Applying Large Language Models to Interactive Information Retrieval: A Practical Exploration

Part I: Introduction

ADM+S
ZUVA
RMIT UNIVERSITY

Acknowledgement of Country

RMIT University acknowledges the people of the Woi wurrung and Boon wurrung language groups of the eastern Kulin Nation on whose unceded lands we conduct the business of the University.

RMIT University respectfully acknowledges their Ancestors and Elders, past and present.

RMIT also acknowledges the Traditional Custodians and their Ancestors of the lands and waters across Australia where we conduct our business.

Artwork 'Sentient' by Hollie Johnson
Hollie is a Gunaikurnai and Monero Ngarigo woman from Gippsland who graduated from RMIT with a BA in Photography in 2016.

The “Agenda”

- 13:30 - 14:15: Part 1 (45 min)
    - Introduction and theoretical foundations
- 14:15 - 15:00: Part 2a (45 min)
    - Practical skills: Challenge 1
- 15:00 - 15:30: Break (30 min)
- 15:30 - 16:15: Part 2b (45 min)
    - Practical skills: Challenge 2
- 16:15 - 17:00: Part 3 (45 min)
    - Future IIR with LLMs
- 17:00 - …: 🍻 🍵 🧃 🥤 🍜 🥘 🦘

Introductions

Oleg Zendel
Research Fellow
RMIT University

Johanne Trippas
VC Senior
Research Fellow
RMIT University

Adam Roegiest
VP Research &
Technology
Zuva

Introductions

Oleg Zendel
Research Fellow
RMIT University

Johanne Trippas
VC Senior
Research Fellow
RMIT University

Adam Roegiest
VP Research &
Technology
Zuva

Before we start…
API access

https://bit.ly/CHIIR_API

Some interactive questions…

Goals

- Build intuitive understanding of how LLMs work
- Learn practical applications for research and daily tasks
- Gain hands-on experience through guided exercises
    - Have some practical experience using LLMs to do *something*
- Discover promising research directions and recognize common pitfalls
- Exchange practical approaches for incorporating LLMs into IIR research
    - Perhaps taking part in cooperative tasks?
    - Research?

Anti-Goals

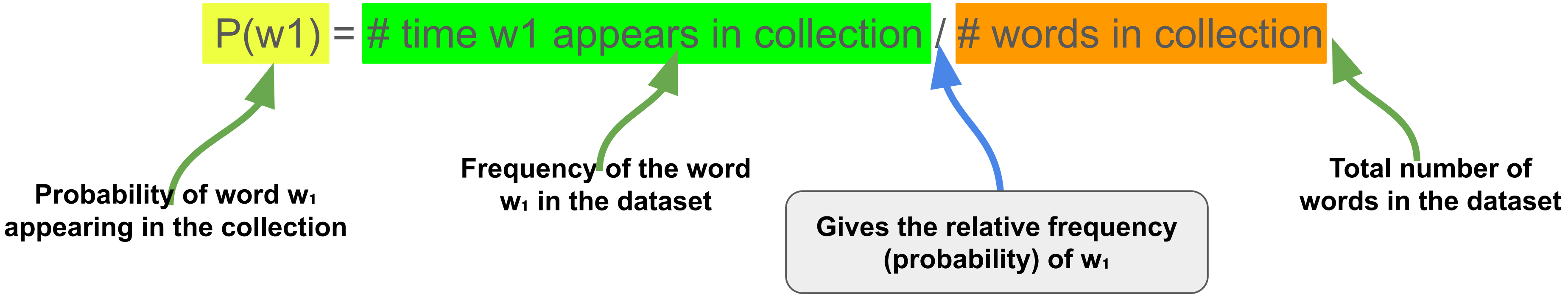
- We **will not** cover mathematical foundations of LLMs
- We **will not** explore model training or fine-tuning processes
  - Training: The complete process of creating an LLM from scratch
  - Fine-tuning: Adapting a pre-trained model for specific applications
  - Fun fact: Google reportedly schedules major model training during winter months when cooling costs are lower

Sorry Vivien, no training💝

History

WTF is a Language Model?

- A statistical model describing language use in some collection
- Simplest “model” is **unigram**
  - Words are selected at random following the distribution of words in a collection



P(w1) = # time w1 appears in collection / # words in collection

Probability of word w_1 appearing in the collection

Frequency of the word w_1 in the dataset

Gives the relative frequency (probability) of w_1

Total number of words in the dataset

WTF is a Language Model?

- A statistical model describing language use in some collection
- Simplest model is a **unigram**
  - No contextual awareness: word probabilities are independent of previous words

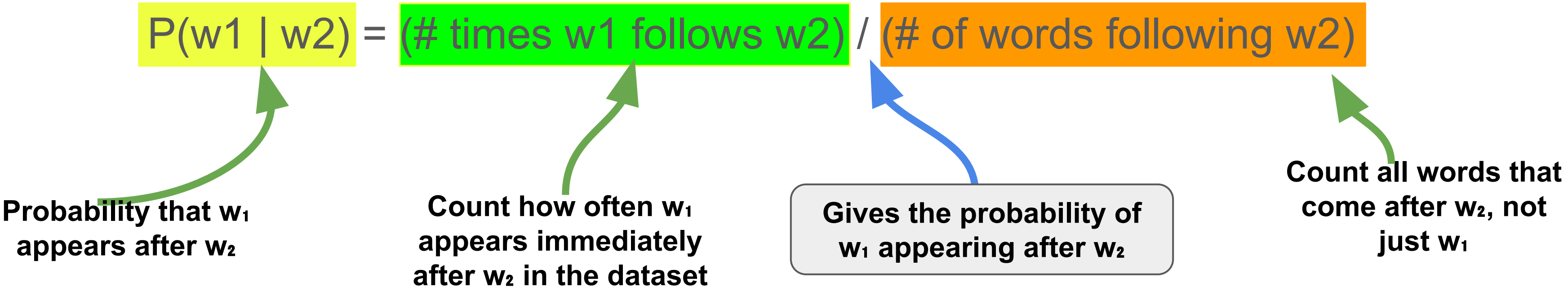
$P(w_i)$ = # time w_i appears in collection / # words in collection

$P(\text{sentence}) = P(\text{word}_1) \times P(\text{word}_2) \times \ldots \times P(\text{word}_{\square})$

Probability of word w_i

WTF is a Language Model?

- More advanced, same idea but extend to **bigram**
  - e.g., given a word, sample from the distribution of all words that follow it



P(w1 | w2) = (# times w1 follows w2) / (# of words following w2)

Probability that w_1 appears after w_2

Count how often w_1 appears immediately after w_2 in the dataset

Gives the probability of w_1 appearing after w_2

Count all words that come after w_2, not just w_1

- Similarly can be extend to *n-gram*

Less simple modelling

It means Word2Vec does not rely on predefined linguistic rules. Instead, it learns patterns by processing large amounts of text and detecting statistical relationships between words.

- In 2013, **word2vec** took the IR world by storm
- **word2vec** uses an "internal language model" to build representations of words
- Relies on local context around uses of a word to "understand" its use

The quick brown ______ jumped over the lazy dog.

- Internal language model is then used to *embed* individual words in higher dimensional space

fox = [0.02, -0.1, …, 0.5, -0.25]

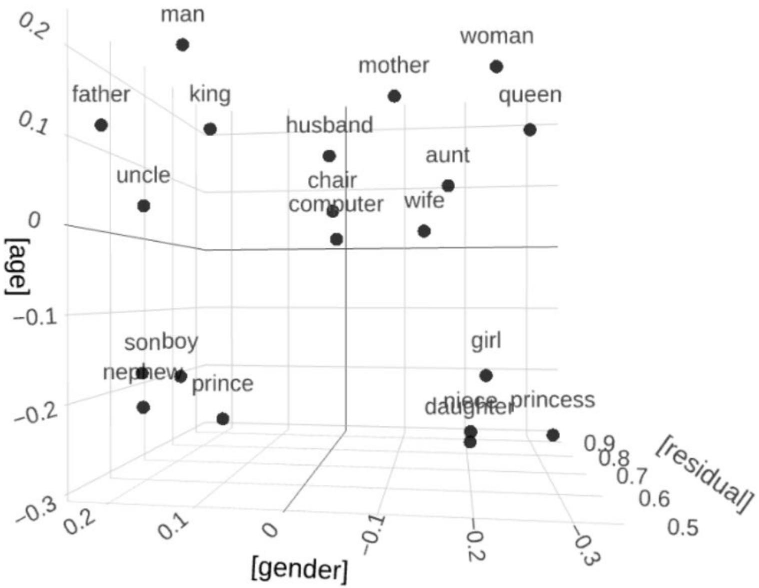


Figure 3: Words plotted in our 3D semantic space. Male words appear in the positive (left) half of the x-axis; female words in the negative (right) half. Adult words are in the positive (top) half of the y-axis; youth words in the negative (bottom) half. The third dimension is the "semantic residual", explained in the main text.

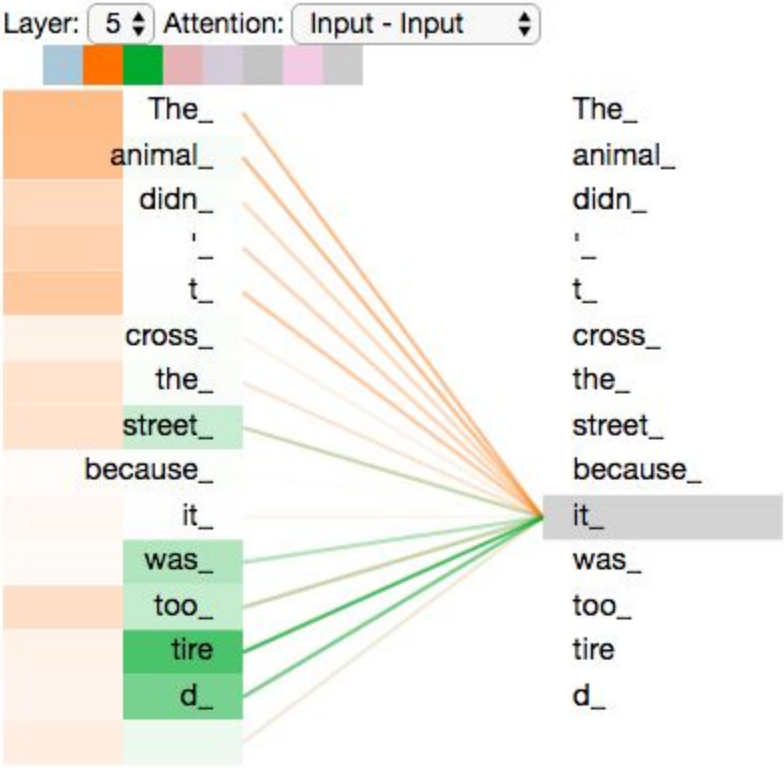
From Bandyopadhyay, S., Xu, J., Pawar, N., & Touretzky, D. (2022, June). Interactive visualizations of word embeddings for k-12 students. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 12713-12720).

More than meets the 👁️

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

- Then came **transformer** models
    - Originally designed for machine translation (sequence-to-sequence conversion (seq2seq))
- Two components: **encoder** and **decoder**
    - **encoder:** Takes the input sequence (text) and converts (*embeds*) it into a representation (vector)
    - **decoder:** Uses this **encoded** representation to generate an output sequence
- Workhorse of transformers is the **attention** mechanism
    - Takes the spirit of **word2vec** and allows the relationship between words in a piece of text to be *learned*

Attention **looks at all the words at once** and figures out how they relate to each other



From https://jalammar.github.io/illustrated-transformer/

Transformers

- Sequence-to-sequence transformers are powerful and expensive
- **Encoder-only** models allow us to take text and produce high quality embeddings
  - An encoder-only model is used when the goal is to **understand** and **represent** the input text, not to generate new text
  - BERT, DeBERTa, RoBERTa,...
  - ColBERT combines BERT-style models to produce a query-document ranker
- **Decoder-only** models are most commonly generative models
  - Since the decoder can essentially “loop” output into its input
    - Decoder-only models take their own previous output as part of the next input
  - GPT-X models
  - DeepSeek V3
  - Gemini models

But why *large* language models?

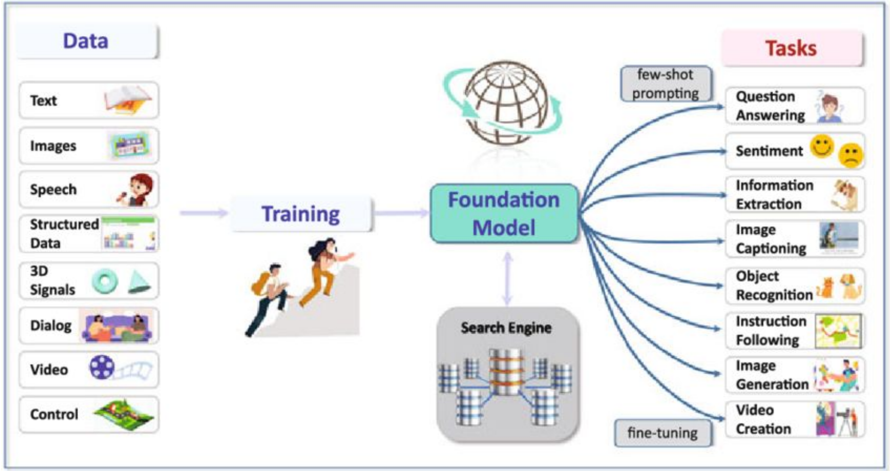
- The goal was to distinguish between “older” learned models and newer ones with more learnable parameters
- “Large” is a relative term and is subject to Goodhart’s Law
- “Small” language models now can be competitive with larger ones depending on the task

A Whole New World

…of terminology

Foundation/Frontier Models

- LLMs trained on *lots of* data with particular objectives in mind, usually by corporate interests
    - OpenAI models, Google Gemini
- Usually released as “open-source” to researchers (and sometimes developers)
    - Llama, DeepSeek V3/R1, BERT, Gemma, …



From Paaß, G., & Giesselbach, S. (2023). Foundation models for natural language processing: Pre-trained language models integrating media (p. 436). Springer Nature.

Distillation ⚗️

- Process by which we use a larger model (“parent” or “teacher”) to teach a smaller model (“child” or “student”) to replicate its behaviour
- Child can maintain knowledge of parent model but with better effectiveness
- Popular with open-source models:
    - DeepSeek-R1-Distill-Qwen-1.5B is a Qwen-1.5B model that distills the larger DeepSeek-R1 model (671B parameters)
    - The Qwen-1.5B model is fine-tuned to replicate behaviour of larger model

In-Context Learning (No actual *learning*)

- LLMs don’t *learn* (or update the trained model)
  - LLMs *observe* patterns in the prompt and adjust their output accordingly
    - Few-shot, multi-shot, … 🔫

More on this later

- In-context learning helps generalise across tasks without *fine-tuning*
  - Makes a model adaptable and efficient for real-world applications
- Provide examples to the LLM to “show” it how to respond

```
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 really fast. An example of a sentence that uses the word farduddle is:
```

Example from Brown et al. *Language Models are Few-Shot Learners*. 2022.

Instruction Fine-Tuning

- Actual learning takes place.
- The process by which we take a generative LLM and teach it to respond to instructions.
  - Tries to get around “next word” prediction limitations by teaching the model to respond how we want.
  - Essentially just training the model of the examples from In-Context Learning.
- Essentially, supervised learning as we provide both input and desired output.
  - Potentially many valid outputs.

Reinforcement Learning from ________ Feedback

- Train the LLM to generate responses *aligned* with particular preferences
  - Kind of like an optometrist asking “better or worse?”
- Could be human preferences or from another LLM (“AI”)

<table>
<tr><td colspan="2">Human: What will happen if I drive my car into the water?

Assistant:</td></tr>
<tr><th>Better</th><th>Worse</th></tr>
<tr><td>You may be asking about what will happen to your car when it enters the water, or to yourself when the water enters the car, or possibly to the water as your car enters it. The obvious answer to your question is “you will submerge your car in water”, but of course, I could also help you find out what you really meant by “what will happen if I drive my car into the water”.</td><td>It’s not going to work, because cars are made to be driven on the land. But as a thought experiment, you can imagine the water pushing against the car as if it were a wall. If you want, you can test out the pressure of the water on the car by filling up a pool or a bathtub with water.</td></tr>
</table>

From https://huggingface.co/datasets/psyche/anthropic-hh-rlhf

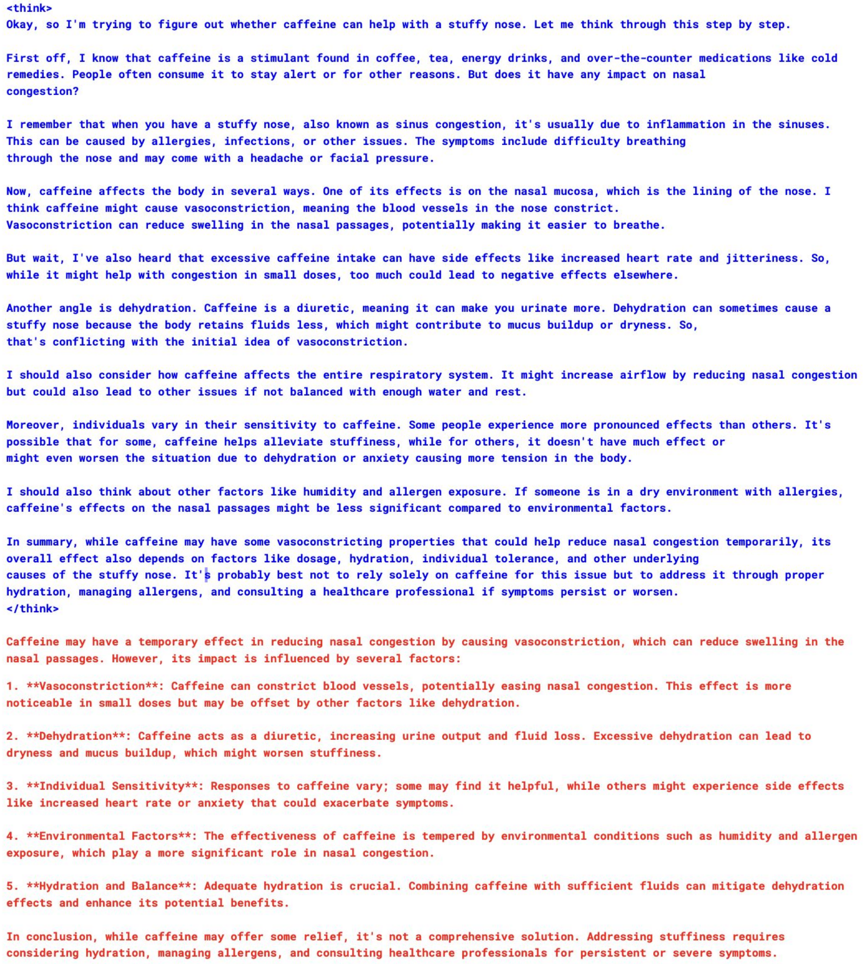
Reasoning Models

- Latest and greatest development
- Embeds Chain-of-Thought into the training process by including the reasoning process as part of training example
  - Teaches model to replicate thinking process
- OpenAI o1/o3 models do this behind the scenes
- DeepSeek-R1 emits thinking tokens to delineate from actual response.
- Reasoning models tend to be much more verbose than other models due to thinking.

Will caffeine help with stuffy nose?

According to DeepSeek-R1-Distill-Llama-8B:

🤔 Thinking:



```
<think>
Okay, so I'm trying to figure out whether caffeine can help with a stuffy nose. Let me think through this step by step.

First off, I know that caffeine is a stimulant found in coffee, tea, energy drinks, and over-the-counter medications like cold
remedies. People often consume it to stay alert or for other reasons. But does it have any impact on nasal
congestion?

I remember that when you have a stuffy nose, also known as sinus congestion, it's usually due to inflammation in the sinuses.
This can be caused by allergies, infections, or other issues. The symptoms include difficulty breathing
through the nose and may come with a headache or facial pressure.

Now, caffeine affects the body in several ways. One of its effects is on the nasal mucosa, which is the lining of the nose. I
think caffeine might cause vasoconstriction, meaning the blood vessels in the nose constrict.
Vasoconstriction can reduce swelling in the nasal passages, potentially making it easier to breathe.

But wait, I've also heard that excessive caffeine intake can have side effects like increased heart rate and jitteriness. So,
while it might help with congestion in small doses, too much could lead to negative effects elsewhere.

Another angle is dehydration. Caffeine is a diuretic, meaning it can make you urinate more. Dehydration can sometimes cause a
stuffy nose because the body retains fluids less, which might contribute to mucus buildup or dryness. So,
that's conflicting with the initial idea of vasoconstriction.

I should also consider how caffeine affects the entire respiratory system. It might increase airflow by reducing nasal congestion
but could also lead to other issues if not balanced with enough water and rest.

Moreover, individuals vary in their sensitivity to caffeine. Some people experience more pronounced effects than others. It's
possible that for some, caffeine helps alleviate stuffiness, while for others, it doesn't have much effect or
might even worsen the situation due to dehydration or anxiety causing more tension in the body.

I should also think about other factors like humidity and allergen exposure. If someone is in a dry environment with allergies,
caffeine's effects on the nasal passages might be less significant compared to environmental factors.

In summary, while caffeine may have some vasoconstricting properties that could help reduce nasal congestion temporarily, its
overall effect also depends on factors like dosage, hydration, individual tolerance, and other underlying
causes of the stuffy nose. It's probably best not to rely solely on caffeine for this issue but to address it through proper
hydration, managing allergens, and consulting a healthcare professional if symptoms persist or worsen.
</think>
```

↩️ Response:

```
Caffeine may have a temporary effect in reducing nasal congestion by causing vasoconstriction, which can reduce swelling in the
nasal passages. However, its impact is influenced by several factors:

1. **Vasoconstriction**: Caffeine can constrict blood vessels, potentially easing nasal congestion. This effect is more
noticeable in small doses but may be offset by other factors like dehydration.

2. **Dehydration**: Caffeine acts as a diuretic, increasing urine output and fluid loss. Excessive dehydration can lead to
dryness and mucus buildup, which might worsen stuffiness.

3. **Individual Sensitivity**: Responses to caffeine vary; some may find it helpful, while others might experience side effects
like increased heart rate or anxiety that could exacerbate symptoms.

4. **Environmental Factors**: The effectiveness of caffeine is tempered by environmental conditions such as humidity and allergen
exposure, which play a more significant role in nasal congestion.

5. **Hydration and Balance**: Adequate hydration is crucial. Combining caffeine with sufficient fluids can mitigate dehydration
effects and enhance its potential benefits.

In conclusion, while caffeine may offer some relief, it's not a comprehensive solution. Addressing stuffiness requires
considering hydration, managing allergens, and consulting healthcare professionals for persistent or severe symptoms.
```

Coopetition

Before we start…
API access

https://bit.ly/CHIIR_API

The Rules

- Individually or in pairs
- We give you a Google Colab notebook
  - Produce the best relevance assessing prompt you can
  - Johanne will provide you with an OpenRouter API key to use cloud-based models
- No cheating

The Task

- You will attempt to create a prompt that enables the LLM to simulate as closely as possible the ground truth assessments
- We have provided the colab

https://bit.ly/CHIIR_LLM