Introduction to LLMs for science of science research @ LMU 2025

Dirk Wulff & Zak Hussain

MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

UNI
BASEL

Goals

Familiarize you with the workings and applications of open-source LLMs and how to implement them using the Hugging Face ecosystem

Software stack

+

+

LLM4SciSci

Public

Unwatch 2 Fork 0 Star 0

main 1 Branch 0 Tags Go to file t Add file Code

Zak-Hussain adding markdown e8876dd · 13 minutes ago 18 Commits

SciSci_articles	add paper pdfs	46 minutes ago
.gitignore	Initial commit	4 days ago
1_pipelines.ipynb	add templates	4 days ago
2_clustering.ipynb	clear outputs	38 minutes ago
3_labeling_retrieval.ipynb	adding markdown	13 minutes ago
LICENSE.txt	add LICENSE.txt	4 days ago
README.md	clear outputs	38 minutes ago
science_of_science.csv	add code first draft	yesterday

README License

LLM4SciSci

About

Materials "LLMs for science of science research" training, LMU, 2025

Readme
View license
Activity
0 stars
2 watching
0 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 2

Today

09:15 AM - 09:45 AM: Welcome & Intro
10:00 AM - 11:00 AM: Talk: Intro to LLMs
11:00 AM - 11:15 AM: Break
11:15 AM - 11:45 PM: Talk: A gentle intro to Hugging Face and Python
11:45 AM - 12:00 PM: Setup Colab
12:00 PM - 12:30 PM: Exercise: Running pipelines
12:30 PM - 01:00 PM: Discussion: Find applications in small groups
01:00 PM - 02:00 PM: Lunch
02:00 PM - 03:00 PM: Talk: Intro to transformers & embeddings
03:00 PM - 03:30 PM: Exercise: Clustering
03:30 PM - 03:45 PM: Break
03:45 PM - 04:30 PM: Talk: Intro to text generation
04:30 PM - 05:00 PM: Exercise: Labeling & retrieval
05:00 PM - 06:00 PM: Open discussion

1. Where are you from, what is your background, and what do you do?
2. What motivates you to learn more about LLMs?
3. R or Python or __ ?
4. How much experiences do you have with programming and machine learning?

Intro LLMs

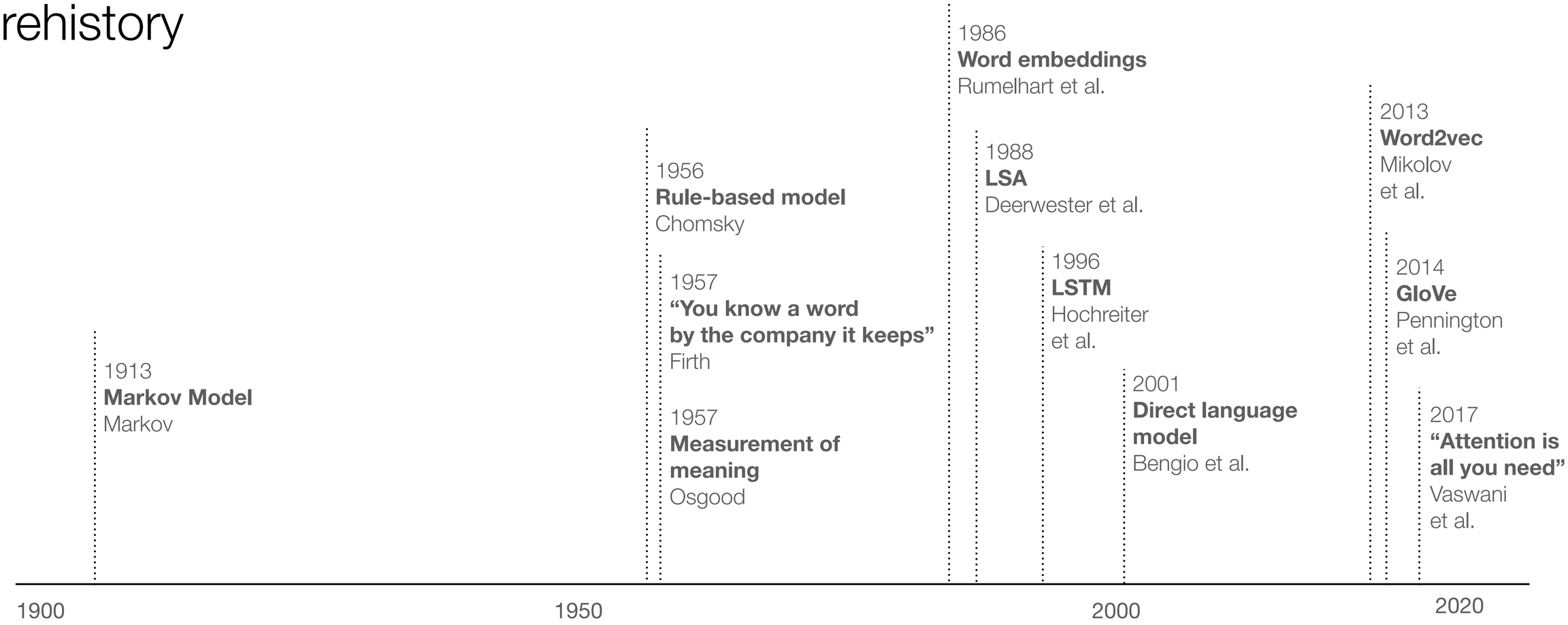
Dirk Wulff & Zak Hussain

MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

UNI
BASEL

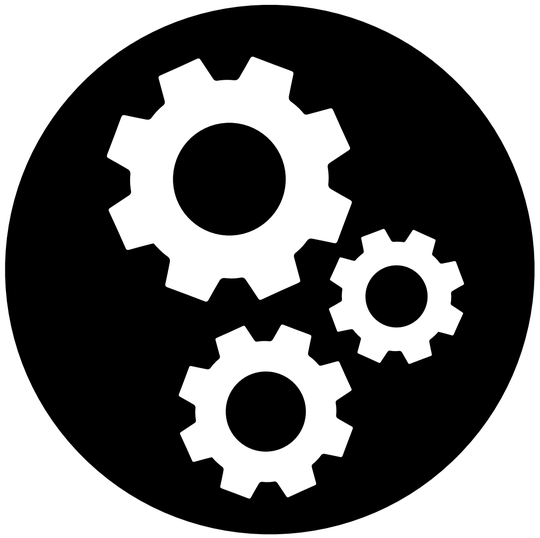
Language models

Prehistory



LLMs as mechanisms

LLM



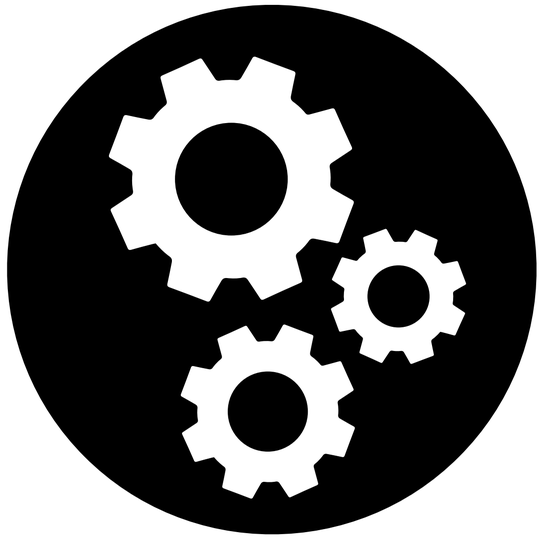
LLMs as mechanisms

Input

"Once upon a time, the"

→

LLM



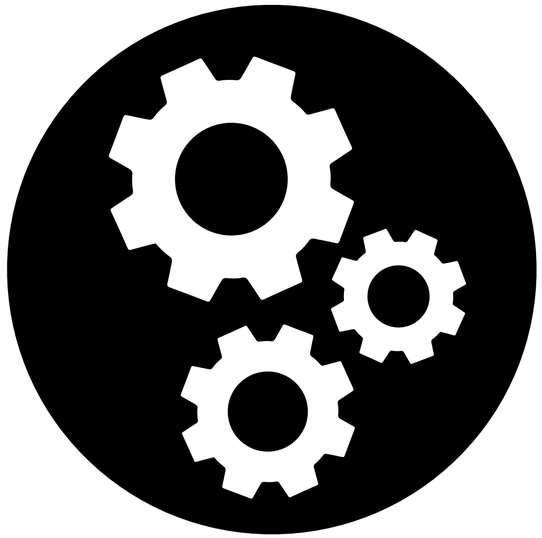
LLMs as mechanisms

Input

"Once upon a time, the"

→

LLM



→

Output

"dog"

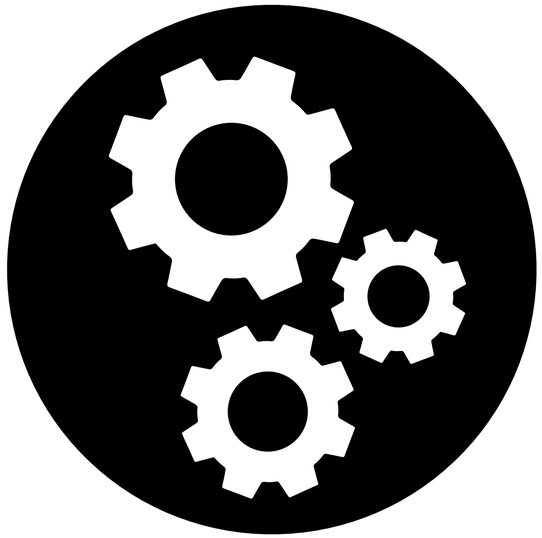
LLMs as mechanisms

Input

"Once upon a time, the"

→

LLM



→

Output

"2"

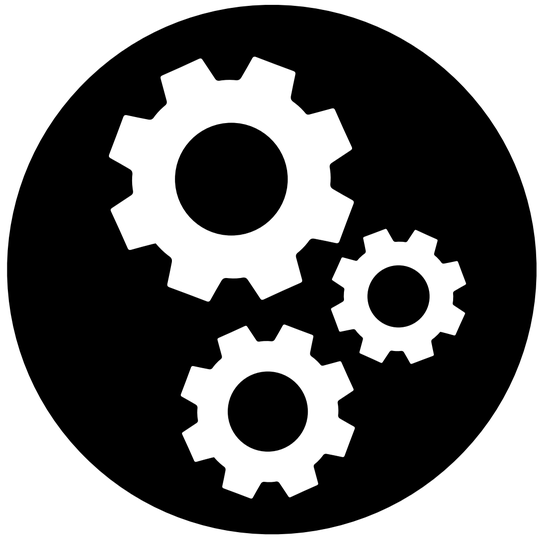
LLMs as mechanisms

Input

"Once upon a time, the"

→

LLM



→

Output

"!"

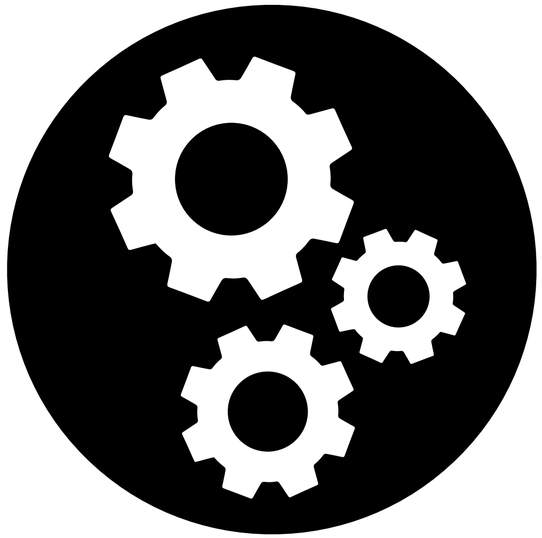
LLMs as mechanisms

Input

"Once upon a time, the"

→

LLM



→

Output

"dog dreamt it was a cat."

Once upon a time, the

Once upon a time, the small village of Elmswood was nestled in a lush valley surrounded by towering mountains. The villagers lived peacefully, their days marked by the rhythms of nature and the changing seasons. However, everything changed when a mysterious old man arrived, carrying with him a locked chest that was said to contain a secret capable of altering the course of history. Intrigued by the stranger and his enigmatic treasure, the people of Elmswood soon found themselves on the brink of an adventure that would bind them together in ways they could never have imagined.

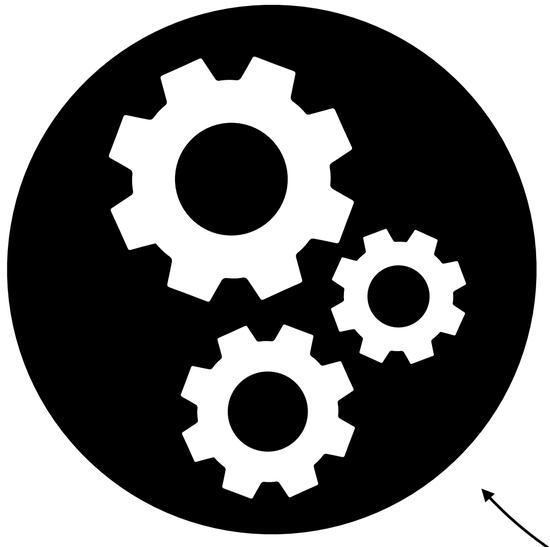
LLMs as trained mechanisms

Input

LLM

Output

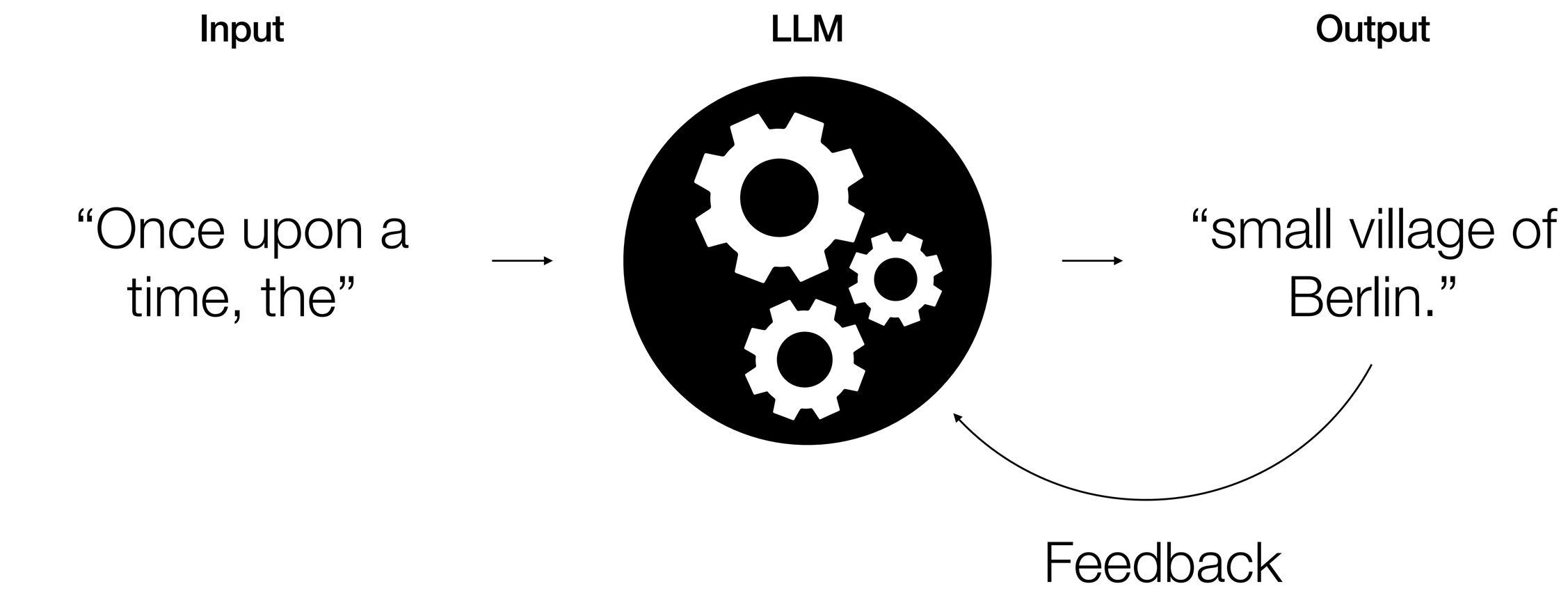
“Once upon a time, the”



“dog dreamt it was a cat.”

Feedback

LLMs as trained mechanisms



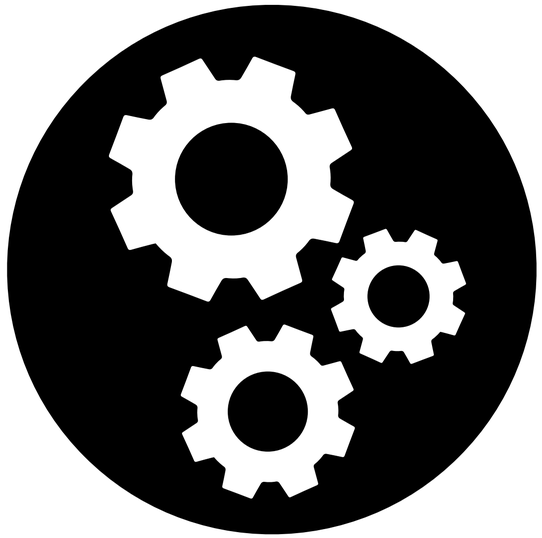
LLMs as trained mechanisms

Input

"Once upon a time, the"

→

LLM



→

Output

"small village of Elmswood…"

Phi-3-mini-4K-Instruct and Phi-3-mini-128K-Instruct were trained over 7 days on 3.3T tokens using 512 H100-80G GPUs for each model. They followed advanced fine-tuning techniques to align with human preferences and safety standards.

The pre-training process followed two distinct and consecutive stages:

- *In the first stage*, the models were primarily exposed to a vast collection of web sources. This data helped the models develop general knowledge and language comprehension.
- *In the second stage*, the models were fine-tuned with a more rigorously selected subset of web data from the first phase, combined with additional synthetic data, to improve their logical reasoning and specialized abilities.

After these 2 stages, the models underwent *additional training*, which included supervised instruction fine-tuning and preference tuning, to enhance their stability and security.

The training dataset, made of 3.3 trillion tokens, is a meticulously curated mix of quality-filtered public documents, select educational materials, code, and newly generated synthetic data generated by LLMs. Specifically, the team filtered the web data to encompass the appropriate degree of *knowledge* and retained a greater number of web pages that may enhance the models' reasoning abilities. Instead of indiscriminately feeding vast amounts of data into the training model, the emphasis was placed on enhancing its reasoning capabilities, rather than one that merely has a vast repository of information.

LLM training = **Pretraining** + **Fine-tuning**

Trillions of tokens
Millions of power consumption
Uses masked/next token prediction

Hand-selected/crafted texts
Quality input-output pairs
Human feedback

Masked/next token prediction

"**Once upon a time**" is a stock phrase used to introduce a narrative of past events, typically in fairy tales and folk tales. It has been used in some form since at least 1380 (according to the *Oxford English Dictionary*) in storytelling in the English language and has started many narratives since 1600. These stories sometimes end with "and they all lived happily ever after", or, originally, "happily until their deaths".

The phrase is common in fairy tales for younger children. It was used in the original translations of the stories of Charles Perrault as a translation for the French "*il était une fois*", of Hans Christian Andersen as a translation for the Danish "*der var engang*" (literally "there was once"), the Brothers Grimm as a translation for the German "*es war einmal*" (literally "it was once") and Joseph Jacobs in English translations and fairy tales.

In *More English Fairy Tales*, Joseph Jacobs notes that:

> "The opening formula are varied enough, but none of them has much play of fancy. 'Once upon a time and a very good time it was, though it wasn't in my time nor in your time nor in any one else's time.' is effective enough for a fairy epoch, and is common, according to Mayhew (London Labour, III), among tramps."[1]

https://en.wikipedia.org/wiki/Once_upon_a_time

LLM

Masked/next token prediction

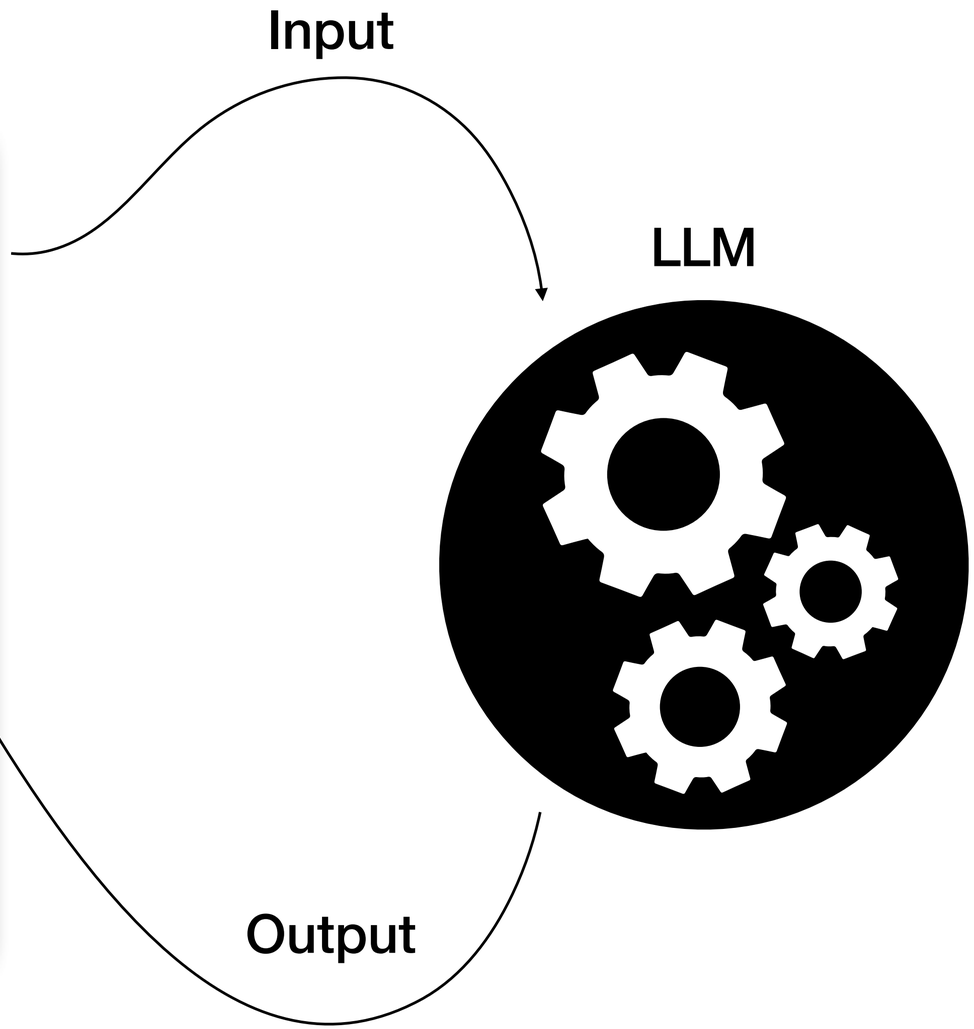
"**Once upon a time**" is a stock phrase used to introduce a narrative of past events, typically in fairy tales and folk tales. It has been used in some form since at least 1380 (according to the *Oxford English Dictionary*) in [redacted] in the English language and has started many narratives since 1600. These stories sometimes end with "and they all lived happily ever after", or, originally, "happily until their deaths".

The phrase is common in fairy tales for younger children. It was used in the original translations of the stories of Charles Perrault as a translation for the French "*il était une fois*", of Hans Christian Andersen as a translation for the Danish "*der var engang*" (literally "there was once"), the Brothers Grimm as a translation for the German "*es war einmal*" (literally "it was once") and Joseph Jacobs in English translations and fairy tales.

In *More English Fairy Tales*, Joseph Jacobs notes that:

> "The opening formula are varied enough, but none of them has much play of fancy. 'Once upon a time and a very good time it was, though it wasn't in my time nor in your time nor in any one else's time.' is effective enough for a fairy epoch, and is common, according to Mayhew (London Labour, III), among tramps."[1]

https://en.wikipedia.org/wiki/Once_upon_a_time



Masked/next token prediction

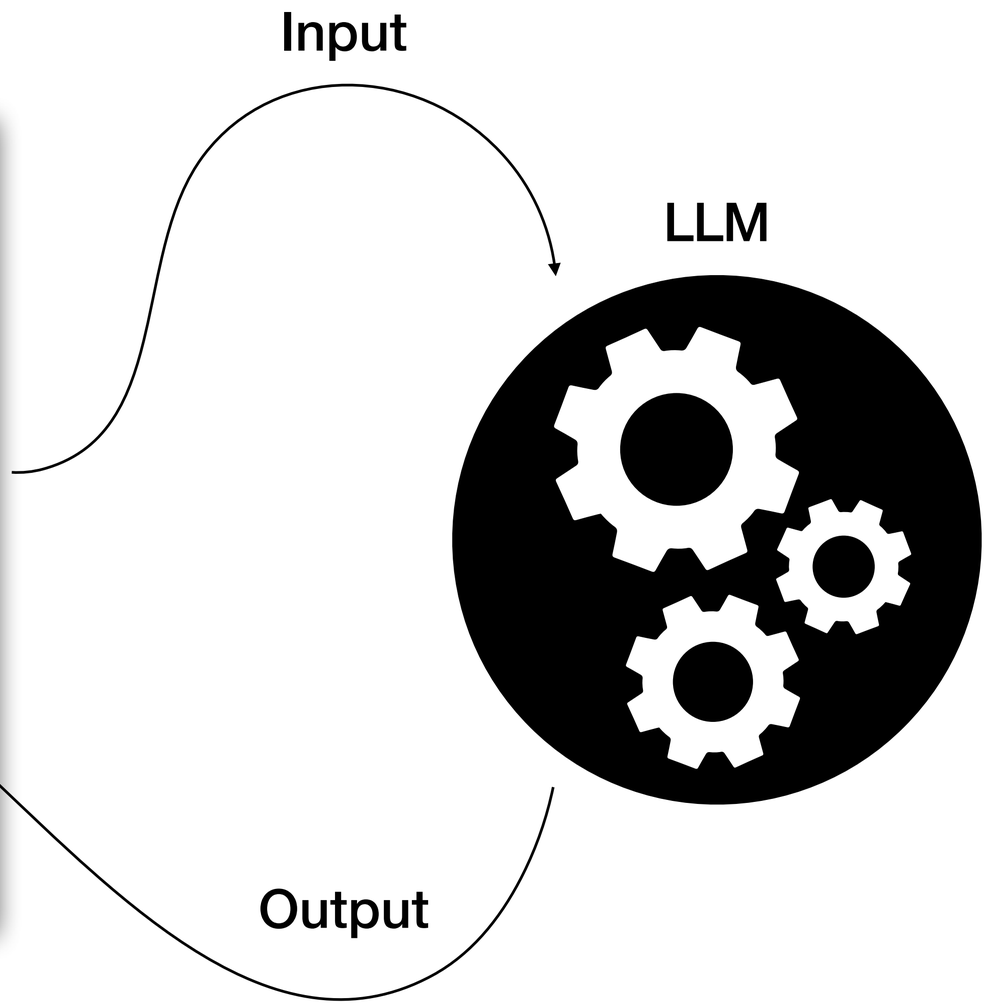
"**Once upon a time**" is a stock phrase used to introduce a narrative of past events, typically in fairy tales and folk tales. It has been used in some form since at least 1380 (according to the *Oxford English Dictionary*) in storytelling in the English language and has started many narratives since 1600. These stories sometimes end with "and they all lived happily ever after", or, originally, "happily until their deaths".

The phrase is common in fairy tales for younger children. It was used in the original translations of the stories of Charles Perrault as a translation for the French "*il était une fois*", of Hans Christian Andersen as a translation for the Danish "*der var engang*" (literally "there was once"), the Brothers [redacted] as a translation for the German "*es war einmal*" (literally "it was once") and Joseph Jacobs in English translations and fairy tales.

In *More English Fairy Tales*, Joseph Jacobs notes that:

> "The opening formula are varied enough, but none of them has much play of fancy. 'Once upon a time and a very good time it was, though it wasn't in my time nor in your time nor in any one else's time.' is effective enough for a fairy epoch, and is common, according to Mayhew (London Labour, III), among tramps."[1]

https://en.wikipedia.org/wiki/Once_upon_a_time



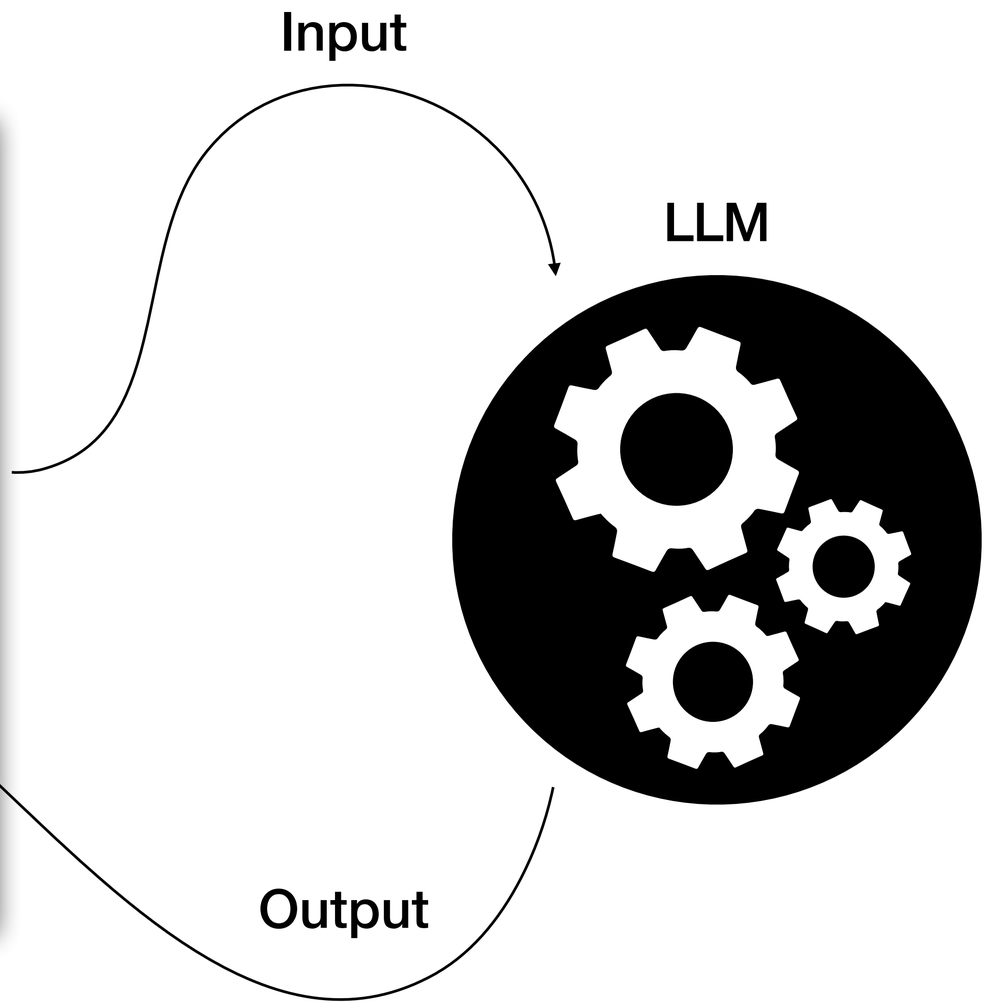
Masked/next token prediction

"**Once upon a time**" is a stock phrase used to introduce a narrative of past events, typically in fairy tales and folk tales. It has been used in some form since at least 1380 (according to the *Oxford English Dictionary*) in storytelling in the English language and has started many narratives since 1600. These stories sometimes end with "and they all lived happily ever after", or, originally, "happily until their deaths".

The phrase is common in fairy tales for younger children. It was used in the original translations of the stories of Charles Perrault as a translation for the French "*il était une fois*", of Hans Christian Andersen as a translation for the Danish "*der var engang*" (literally "there was once"), the Brothers [redacted] as a translation for the German "*es war einmal*" (literally "it was once") and Joseph Jacobs in English translations and fairy tales.

In *More English Fairy Tales*, Joseph Jacobs notes that:

> "The opening formula are varied enough, but none of them has much play of fancy. 'Once upon a time and a very good time it was, though it wasn't in my time nor in your time nor in any one else's time.' is effective enough for a fairy epoch, and is common, according to Mayhew (London Labour, III), among tramps."[1]



https://en.wikipedia.org/wiki/Once_upon_a_time

Masked/next token prediction

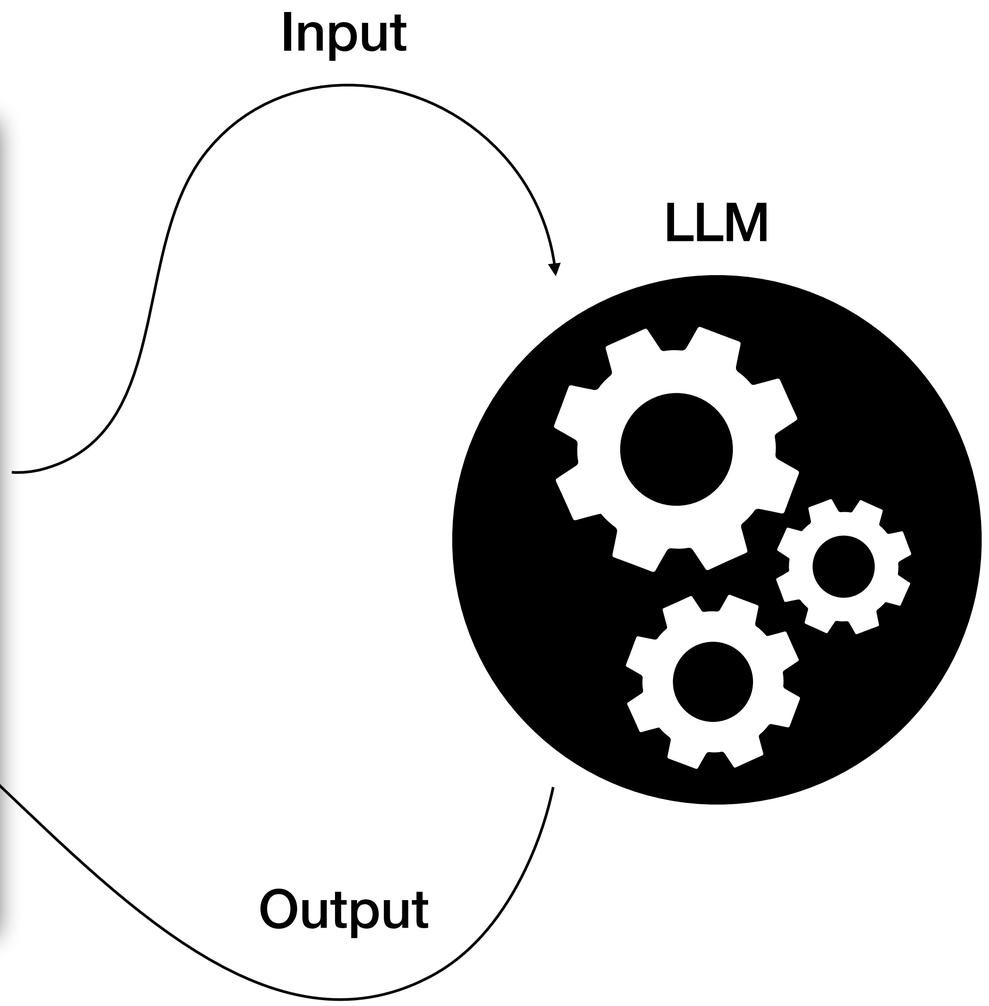
"**Once upon a time**" is a stock phrase used to introduce a narrative of past events, typically in fairy tales and folk tales. It has been used in some form since at least 1380 (according to the *Oxford English Dictionary*) in storytelling in the English language and has started many narratives since 1600. These stories sometimes end with "and they all lived happily ever after", or, originally, "happily until their deaths".

The phrase is common in fairy tales for younger children. It was used in the original translations of the stories of Charles Perrault as a translation for the French "*il était une fois*", of Hans Christian Andersen as a translation for the Danish "*der var engang*" (literally "there was once"), the Brothers Grimm as a translation for the German "*es war einmal*" (literally "it was once") and Joseph Jacobs in English translations and fairy tales.

In *More English Fairy Tales*, Joseph Jacobs notes that:

> "The opening formula are varied enough, but none of them has much play of fancy. 'Once upon a time and a very good time it was, though it wasn't in my time nor in your time nor in any one else's time.' is effective enough for a fairy epoch, and is common, according to Mayhew (London Labour, III), among tramps."[1]

https://en.wikipedia.org/wiki/Once_upon_a_time



Masked/next token prediction

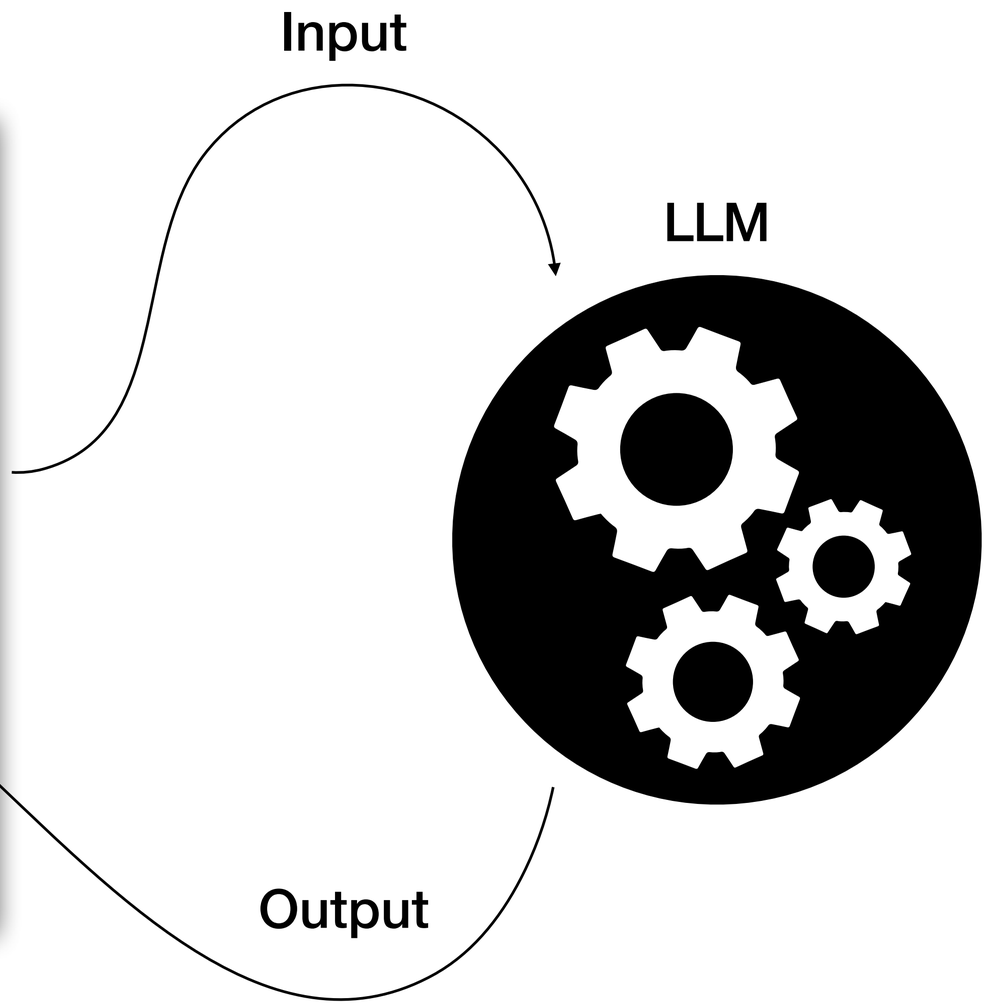
"**Once upon a time**" is a stock phrase used to introduce a narrative of past events, typically in fairy tales and folk tales. It has been used in some form since at least 1380 (according to the *Oxford English Dictionary*) in storytelling in the English language and has started many narratives since 1600. These stories sometimes end with "and they all lived happily ever after", or, originally, "happily until their deaths".

The phrase is common in fairy tales for younger children. It was used in the original translations of the stories of Charles Perrault as a translation for the French "*il était une fois*", of Hans Christian Andersen as a translation for the Danish "*der var engang*" (literally "there was once"), the Brothers Grimm as a translation for the German "*es war einmal*" (literally "it was once") and Joseph Jacobs in English translations and fairy tales.

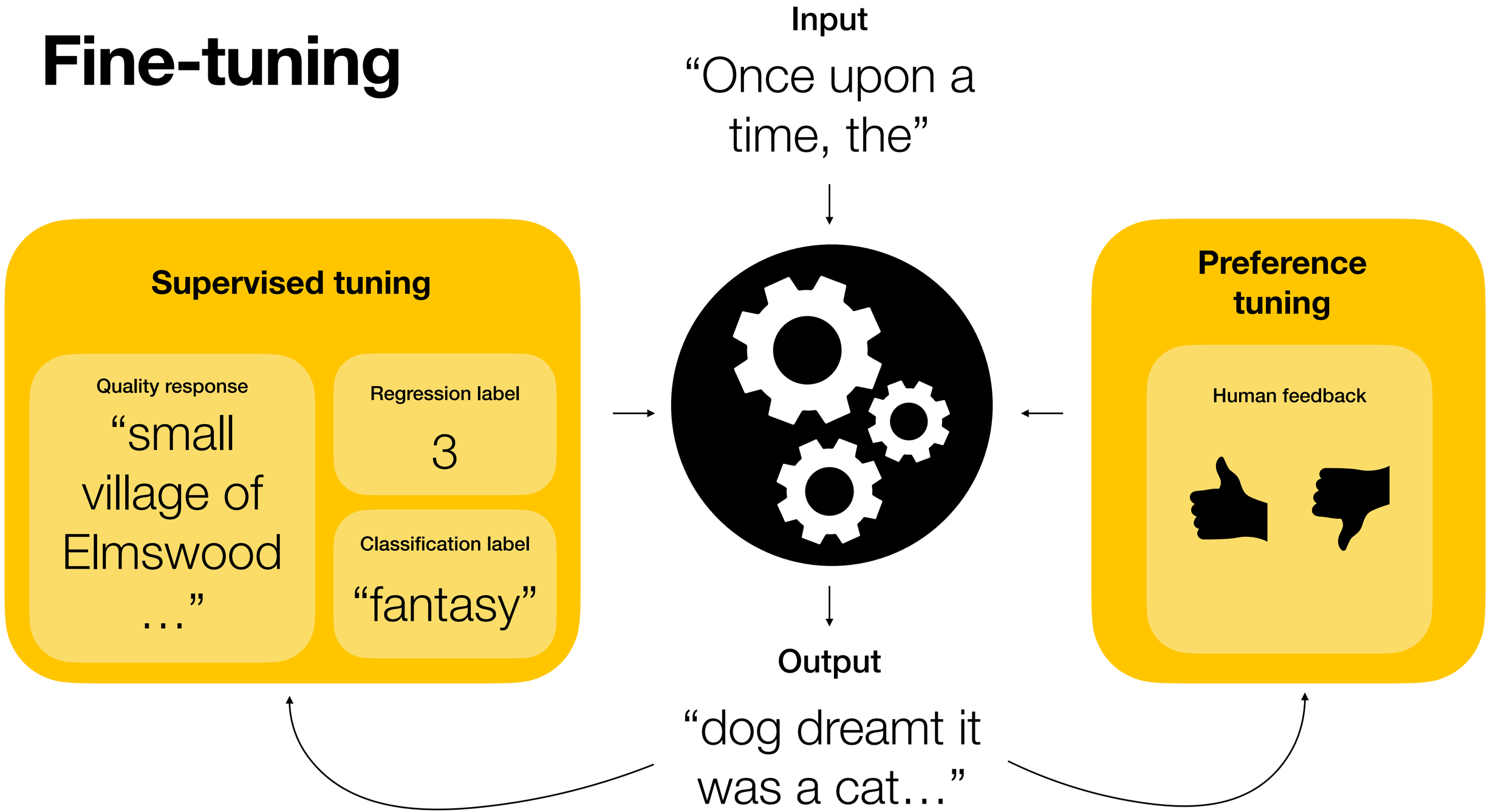
In *More English Fairy Tales*, Joseph Jacobs notes that:

> "The opening formula are varied enough, but none of them has much play of fancy. 'Once upon a time and a very good time it was, though it wasn't in my time nor in your time nor in any one else's time.' is effective enough for a fairy epoch, and is common, according to Mayhew (London Labour, III), among tramps."[1]

https://en.wikipedia.org/wiki/Once_upon_a_time



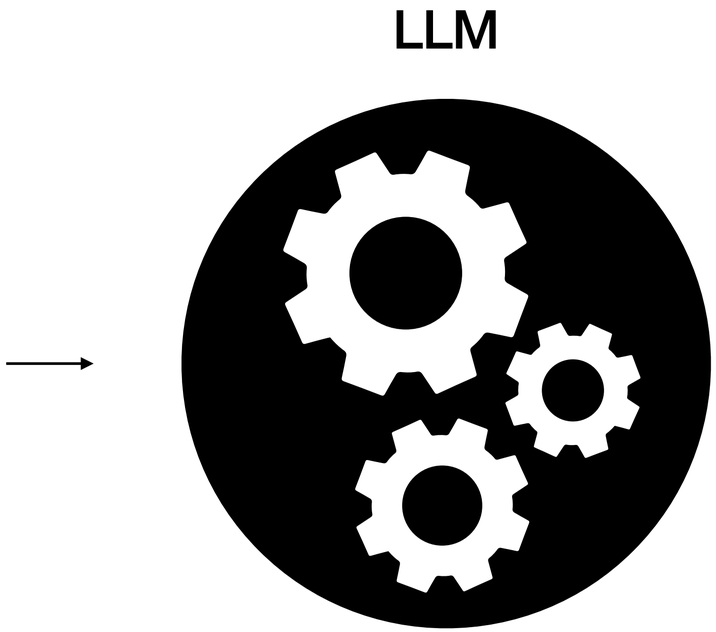
Fine-tuning


Input
“Once upon a time, the”
Supervised tuning
Quality response
“small village of Elmswood …”
Regression label
3
Classification label
“fantasy”
Preference tuning
Human feedback
Output
“dog dreamt it was a cat…”

Two major applications

Input

"Once upon a time, the"



LLM

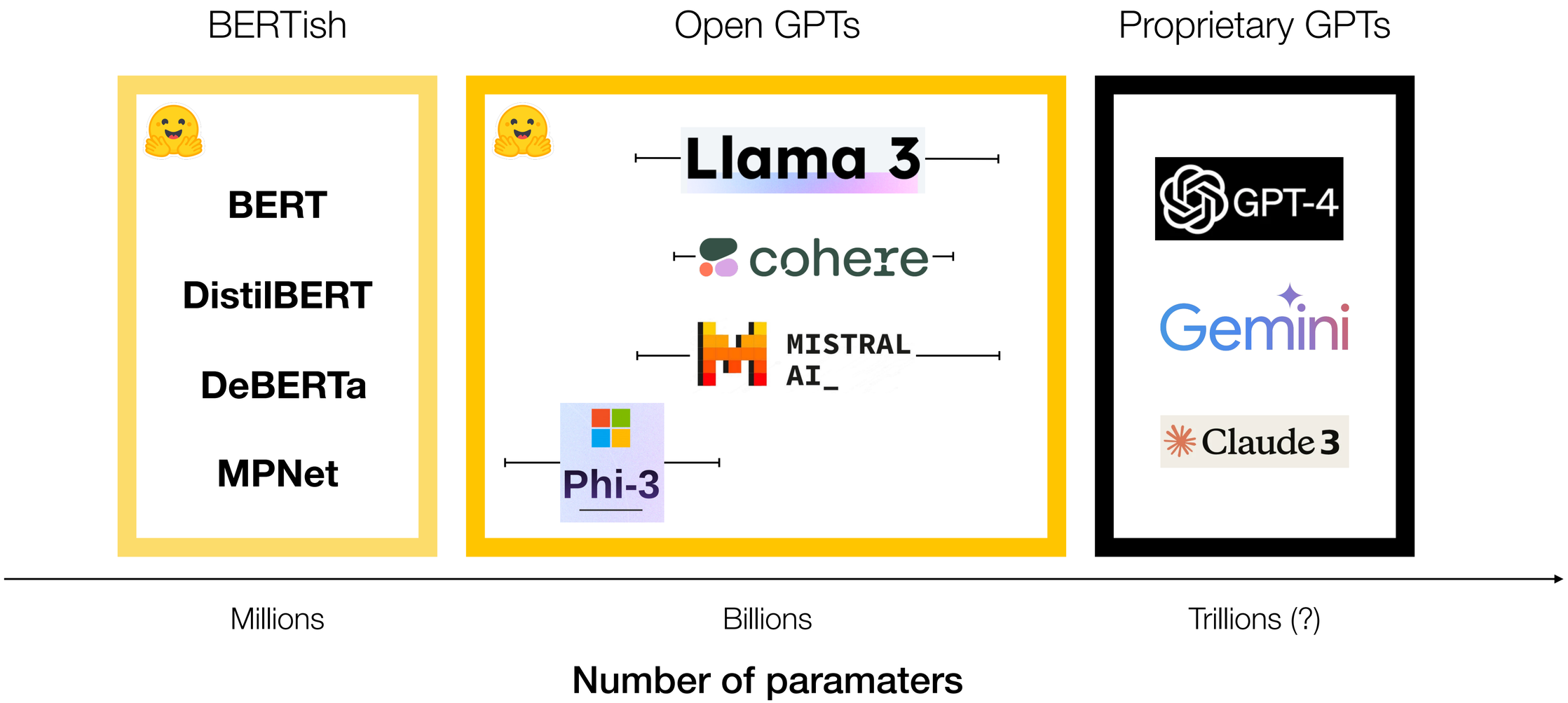
Text generation

"small village of Elmswood…"

Feature extraction

-.23, 1.23, .24, -.12, .34, .32, …

Models


BERTish
Open GPTs
Proprietary GPTs
BERT
DistilBERT
DeBERTa
MPNet
Llama 3
cohere
MISTRAL AI_
Phi-3
GPT-4
Gemini
Claude 3
Millions
Billions
Trillions (?)
Number of paramaters

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Audio-Text-to-Text
- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering
- Video-Text-to-Text
- Any-to-Any

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Text-to-Image
- Image-to-Text
- Image-to-Image
- Image-to-Video
- Unconditional Image Generation
- Video Classification
- Text-to-Video
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection
- Text-to-3D
- Image-to-3D
- Image Feature Extraction
- Keypoint Detection

Models 1,304,249

Filter by name | Full-text search | Sort: Trending

hexgrad/Kokoro-82M
Text-to-Speech • Updated 1 day ago • ↓ 25k • ♡ 2.01k

openbmb/MiniCPM-o-2_6
Any-to-Any • Updated about 5 hours ago • ↓ 15k • ♡ 609

microsoft/phi-4
Text Generation • Updated 11 days ago • ↓ 124k • ♡ 1.44k

MiniMaxAI/MiniMax-Text-01
Text Generation • Updated 3 days ago • ↓ 2.38k • ♡ 406

deepseek-ai/DeepSeek-V3
Updated 21 days ago • ↓ 155k • ♡ 2.04k

NovaSky-AI/Sky-T1-32B-Preview
Text Generation • Updated 6 days ago • ↓ 7.51k • ♡ 476

jinaai/ReaderLM-v2
Text Generation • Updated 3 days ago • ↓ 2.64k • ♡ 258

MiniMaxAI/MiniMax-VL-01
Text Generation • Updated 4 days ago • ↓ 635 • ♡ 196

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **1,000,000** human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper.

Arena | Full Leaderboard

Total #models: **99**. Total #votes: **1,170,955**. Last updated: 2024-05-20.

📣 **NEW!** View leaderboard for different categories (e.g., coding, long user query)! This is still in preview and subject to change.

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-Exp-1206	1374	+4/-5	20227	Google	Proprietary
1	1	ChatGPT-4o-latest (2024-11-20)	1365	+4/-3	33383	OpenAI	Proprietary
1	4	Gemini-2.0-Flash-Thinking-Exp-1219	1364	+5/-6	15728	Google	Proprietary
2	4	Gemini-2.0-Flash-Exp	1357	+6/-4	19030	Google	Proprietary
3	1	o1-2024-12-17	1351	+7/-7	7289	OpenAI	Proprietary
6	4	o1-preview	1335	+4/-4	33194	OpenAI	Proprietary
7	7	DeepSeek-V3	1319	+6/-6	10510	DeepSeek	DeepSeek
7	10	Step-2-16K-Exp	1305	+8/-9	3374	StepFun	Proprietary

Massive Text Embedding Benchmark (MTEB) Leaderboard. To submit, refer to the MTEB GitHub repository 🤗 Refer to the MTEB paper for details on metrics, tasks and models.

Overall | Bitext Mining | Classification | Clustering | Pair Classification | Reranking | Retrieval | STS | Summarization | Retrieval w/Instructions

English | Chinese | French | Polish

Overall MTEB English leaderboard 🔮

- **Metric:** Various, refer to task tabs
- **Languages:** English

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)
1	NV-Embed-v1					69.32	87.35	52.8
2	voyage-large-2-instruct			1024	16000	68.28	81.49	53.35
3	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67
4	gte-Qwen1.5-7B-instruct	7099	26.45	4096	32768	67.34	79.6	55.83
5	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13	79.25	52.42
6	GritLM-7B	7242	26.98	4096	32768	66.76	79.46	50.61
7	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63	78.47	50.26
8	google-gecko.text-embedding-p	1200	4.47	768	2048	66.31	81.17	47.48
9	SE_v1					65.66	76.8	47.38
10	GritLM-8x7B	46703	173.98	4096	32768	65.66	78.53	50.14

Models

BERTish

all-MiniLM-L6-v2

Feature extraction
(fine-tuning)

Open GPTs

Llama 3.2 1B/3B

Text-generation
(in-context learning)

Analyzing science of science research

ELSEVIER

Scopus

PNAS RESEARCH ARTICLE | DEMOGRAPHY OPEN ACCESS

A gender perspective on the global migration of scholars

Xinyi Zhao[a,b,1], Allakbar Akbaritabar[a], Ridhi Kashyap[b,c], and Emilio Zagheni[a]

Edited by Douglas Massey, Princeton University, Princeton, NJ; received August 26, 2022; accepted January 5, 2023

Significance

Within a globalizing scientific system, international migration is increasingly recognized as a strategy for scientists to advance their careers. The migration literature more broadly has suggested a process of feminization, with an increasing share of women among all international migrants. With respect to the migration of scholars, however, whether male and female scholars participate equally in transnational mobility and how these patterns have shifted over time from a global perspective are not known. Using bibliometric data that cover the past two decades, we show that, while female researchers continued to be underrepresented among internationally mobile researchers, and migrated over shorter distances, this gender gap has been narrowing at a faster rate than the gender gap in the population of general researchers.

Does the article use bibliometric analysis?

Statistical reasoning

Spatial search

Cognitive neuro-science

Performance monitoring

“science of science” OR “metascience” OR “meta science”

1,124 titles, abstracts, etc.

PDF articles