AI in News

Applying Generative AI in Production with Confidence

Thomson Reuters™

Agenda

- 9:00 – 10:30: Improve generation results with RAG, prompt engineering, and Chain-of-thought
- 11:00 – 12:30: Comprehensive Evaluation and Hallucination Detection
- 14:00 – 15:30: Explore instruction fine-tuning techniques and understand QLoRa
- 16:00 – 17:30: Bridging the Gap in News Production: Aligning AI Models

RAG and prompt engineering

Thomson Reuters™

Information retrieval for RAG

- Task: Identify and retrieve information (text, document) relevant to the given query
- Essential for good performance of the system
- Finds relevant, supporting information

Semantic search

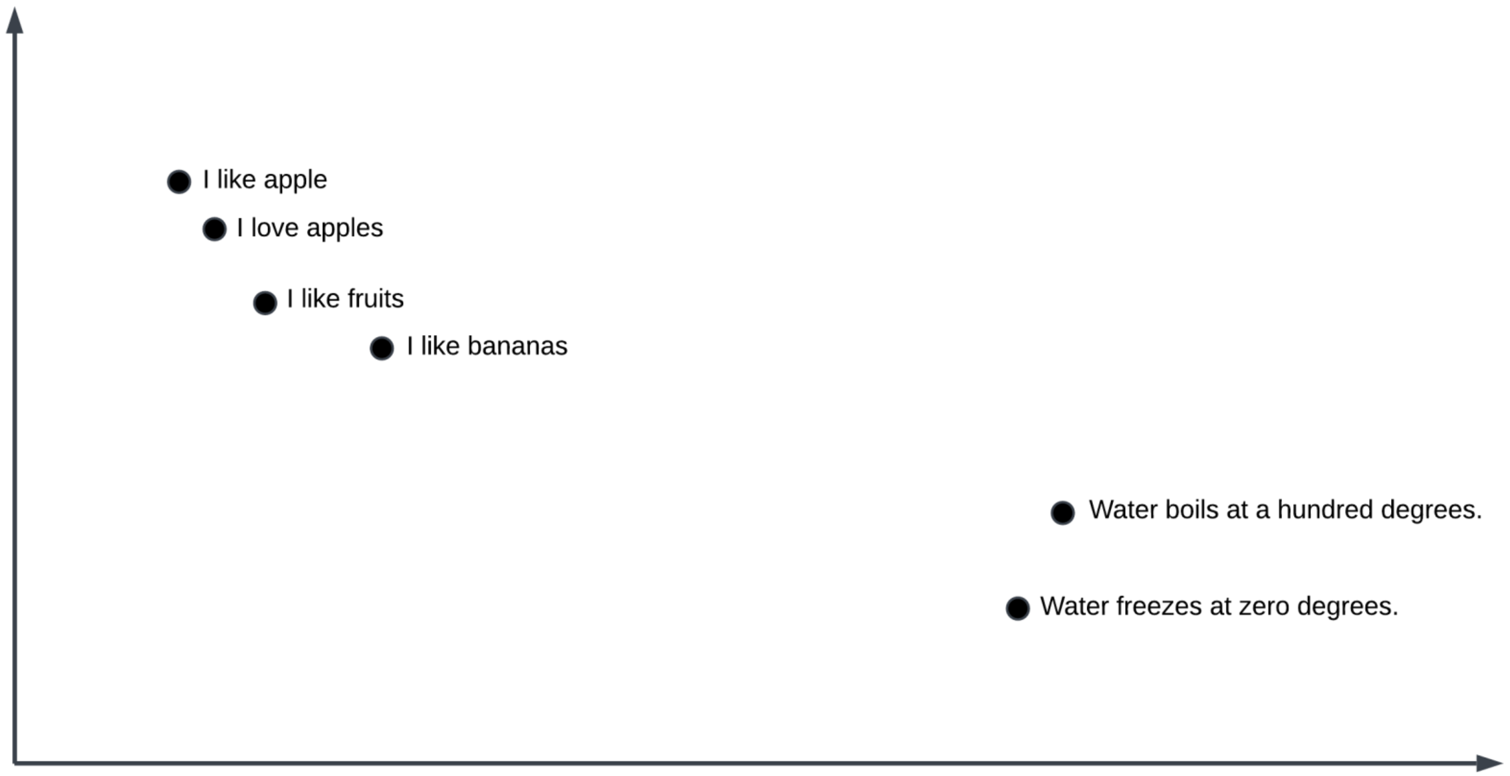
Idea: Instead of looking for the same words look for the same meaning

1. Transform each text you want to use to a vector representation
2. Transform query to a vector representation
3. Compare query's vector with vectors in your database
4. Select closest ones.

Semantic search – text embeddings

- Numerical representation of the text
- You can obtain it using one of many language models:
    - Small ones like BERT-based:
        - Useful library: SentenceTransformers
        - Easy to run locally
    - From major LLM providers:
        - Just make any API call
        - Paid service

Semantic search – text embeddings



Semantic search – distance

What does it mean: Select closest ones?

- To find similar vectors we need to define similarity metric first
- Most commonly used: cosine similarity
- $cosine\ similarity = \frac{ab}{||a||\ ||b||}$
- It’s a cosine of the angle between the vectors
- The higher cosine similarity the more similar the vectors are

Thomson Reuters™

Vector store (DB)

- Optimized for vector search
- Often use approximate k-NN search for low latency (e.g. HNSW)
- Supports additional fields for traditional filtering

Hybrid search

- Semantic search is cool, but exact word match can be useful
- Let’s add key-word method to a semantic search and merge it
- Helps when we have unusual words in our corpus

Reciprocal Rank Fusion

How to combine two search results?

- Create ranked lists with results for both methods (sorted from the most to the least similar)
- Calculate reciprocal rank for each item
  - $reciprocal\ rank = \frac{1}{rank(i)+k} \quad \text{for i} - \text{th item}$
- Sum ranks for each item, the sum is a final score
- Sort the list by final score

Thomson Reuters™

Generation

- Zero-shot prompting
- In-context learning: one or few-shot learning:
    - Provide new data during the inference instead of retraining the model
    - Use high quality data to improve quality or relevancy of the response
    - Easy method for domain adaptation

Chain of Thoughts

- Instructs model to break down (“think about”) the problem before returning the answer
- Prompt includes instruction like “Let’s think step by step”
- Can include multiple interactions with the model
- Zero or few-shot

… of Thoughts

Chain of thoughts is just the beginning

- Many thought generation techniques were proposed since
    - Multiple chain of thoughts
    - Tree of thoughts
    - Graph of thoughts
    - …

The Prompt Report: A Systematic Survey of Prompting Techniques

Sander Schulhoff[1,2*] **Michael Ilie**[1*] **Nishant Balepur**[1] **Konstantine Kahadze**[1]
Amanda Liu[1] **Chenglei Si**[4] **Yinheng Li**[5] **Aayush Gupta**[1] **HyoJung Han**[1] **Sevien Schulhoff**[1]
Pranav Sandeep Dulepet[1] **Saurav Vidyadhara**[1] **Dayeon Ki**[1] **Sweta Agrawal**[12] **Chau Pham**[13]
Gerson Kroiz **Feileen Li**[1] **Hudson Tao**[1] **Ashay Srivastava**[1] **Hevander Da Costa**[1] **Saloni Gupta**[1]
Megan L. Rogers[8] **Inna Goncearenco**[9] **Giuseppe Sarli**[9,10] **Igor Galynker**[11]
Denis Peskoff[7] **Marine Carpuat**[1] **Jules White**[6] **Shyamal Anadkat**[3] **Alexander Hoyle**[1] **Philip Resnik**[1]
[1] University of Maryland [2] Learn Prompting [3] OpenAI [4] Stanford [5] Microsoft [6] Vanderbilt [7] Princeton
[8] Texas State University [9] Icahn School of Medicine [10] ASST Brianza
[11] Mount Sinai Beth Israel [12] Instituto de Telecomunicações [13] University of Massachusetts Amherst
sschulho@umd.edu milie@umd.edu resnik@umd.edu

Paper: The Prompt Report: A Systematic Survey of Prompting Techniques, https://arxiv.org/abs/2406.06608

Query expansion

- Query and texts in our database can be quite different
- This can harm a semantic search
- Ideas:
    - Use LLM to generate new, more similar text based on the query
    - Use LLM to paraphrase query and use all of those to run a search

Practical notes

- Use LLMs to create great projects!
- Leverage their functionalities and external libraries
    - Structured output for communication with your program
    - Human in the loop approach to interact with users
    - …

Thank you!