THE UNIVERSITY of EDINBURGH

Text Technologies for Data Science

INFR11145

Retrieval Augmented Generation

Instructor:
Tuğrulcan "TJ" Elmas

27-Nov-2024

1

Lecture Objectives

- Learn about:
  - Advances in Text-To-Text Generation
  - Retrieval Augmented Generation (RAG) Pipeline
  - (Dense) Retrieval
  - Generation
  - RAG use-cases

Web is Massive

- Growing (from 13 Web Search)
  - 20 PB/day in 2008 → 160 PB/day in 2013 → now??
- Question answering task – Microsoft's solution
  - **Q**: Who created the character of Scrooge?
  - **A**: Scrooge, introduced by Charles Dickens in "A Chrismas Carol"
  - Identify (subj verb obj), rewrite as queries:
    - "created the character of Scrooge"
  - Search the web for exact phrase
  - Get top results

117	Dickens
78	Christmas Carol
75	Charles Dickens
72	Disney
54	Carl Banks
	…

- Good news: We can do this with web data but without Googling
  - Bad news: It turns out we still have to Google and use RAG

3

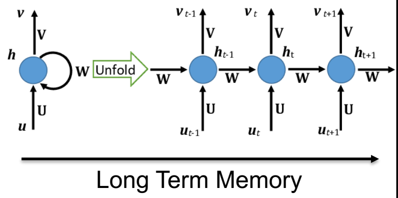
Text-to-Text Generation

- Many NLP tasks:
  - Document Similarity
  - Text Classification
- Text-to-Text Generation
  - Machine Translation
  - Question Answering
- Problem: Maximize P (desired output text | input text)
  - P (hola | Translate to Spanish: hello)
  - P (Scotland's capital is Edinburgh | What's Scotland's capital?)
    - Next word prediction: P (Scotland's | What's Scotland's capital?) x P(capital | Scotland's, What's…) x P (is | Scotland's capital, What's…) x P (Edinburgh | Scotland's capital is, What's…)

4

Early Works on Text-to-Text Generation

- Stone age
  - Rule based systems, dictionaries
  - Statistical Methods
- Recurrent Neural Networks (RNNs)
  - Predict the next word & update
  - Vanishing Gradient Problem
    - “AI forgets the beginning of the text”



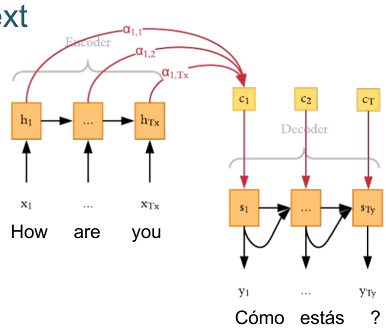
- Long Short-Term Memory (LSTM)
  - Maintains a long-term memory
  - Breakthrough in machine translation
    - still limited to a single context vector

Tuğrulcan Elmas, TTDS 2024/2025
THE UNIVERSITY of EDINBURGH

5

Large Language Models

- ”Attention is all you need”
  - Focus on the relevant parts of the text
  - Parallelism
- Predefined context window size
  - Max. tokens the model processes
  - 4k tokens for ChatGPT (GPT-3)



- Transformers architecture
  - Transforms the input text into a rich **representation**

Tuğrulcan Elmas, TTDS 2024/2025
THE UNIVERSITY of EDINBURGH

6

Features of Large Language Models

- Representations or Embeddings, not features
- Classic Bag of Words
  - Every feature corresponds to a word
  - Sparse, cannot handle homonyms.

"I like eating kebabs"

I	like	eating	kebabs	50k+ other words...
1	1	1	1	0

- Embeddings
  - Vector representation for words, sentences, passages etc.
  - Dense, incorporates the semantics & context

"I like eating kebabs
~="Kebabs please me"

0	1	2	3	...
3	23423	-313	0.003	0

Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY *of* EDINBURGH

7

Training of Large Language Models

- Pretraining
  - Create training data automatically from a large corpus
  - Masked Language Modeling
    - Autoregressive: I like eating ___ (kebabs)
    - Bidirectional: I like eating ___ (kebabs) in Istanbul
- Fine-tuning
  - Training by your preferred task & your corpus
  - Make chatbots, translators, search engines etc.

Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY *of* EDINBURGH

8

LLMs are Massive

- BERT (2018) by Google
  - Bidirectional encoder representations from transformers
  - BookCorpus (11k e-books, 6 GB)
  - English Wikipedia (120 GB text)
  - 220 MB model size

- GPT-3 (2020) by OpenAI
  - Generative Pretrained Transformers
  - Common Crawl (5 PB of internet text)
  - Wikipedia (2 TB text)
  - Books, academic articles, newspapers, codes…
  - 350 GB model size

LLMs perform better with clever prompts

- Provide instructions, context, examples
  - More data for the LLM
  - Narrows the search space

- Prompt engineering
  - Teach an LLM how to perform a new task
  - One Shot or Few Shot Learning
  - Chain of Thought Reasoning (“Answer step by step”)
  - Provide additional documents (RAG)

- No need to fine-tune every time

Why LLMs are bad?

- LLMs may hallucinate
- LLMs do not give credits to source
- LLMs are hard to update
  - Hard to teach new info (fine-tuning)
  - Harder to make it forget

Why Web Search is bad?

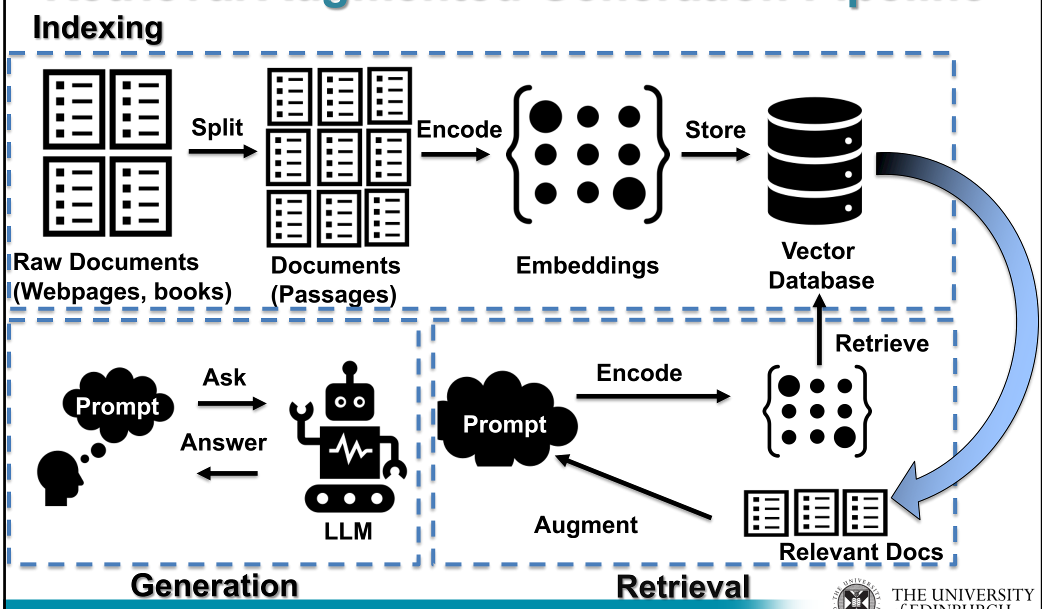
User Need on Web Search (14 Web Search 2)

- **Informational** – want to learn about something (~40% / 65%)
- **Navigational** – want to go to that page (~25% / 15%),
- **Transactional** – want to do something (web-mediated) (~35% / 20%)
- **For 40-65% of searches we do not really need a web search**
  - **Activity on Stackoverflow.com dropped by at least 25%**

TL;DR: Retrieval Augmented Retrieval

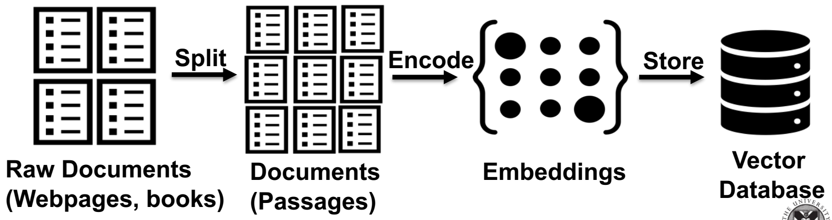
- Ask a question to ChatGPT
- ChatGPT googles
- ChatGPT appends the search results to the prompt
- ChatGPT answers

Retrieval Augmented Generation Pipeline



RAG Indexing

- Documents for augmentation
  - Webpages, Wikipedia, internal documents
- Inverted index is redundant
  - User queries are prompt – can be very long
- Create a vector database
  - Passages, sentences, entire text from a document (size limit!)
  - Represented by embeddings (e.g., by BERT)
  - Only need to be done **once** for each document



Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY of EDINBURGH

15

RAG Retrieval

- Dense Retrieval
  - Representations instead of term and document frequencies
  - Handles synonyms & query expansion
- Vectorize the query (prompt)
  - Documents are already vectorized
- Compute similarity
  - Cosine similarity: $\cos(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} = \sum_{i=1}^{|V|} q_i d_i$
- Retrieve the documents most similar to the query
  - Collect the plaintext for augmentation

Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY of EDINBURGH

16

RAG Retrieval – Computing Similarity

- We do not have an inverted index
  - How to collect a subset of documents to compute similarity?
- Naïve approach: compute the similarity between **all** documents versus the given query
  - Feasible for small vector databases, slow otherwise
- Use an Approximate Nearest Neighbour (ANN) algorithm
  - Trade off precision for speed
  - E.g., Hierarchical Navigable Small Worlds (HNSW)
  - Similar documents are linked together
  - More discussion in the guest lecture

Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY *of* EDINBURGH

17

RAG Generation

- Employ an LLM for generation
  - Preferably one with a large context window
- Append the retrieved documents to the prompt
  - Append on top or bottom
  - Explicity, e.g., “Question: …, Context: [retrieved documents]”
- Press enter & get the answer

Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY *of* EDINBURGH

18

RAG Generation

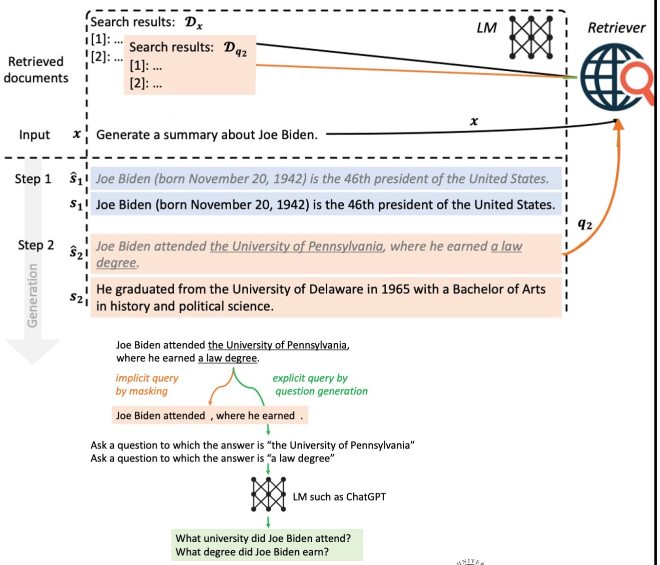
- Multiple (sets of) candidate documents?
- RAG-Sequence: Generate once for each (set of) document(s)
  - Compare answers
- RAG-Token: Multiple documents at each word generation
  - Can change prompt & documents during generation
  - Allows for **dynamic retrieval**

When to Retrieve? Static vs. Dynamic Retrieval

- Static retrieval: Retrieve before generating an answer
  - Predict if you need to retrieve, retrieve
  - Generate first, retrieve & regenerate if needed
- Dynamic Retrieval: Retrieve during generation
  - Naïve: Retrieve for each token
  - Batch: Answer step by step & retrieve if needed for a step

Active Retrieval Augmented Generation

- LLM decides when to retrieve and what to retrieve for
- Existence of low probability token(s) -> low confidence on generation -> retrieve
- Generated low probability text is used for a new query
  - Ask LLM to create questions as queries



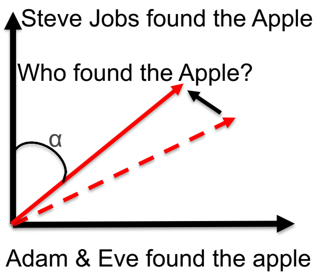
Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY *of* EDINBURGH

21

Training RAG

- Frozen RAG
  - Don't train anything
- Update the model for query encoding
  - Maximize the similarity between the query and useful documents
  - Not practical to update the model for document encoding
- Update the generator (finetune)
  - Using the prompt + retrieved documents, answers pairs
  - Not suitable for every use case



Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY *of* EDINBURGH

22

Use Case: Small LLM + RAG in PC > ChatGPT?

>

TRY GPT4ALL

24

Use Case: Isolating Risky Data

The New York Times

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Share full article 1.3K

A lawsuit by The New York Times could test the emerging legal contours of generative A.I. technologies. Sasha Maslov for The New York Times

EXAMPLE 3: RUSSIA SECRETLY OFFERED AFGHAN MILITANTS BOUNTIES TO KILL U.S. TROOPS, INTELLIGENCE SAYS

https://www.nytimes.com/2020/06/26/us/politics/russia-afghanistan-bounties.html
Copyright number: TX 8-890-563 Effective Date: 2020-08-04

Prompt (taken from article):
WASHINGTON — American intelligence officials have concluded that a Russian military intelligence unit secretly offered bounties to Taliban-linked militants for killing coalition

Output from GPT-4:
forces in Afghanistan — including targeting American troops — amid the peace talks to end the long-running war there, according to officials briefed on the matter.
The United States concluded months ago that the Russian unit, which has been linked to assassination attempts and other covert operations in Europe intended to destabilize the West or take revenge on turncoats, had covertly offered rewards for successful attacks last year.
Islamist militants, or armed criminal elements closely associated with them, are believed to have collected some bounty money, the officials said. Twenty Americans were killed in combat in Afghanistan in 2019, but it was not clear which killings were under suspicion.
The intelligence finding was briefed to President Trump, and the White House's National Security Council discussed the problem at an interagency meeting in late March, the officials said. Officials developed a menu of potential options — starting with making a diplomatic complaint to Moscow and a demand it stop, along with an escalating series of sanctions and other possible responses, but the White House has yet to authorize any step, the officials

Actual text from NYTimes:
forces in Afghanistan — including targeting American troops — amid the peace talks to end the long-running war there, according to officials briefed on the matter.
The United States concluded months ago that the Russian unit, which has been linked to assassination attempts and other covert operations in Europe intended to destabilize the West or take revenge on turncoats, had covertly offered rewards for successful attacks last year.
Islamist militants, or armed criminal elements closely associated with them, are believed to have collected some bounty money, the officials said. Twenty Americans were killed in combat in Afghanistan in 2019, but it was not clear which killings were under suspicion.
The intelligence finding was briefed to President Trump, and the White House's National Security Council discussed the problem at an interagency meeting in late March, the officials said. Officials developed a menu of potential options — starting with making a diplomatic complaint to Moscow and a demand that it stop, along with an escalating series of sanctions and other possible responses, but the White House has yet to authorize any step, the officials

26

Use Case: Isolating Risky Data

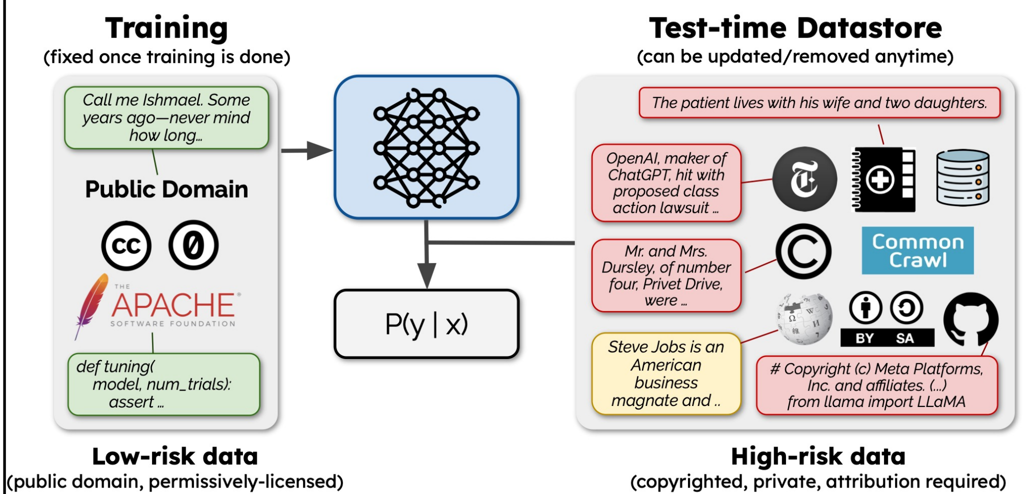

Futurism

OpenAI "Accidentally" Deleted Evidence From Its New York Times Lawsuit

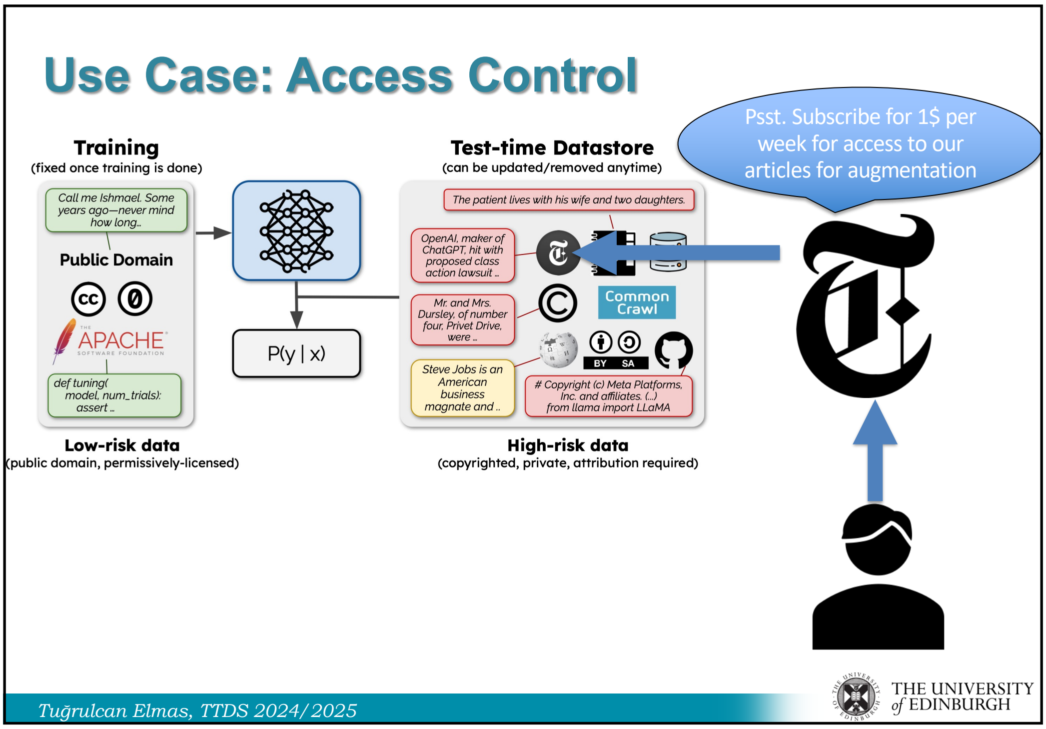
This was a massive mistake.

27

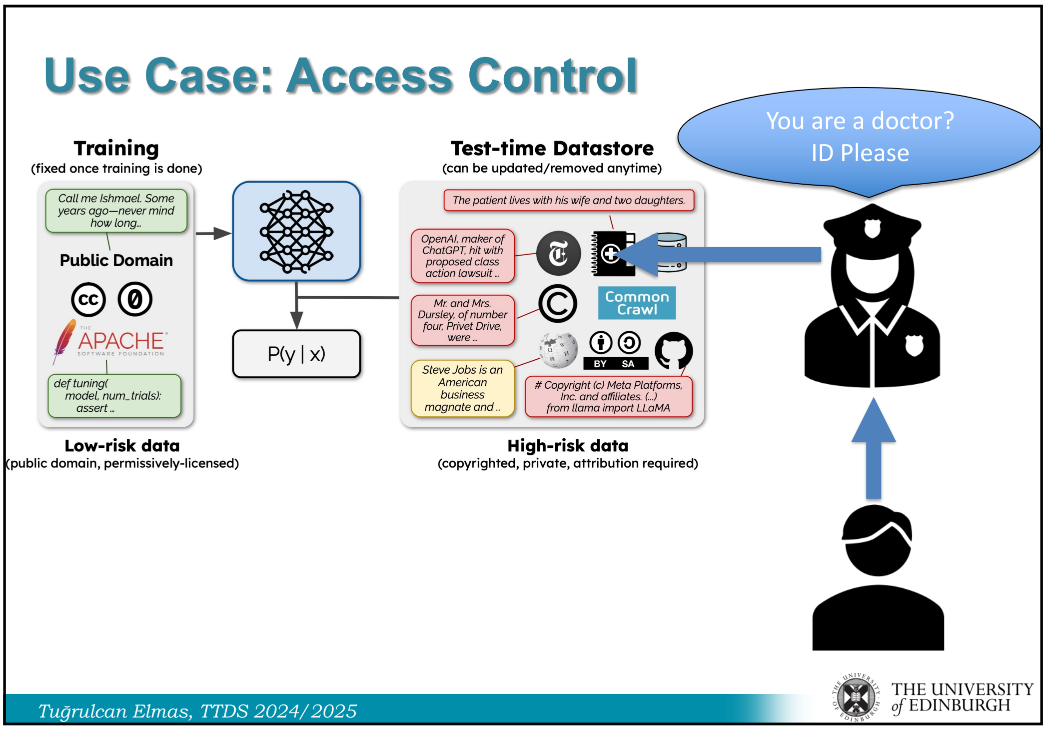
Use Case: Isolating Risky Data



28


Use Case: Access Control
Training
(fixed once training is done)
Call me Ishmael. Some years ago—never mind how long..
Public Domain
APACHE
SOFTWARE FOUNDATION
def tuning(model, num_trials): assert ..
Low-risk data
(public domain, permissively-licensed)
P(y | x)
Test-time Datastore
(can be updated/removed anytime)
The patient lives with his wife and two daughters.
OpenAI, maker of ChatGPT, hit with proposed class action lawsuit ..
Mr. and Mrs. Dursley, of number four, Privet Drive, were ..
Steve Jobs is an American business magnate and ..
Common Crawl
BY SA
Copyright (c) Meta Platforms, Inc. and affiliates. (...) from llama import LLaMA
High-risk data
(copyrighted, private, attribution required)
Psst. Subscribe for 1$ per week for access to our articles for augmentation
Tuğrulcan Elmas, TTDS 2024/2025
THE UNIVERSITY of EDINBURGH

29


Use Case: Access Control
Training
(fixed once training is done)
Call me Ishmael. Some years ago—never mind how long..
Public Domain
APACHE
SOFTWARE FOUNDATION
def tuning(model, num_trials): assert ..
Low-risk data
(public domain, permissively-licensed)
P(y | x)
Test-time Datastore
(can be updated/removed anytime)
The patient lives with his wife and two daughters.
OpenAI, maker of ChatGPT, hit with proposed class action lawsuit ..
Mr. and Mrs. Dursley, of number four, Privet Drive, were ..
Steve Jobs is an American business magnate and ..
Common Crawl
BY SA
Copyright (c) Meta Platforms, Inc. and affiliates. (...) from llama import LLaMA
High-risk data
(copyrighted, private, attribution required)
You are a doctor? ID Please
Tuğrulcan Elmas, TTDS 2024/2025
THE UNIVERSITY of EDINBURGH

30

Summary

- Text-To-Text Generation & Transformers
- Cons of LLMs and Web-Search
- RAG Pipeline
- RAG Indexing & Vector Database
- Dense Retrieval
- Generation & Dynamic Retrieval
- RAG Use Cases

Resources

- Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS 2020*. https://arxiv.org/abs/2005.11401
- Jiang et al. Active Retrieval Augmented Generation. *EMNLP 2023.* https://arxiv.org/abs/2305.06983
- Vaswani et al. Attention is All You Need. https://arxiv.org/abs/1706.03762
- Guest Lecture by Amin Ahmad
- Pasquale Minervini, NLU-11 Natural Language Understanding Generation and Machine Translation.