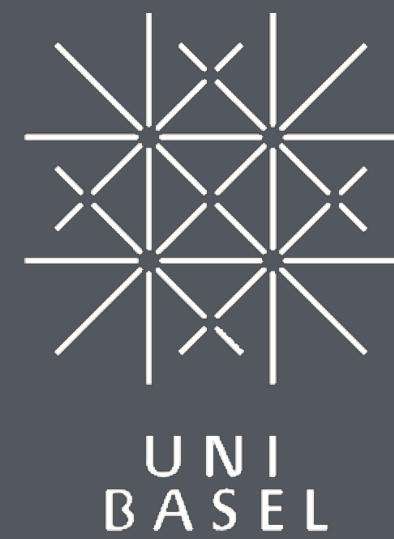


# Introduction to LLMs for science of science research @ LMU 2025

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



# Software stack



+



+





Zak-Hussain

/ LLM4SciSci

 Type / to search[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

LLM4SciSci

Public

[Unwatch](#) 2[Fork](#) 0[Star](#) 0[main](#) ▾[1 Branch](#) [0 Tags](#) Go to file[Add file](#) ▾[Code](#) ▾**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](#)

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****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

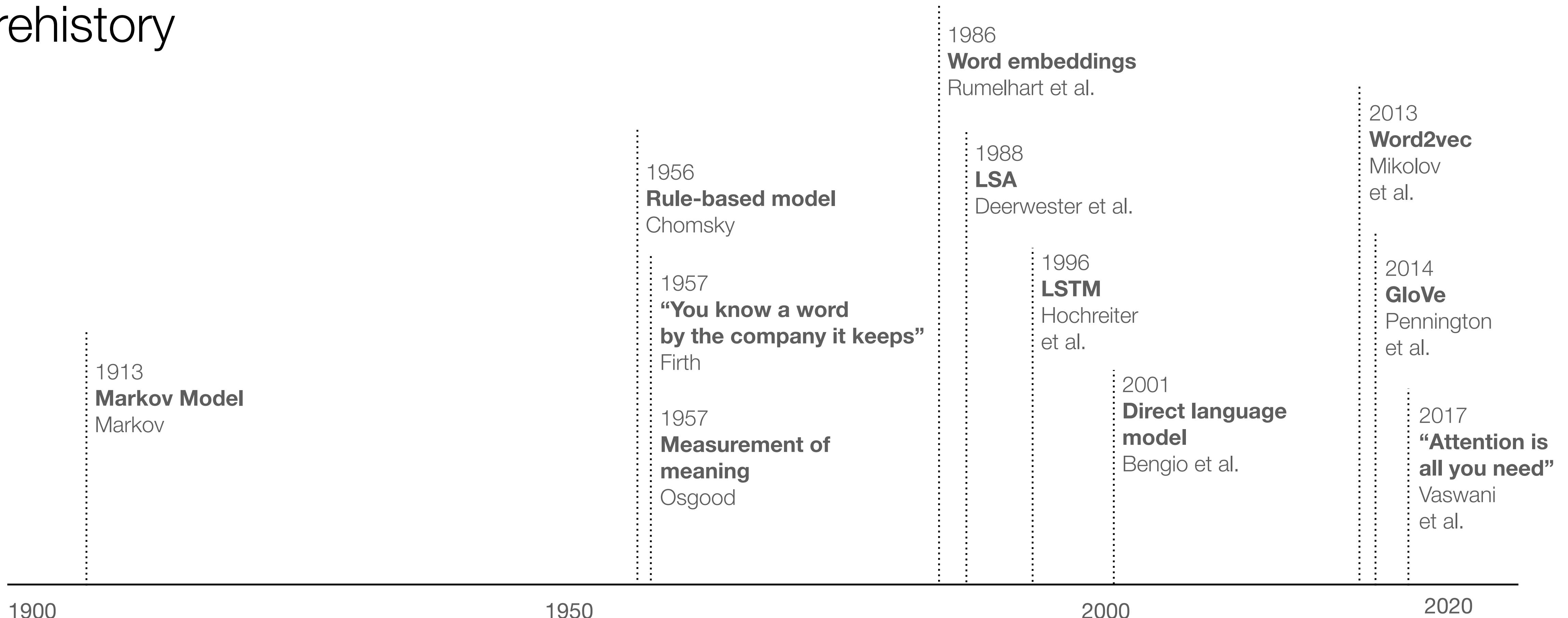


MAX PLANCK INSTITUTE  
FOR HUMAN DEVELOPMENT



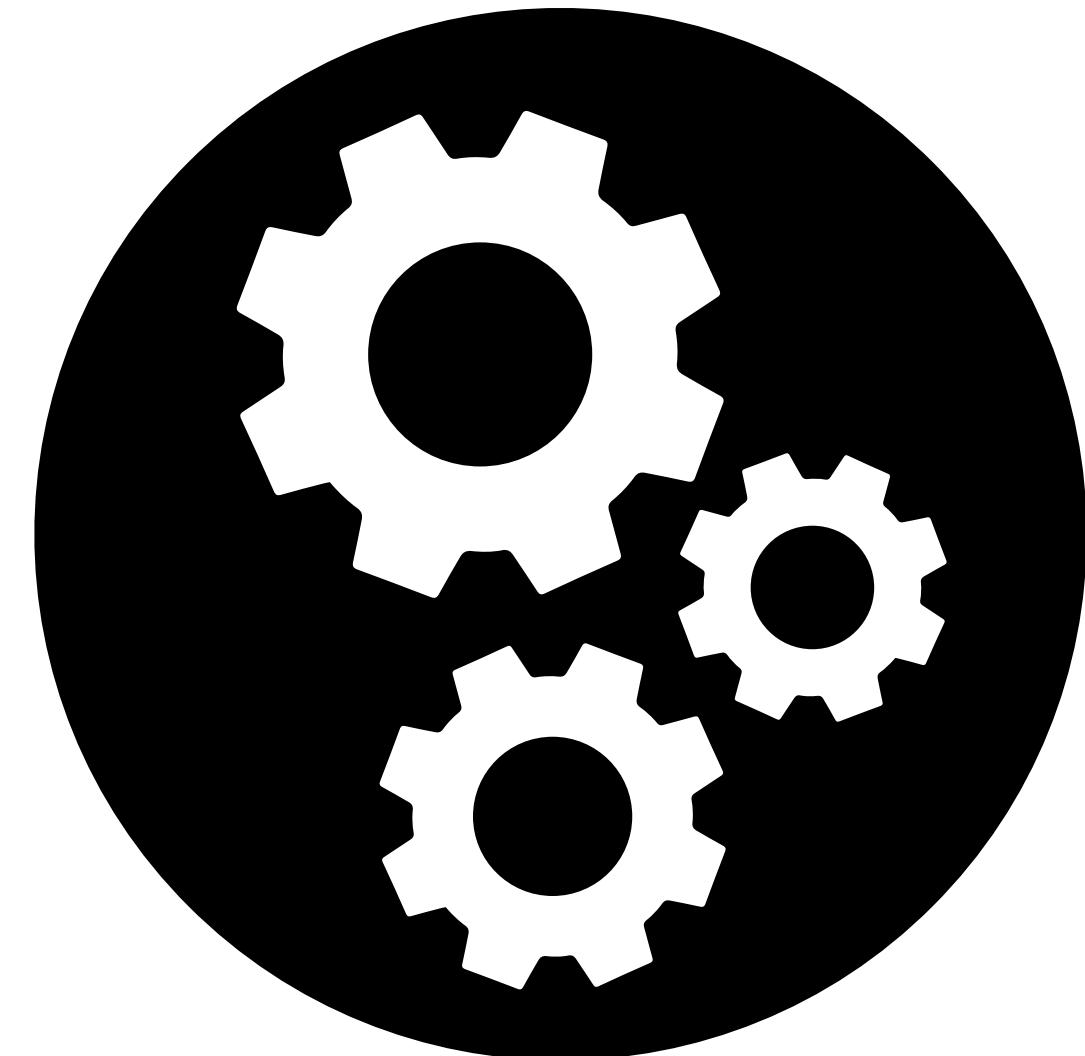
# Language models

## Prehistory

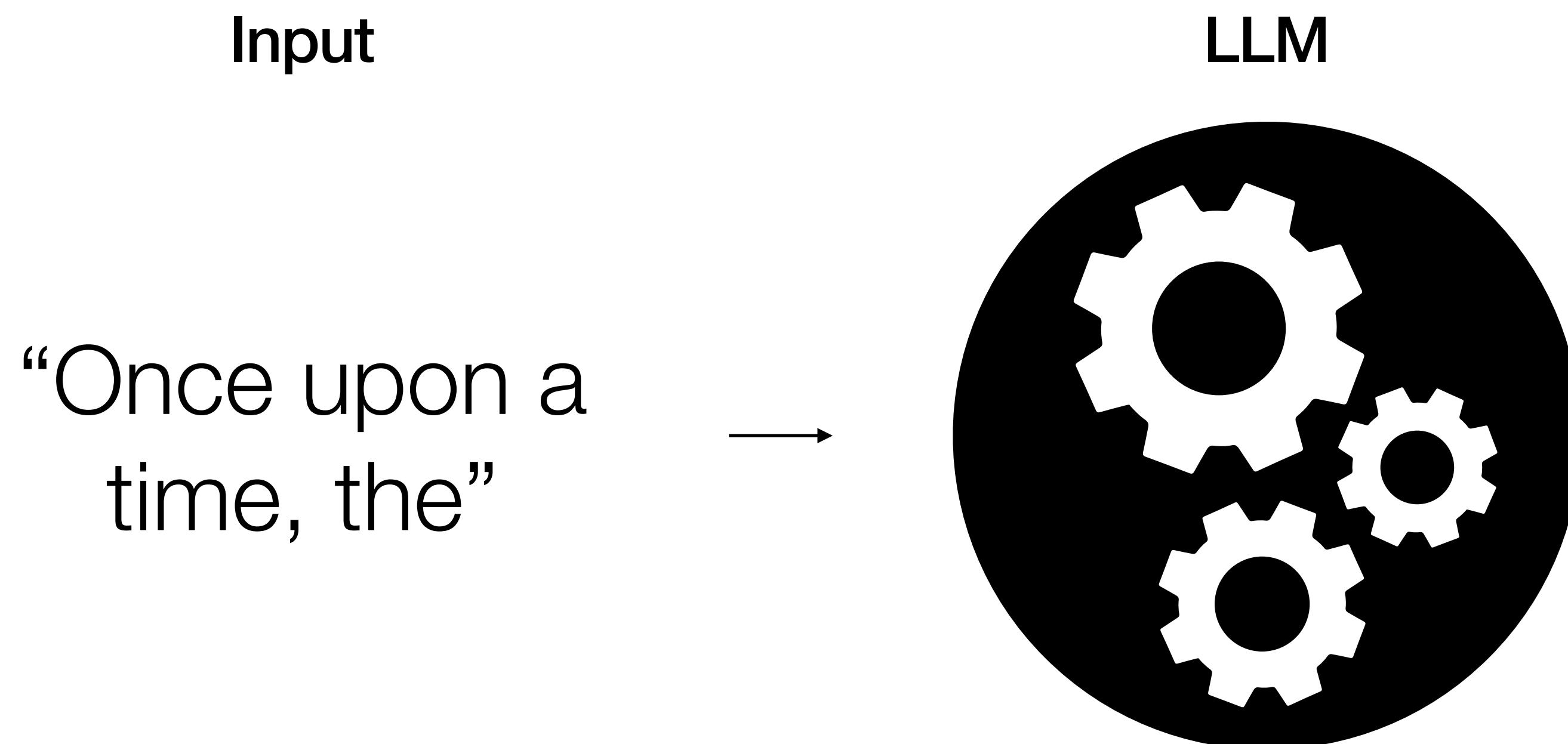


# LLMs as mechanisms

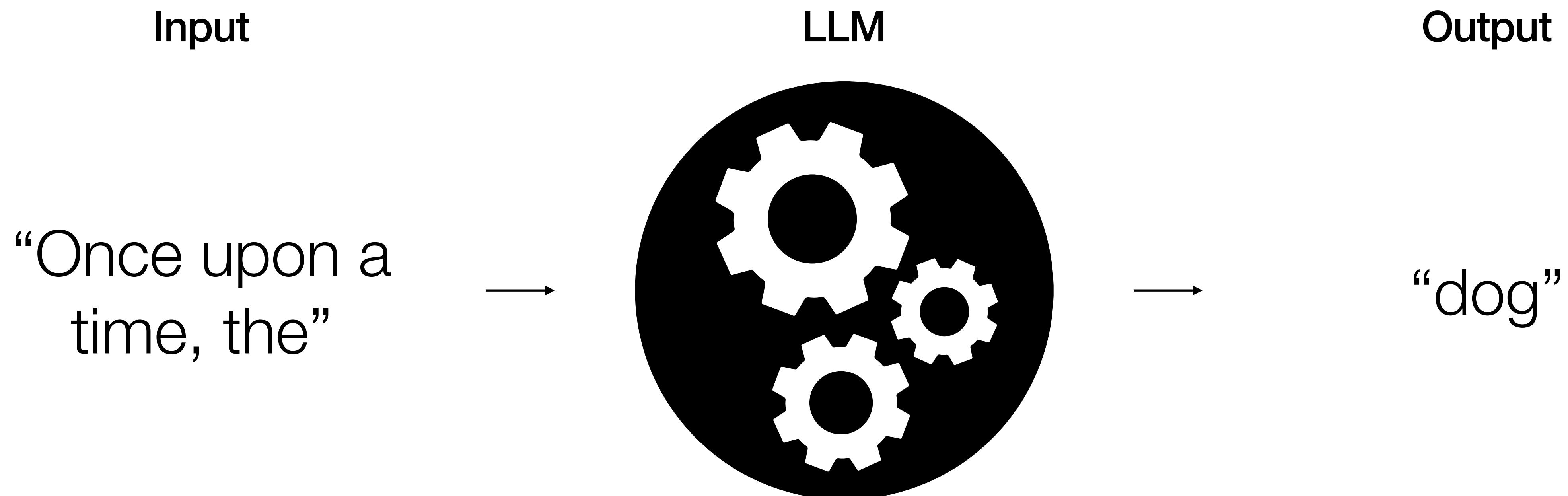
LLM



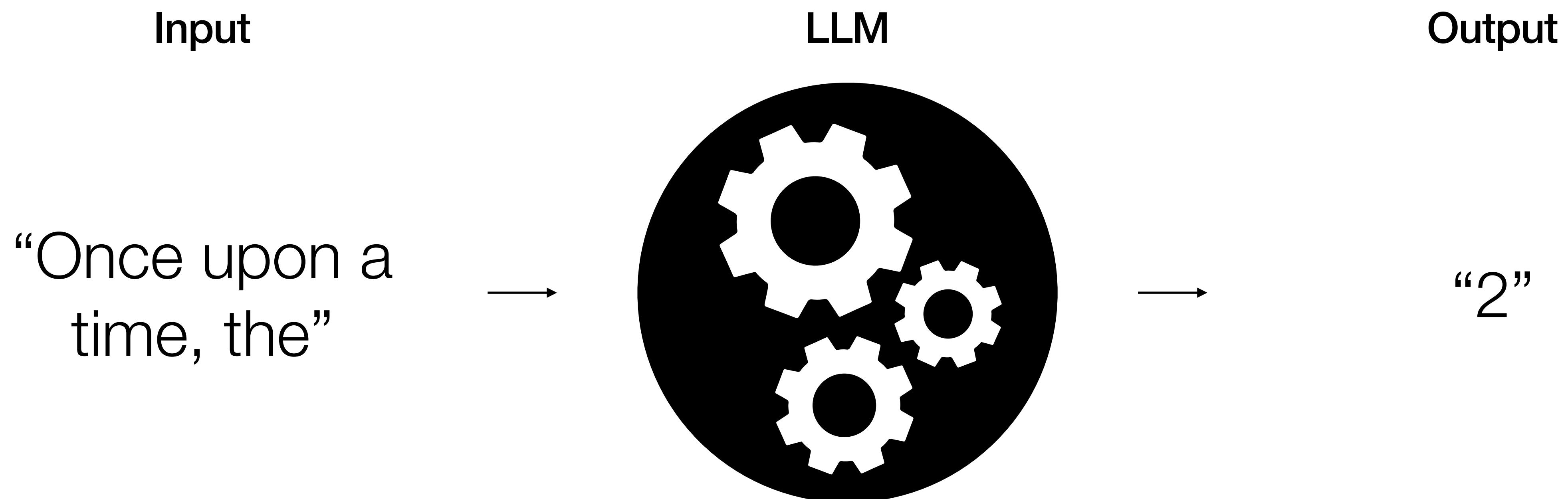
# LLMs as mechanisms



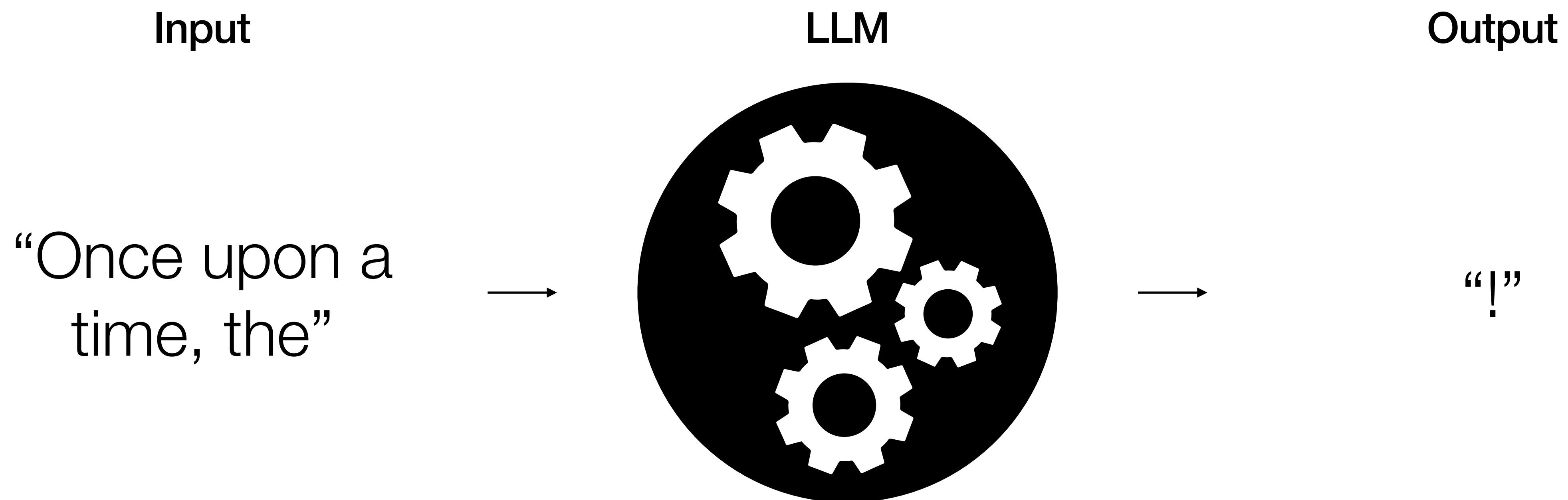
# LLMs as mechanisms



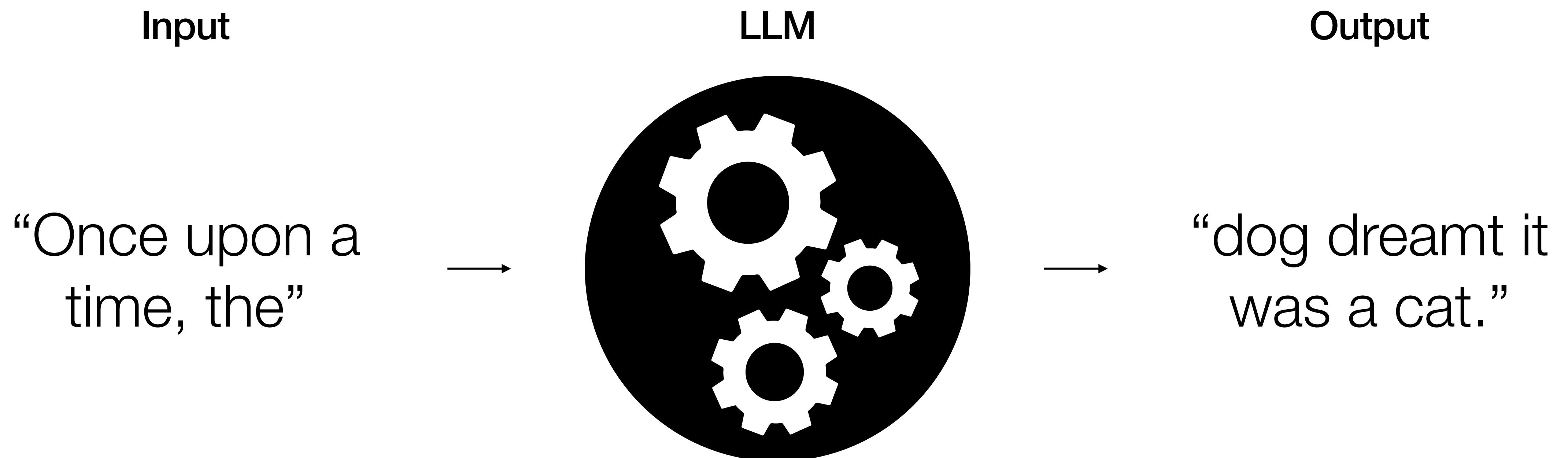
# LLMs as mechanisms



# LLMs as mechanisms



# LLMs as mechanisms



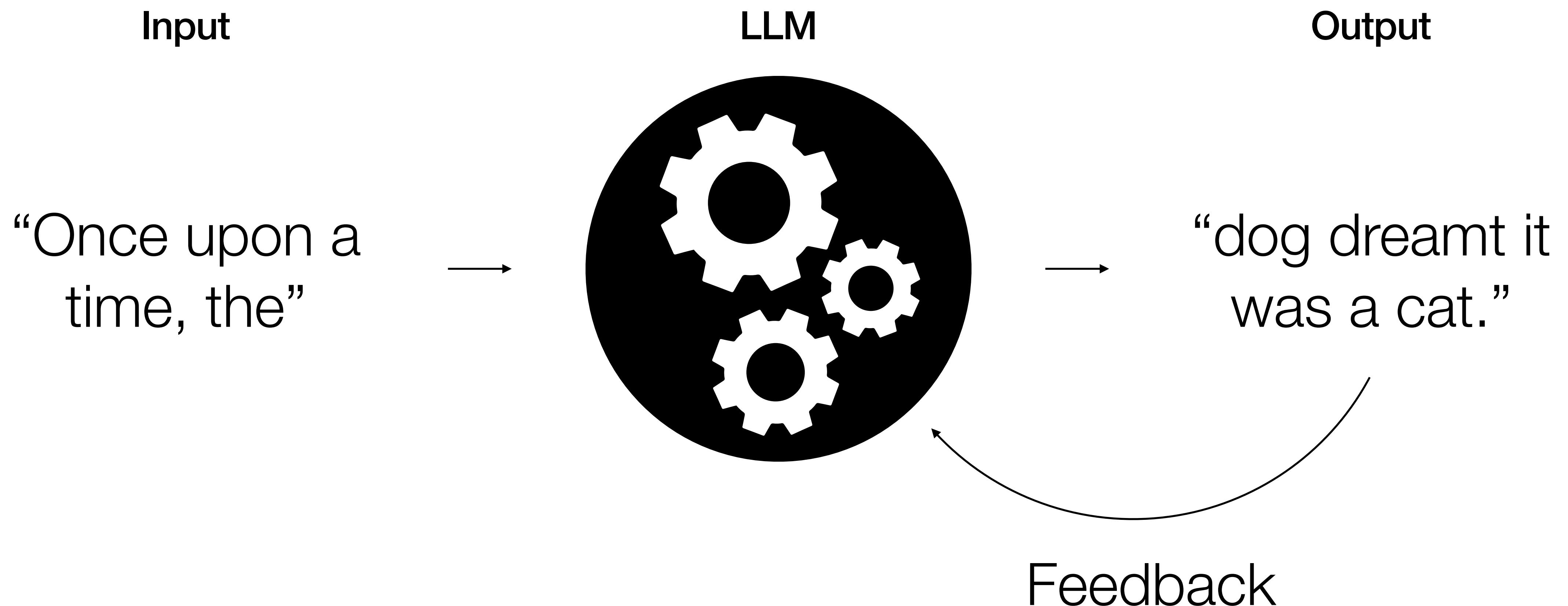
Once upon a time, the



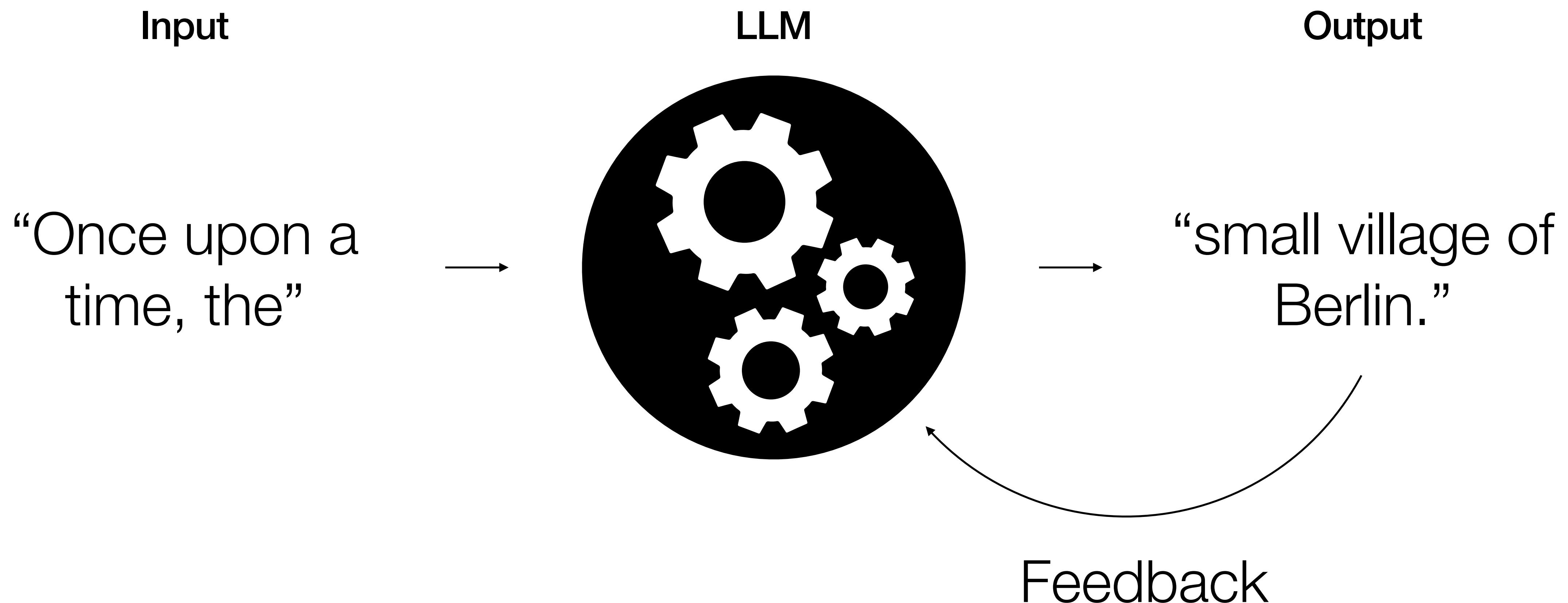
Once upon a time, the small village of Elmswood was nestled in a lush valley surrounded by towering mountains. The villagers lived peacefully, their days marked by the rhythms of nature and the changing seasons. However, everything changed when a mysterious old man arrived, carrying with him a locked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



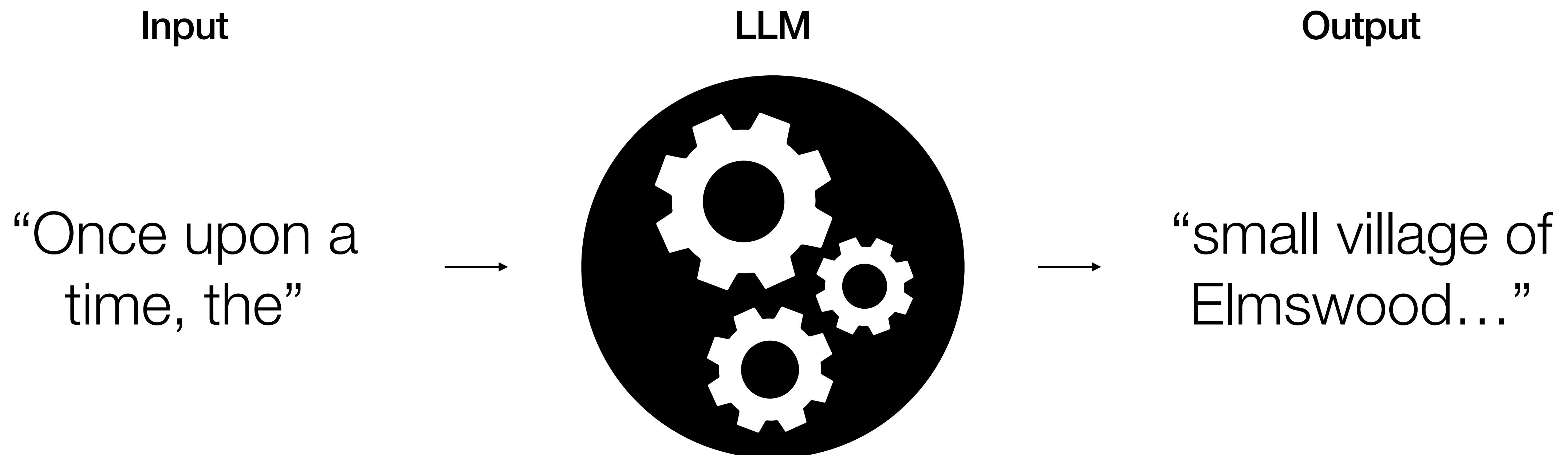
# LLMs as **trained** mechanisms



# 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

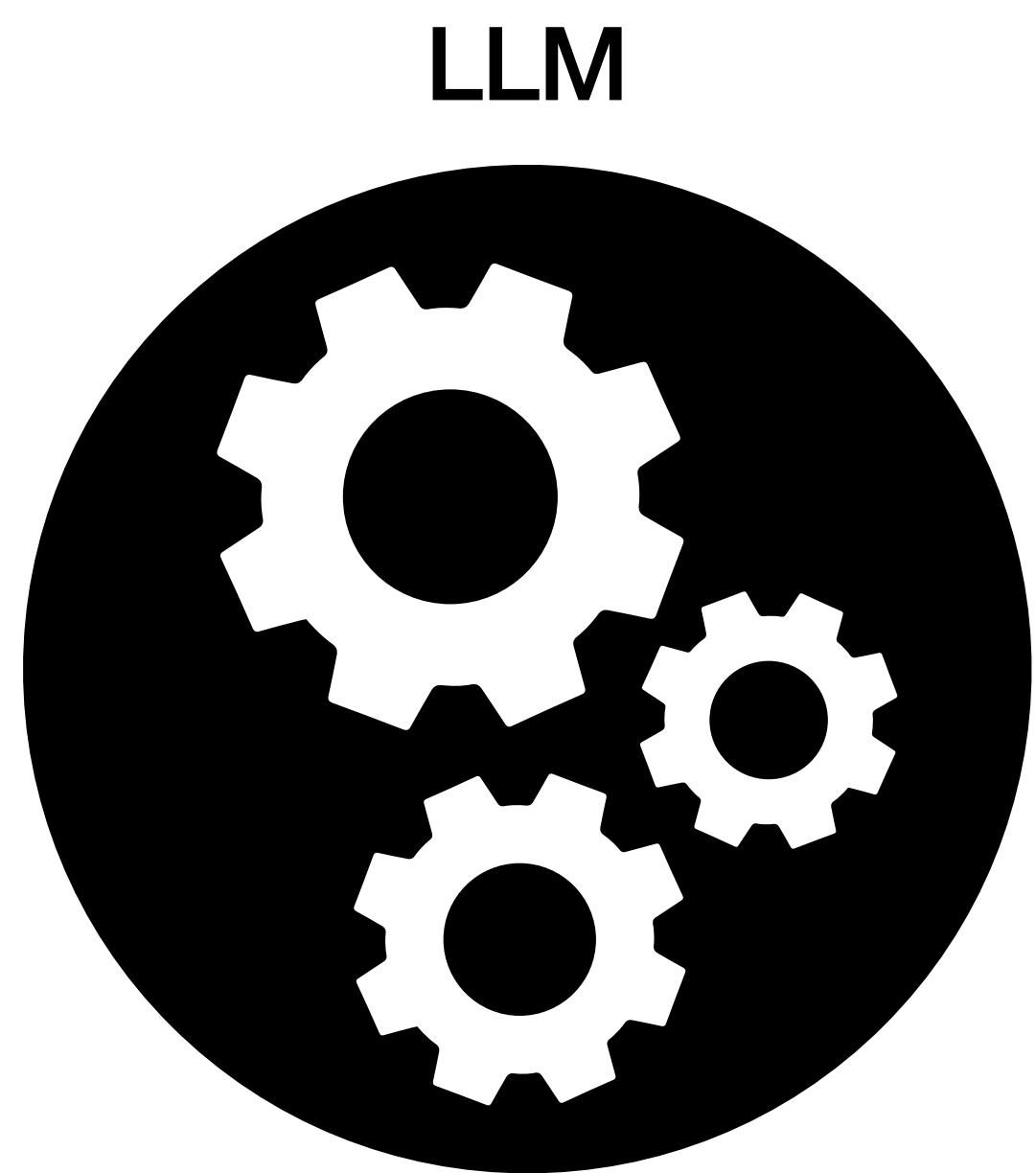
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many 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."<sup>[1]</sup>

[https://en.wikipedia.org/wiki/Once\\_upon\\_a\\_time](https://en.wikipedia.org/wiki/Once_upon_a_time)



# Masked/next token prediction

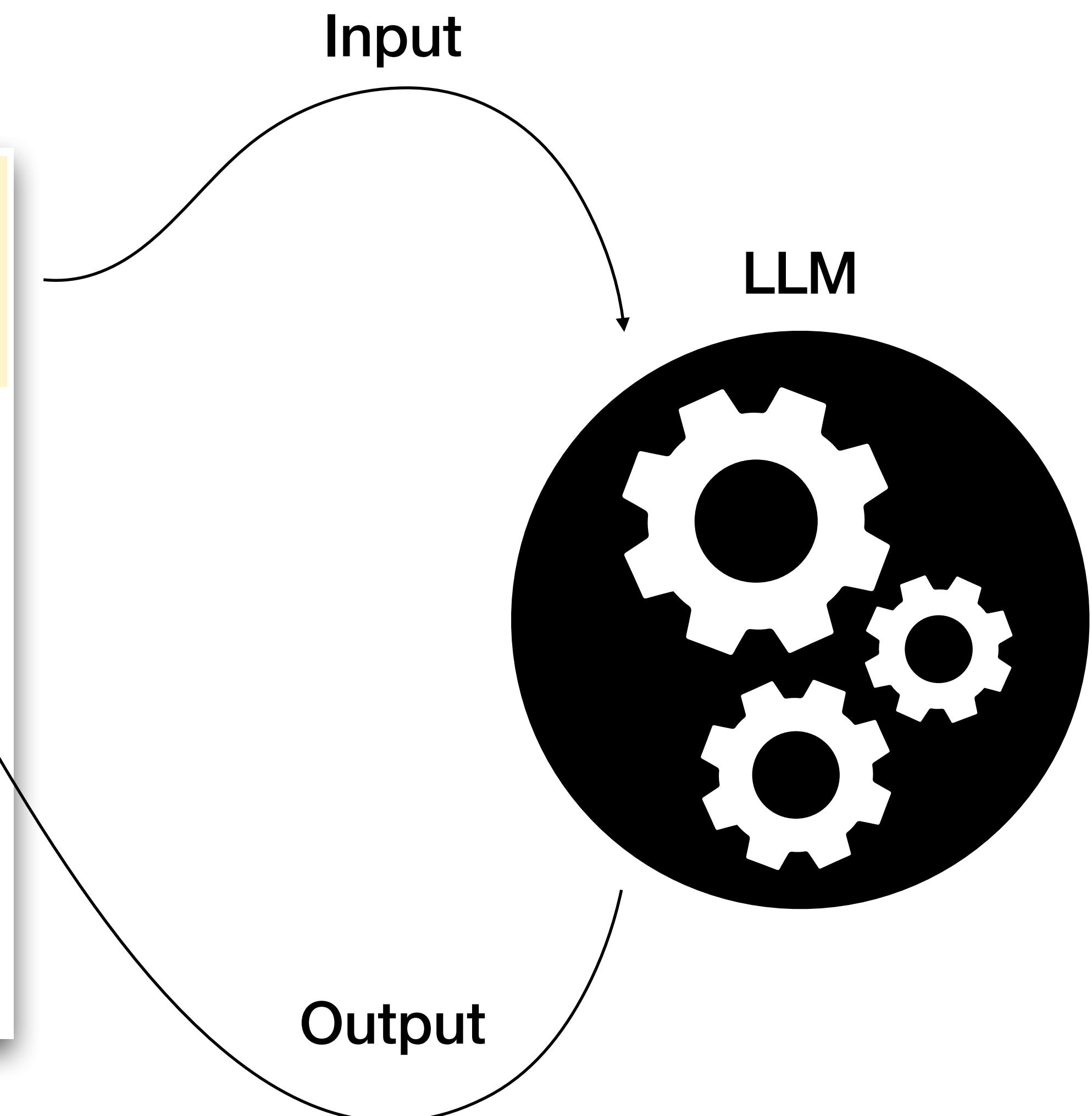
"Once upon a time" is a stock phrase used to introduce a narrative of past events, typically in fairy tales and folk tales. It has been used in some form since at least 1380 (according to the *Oxford English Dictionary*) in [REDACTED] in the English language and has started many narratives since 1600. 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."<sup>[1]</sup>

[https://en.wikipedia.org/wiki/Once\\_upon\\_a\\_time](https://en.wikipedia.org/wiki/Once_upon_a_time)



# Masked/next token prediction

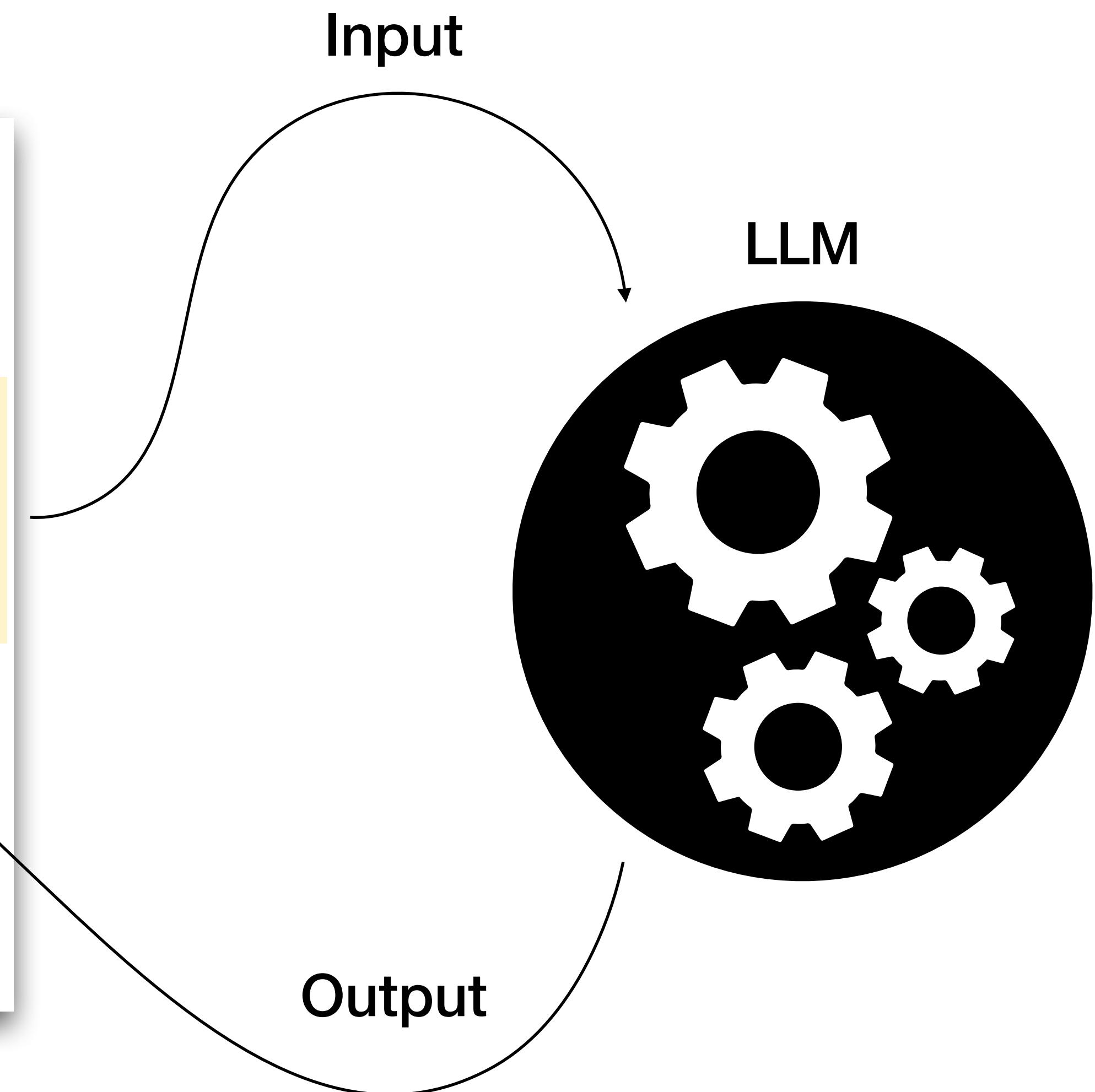
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many 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."<sup>[1]</sup>

[https://en.wikipedia.org/wiki/Once\\_upon\\_a\\_time](https://en.wikipedia.org/wiki/Once_upon_a_time)



# Masked/next token prediction

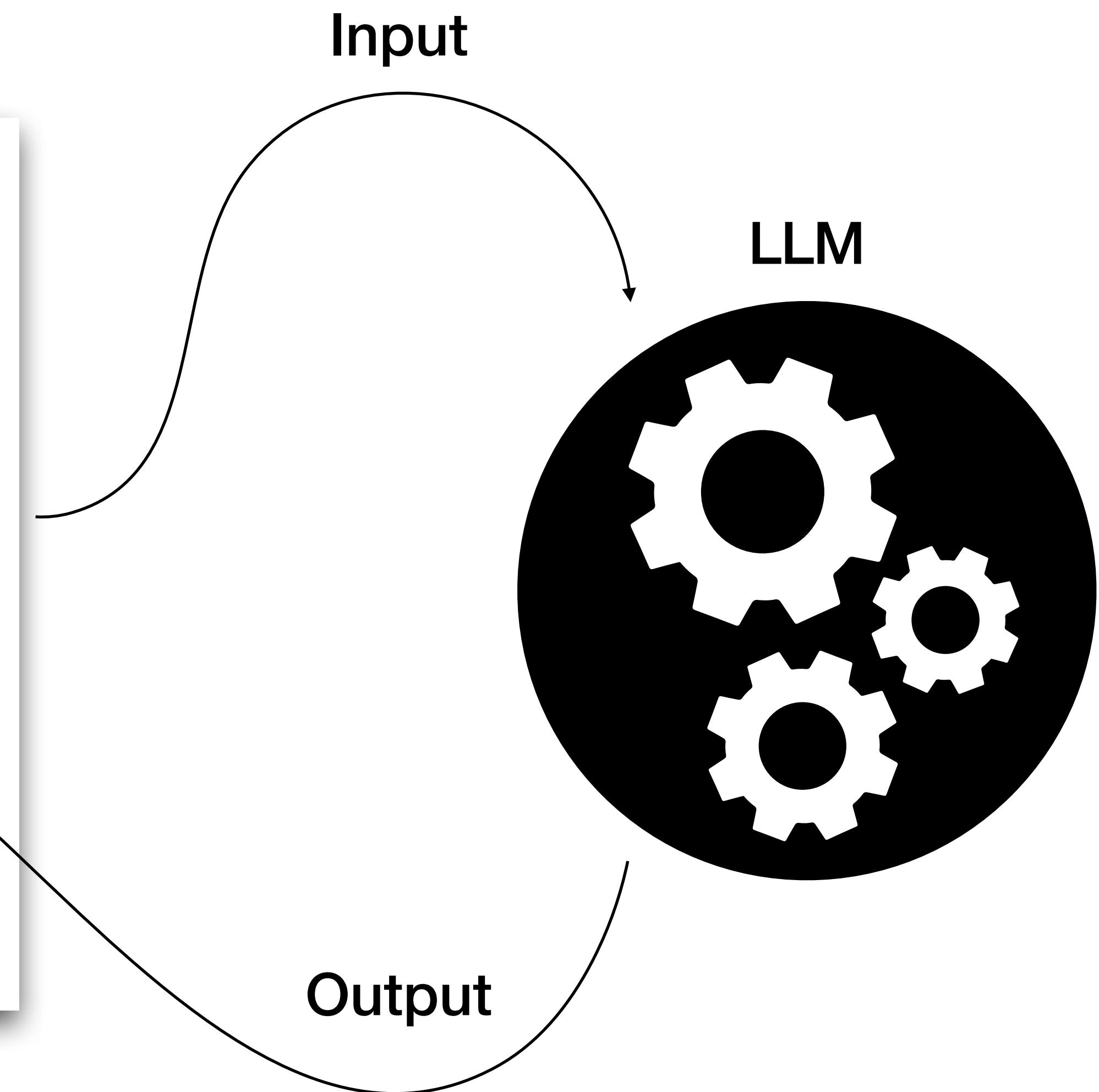
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many 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."<sup>[1]</sup>

[https://en.wikipedia.org/wiki/Once\\_upon\\_a\\_time](https://en.wikipedia.org/wiki/Once_upon_a_time)



# Masked/next token prediction

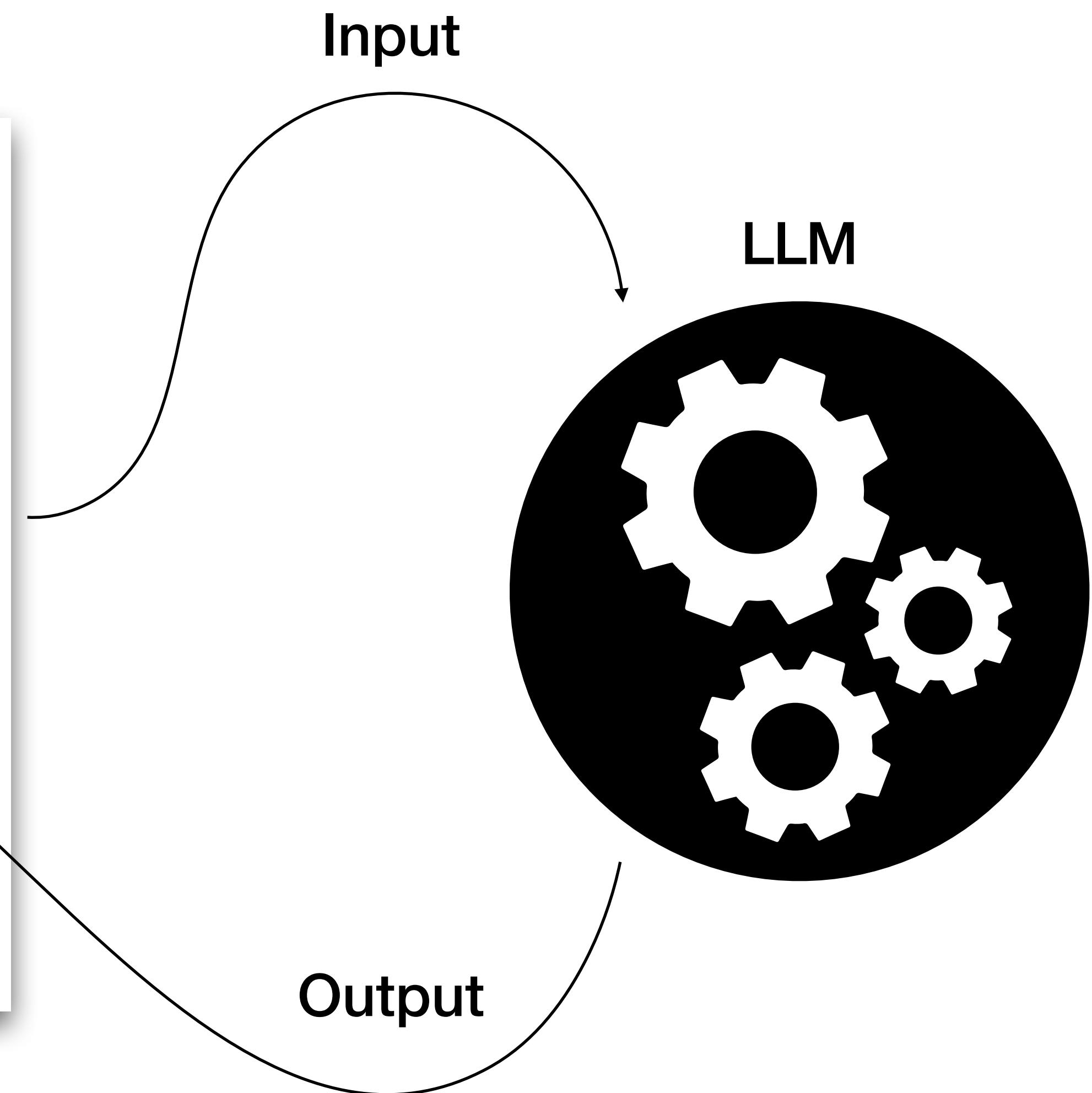
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many 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."<sup>[1]</sup>

[https://en.wikipedia.org/wiki/Once\\_upon\\_a\\_time](https://en.wikipedia.org/wiki/Once_upon_a_time)



# Masked/next token prediction

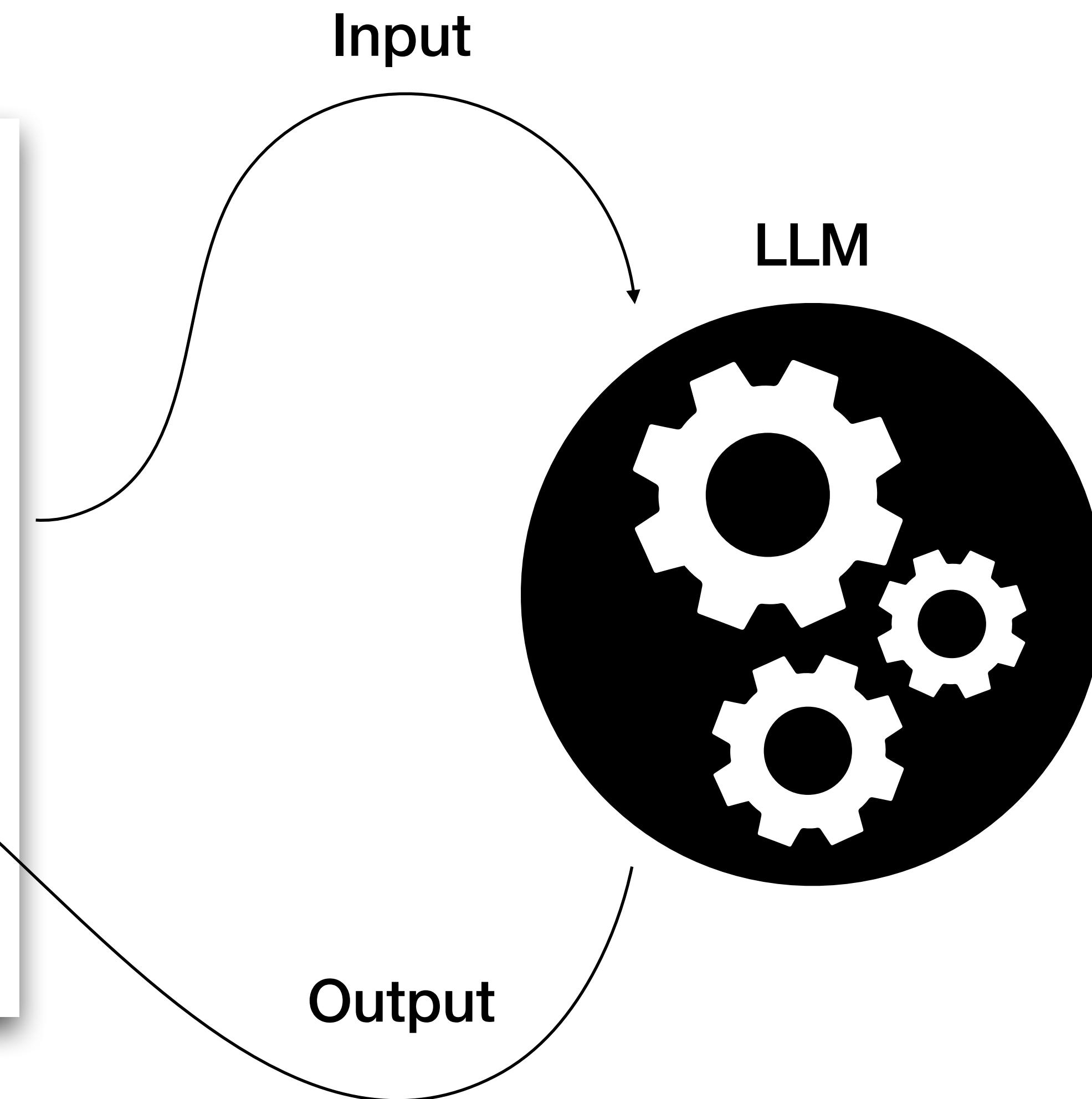
"Once upon a time" is a [stock phrase](#) used to introduce a narrative of past events, typically in [fairy tales](#) and folk tales. It has been used in some form since at least 1380 (according to the [Oxford English Dictionary](#)) in [storytelling](#) in the [English language](#) and has started many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

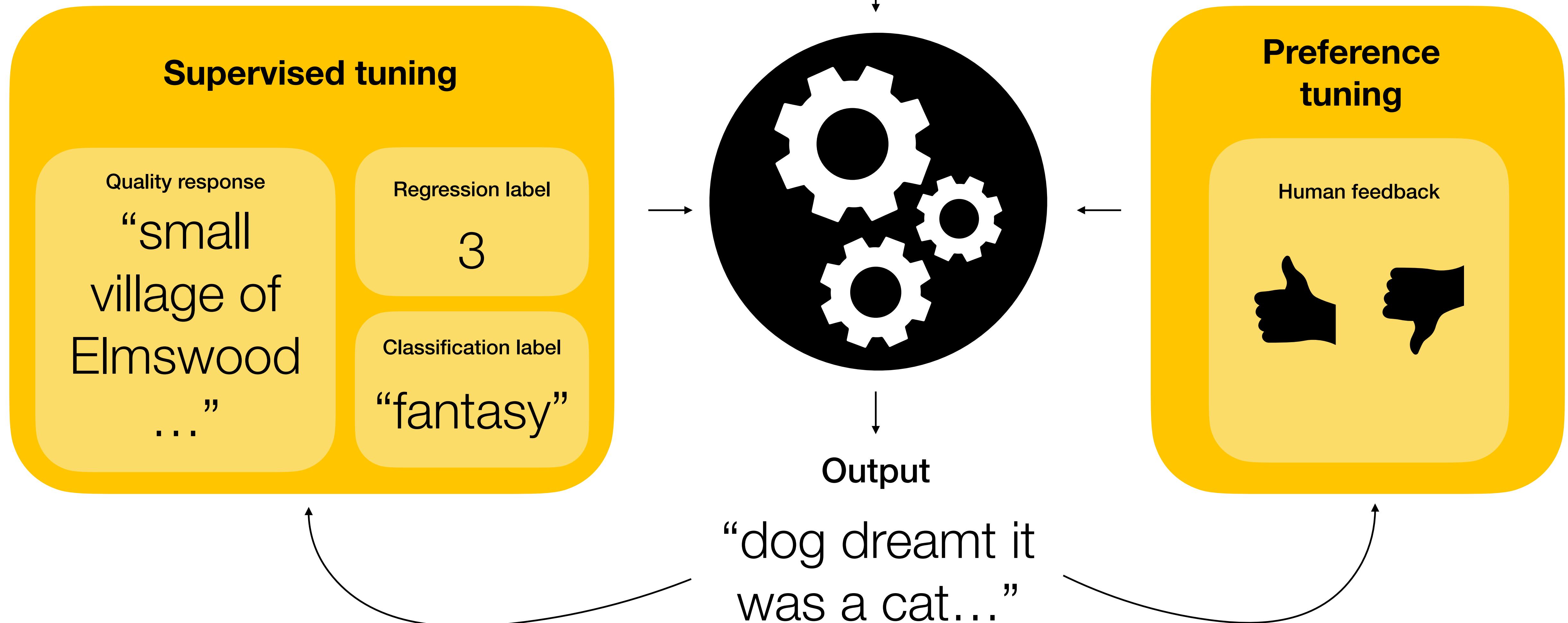
In *More English Fairy Tales*, Joseph Jacobs notes that:

"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."<sup>[1]</sup>

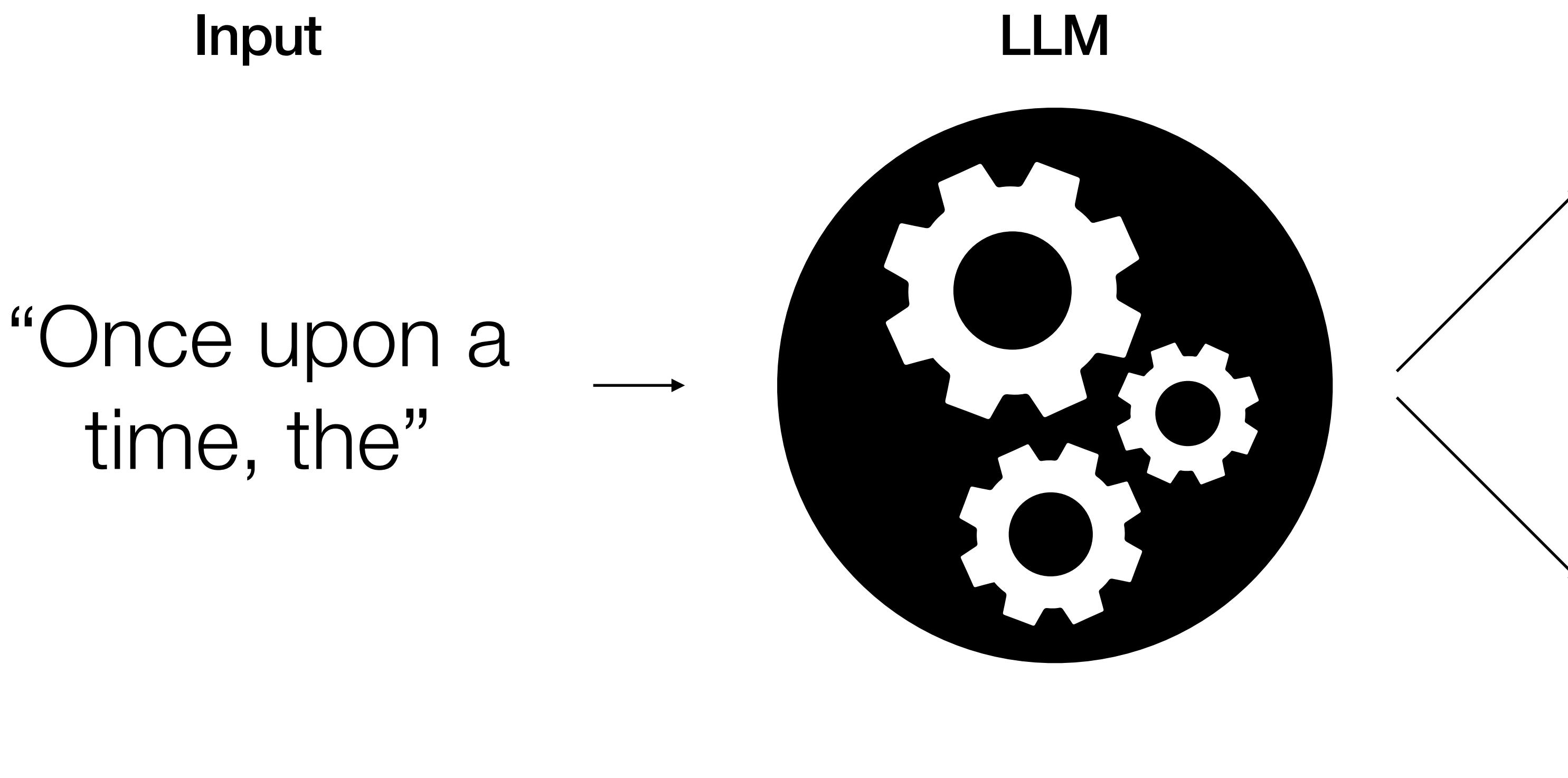
[https://en.wikipedia.org/wiki/Once\\_upon\\_a\\_time](https://en.wikipedia.org/wiki/Once_upon_a_time)



# Fine-tuning



# Two major applications



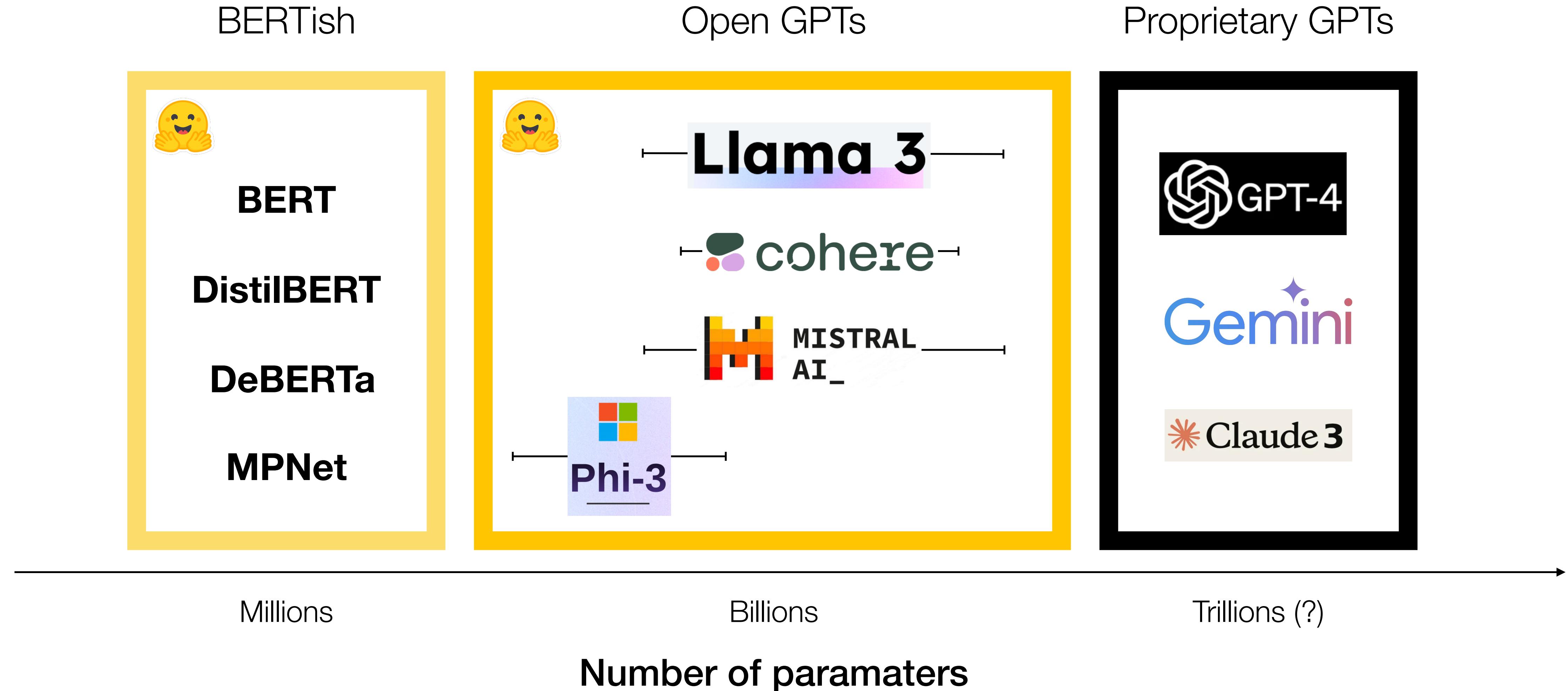
## Text generation

“small village of  
Elmswood...”

## Feature extraction

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

# Models



[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	<a href="#">Gemini-Exp-1206</a>	1374	+4/-5	20227	Google	Proprietary
1	1	<a href="#">ChatGPT-4o-latest...(2024-11-20)</a>	1365	+4/-3	33383	OpenAI	Proprietary
1	4	<a href="#">Gemini-2.0-Flash-Thinking-Exp-1219</a>	1364	+5/-6	15728	Google	Proprietary
2	4	<a href="#">Gemini-2.0-Flash-Exp</a>	1357	+6/-4	19030	Google	Proprietary
3	1	<a href="#">o1-2024-12-17</a>	1351	+7/-7	7289	OpenAI	Proprietary
6	4	<a href="#">o1-preview</a>	1335	+4/-4	33194	OpenAI	Proprietary
7	7	<a href="#">DeepSeek-V3</a>	1319	+6/-6	10510	DeepSeek	DeepSeek
7	10	<a href="#">Step-2-16K-Exp</a>	1305	+8/-9	3374	StepFun	Proprietary

Overall	Bitext Mining	Classification	Clustering	Pair Classification	Reranking	Retrieval	STS	Summarization	Retrieval w/Instructions
English	Chinese	French	Polish						
<b>Overall MTEB English leaderboard 🎉</b>									
<ul style="list-style-type: none"> <li><b>Metric:</b> Various, refer to task tabs</li> <li><b>Languages:</b> English</li> </ul>									
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	<a href="#">NV-Embed-v1</a>					69.32	87.35	52.8	
2	<a href="#">voyage-large-2-instruct</a>			1024	16000	68.28	81.49	53.35	
3	<a href="#">SFR-Embedding-Mistral</a>	7111	26.49	4096	32768	67.56	78.33	51.67	
4	<a href="#">gte-Qwen1.5-7B-instruct</a>	7099	26.45	4096	32768	67.34	79.6	55.83	
5	<a href="#">voyage-lite-02-instruct</a>	1220	4.54	1024	4000	67.13	79.25	52.42	
6	<a href="#">GritLM-7B</a>	7242	26.98	4096	32768	66.76	79.46	50.61	
7	<a href="#">e5-mistral-7b-instruct</a>	7111	26.49	4096	32768	66.63	78.47	50.26	
8	<a href="#">google-gecko.text-embedding-p</a>	1200	4.47	768	2048	66.31	81.17	47.48	
9	<a href="#">SE_v1</a>					65.66	76.8	47.38	
10	<a href="#">GritLM-8x7B</a>	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 Check for updates

**A gender perspective on the global migration of scholars**

Xinyi Zhao<sup>a,b,1</sup>, Aliakbar Akbaritabar<sup>a</sup>, Ridhi Kashyap<sup>b,c</sup>, and Emilio Zagheni<sup>a</sup>

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

Although considerable progress toward gender equality in science has been made in recent decades, female researchers continue to face significant barriers in the academic labor market. International mobility has been increasingly recognized as a strategy for scientists to expand their professional networks, and that could help narrow the gender gap in academic careers. Using bibliometric data on over 33 million Scopus publications, we provide a global and dynamic view of gendered patterns of transnational scholarly mobility, as measured by volume, distance, diversity, and distribution, from 1998 to 2017. We find that, while female researchers continued to be underrepresented among internationally mobile researchers and migrated over shorter distances, this gender gap was narrowing at a faster rate than the gender gap in the population of general active researchers. Globally, the origin and destination countries of both female and male mobile researchers became increasingly diversified, which suggests that scholarly migration has become less skewed and more globalized. However, the range of origin and destination countries continued to be narrower for women than for men. While the United States remained the leading academic destination worldwide, the shares of both female and male scholarly inflows to that country declined from around 25% to 20% over the study period, partially due to the growing relevance of China. This study offers a cross-national measurement of gender inequality in global scholarly migration that is essential for promoting gender-equitable science policies and for monitoring the impact of such interventions.

global migration of scholars | gender gap | bibliometric data | science of science | feminization of global migration

Over the past 50 y, women have made enormous strides in scientific research, including in the fields of science, technology, engineering, and mathematics (STEM) (1, 2). Nonetheless, women continue to face a number of barriers to participation, recognition, and progression in the scientific arena (3–5). In the current era of globalization, international mobility is increasingly recognized as a key strategy for scientists seeking to participate in global scientific networks and collaborations and to advance their careers (6, 7). However, less attention has been paid to gender differences in international scholarly migration, especially on a global basis (3, 5, 8, 9). Our study considers the interplay between the globalization of scientific knowledge, the internationalization of academia, and gender inequalities in the academic labor market (10–12), with the aim of providing substantive support for policies that advance gender equality in academia.

While the population of female scientists and scholars has more than doubled since 1993 and a wide array of programs promoting gender equality in academia have been launched, gender disparities persist in nearly all facets of academia and sciences (8, 13). In 2016, women researchers held 41% of academic positions across the 28 countries of the European Union (EU-28). However, in many European countries, including in the Netherlands and Germany, women held fewer than one in five senior academic positions (13). Women are also underrepresented as researchers in Asian countries such as Japan, where they account for only approximately one in four full-time faculty members (14). Female researchers in the Global South are relatively “invisible” compared to those in the Global North (15), and their representation among researchers in Guinea (6%), Ethiopia (7.6%), and Mali (10.6%) shows more alarming gender disparities (16). While it is clear that the sciences and academia continue to be dominated by males at the global scale, there is also substantial variation in levels of gender inequality across countries. Unfortunately, unified and comprehensive statistics suitable for making cross-national comparisons of gender disparities in the sciences do not exist (17–19), let alone statistics on gender disparities in global brain circulation. The first goal of our study is to document cross-national trends in a systematic way.

Existing research that has considered the gender dimension in international scholarly migration has mainly focused on either emigrants from an origin country perspective (20, 21) or on immigrants from a destination country perspective (11, 22, 23). Although

1To whom correspondence may be addressed. Email: zhao@demogr.mpg.de or xinyi.zhao@st-hughs.ox.ac.uk.  
This article contains supporting information online at <http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2214664120>.  
Published February 27, 2023.

Statistical  
reasoning

Spatial  
search

Does the  
article use  
bibliometric  
analysis?

Cognitive  
neuro-  
science

Performance  
monitoring

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

1,124 titles, abstracts, etc.

PDF articles