AdvNLE Seminar 10

Using LLMs and the Future

Dr Julie Weeds, Spring 2024

Previously …

- Distributional representations of meaning
- Neural Language modelling
- Contextualised word embeddings
- Large language models
  - BERT (Bi-directional Encoder Representations from Transformers)
  - Pre-training
  - Fine-tuning

Today

- More distant relatives of BERT
  - GPT and ChatGPT
- Generative applications with LLMs
  - MT
  - Summarization
  - Question-answering
- Trustworthy and responsible AI
- Environmental impact of LLMs
- Revision

More Distant Relatives of BERT

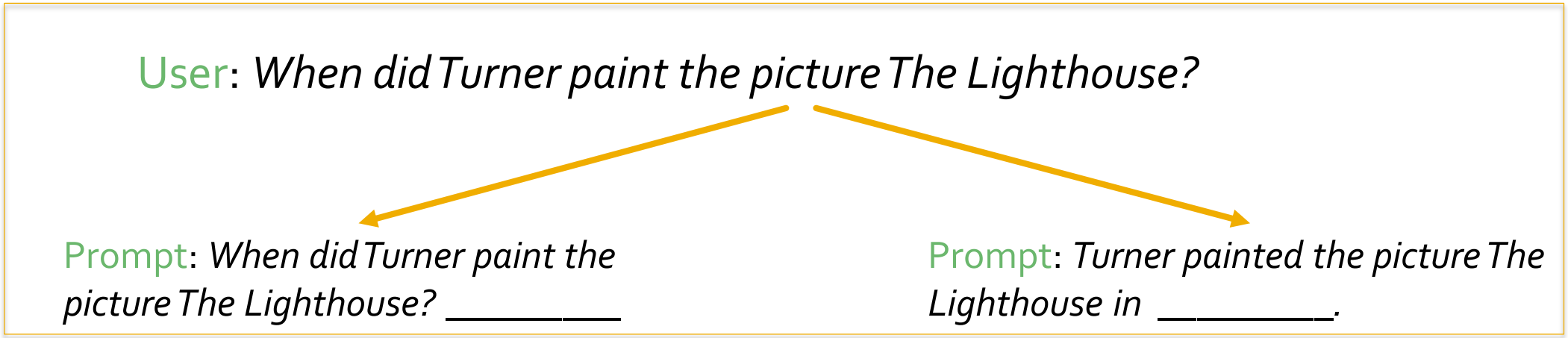
- Other Pretrained Large Language Models, generally still based on transformers e.g.,
  - GPT (GPT-2, GPT-3, ChatGPT GPT-4 …)
  - Turing-NLG,
  - XLNet,
  - Electra
  - Dolly
  - NeMo
  - BLOOM
  - LLaMa
  - PaLM2

Generative Pre-trained Transformer 3 (GPT-3)

- Brown et al. 2020: *Language Models are Few Shot Learners*
- autoregressive language model
  - this means it predicts the next token rather than masked tokens
- variable length inputs but uni-directional in nature
- largest non-sparse language model: 175 billion parameters, 10x bigger than competitors
- Trained on Common Crawl, WebText2, Books1, Books2 and Wikipedia
- No fine-tuning. Used to generate answers using a few-shot training / prompting paradigm

Generating responses

- Take a ***prompt*** and use a language model to predict what comes next or what fills in the gaps.
- Working out what is the best prompt strategy (how to convert the user utterance into a prompt for the LLM) is known as **prompt engineering**



- Find the word(s) that fills in the blank(s) with the highest probability according to the large language model (pretrained on the very large corpus)

What could possibly go wrong?

What could possibly go wrong?

- Large Language Models are trained to give plausible / believable answers based on large text corpora
- If 2 (or more words) often occur together in the training data, they may lead to a higher probability response even if factually incorrect in the specific context –
  - This is referred to as "***hallucination***" where a large language model adds information which has no basis in say the text being summarized or translated
- Text corpora may contain factually incorrect documents (e.g., fiction) and / or text which exhibits ***biases*** so these can be replicated in generated texts

Common hallucinations

- Based on your knowledge of LLMs, what do you think the most common types of errors and hallucinations are likely to be?

Genuine example (Feb 24)

You

Where in Rotherfield is the Six Bells pub?

ChatGPT

The Six Bells pub in Rotherfield is located at Church Street, Rotherfield, Crowborough, East Sussex, TN6 3LQ, United Kingdom.

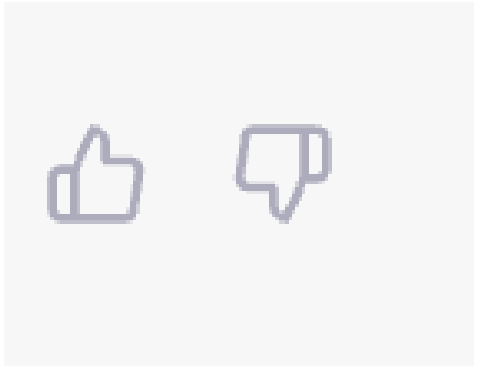
- There is no Six Bells pub in Rotherfield
- There is no Church Street in Rotherfield (only a Church Road)
- TN6 3LQ is Court Meadow, Rotherfield

ChatGPT (OpenAI, 2023)

- InstructGPT models
  - Trained with humans in the loop
  - Deployed as default language models on OpenAI's API
  - Better at following user intentions than GPT-3
  - More truthful and less toxic
  - Uses "alignment" technique
    - Reinforcement learning from human feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF)

- Customers submit prompts to the API
- Human labelers provide demonstrations of desired model behaviours
  - This data is used to fine-tune GPT-3
- Human labelers then rank different model outputs
  - This data is use to train a **reward model** to predict which output labellers would prefer
- GPT-3 has an additional input known as its "policy"
  - this is fine-tuned to maximise the reward using proximal policy optimization (PPO) (Schulman et al. 2017)



Training ChatGPT (Ouyang et al. 2022)

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

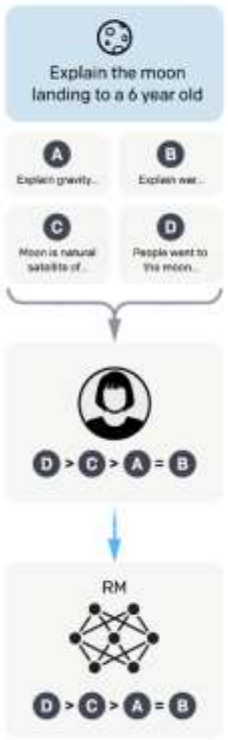
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

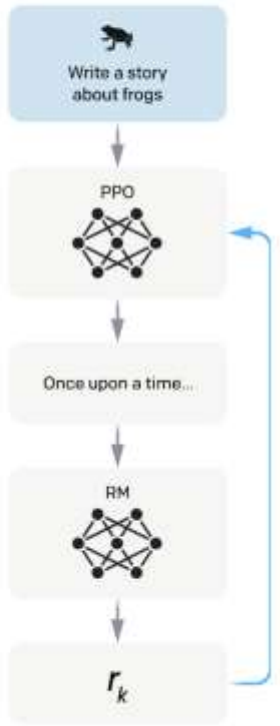
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

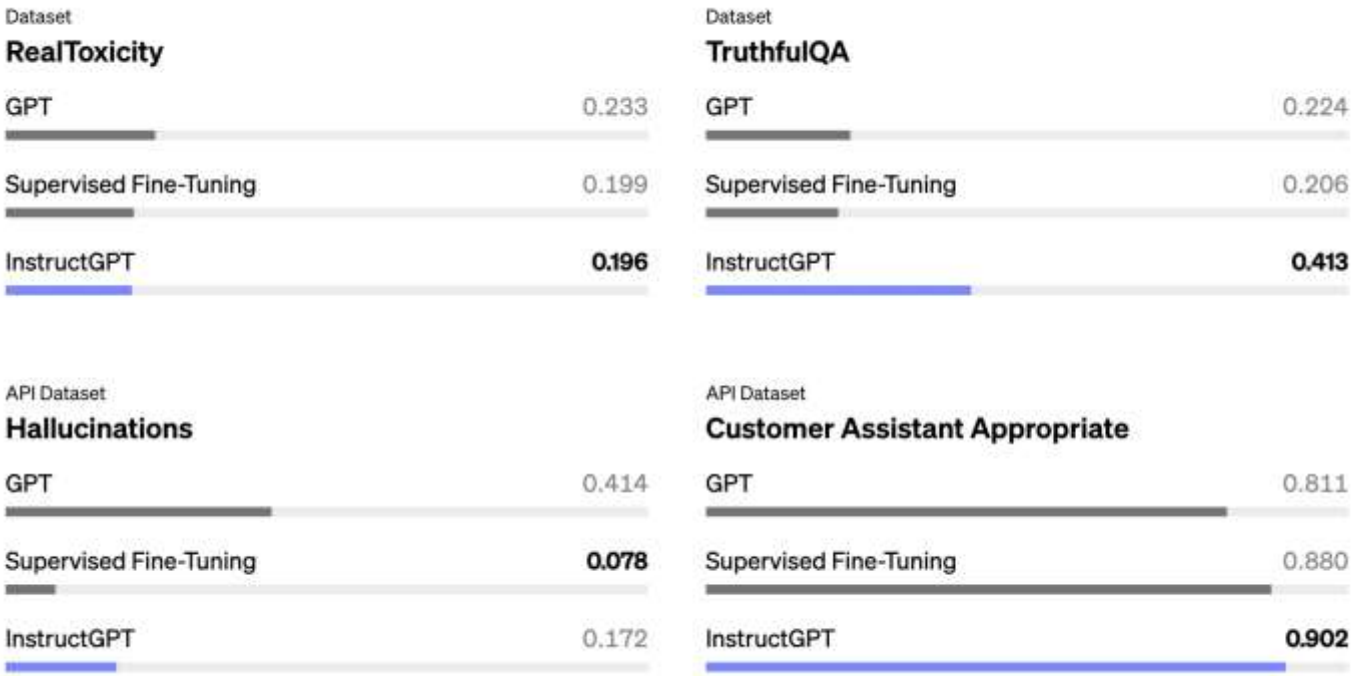
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Does it work?



According to OpenAI

- Less toxic
- More truthful
- Less hallucinations
- More appropriate

Yes, but …?

Weaknesses

ChatGPT is a pre-trained large language model

- It doesn't know or understand anything
- It doesn't look things up or carry out inferences
- Words which mean "similar" things are easy for it to mix up especially numbers and dates
- It has been trained on a very large corpus which is biased towards
  - a certain time period, particular geographic locations, culture and ways of thinking
- Fine-tuning process could be subverted

Using Generative Models

- Nearly all NLP applications can be posed as a prompt to a generative model:
  - Translate the following text from English to French: The cat sat on the mat
  - Summarise the following information in 3 sentences: …..
  - Who played Hans Solo in Star Wars?
  - What is the sentiment of the following review:
- **Encode** the prompt
- **Decode** and generate a response

Document – level MT (Wang et al. 2023)

Figure 1: An example of translating a document-level text from English to Chinese using GPT-4 (Date: 2023.03.17). We highlight the discourse phenomena using figures and lines, which are invisible to GPT-4.

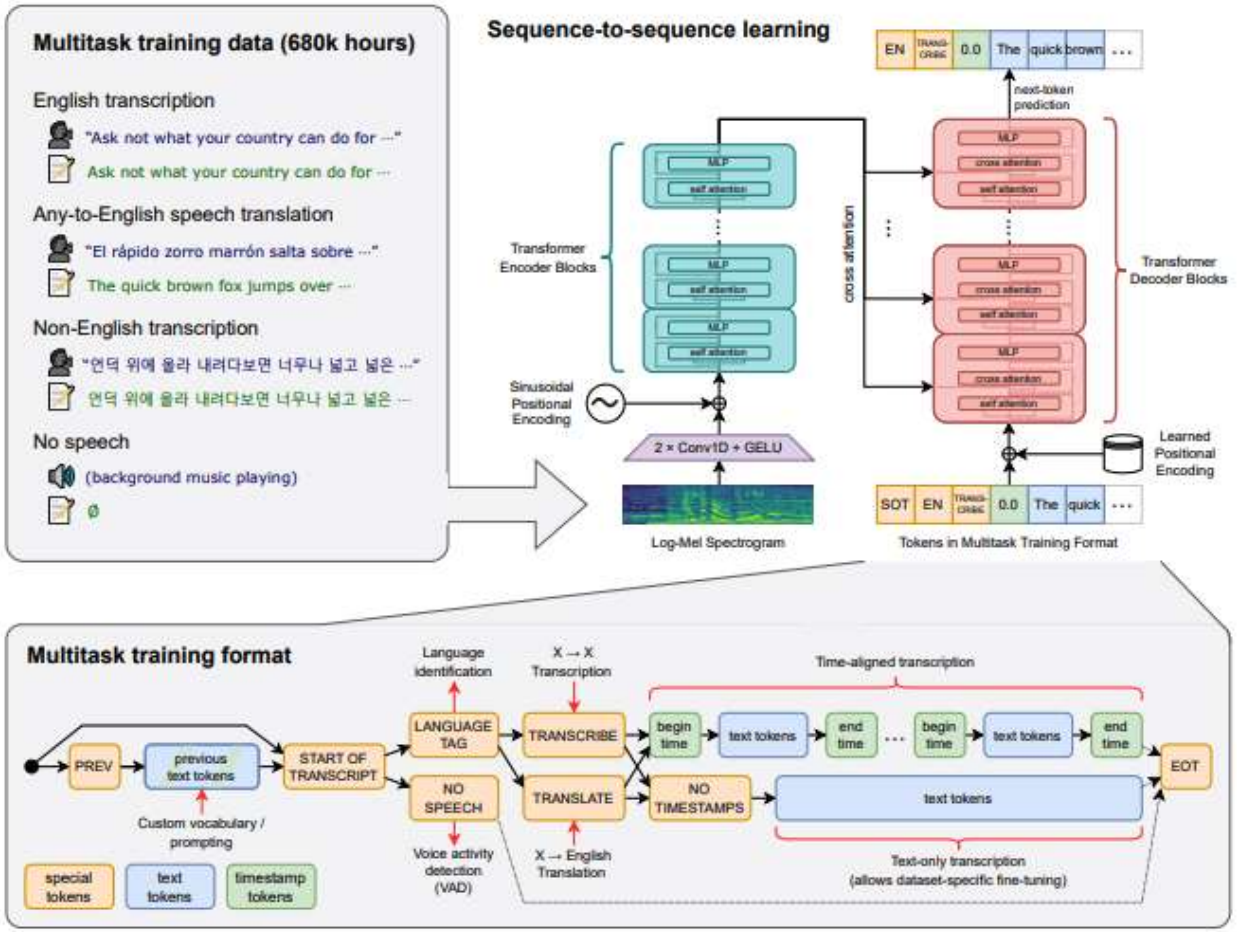
Model	Automatic (d-BLEU)					Human (General/Discourse)				
	News	Social	Fiction	Q&A	Ave.	News	Social	Fiction	Q&A	Ave.
Google	27.7	35.4	16.0	12.0	22.8	1.9/2.0	1.2/1.3	2.1/2.4	1.5/1.5	1.7/1.8
DeepL	**30.3**	33.4	16.1	11.9	22.9	2.2/2.2	1.3/1.1	2.4/2.6	1.6/1.5	1.9/1.9
Tencent	29.3	**38.8**	**20.7**	15.0	**26.0**	2.3/2.2	1.5/1.5	2.6/2.8	1.8/1.7	2.1/2.1
GPT-3.5	29.1	35.5	17.4	17.4	24.9	2.8/2.8	2.5/2.7	**2.8**/**2.9**	2.9/2.9	2.8/2.8
GPT-4	29.7	34.4	18.8	**19.0**	25.5	**3.3**/**3.4**	**2.9**/**2.9**	2.6/2.8	**3.1**/**3.2**	**3.0**/**3.1**

- Introduces evaluation metric which are “discourse aware”
- Shows the superiority of LLMs over other advanced MT systems
- Demonstrates the need for context-aware prompts
  - ChatGPT can translate sentence-by-sentence but remember the discourse-level context via the chat box
  - But best results generally by providing the whole document

Whisper (Radford et al. 2023)

- Trained to transcribe and translate audio
- Dataset constructed from audio paired with human-generated transcripts on the Internet
- Minimal pre-processing
- Audio language detection
- Multi-task format
- Zero-shot evaluation
- Available in Huggingface

Overview of Whisper



X → English	High	Mid	Low	All
XMEF-X	34.2	20.2	5.9	14.7
XLS-R (2B)	36.1	27.7	15.1	22.1
mSLAM-CTC (2B)	37.8	29.6	18.5	24.8
Maestro	**38.2**	31.3	18.4	25.2
Zero-Shot Whisper	36.2	**32.6**	**25.2**	**29.1**

Table 3. **X→en Speech translation performance.** Zero-shot Whisper outperforms existing models on CoVoST2 in the overall, medium, and low resource settings but still moderately underperforms on high-resource languages compared to prior directly supervised work.

Machine Translation: remaining challenges

- Long documents can still pose problems
- Stylised translation
- Interactivity
- Multi-modal
- Evaluation – quality and interpretability / explanations

Benchmarking LLMs for News Summarization (Zhang et al. 2024)

Article: https://www.dailymail.co.uk/femail/article-3052124

Reference summary:

Rare tote is so exclusive that it can't even be brought in store. Experts say it's an extremely good investment as only increases in value. Kim Kardashian and Victoria Beckham are both fans of Hermes.

0-shot GPT-3 Davinci summary:

The most expensive handbag in the world is a pink Hermes Birkin bag, which has been put on sale for £140,000. The bag is so exclusive that it can't even be brought in store and experts say it's an extremely good investment. With the same amount of money that the handbag costs, you could splash out on an Aston Martin DB9 Volante.

What is the main idea of the article?

0-shot Instruct Davinci summary:

The article discusses a pink Hermes Birkin bag that is on sale for £140,000. The bag is made-to-measure and embellished with rubies. The article also discusses how the bag is a collector's item and an investment.

Freelance writer summary:

Hermes is a brand rich in history and value, so much so that experts are saying their new £140,000 handbag is an extremely good investment. The bag, studded in rubies, is only the latest in Hermes Birkin's long line of highly expensive handbags for highly stylish women, which include celebrities and royalty.

Figure 2: Examples summaries generated by GPT-3 models (Section 3) or written by freelance writers (Section 4) of an article from the CNN/DM dataset. We find that the instruction-tuned GPT-3 model can generate a much better summary compared to the non-instruction-tuned variant. The reference summary from CNN/DM is not coherent whereas the freelance writer summary is both coherent and relevant.

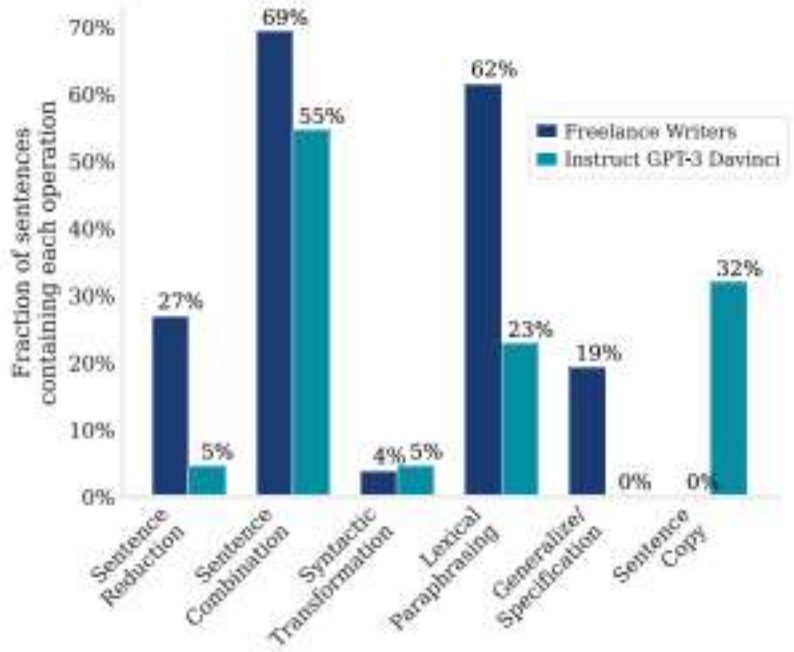


Figure 4: Distributions of cut and paste operations in the summaries written by freelance writers and by Instruct Davinci. By comparison, human-written summaries contain more lexical paraphrasing and sentence reduction whereas the Instruct Davinci model has more direct copying from the article.

Retrieval Augmented Generation (Gao et al. 2023)

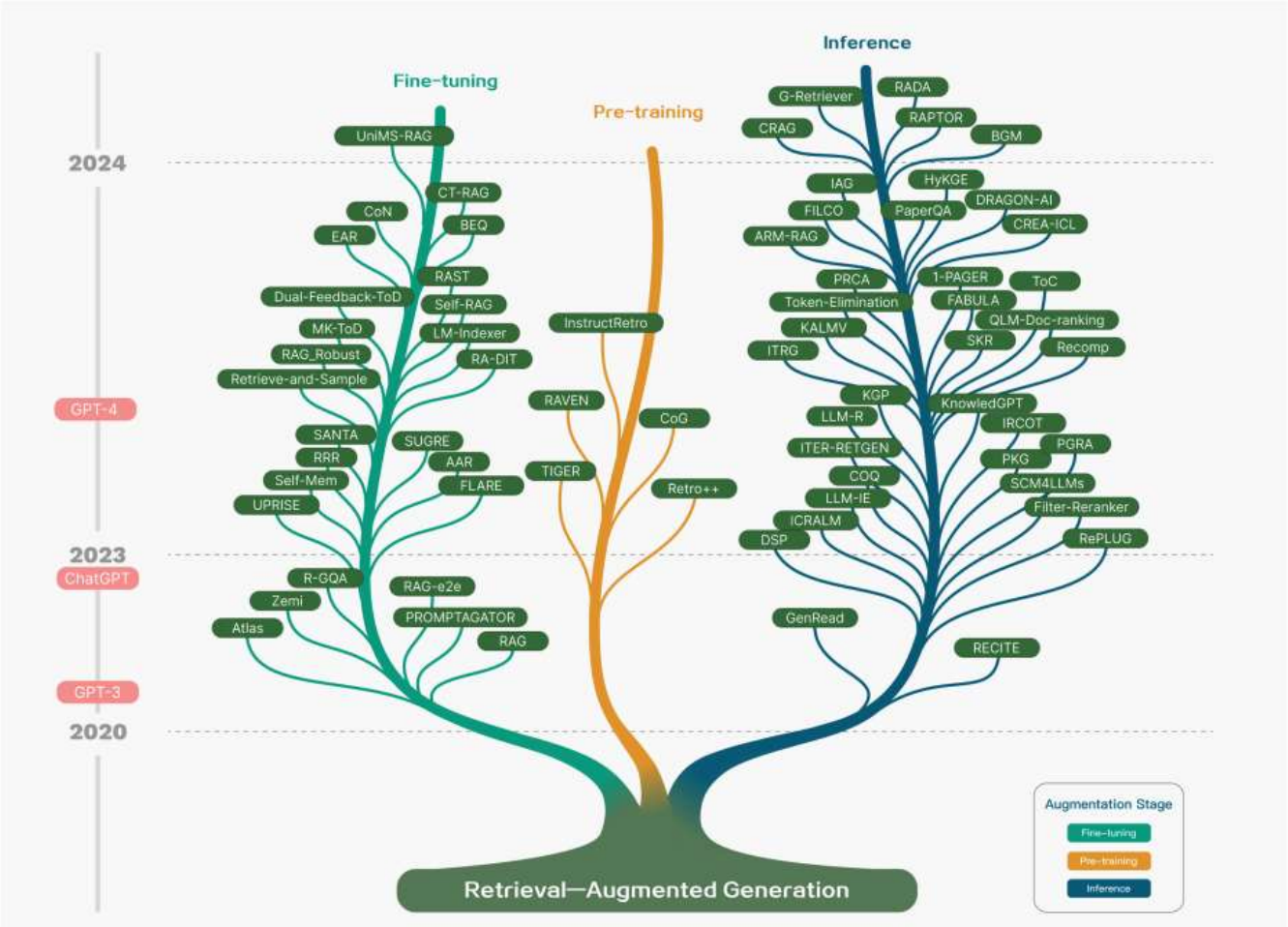


Fig. 1. Technology tree of RAG research. The stages of involving RAG mainly include pre-training, fine-tuning, and inference. With the emergence of LLMs,

Retrieve-Read Framework for RAG

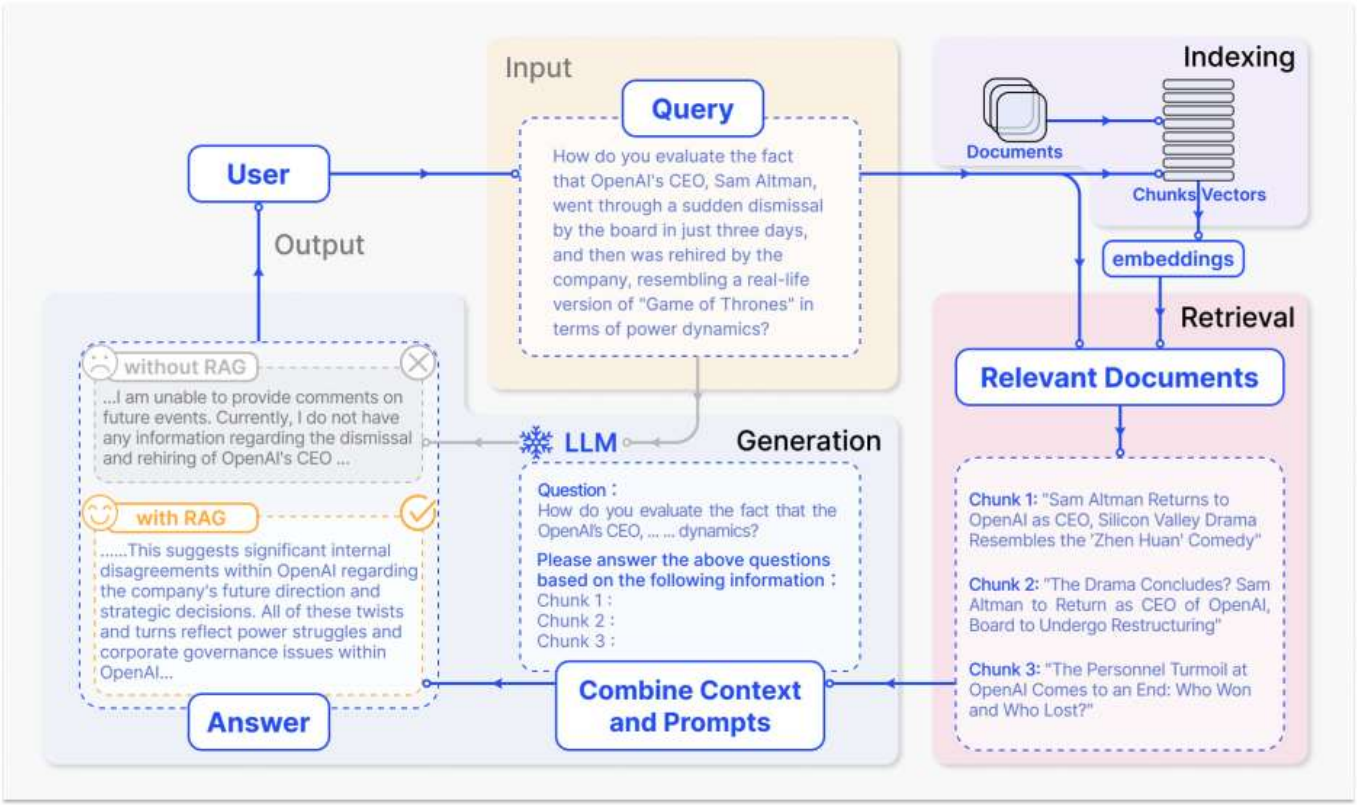


Fig. 2. A representative instance of the RAG process applied to question answering. It mainly consists of 3 steps. 1) Indexing. Documents are split into chunks, encoded into vectors, and stored in a vector database. 2) Retrieval. Retrieve the Top k chunks most relevant to the question based on semantic similarity. 3) Generation. Input the original question and the retrieved chunks together into LLM to generate the final answer.

More Advanced RAG

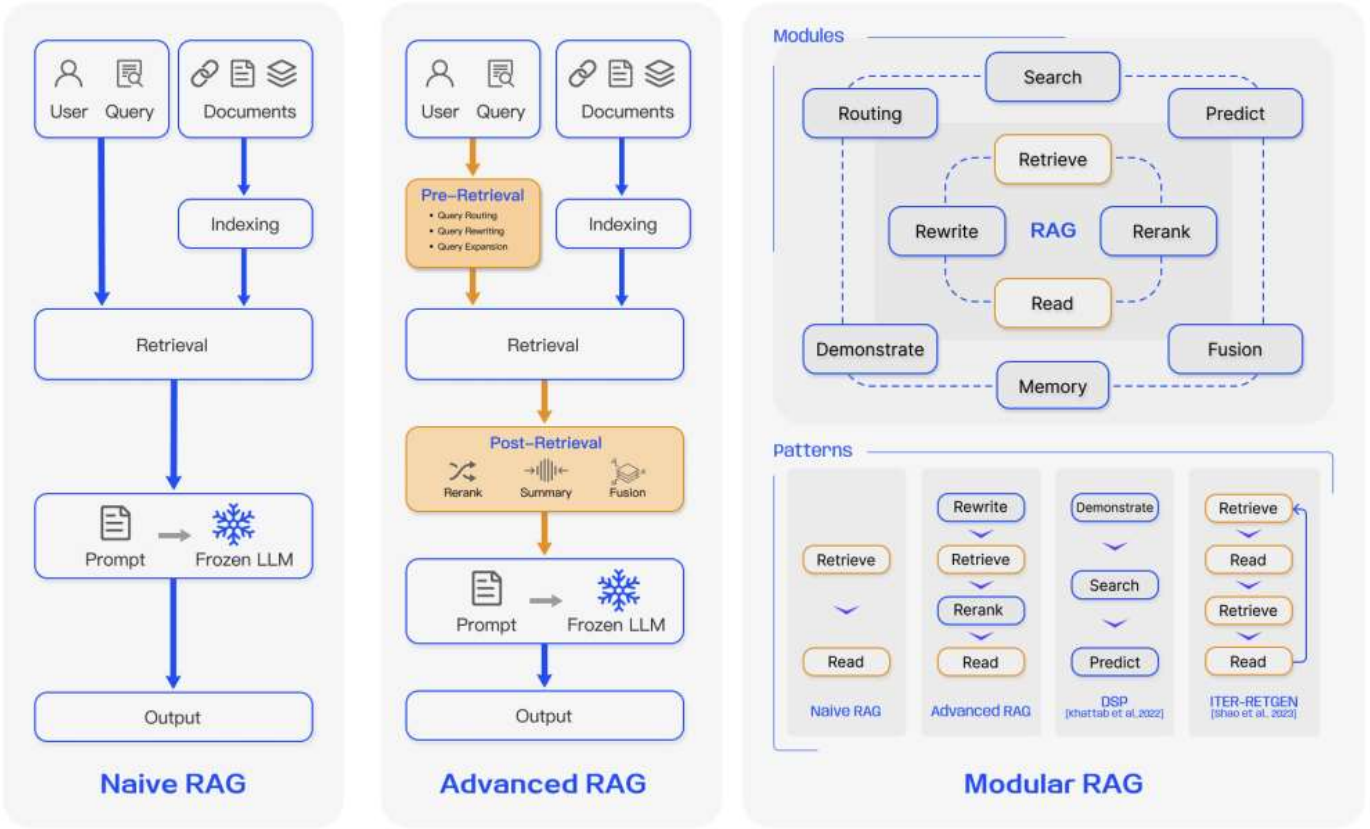


Fig. 3. Comparison between the three paradigms of RAG. (Left) Naive RAG mainly consists of three parts: indexing, retrieval and generation. (Middle) Advanced RAG proposes multiple optimization strategies around pre-retrieval and post-retrieval, with a process similar to the Naive RAG, still following a chain-like structure. (Right) Modular RAG inherits and develops from the previous paradigm, showcasing greater flexibility overall. This is evident in the introduction of multiple specific functional modules and the replacement of existing modules. The overall process is not limited to sequential retrieval and generation; it includes methods such as iterative and adaptive retrieval.

Ethical considerations

- What else do you need to consider before using an advanced LLM system to generate text for you?

Accountability

By Maria Yagoda 23rd February 2024

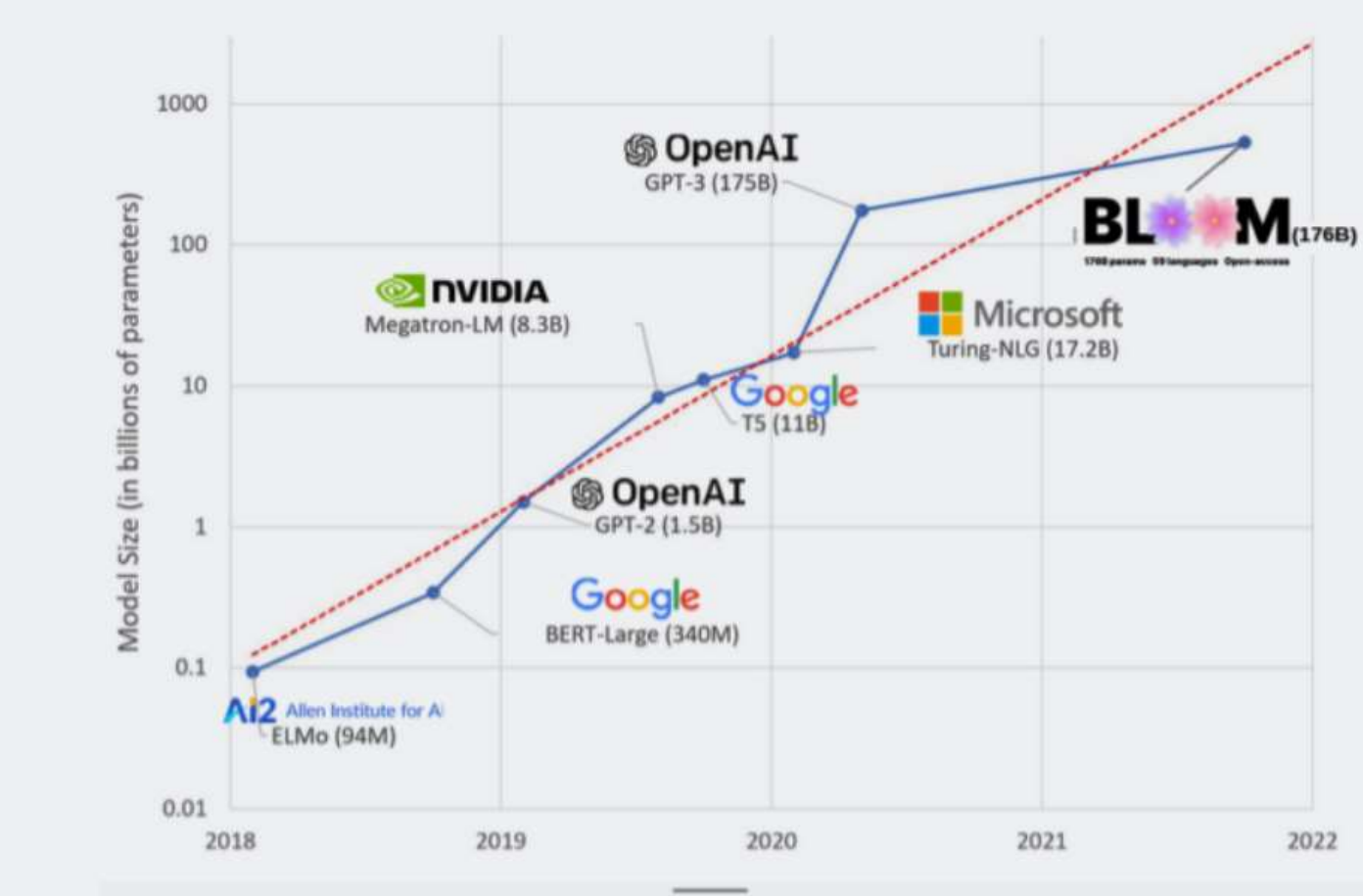
When Air Canada's chatbot gave incorrect information to a traveller, the airline argued its chatbot is "responsible for its own actions".

Artificial intelligence **is having a growing impact on the way we travel,** and a remarkable new case shows what AI-powered chatbots can get wrong – and who should pay. In 2022, Air Canada's chatbot **promised a discount** that wasn't available to passenger Jake Moffatt, who was assured that he could book a full-fare flight for his grandmother's funeral and then apply for a bereavement fare after the fact.

According to a civil-resolutions tribunal **decision** last Wednesday, when Moffatt applied for the discount, the airline said the chatbot had been wrong – the request needed to be submitted before the flight – and it wouldn't offer the discount. Instead, the airline said the chatbot was a "separate legal entity that is responsible for its own actions". Air Canada argued that Moffatt should have gone to the link provided by the chatbot, where he would have seen the correct policy.

The British Columbia Civil Resolution Tribunal rejected that argument, ruling that Air Canada had to pay Moffatt $812.02 (£642.64) in damages and tribunal fees. "It should be obvious to Air Canada that it is responsible for all the information on its website," read tribunal member Christopher Rivers' written response. "It makes no difference whether the information comes from a static page or a chatbot." The BBC reached out to Air Canada for additional comment and will update this article if and when we receive a response.

Model Growth



Enlarge / Model size growth over the years.

Source: https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/

Environmental Impact of LLMs

Model name	Number of parameters	Datacenter PUE	Carbon intensity of grid used	Power consumption	CO_2eq emissions	CO_2eq emissions × PUE
GPT-3	175B	1.1	429 gCO_2eq/kWh	1,287 MWh	*502 tonnes*	552 tonnes
Gopher	280B	1.08	330 gCO_2eq/kWh	*1,066 MWh*	*352 tonnes*	380 tonnes
OPT	175B	*1.09* [2]	*231 gCO_2eq/kWh*	*324 MWh*	70 tonnes	*76.3 tonnes* [3]
BLOOM	176B	1.2	57 gCO_2eq/kWh	433 MWh	25 tonnes	30 tonnes

Table 4: Comparison of carbon emissions between BLOOM and similar LLMs. Numbers in *italics* have been inferred based on data provided in the papers describing the models.

Sasha Luccioni, et al.

- high levels of energy use and CO2 emissions
- Training GPT-3 and Llama 2 required around 1.3 GWH → 552 tonnes of CO2 → 2 or 3 full Boeing 767s flying round-trip from New York to San Francisco
- Electricity use of ChatGPT in inference likely surpasses that of training within weeks or days (inference counts for 90% of AI workloads) → more like using a Boeing 767 to transport a single passenger at a time

What should we do?

Part 2 : Revision

Revision questions

1. Name and give examples of different semantic relationships which can hold between *distributionally similar* words.
2. What is Zipf's Law and why is it a problem?
3. What are the main similarities and differences between word2vec and GLoVE?
4. Explain how to use a trigram model to compute the probability of a sentence.
5. What is perplexity?
6. What advantage do LSTMs have over vanilla RNNs in language modelling?
7. How and why might you combine a character-based network with a word-based network in language modelling?
8. For what types of problems are CRFs typically used?
9. What's the difference between a generative statistical classification model and a discriminative statistical classification model?
10. When and why might it be better to use F1 rather than accuracy as an evaluation metric?

More questions

1. In a multi-class scenario, how would you calculate micro-average F1 and macro-average F1. Which is better and why?
2. Describe 2 different ways word embeddings might be combined to make sentence embeddings
3. Give an example of structural differences between languages.
4. Outline 2 different methods for evaluating machine translation systems.
5. How might an encoder-decoder network be used for MT?
6. What is subword tokenization?
7. In an attention head, what are the 3 different vectors which are created? How are they created and how are they used?
8. What is the input representation used by BERT?
9. What is the difference between masked language modelling and autoregressive language modelling?
10. What is transfer learning? Explain with reference to BERT