CS 2731
Human Language Technologies

Session 15: LLMs, project peer group feedback

Michael Miller Yoder

October 15, 2025

University of Pittsburgh | School of Computing and Information

Course logistics: project

- Next project deliverable: project proposal due **Fri Oct 17**
  - Include plans for **task, data, methods, evaluation**
  - Compare multiple approaches, including an LLM-based approach
  - Literature review of at least 3 related papers
    - How have others approached your task or similar tasks? What are other NLP papers that use the same dataset or domain as your project?
    - Baselines to compare your approach to
  - Feel free to email or book office hours with Michael to discuss

Course logistics: project

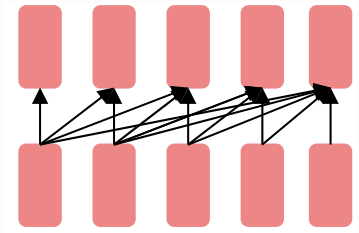
- Project proposal presentations in-class **next Mon Oct 20**
  - Add your slides to this shared PowerPoint
  - Ungraded
- LLM access
  - We have $150 total as a class to use on OpenAI LLM credits
  - Access to open-source LLM set up on School of Computing and Information servers for API access has been set up!
    - We will only have access to Gemma 3
    - Michael will give a how-to in a future class session

Overview: LLMs, project peer group feedback

- Types of LLMs: encoders, encoder-decoders, decoders
- Sampling for LLM generation
- Harms from LLMs
- Project peer group feedback

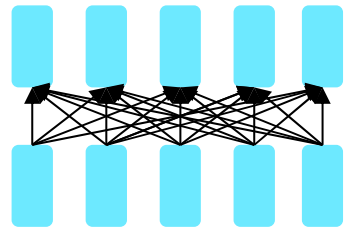
3 types of LLMs:
encoders, encoder-decoders, decoders

Three architectures for large language models



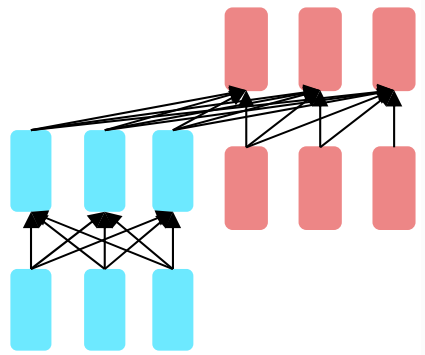
Decoders

GPT, Claude, Llama, Mixtral



Encoders

BERT family, HuBERT



Encoder-decoders

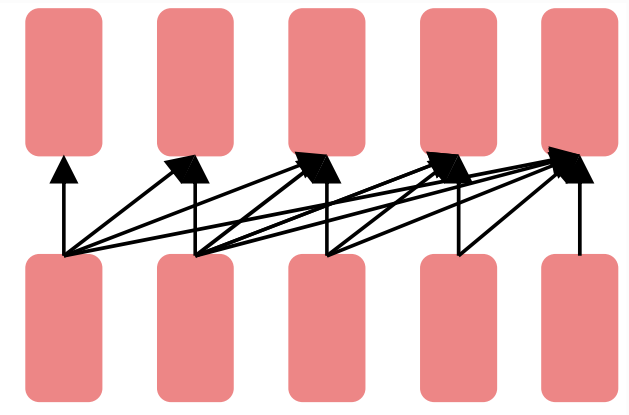
Flan-T5, Whisper

Slide adapted from Jurafsky and Martin

Decoder-only models

Also called:

- Causal LLMs
- Autoregressive LLMs
- Left-to-right LLMs



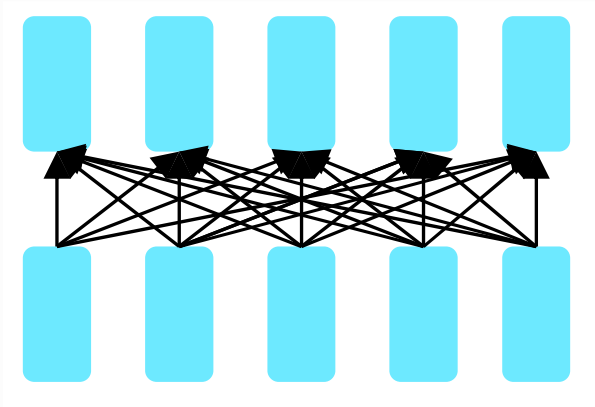
- Predict words left to right

Slide adapted from Jurafsky and Martin

Encoders

Many varieties!

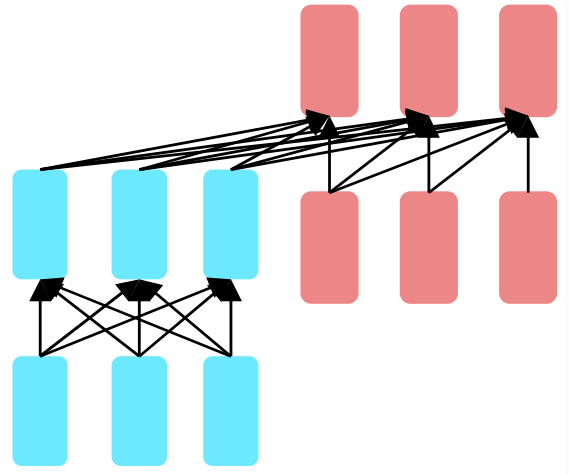
- Popular: Masked Language Models (MLMs)
- BERT family
- Trained by predicting words from surrounding words on both sides
- Are usually **finetuned** (trained on supervised data) for classification tasks.



Slide adapted from Jurafsky and Martin

Encoder-Decoders

- Trained to map from one sequence to another (sequence to sequence)
- Popular for:
  - machine translation: map from one language to another
  - speech recognition: map from acoustics to words



Slide adapted from Jurafsky and Martin

Decoder LLMs

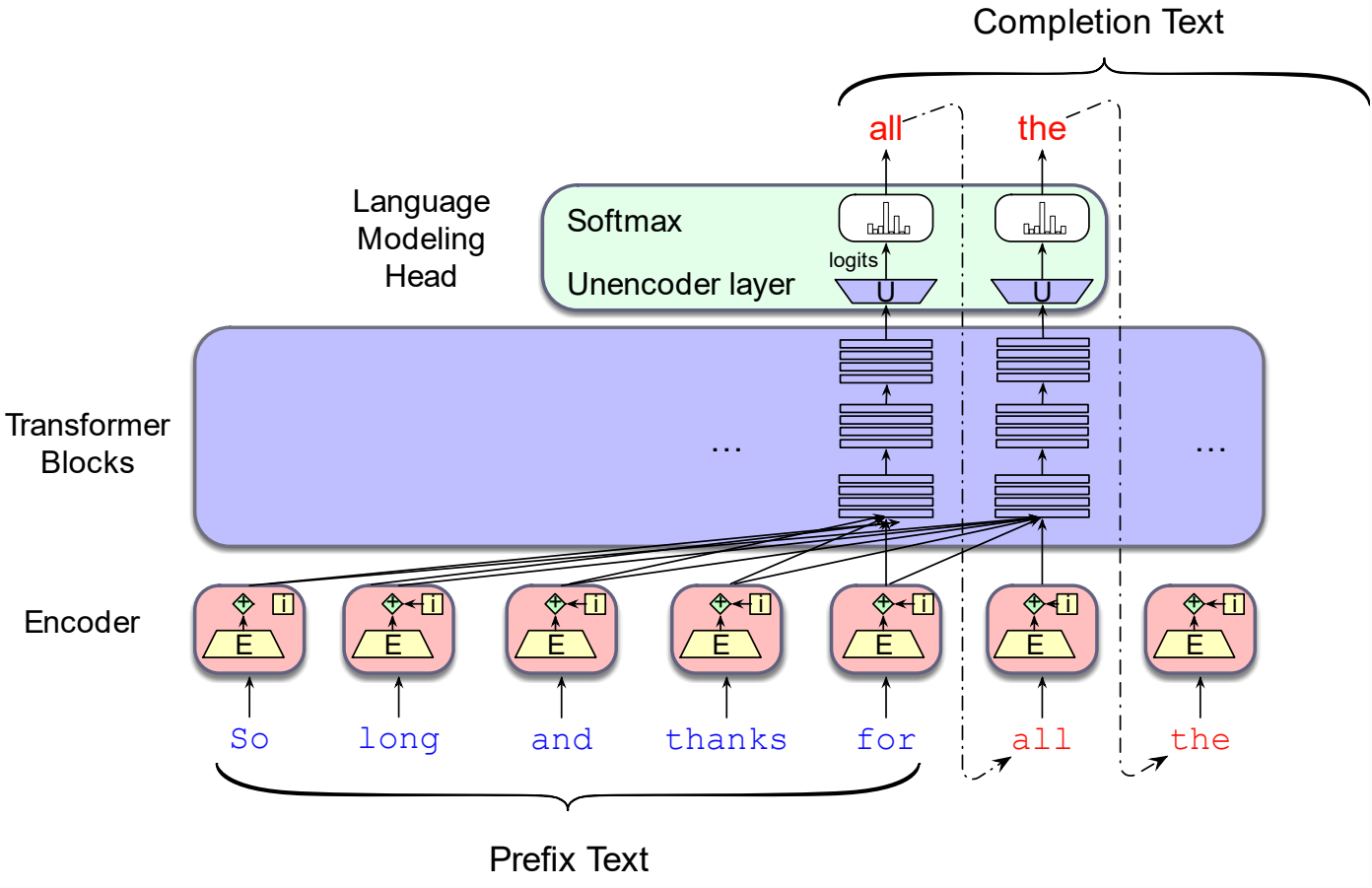
Decoder-only models can handle many tasks

- Many tasks can be turned into tasks of predicting words!

Slide adapted from Jurafsky and Martin

Conditional generation

Generating text conditioned on previous text!



Slide adapted from Jurafsky and Martin

Many practical NLP tasks can be cast as word prediction!

Sentiment analysis: “I like Jackie Chan”

1. We give the language model this string:
   `The sentiment of the sentence "I like Jackie Chan" is:`
2. And see what word it thinks comes next:

$P(\text{positive} \mid \texttt{The sentiment of the sentence ``I like Jackie Chan" is:})$

$P(\text{negative} \mid \texttt{The sentiment of the sentence ``I like Jackie Chan" is:})$

Slide adapted from Jurafsky and Martin

Framing lots of tasks as conditional generation

QA: “Who wrote The Origin of Species”

1. We give the language model this string:

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

2. And see what word it thinks comes next:

$P(w|$`Q: Who wrote the book ``The Origin of Species"? A:`$)$

3. And iterate:

$P(w|$`Q: Who wrote the book ``The Origin of Species"? A: Charles`$)$

Slide adapted from Jurafsky and Martin

Summarization

Original

The only thing crazier than a guy in snowbound Massachusetts boxing up the powdery white stuff and offering it for sale online? People are actually buying it. For $89, self-styled entrepreneur Kyle Waring will ship you 6 pounds of Boston-area snow in an insulated Styrofoam box – enough for 10 to 15 snowballs, he says.

But not if you live in New England or surrounding states. “We will not ship snow to any states in the northeast!” says Waring’s website, ShipSnowYo.com. “We’re in the business of expunging snow!”

His website and social media accounts claim to have filled more than 133 orders for snow – more than 30 on Tuesday alone, his busiest day yet. With more than 45 total inches, Boston has set a record this winter for the snowiest month in its history. Most residents see the huge piles of snow choking their yards and sidewalks as a nuisance, but Waring saw an opportunity.

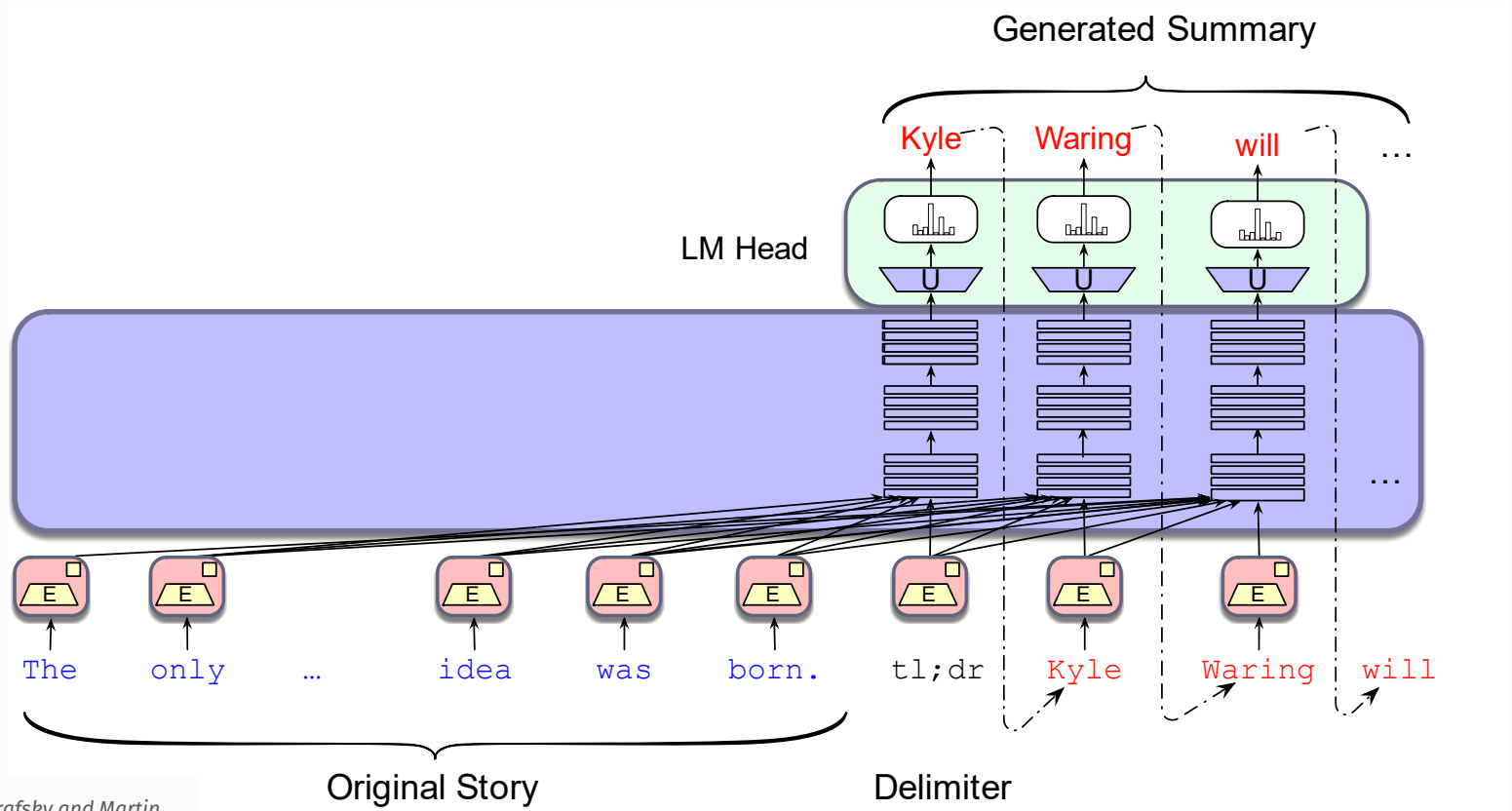
According to Boston.com, it all started a few weeks ago, when Waring and his wife were shoveling deep snow from their yard in Manchester-by-the-Sea, a coastal suburb north of Boston. He joked about shipping the stuff to friends and family in warmer states, and an idea was born. [...]

Summary

Kyle Waring will ship you 6 pounds of Boston-area snow in an insulated Styrofoam box – enough for 10 to 15 snowballs, he says. But not if you live in New England or surrounding states.

Slide adapted from Jurafsky and Martin

LLMs for summarization (using tl;dr)



Slide adapted from Jurafsky and Martin

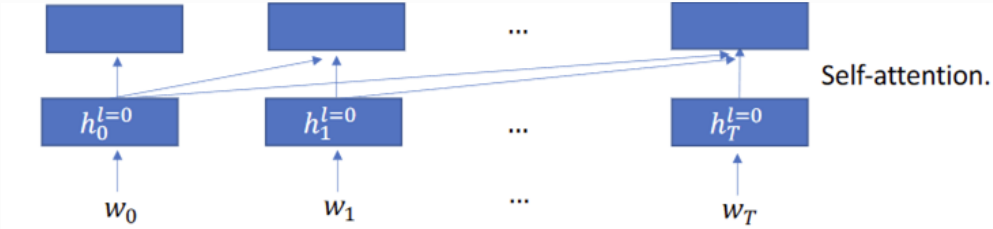
Pretraining decoder LLMs

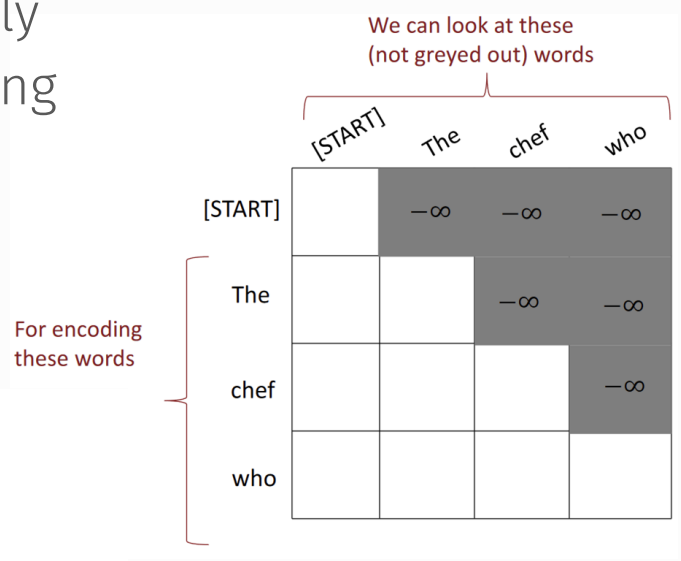
- Take a corpus and ask the model to predict the next word!
- Train the model using gradient descent to minimize the error
- Same loss function as other neural models: cross-entropy loss
- Move the weights in the direction that assigns a higher probability to the true next word

Slide adapted from Jurafsky and Martin

Decoding: apply a “causal mask” for self-attention

- To do auto-regressive LM, we need to apply a “causal” mask to self-attention, forbidding it from getting future context.
- At timestep t, we set a_i = 0 for $i > t$

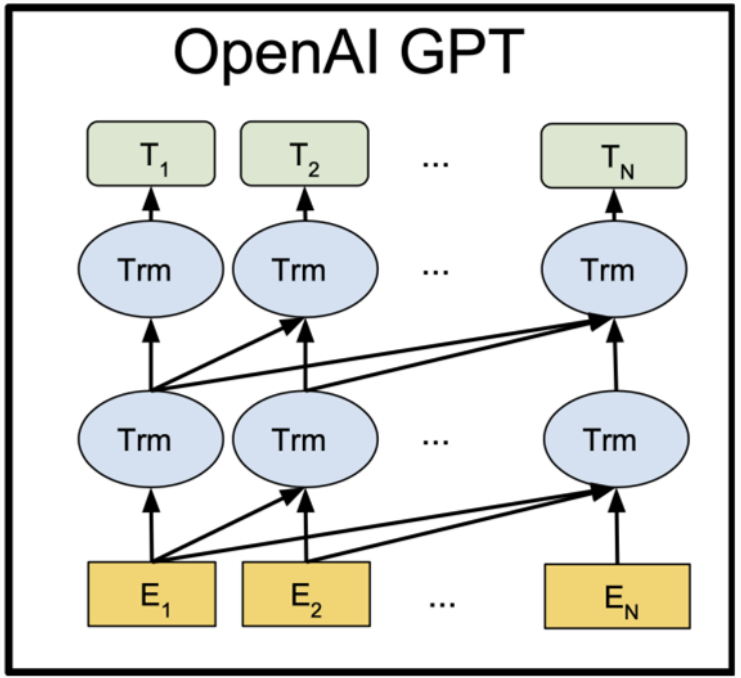




Slide adapted from Tianxing He, John Hewitt

Generative Pretrained Transformer (GPT; Radford et al. 2018)

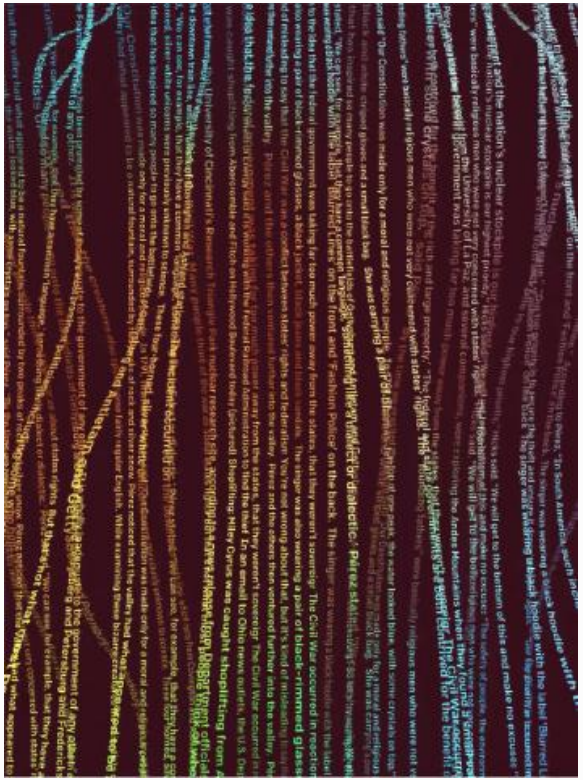
- 2018's GPT was a big success in pretraining a decoder!
- Transformer decoder with 12 layers, 117M parameters.
- 768-dimensional hidden states, 3072-dimensional feed-forward hidden layers.
- Trained on BooksCorpus: over 7000 unique books.
    - Contains long spans of contiguous text, for learning long-distance dependencies.



Slide adapted from John Hewitt, David Mortensen

GPT-2, GPT-3, GPT-4, GPT-5 from OpenAI

- They are basically larger and larger autoregressive transformer LMs trained on larger and larger amounts of data
- They have shown amazing language generation capability when you give it a prompt (aka. prefix, the beginning of a paragraph)



Slide adapted from Tianxing He

Generation example from the GPT-2 model

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

A sample from GPT2 (with top-k sampling)

Slide credit: Tianxing He

Sampling for LLM generation

Decoding and Sampling

- This task of choosing a word to generate based on the model's probabilities is called **decoding**.
- The most common method for decoding in LLMs: **sampling**.
- Sampling from a model's distribution over words:
  - choose random words according to their probability assigned by the model.
- After each token we'll sample words to generate according to their probability *conditioned on our previous choices*,
  - A transformer language model will give the probability

Slide adapted from Jurafsky and Martin

Random sampling

$\mathrm{i} \leftarrow 1$

$w_i \sim \mathrm{p}(w)$

while w_i != EOS

$\quad \mathrm{i} \leftarrow \mathrm{i} + 1$

$\quad w_i \sim \mathrm{p}(w_i \mid w_{<i})$

Slide adapted from Jurafsky and Martin

Random sampling doesn't work very well

- Even though random sampling mostly generate sensible, high-probable words,
- There are many odd, low- probability words in the tail of the distribution
- Each one is low- probability but added up they constitute a large portion of the distribution
- So they get picked enough to generate weird sentences

Slide adapted from Jurafsky and Martin

Factors in word sampling: **quality** and **diversity**

Emphasize **high-probability** words

\+ **quality**: more accurate, coherent, and factual,

\- **diversity**: boring, repetitive.

Emphasize **middle-probability** words

\+ **diversity**: more creative, diverse,

\- **quality**: less factual, incoherent

Slide adapted from Jurafsky and Martin

Top-k sampling:

1. Choose # of words k
2. For each word in the vocabulary V , use the language model to compute the likelihood of this word given the context $p(w_t \mid w_{<t})$
3. Sort the words by likelihood, keep only the top k most probable words.
4. Renormalize the scores of the k words to be a legitimate probability distribution.
5. Randomly sample a word from within these remaining k most-probable words according to its probability.

Slide adapted from Jurafsky and Martin

Temperature sampling

Reshape the distribution instead of truncating it

Intuition from thermodynamics,

- a system at high temperature is flexible and can explore many possible states,
- a system at lower temperature is likely to explore a subset of lower energy (better) states.

In **low-temperature sampling**, ($\tau \leq 1$) we smoothly

- increase the probability of the most probable words
- decrease the probability of the rare words.

Slide adapted from Jurafsky and Martin

Temperature sampling

Divide the output by a temperature parameter τ before passing it through the softmax.

Instead of

$$\mathbf{y} = \text{softmax}(u)$$

We do

$$\mathbf{y} = \text{softmax}(u/\tau)$$

A lower τ pushes high-probability words higher and low probability word lower due to the way softmax works

Slide adapted from Jurafsky and Martin

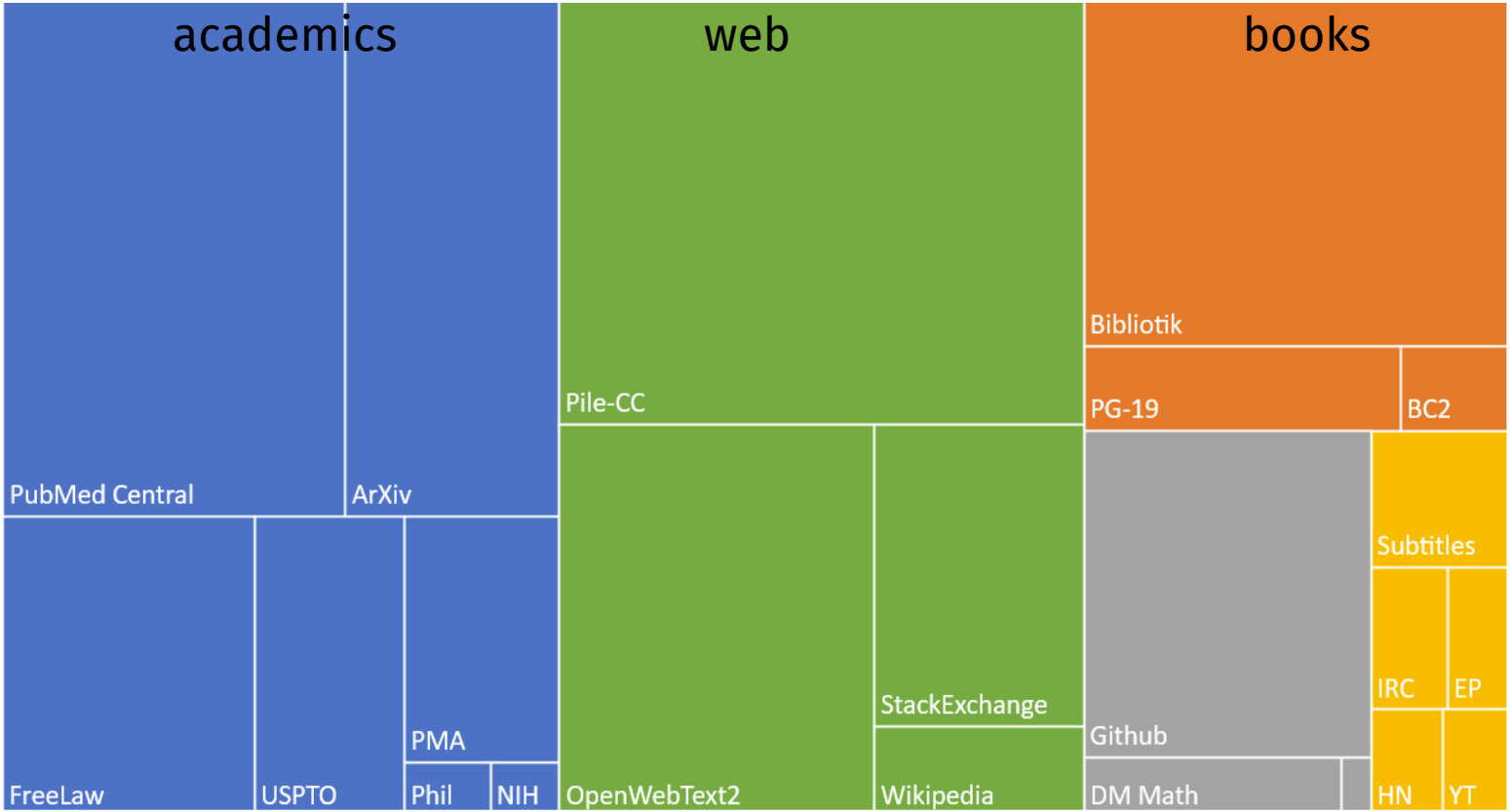
Pretraining data and harms of LLMs

LLMs are mainly trained on the web

- Common crawl, snapshots of the entire web produced by the non-profit Common Crawl with billions of pages
- Colossal Clean Crawled Corpus (C4; Raffel et al. 2020), 156 billion tokens of English, filtered
- What's in it? Mostly patent text documents, Wikipedia, and news sites

Slide adapted from Jurafsky and Martin

The Pile: a pretraining corpus



dialog

Slide adapted from Jurafsky and Martin

Big idea

- Text contains enormous amounts of knowledge
- Pretraining on lots of text with all that knowledge is what gives language models their ability to do so much

Slide adapted from Jurafsky and Martin

But there are problems with scraping from the web

- Copyright: much of the text in these datasets is copyrighted
  - Not clear if fair use doctrine in US allows for this use
  - This remains an open legal question
- Data consent
  - Website owners can indicate they don't want their site crawled
- Privacy:
  - Websites can contain private IP addresses and phone numbers

Slide adapted from Jurafsky and Martin

Harms from LLMs

Hallucination

What Can You Do When A.I. Lies About You?

People have little protection or recourse when the technology creates and spreads falsehoods about them.

Air Canada loses court case after its chatbot hallucinated fake policies to a customer

The airline argued that the chatbot itself was liable. The court disagreed.

Copyright

Authors Sue OpenAI Claiming Mass Copyright Infringement of Hundreds of Thousands of Novels

Privacy

Slide adapted from Jurafsky and Martin

Harms from LLMs

Toxicity and abuse

The New AI-Powered Bing Is Threatening Users.

Cleaning Up ChatGPT Takes Heavy Toll on Human Workers

Contractors in Kenya say they were traumatized by effort to screen out descriptions of violence and sexual abuse during run-up to OpenAI's hit chatbot

Misinformation

Chatbots are generating false and misleading information about U.S. elections

Slide adapted from Jurafsky and Martin

Conclusion

- Transformer-based language models pretrained on lots of text are called **large language models (LLMs)**
- LLMs can have decoder-only, encoder-only, or encoder-decoder architectures
- Decoder-only LLMs can cast many different NLP tasks as word prediction
- There are many different sampling approaches that balance diversity and quality in text generation from LLMs
- Harms from LLMs include hallucinating false information, leaking private information from training data, generating abuse and misinformation

Project peer group feedback

Project peer group feedback

1. Find another group to work with
2. Present an overview of your project: 5 min
3. Other group asks clarifying questions
4. Presenting group asks for any advice or guidance from the other group on lingering questions about the project proposal
5. Switch groups when Michael says