QA Systems + RAG

CS 4740 (and crosslists): Introduction to Natural Language Processing
https://courses.cs.cornell.edu/cs4740/2025sp

Slides developed by:
Magd Bayoumi, Claire Cardie, Tanya Goyal, Dan Jurafsky, Lillian Lee, James Martin, Marten van Schijndel

Announcements

- Practice questions
  - Zoom session (mostly) recorded
  - 2nd Zoom session TBD

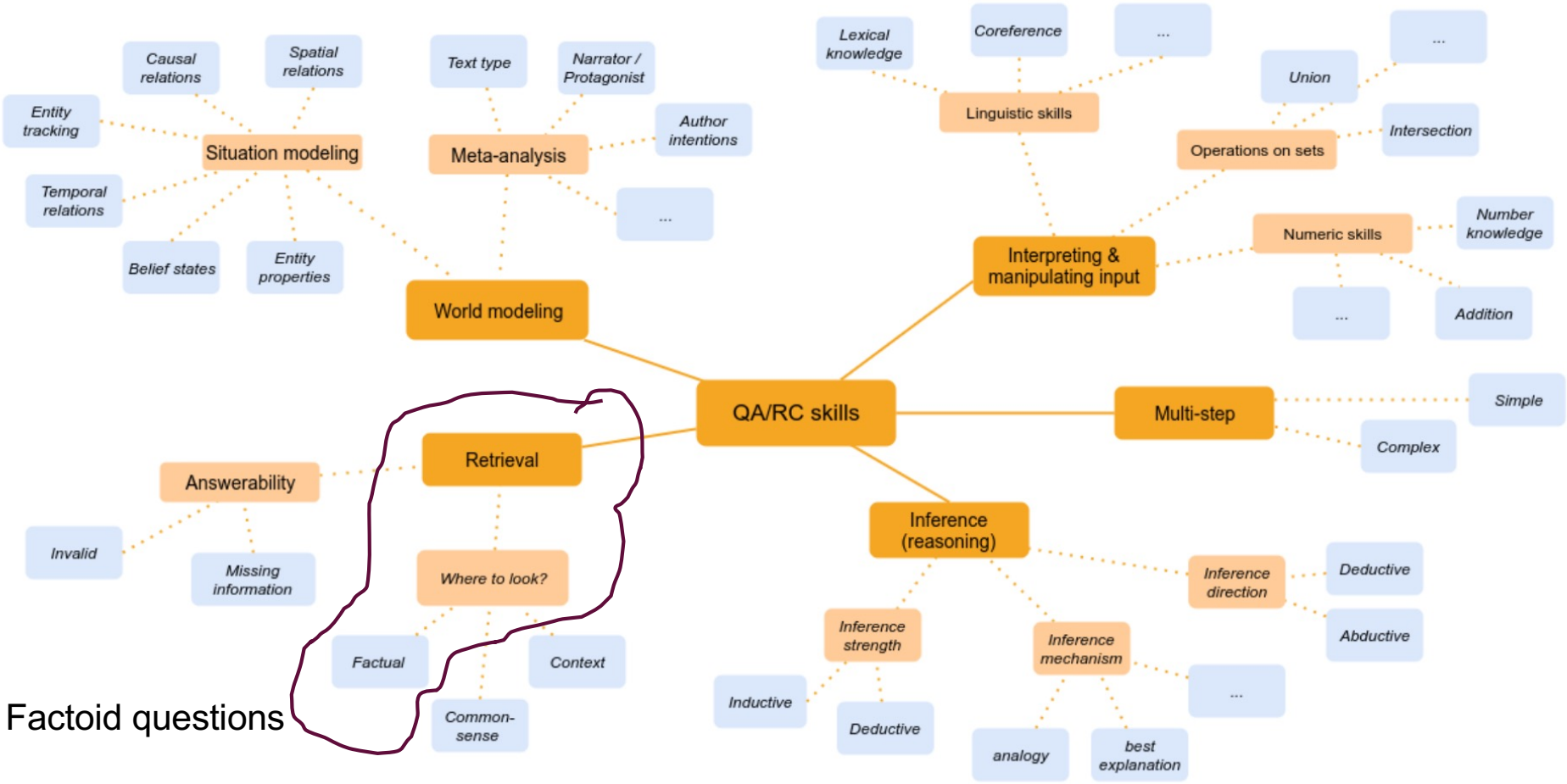
Today: QA systems (part 2) + RAG

- There are many types of QA paradigms
  - QA Datasets
  - Exercises: How to build systems for them
- Training LLM-based chatbots for QA
- QA using Retrieval-Augmented Generation

Traditional Focus: Factoid Questions

Questions of fact or reasoning that can be answered with simple facts expressed in a short or medium-length text.

Question	Answer
Where is the Louvre Museum located?	in Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	the yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What's the official language of Algeria?	Arabic
How many pounds are there in a stone?	14



Factoid questions

Types of QA tasks

- Information-seeking

  Communicative intent of the author of the question is **to seek information they did not have** [E.g., factoid questions]

- Probing

  Communicative intent of the author of the question is **to test the knowledge of another person or machine**

Data set: SQuAd

- Example of a “probing task”
- 100,000+ **reading comprehension** questions posed by crowdworkers on a set of Wikipedia articles
- Answer is a segment of text from the corresponding reading passage

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

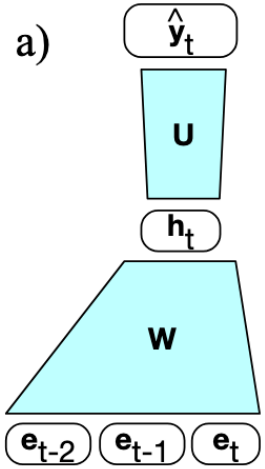
Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

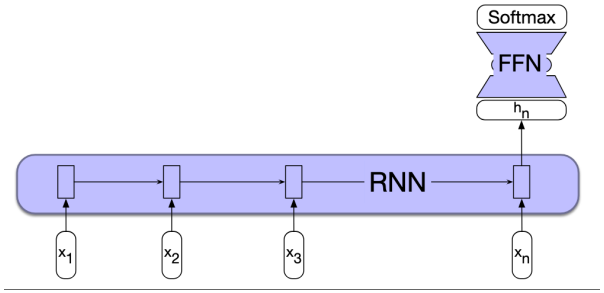
Data set: SQuAd

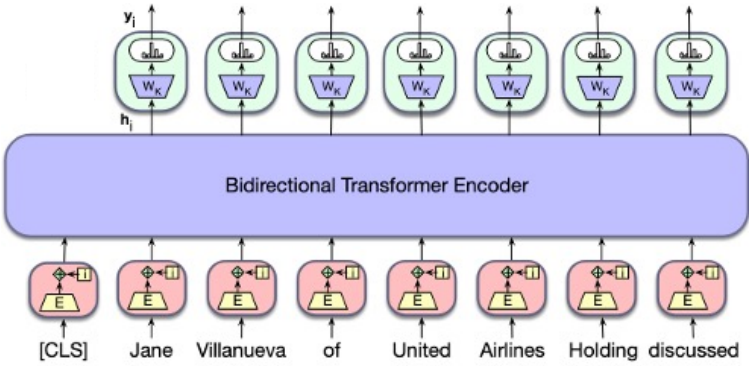
- Each **test set** example is:
  - Text passage
  - Question
- Each **training set** example is:
  - Question
  - Text passage with **answer span** marked/annotated

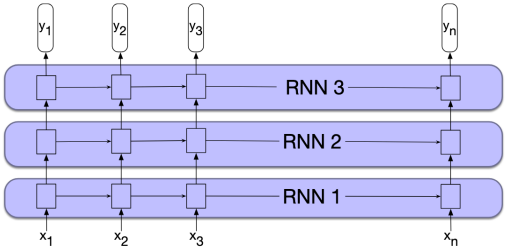
How should we model this task as a ML problem?

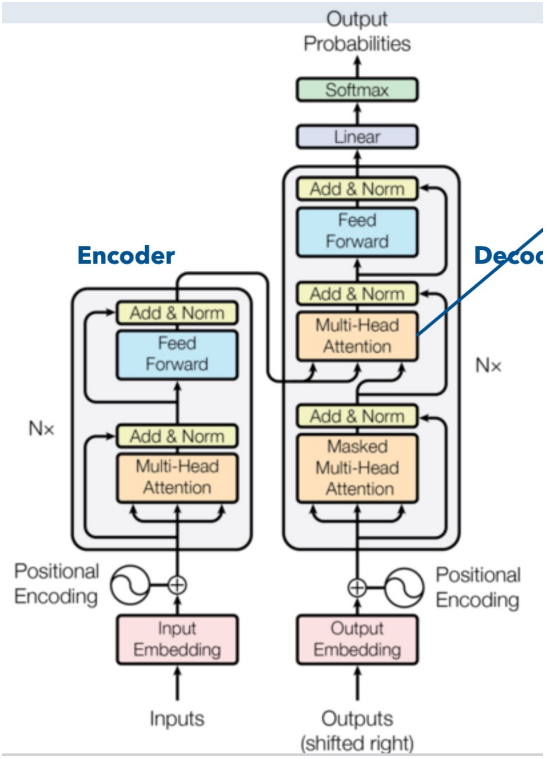
- Some of the models we've covered this semester...see the next slide


a)
$\hat{y}_t$
U
h_t
W
e_{t-2}
e_{t-1}
e_t

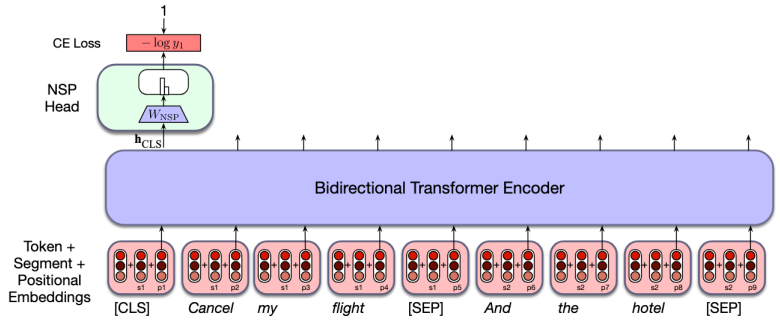

Softmax
FFN
h_n
RNN
x_1
x_2
x_3
x_n


y_i
h_i
W_K
Bidirectional Transformer Encoder
E
[CLS]
Jane
Villanueva
of
United
Airlines
Holding
discussed


y_1
y_2
y_3
y_n
RNN 3
RNN 2
RNN 1
x_1
x_2
x_3
x_n

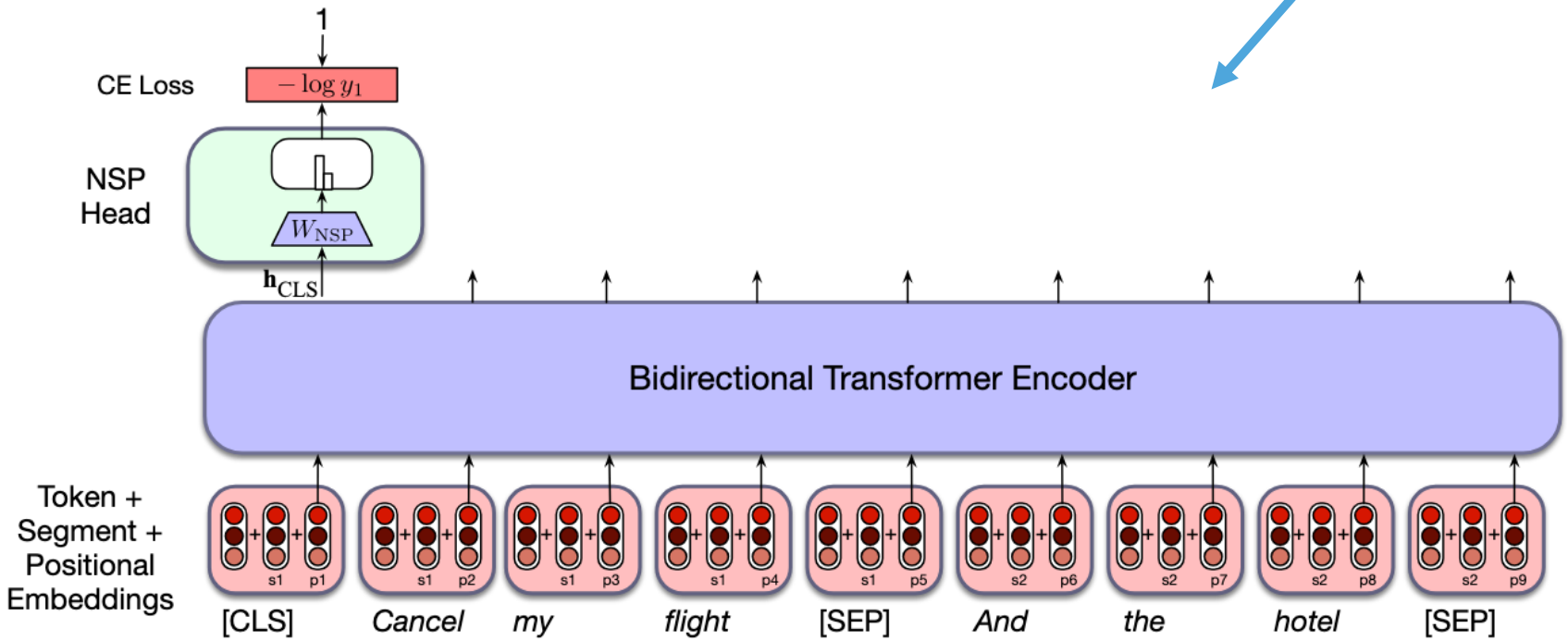

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Encoder
Decoc
N×
N×
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Bert


1
CE Loss
$-\log y_{i1}$
NSP Head
W_{NSP}
h_{CLS}
Bidirectional Transformer Encoder
Token + Segment + Positional Embeddings
[CLS]
Cancel
my
flight
[SEP]
And
the
hotel
[SEP]

Data set: SQuAd

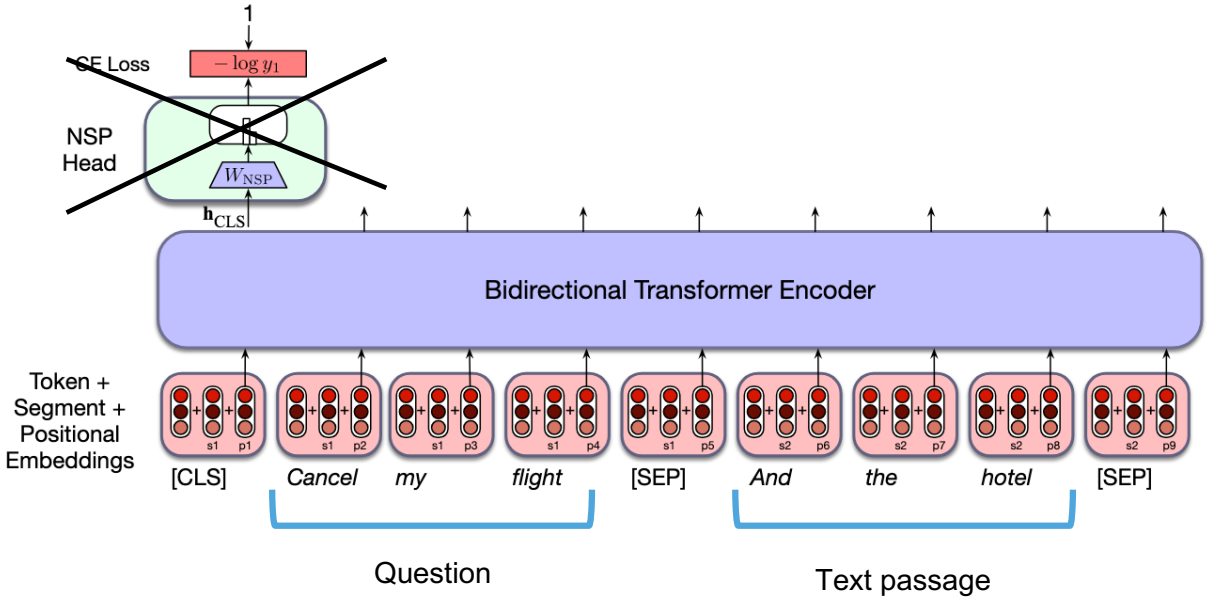
Approach: BERT



Data set: SQuAd

Approach: BERT

- Each **training set** example is:
  - Answer: text passage with **answer span** marked/annotated
  - Question
- For BERT
  - Label text passage to indicate answer



BERT for sequence tagging

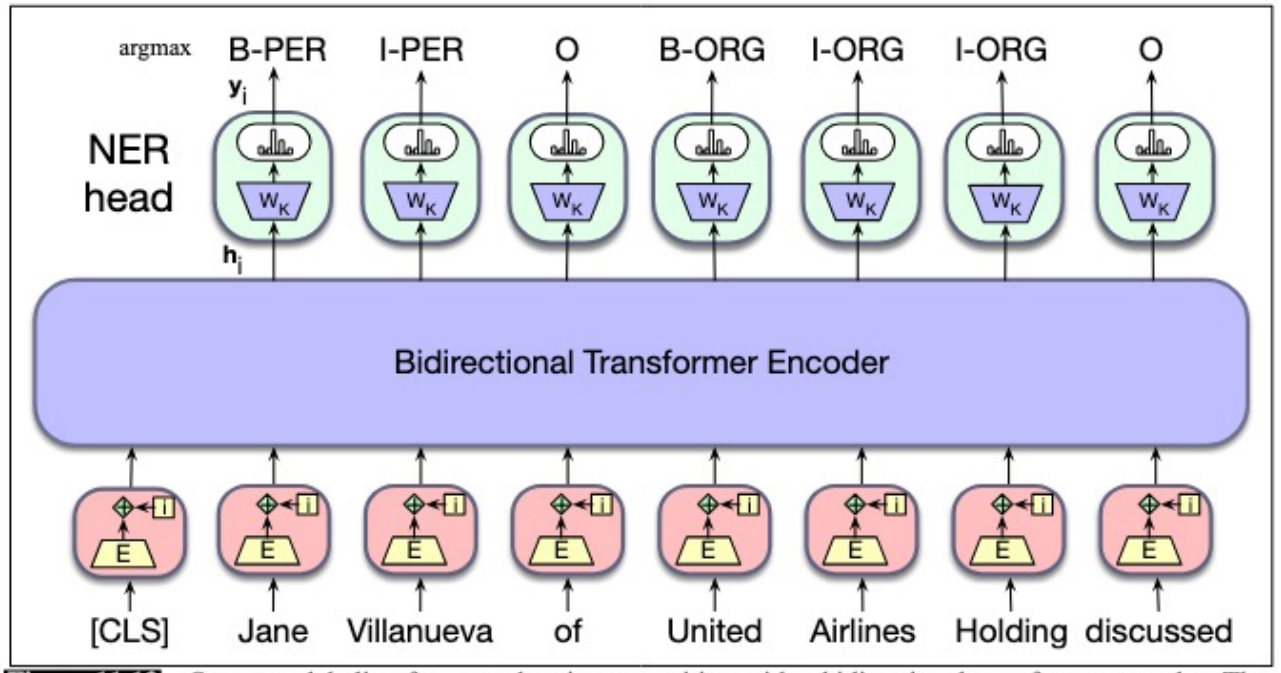


Figure 11.13 Sequence labeling for named entity recognition with a bidirectional transformer encoder. The output vector for each input token is passed to a simple k-way classifier.

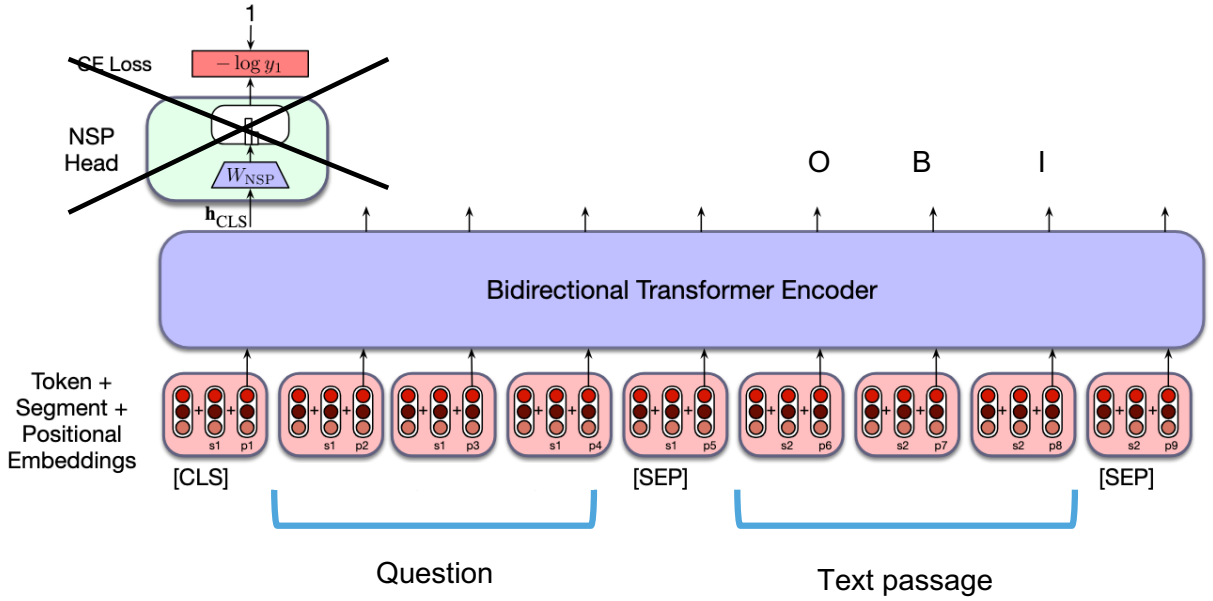
Data set: SQuAd

Approach: BERT

- For BERT

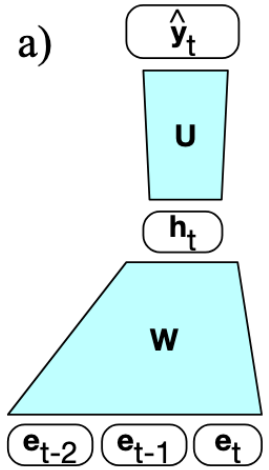
  Train FFNN to label text passage tokens as

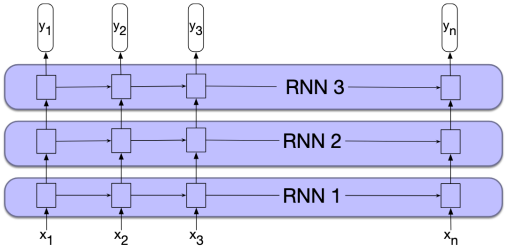
START		B
END	OR	I
Neither		O

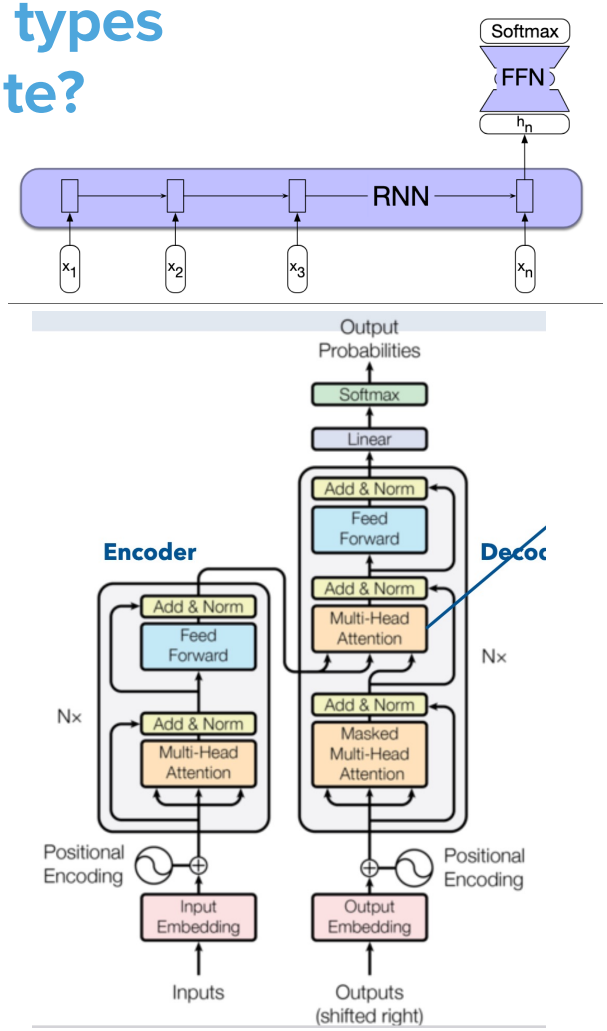


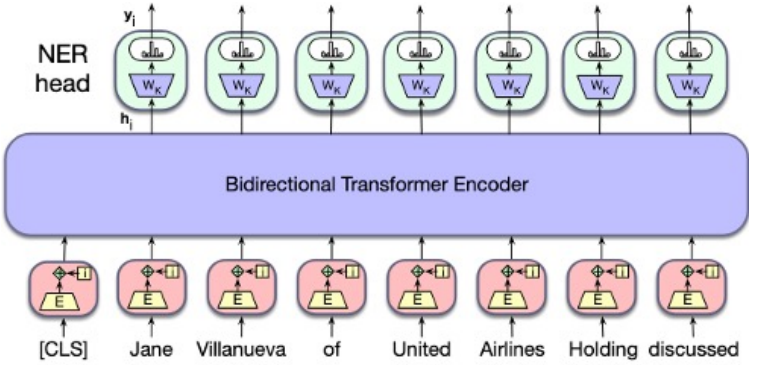
Which other model types might be appropriate?

Which other model types might be appropriate?

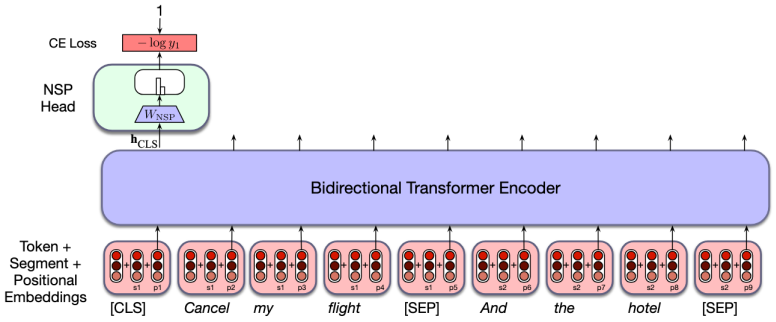








Bert



Data set: GSM8K [2021]

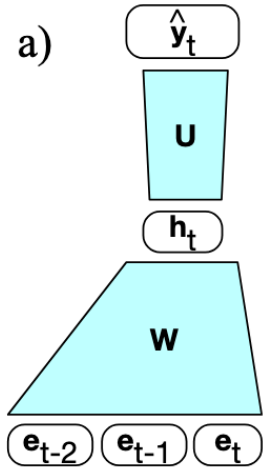
- 8,000 grade school math problems
- Take between 2 and 8 steps to solve
- Example of a “probing” QA task

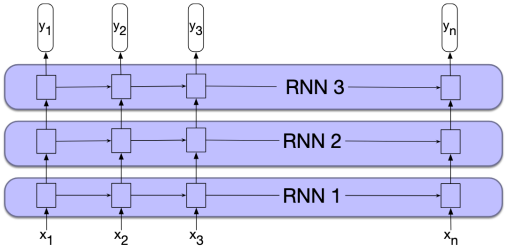
Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

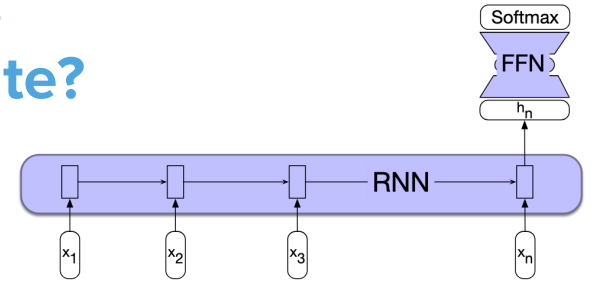
Data set: GSM8k [2021]

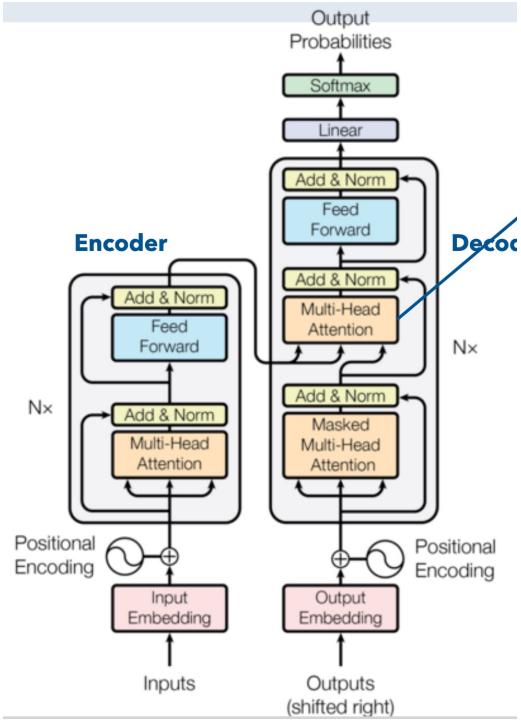
- Each **training set** example is:
  - Question (math word problem) with (optional) calculations annotated
  - Answer
- Each **test set** example is:
  - Question (math word problem) with (optional) calculations annotated

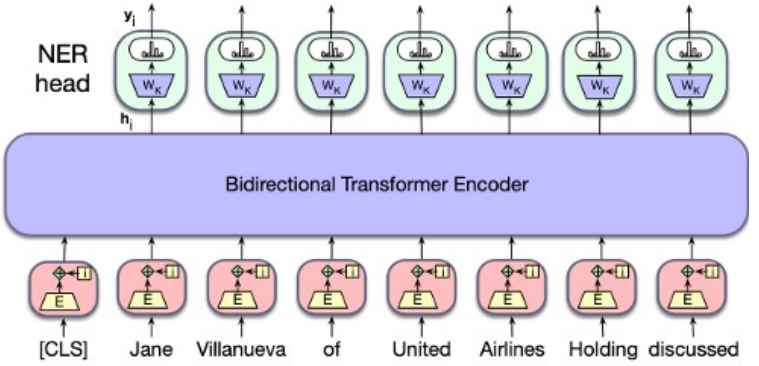
Which model types might be appropriate?



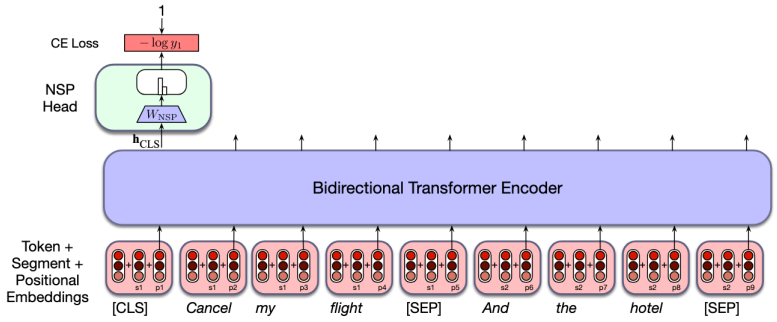








Bert



Data set: HotPotQA

- Wikipedia-based question-answer pairs
  - Questions require finding and reasoning over multiple supporting documents to answer
  - Answers can be “yes”, “no” or a string
  - Full Wikipedia setting vs. 10 paragraphs vs. gold standard paragraphs
  - With and without supporting facts identification

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 1: An example of the multi-hop questions in HotpotQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

Data set: HotPotQA

- One evaluation setting
  - Given question and gold standard paragraphs
  - Identify answer
- More difficult setting
  - Given question and all Wikipedia articles
  - Identify answer

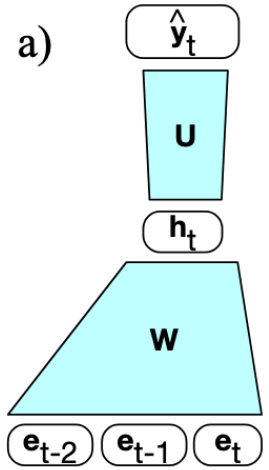
Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

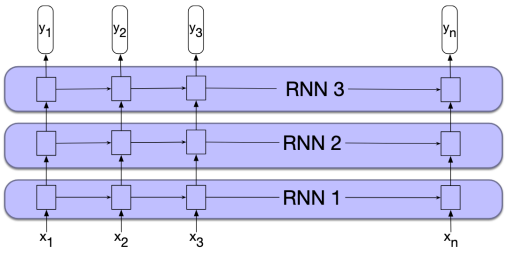
Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

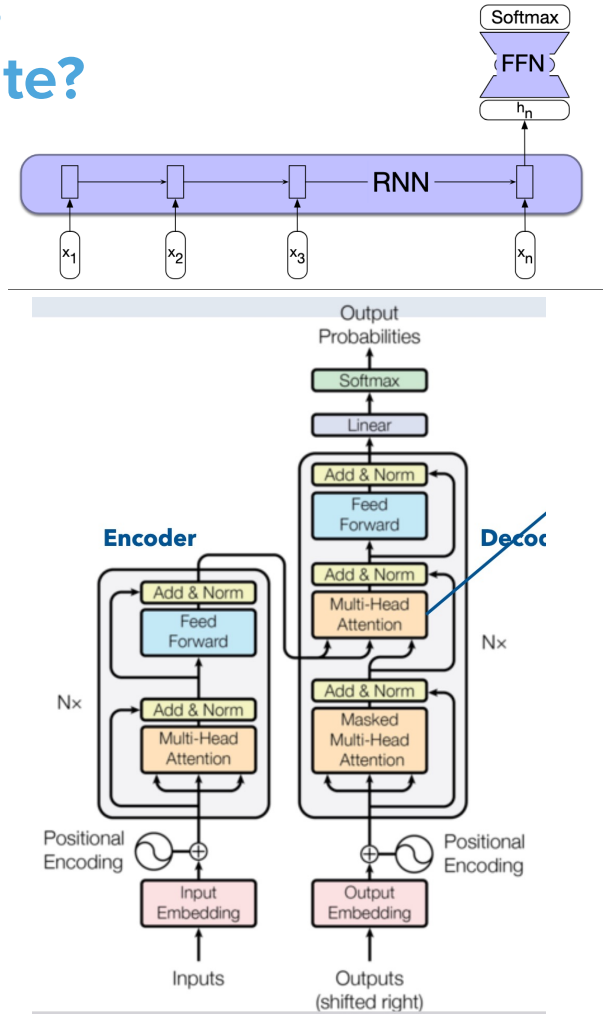
Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

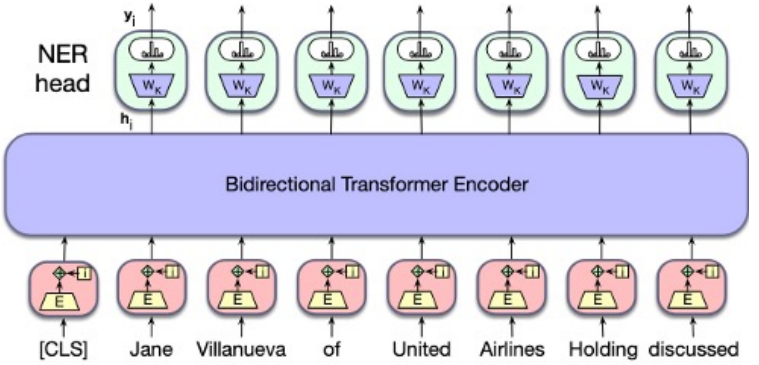
Figure 1: An example of the multi-hop questions in HotpotQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

Which model types might be appropriate?

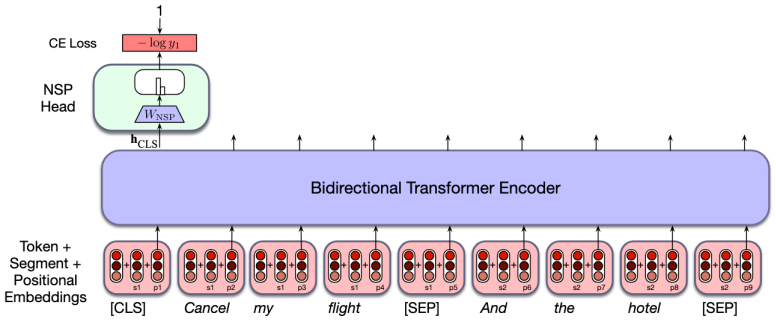








Bert



Better handled using hybrid system

1. Document/paragraph retrieval
2. Decoder-only transformer

Retrieval augmented transformers (RAG)

Today: QA systems (part 2) + RAG

- There are many types of QA paradigms
  - QA Datasets
  - Exercises: How to build systems for them
- Training LLM-based chatbots for QA
- QA using Retrieval-Augmented Generation

Recall: Modern QA approaches

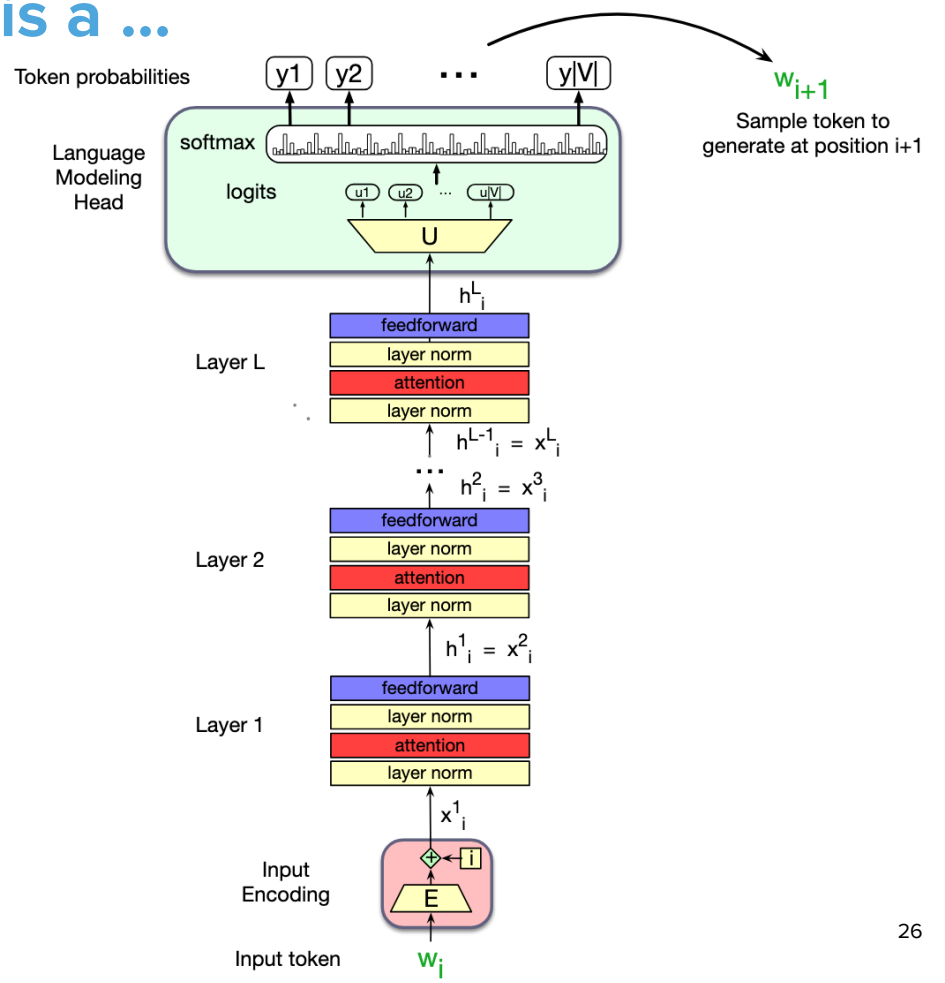
- LLM-based (using transformers) chatbots, e.g. chatGPT
  - Just ask a question
  - System returns the answer

Q: Where is the Louvre Museum located? A: Paris, France

But how are LLMs constructed?

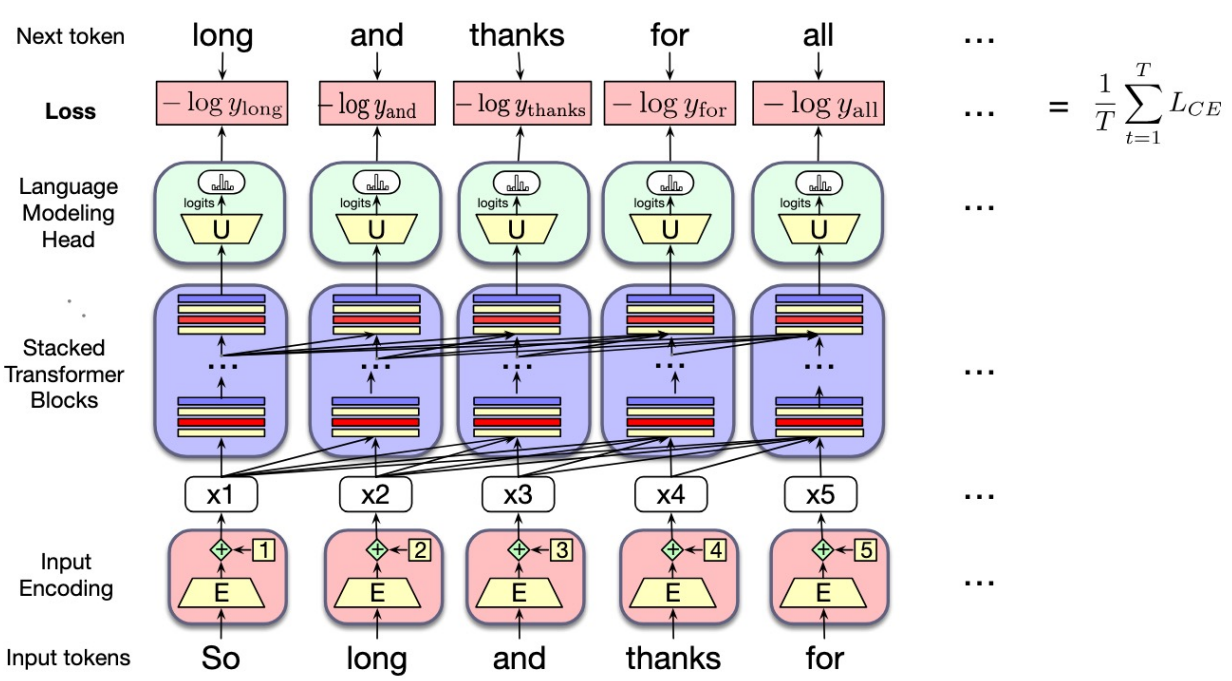
An LLM (large language model) is a ...

- Decoder-only transformer LM



An LLM is a …

- Decoder-only transformer LM
- **Pretrained** with massive amounts of text
  - To **perform next word prediction**
  - Using **self-supervision** (self-training) and **teacher forcing**
  - Cross-entropy is the loss function

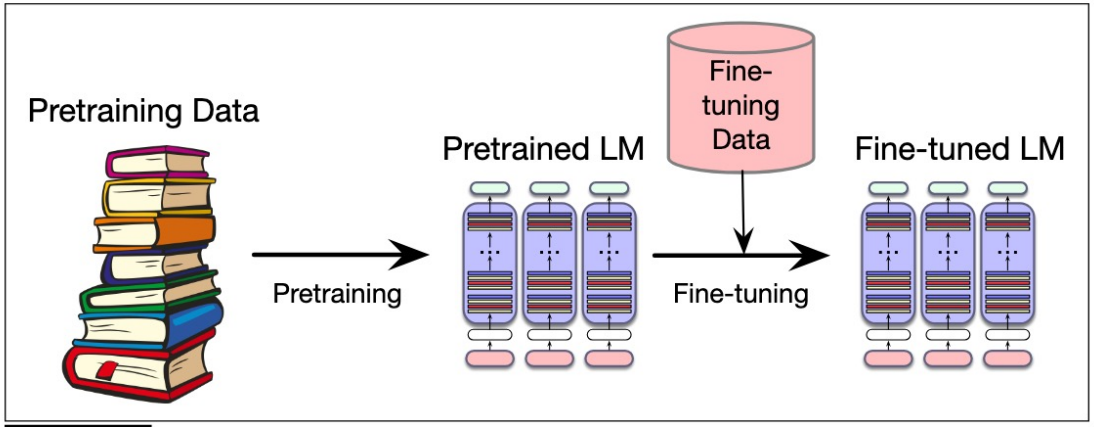


LLMs

- Trained by filling the full context window with text
  - E.g., 4096 tokens for GPT4 or 8192 for Llama 3
  - If documents are shorter
- Multiple documents packed into the window with a special end-of-text token separating them
- Data
  - Web text from the **common crawl**, a series of snapshots of the entire web (each has billions of webpages)
  - **The Pile** (825 GB English text corpus: web, books, Wikipedia)
  - **Dolma** (even larger)

LLMs for QA

- Finetuned for QA using **supervised finetuning (SFT)**
  - This is just one type of finetuning
  - Falls under the umbrella of **instruction finetuning** (i.e., want a pretrained language model to learn to follow text instructions)



LLMs for QA

- Instruction finetuning for QA using **supervised finetuning (SFT)**
  - Create a dataset of example questions and their answers
  - Train the language model on this data using the normal cross-entropy loss to predict each token in the instruction prompt (i.e. the QA pair) iteratively

  Q: Where is the Louvre? **A:** Paris, France

- QA then becomes an instance of **conditional** (text) **generation**
  - Given: **Q:** <question> **A:**
  - LLM should generate the answer

Recall: Problems with current LLMs for QA

- **Hallucinate** responses
- **Not well-calibrated** (confidently produce a very wrong answer)
- Cannot be used on proprietary data
- Cannot deal with rapidly changing information

Today: QA systems (part 2) + RAG

- There are many types of QA paradigms
  - QA Datasets
  - Exercises: How to build systems for them
- Training LLM-based chatbots for QA
- QA using Retrieval-Augmented Generation

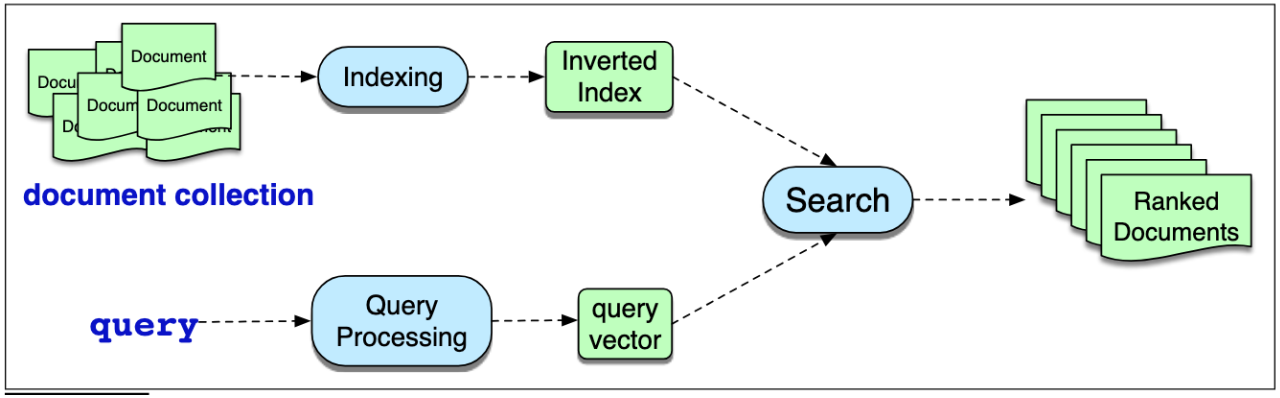
Retrieval-Augmented Generation (RAG)

- Use information retrieval (IR) techniques to retrieve documents that are likely to have information that might help answer the question.
- Then use an LLM to generate an answer given these documents.

How does this alleviate the problems we listed two slides back???

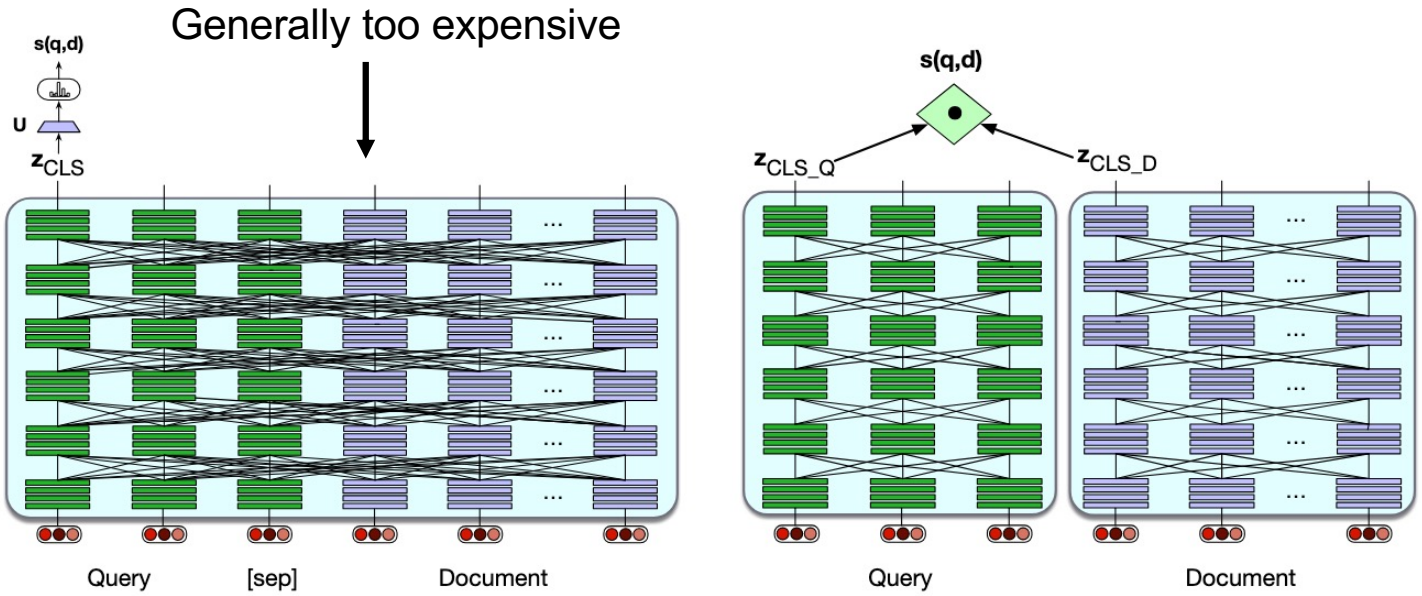
Information retrieval

- Traditionally used **bag-of-words representation** for both the ***query*** and the ***documents***
- Retrieve the documents that are most similar to the query



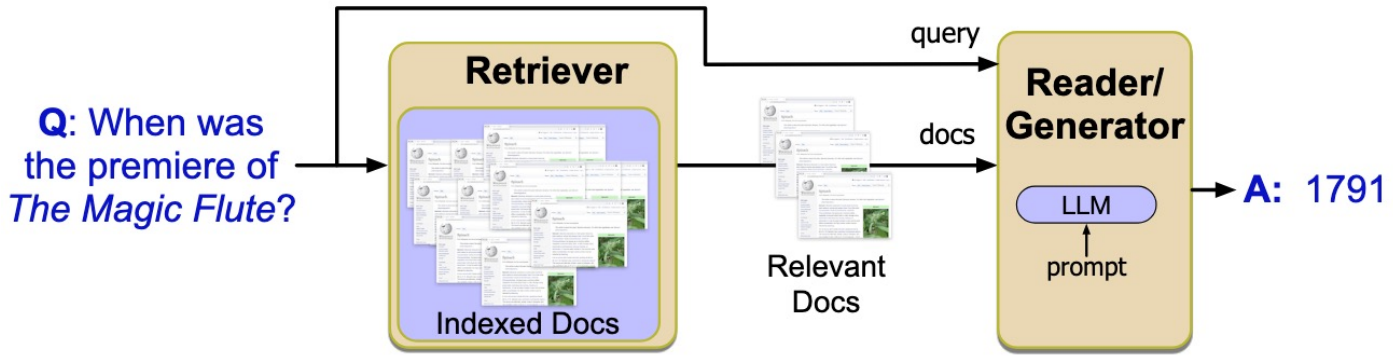
IR with dense retrievers

- Use **dense vectors** to represent both the ***query*** and the ***documents***



Retrieval-augmented Generation

- Two steps



Retrieval-augmented Generation

- Condition on the retrieved passages as part of the QA prompt

  Q: Who wrote the book “The Origin of Species"?

Schematic of a RAG Prompt

```
retrieved passage 1

retrieved passage 2

...

retrieved passage n

Based on these texts, answer this question:  Q: Who wrote
the book ‘‘The Origin of Species"?  A:
```

Evaluating QA systems

- Multiple choice and some math data sets
  - Accuracy according to exact match
- Free form answers
  - Token-level F1 score
- Ranked list of answers
  - Mean reciprocal rank (MRR)
    - Each test set question is scored with the reciprocal of the rank of the first correct answer
    - The score for questions that return no correct answer is 0.
    - Average the scores over all questions

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$