Large Language Models

How Do Large Language Models (LLMs) Work?

- At their most basic LLMs are statistical pattern-recognition and prediction systems

- LLMs output the next likely word (“token”) in a sentence (“sequence”)
  - token: unit of text e.g. word, character. 1 word ~ 0.75 token
  - sequence: context - section (“window”) of text e.g. sentence, paragraph, book
  - input into chatGPT is 4096 tokens; Claude 2 is 100K tokens

- The likelihood of the next work appearing is determined by
  - the context in which the words are seen in a larger body of text (“corpus”) and
  - the input to the chat

LLMs “Understand” “Meaning”

- Learning from a large corpus allows LLMs to understand the meaning of words.

For example

  - the training data may consist of many sentences beginning with “my favourite colour is…”
  - the next word will be a colour, allowing LLMs to cluster the words “red, blue, green…” into a set that represents the concept of “colour”

- It’s important to note that LLMs don’t really understand anything. They create statistical patterns that groups similar tokens based on a complex measure of how similar or dissimilar they are.

Data used to Train LLMs

- LLMs are trained in an unsupervised manner on vast quantities of open source and licensed data e.g.
- Responses are refined using question-response pairs (“InstructGPT”) from the web, humans or bootstrapped (i.e. the LLM outputs its own pairs)

- reinforcement learning with human feedback (RLHF) is used to reward LLMs to give appropriate responses (“guardrails”)

- “Constitutional AI” – trained to filter responses based on e.g. Universal Declaration of Human Rights (Claude 2)

Next Word Prediction

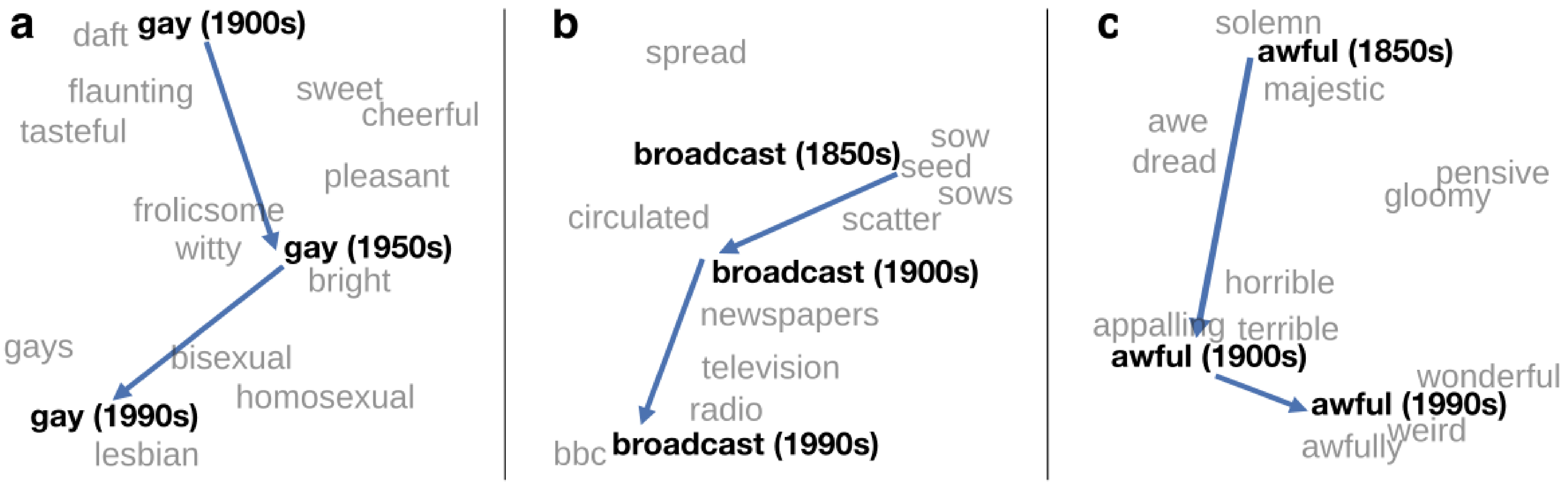
- is influenced by the frequency the word is seen in various contexts
- but there is a degree of randomness so that the word with the highest probability isn’t always seen

My favourite colour is

green	9.7%
red	15%
pink	11.6%
puce	2.3%

Meanings can change based on context

In each these examples the meaning of the same word changes over time



LLMs can (seem to) be creative

- A consequence of context-based learning and randomness allows the LLMs to generate surprising outputs.
  - Note though that they’re not creative in a human sense, but driven by pattern recognition and prediction algorithms

We can use LLMs to:

- identify weakly similar concepts from different disciplines and help understand different disciplines
- generate diverse narratives
- help with ambiguity
- role playing

Beware!

- LLMs may seem to “lie” and “hallucinate” i.e. give what are factually-incorrect responses to questions*
  - as you now know, they’re not trained to do give you an objectively correct answer!
- this is some function of training data (e.g. bias), learning, search and probability
- don’t believe the outputs – they **always** need checking, at least for now

Interacting with LLMs using “Prompt Engineering”

- Remember that the output of a LLM is determined by both what the system has been trained on and what information you give it

- Prompt engineering means tailoring your questions and input so you can get the most out of an LLM

- Prompts can take many forms, from instructing the LLM to take on a role (e.g. a helpful teacher, a pirate) or guiding the way it should process its output (e.g. “chain of thought” or a particular method).

LLMs can help you engineer prompts

- prompts shouldn’t be too precise (“What’s the capital of England?”), or too vague (“Tell me about sustainability”)

- sometimes you may not know how to ask an LLM to do a task
- ask it what it needs and collaborate with it

E.g.

- “What could I ask you to help me refine my aims for an essay?”
- “Do you need any more information?”

What you can’t do, some things you could do, and why you should do

Can’t

- cAI cannot be an author or co-author; you can’t get it to write for you in whole or in part
- cAI cannot be cited or referenced
- cAI cannot think for you!

Could

- proof-reading, but why not use Grammarly?
- identify publications, but why not use Scopus/Scholar/Elicit/ScholarAI
- summarising ideas, but why not use Wikipedia?
- it can suggest ways to restructure, but why not speak to your supervisor?