CS224C: NLP for CSS

Deep Learning Highlights for CSS

Diyi Yang
Stanford CS
Lecture Overview

- BERT for Classification
- Prompting LLMs
- Using Prompting in CSS
BERT for Classification

Bidirectional encoder representations from Transformers

**Context is the key**

\[ p(\text{play} | \text{Elmo and Cookie Monster play a game .}) \neq p(\text{play} | \text{The Broadway play premiered yesterday .}) \]
BERT demonstrated strong performances on a wide range of NLP tasks!

**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

Jacob Devlin    Ming-Wei Chang    Kenton Lee    Kristina Toutanova
Google AI Language
{jacobdevlin, mingweichang, kentonl, kristout}@google.com
Masked Language Modeling

Mask out k% of the input words, and then predict the masked words (k=15%)

**Input:** The man went to the [MASK]. He bought a [MASK] of milk.

**Labels:** [MASK] = store; [MASK] = gallon.
Next Sentence Prediction

To learn relationship between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence.

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence
Input Representation

Each token is the sum of three embeddings

<table>
<thead>
<tr>
<th>Input</th>
<th>[CLS]</th>
<th>my</th>
<th>dog</th>
<th>is</th>
<th>cute</th>
<th>[SEP]</th>
<th>he</th>
<th>likes</th>
<th>play</th>
<th># #ing</th>
<th>[SEP]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Token Embeddings</strong></td>
<td>$E_{[CLS]}$</td>
<td>$E_{my}$</td>
<td>$E_{dog}$</td>
<td>$E_{is}$</td>
<td>$E_{cute}$</td>
<td>$E_{[SEP]}$</td>
<td>$E_{he}$</td>
<td>$E_{likes}$</td>
<td>$E_{play}$</td>
<td>$E_{# #ing}$</td>
<td>$E_{[SEP]}$</td>
</tr>
<tr>
<td><strong>Segment Embeddings</strong></td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_B$</td>
<td>$E_B$</td>
<td>$E_B$</td>
<td>$E_B$</td>
<td>$E_B$</td>
<td>$E_B$</td>
</tr>
<tr>
<td><strong>Position Embeddings</strong></td>
<td>$E_0$</td>
<td>$E_1$</td>
<td>$E_2$</td>
<td>$E_3$</td>
<td>$E_4$</td>
<td>$E_5$</td>
<td>$E_6$</td>
<td>$E_7$</td>
<td>$E_8$</td>
<td>$E_9$</td>
<td>$E_{10}$</td>
</tr>
</tbody>
</table>
Model Architecture: Transformer

Multi-headed self attention to model context

Feed-forward layers to compute non-linear hierarchical features

Positional embeddings to allow model to learn relative positioning

Link: https://nlp.stanford.edu/seminar/details/jdevlin.pdf
1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

**Semi-supervised Learning Step**

**Model:** BERT

**Dataset:**

![Books](https://cdn.pixabay.com/photo/2015/12/15/18/26/book-1211599_960_720.png)

![Wikipedia](https://cdn.pixabay.com/photo/2015/10/05/21/28/wikipedia-998214_960_720.png)

**Objective:** Predict the masked word (language modeling)

https://jalammar.github.io/illustrated-bert/

2 - **Supervised** training on a specific task with a labeled dataset.

**Supervised Learning Step**

**Model:** (pre-trained in step #1)

**Dataset:**

<table>
<thead>
<tr>
<th>Email message</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buy these pills</td>
<td>Spam</td>
</tr>
<tr>
<td>Win cash prizes</td>
<td>Spam</td>
</tr>
<tr>
<td>Dear Mr. Atreides, please find attached…</td>
<td>Not Spam</td>
</tr>
</tbody>
</table>

75% Spam  
25% Not Spam
How to use BERT for Classification
(e.g., sentiment, fact-checking, rumors)
Pros and Cons of BERT for CSS

👍 Strong prediction performance

👍 Fine-tuning on top of pertained representations

👎 Prediction and representation can be hard to interpret

👎 Subject to biases in these learned representations

👎 Require computational resources
Lecture Overview

✦ BERT for Classification

✦ Prompting LLMs
Prompt for LLMs

Fine-tuning LLMs (e.g., GPT-3 175B) is often not feasible due to its large size.

Prompts (or **in-context learning**) were then introduced and used.
**Prompting**: encourage a pre-trained model to make particular predictions by providing a "prompt" specifying the task to be done.

Intuition of Prompting

**Sentiment**

The value I got was the sum total of the popcorn and the drink. Overall, it was a **boring** movie!

**World knowledge**

Peking University is located in **Beijing**, China.

**Syntactic categories**

I put **the** fork down on the table.

**Coreference**

The woman walked across the street, checking for traffic over **her** shoulder.

**Semantic categories**

I went to the ocean to see the fish, turtles, seals, and **crabs**.

**Reasoning**

Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the **kitchen**.

Intuition of Prompting

- Sentiment
- World knowledge
- Syntactic categories
- Coreference
- Semantic categories
- Reasoning

Pre-trained models learn **various types of knowledge**.
The knowledge is useful **across NLP tasks**.
These knowledge **can be surfaced** with “templates” (prompt).
Prompting

Zero/few-shot Prompting

1. Translate English to French:
   1. sea otter => loutre de mer
   2. peppermint => menthe poivrée
   3. plush giraffe => girafe peluche
   4. cheese => ...........................................

[Brown et al., 2020]

Traditional Fine-tuning

1. sea otter => loutre de mer
   ↓
   gradient update

1. peppermint => menthe poivrée
   ↓
   gradient update

1. cheese => ...........................................
One model, N tasks (e.g., GPT-3, 175B)

Sentiment analysis classifier

Decide whether a Tweet's sentiment is positive, neutral, or negative.

Tweet: "I loved the new Batman movie!"
Sentiment: Positive

Q&A

Chatbot: I am a ML/AI language model tutor
You: What is a language model?
Chatbot: A language model is a statistical model that describes the probability of a word given the previous words.

Translation / NL2code

Create a SQL request to find all users who live in California and have over 1000 credits:

```sql
SELECT * FROM users WHERE state='CA' AND credits > 1000;
```

Semantic of query Syntax of code

Summarization

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.[1] Neutron stars are the smallest and densest stellar objects, excluding black holes and hypothetical white holes, quark stars, and strange stars.[2] Neutron stars have a radius on the order of 10 kilometres (6.2 mi) and a mass of about 1.4 solar masses.[3] They result from the supernova explosion of a massive star, combined with gravitational collapse, that compresses the core past white dwarf star density to that of atomic nuclei.

TL;DR: A neutron star is the collapsed core of a massive supergiant star. These ultra-dense objects are incredibly fascinating due to their strange properties and their potential for phenomena such as extreme gravitational forces and a strong magnetic field.

The semantic of TL;DR

GPT-3 API, from OpenAI: https://beta.openai.com/examples
Emergent Zero-shot Learning

GPT-2 beats SoTA on language modeling benchmarks with no task-specific fine-tuning

*Context:* “Why?” “I would have thought you’d find him rather dry,” she said. “I don’t know about that,” said Gabriel.

“He was a great craftsman,” said Heather. “That he was,” said Flannery.

*Target sentence:* “And Polish, to boot,” said _____.

*Target word:* Gabriel

<table>
<thead>
<tr>
<th></th>
<th>LAMBADA (PPL)</th>
<th>LAMBADA (ACC)</th>
<th>CBT-CN (ACC)</th>
<th>CBT-NE (ACC)</th>
<th>WikiText2 (PPL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA</td>
<td>99.8</td>
<td>59.23</td>
<td>85.7</td>
<td>82.3</td>
<td>39.14</td>
</tr>
<tr>
<td>117M</td>
<td>35.13</td>
<td>45.99</td>
<td>87.65</td>
<td>83.4</td>
<td>29.41</td>
</tr>
<tr>
<td>345M</td>
<td>15.60</td>
<td>55.48</td>
<td>92.35</td>
<td>87.1</td>
<td>22.76</td>
</tr>
<tr>
<td>762M</td>
<td>10.87</td>
<td>60.12</td>
<td>93.45</td>
<td>88.0</td>
<td>19.93</td>
</tr>
<tr>
<td>1542M</td>
<td>8.63</td>
<td>63.24</td>
<td>93.30</td>
<td>89.05</td>
<td>18.34</td>
</tr>
</tbody>
</table>

*LAMBADA* (language modeling w/ long discourse dependencies) [Paperno et al., 2016]
Emergent few-shot learning

Specify a task by simply prepending examples of the task before your example.

Also called in-context learning, to stress that no gradient updates are performed when learning a new task.
Emergent few-shot learning

Zero-shot

1. Translate English to French:
   cheese =>

[Brown et al., 2020]
Emergent few-shot learning

One-shot

1. Translate English to French:
   sea otter => loutre de mer

2. cheese =>

[Diagram showing In-Context Learning on SuperGLUE with human, fine-tuned SOTA, fine-tuned BERT++, fine-tuned BERT Large, and random guessing performance.]

[Reference: Brown et al., 2020]
Emergent few-shot learning

**Few-shot**

1. Translate English to French:
   - sea otter => loutre de mer
   - peppermint => menthe poivrée
   - plush giraffe => girafe peluche
   - cheese =>

   [Brown et al., 2020]
Limits of Prompting for Harder Tasks

Some tasks seem too hard for even large LMs to learn through prompting alone. Especially tasks involving richer, multi-step reasoning. (Humans struggle at these tasks too!)

\[
\begin{align*}
19583 & + 29534 = 49117 \\
98394 & + 49384 = 147778 \\
29382 & + 12347 = 41729 \\
93847 & + 39299 = ?
\end{align*}
\]

Solution: change the prompt!
Chain-of-thought Prompting

Standard Prompting
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
A: The answer is 27. ❌

Chain-of-Thought Prompting
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✔️

[Wei et al., 2022; also see Nye et al., 2021]
Chain-of-thought prompting is an emergent property of model scale

[Wei et al., 2022; also see Nye et al., 2021]
Chain-of-thought Prompting

**Model Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

**Model Output**

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

Do we even need examples of reasoning? Can we just ask the model to reason through things?

[Wei et al., 2022; also see Nye et al., 2021]
Zero-shot Chain-of-thought Prompting

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let’s think step by step. There are 16 balls in total. Half of the balls are golf balls. That means there are 8 golf balls. Half of the golf balls are blue. That means there are 4 blue golf balls.
Zero-shot COT prompting

<table>
<thead>
<tr>
<th></th>
<th>MultiArith</th>
<th>GSM8K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot</td>
<td>17.7</td>
<td>10.4</td>
</tr>
<tr>
<td>Few-Shot (2 samples)</td>
<td>33.7</td>
<td>15.6</td>
</tr>
<tr>
<td>Few-Shot (8 samples)</td>
<td>33.8</td>
<td>15.6</td>
</tr>
<tr>
<td>Zero-Shot-CoT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Few-Shot-CoT (2 samples)</td>
<td>84.8</td>
<td>41.3</td>
</tr>
<tr>
<td>Few-Shot-CoT (4 samples : First) (*1)</td>
<td>89.2</td>
<td>-</td>
</tr>
<tr>
<td>Few-Shot-CoT (4 samples : Second) (*1)</td>
<td>90.5</td>
<td>-</td>
</tr>
<tr>
<td>Few-Shot-CoT (8 samples)</td>
<td>93.0</td>
<td>48.7</td>
</tr>
</tbody>
</table>

Greatly outperforms zero-shot

Manual CoT still better

[Kojima et al., 2022]
# Zero-shot Chain-of-thought prompting

<table>
<thead>
<tr>
<th></th>
<th>Method</th>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LM-Designed</td>
<td>Let’s work this out in a step by step way to be sure we have the right answer.</td>
<td>82.0</td>
</tr>
<tr>
<td>2</td>
<td>Human-Designed</td>
<td>First, (1) Let’s think step by step.</td>
<td>78.7</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>(2) Let’s think about this logically.</td>
<td>77.3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>(3) Let’s solve this problem by splitting it into steps.</td>
<td>74.5</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>(4) Let’s be realistic and think step by step.</td>
<td>72.2</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>(5) Let’s think like a detective step by step.</td>
<td>70.8</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>(6) Let’s think</td>
<td>70.3</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>(7) Before we dive into the answer,</td>
<td>57.5</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>(8) The answer is after the proof.</td>
<td>55.7</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>(Zero-shot)</td>
<td>45.7</td>
</tr>
</tbody>
</table>

[Zhou et al., 2022; Kojima et al., 2022]
Self-Consistency Further Improves Reasoning!

Figure 1: The self-consistency method contains three steps: (1) prompt a language model using example chains of thought; (2) sample from the language model’s decoder to generate a diverse set of reasoning paths; and (3) choose the most consistent answer using the majority/plurality vote.

Hallucination

What is the world record for crossing the English Channel entirely on foot?
Hallucination

No fact check, e.g., summarizing a non-existent news article.
No explicit reasoning mechanism, leading to stupid mistakes
Easy to be manipulated, when the prompt is contaminated.
Downside of Prompt-based Learning

• **Inefficiency**: The prompt needs to be processed every time the model makes a prediction.

• **Poor performance**: Prompting generally performs worse than fine-tuning [Brown et al., 2020].

• **Sensitivity** to the wording of the prompt [Webson & Pavlick, 2022], order of examples [Zhao et al., 2021; Lu et al., 2022], etc.

• **Lack of clarity** regarding what the model learns from the prompt. Even random labels work [Zhang et al., 2022; Min et al., 2022]
Lecture Overview

✧ BERT for Classification
✧ Prompting LLMs
✧ Using Prompting in CSS
Prompting for CSS

Can Large Language Models Transform Computational Social Science?

Caleb Ziems*
Stanford University

William Held
Georgia Institute of Technology

Omar Shaikh
Stanford University

Jiaao Chen
Georgia Institute of Technology

Zhehao Zhang
Dartmouth College

Diyi Yang**
Stanford University
Are LLMs feasible tools for CSS?

emotion recognition
humor recognition
politeness recognition
empathy classification
issue frames extraction
ideology detection
agent framing
relationship dynamics
event extraction
power relations identification
social role detection
dialect feature identification

Psychology
Political Science
Literature
History
Sociology
Linguistics
Are LLMs feasible tools for CSS?

- emotion recognition
- humor recognition
- politeness recognition
- empathy classification
- issue frames extraction
- ideology detection
- agent framing
- relationship dynamics
- event extraction
- power relations identification
- social role detection
- dialect feature identification

<table>
<thead>
<tr>
<th>CARER</th>
<th>(Saravia et al. 2018)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r/Jokes + Pun of the Day</td>
<td>(Weiler and Sopož 2019)</td>
</tr>
<tr>
<td>Stanford Politeness Corpus</td>
<td>(Danesico-Nicolae-Mişil et al., 2013)</td>
</tr>
<tr>
<td>EPITOME</td>
<td>(Sharma et al., 2020)</td>
</tr>
<tr>
<td>SemEval-2016 Stance Dataset</td>
<td>(Mohammed et al., 2016)</td>
</tr>
<tr>
<td>Ideological Books Corpus</td>
<td>(Grass et al., 2013)</td>
</tr>
<tr>
<td>Article Bias Corpus</td>
<td>(Baly et al. 2010)</td>
</tr>
<tr>
<td>WikiEvents</td>
<td>(Li et al., 2021)</td>
</tr>
<tr>
<td>Hippocorpus</td>
<td>(Say et al. 2020)</td>
</tr>
<tr>
<td>Wikipedia Talk Pages</td>
<td>(Danesicu-Nicolae-Mişil et al., 2012)</td>
</tr>
<tr>
<td>CMU Movie Corpus</td>
<td>(Barnes et al. 2013)</td>
</tr>
<tr>
<td>Indian English Minimal Pairs</td>
<td>(Demosky et al. 2019)</td>
</tr>
</tbody>
</table>

*Green circle: utterance-level  Brown circle: conversation-level  Pink circle: document-level*
Zero-Shot Prompted LLMs

Autoregressive
- GPT-3
  - text-ada
  - text-babbage
  - text-curie
  - text-davinci
    - 001
    - 002
    - 003
    - ChatGPT
    - GPT-4

Encoder-Decoder
- Flan-T5
  - small
  - base
  - large
  - XL
  - XXL
  (instruction-tuned)
    - 80M
    - 250M
    - 780M
    - 3B
    - 11B

Fine-Tuned LLMs

MLM
  (Classification)

RoBERTa-large

Encoder-Decoder (Generation)
- T5

Note: GPT model sizes are estimates
Prompt Engineering

**Best Practice:** multiple choice

White House Ousts Top Climate Change Official
Which of the following describes the above news headline?
Prompt Engineering: multiple choices

Best Practice: multiple choice (Hendrycks et al. 2021)

White House Ousts Top Climate Change Official
Which of the following describes the above news headline?

A: Misinformation
B: Trustworthy
Prompt Engineering: newlines

**Best Practice:** newlines (see Inverse Scaling Prize)

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation
B: Trustworthy
Prompt Engineering: Give instructions

**Best Practice:** Give instructions after the context (Child et al. 2019)

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation

B: Trustworthy

giving instructions or questions after the context
Prompt Engineering: Clarify the expected output

**Best Practice:** multiple choice (Hendrycks et al. 2021)

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation

B: Trustworthy

Constraint: Answer with only the option above that is most accurate and nothing else.
Prompt Engineering: Request Structured Output

**Best Practice:** request structured responses in JSON format (see MiniChain)

`{'Victim': 'BLANK', 'Place': 'BLANK', 'Killer': 'BLANK', 'MedicalIssue': 'Blank'}`

Replace the BLANKs with the extracted information about the event in `<tgr>`. Leave the keys of the JSON unchanged.

**JSON Output:**
Classification Evaluation

Which of the following leanings would a political scientist say that the above article has?
A: Liberal
B: Conservative
C: Neutral

Prompt templates constructed per task
x 500 test examples

Logit bias {“A”, “B”, “C”}

Auto Eval: F1
<table>
<thead>
<tr>
<th>Model Data</th>
<th>Baselines</th>
<th>FLAN-T5</th>
<th>FLAN</th>
<th>text-001</th>
<th>text-002</th>
<th>text-003</th>
<th>Chat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rand</td>
<td>Finetune</td>
<td>Small</td>
<td>Base</td>
<td>Large</td>
<td>XL</td>
<td>XXL</td>
</tr>
<tr>
<td>Dialect</td>
<td>3.3</td>
<td>3.0</td>
<td>0.2</td>
<td>4.5</td>
<td>23.4</td>
<td>24.8</td>
<td>30.3</td>
</tr>
<tr>
<td>Emotion</td>
<td>16.7</td>
<td>71.6</td>
<td>19.8</td>
<td>63.8</td>
<td>69.7</td>
<td>65.7</td>
<td>66.2</td>
</tr>
<tr>
<td>Figurative</td>
<td>25.0</td>
<td>99.2</td>
<td>16.6</td>
<td>23.2</td>
<td>18.0</td>
<td>32.2</td>
<td>53.2</td>
</tr>
<tr>
<td>Humor</td>
<td>49.5</td>
<td>73.1</td>
<td>51.8</td>
<td>37.1</td>
<td>54.9</td>
<td>56.9</td>
<td>29.9</td>
</tr>
<tr>
<td>Ideology</td>
<td>33.3</td>
<td>64.8</td>
<td>18.6</td>
<td>23.7</td>
<td>43.0</td>
<td>47.6</td>
<td>53.1</td>
</tr>
<tr>
<td>Impl. Hate</td>
<td>16.7</td>
<td>62.5</td>
<td>7.4</td>
<td>14.4</td>
<td>7.2</td>
<td>32.3</td>
<td>29.6</td>
</tr>
<tr>
<td>Misinfo</td>
<td>50.0</td>
<td>81.6</td>
<td>33.3</td>
<td>53.2</td>
<td>64.8</td>
<td>68.7</td>
<td>69.6</td>
</tr>
<tr>
<td>Persuasion</td>
<td>14.3</td>
<td>52.0</td>
<td>3.6</td>
<td>10.4</td>
<td>37.5</td>
<td>32.1</td>
<td>45.7</td>
</tr>
<tr>
<td>Sem. Chng.</td>
<td>50.0</td>
<td>62.3</td>
<td>33.5</td>
<td>41.0</td>
<td>56.9</td>
<td>52.0</td>
<td>36.3</td>
</tr>
<tr>
<td>Stance</td>
<td>33.3</td>
<td>36.1</td>
<td>25.2</td>
<td>36.6</td>
<td>42.2</td>
<td>43.2</td>
<td>49.1</td>
</tr>
</tbody>
</table>

### Conversation Level Tasks

<table>
<thead>
<tr>
<th></th>
<th>Baselines</th>
<th>FLAN-T5</th>
<th>FLAN</th>
<th>text-001</th>
<th>text-002</th>
<th>text-003</th>
<th>Chat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rand</td>
<td>Finetune</td>
<td>Small</td>
<td>Base</td>
<td>Large</td>
<td>XL</td>
<td>XXL</td>
</tr>
<tr>
<td>Discourse</td>
<td>14.3</td>
<td>49.6</td>
<td>4.2</td>
<td>21.5</td>
<td>33.6</td>
<td>37.8</td>
<td>50.6</td>
</tr>
<tr>
<td>Empathy</td>
<td>33.3</td>
<td>71.6</td>
<td>16.7</td>
<td>16.7</td>
<td>22.1</td>
<td>21.2</td>
<td>35.9</td>
</tr>
<tr>
<td>Persuasion</td>
<td>50.0</td>
<td>33.3</td>
<td>9.2</td>
<td>11.0</td>
<td>11.3</td>
<td>8.4</td>
<td>41.8</td>
</tr>
<tr>
<td>Politeness</td>
<td>33.3</td>
<td>75.8</td>
<td>22.4</td>
<td>42.4</td>
<td>44.7</td>
<td>57.2</td>
<td>51.9</td>
</tr>
<tr>
<td>Power</td>
<td>49.5</td>
<td>72.7</td>
<td>46.6</td>
<td>48.0</td>
<td>40.8</td>
<td>55.6</td>
<td>52.6</td>
</tr>
<tr>
<td>Toxicity</td>
<td>50.0</td>
<td>64.6</td>
<td>43.8</td>
<td>40.4</td>
<td>42.5</td>
<td>43.4</td>
<td>34.0</td>
</tr>
</tbody>
</table>

### Document Level Tasks

<table>
<thead>
<tr>
<th></th>
<th>Baselines</th>
<th>FLAN-T5</th>
<th>FLAN</th>
<th>text-001</th>
<th>text-002</th>
<th>text-003</th>
<th>Chat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rand</td>
<td>Finetune</td>
<td>Small</td>
<td>Base</td>
<td>Large</td>
<td>XL</td>
<td>XXL</td>
</tr>
<tr>
<td>Event Arg.</td>
<td>22.3</td>
<td>65.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Event Det.</td>
<td>0.4</td>
<td>75.8</td>
<td>9.8</td>
<td>7.0</td>
<td>1.0</td>
<td>10.9</td>
<td>41.8</td>
</tr>
<tr>
<td>Ideology</td>
<td>33.3</td>
<td>85.1</td>
<td>24.0</td>
<td>19.2</td>
<td>28.3</td>
<td>29.0</td>
<td>42.4</td>
</tr>
<tr>
<td>Tropes</td>
<td>36.9</td>
<td>-</td>
<td>1.7</td>
<td>8.4</td>
<td>13.7</td>
<td>14.6</td>
<td>19.0</td>
</tr>
<tr>
<td>Model</td>
<td>Data</td>
<td>Baselines</td>
<td>FLAN-T5</td>
<td>FLAN</td>
<td>text-001</td>
<td>text-002</td>
<td>text-003</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-----------</td>
<td>---------</td>
<td>-------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rand</td>
<td>Finetune</td>
<td>Small</td>
<td>Base</td>
<td>Large</td>
<td>XL</td>
</tr>
<tr>
<td>Dialect</td>
<td></td>
<td>3.3</td>
<td>3.0</td>
<td>0.2</td>
<td>4.5</td>
<td>23.4</td>
<td>24.8</td>
</tr>
<tr>
<td>Emotion</td>
<td></td>
<td>16.7</td>
<td>71.6</td>
<td>19.8</td>
<td>63.8</td>
<td>69.7</td>
<td>65.7</td>
</tr>
<tr>
<td>Figurative</td>
<td></td>
<td>25.0</td>
<td>99.2</td>
<td>16.6</td>
<td>23.2</td>
<td>18.0</td>
<td>32.2</td>
</tr>
<tr>
<td>Humor</td>
<td></td>
<td>49.5</td>
<td>73.1</td>
<td>51.8</td>
<td>37.1</td>
<td>54.9</td>
<td>56.9</td>
</tr>
<tr>
<td>Ideology</td>
<td></td>
<td>33.3</td>
<td>64.8</td>
<td>18.6</td>
<td>23.7</td>
<td>43.0</td>
<td>47.6</td>
</tr>
<tr>
<td>Impl. Hate</td>
<td></td>
<td>16.7</td>
<td>62.5</td>
<td>7.4</td>
<td>14.4</td>
<td>7.2</td>
<td>32.3</td>
</tr>
<tr>
<td>Misinfo</td>
<td></td>
<td>50.0</td>
<td>81.6</td>
<td>33.3</td>
<td>53.2</td>
<td>64.8</td>
<td>68.7</td>
</tr>
<tr>
<td>Persuasion</td>
<td></td>
<td>14.3</td>
<td>52.0</td>
<td>3.6</td>
<td>10.4</td>
<td>37.5</td>
<td>32.1</td>
</tr>
<tr>
<td>Sem. Chng.</td>
<td></td>
<td>50.0</td>
<td>62.3</td>
<td>33.5</td>
<td>41.0</td>
<td>56.9</td>
<td>52.0</td>
</tr>
<tr>
<td>Stance</td>
<td></td>
<td>33.3</td>
<td>36.1</td>
<td>25.2</td>
<td>36.6</td>
<td>42.2</td>
<td>43.2</td>
</tr>
</tbody>
</table>

**Utterance Level Tasks**

**Conversation Level Tasks**

**Document Level Tasks**
How does model size affect CSS tasks?

The diagram shows a scatter plot with model size on the x-axis and the mean CSS Task F1 Score on the y-axis. Different models are represented by different markers and colors. The models include flan-t5-small, flan-t5-base, flan-t5-large, flan-t5-xl, flan-t5-xxl, flan-UL2, davinci-001, davinci-002, davinci-003, gpt3.5, and gpt-4. The graph indicates a positive correlation between model size and F1 Score.
Are LLMs better adapted for some subfields?

Performance is not tied to academic discipline but rather by the complexity of the input.
Lecture Overview

♦ BERT for Classification
♦ Prompting LLMs
♦ Using Prompting in CSS