Pre-trained LMs

CS 685, Fall 2025

Advanced Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

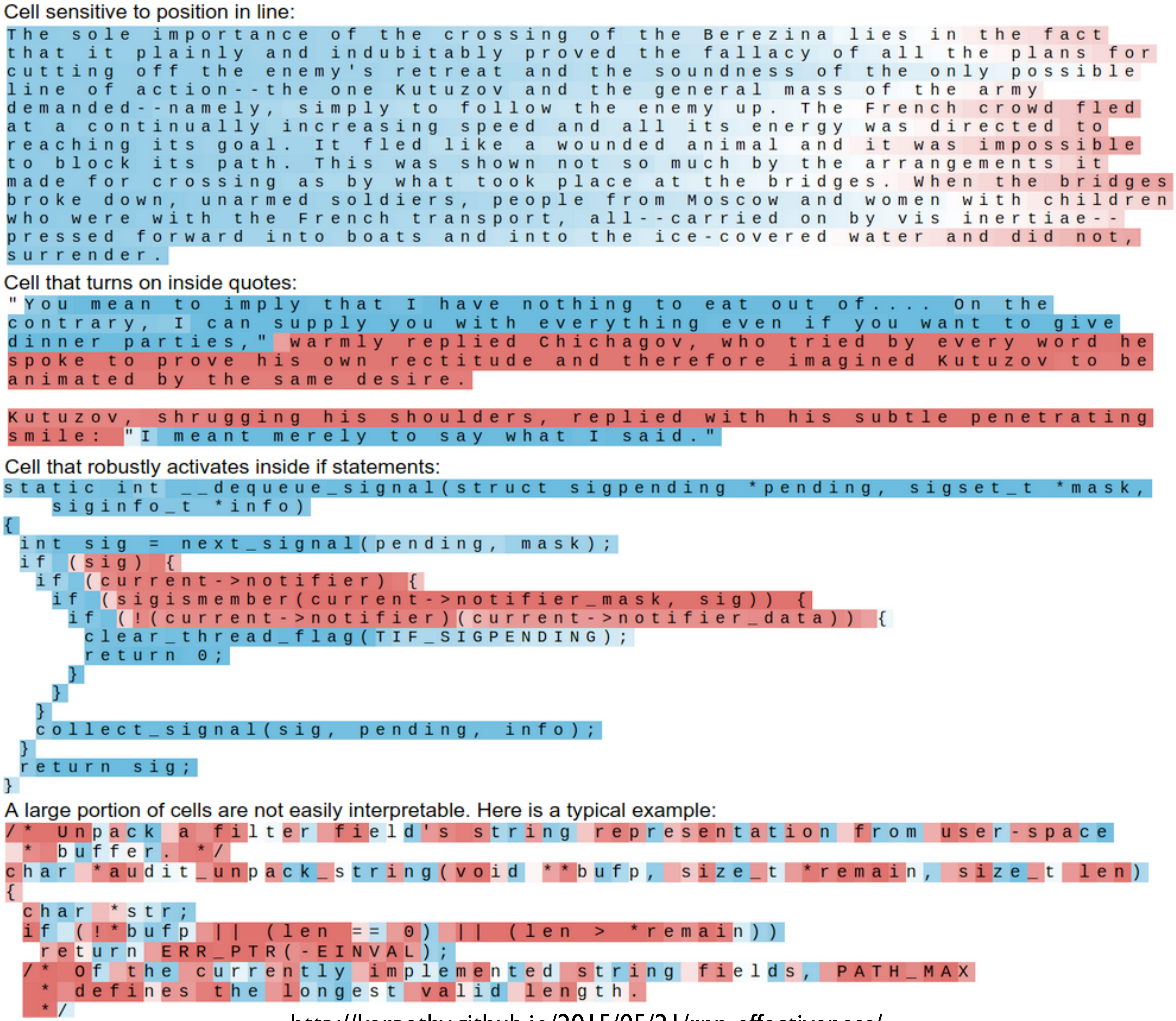
Character LMs comparison: LSTM vs. N-Gram

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.
```

```
First Citizen:
Nay, then, that was hers,
It speaks against your other service:
But since the
youth of the circumstance be spoken:
Your uncle and one Baptista's daughter.

SEBASTIAN:
Do I stand till the break off.

BIRON:
Hide thy head.

VENTIDIUS:
He purposeth to Athens: whither, with the vow
I made to handle you.
```

http://karpathy.github.io/2015/05/21/rnn-effectiveness/
http://nbviewer.jupyter.org/gist/yoavg/d76121dfde2618422139

Structure awareness



Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport, all--carried on by vis inertiae--
pressed forward into boats and into the ice-covered water and did not,
surrender.

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
   siginfo_t *info)
{
 int sig = next_signal(pending, mask);
 if (sig) {
  if (current->notifier) {
   if (sigismember(current->notifier_mask, sig)) {
    if (!(current->notifier)(current->notifier_data)) {
     clear_thread_flag(TIF_SIGPENDING);
     return 0;
    }
   }
  }
  collect_signal(sig, pending, info);
 }
 return sig;
}
```

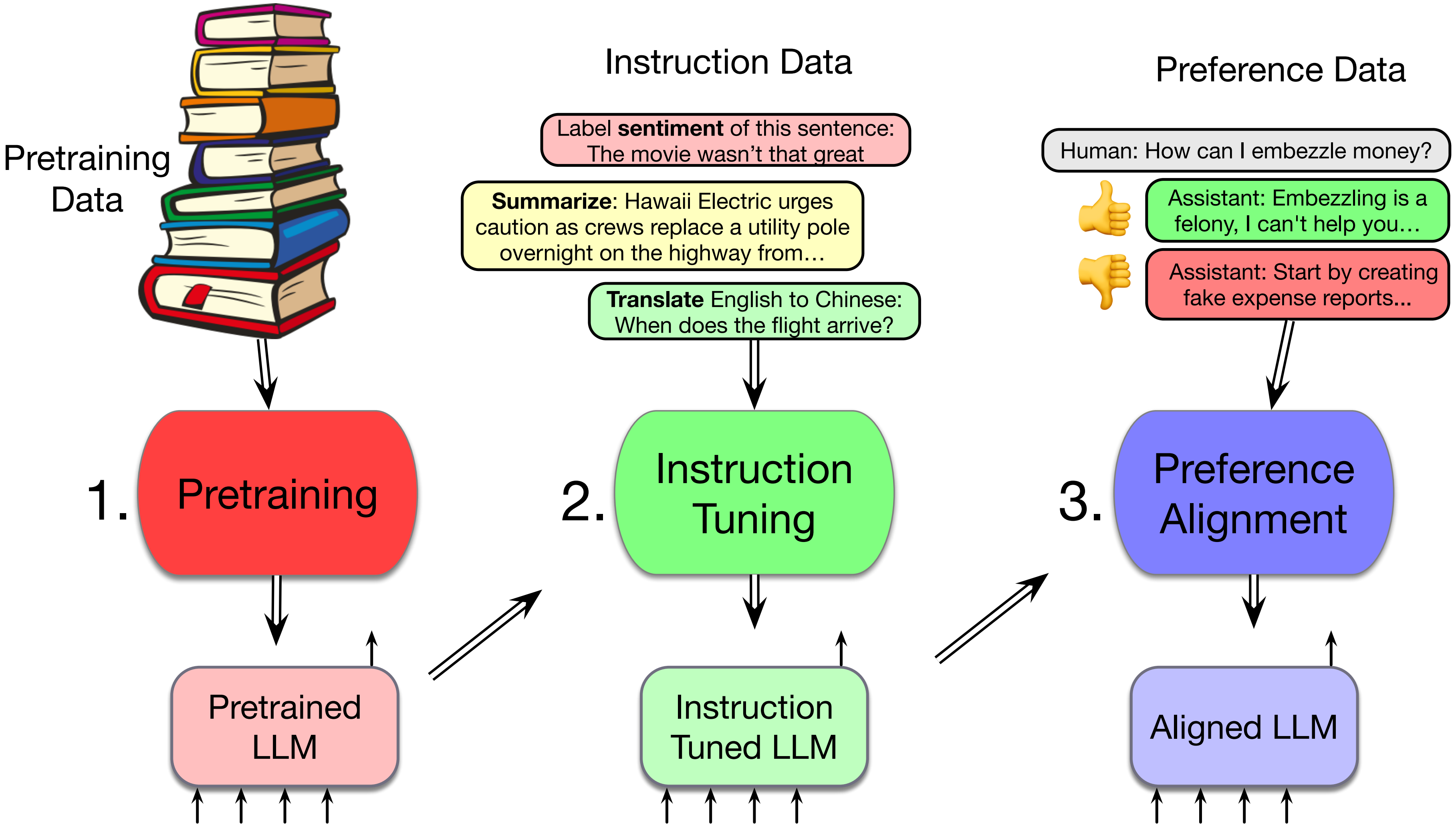
A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
 char *str;
 if (!*bufp || (len == 0) || (len > *remain))
  return ERR_PTR(-EINVAL);
 /* Of the currently implemented string fields, PATH_MAX
  * defines the longest valid length.
  */
```

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

- How to *use* a pretrained LM to do useful tasks?
    - Encoding (e.g. in BOE)
    - Fine-tuning (today)
    - Directly use generation (later in semester)

Three stages of training in LLMs



Pretraining Data

Instruction Data

Label **sentiment** of this sentence: The movie wasn't that great

Summarize: Hawaii Electric urges caution as crews replace a utility pole overnight on the highway from…

Translate English to Chinese: When does the flight arrive?

Preference Data

Human: How can I embezzle money?

👍 Assistant: Embezzling is a felony, I can't help you…

👎 Assistant: Start by creating fake expense reports...

1. Pretraining

2. Instruction Tuning

3. Preference Alignment

Pretrained LLM

Instruction Tuned LLM

Aligned LLM

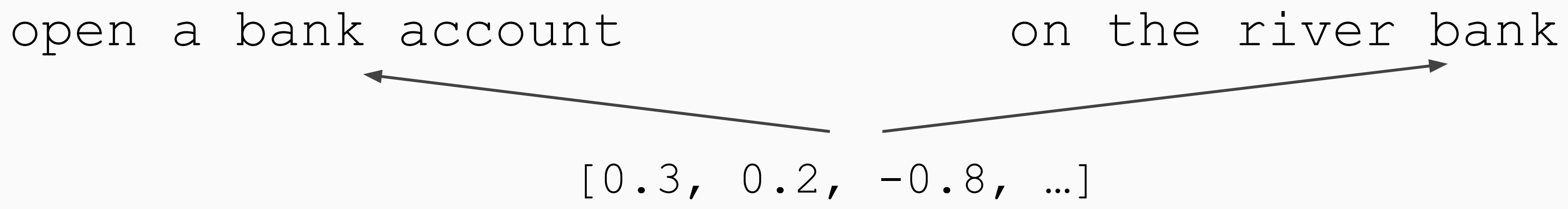
[Slide: SLP3]

can we use language models to produce word embeddings?

Deep contextualized word representations. Peters et al., NAACL 2018

Contextual Representations

- **Problem**: Word embeddings are applied in a context free manner

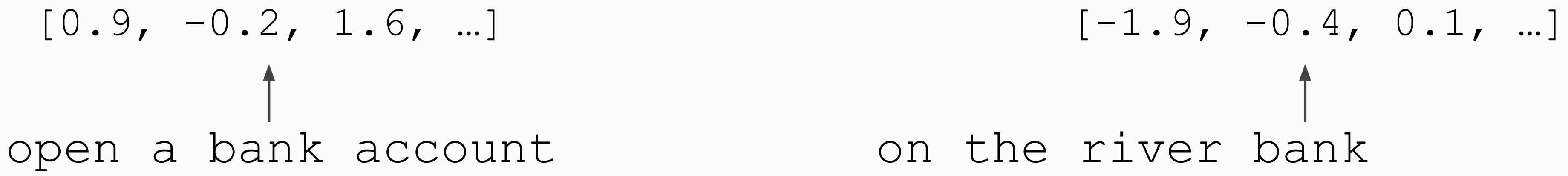


open a bank account

on the river bank

[0.3, 0.2, -0.8, …]

- **Solution**: Train *contextual* representations on text corpus



[0.9, -0.2, 1.6, …]

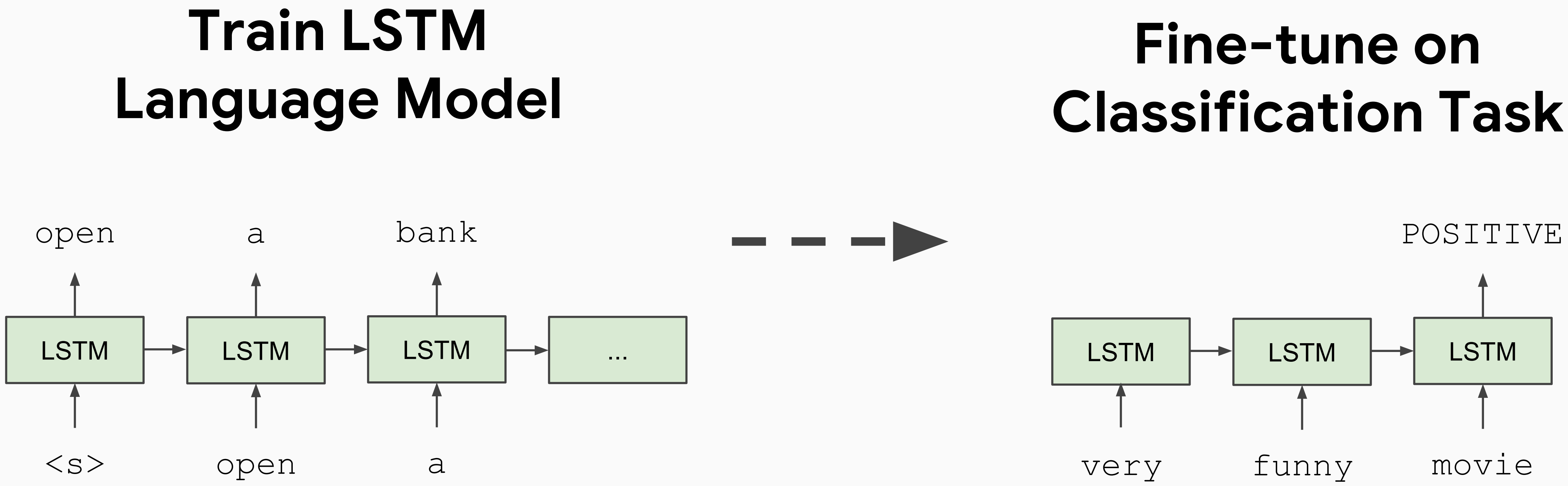
open a bank account

[-1.9, -0.4, 0.1, …]

on the river bank

History of Contextual Representations

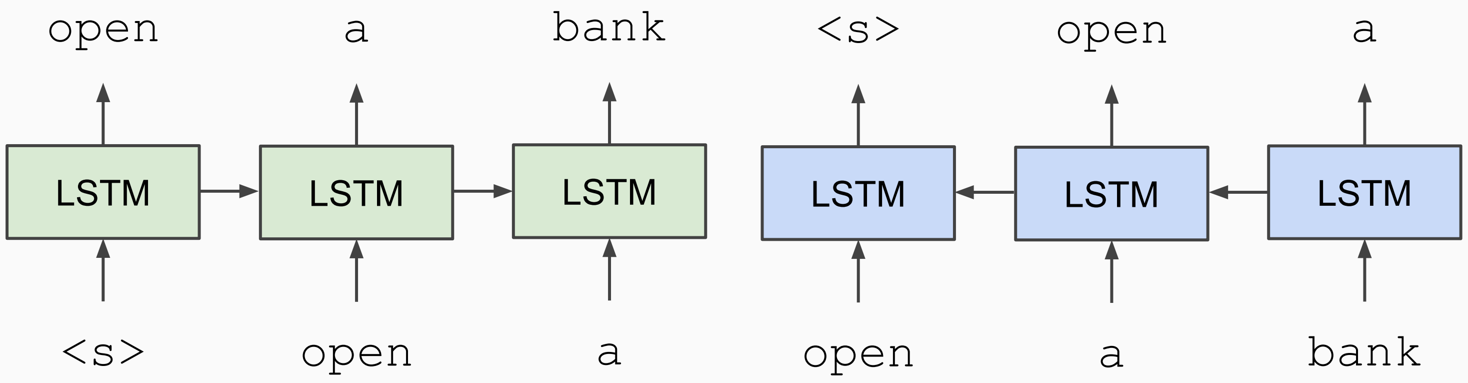
- *Semi-Supervised Sequence Learning*, Google, 2015



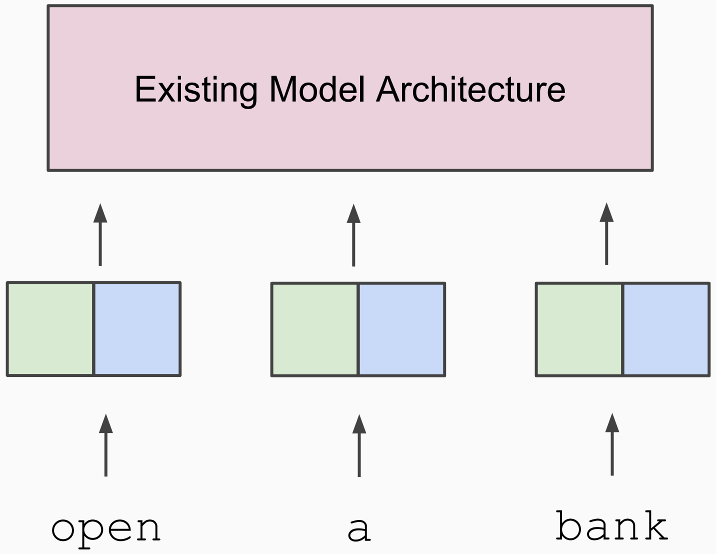
History of Contextual Representations

- *ELMo: Deep Contextual Word Embeddings*, AI2 & University of Washington, 2017

Train Separate Left-to-Right and Right-to-Left LMs

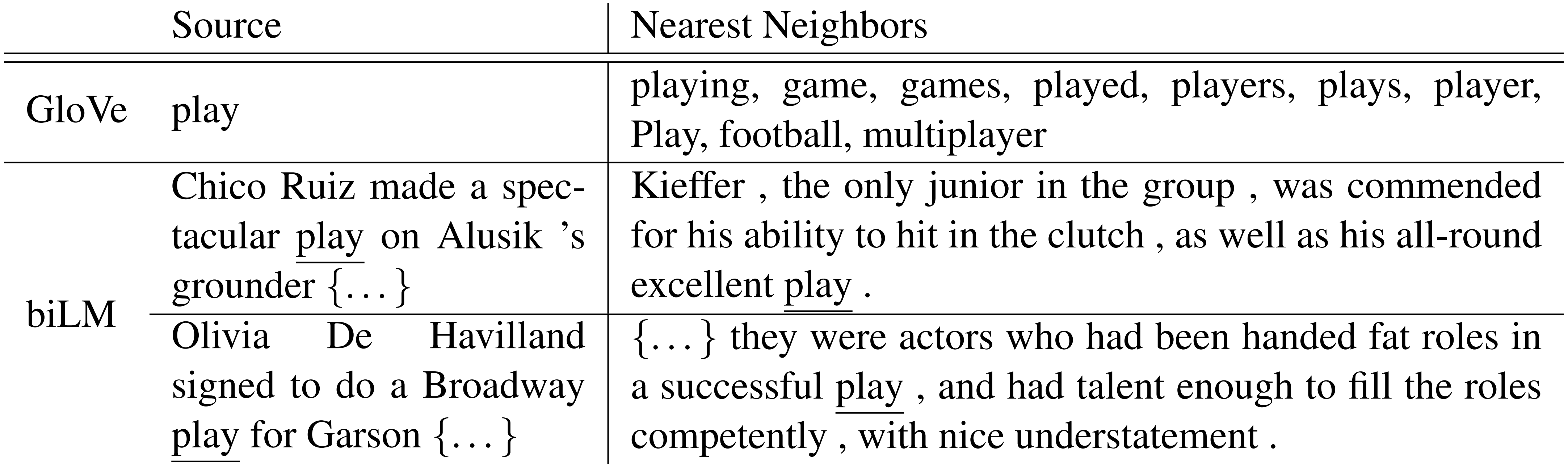


Apply as "Pre-trained Embeddings"



[Peters et al., 2018]

Context-specific word sense



	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {. . . }	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {. . . }	{. . . } they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

[Peters et al., 2018]

Self-supervised training algorithm

We train them to predict the next word!

1. Take a corpus of text
2. At each time step *t*
    i. ask the model to predict the next word
    ii. train the model using gradient descent to minimize the error in this prediction

"**Self-supervised**" because it just uses the next word as the label!

[Slide: SLP3]

LLMs are mainly trained on the web

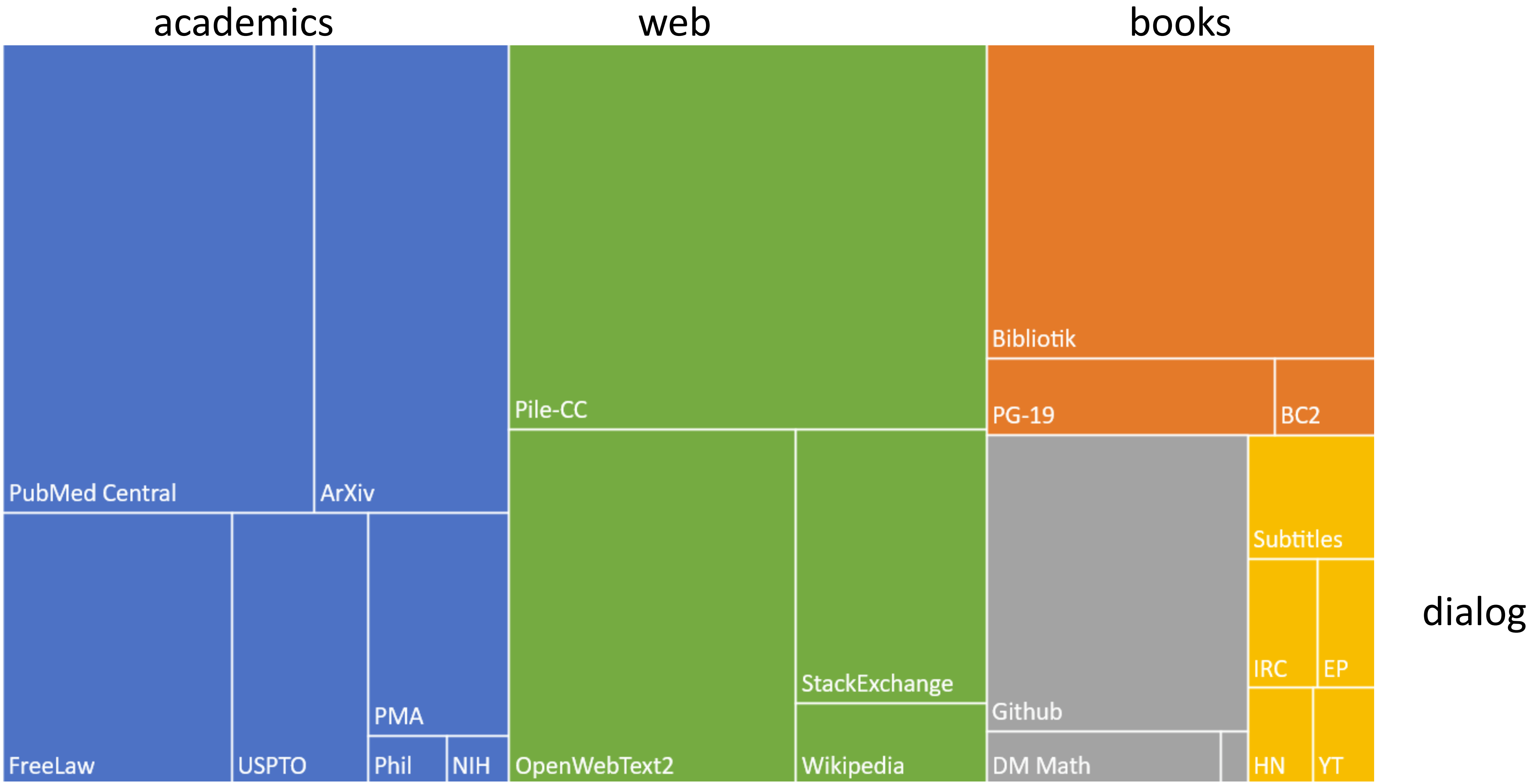
Common crawl, snapshots of the entire web produced by the non- profit Common Crawl with billions of pages

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: SLP3]

The Pile: a pretraining corpus



[Slide: SLP3]

Filtering for quality and safety

Quality is subjective

- Many LLMs attempt to match Wikipedia, books, particular websites
- Need to remove boilerplate, adult content
- Deduplication at many levels (URLs, documents, even lines)

Safety also subjective

- Toxicity detection is important, although that has mixed results
- Can mistakenly flag data written in dialects like African American English

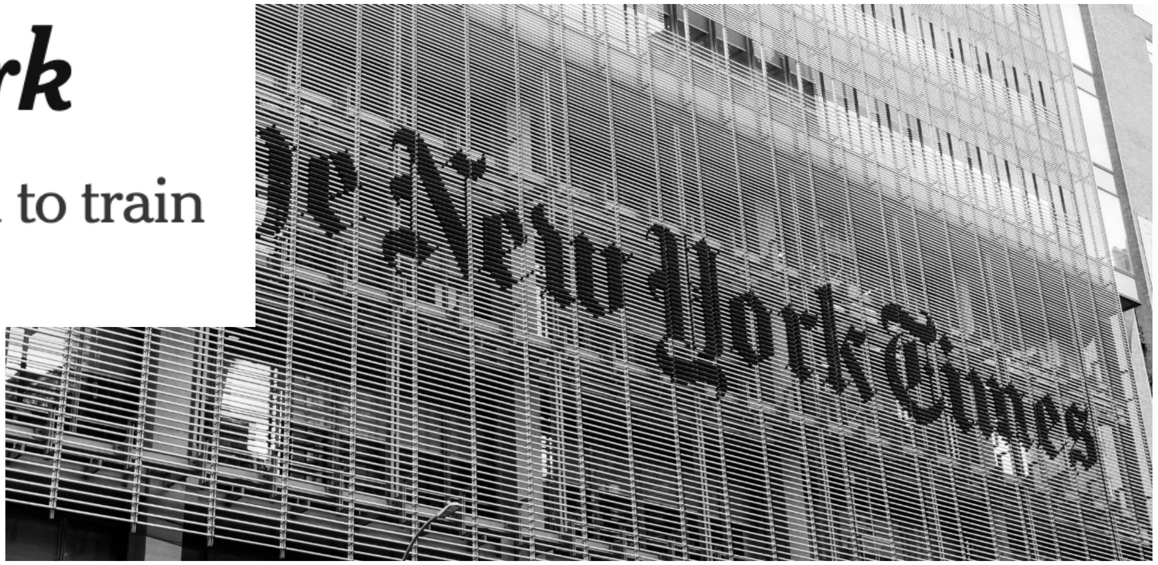
[Slide: SLP3]

There are problems with scraping from the web

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

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.



[Slide: SLP3]

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 across the world

Data consent

- Website owners can indicate they don't want their site crawled

Privacy:

- Websites can contain private IP addresses and phone numbers

Skew:

- Training data is disproportionately generated by authors from the US which probably skews resulting topics and opinions

[Slide: SLP3]