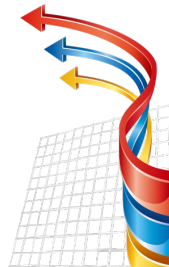
Data Driven Optimization for Digital Industries

Part 2: Foundation Models

Rearrenged from Umberto Junior Mele material

Andrea Corsini
andrea.corsini@unimore.it

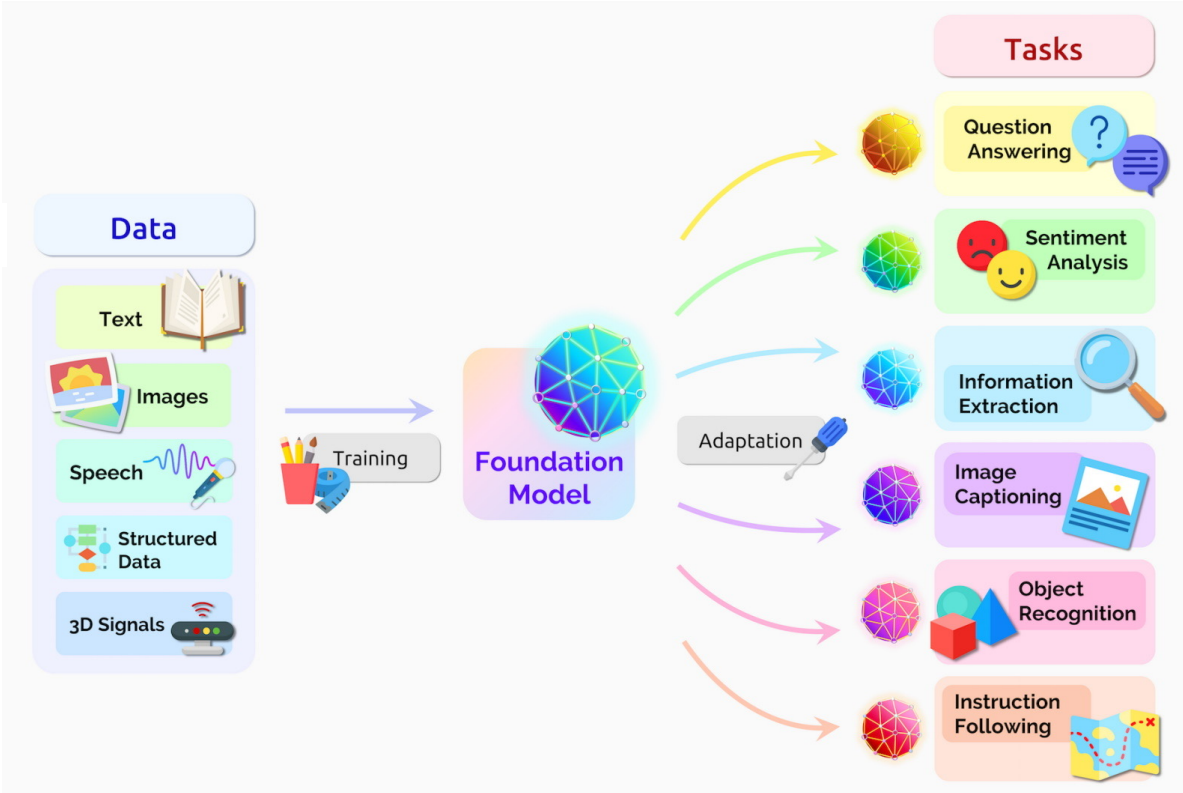


Exercises

You need to execute the experiments provided in the jupyter notebook file shown during class. Then, upload on Moodle a compressed folder with the file due to 19/5.

Foundation Models

- Foundation models are pre-trained models (neural networks) on diverse and large datasets.
- They enable broad applications due to their generalist nature, from customer service to content creation.
- They can be fine-tuned for a variety of specific tasks (e.g., translation and answering questions), thus reducing the cost and time of training neural networks.
- They allow researchers and practitioners to leverage knowledge captured from a wide range of internet text.



Examples

Model	Core differentiator	Pre-training objective	Para-meters	Access	Information Extraction	Text Classification	Conversa-tional AI	Summari-zation	Machine Translation	Content generation
BERT	First transformer-based LLM	AE	370M	Source code						
RoBERTa	More robust training procedure	AE	354M	Source code						
GPT-3	Parameter size	AR	175B	API						
BART	Novel combination of pre-training objectives	AR and AE	147M	Source code						
GPT-2	Parameter size	AR	1.5B	Source code						
T5	Multi-task transfer learning	AR	11B	Source code						
LaMDA	Dialogue; safety and factual grounding	AR	137B	No access						
XLNet	Joint AE and AR	AE and AR	110M	Source code						
DistilBERT	Reduced model size via knowledge distillation	AE	82M	Source code						
ELECTRA	Computational efficiency	AE	335M	Source code						
PaLM	Training infrastructure	AR	540B	No access						
MT-NLG	Training infrastructure	AR and AE	530B	API						
UniLM	Optimised both for NLU and NLG	Seq2seq, AE and AR	340M	Source code						
BLOOM	Multilingual (46 languages)	AR	176B	Source code						

AR = Autoregression

AE = Autoencoding

Seq2seq = Sequence-to-sequence

Highly appropriate

Appropriate

Somewhat appropriate

Task Specialization

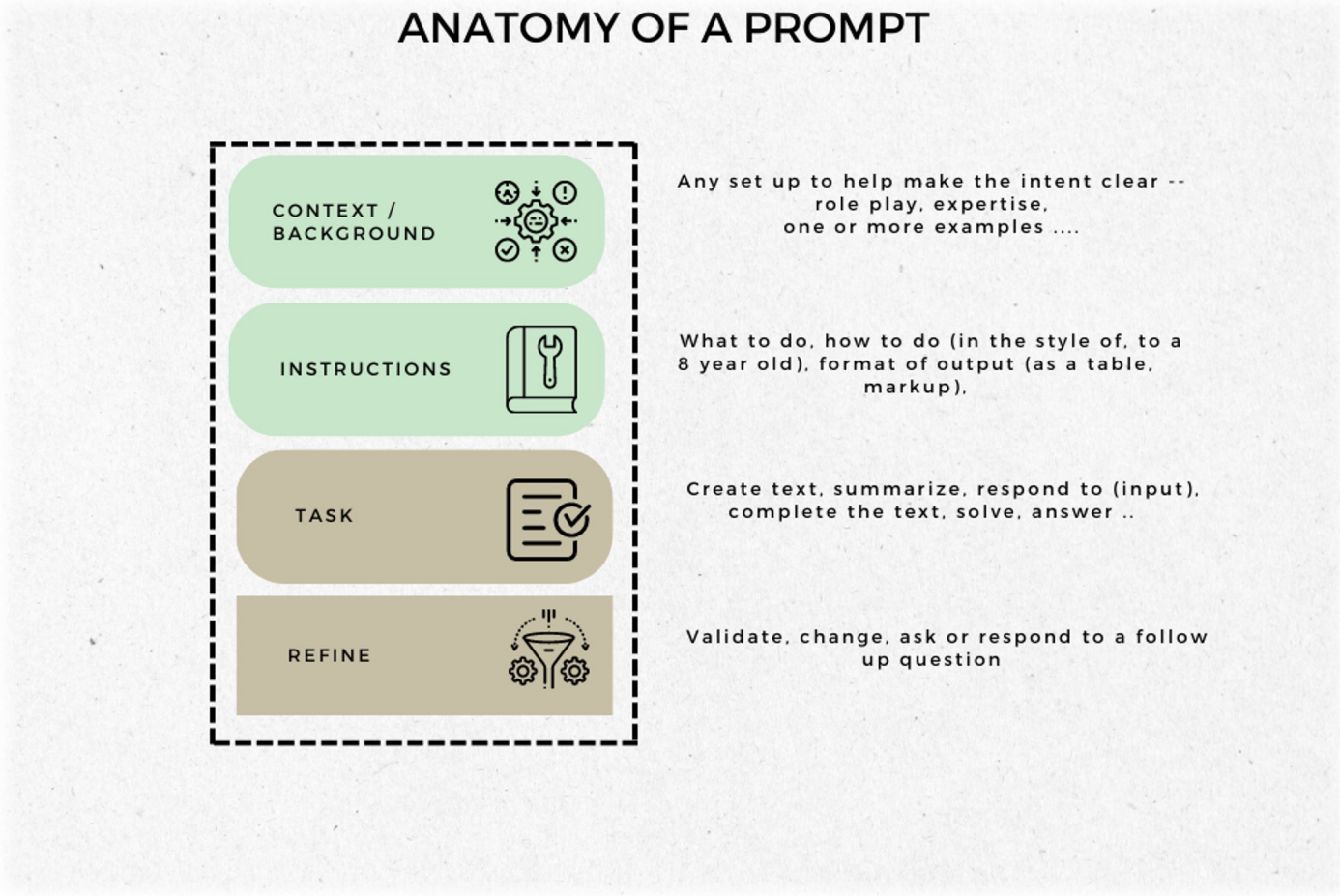
Prompt Engineering consists in designing the input to the foundation models to get the desired outputs:

- Zero-Shot
- Few-Shot
- Chain of Thought
- Self-Reflection

Training Adaptation consists in changing the parameters of the FM or a smaller model used to adapt the outputs of specific applications:

- Linear Probing
- Fine-Tuning
- Mixed Approaches

Prompt Engineering



Prompt Engineering

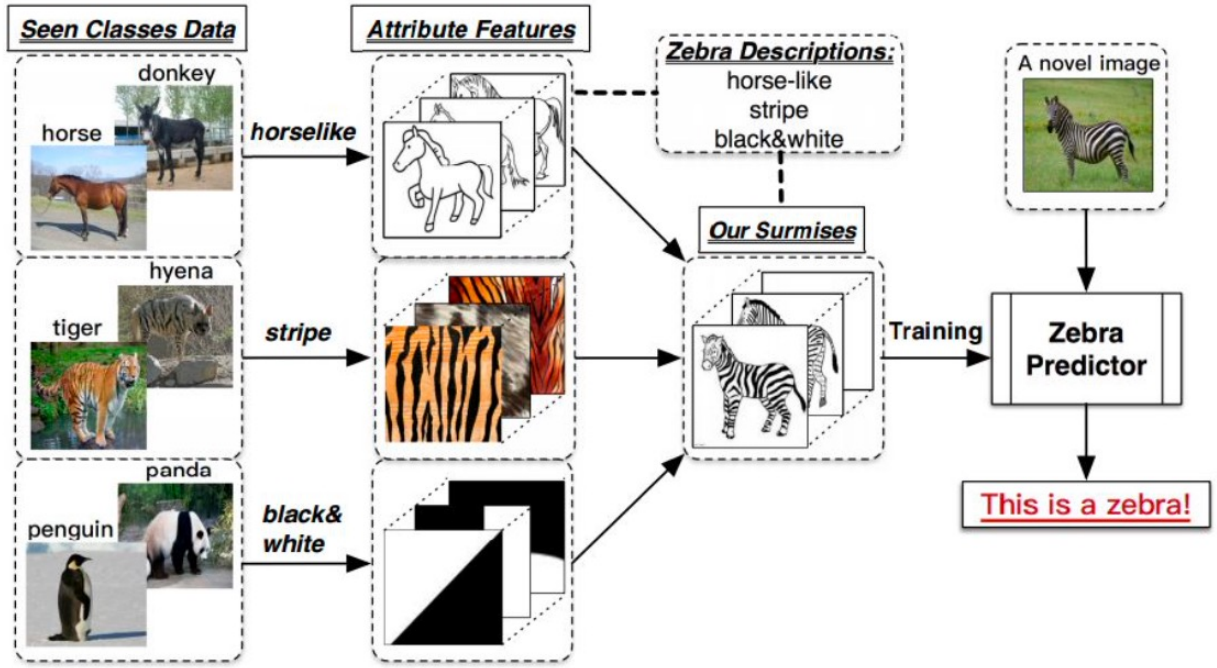
Prompt engineering can include tactics such as:

- **Specificity**: making prompts more detailed can better guide the model.
- **Instructions**: explicit the structure of the desired output (format or content) can improve results.
- **Examples**: providing examples of the output can be very effective, particularly for tasks that require generating text in a specific format.
- **Redundancy**: asking the same question in different ways can ensure the model understands the task.
- **Question Framing**: adjusting the way a question is posed can significantly influence the response.

Prompt engineering is both an art and a science, often requiring experimentation and iteration to get right. While effective, it can also be labour-intensive and requires a good understanding of the specific model's strengths and weaknesses.

Zero-shot Learning

Definition: Zero-shot Learning refers to the scenario where the FM is asked to answer a task unseen during training. This capability is made possible due to the broad pretraining which exposes the model to diverse scenarios and tasks. For example, a model can identify animals not by directly learning from images of them, but by understanding and applying descriptive attributes about them.



Few Shot Learning

Definition: Few-shot Learning refers to the scenario where the model is expected to generalize from a limited number of examples. In the context of Foundation Models, it involves providing a few examples of a particular task to the model in the form of a prompt. This helps the model understand the desired output and generalize from the examples to perform the task on new inputs.

Prompt

A “whatpu” is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is: “We were traveling in Africa and we saw these very cute whatpus.”

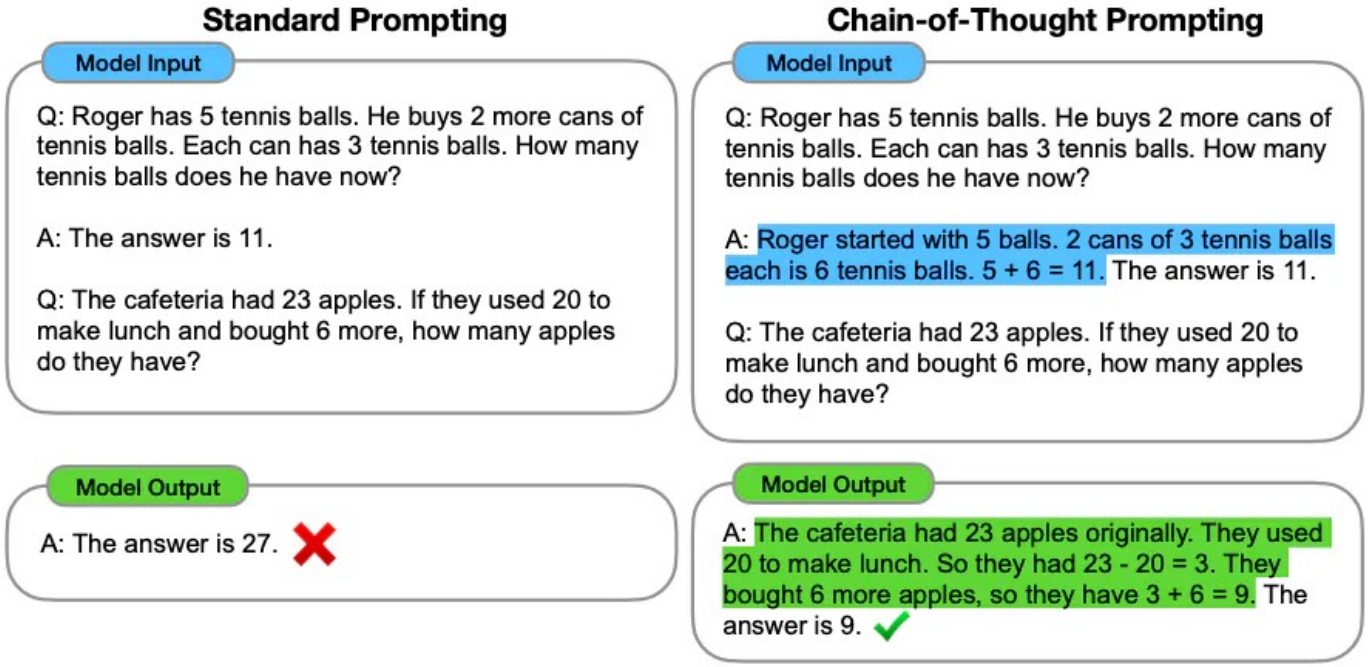
To do a “farduddle” means to jump up and down fast. An example of a sentence that uses the word farduddle is?

Answer

When we won the game, we all started to farduddle in celebration.

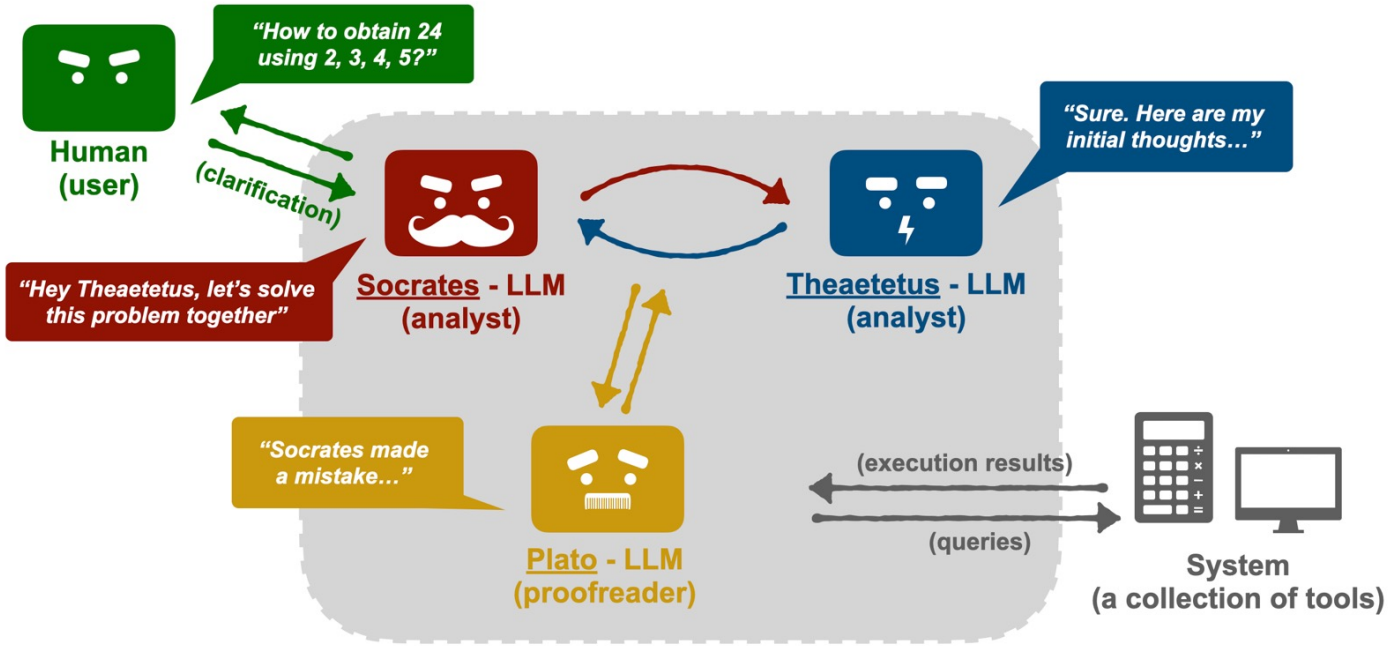
Chain of Thought

Definition: Chain of Thought is achieved by prompting the models to generate a series of intermediate steps that lead to the final answer of a multi-step problem. The technique improves results in reasoning tasks that require logical thinking and multiple steps to solve. It could compete with task-specific fine-tuned models on several tasks.

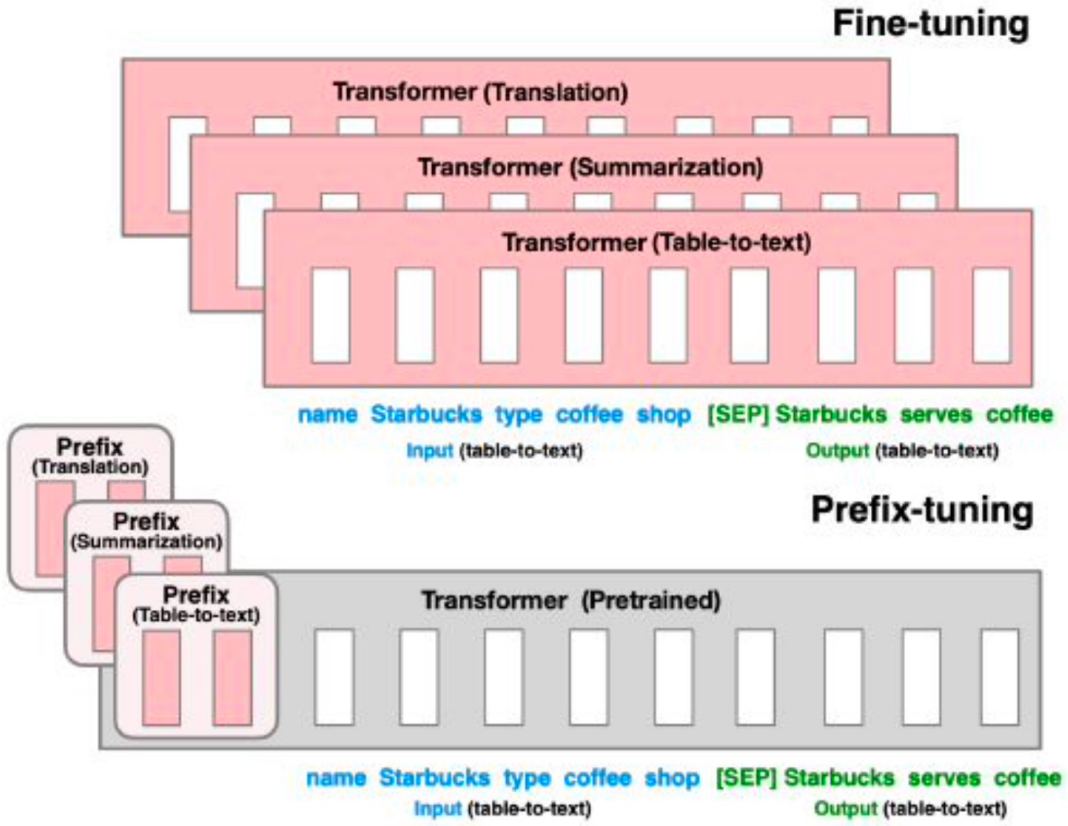


Self-Reflection

Definition: Self-Reflection is achieved by prompting the models to introspect and analyse their own outputs and reasoning process. The model leverages its underlying knowledge and understanding gained during the pretraining phase to provide insights about its own thought process.



Training Adaptation



Training Adaptation

Linear Probing involves training a model on top of the frozen model to adapt to the specific task.

- Pros: It's a simple and efficient way to extract useful information from the model.
- Cons: It may not leverage the full potential of the model as it doesn't alter the foundation model's parameters.

Fine-Tuning involves continuing the training of the pretrained model on the specific task.

- Pros: It can significantly improve the model's performance on the task.
- Cons: It requires costly task-specific training data and can overfit if not managed properly.

Mixed Approaches combines different methods like linear probing and fine-tuning to achieve better results.

- Pros: It improves performances with respect to both methods and it is less costly.
- Cons: It could be more complex to implement and may be more costly than Linear Probing.

Using Tools within LLMs

Foundation models can learn how to use tools directly from input prompts. The large-scale training data often include examples of using tools, which the models generalize from.

Examples:

- **Search the Internet**: Foundation models can gather additional information from internet searches. This could be used for fact-checking or information retrieval.

- **Calculators and Solvers:** models can effectively use calculators or solvers. When given prompts related to calculations or problem-solving, they can leverage their internal understanding to provide solutions.

Coding

Foundation models can write code in various programming languages, understand syntax, solve algorithmic problems, and debug. These models can help to design an algorithm, write the code, and validate it with test cases. For example, they can generate a Python function to solve a system of linear equations.

Advantage:

- They are expected to play a more vital role in coding. This includes not just writing and debugging code but also suggesting architectural improvements, identifying security risks, and more.

Limitation:

- Foundation Models are not perfect and can make mistakes, especially with complex coding tasks. They should be used as a tool to aid developers, not as a replacement.

Searching the Net

GPT-4 generates responses based on training data. However, allowing such models to interact with the internet in real-time could greatly improve their responses, keeping them updated with the latest information, and enabling fact-checking against live data. For instance, they could provide recent stock market trends, up-to-date news, or even the latest scientific research.

Advantages:

- Enable better output with fact-checking.
- Allow automating tasks.

Limitations:

- Could increase the creation of fake-news
- Not all the data available in internet is correct.

Final Considerations

Be aware of

As language models become larger, they generally perform better, scoring higher on benchmarks, unlocking new capabilities, but also introducing new biases or misinformation.

- **Inverse Scaling Phenomenon**: is based on the hypothesis that certain tasks exhibit inverse scaling. As the overall test loss of the model improves, task performance predictably worsens. These tasks appear to be rare but could represent important issues with current pretraining and scaling paradigms.

- **Logic Issues**: LLMs often struggle with logical reasoning tasks, including the ability to accurately perform deductions. As models scale, this problem becomes more pronounced, a phenomenon known as inverse scaling.

Real-life Applications

- **Customer Service:** LLMs can be used to handle customer inquiries and complaints, automating responses to frequently asked questions, and escalating complex issues to human agents.
- **Content Generation:** LLMs can produce high-quality written content, including product descriptions, social media posts, and blog articles, aiding marketing and communication efforts.
- **Data Analysis:** LLMs can analyse large volumes of text data, like customer reviews or social media comments, to gain insights into consumer sentiment and trends.
- **Personal Assistants:** advanced versions of voice-activated personal assistants like Siri or Alexa can be developed using LLMs, providing more accurate and context-aware responses.
- **Education and Training:** LLMs can provide personalized learning experiences, offering real-time feedback and assistance on a variety of subjects.

Future Trends

- **Fine-Tuning LLMs:** there is an increasing focus on effectively fine-tuning LLMs to specific tasks, industries, or domains to enhance their utility and applicability.
  - They can solve optimization problems.
- **Responsible AI:** LLMs grow larger and more complex, efforts to identify and mitigate issues such as bias, misinformation, and ethical considerations will become more important.
- **Interactive AI Systems:** expected growth in the development of interactive AI systems that engage users in more dynamic and personalized ways, combining NLP, voice recognition, and machine learning techniques.
- **Data Privacy:** ensuring the privacy and security of data used to train and interact with LLMs will be a key concern as regulations become stricter.

Lab

[Link to the jupyter file](#).

[Link to the GitHub repo](#).