

Applying open-source LLMs in Social and Behavioral Sciences

Dirk Wulff & Zak Hussain



MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

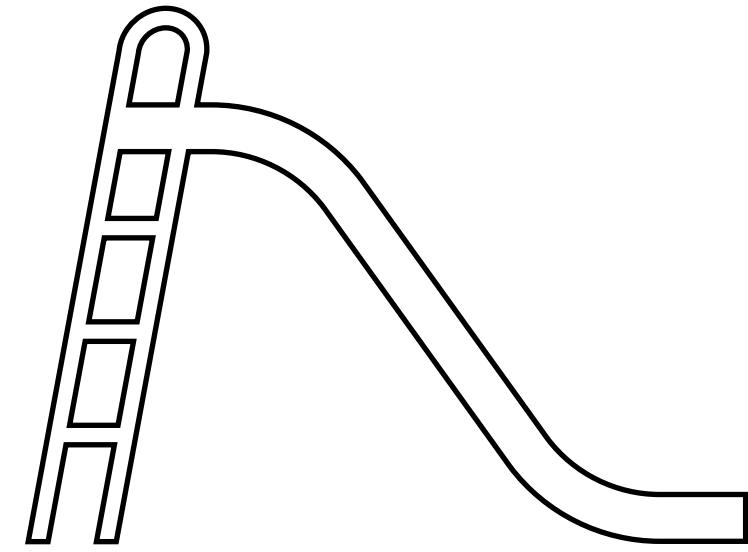


Goals

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

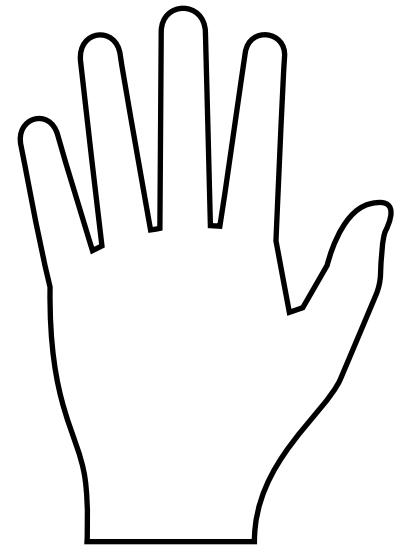


Components



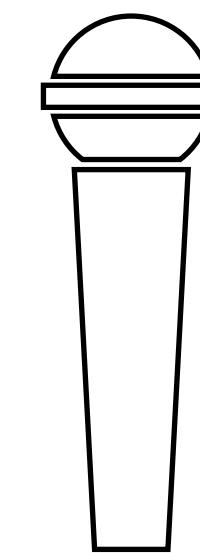
Slide-based introductions

Introduces core concepts and code



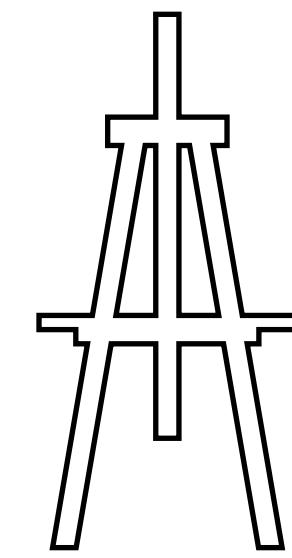
Hands-on exercises

Work through ready-made notebooks to carry out analyses step-by-step



Discussions

Discuss and reflect on applications of LLMs

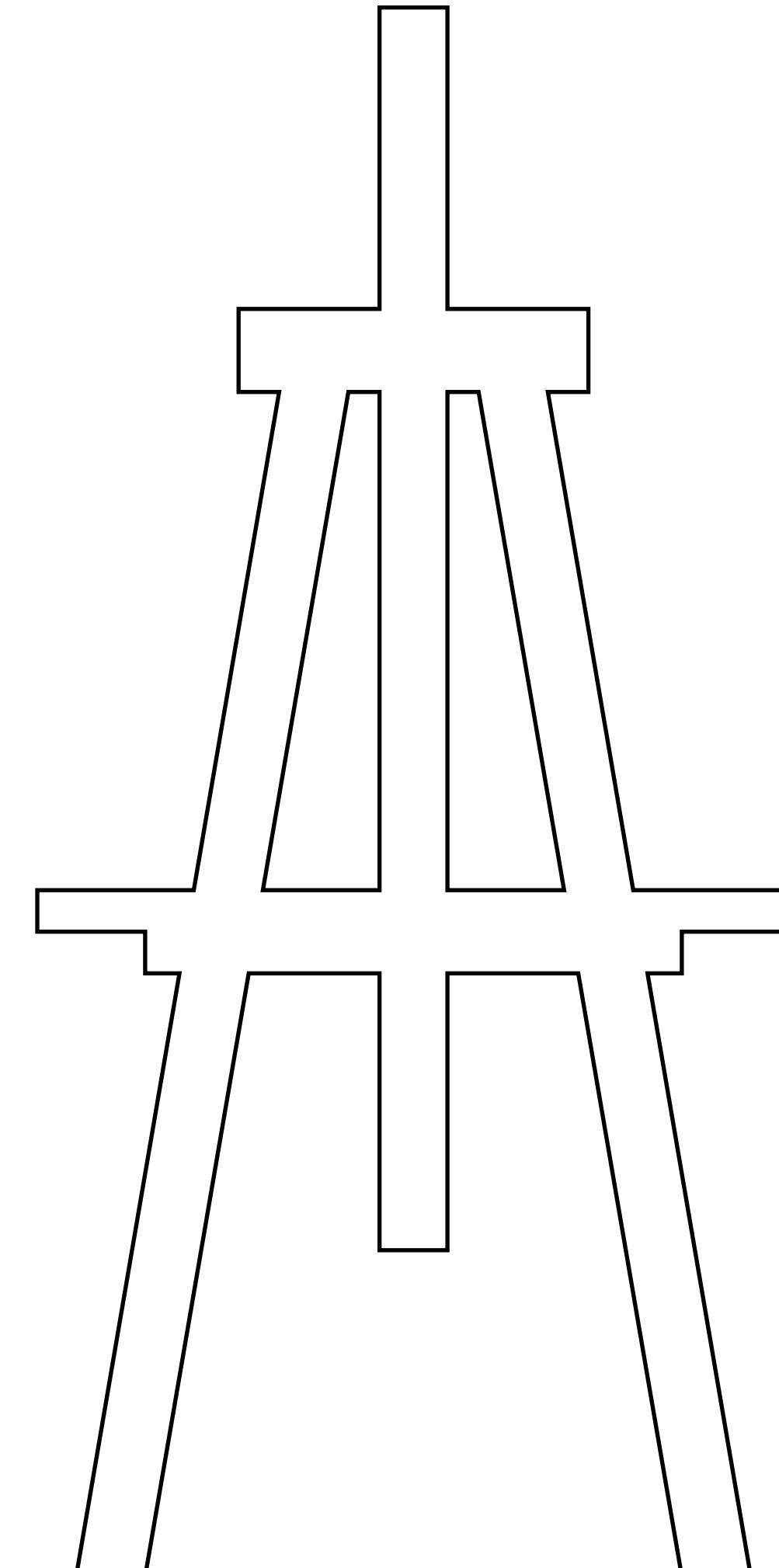


Project sketching

Sketch and pitch an LLM project of your own

Project

- † Any project using LLMs
- † Needs to be motivated (i.e., why is it interesting?)
- † Can be based on the exercises, but must go beyond (i.e., different data, different models, etc.)
- † Is presented at the end (max. 3 slides and 4 minutes)
- † Carried out after the event and documented in a two-page research article **submitted by June 28**



Software stack



+



+





Zak-Hussain / LLM4BeSci_GSERM2024

Type / to search

[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

LLM4BeSci_GSERM2024

Public

Unwatch 2

Fork 0

Star 0

main

1 Branch

0 Tags

Go to file

Add file

Code

About

Zak-Hussain	remove gpu instruction ✓	3672619 · 1 hour ago	112 Commits
day_1	final tweaking	8 hours ago	
day_2	final tweaking	8 hours ago	
day_3	final tweaking	8 hours ago	
day_4	add text	10 hours ago	
day_5	final tweaking	8 hours ago	
.gitignore	Initial commit	last month	
LICENSE.txt	add LICENSE.txt	last month	
README.md	remove gpu instruction	1 hour ago	

README	License		
--------	---------	--	--

LLM4BeSci at GSERM2024

The course introduces the use of open source large language models (LLMs) from the Hugging Face ecosystem for research in the behavioral and social sciences.

[Readme](#)[View license](#)[Activity](#)[0 stars](#)[2 watching](#)[0 forks](#)[Report repository](#)

Releases

No releases published

[Create a new release](#)

Packages

No packages published

Schedule

Day 1	Day 2	Day 3	Day 4	Day 5
<i>Morning</i>	<i>Morning</i>	<i>Morning</i>	<i>Morning</i>	<i>Morning</i>
Welcome and intro to large language models	Intro to feature extraction and embedding models	Intro to classification and regression	Intro to token and text generation	Additional applications of LLMs for qualitative data analysis
<i>Afternoon</i>	<i>Afternoon</i>	<i>Afternoon</i>	<i>Afternoon</i>	<i>Afternoon</i>
Applying the Hugging Face ecosystem for open-source large language models	Applying large language models to predict the relationships between survey items and questionnaires in personality psychology	Applying large language models to evaluate and classify texts in political science	Applying large language models to predict the relationships between survey items and questionnaires in personality psychology	Project pitches

Today

09:15 am - 09:45 am	Welcome & Intro
09:45 am - 10:45 am	Talk: Intro to LLMs
10:45 am - 11:00 am	Break
11:00 am - 12:00 pm	Exercise: Identify LLM applications in small groups
12:00 pm - 01:00 pm	Lunch
01:00 pm - 02:00 pm	Talk: A gentle intro to Hugging Face and Python
02:00 pm - 02:15 pm	Setup Colab
02:15 pm - 02:45 pm	Exercise: Run pipelines
02:45 pm - 03:15 pm	Walkthrough



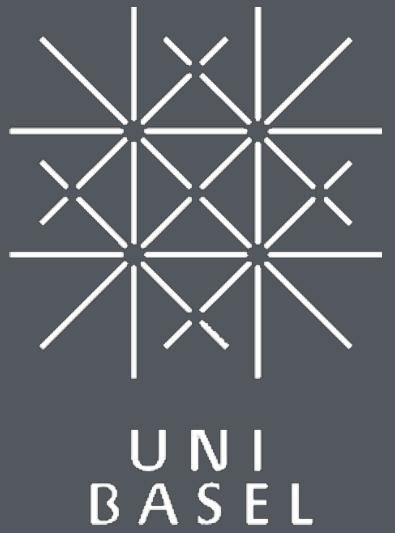
1. Where are you from, what is your background, and what do you do?
2. What motivates you to learn more about LLMs?
3. R or Python or __ ?
4. How much experiences do you have with programming and machine learning?

Intro LLMs

Dirk Wulff & Zak Hussain

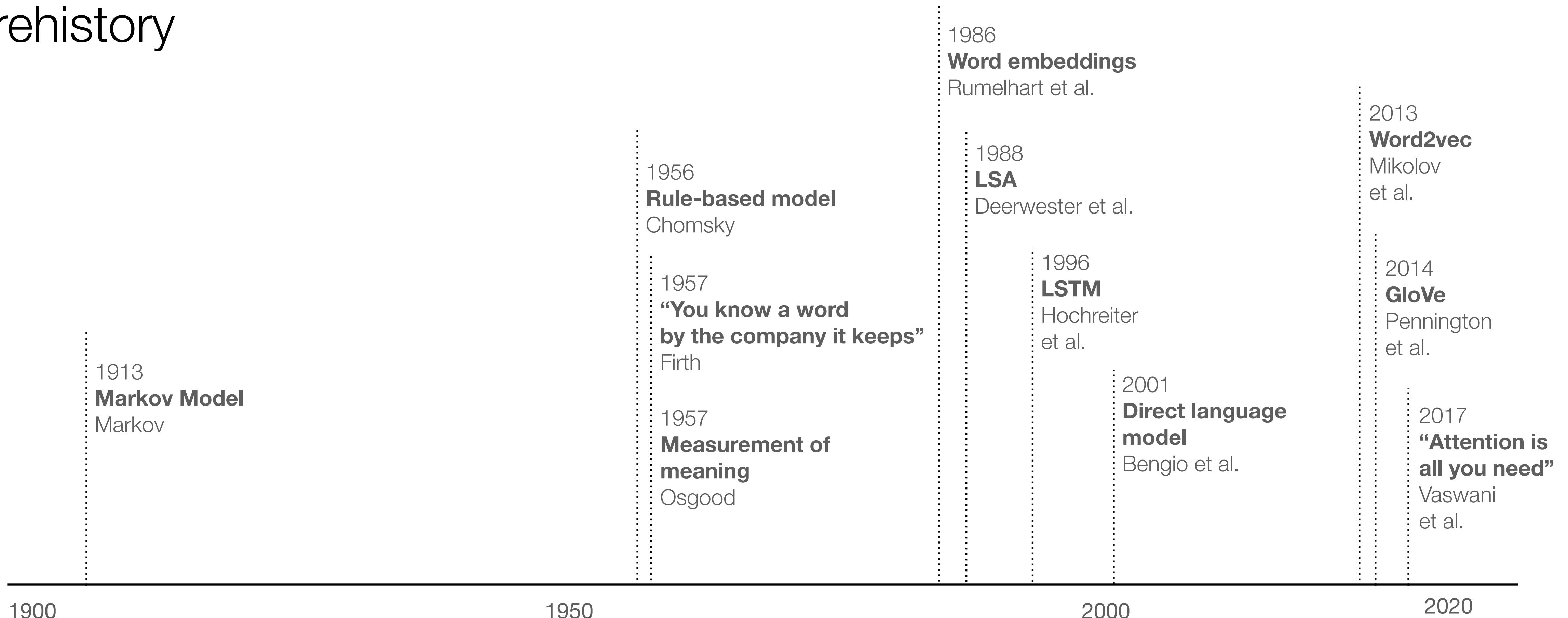


MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT



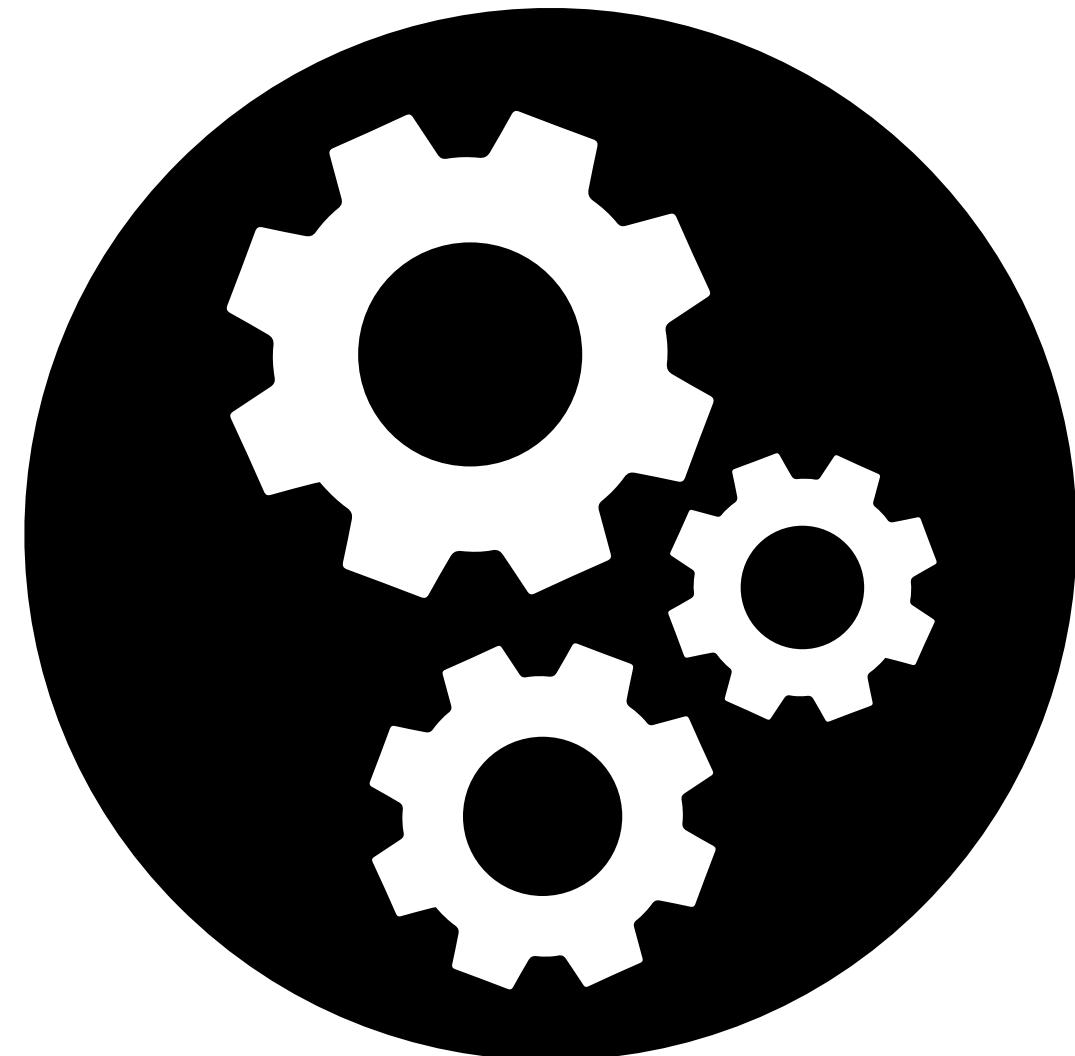
Language models

Prehistory

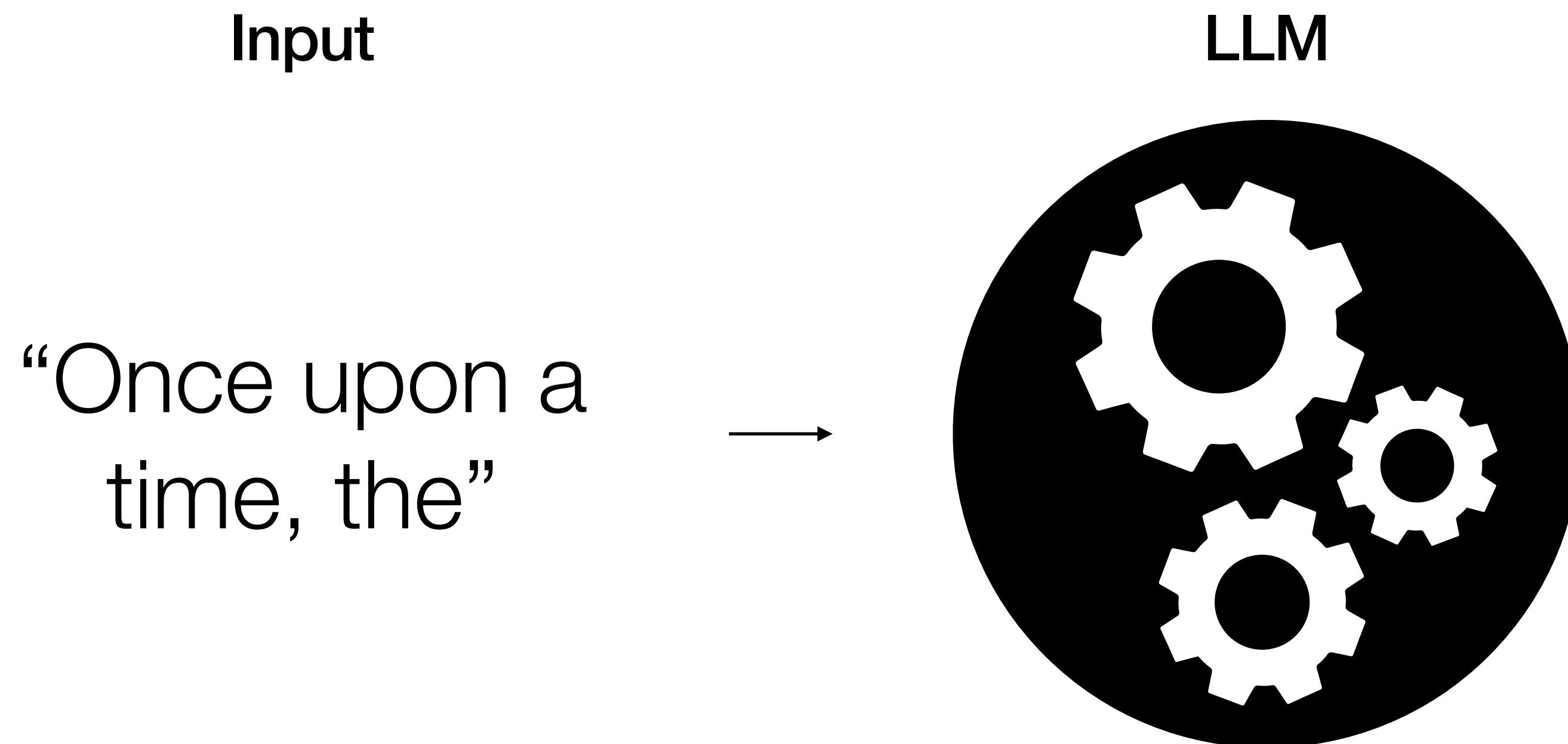


LLMs as mechanisms

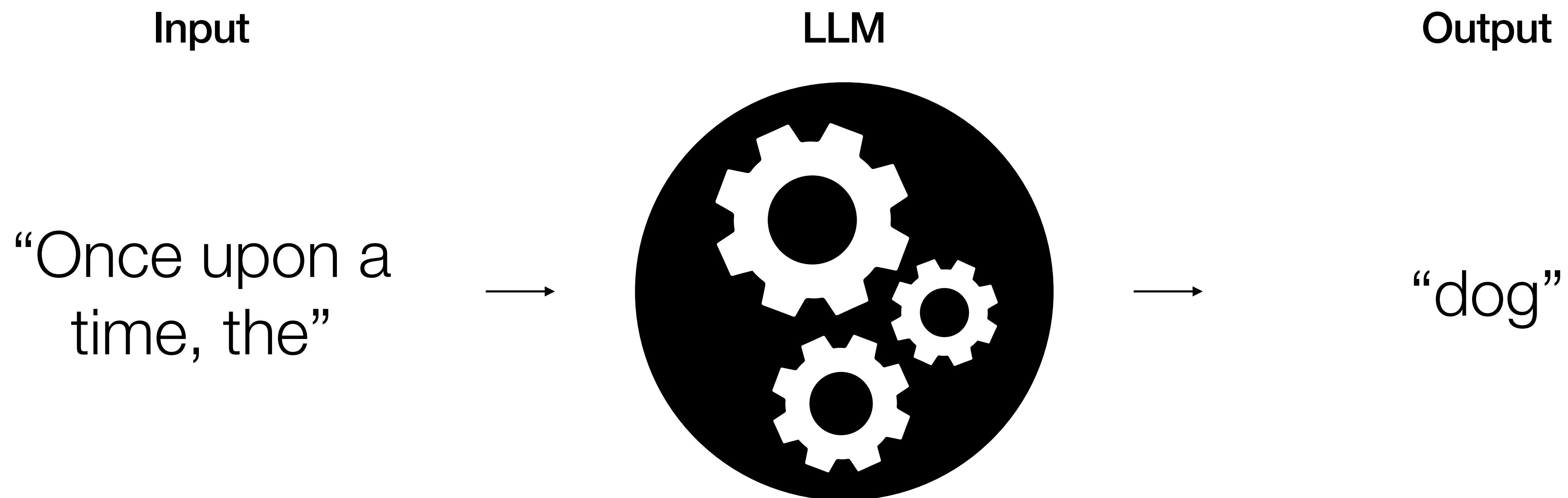
LLM



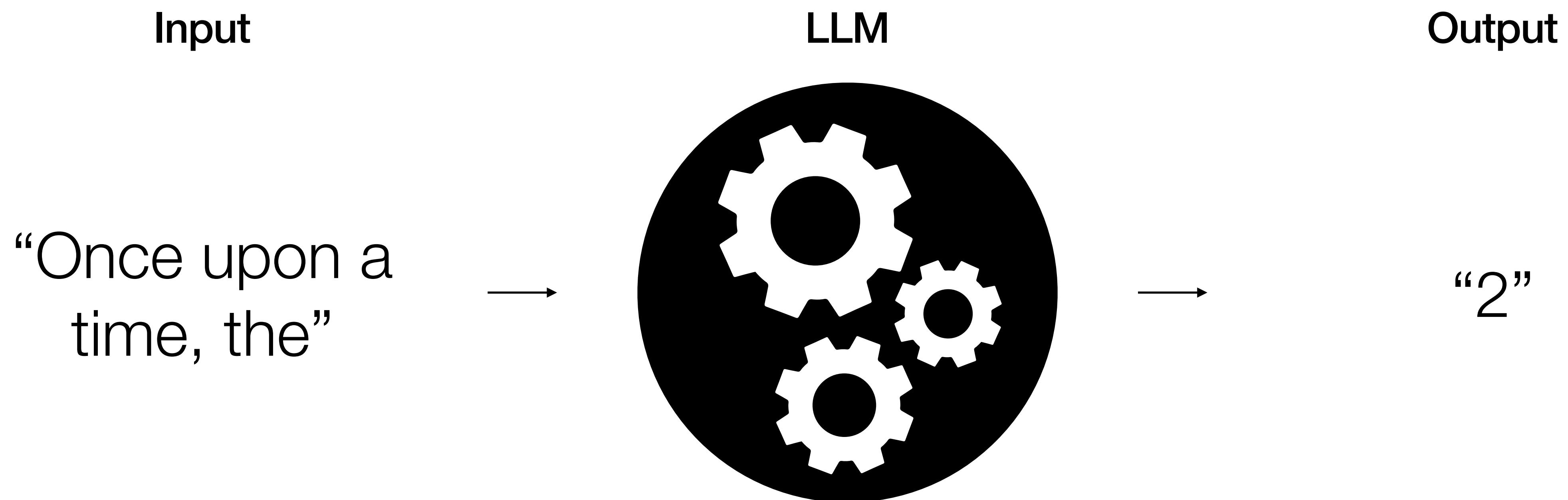
LLMs as mechanisms



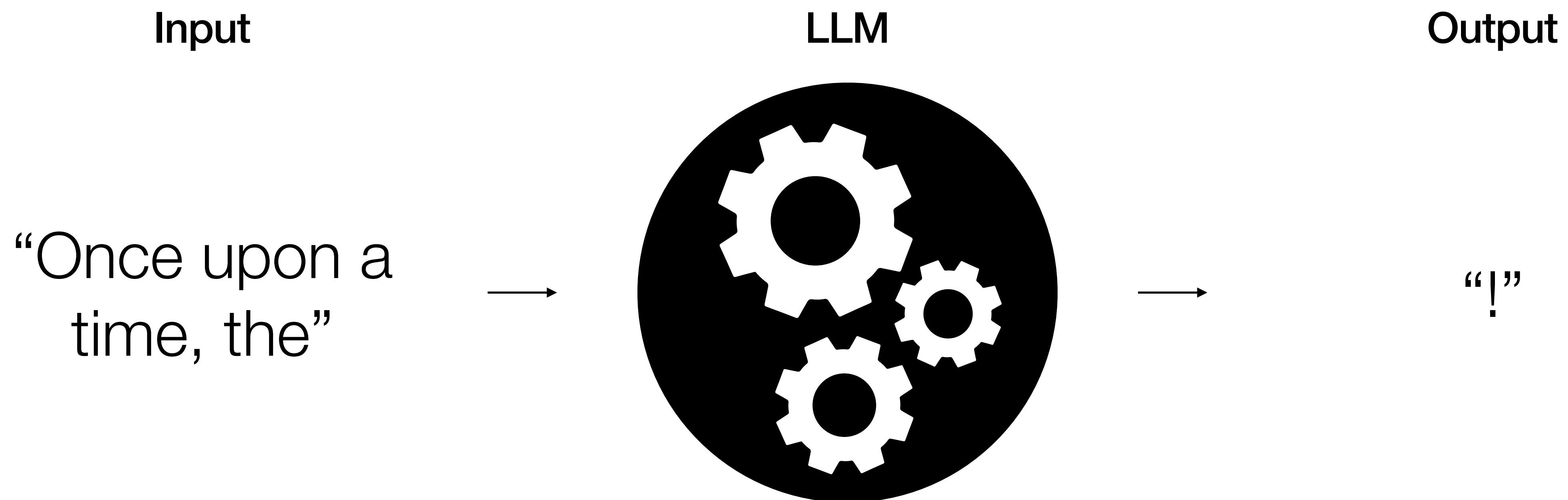
LLMs as mechanisms



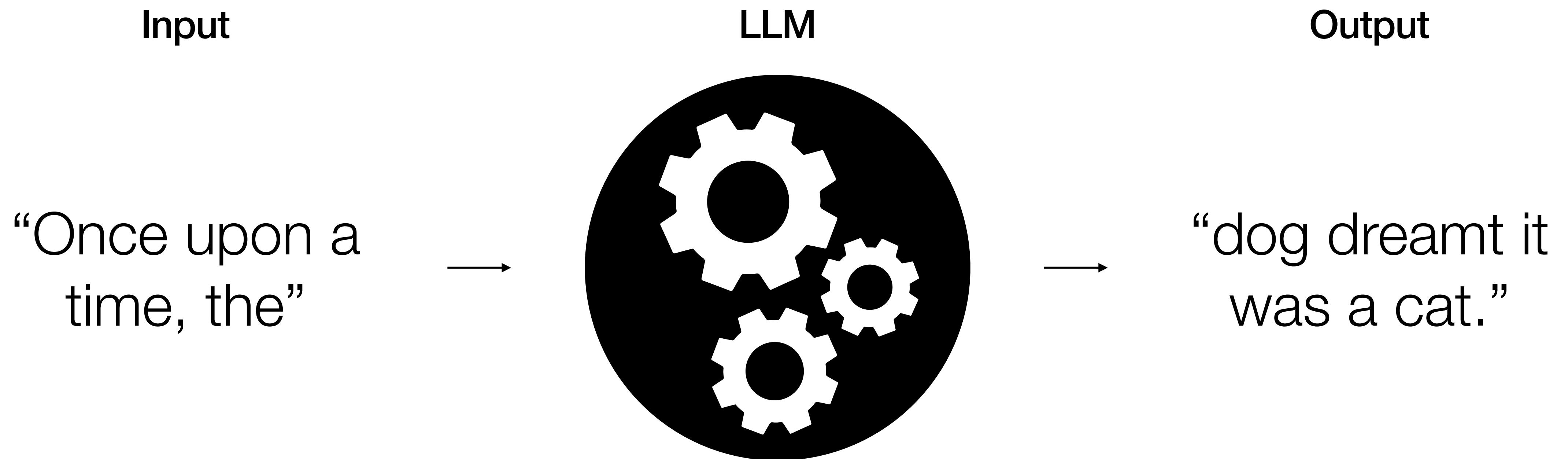
LLMs as mechanisms



LLMs as mechanisms



LLMs as mechanisms



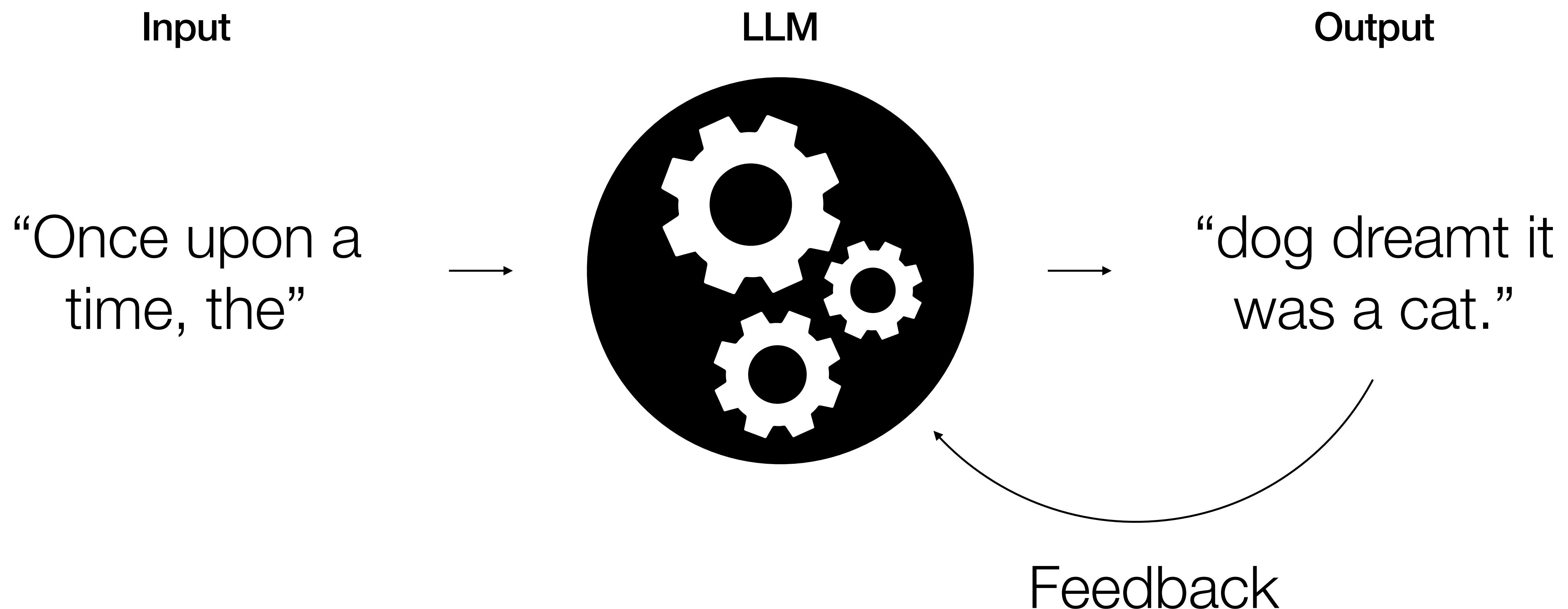
Once upon a time, the



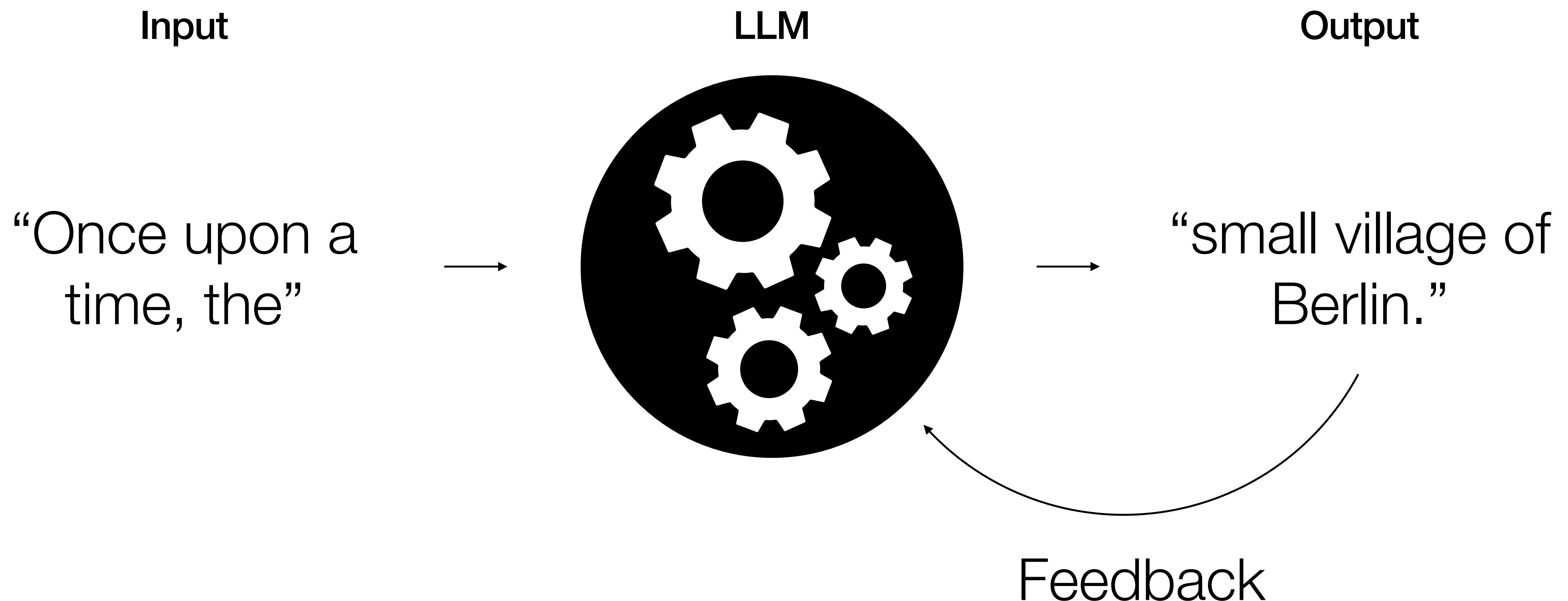
Once upon a time, the small village of Elmswood was nestled in a lush valley surrounded by towering mountains. The villagers lived peacefully, their days marked by the rhythms of nature and the changing seasons. However, everything changed when a mysterious old man arrived, carrying with him a locked 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.



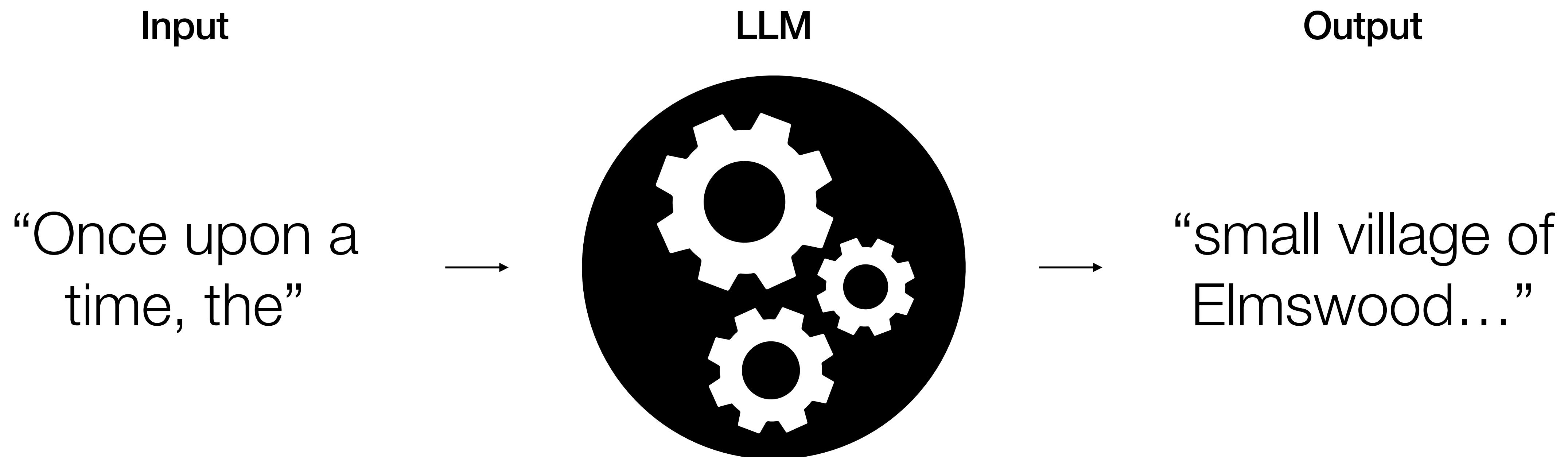
LLMs as **trained** mechanisms



LLMs as **trained** mechanisms



LLMs as **trained** mechanisms



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.



Phi-3

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".

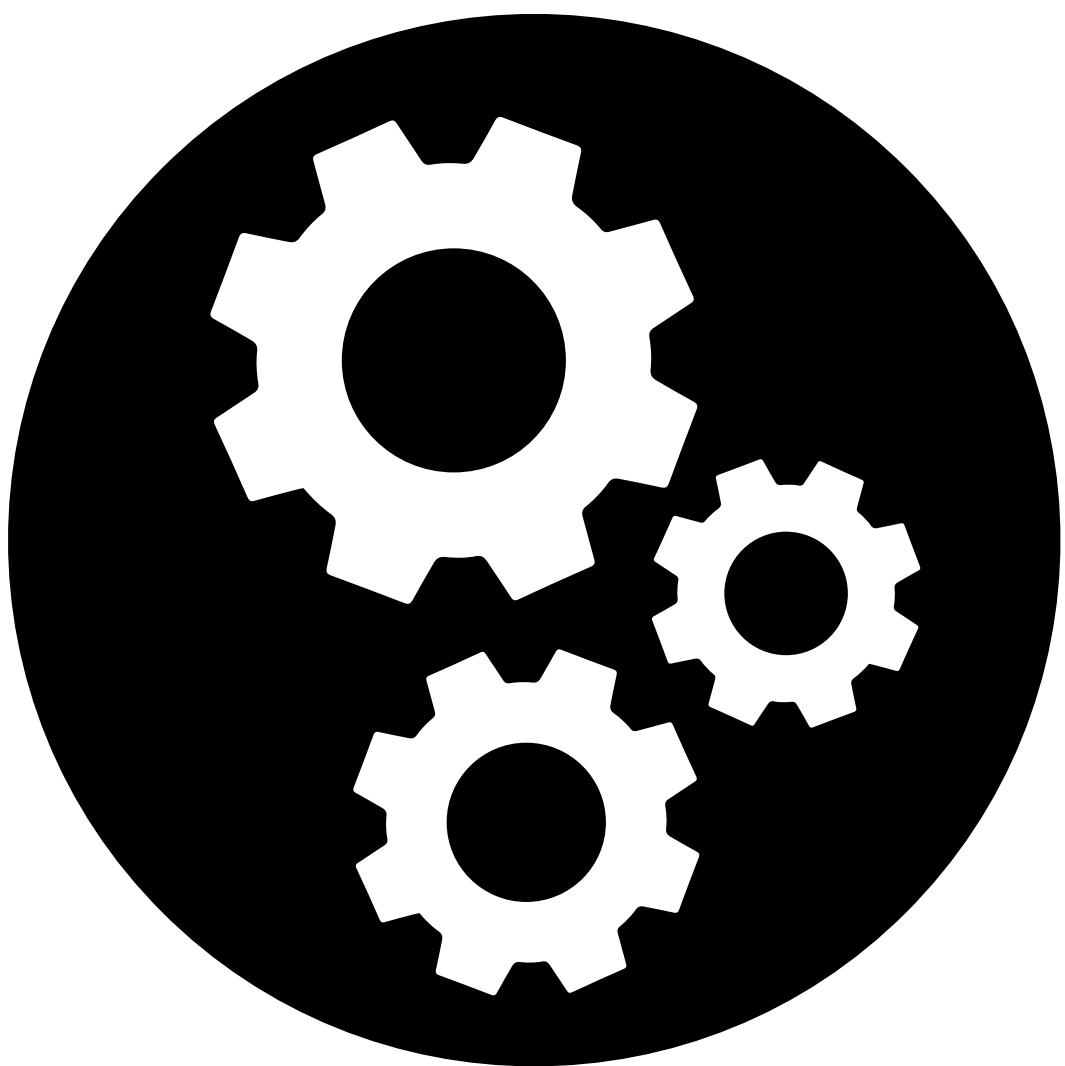
The phrase is common in [fairy tales](#) for younger children. It was used in the original translations of the stories of [Charles Perrault](#) as a translation for the [French](#) "*il était une fois*", of [Hans Christian Andersen](#) as a translation for the [Danish](#) "*der var engang*" (literally "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 formulae 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

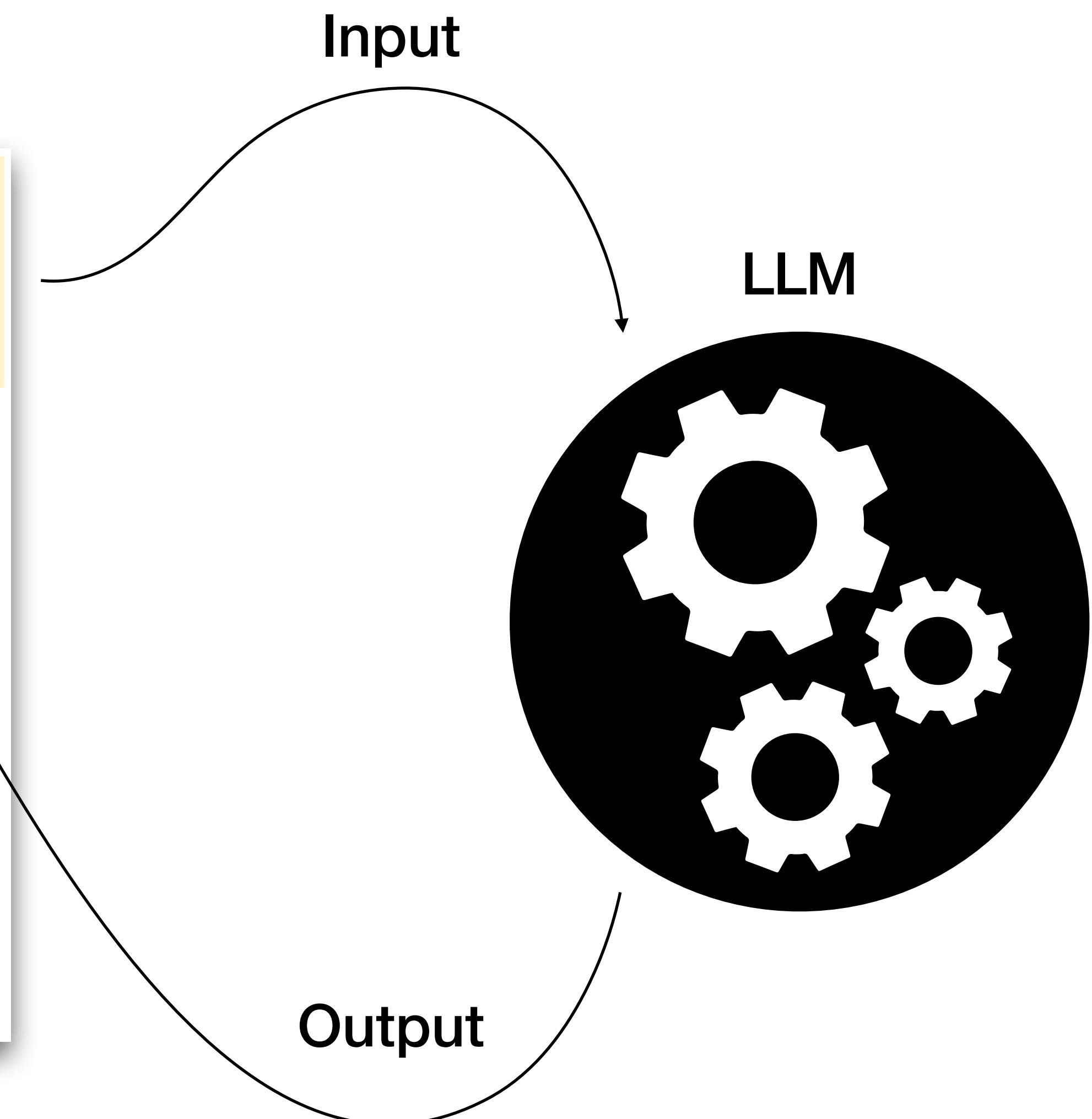
"Once upon a time" is a stock phrase used to introduce a narrative of past events, typically in fairy tales and folk tales. It has been used in some form since at least 1380 (according to the *Oxford English Dictionary*) in [REDACTED] in the English language and has started many narratives since 1600. 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 formulae 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

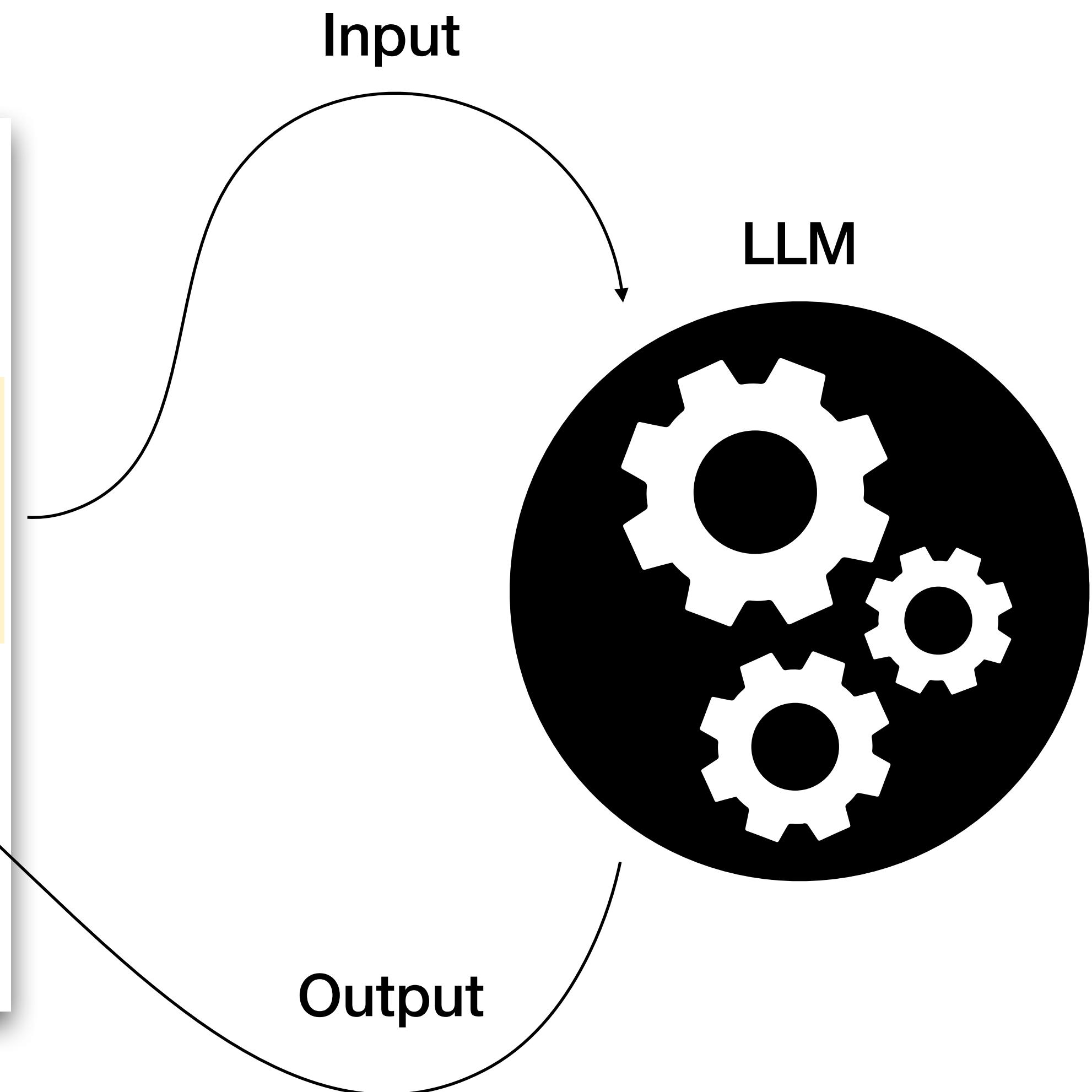
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many 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 formulae 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

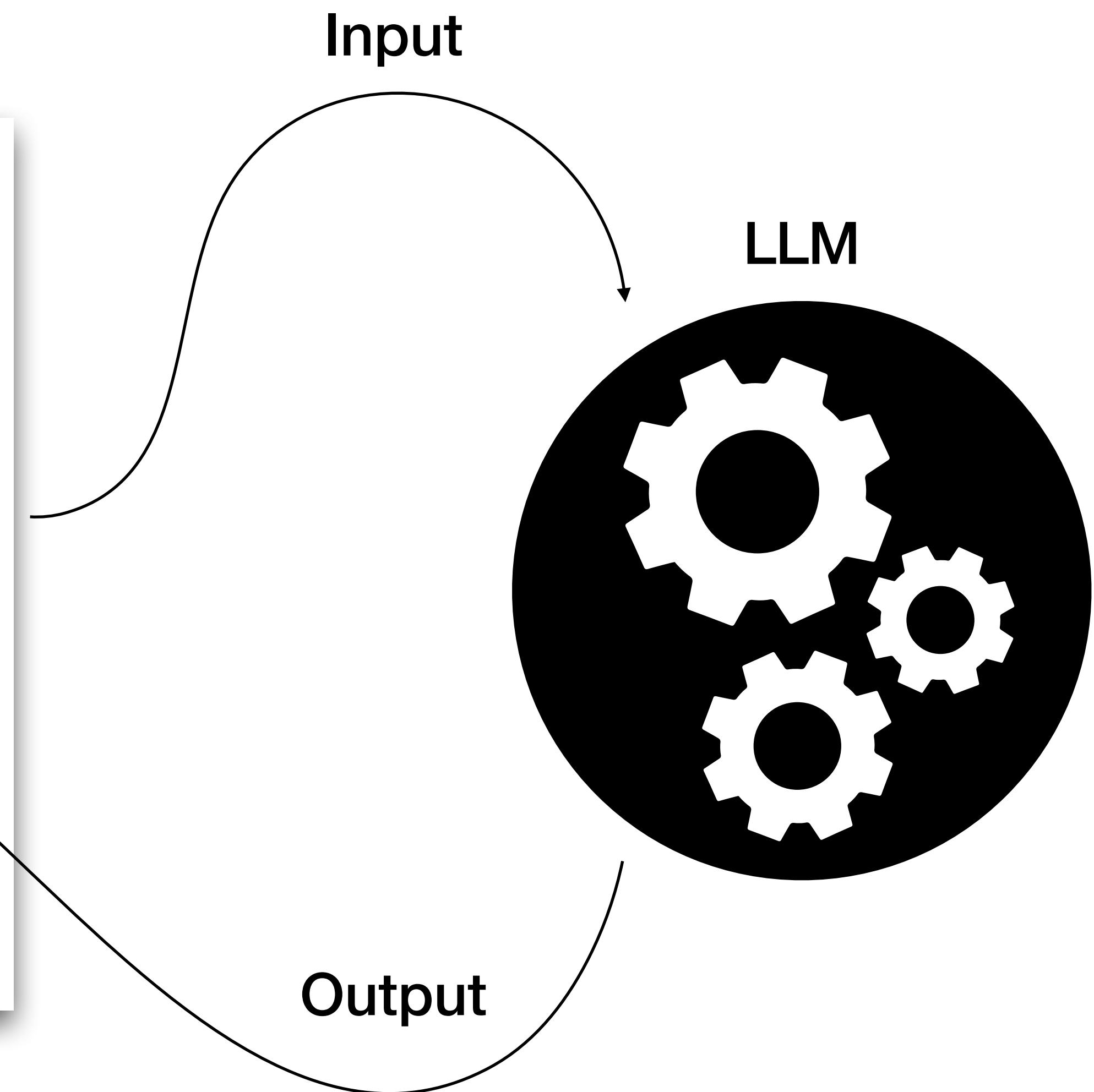
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many 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 formulae 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

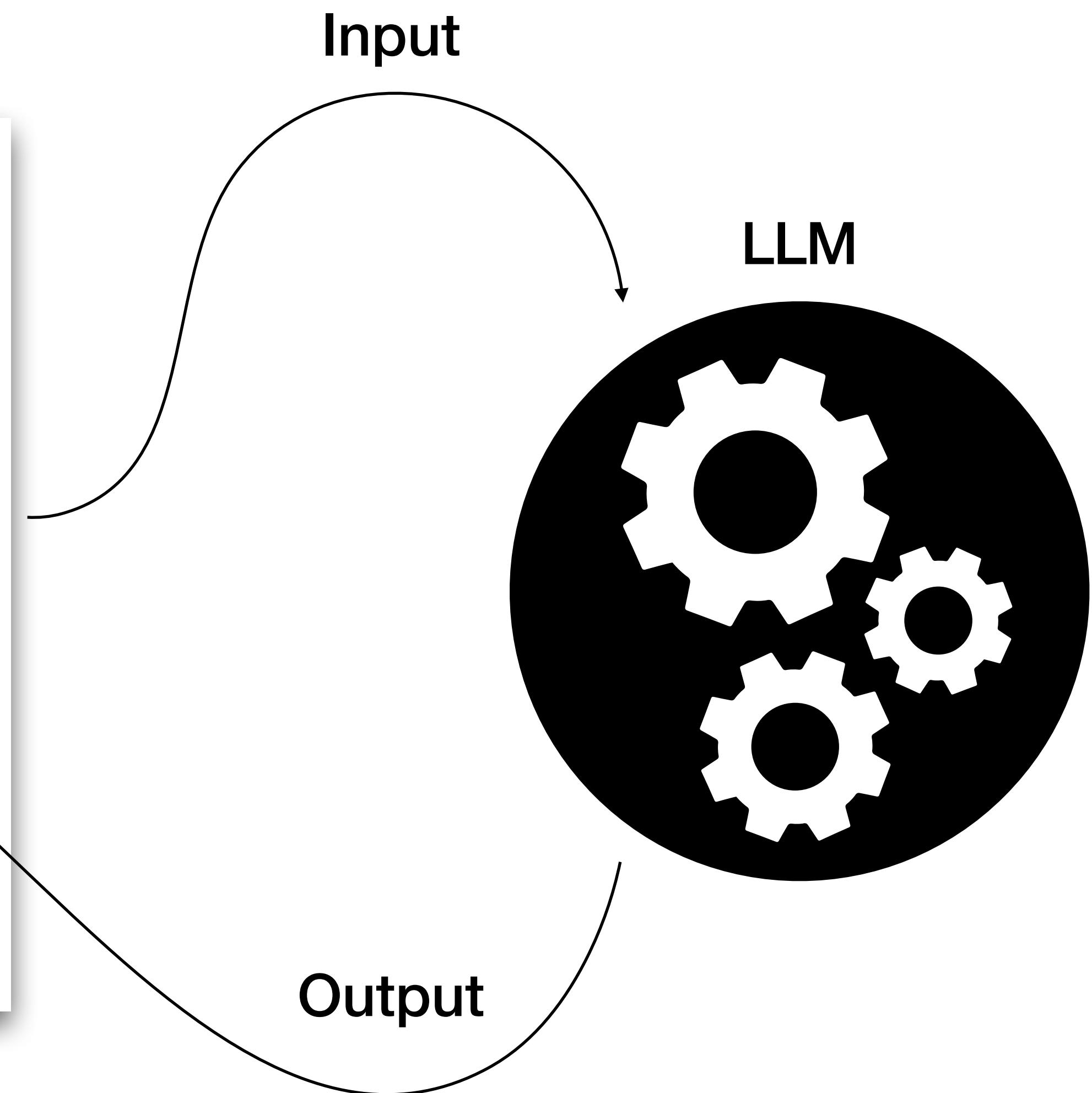
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many 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](#)  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 formulae 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

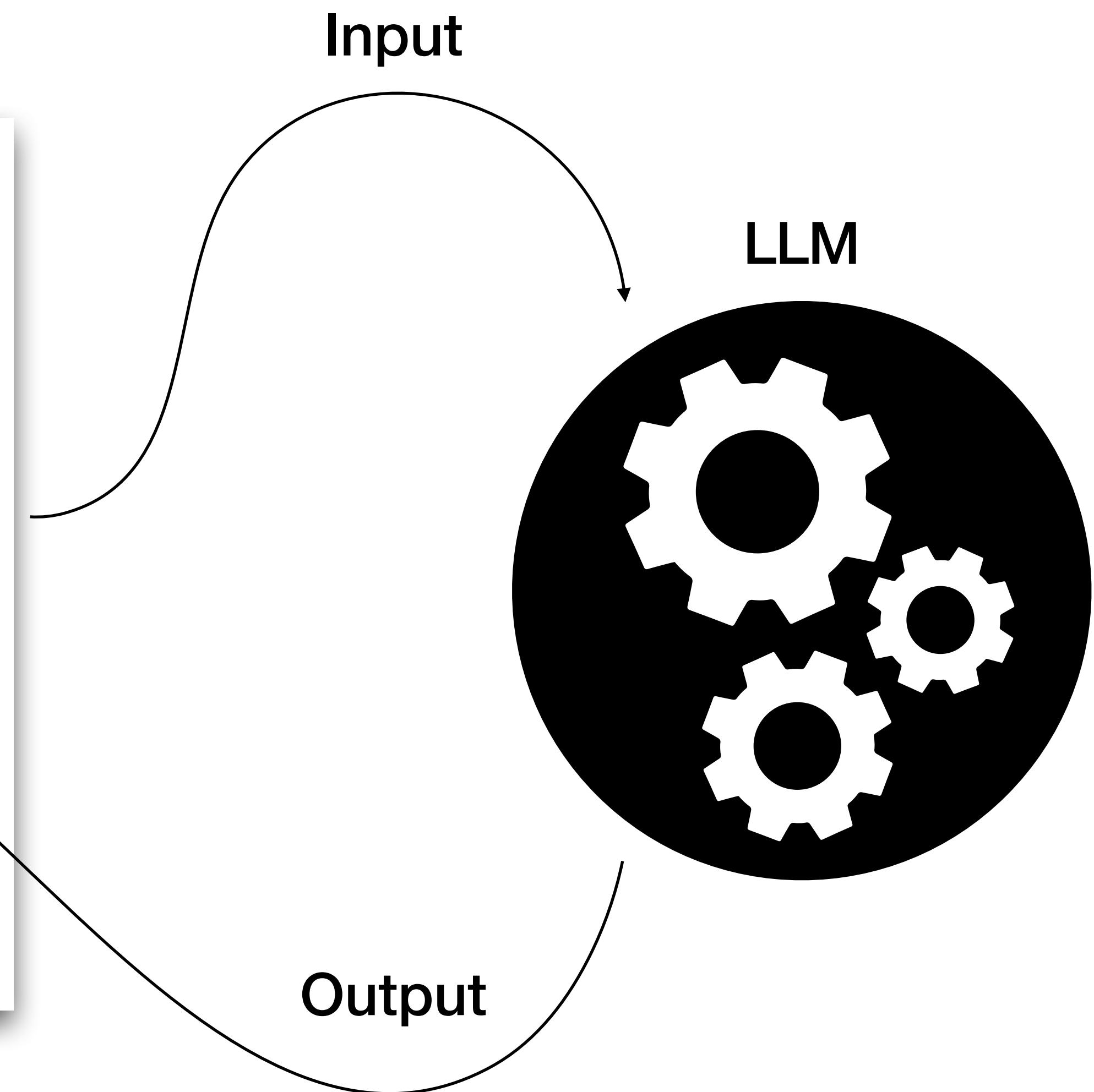
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many 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](#)  a translation for the [German](#) "*es war einmal*" (literally "it was once") and [Joseph Jacobs](#) in [English](#) translations and fairy tales.

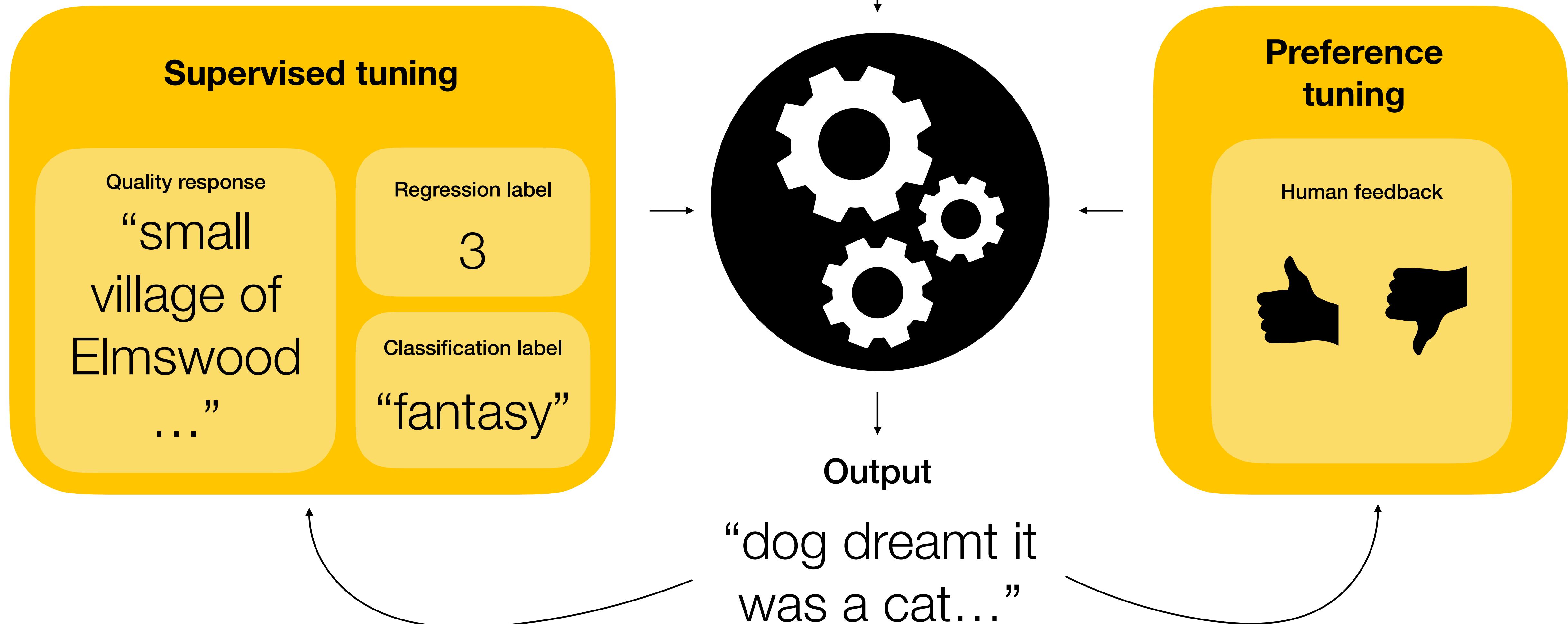
In *More English Fairy Tales*, Joseph Jacobs notes that:

"The opening formulae 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]

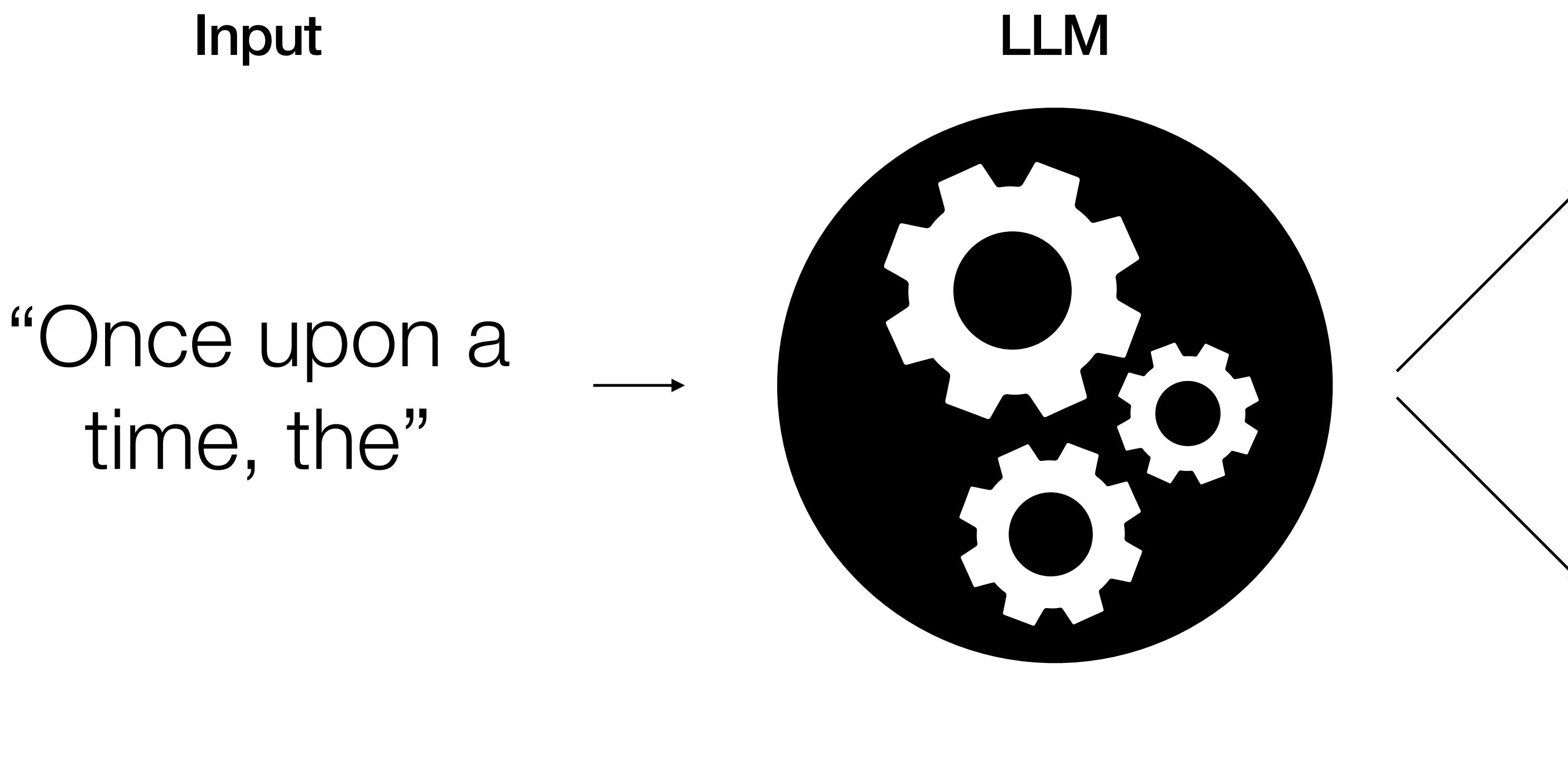
https://en.wikipedia.org/wiki/Once_upon_a_time



Fine-tuning

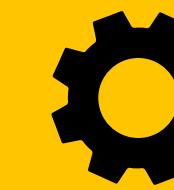


Two major applications



Text generation

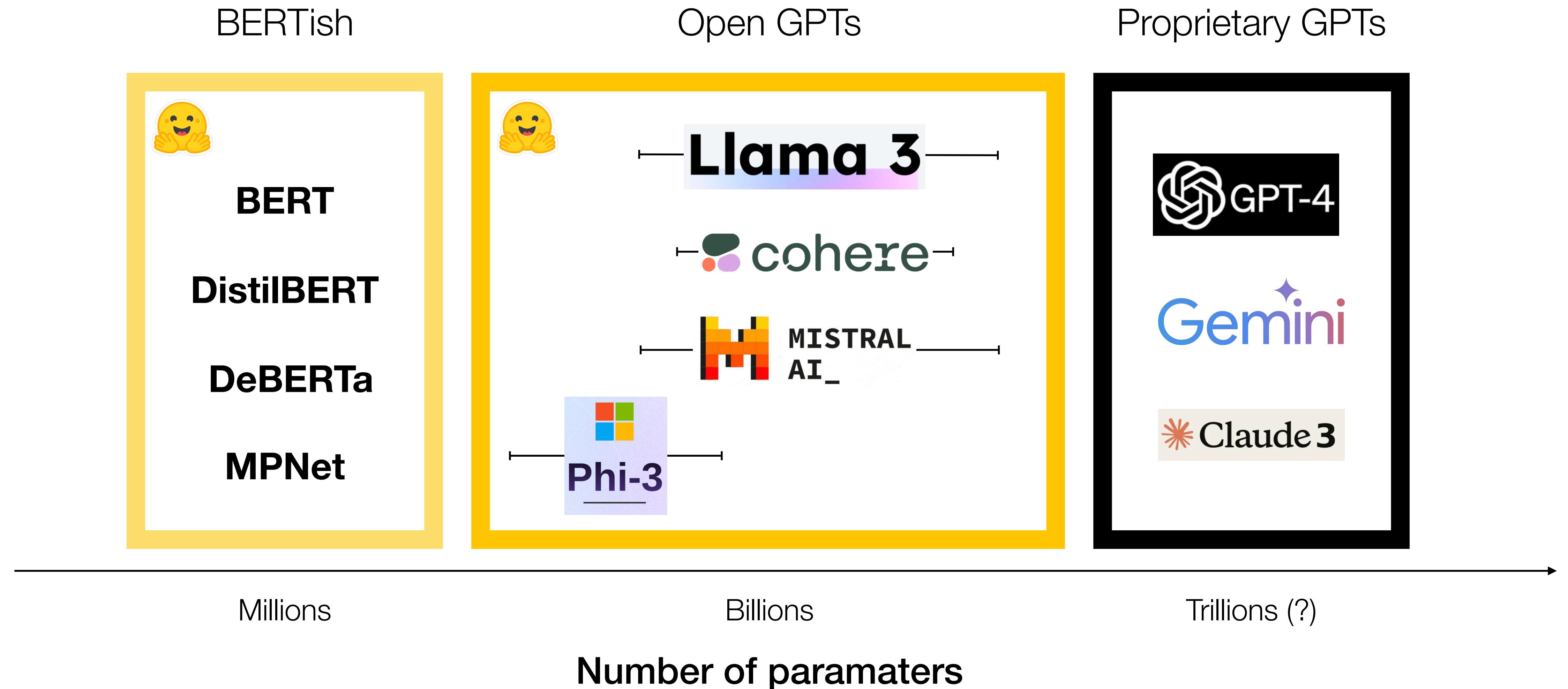
“small village of
Elmswood...”



Feature extraction

-.23, 1.23, .24,
-.12, .34, .32, ...

Models



Hugging Face is way more fun with friends and colleagues! 😊 [Join an organization](#)

[Dismiss this message](#)[Tasks](#) [Libraries](#) [Datasets](#) [Languages](#) [Licenses](#)[Other](#)[Filter Tasks by name](#)[Multimodal](#)[Image-Text-to-Text](#)[Visual Question Answering](#)[Document Question Answering](#)[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](#)**Models** 674,462[Filter by name](#)[Full-text search](#)[↑↓ Sort: Trending](#)[openbmb/MiniCPM-Llama3-V-2_5](#)[Visual Question Answering](#) • Updated about 5 hours ago • ↓ 14k • ❤ 552[microsoft/Phi-3-vision-128k-instruct](#)[Text Generation](#) • Updated 1 day ago • ↓ 12.4k • ❤ 469[mistralai/Mistral-7B-Instruct-v0.3](#)[Text Generation](#) • Updated 3 days ago • ↓ 16.7k • ❤ 377[meta-llama/Meta-Llama-3-8B](#)[Text Generation](#) • Updated 13 days ago • ↓ 978k • ❤ 4.19k[microsoft/Phi-3-medium-128k-instruct](#)[Text Generation](#) • Updated 3 days ago • ↓ 17.6k • ❤ 233[mistralai/Mistral-7B-v0.3](#)[Text Generation](#) • Updated 3 days ago • ↓ 19.1k • ❤ 144[CohereForAI/ay-a-23-8B](#)[Text Generation](#) • Updated 2 days ago • ↓ 6.63k • ❤ 138

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)	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1287	+5/-3	20156	OpenAI	Proprietary	2023/10
2	GPT-4-Turbo-2024-04-09	1252	+3/-3	62203	OpenAI	Proprietary	2023/12
2	GPT-4-1106-preview	1250	+3/-3	82286	OpenAI	Proprietary	2023/4
2	Gemini 1.5 Pro API-0409-Preview	1248	+3/-3	62929	Google	Proprietary	2023/11
3	Claude 3 Opus	1246	+2/-2	121218	Anthropic	Proprietary	2023/8
4	GPT-4-0125-preview	1244	+3/-3	76435	OpenAI	Proprietary	2023/12
7	Yi-Large-preview	1236	+4/-4	15671	01 AI	Proprietary	Unknown
8	Bard (Gemini Pro)	1208	+6/-7	12387	Google	Proprietary	Online
8	Llama-3-70b-Instruct	1203	+2/-2	129016	Meta	Llama 3 Community	2023/12
8	Claude 3 Sonnet	1199	+3/-2	97268	Anthropic	Proprietary	2023/8
10	Reka-Core-20240501	1195	+3/-3	37076	Reka AI	Proprietary	Unknown

Overall	Bitext Mining	Classification	Clustering	Pair Classification	Reranking	Retrieval	STS	Summarization	Retrieval w/Instructions
English	Chinese	French	Polish						
Overall MTEB English leaderboard 🎉									
<ul style="list-style-type: none"> 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
MPNet
MPNet-personality



Feature extraction
(fine-tuning)

Open GPTs

Phi-3-mini



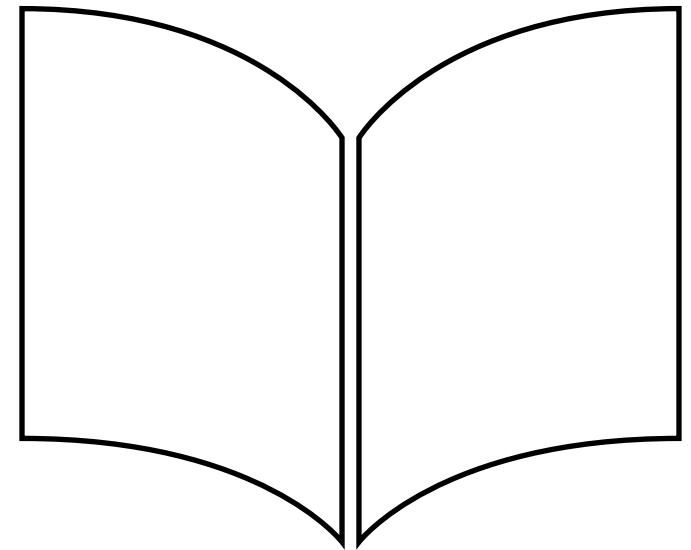
Text-generation
(in-context learning)

Applications



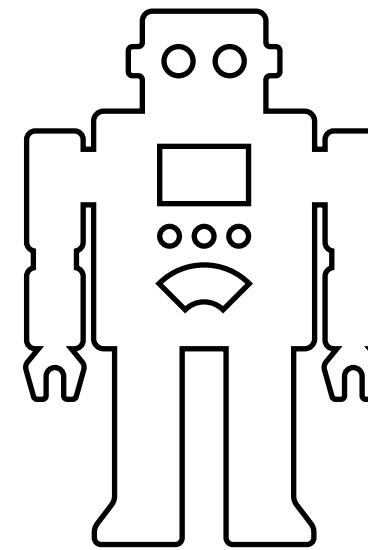
Personality

Predict the relatedness among survey items and between items and constructs based on **feature extraction**



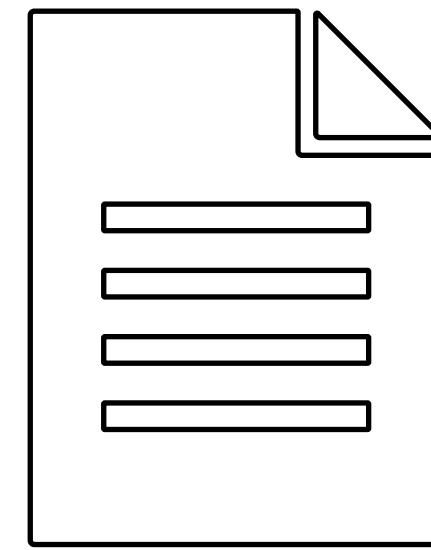
Media bias

Predict media bias based on headline texts using **text generation** and **fine-tuning**



Synthetic samples

Generate human-like responses using **text generation** to simulate human participants



Querying articles

Process articles using **in-context learning** and **text generation** to extract and summarize information

Discussion

- Gather in small groups (2-3 people)
- Identify applications of LLMs
- Report 1-3 applications

