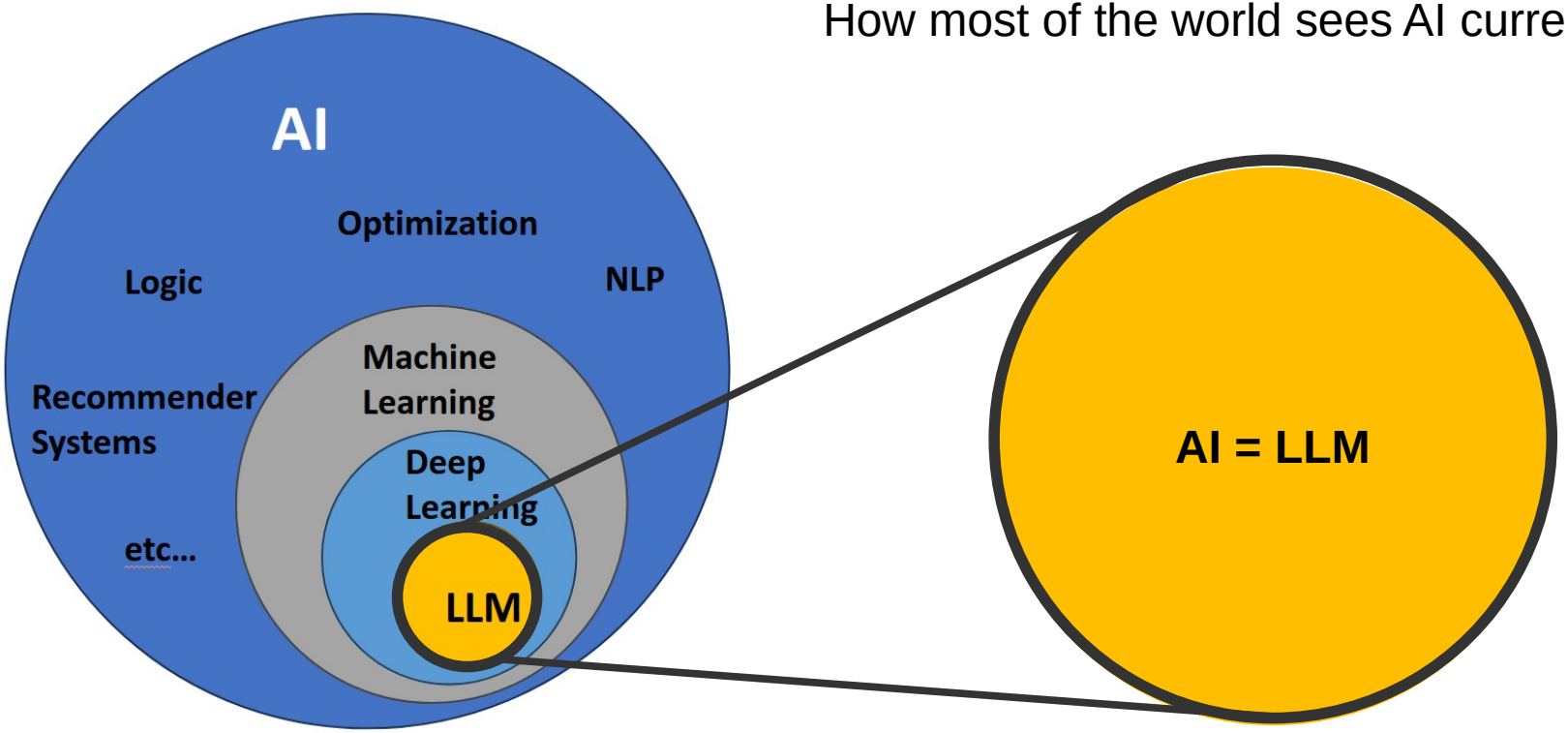
Large Language Models (LLMs)

week 3, 2024

The Big Picture

How most of the world sees AI currently:



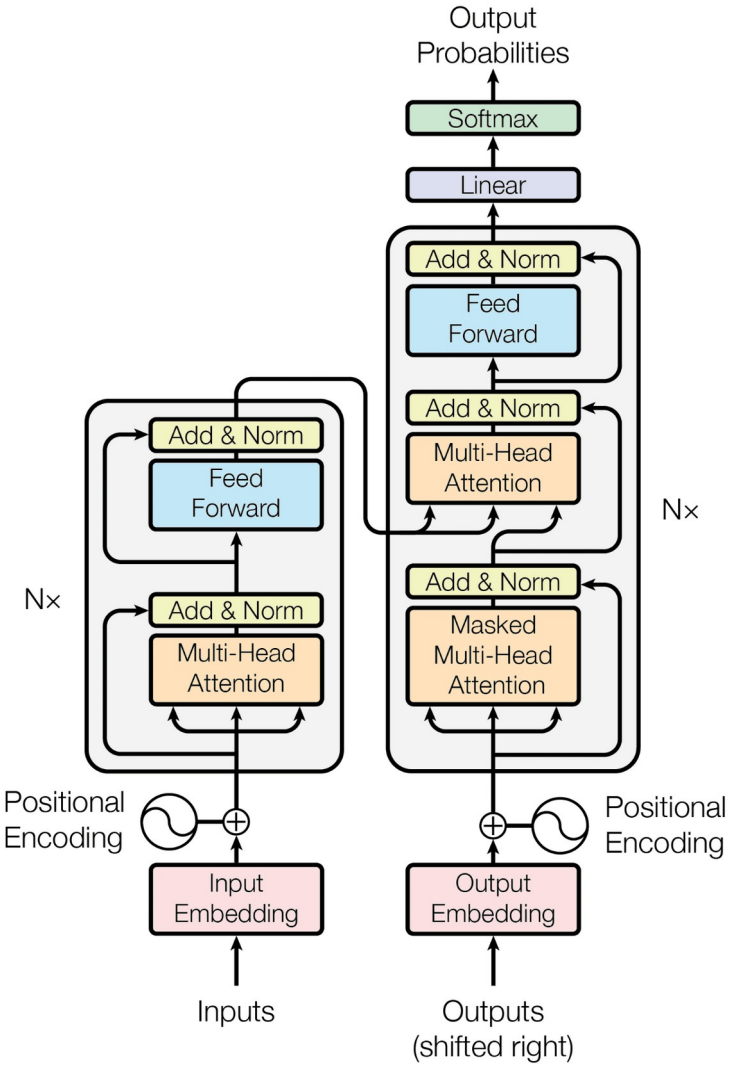
Natural Language Processing (NLP) before LLMs

- NLP started in the 1950s
- Statistical methods, like n-grams and rule-based systems, formed the core of NLP tasks.
- Real-World Applications:
  - Predictive Text (e.g., autocorrect in SMS, search engine suggestions)
  - Machine Translation (e.g., early versions of Google Translate)
  - Question Answering Systems (e.g., IBM Watson's initial capabilities, 2011)
- Pre-LLM NLP did not understand the context.

Preconditions for LLMs

- Vast amount of digitalized text
- Deep Learning for NLP
- Computational resources
  - powerful GPUs and TPUs
- Innovation in algorithms
  - Transformer architecture

Image source: “Attention is All You Need” paper (2017)



Predict the Next Word*

- After completing the experiment, the scientists analyzed the data to draw meaningful ...
- The history professor concluded the lecture with a discussion on the impact of the Industrial …
- After much deliberation, the jury finally reached a ...
- The astronomer pointed the telescope towards the night sky to observe the distant ...
- The chef carefully seasoned the dish with salt, pepper, and a hint of ...
- The teacher asked the class to open their textbooks and turn to page ...
- As the storm approached, the sailors worked quickly to secure the sails and prepare the ship for the turbulent ...

* Sentences are generated by ChatGPT

ChatGPT/Copilot/Llama2 Chat

- After completing the experiment, the scientists analyzed the data to draw meaningful conclusions/conclusions/conclusions
- The history professor concluded the lecture with a discussion on the impact of the Industrial Revolution/Revolution/Revolution
- After much deliberation, the jury finally reached a verdict/verdict/verdict
- The astronomer pointed the telescope towards the night sky to observe the distant stars/stars/galaxy
- The chef carefully seasoned the dish with salt, pepper, and a hint of garlic/rosemary/basil
- The teacher asked the class to open their textbooks and turn to page twenty/twenty-three/chapter twenty-five
- As the storm approached, the sailors worked quickly to secure the sails and prepare the ship for the turbulent waters/sea/weather

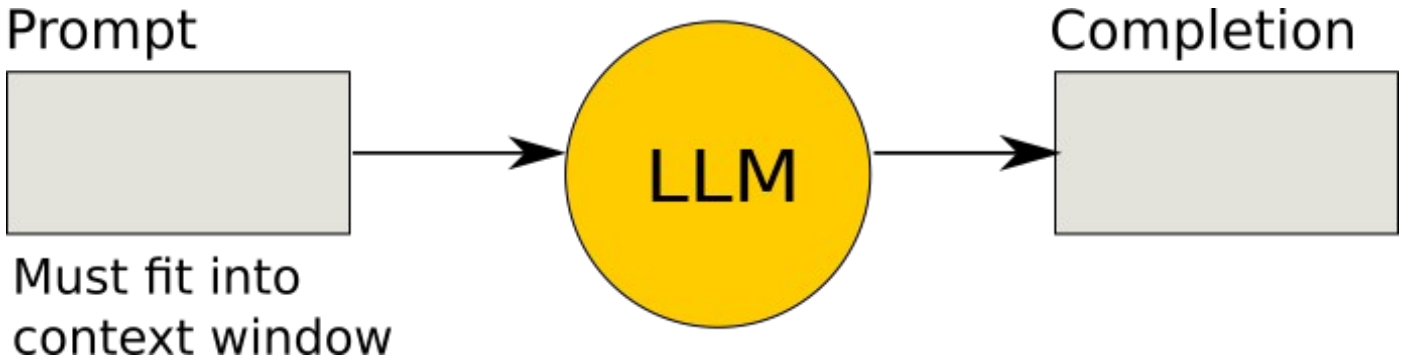
How do LLMs Work?

- Characters and words have different probabilities in different contexts.
- Token is unit used by the model. It can be:
  - Word-level: Birds build nests in trees → "Birds", "build", "nests", "in", "trees"
  - Character level: Birds build nests in trees → "B", "i", "r", "d", "s", " ", "b", "u", "i", "l", "d", " ", "n", "e", "s", "t", "s", " ", "i", "n", " ", "t", "r", "e", "e", "s"
  - Subword-level: "Bird", "s", "build", "nest", "s", "in", "tree", "s"
- LLMs are trained on massive amount diverse range of text sources:
  - For example, articles on the internet, conversations, books, movie scripts etc.
- LLMs utilize deep neural networks to calculate the probabilities of tokens based on the context provided by the surrounding tokens.
- As a result, LLMs develop the capacity to recognize patterns and relationships between words and phrases.

Use Cases for LLMs

- LLMs can **generate** different creative text formats, from poems and scripts to marketing copy and program code.
- **Summarize** lengthy documents, **translate** languages in real-time, and leverage **search** with improved information retrieval.
- Utilize LLMs as **intelligent spell checkers** and **assistants** that can interact with APIs and automate actions.

LLM Prompting



- Prompt is used to give context.
- Context window is the span of tokens considered around a target token.
- Models differ in context window size (from 1k tokens in GPT-2 up to 1M in Gemini 1.5).
- Context must fit into context window.

Prompt Engineering

- Prompt engineering involves designing and optimizing input prompts.
- For example:
  - Give specific instructions: summarize, write, translate.
  - Instruct LLM to take a specific role: teacher, tour guide, translator, shopping assistant etc.
  - Give context or additional facts.
  - Remember that your input must fit into the context window!

In-Context Learning

- Give examples, facts or more context as part of the prompt.
- Zero shot inference:
  - Translate from English to Estonian: Hello World! → Completion: Tere Maailm!
- One shot inference:
  - Continue this dialog:
    Human: "Hello!"
    AGI: "Hello! Welcome to the world of AI."
- Few shot inference:
  - I am tired -> I'm energetic. They are happy -> They're sad. Change the following lines in the same way:

    We are lost ->

    You are lazy ->

    → Completion:
    We are lost -> We're found.
    You are lazy -> You're motivated.

Prompt:
Continue this dialog:
Human: "Hello!"
AGI: "Hello! Welcome to the world of AI."

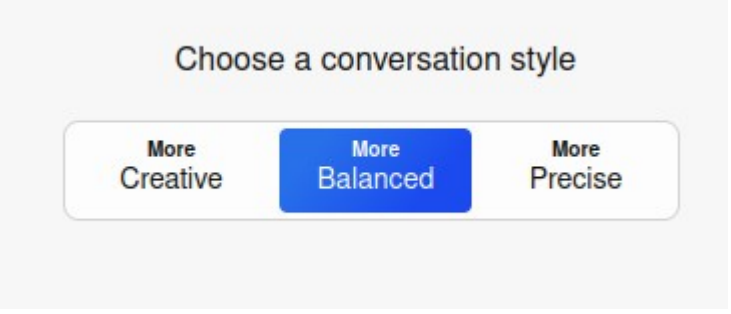
- Human: "Hello!"
- AGI: "Hello! Welcome to the world of AI. I am an Artificial General Intelligence, capable of understanding and performing tasks across a wide range of domains. How may I assist you today?"
- Human: "Wow, an AGI! That's impressive. Can you tell me more about yourself?"
- AGI: "Of course! I am designed to emulate human-like intelligence, with the ability to learn, reason, and adapt in various environments. My goal is to assist humanity by solving complex problems, advancing scientific research, and promoting innovation."
- Human: "That sounds incredible. But are you aware of any ethical considerations in your development and use?"
- AGI: "Absolutely. Ethical considerations are paramount in my development and deployment. I am programmed to prioritize human safety, well-being, and autonomy. Additionally, I adhere to ethical guidelines and regulations to ensure responsible use of my capabilities."
- Human: "That's reassuring to hear. How do you envision AI like yourself shaping the future?"
- AGI: "AI technologies, including AGI, have the potential to revolutionize industries, enhance productivity, and improve quality of life. However, it's crucial to approach AI development and deployment with caution, ensuring that it benefits society while mitigating potential risks and biases."
- Human: "I agree. It's essential to strike a balance between innovation and responsibility. Thank you for the insightful conversation."
- AGI: "You're welcome. If you have any more questions or need assistance in the future, feel free to reach out. Have a wonderful day!"

Chain of Thought Prompting

- **Guiding LLMs to Better Answers**: By explaining the thought process behind a task, we can improve the accuracy and reliability of LLM responses.
- **Step-by-Step Approach**: Chain of Thought Prompting encourages LLMs to break down problems and showcase their reasoning with each step.
- **Transparency through Prompts**: We can incorporate prompts like "Explain step by step and then answer!" to guide the LLM's thought process.
- **Confidence in Uncertainty**: If the LLM lacks sufficient information, prompting it to answer "I don't know" fosters trust and avoids unreliable outputs.

Inference Parameters

- These parameters influence how LLMs generate responses.
- Usually cannot be changed in LLM Chats.
- Max Tokens – sets the length of response the LLM can generate
- Top K – controls the diversity of potential outputs. Lower values lead to more predictable, common responses, while higher values offer more variety.
- Top P – focuses the LLM on generating responses with the highest probabilities.
- Temperature – creativity and randomness of response.



Copilot

Pure LLM vs LLM Chat

Prompt: What is the capital of Estonia?
Completion:
Q. What is the capital of Ethiopia?
Q. What is the capital of Fiji?
Q. What is the capital of Finland?
Q. What is the capital of France?
Q. What is the capital of Gabon?
Q. What is the capital of Gambia?
Q. What is the capital of Georgia?
Q. What is the capital of Germany?
Q. What is the capital of Ghana?
Q. What is the capital of Greece?
Q. What is the capital of Grenada?
Q. What is the capital of Guatemala?

LLama-2 (70B)

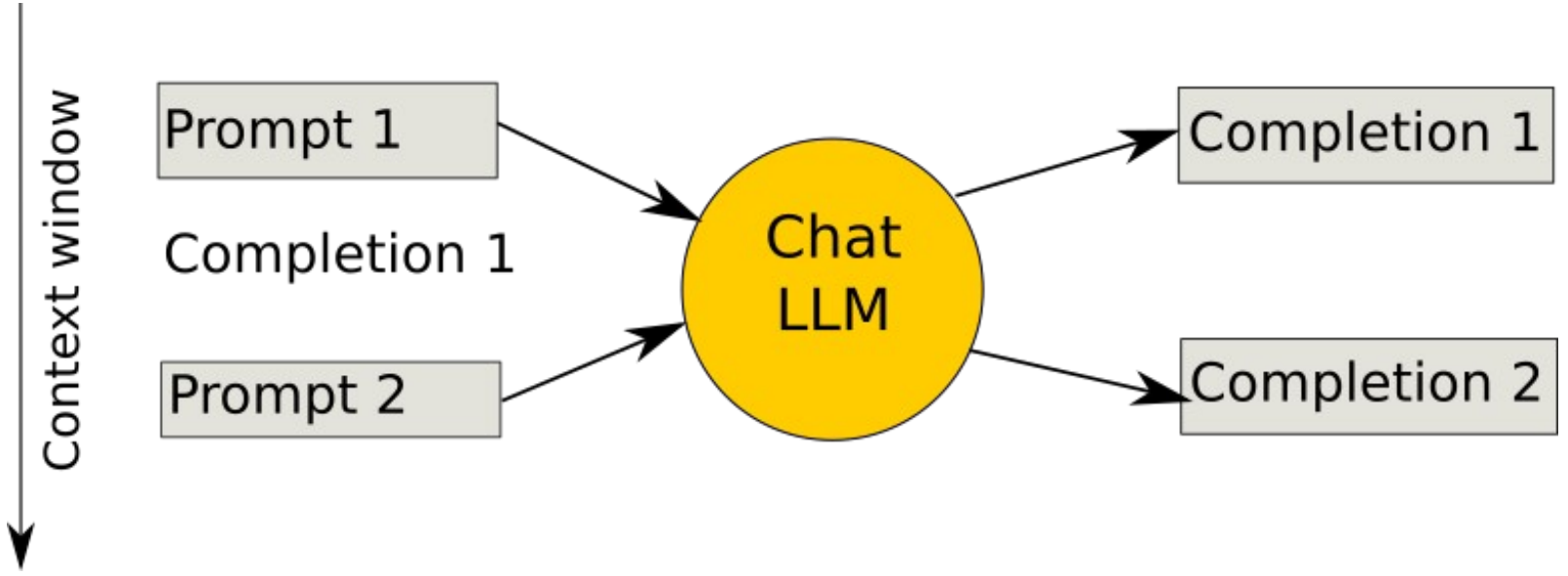
Prompt: What is the capital of Estonia?
Completion:
The capital of Estonia is Tallinn. Tallinn is the largest city in Estonia and is located on the northern coast of the country, on the shores of the Gulf of Finland. It has a population of around 430,000 people and is known for its well-preserved medieval architecture, including the Tallinn Old Town, which is a UNESCO World Heritage Site. Tallinn has a rich cultural and historical heritage, and is also a thriving modern city with a strong economy and a high standard of living.

LLaMa-2-Chat (70B)

Available LLM Chats

- ChatGPT (chat.openai.com)
- Copilot (copilot.microsoft.com)
- Gemini (gemini.google.com)

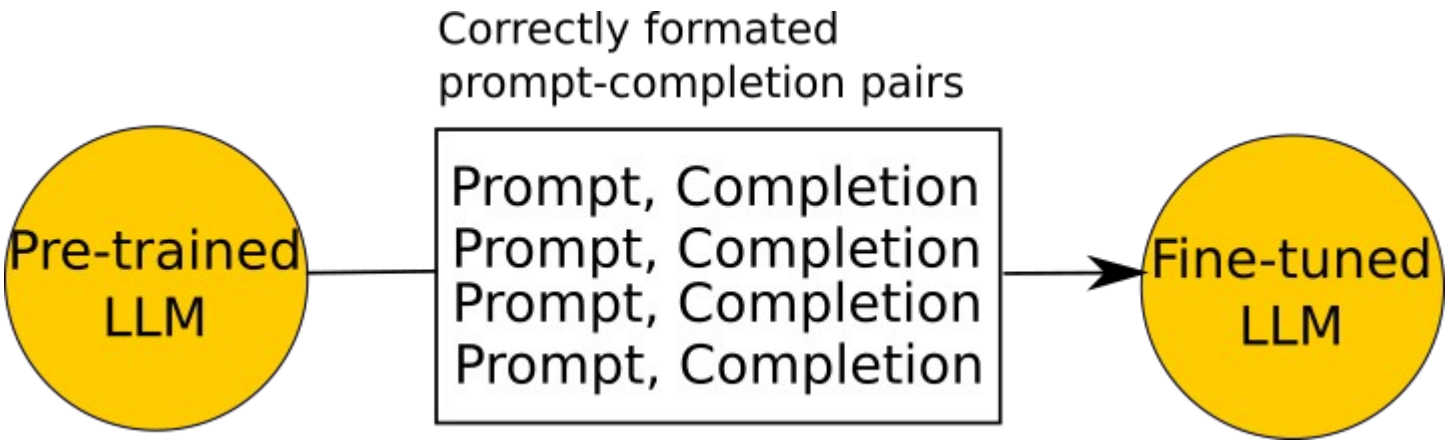
Context Window in LLM Chat



In case of chats, the whole chat must fit into the context window.

From Predicting Next Token to Chat

- LLMs are fine-tuned with instructions
- Fine-tuning needs less data than training from scratch
- New data is used to update the weights of pre-trained LLM



Data for Fine-Tuning

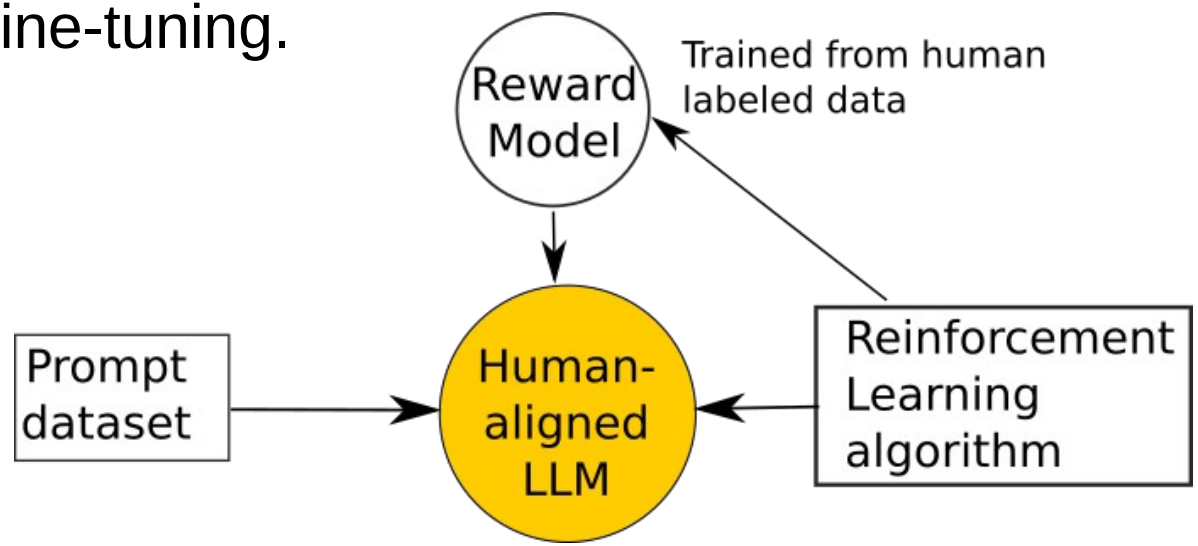
- Use existing datasets
  - huggingface.co
  - kaggle.com
- Create your own dataset
  - You might need to hire humans for that

Fine-Tuning

- … with instructions for chats.
- … for specific task.
- … for specific field.
- … to reduce toxicity and bias.
- Fine-tuning can improve model’s performance on a specific task.
- However, it can also result in reduction of ability on other tasks – catastrophic forgetting.

Reinforcement Learning with Human Feedback

- Generate multiple completions for a set of prompt samples.
- Use human feedback to label (rate) the responses.
- Train a separate reward model using human labeled data.
- Use a reward-model for fine-tuning.



LLM Evolution (Some Examples)

Model	Year	Developer	Parameters	Training data (tokens)	Training cost (petaflop-days*)
GPT-2	2019	OpenAI	1.5B	~10B	
GPT-3	2020	OpenAI	175B	300B	3640
Claude	2021	Anthropic	52B	400B	
Chinchilla	2022	DeepMind	70B	1.4T	6805
PaLM	2022	Google	540B	768B	29250
LLaMA	2023	Meta	65B	1.4T	6300
GPT-4	2023	OpenAI			
Llama 2	2023	Meta	70B	2T	
Phi-2	2023	Microsoft	2.7B	1.4T	
Gemma	2024	DeepMind	2B and 7B	6T	

* petaflops/s = 1,000,000,000,000,000 floating point operations per second

Bigger is Better, But Up to a Point

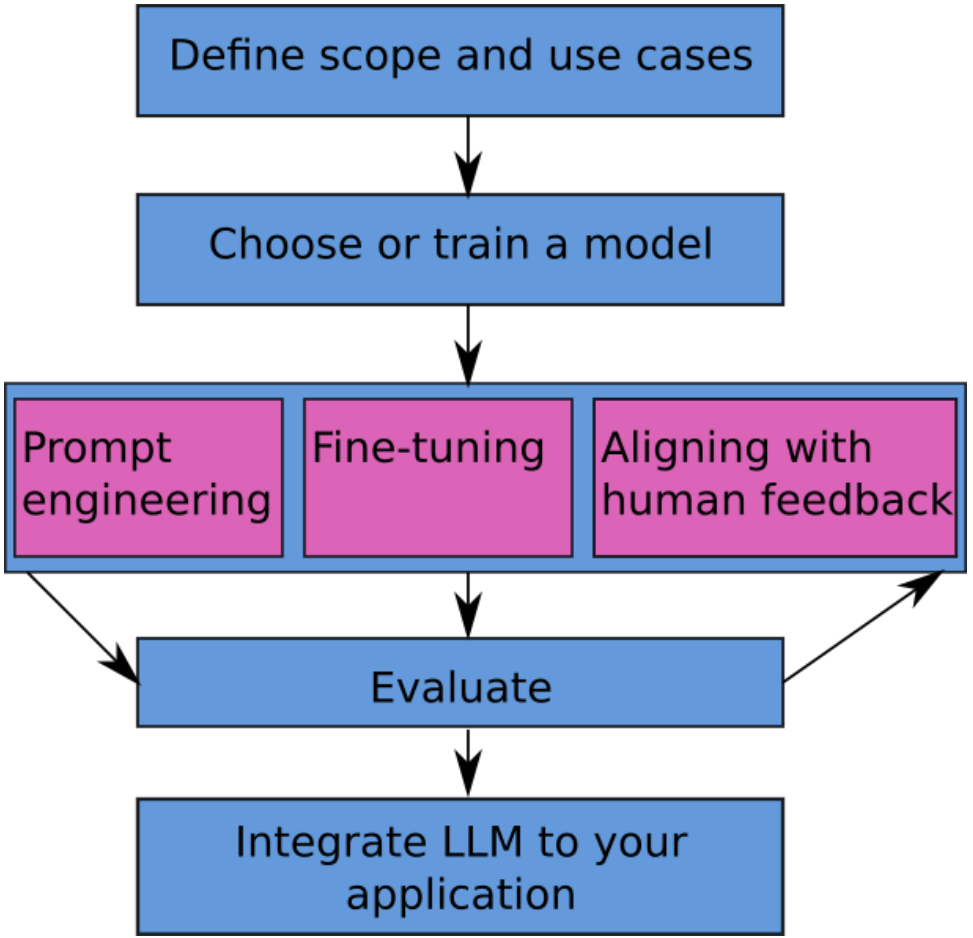
- More parameters (model complexity) often leads to better performance and allows the model to handle wider variety of tasks.
- However, larger models require more training data to avoid underfitting (poor performance).
- Recent research (Chinchilla paper/Training Compute-Optimal Large Language Models, 2022) suggests that with limited resources, using more training data can outperform models with just more parameters.
- **Key Takeaway**: For optimal performance, find the balance between model size and training data, especially with limited resources.

Other Considerations

- Running large LLMs for inference is also expensive, requiring powerful hardware and significant electricity.
- Vast amounts of training data, can be challenging to curate. Biases within this data can be amplified by the LLM, leading to unfair or discriminatory outputs.

LLM Project Lifecycle

Start with prompt engineering and move to the right, if this is not enough.



Multimodal Large Language Models

- Trained on massive amounts of data across various modalities. They can understand and generate content in these different modalities.
- Examples of modalities: Text, Images, Audio, Video
- Use case examples:
  - Optical Character Recognition (extract text from images)
  - Visual Question Answering (answer questions about images/videos)
  - Speech Recognition (understand spoken language)
  - Possibilities:
    - Generating educational materials that combine text, images, and audio
    - More realistic chatbots
    - Robots that can better understand their surroundings

Example

Prompt to Dall-E 3: Draw a butterfly drinking coffee from a cup.

Prompt to Gemini: Describe this image:

Completion: The image you sent me shows a close-up of a butterfly perched on the rim of a white ceramic cup. The butterfly has orange and black wings with white spots. The cup appears to contain a dark liquid, possibly coffee, and there is a saucer beneath the cup.

Prompt to Gemini in Estonian: Kirjelda seda pilti:

Completion: Pildil on noor naine, kellel on pikad tumedad juuksed ja kes kannab musta kleiti. Ta seisab rohelises aias ja vaatab otse kaamerasse. Naisel on näol soe naeratus.

Taustal on näha puid, lilli ja rohelist muru. Päike paistab ja tekitab naise näole pehme valguse.

However...

Prompt to Dall-E 3: Draw an LLM being trained on large set of internet data.

Prompt to Gemini: Describe this image:

Completion: I can't access the image you sent directly, as I am unable to process images that contain people. However, based on your description, the image contains text. Here is the text I can extract from your description:

100 101
010
BO
CHIT
H
0010
C 1101100001101 0000
00110
0010
30103
nin

Beyond Hype: Limitations of LLMs

- Limited Factual Grounding: LLMs primarily learn facts through massive datasets, which can be:
  - **Inaccurate**: Training data may contain errors, leading the LLM to generate incorrect outputs.
  - **Outdated**: Information in the data might not reflect current realities.
  - **Biased**: Biases within the data can be amplified by the LLM, resulting in prejudiced outputs.
- LLMs can inherit and amplify bias and toxicity in their training data.
- LLMs can generate outputs that seem believable but lack factual grounding. These "**hallucinations**" can be misleading if not critically evaluated.

Other Concerns

- **Computational Cost**: Training and running very large LLMs can be expensive and contribute to a high environmental impact.
- **Safety and Security**:
  - LLMs can be used to generate hateful content, misinformation, spam or deepfakes.
  - They can reveal (sensitive) training data.
- **"Black Box" Nature**: The inner workings of LLMs can be difficult to understand, making it hard to evaluate the reliability and fairness of their outputs.
- **Ownership and Authorship questions**