AI in Qualitative Analysis
CPDE Lunch & Learn

Simon Stone

Research Software Engineer for HPC and AI

Research Computing @ ITC, Dartmouth College

Introducing Research Software Engineering

Collaborative expertise in software engineering, designed to bridge the gap between innovative ideas and impactful outcomes. Our services include:

- 🤝 **Grant Proposal Consulting** to ensure accurate resource estimations and project feasibility.
- 🚀 **Rapid Prototyping** to refine concepts and explore solutions.
- ⛑️ **Ongoing Application Support** and **Application Rehabilitation** for existing applications.
- 🌎 **Open-Source Releases** to share knowledge and contribute to the wider community.

Contact us today to discuss your project and discover how Research Software Engineering can be your trusted partner in innovation.

Why talk about this?

- 🎉 Artificial Intelligence (AI) seems to be everywhere, all of a sudden
- 🏋️ Large Language Models' (LLM) capabilities are impressive
- 🌊 A wave of new and shiny AI tools is flooding the scene
- 🤖 AI components are making their way into many analysis workflows and tools
- 🤔 AI literacy is a critical skill for everyone

Objectives of this talk

- 💬 Explain how Large Language Models (de-)construct language
- 🧭 Show how AI can help explore qualitative data
- 🔬 Equip you with critical knowledge to assess new tools and techniques
- Spark ideas and conversation

How Large Language Model's (de-)construct language

What *can't* AI do?

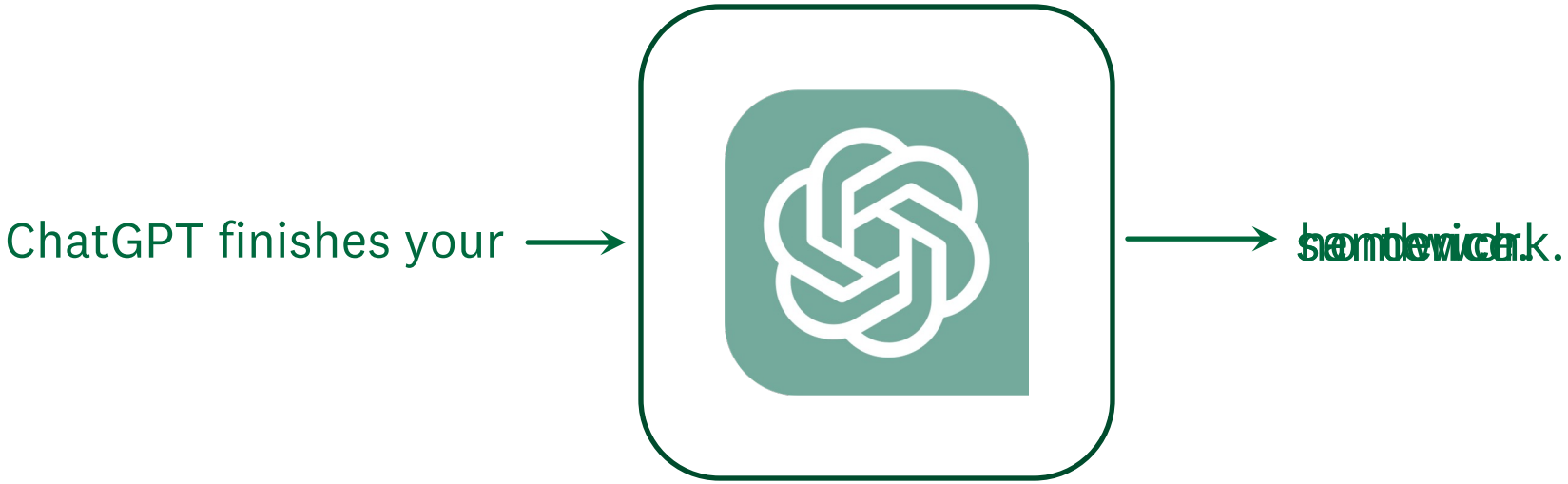
🥇 LLMs can do exactly one thing:

- 🔮 Predict the next word given a sequence of input words

🦜 Everything else they do is a clever use of this singular ability!

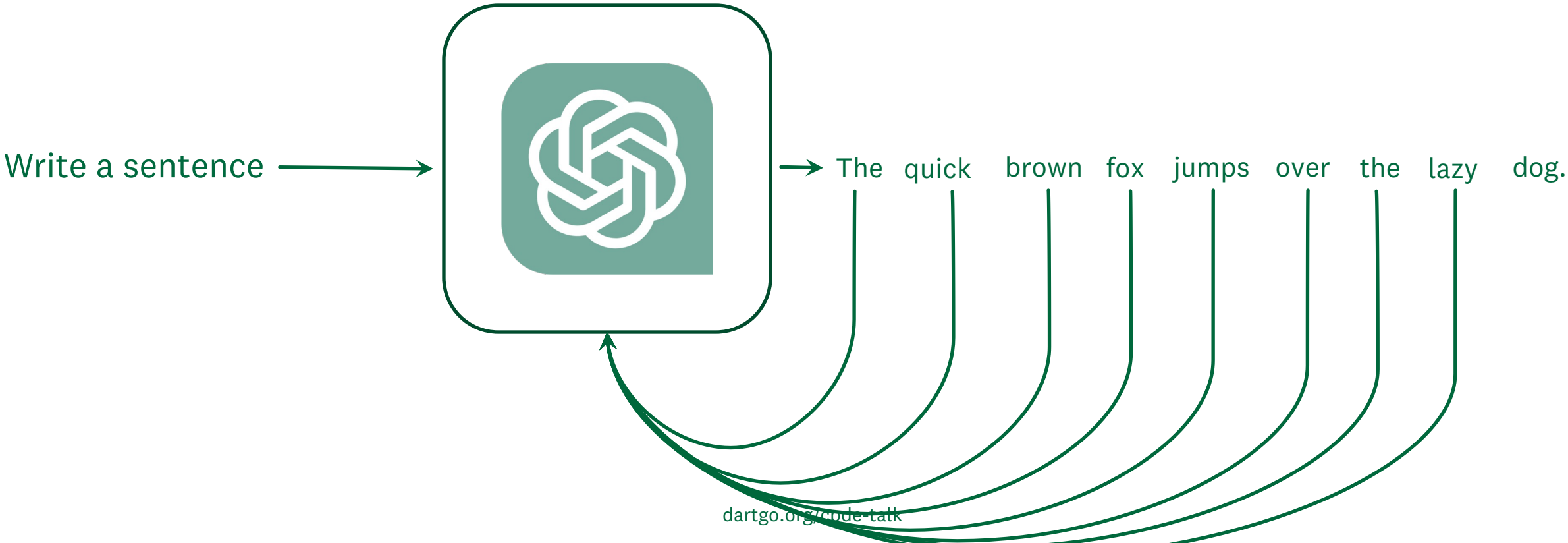
How Large Language Model's (de-)construct language

The fundamental case



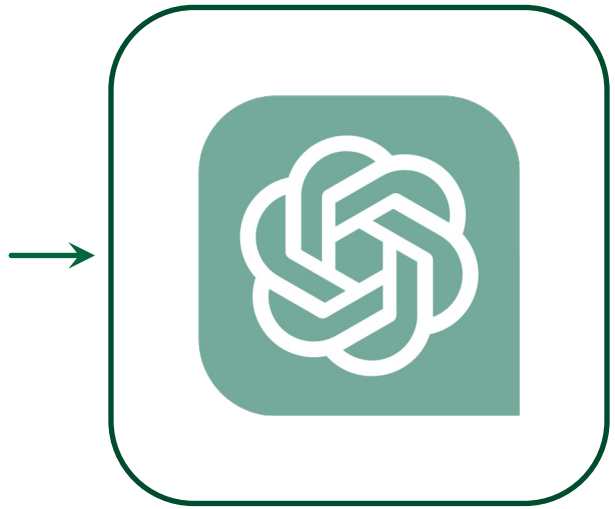
How Large Language Model's (de-)construct language

Predicting in a loop



How Large Language Model's (de-)construct language
Predicting a conversational turn

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
Cutting Knowledge Date: December 2023 Today
Date: 23 July 2024 You are a helpful assistant
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
What is the capital of France?
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```



```
The capital of France is Paris.
<|eot_id|>
```

*The template shown is the template used for Llama 3.2

How Large Language Model's (de-)construct language

How Large Language Models construct language

🗣️ How many words are in your response?

There are 12 words in my response.

🗣️ List them out in a numbered list.

Sure, here are the words in my previous response listed in a numbered list:
1. There
2. are
3. 12
4. words
5. in
6. my
7. response.

How Large Language Model's (de-)construct language

LLMs lie? - It's complicated

🤷‍♂️ LLMs have no concept of facts or truth, no internal database, no search mechanism

🦜 They only "know" patterns reinforced through their training data

- Example in Dartmouth Chat: Timothée Chalamet's mother

🤞 Factually correct responses are patterns that happen to be correct

🤔 "Hallucination" is a bit of a misnomer

Emily M. Bender *et al*. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

How Large Language Model's (de-)construct language

How to find the next word?

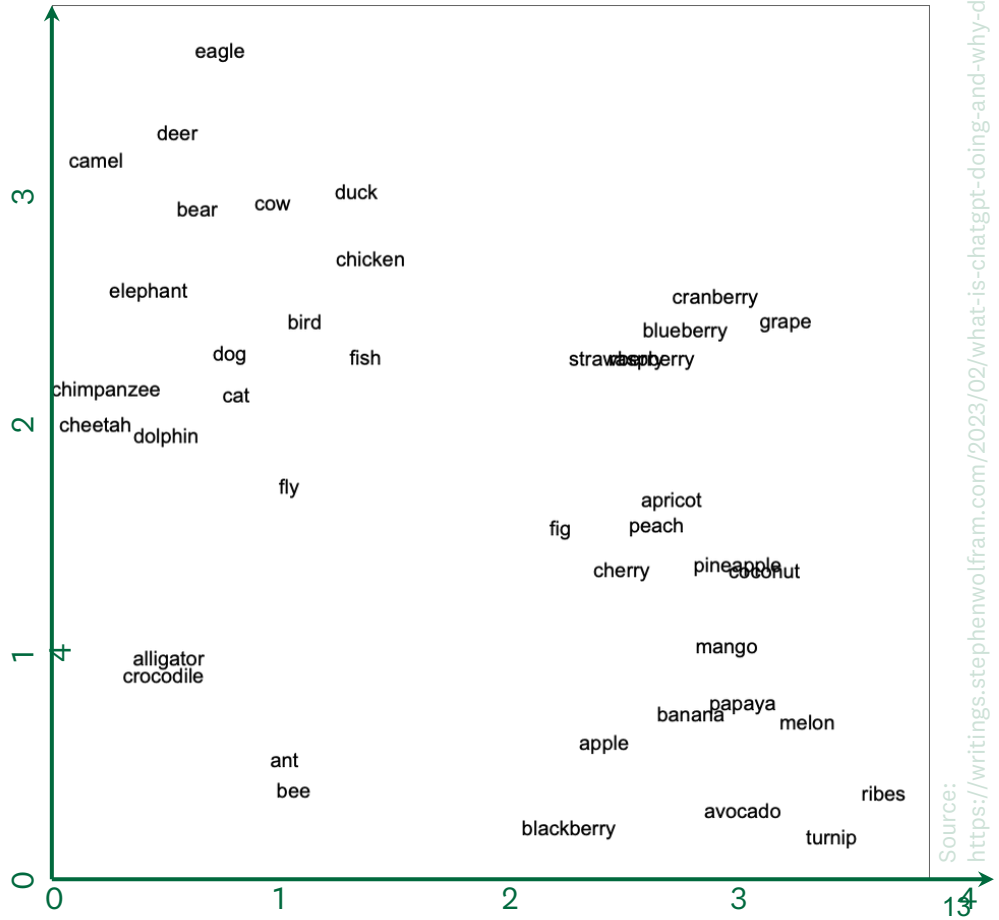
- 🏋️ The patterns in written language are learned from text data in a training phase
- 📚 GPT 3.5, for example, has been trained on about 570 GB of texts or 300 billion words (English Wikipedia: 4.2 billion words)
- 🤯 This would still not be enough data if we were looking for examples of longer specific word sequences!

How Large Language Model's (de-)construct language

How to find the next word?

Two important "tricks" solve this problem:

- The model learns to map every word into a numerical "meaning space", where words with similar meaning are represented by similar numbers
- That way the model can base the probability of the next word on sequences of similar words
- GPT 3.5 uses 2048 dimensional embeddings (probably)



Source: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Background

How to find the next word?

Two important “tricks” solve this problem:

1. Words in “meaning space”
2. The model uses a mechanism called “attention” to emphasize the most important words in the sequence
   - 🏋️ These weights are learned during training
   -  This structure is called a “transformer”
   - 💡 GPT: **G**enerative **P**re-trained **T**ransformer

What can AI do for qualitative analysis?

- 🧨 Concept of “semantic similarity” very powerful
- 🦾 LLM’s are excellent at picking up patterns in language
- 👉 Let’s look at an example!

- 🤖 By the way: All responses for the following demos were generated using Claude 3.7 Sonnet

Summary

- 📚 Large Language Models are excellent text processors
- 🎯 Semantic embeddings are a useful tool to capture contextual information in text
- 🤝 Using a Human-in-the-Loop approach, AI holds a lot of potential to speed up and scale qualitative analysis

Discussion

Thank you

dartgo.org/cpde-talk