50.039 Theory and Practice of Deep Learning

W11S3-bis (Special) ChatGPT

Matthieu De Mari

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

About this week (Week 11)

1. What are **Large Language Models**?
2. What is **ChatGPT?**
3. What are the **different elements** composing the ChatGPT model and **how do they relate to the concepts we have seen in previous weeks?**
4. How was ChatGPT **trained** using a **mix of Deep Learning and Reinforcement Learning with Human Feedback**?

Large Language Models

Definition (Large Language Models):

Large Language Models (LLMs) are a class of machine learning algorithms designed to understand, interpret, and generate human language. They are based on probabilistic methods that capture the statistical structure of text data, allowing them to generate text that appears coherent and contextually relevant.

Similar to the Word and Sentence Embedding Problem in Week 8!

Large Language Models

The development of language models has been significantly influenced by the introduction of Deep Learning techniques.

- DL facilitates the creation of more sophisticated and expressive models, capable of generating text that resembles human-written content.
- These LLM models have become the essential components in various NLP tasks (machine translation, sentiment analysis, text summarization, text generation, and conversational AI systems).
- Typically used for embedding text, enabling machines to understand and communicate with humans more effectively.

The ChatGPT ML problem

ChatGPT is a model used in a typical task of NLP, which is designing a query chatbot, answering user requests as efficiently as possible.

Like any ML problem, it consists of four elements

1. **Task**
2. **Dataset**
3. **Model**
4. **Loss and Training Procedure**

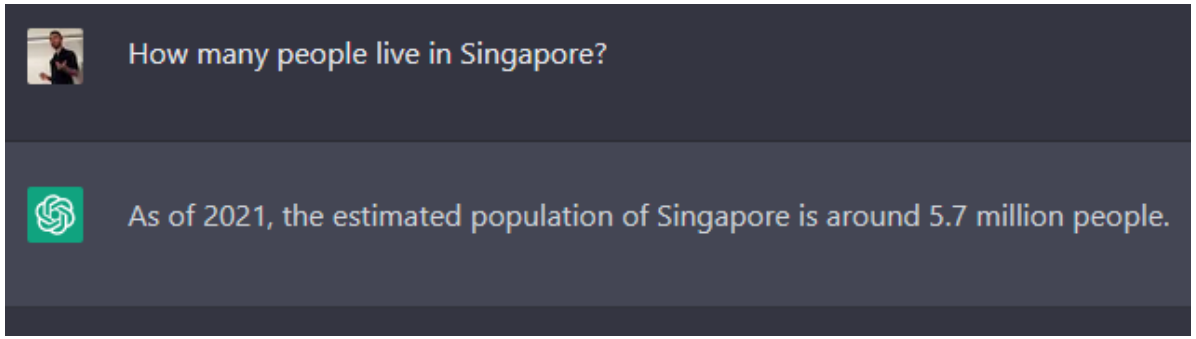
The ChatGPT ML problem: 1. Task

1. Task

The ChatGPT model is one of the many models, specifically designed to **engage in meaningful conversations with users, allowing for the generation of coherent, contextually relevant, and informative responses.**

Other examples: Bard AI (Google), LLaMA (Facebook), Claude, etc.

- LLaMA: https://ai.facebook.com/blog/large-language-model-llama-meta-ai/
- Google Bard: https://bard.google.com/
- ChatGPT: https://chat.openai.com/chat



The ChatGPT ML problem: 2. Dataset

2. Dataset

The pre-training dataset for ChatGPT consists of a vast collection of text data from diverse sources, such as **books, articles, websites, and social media posts**.

The purpose of using a large and varied dataset is to ensure that the model learns a broad understanding of language, encompassing different styles, topics, and contexts.

- Must be designed to expose the model to a wide variety of linguistic patterns, contexts, and domains, enabling it to learn the complexities and nuances of human language.
- To achieve this, the dataset must be large and diverse, containing text from numerous sources that reflect the many ways language is used.

The ChatGPT ML problem: 2. Dataset

2. Dataset

The primary source of data for ChatGPT is the **WebText** dataset, which consists of approximately 8 million web pages collected from the internet, such as news articles, websites, and online forums.

- This diversity helps to ensure that the model is exposed to a broad range of language and writing styles, which is essential for generating human-like text.
- Typically, we want the text in the dataset to be written by people for different purposes and with different writing styles.
- Also, constantly updated with new data. When new web pages are added to tracked websites, they are included in the dataset, which helps to ensure that the model is trained on the most recent and relevant language data.

The ChatGPT ML problem: 2. Dataset

2. Dataset

Also, some additional datasets were used. These include a variety of text sources such as books, articles, and other written works.

The exact sources of these datasets were not publicly disclosed by OpenAI, but we suspect something similar to what was used for the other models we discussed in W8S1-2, i.e. Word2Vec, FastText, etc.

The ChatGPT ML problem: 2. Dataset

2. Dataset

Should also carefully curate and filter the dataset to **minimize biases (racism, violence, etc.)**, aiming to provide a representative sample of human language that is both balanced, comprehensive and safe.

Remember W8S3 about Word Embedding biases!

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

The ChatGPT ML problem: 2. Dataset

2. Dataset

Also used **Data Augmentation**!

Typically, **text manipulation,** e.g.

- **Synonym replacement,**
- **Random insertion,**
- **Random deletion,**
- **Random swapping of words,**
- **Text-to-speech conversion,**
- Etc.

(Discussed on W4 for images)

- Also, **back-translation**, which is the process of translating a sentence from one language to another, and then translating it back to the original language.
- Can be especially useful for training models on multilingual data, as it can help the model to learn the nuances of different languages and cultures.
- Allow the model to learn from a wider range of language styles and structures.

The ChatGPT ML problem: 3. Model

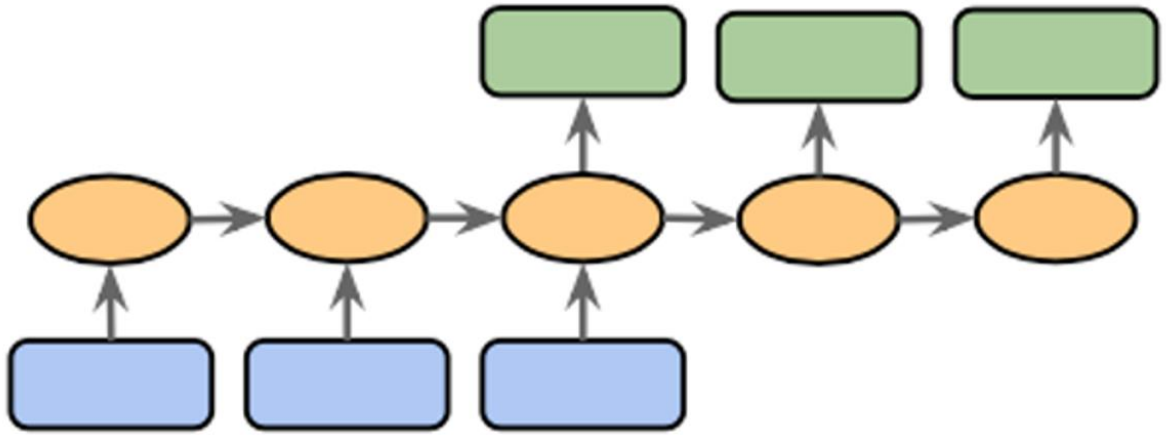
3. Model

ChatGPT is another typical **encoder-decoder model**, or **Seq2Seq model.**

Both the inputs and the outputs can be seen as sequential data.

Typically discussed in W6S2 and Week 8 lectures.

Eventually both the **encoder** and the **decoder** models will assemble into a **Seq2Seq** model.



The ChatGPT ML problem: 3. Model

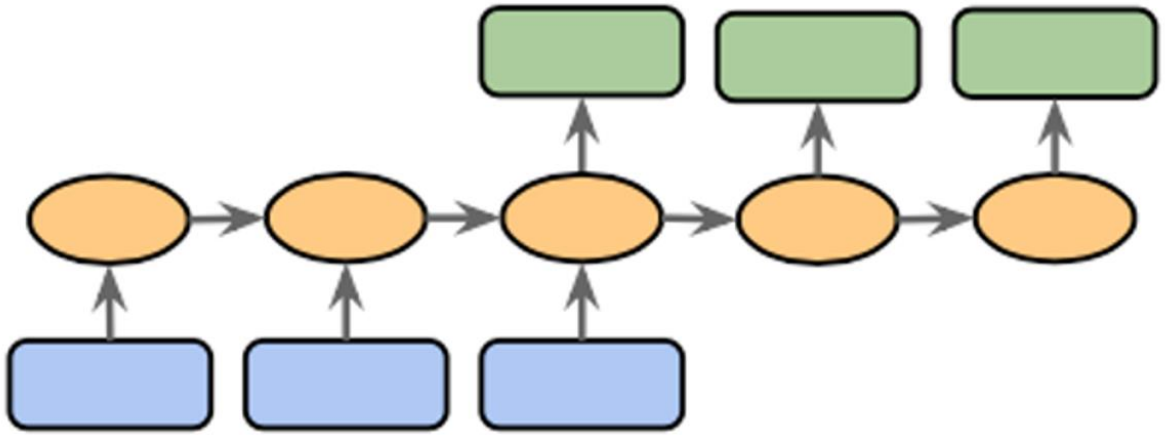
3. Model

The model input here consists of the user prompt, in the form of sequential text.

It also includes a memory vector, containing information about previous queries and answers.

The **encoder part** should combine all elements and encode the meaning of the user query into a numerical vector describing the meaning of that request.

Eventually both the **encoder** and the **decoder** models will assemble into a **Seq2Seq** model.



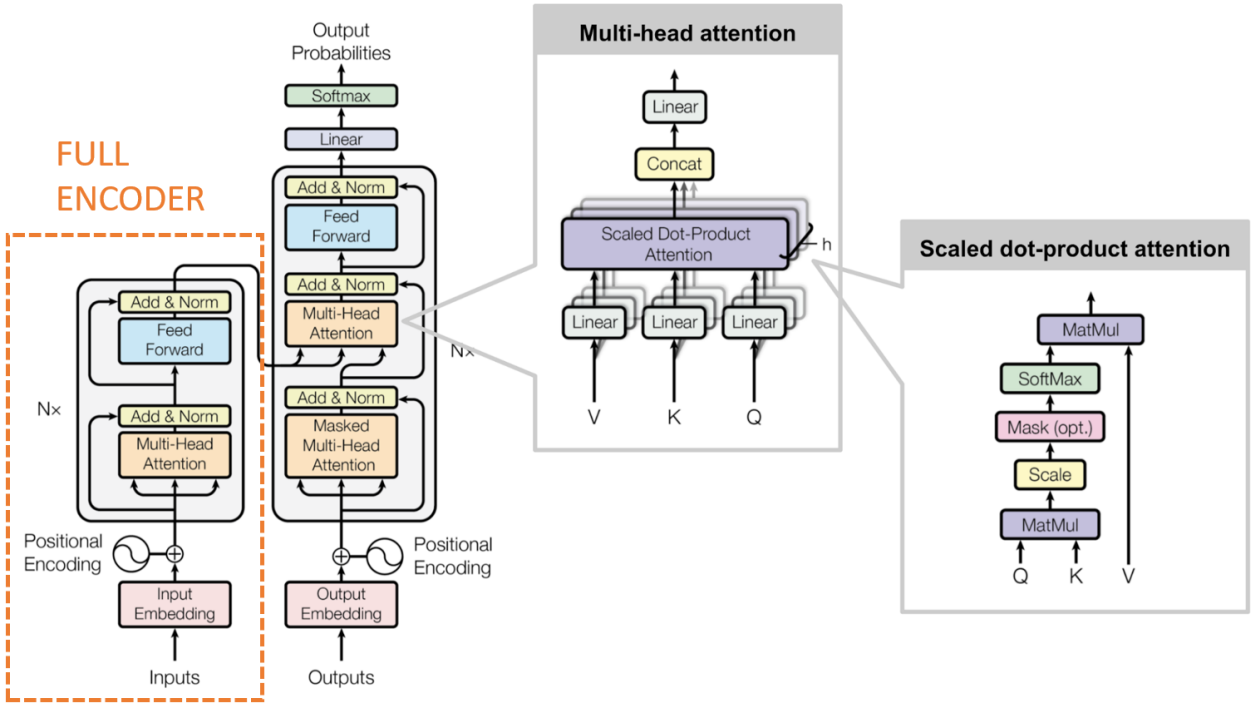
The ChatGPT ML problem: 3. Model

3. Model (Encoder)

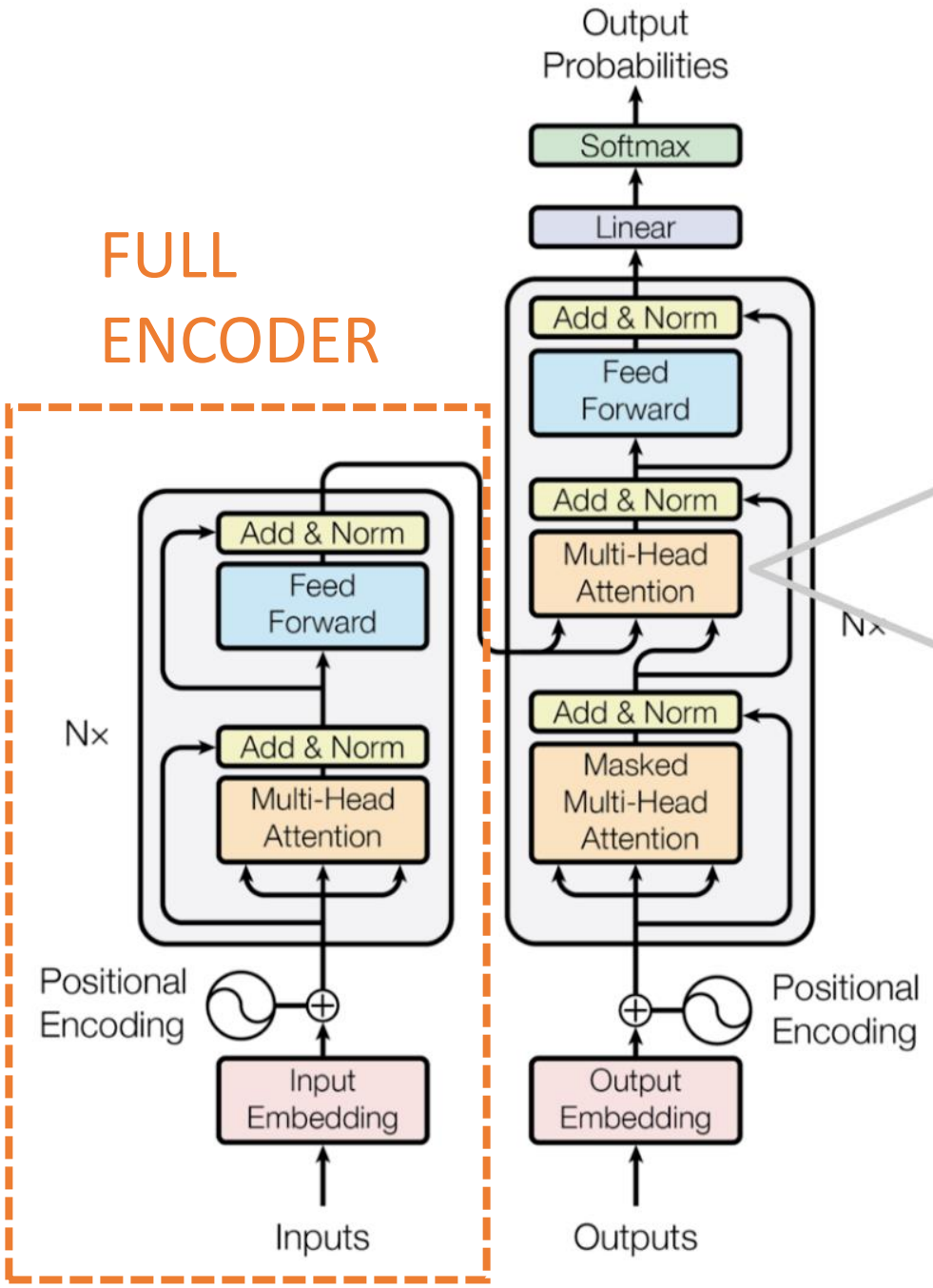
The encoder part of ChatGPT relies on the GPT idea discussed in W8S3.

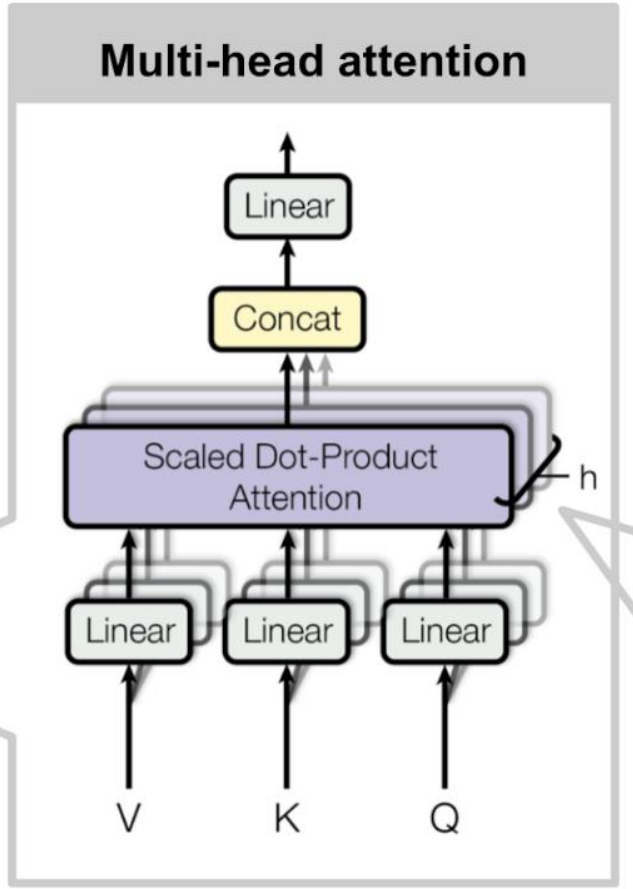
- **Transformer-based model,**
- Will process an input consisting of a sequence of words,
- And will produce an encoding vector.

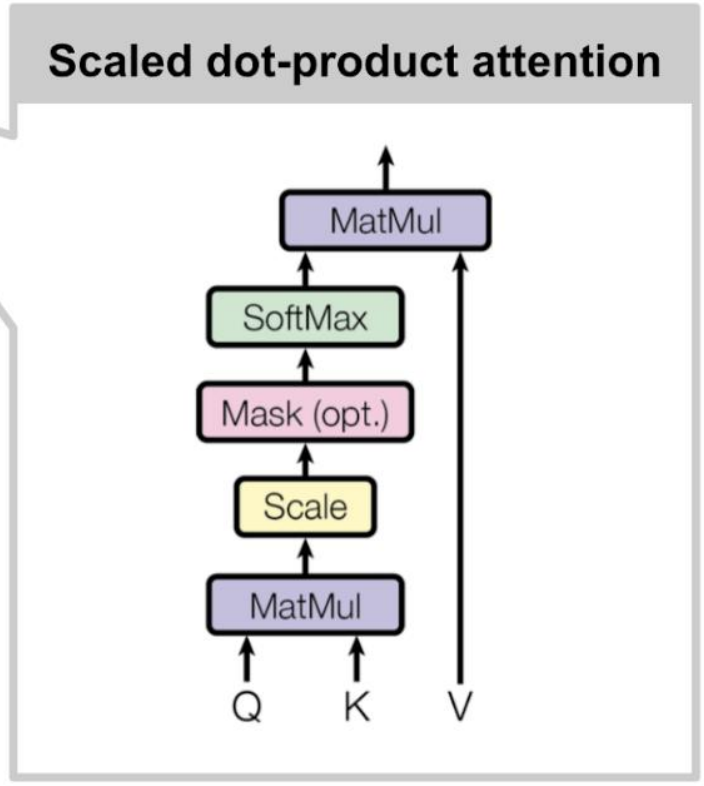
- As in W8S3 on Transformers architectures (Encoder part).



FULL ENCODER





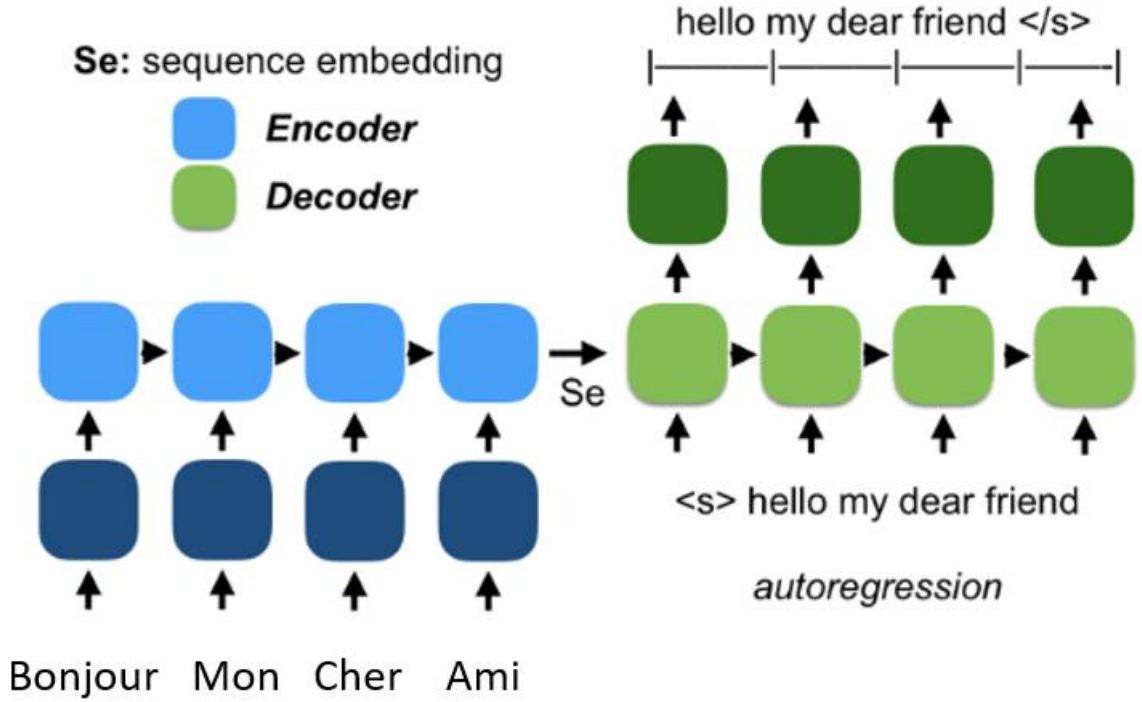


The ChatGPT ML problem: 3. Model

3. Model (Decoder)

The decoder part of ChatGPT relies on the auto-regressive models, that are trying to predict the next word to use in sentence and constructs an answer, one word at a time.

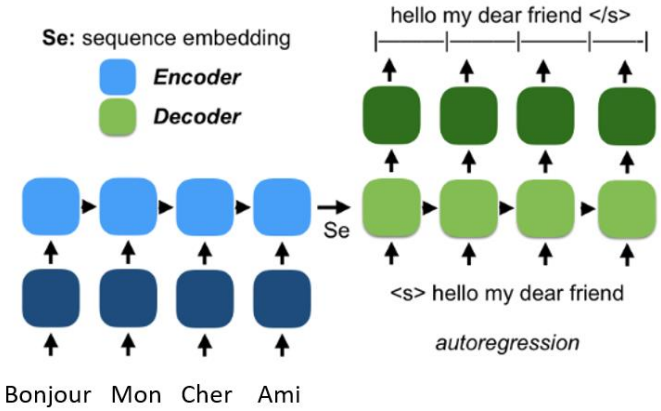
Will use the encoding vector as a memory vector for the autoregressive RNN used in the decoder part!



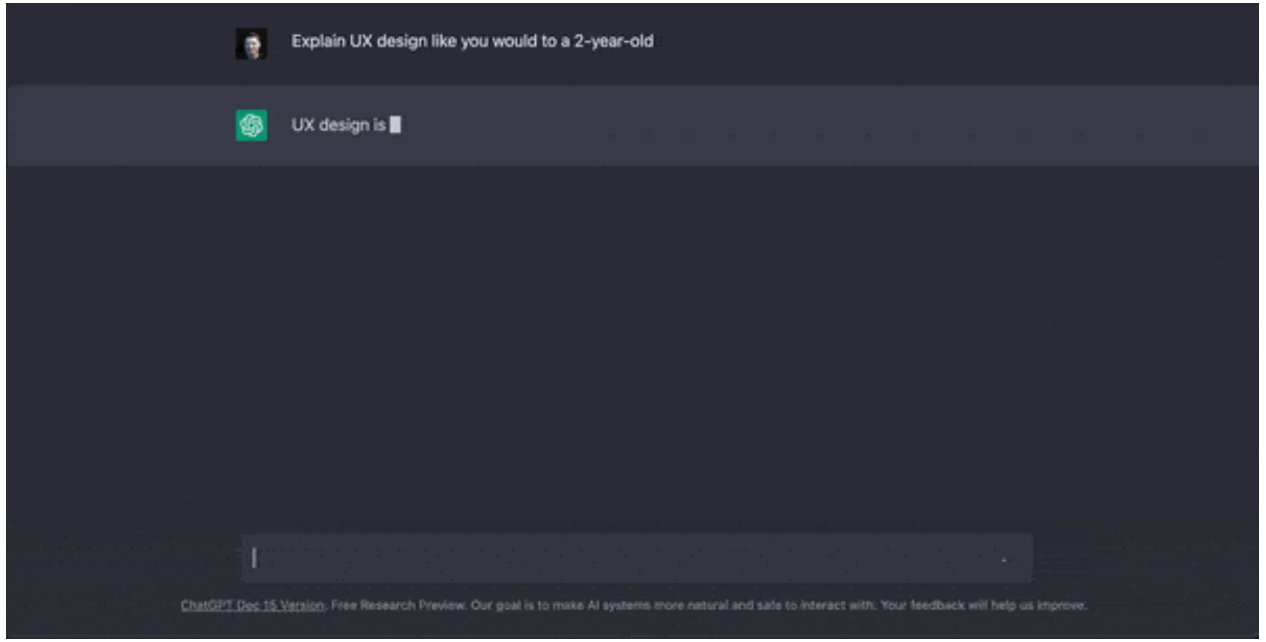
The ChatGPT ML problem: 3. Model

3. Model (Decoder)

Typically, **ChatGPT** uses **autoregressive RNNs** in its decoder part!



- Takes a sequence of words as input (user query), encodes it as a vector using a transformer-based model.
- **And then reconstructs an answer, a sentence, one word after the other!**

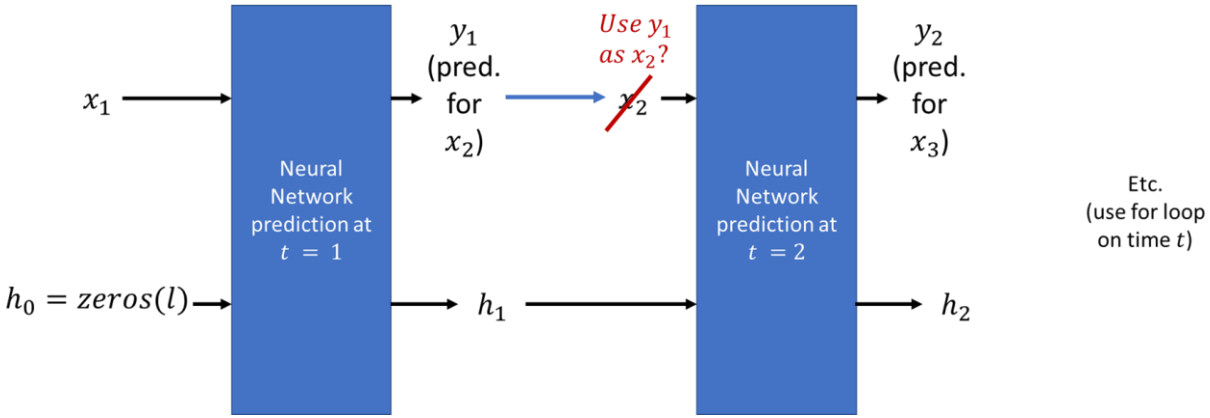


The ChatGPT ML problem: 3. Model

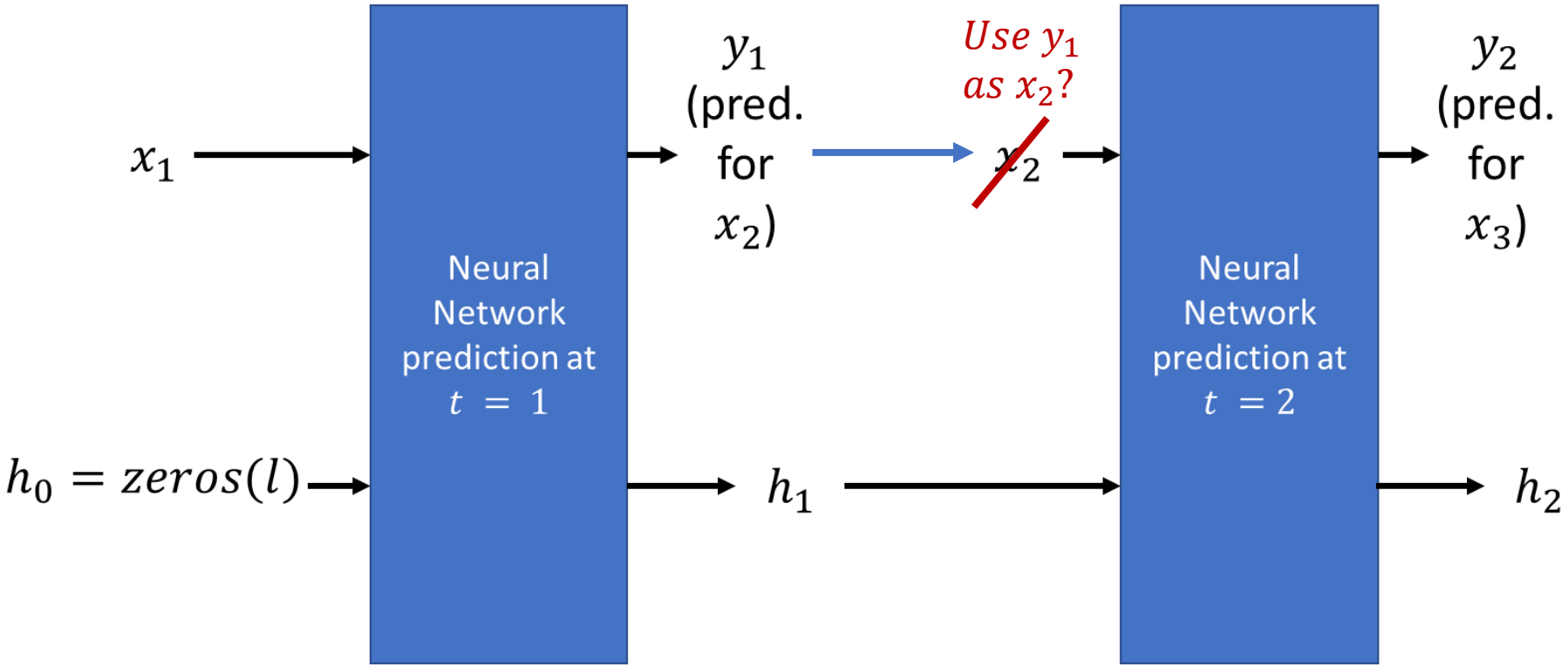
3. Model (Decoder)

Commonly referred to as **token prediction**:

- At each time step, the model predicts the next token in the sequence based on the context provided by the previous tokens.
- The final layer of the Transformer produces a probability distribution over the entire vocabulary, and the next token is selected based on this distribution.
- Also updates the encoding vector, to update which parts of the query the answer as already covered, RNN-style! (W6S2).



A quick word on autoregressive RNNs

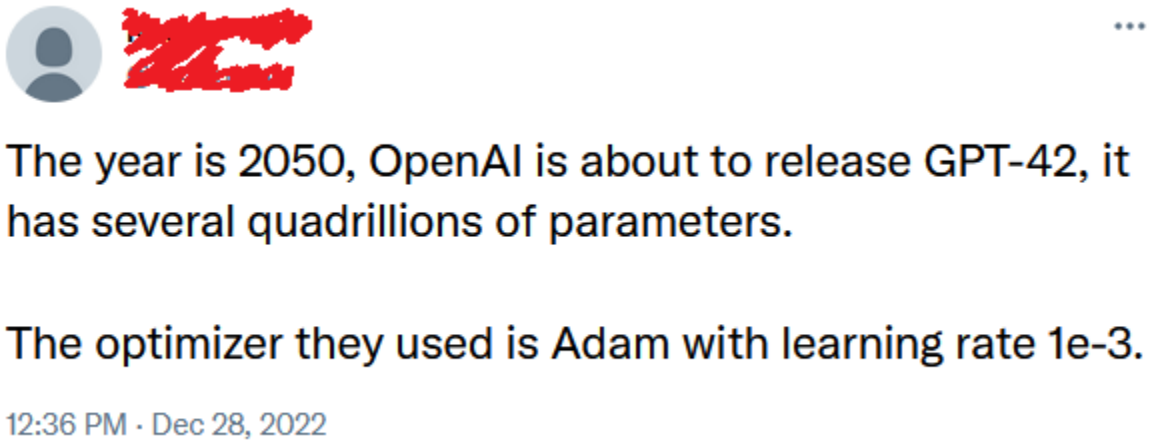


Etc.
(use for loop on time t)

The ChatGPT ML problem: 4. Loss/Training

4. Loss and Training

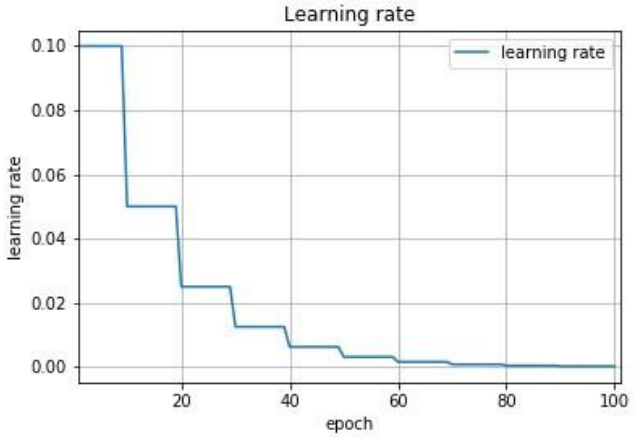
- At the end of the day, the token prediction task is a **classification** one (predicts from a finite number of words/characters)
- The primary objective function for training is most likely a **Maximum Likelihood Estimation** (MLE) or something similar.
- ChatGPT relies on advanced optimizers, such as the **Adam** optimizer. (Week 2!)



The ChatGPT ML problem: 4. Loss/Training

4. Loss and Training

- Also used some **learning rate scheduling/decay** (Week 2), according to their paper.
- **Regularization** (Week 1) also used, in the form of weight decay, which adds a penalty term to the loss function based on the magnitude of the model's parameters.
- Also used **dropout** (Week 4) as regularization, during training.



L1 Regularization

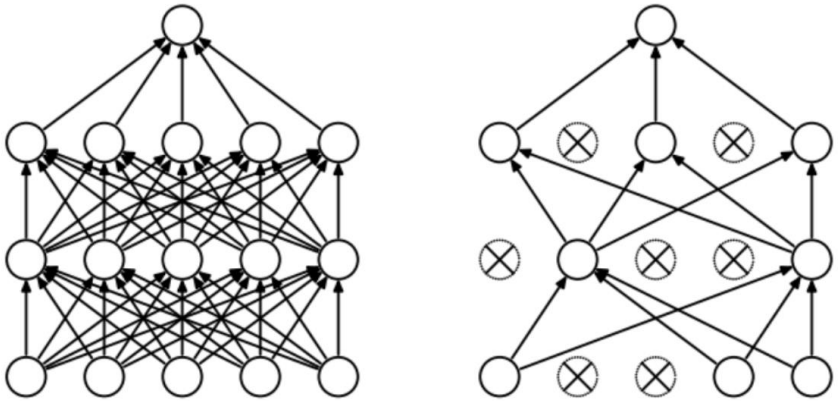
$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} W_j^2$$

Loss function

Regularization Term



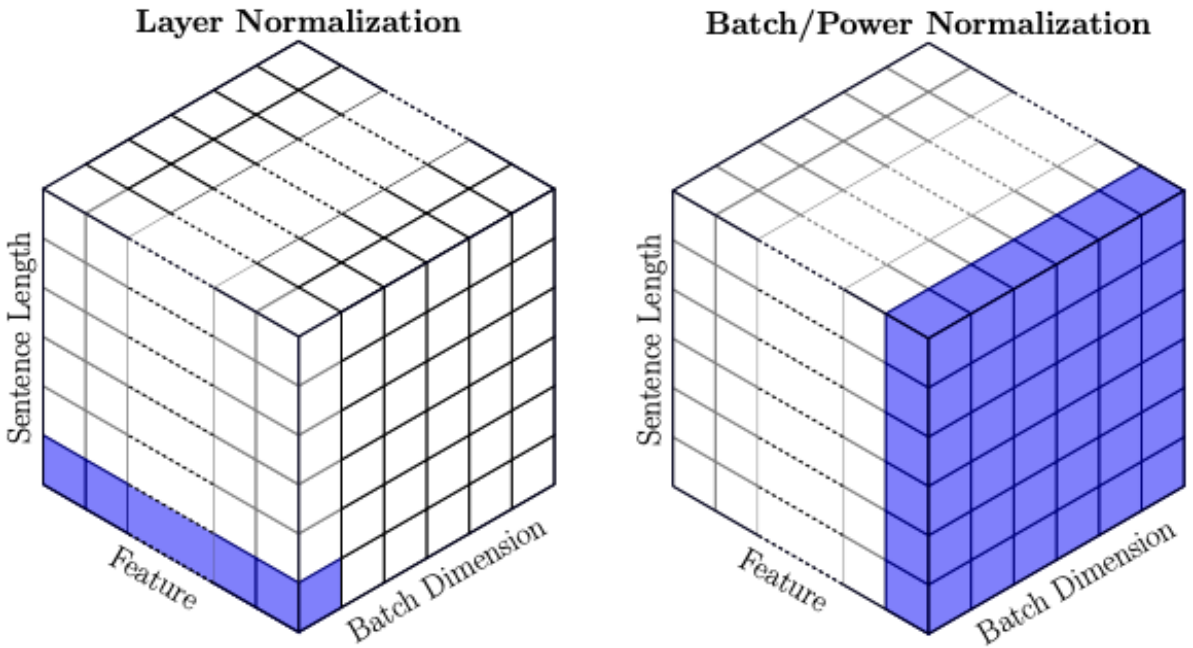
(a) Standard Neural Net

(b) After applying dropout.

The ChatGPT ML problem: 4. Loss/Training

4. Loss and Training

- **Layer normalization** (Week 4) is also applied within the Transformer architecture to stabilize the training process.
- Similar to **BatchNorm** (Week 4), it normalizes the inputs to each layer, ensuring that they have a consistent mean and variance.
- This approach helps prevent the **vanishing or exploding gradient problems (Week 2)**.



The ChatGPT ML problem: 4. Loss/Training

4. Loss and Training

As for **evaluation**…

During the training process, various **evaluation metrics** are used to monitor the model's performance and progress.

These typical NLP metrics include

- **Perplexity**, which is used to measure of how well the model predicts the next token in a sequence,
- and task-specific NLP metrics for fine-tuning, such as **F1 score**, **BLEU** score, or **ROUGE** score.

You will learn more about these metrics in NLP Term 7 course!

Fine-tuning the model

Fine-tuning is a crucial step in the development of models like ChatGPT, as it adapts the pretrained model to specific tasks or domains.

- **Pretraining** exposes the model to a wide range of language patterns and helps it learn general language understanding.
- **Fine-tuning**, on the other hand, tailors the model to perform specific tasks, such as answering questions, summarizing text, or generating contextually relevant responses in a conversation.

By fine-tuning the model on a more narrow and controlled dataset, it becomes more accurate, reliable, and safe to use in specific applications.

Fine-tuning the model

Models like ChatGPT have typically benefited from fine-tuning using **reinforcement learning techniques that leverage human feedback (W11S3)**.

RLHF involves (re-)training the (pre-trained) model using a reward function that is based on human preferences or judgments.

By collecting human feedback on model-generated outputs, a reward model can be created to evaluate the quality of the generated text.

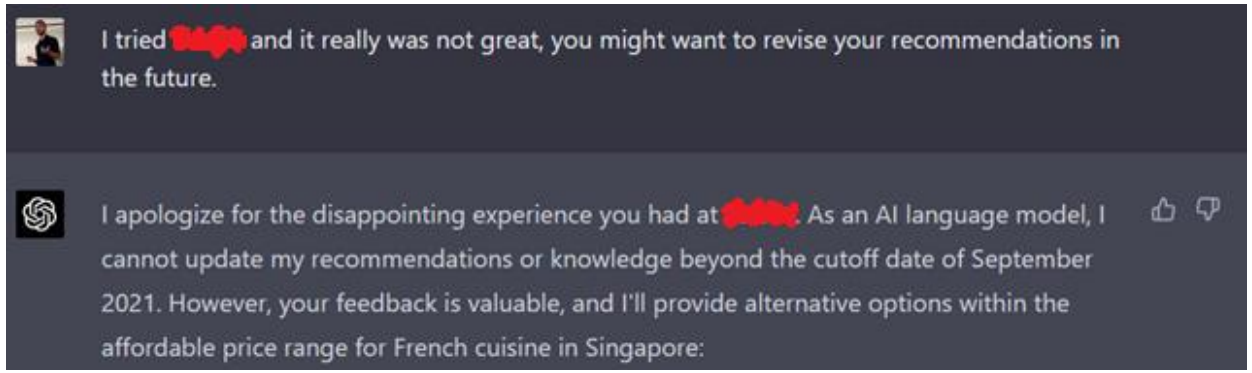
This approach helps the model learn from real-world feedback and further improves its performance, safety, and usefulness in generating human-like responses.

Fine-tuning the model

Typically, this also happen on-the-fly when you give feedback to the model based on its given answers!

- **Using the thumbs-up or thumbs-down icons,**
- **Or by following up on the answer.**

Feedback is probably integrated with a critic engine (as in W10S3 and W11S3) that tells the model how to retrain and improve!



Fine-tuning the model

Definition (Online Learning):

The process of **fine-tuning a model based on instantaneous feedback** is often referred to as **online learning** or **incremental learning**.

In online learning, the model **updates its parameters as new data samples/examples arrive**, allowing it to adapt to changing data distributions or to learn from new examples without the need to retrain on the entire dataset.

Online learning is particularly useful in situations where data is continuously generated or collected (e.g. a chatbot!).

It enables models to quickly adapt to new information and maintain their performance in dynamic environments.

ChatGPT is multimodal

ChatGPT exhibits **multimodality**.

- Can operate in different languages (either human, programming, etc.).
- Embedding function/model to be used will most likely differ from one language/task to another.

- From W8S1, about why we need different embeddings for different languages (English, French, Python, etc.) and different tasks.

Problem #3: Two different words could have really close meanings, but their embeddings may or may not need to be… and this decision could be task-specific.

- If the embedding is used for general language (non-medical), then we are probably fine with

$$\langle e_{covid}, e_{cold} \rangle \approx 1$$

As the only important information from these words is that they are both respiratory diseases.

- However, if the embedding will be used in a very specialized medical context, then it is probably better to have less similarity between the two words, i.e.

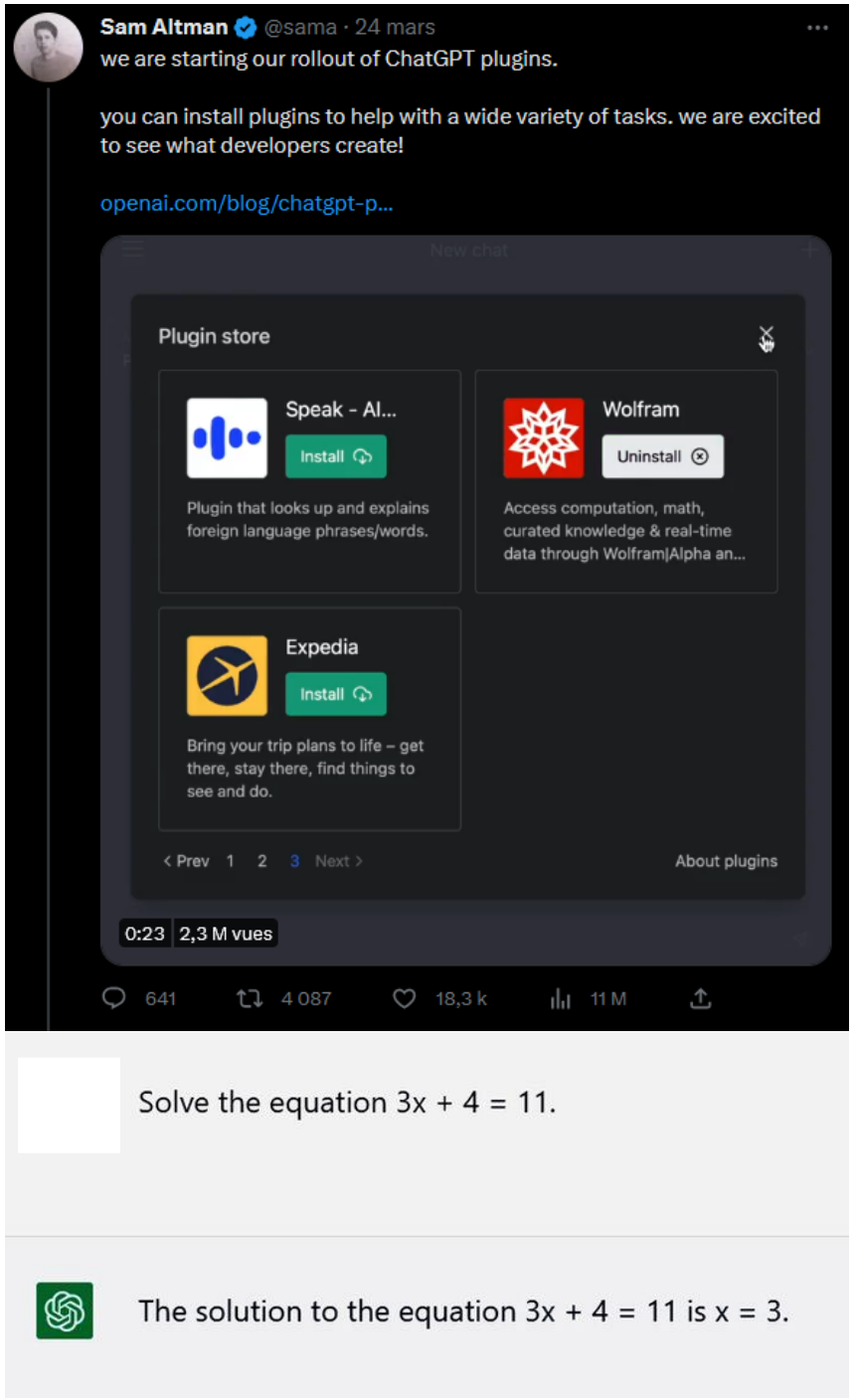
$$\langle e_{covid}, e_{cold} \rangle \in [0.5, 0.8],$$

- That is because, covid and cold are two different diseases, maybe sharing some similar symptoms, but to be treated very differently.

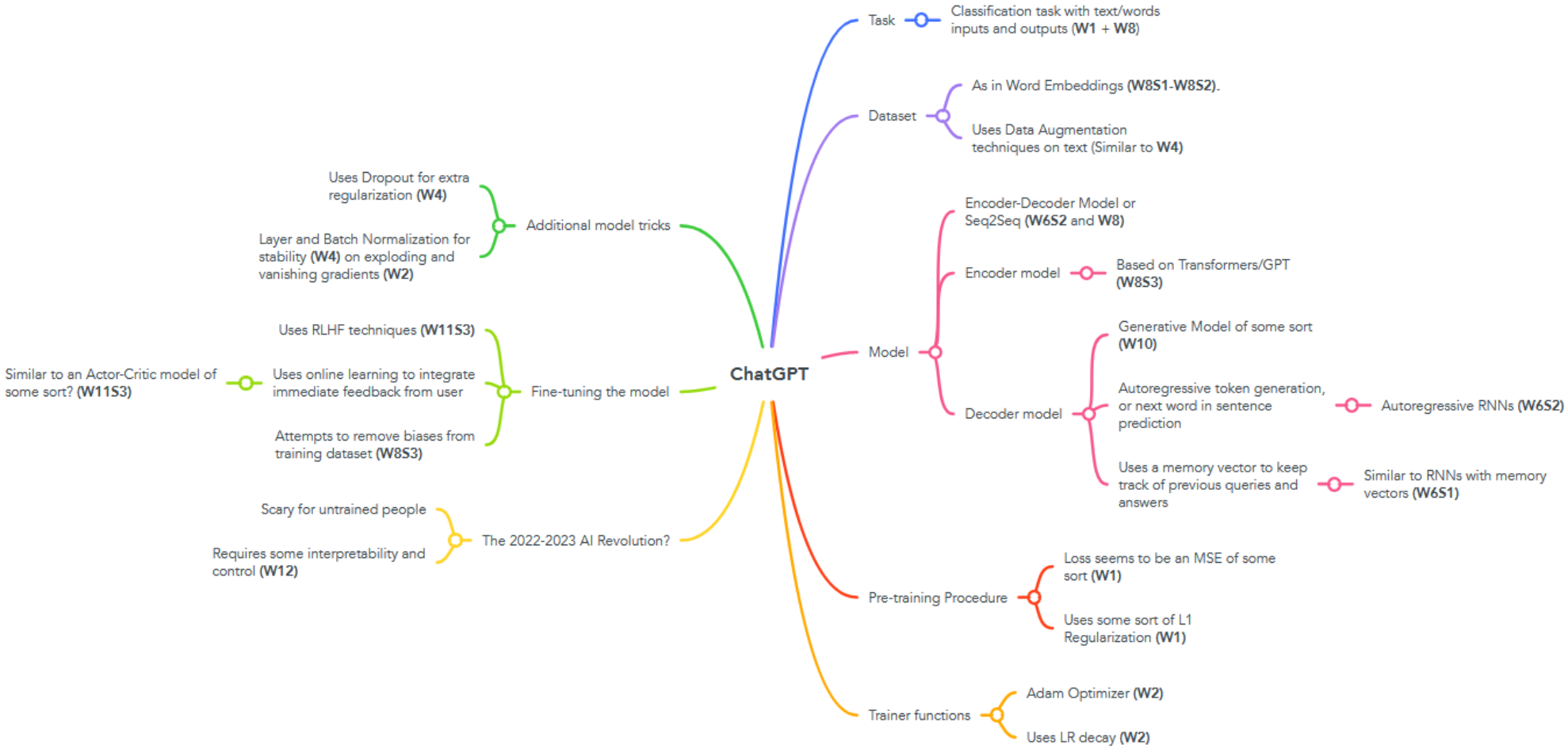
ChatGPT is multimodal

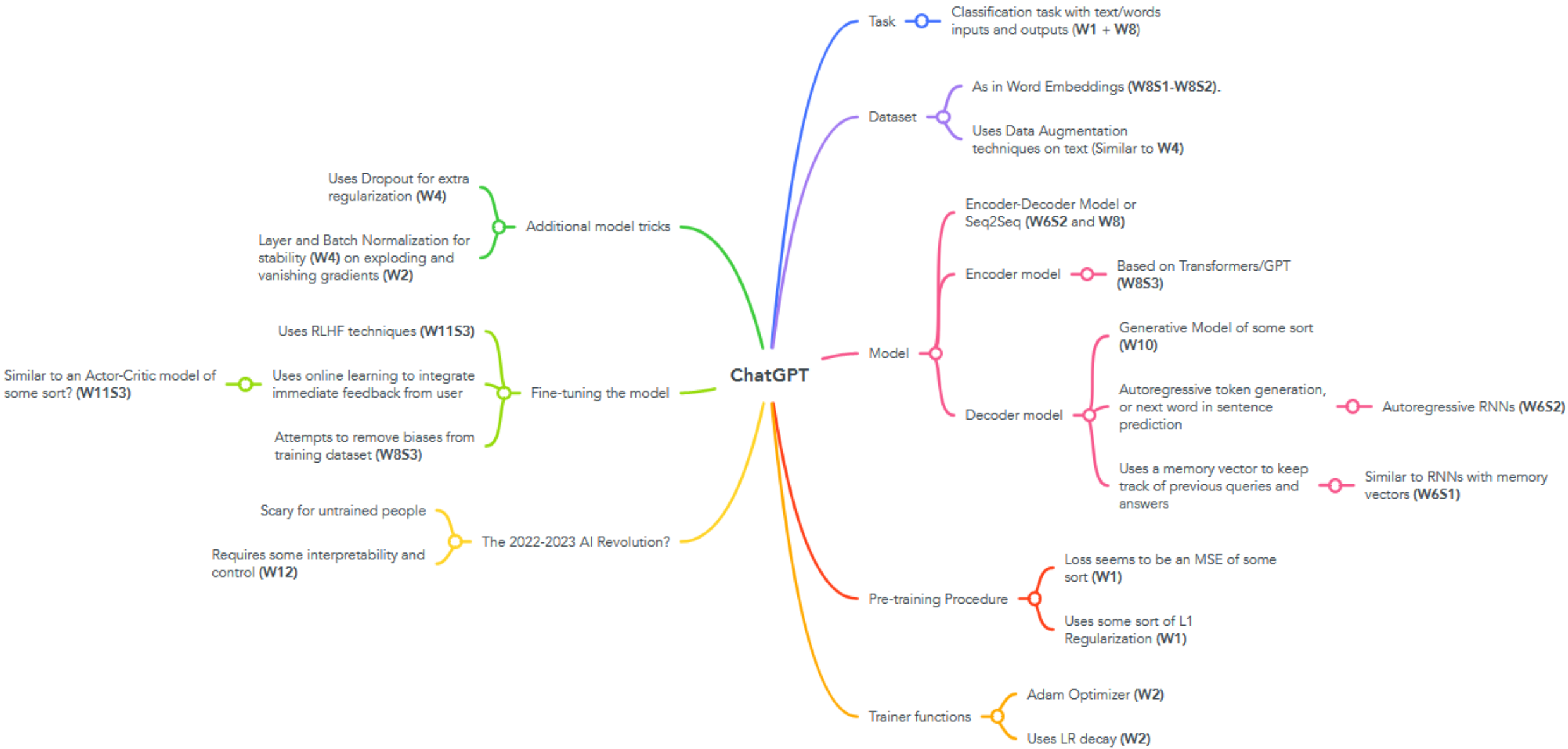
ChatGPT and other LLM chatbots are expected to become **even more multimodal**, for instance

- By incorporating other data types, such as images or audio, alongside text.
- By incorporating models trained explicitly for other tasks (e.g. Wolfram math engine for math questions/queries).
- *(Typically, math used to be a big issue for ChatGPT! Improving…)*



ChatGPT requires to assemble everything we know and learnt about Deep Learning this term





Can I recreate ChatGPT and run it?

- No, because 175 billion parameters in GPT3.5! And we suspect ~100 trillion for GPT4!
- Your laptop does not have this kind of memory!

- But many “nano” versions of ChatGPT out there, with reduced performance.
- For instance (but seriously don’t) https://github.com/karpathy/nanoGPT
- Also, Llama, released by Facebook, with “only” ~7 billion parameters https://github.com/facebookresearch/llama

A quick word on “Prompt Engineering”

Definition (Prompt Engineering):

Prompt engineering refers to the process of designing and refining the prompts used to interact with machine learning models like ChatGPT.

It aims to optimize the way queries are phrased to achieve better responses from the model.

The idea is that subtle changes in how a question or command is presented can result in significantly different outputs.

Many shenanigans…

Prompting Principles

- **Principle 1: Write clear and specific instructions**
- **Principle 2: Give the model time to “think”**

Tactics

Tactic 1: Use delimiters to clearly indicate distinct parts of the input

- Delimiters can be anything like: ```, """, < >, `<tag> </tag>`, `:`

In [ ]:

```
text = f"""
You should express what you want a model to do by \
providing instructions that are as clear and \
specific as you can possibly make them. \
This will guide the model towards the desired output, \
and reduce the chances of receiving irrelevant \
or incorrect responses. Don't confuse writing a \
clear prompt with writing a short prompt. \
In many cases, longer prompts provide more clarity \
and context for the model, which can lead to \
more detailed and relevant outputs.
"""
prompt = f"""
Summarize the text delimited by triple backticks \
into a single sentence.
```{text}```
"""
response = get_completion(prompt)
print(response)
```

Tactic 2: Ask for a structured output
```

A quick word on “Prompt Engineering”

Definition (Prompt Engineering):

Prompt engineering refers to the process of designing and refining the prompts used to interact with machine learning models like ChatGPT.

It aims to optimize the way queries are phrased to achieve better responses from the model.

The idea is that subtle changes in how a question or command is presented can result in significantly different outputs.

Many shenanigans…

- Learn more about with this short (1h) course? https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/

Is ChatGPT a threat?

- ChatGPT is possible the AI revolution for Year 2022-2023.
- *(We get one every 3-5 years or so, e.g. transformers, computer vision, etc.)*
- Revolutions are scary, and might have people freak out.
- While these reactions/fears are understandable, be smart about it.

Letter signed by Elon Musk demanding AI research pause sparks controversy

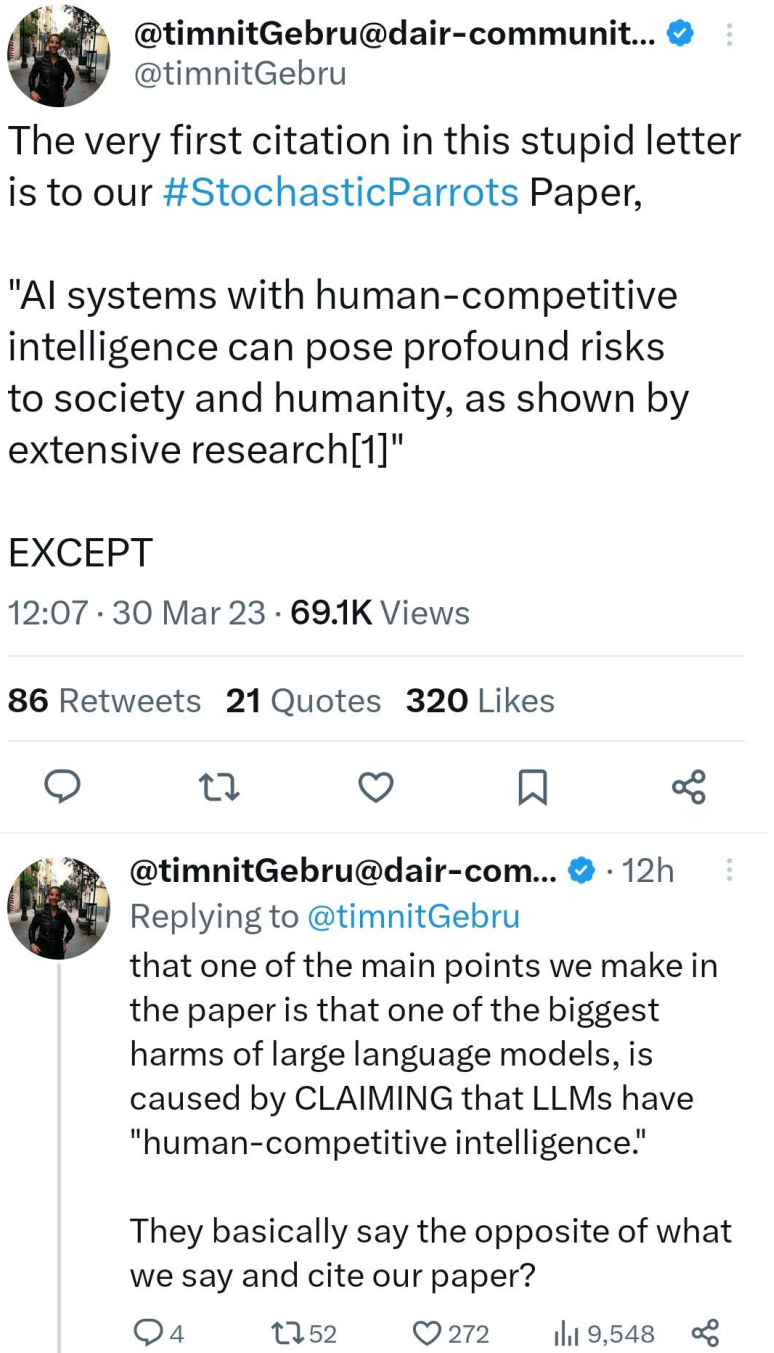
The statement has been revealed to have false signatures and researchers have condemned its use of their work

A letter called for a six-month pause on development of systems more powerful than GPT-4, developed by OpenAI, a company co-founded by Elon Musk. Photograph: Michael Dwyer/AP

A letter co-signed by Elon Musk and thousands of others demanding a pause in artificial intelligence research has created a firestorm, after the researchers

Is ChatGPT a threat?

- While these reactions/fears are understandable, be smart about it.
- **A letter that incorrectly cites papers, without checking on their authors is usually “sus”.**



@timnitGebru@dair-communit...
@timnitGebru

The very first citation in this stupid letter is to our #StochasticParrots Paper,

"AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1]"

EXCEPT

12:07 · 30 Mar 23 · **69.1K** Views

86 Retweets **21** Quotes **320** Likes

@timnitGebru@dair-com... · 12h
Replying to @timnitGebru

that one of the main points we make in the paper is that one of the biggest harms of large language models, is caused by CLAIMING that LLMs have "human-competitive intelligence."

They basically say the opposite of what we say and cite our paper?

4 52 272 9,548

Is ChatGPT a threat?

- While these reactions/fears are understandable, be smart about it.
- **A letter that incorrectly cites papers, without checking on their authors is usually “sus”.**
- **Many big names of DL have reported that their names appear in the list, but they never signed it.**

Letter signed by Elon Musk demanding AI research pause sparks controversy

The statement has been revealed to have false signatures and researchers have condemned its use of their work

A letter called for a six-month pause on development of systems more powerful than GPT-4, developed by OpenAI, a company co-founded by Elon Musk. Photograph: Michael Dwyer/AP

A letter co-signed by Elon Musk and thousands of others demanding a pause in artificial intelligence research has created a firestorm, after the researchers

Is ChatGPT a threat?

- While these reactions/fears are understandable, be smart about it.
- **A letter that incorrectly cites papers, without checking on their authors is usually “sus”.**
- **Many big names of DL have reported that their names appear in the list, but they never signed it.**
- **Be smart.**

Yann LeCun
@ylecun

The year is 1440 and the Catholic Church has called for a 6 months moratorium on the use of the printing press and the movable type.
Imagine what could happen if commoners get access to books!
They could read the Bible for themselves and society would be destroyed.

11:00 · 30 Mar 23 · **892K** Views

1,305 Retweets **265** Quotes **6,751** Likes

Yann LeCun @ylecun · 1d
Replying to @ylecun

Society *was* destroyed...
...for the better.
Printed books enabled the Protestant movement, and 200 years of religious conflicts in Europe.
But printed books also enabled the Enlightenment: literacy, education, science, philosophy, secularism, and democracy.

44 128 1,365 148K

In reality, the issue is...

Definition (AI alignment):

AI alignment refers to the problem of **designing AI systems that reliably and safely pursue the goals and objectives set by their human creators**.

The goal of **AI alignment** is to ensure that AI systems act in accordance with human values and do not cause harm, intentionally or unintentionally.

It involves **developing AI systems that are robust, interpretable, and aligned with the preferences and values of their human stakeholders**.

Achieving **AI alignment** is critical for ensuring that AI technology is used in ways that are beneficial to society and does not lead to unintended negative consequences. More on this on W12S1 - Explainability!

Is ChatGPT a threat?

My two cents: I do not think so.

- If anything, I prefer to see it as an opportunity.
- But I believe it will irremediably question/challenge/improve how we do things in many different fields (including teaching in universities!).

Learn more for yourself (below are some not so stupid yet accessible videos about the ChatGPT possible revolution):

- https://www.youtube.com/watch?v=Puo3VkPkNZ4
- https://www.youtube.com/watch?v=Sqa8Zo2XWc4