such as LLMs. In our work, we demonstrate that training on samples from another generative model can induce a distribution shift, which—over time—causes model collapse.”

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755-759. <https://doi.org/10.1038/s41586-024-07566-y>

And when it does work

Incorrect assumptions

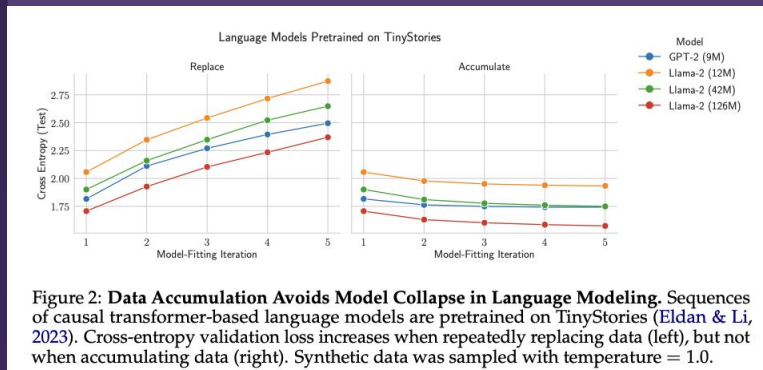
The nature paper assumes that the existing dataset is replaced with a new fully (or almost fully) synthetic dataset.

Accumulate

Accumulating data instead of replacing it avoids the model collapse as seen in the previous paper

Filter

Labs today usually also applies different methods of filtering the synthetic data to improve the results further.



Hybrid Synthetic

AI Feedback

An AI gives feedback to a human annotator and improvement suggestions that they may or may not use

Rewriting

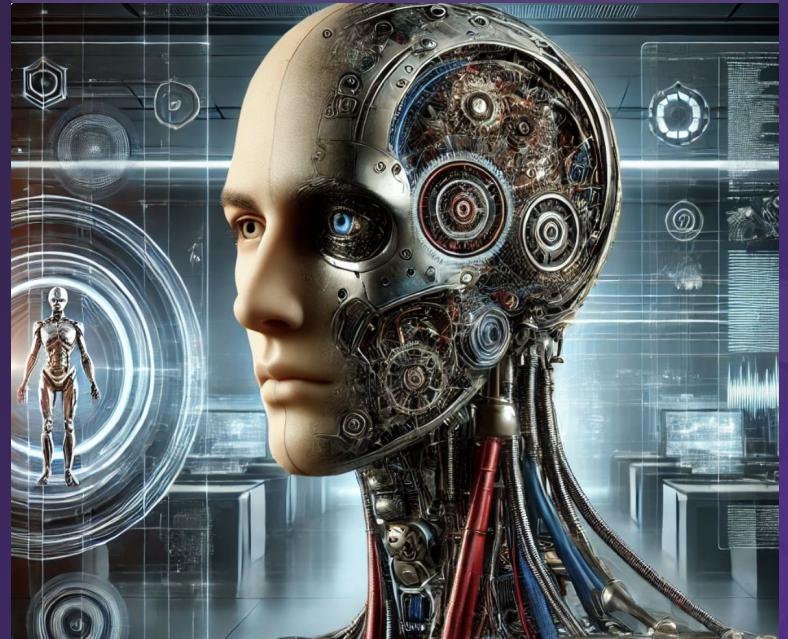
An AI writes the original response and a human creates a new one based on it.

AI QA

AI tags some prompt that seem to be of lower quality for review by an expert human

Selection

A model trained with human data is used to select the best of x synthetic data.



Human vs AI Judge

LLM as a judge

Using a prompted LLM to rate and rank responses instead of a human.

Mixed results

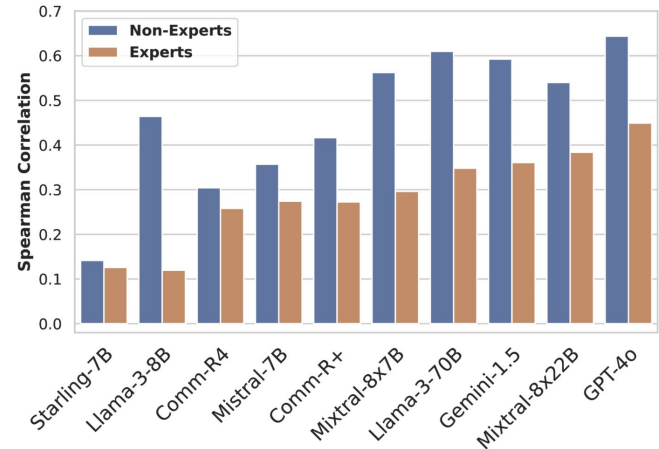
Various studies indicate various rates of alignment between human and ai judges. LLama 2 found that their reward model was better than any LLM Judge

Judge Bench

New benchmark developed by a set of researchers. Showed lower alignment with humans

Data Collection

Crowdsourcing vs expertly managed annotators can produce vastly different results



	Meta Helpful.	Meta Safety	Anthropic Helpful	Anthropic Harmless	OpenAI Summ.	Stanford SHP	Avg
SteamSHP-XL	52.8	43.8	66.8	34.2	54.7	75.7	55.3
Open Assistant	53.8	53.4	67.7	68.4	71.7	55.0	63.0
GPT4	58.6	58.1	-	-	-	-	-
Safety RM	56.2	64.5	55.4	74.7	71.7	65.2	64.3
Helpfulness RM	63.2	62.8	72.0	71.0	75.5	80.0	70.6

Table 7: Reward model results. Performance of our final helpfulness and safety reward models on a diverse set of human preference benchmarks. Note that our model is fine-tuned on our collected data, as opposed to the other baselines that we report.



How top labs do training today



Commonalities in new models

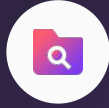


RLHF > SFT

Growing focus on RLHF over SFT

RLHF is easier to scale and more cost-effective.

SFT addresses gaps in specialized tasks.



Synthetic(ish) data

AI-generated data surpasses human performance in many tasks.

Rejection Sampling uses human preference data trained reward model for Synthetic Data filtration



Data Centric

Data-driven approach to model improvement

Data cleaning, filtration, improvement etc.



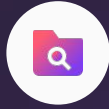
And its implications for “data foundries”



RLHF > SFT

With synthetic data gaining dominance, data providers are shifting focus from SFT data to more valuable preference data and evaluation.

Decreased demand for general human written training data can hit a lot of creative workers



Specialized Data

Focus on highly demanding tasks (e.g., scientific, legal, or medical data)..

Tasks like scientific reports are harder to source and execute well, leading to increased operational costs for data providers and customers



Platform

In some cases companies might have all competence in house and in this case platforms that enable them to easily collect data themselves can be useful



So LLM training is pretty much a solved problem right?

Not so fast ...

Products built with GenAI are more than just LLMs. We are getting multimodal models (LFMs) as well as different types of agents that receives data input, uses tools and more. This makes data for training and evaluation much more complex

There is still a frontier ahead

Two focuses

Model Builders

Well funded scale-ups and enterprises that can keep up with dataset sizes and spending on preference and fine-tuning data.

Focused on building the best foundation models or differentiate with focus on narrow field.

Model adapters

Everyone from small startups to large enterprises leveraging LLMs to build new products and/or improving operations.

Focused on building high performing systems incorporating LLMs and data using agents, fine-tuning, RAG and more.



Three areas where data annotation is challenging

Multimodal

Beyond just text multimodal models can require any type of data input.

Current annotation platforms mainly support text based SFT/RLHF.

Agents

Systems with multiple LLM steps and usage of tools or databases

Creating eval or training datasets require visibility into all behind the scenes reasoning and tool usage steps

Advanced Data

LMFs increasingly require advanced and difficult datasets to keep improving

Building these datasets require domain experts and support for more advanced qa workflows.

Agents



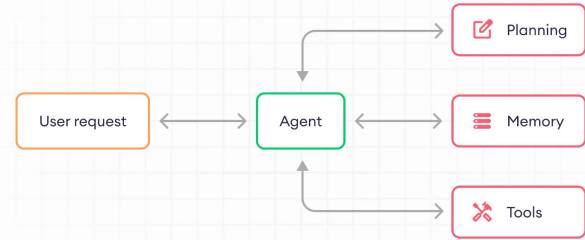
Agents

Definition:

Autonomous systems powered by large language models, designed to perform tasks with minimal human intervention.

UX:

Does not include all the different actions, unlike LLMs where you see the input and the output. Masking a lot of the work that happens and needs to be review



Agents

Sequential Complexity:

Even the simplest agents handle more than just input and output data. For example, a task like booking a flight involves multiple actions, choices, API/tool uses, and user inputs.

Accuracy Loss in Chains:

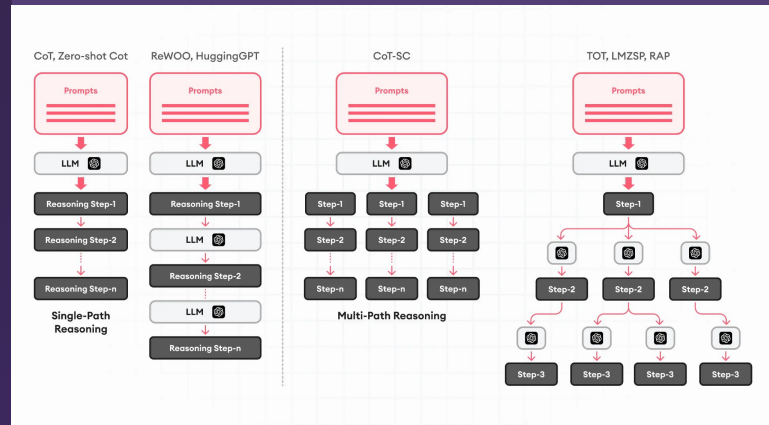
Each step in these sequential tasks introduces a chance for accuracy loss. A model trained on short text dialogues and tool use may struggle with longer, more complex workflows.

Multiple Failure Points:

The involvement of APIs, external tools, and human inputs creates multiple potential points of failure within the process.

Challenges in Review:

Reviewing an agent's overall performance gives insight into success or failure but often lacks the granular detail needed to identify where improvements should be made.



Example – RAG

Query Rewriting

Before searching the initial query often needs to be rewritten for clarity or specificity,

Retrieval

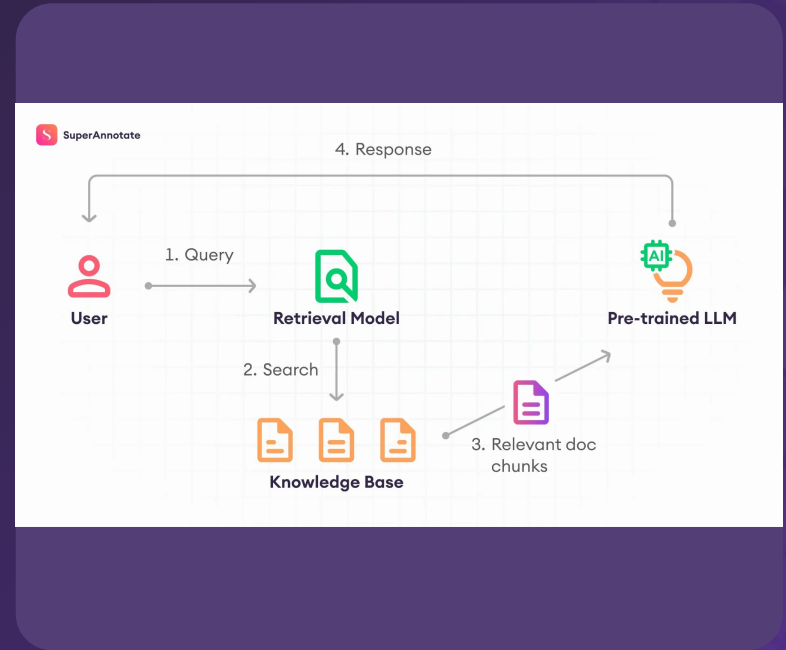
Relevant information must be retrieved. Poor retrieval can drastically affect the quality of the response.

Reranking

Retrieved documents are then reranked based on relevance. Missteps here can lead to incorrect or less optimal answers being prioritized.

Map-reduce response

In some cases initial answers might be produced for each retrieved document and then combined together



LMM (Multimodal)



Multimodal Models

One-to-one

Models that take one modality as input and another as output.

Many-to-one

A mix of modalities both in the input but one in the output

Many-to-many

A mix of modalities both in the input and the output

New algorithms

Both of these require different algorithms than standard LLMs

New training data

Different types of training data is needed as well.



Use-Cases

Content generation/editing

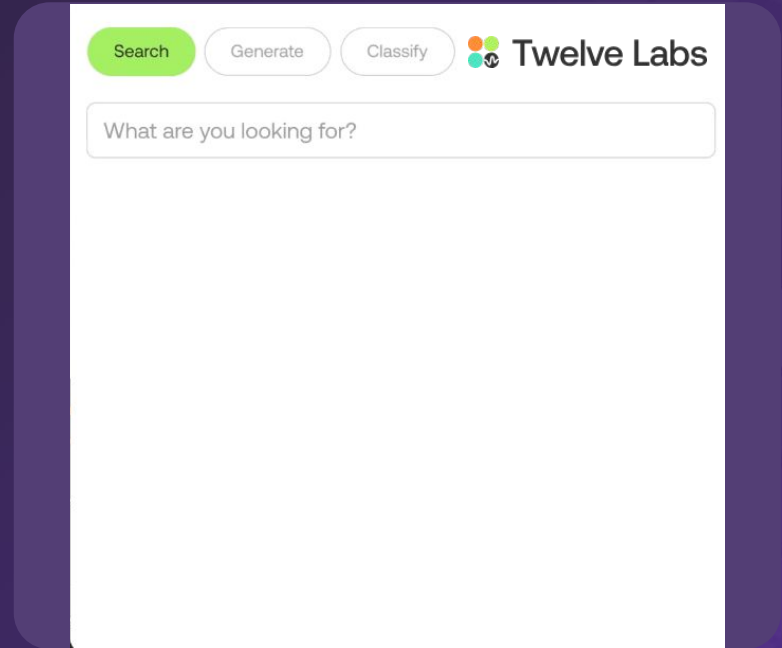
Models that generate/edit content such as images or video based on input of one or more modalities

Embedding

Models that encode different modalities to the same embedding space.

Processing

Models that analyze fully multimodal input for purposes such as conversations or analysis



Highly Specialized data



Three areas where data annotation is challenging

Multimodal

Beyond just text multimodal models can require any type of data input.

Current annotation platforms mainly support text based SFT/RLHF.

Agents

Systems with multiple LLM steps and usage of tools or databases

Creating eval or training datasets require visibility into all behind the scenes reasoning and tool usage steps

Advanced Data

LMFs increasingly require advanced and difficult datasets to keep improving

Building these datasets require domain experts and support for more advanced qa workflows.

Problems with building datasets today



Inflexible annotation tooling

Current tools for dataset building are built with tasks like simple SFT and RLHF in mind,



Managing annotation setups add overhead

Building and managing an own data and annotation pipelines takes time away from working on ml projects

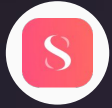


Annotators need to be more specialized

More and more specific data leads to lack of data trainers.

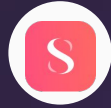


~~Problems~~ Solutions with data today



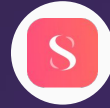
Multimodal tooling

Custom toolset to build and manage complex AI datasets, measure quality, etc



Data Orchestration

Operationalize complex data and annotation pipelines and ensure the data flows from pretraining to fine-tuning applications smoothly



Domain expertise

Manage a diverse network of expert so that enterprises can focus on building their models rather than tedious hiring plans

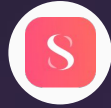


SuperAnnotate



Algorithm (AI Labs)

Algorithms are the core of LLMs, determining how models learn, process, and generate language. The transformer architecture is critical to their success.



Data (Data Foundries)

High-quality data is crucial for LLMs. Custom datasets from foundries like SuperAnnotate ensure models are trained on relevant and diverse data.



Compute (Hardware)

LLMs need massive computing power. GPUs and AI-specific hardware, like those from NVIDIA, make it possible to train these complex models efficiently.

The future is Multimodal, Its Specialized,
Its Complex, Its Customizable, Its 

