LLM Prompting

Johns Hopkins
Whiting School of Engineering

Syllabus

- Today 4/14: LLM Prompting
- Wednesday 4/16: Guest Lecture from Ziang Xiao
- Monday 4/21: Analysis of user-LLM interactions (Miriam)
- Wednesday 4/32: Social Simulations
- Monday 4/28: No class

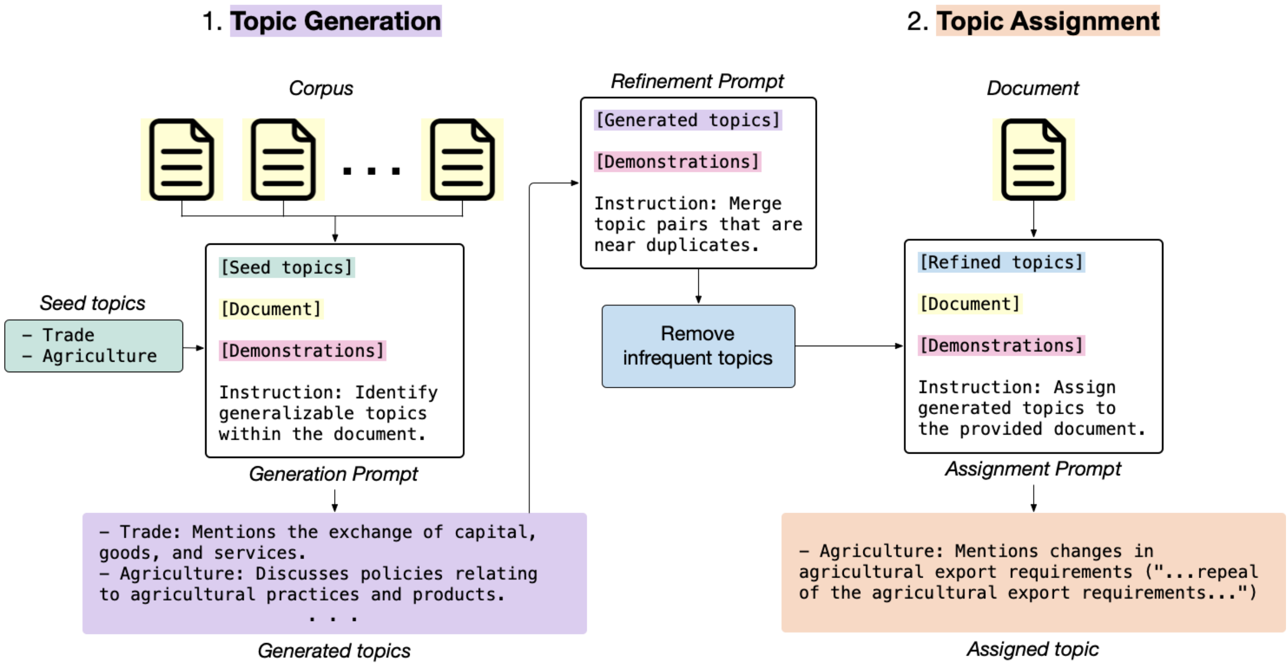
Recap

- (L)LM use cases in NLP for social science:
  - BERT-style models are effective classifiers
  - Metaphorical language
  - Neural topic models (ProdLDA, BERTopic, TopicGPT)

- This class:
  - LLMs as classifiers and data labelers

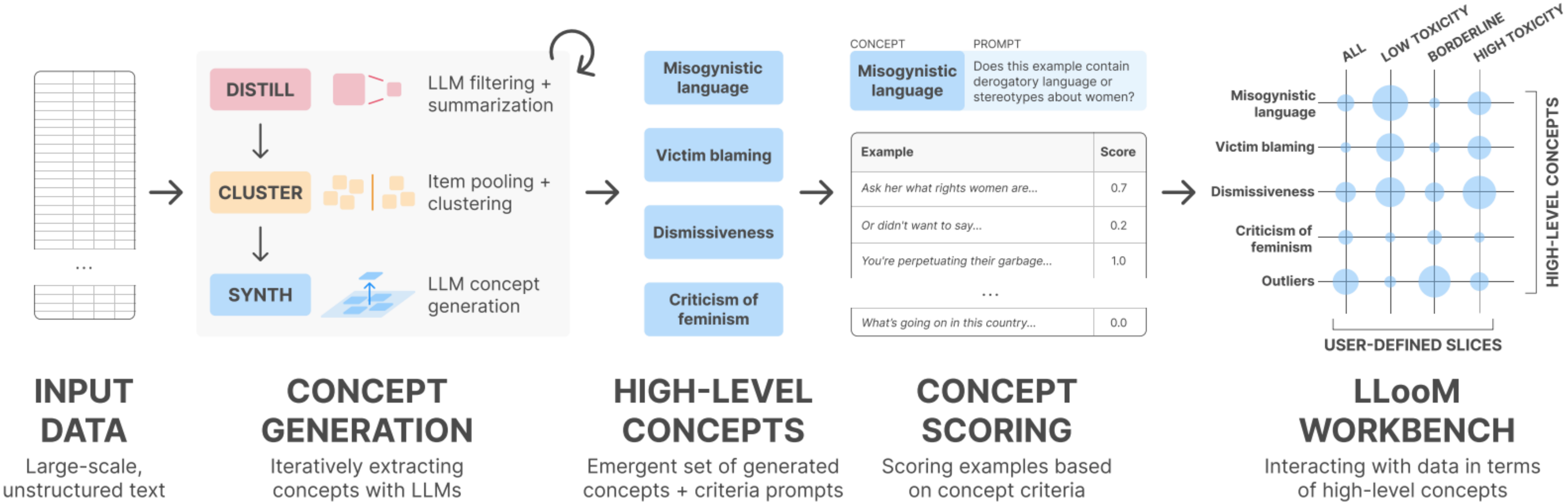
- Next class:
  - Social simulations: using LLMs to simulate people

TopicGPT



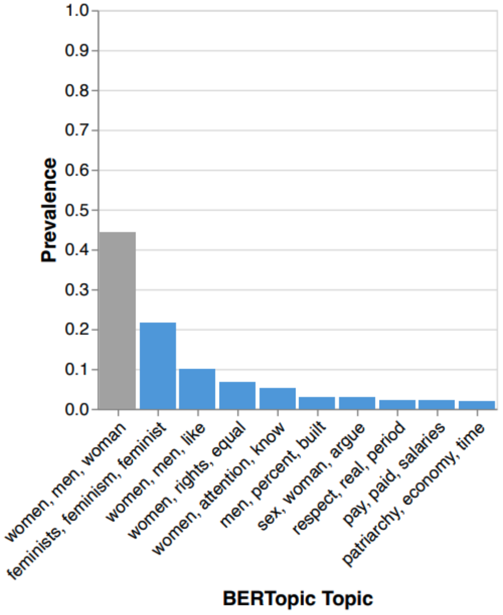
Pham, Chau Minh, et al. "TopicGPT: A prompt-based topic modeling framework." NAACL (2024).
https://arxiv.org/abs/2311.01449

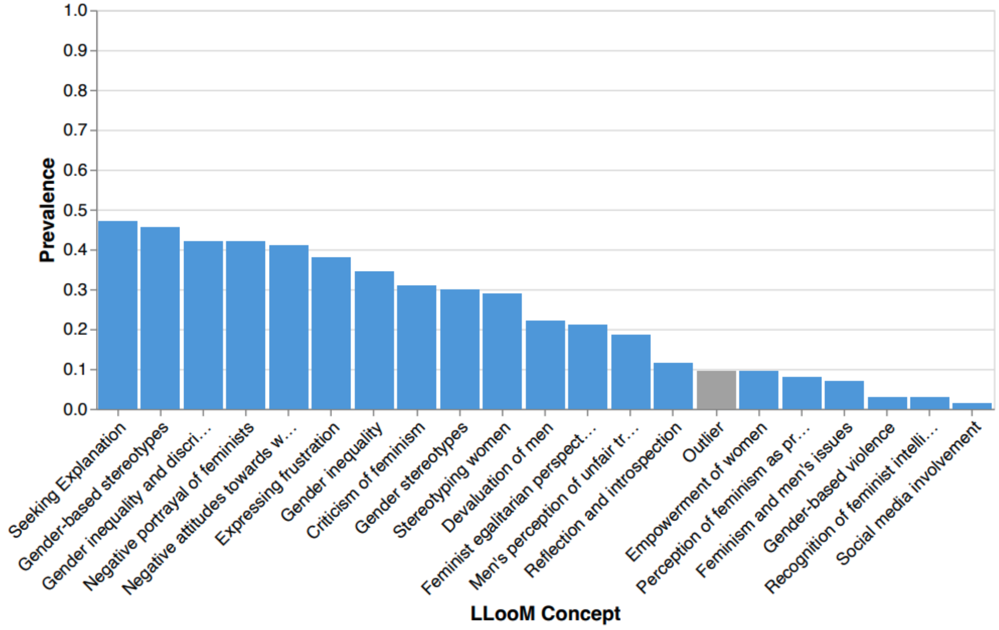
A different approach: LLooM



Lam, Michelle S., et al. "Concept induction: Analyzing unstructured text with high-level concepts using lloom." *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024.

Example evaluation





Recap

- (L)LM use cases in NLP for social science:
  - BERT-style models are effective classifiers
  - Metaphorical language
  - Neural topic models (ProdLDA, BERTopic, TopicGPT)

- This class:
  - LLMs as classifiers and data labelers

- Next class:
  - Social simulations: using LLMs to simulate people

General NLP methods for CSS

- Unsupervised approaches
  - Word statistics, topic modeling
- Semi-supervised approaches
  - Word embeddings, lexicons
- Supervised approaches:
  - Data annotating, classification models, interpreting model outputs

- Typically supervised approaches yield best results (e.g. ability to measure the values we care about) but data annotation is difficult and costly

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Zero-shot approaches

Zero-shot

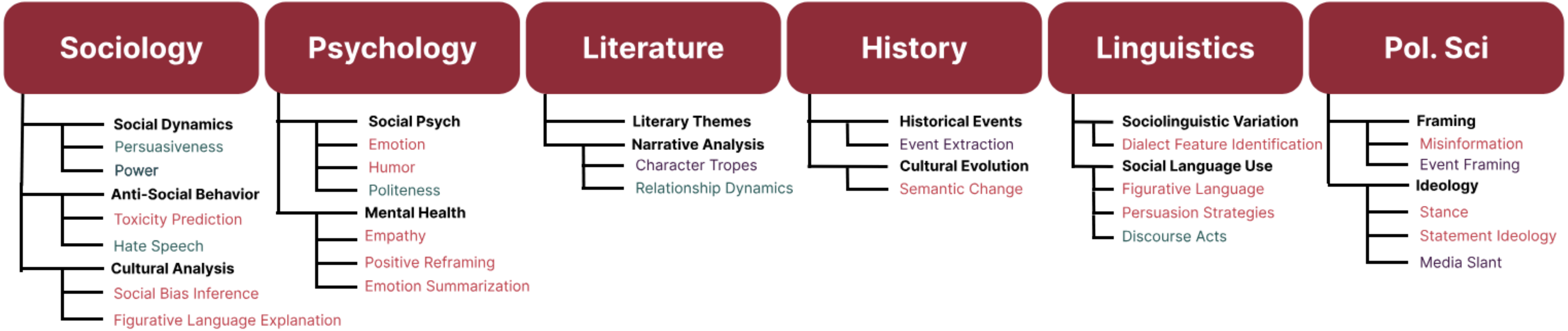
- Take model that has been instructed-tuned (with or without RLHF) and prompt it to label data

Can Large Language Models Transform Computational Social Science?

- Goal:
  - Evaluate zero-shot performance of LLMs over a diverse range of CSS text processing tasks
- Categories of tasks:
  - Utterance level
  - Conversation level
  - Document level
- [Primarily classification tasks, also information extraction and some text generation]

Ziems et al. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics, 2024*

Core subject areas in CSS (and digital humanities)



Ziems et al. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics, 2024*

Models Evaluated

- FLAN-T5 (Chung et al. 2022)
  - Encoder-decoder architecture
  - Instruction Tuned
  - [Open source model with strong zero/few shot performance]
- GPT-3
  - Some variants with only instruction-tuning and some +RLHF
- GPT-4
  - Multimodal model, substantially scales up GPT-3 architecture
- Baselines:
  - Supervised models trained for the specific task (RoBERTa for classification and T5 for generation)

Challenge: Models are sensitive to exact prompt

ChatGPT 3.5

You
What is the best topic model?

ChatGPT
Determining the "best" topic model depends on various factors such as the specific requirements of your task, the characteristics of your data, computational resources available, and personal preferences. However, some popular and effective topic modeling algorithms include Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF).

- Latent Dirichlet Allocation (LDA): LDA is a generative probabilistic model that assumes documents are a mixture of topics, and each topic is a mixture of words. It's widely used and relatively easy to implement. However, it may not capture correlations between words as well as other models.
- Latent Semantic Analysis (LSA): LSA is a technique based on Singular Value Decomposition (SVD) that analyzes relationships between terms and documents by producing a set of concepts related to the terms and documents. LSA can handle synonymy and polysemy to some extent but may not capture more complex semantic relationships.
- Non-Negative Matrix Factorization (NMF): NMF decomposes a matrix into the product of two lower-dimensional matrices, where all elements are non-negative. It's often used for dimensionality reduction and feature extraction. NMF has been shown to perform well for topic modeling tasks, particularly when interpretability of topics is important.

Each of these models has its strengths and weaknesses, and the choice depends on the specific needs of your project. It's often helpful to experiment with different models and evaluate their performance based on metrics such as coherence scores, topic interpretability, and relevance to your application.

Message ChatGPT...

ChatGPT can make mistakes. Consider checking important information.

You
Which topic model is the best for my data?

ChatGPT
Determining the best topic model for your data depends on various factors such as the nature of your data, your specific objectives, the size of your dataset, and the computational resources available to you. Here are a few popular topic modeling algorithms and when they might be suitable:

1. **Latent Dirichlet Allocation (LDA)**:
   - LDA is one of the most commonly used topic modeling algorithms.
   - It assumes that documents are represented as a mixture of topics and that each word's presence is attributable to one of the document's topics.
   - Suitable for large corpora and when interpretability of topics is important.
2. **Latent Semantic Analysis (LSA)**:
   - LSA is a classical method for topic modeling that utilizes singular value decomposition (SVD).
   - It works well with datasets where the relationships between terms and documents can be captured using linear algebra techniques.
   - Suitable when you have a smaller dataset and want a simple approach.
3. **Non-Negative Matrix Factorization (NMF)**:
   - NMF factorizes a non-negative matrix into two non-negative matrices.
   - It's particularly useful when interpretability of topics is important and when your dataset is relatively small.
   - NMF often leads to more easily interpretable topics compared to LDA.
4. **Word Embedding-Based Models**:
   - Models like Word2Vec, Doc2Vec, and GloVe can be used for topic modeling by clustering word vectors or document vectors.
   - They are effective when your data is primarily composed of short texts like tweets or product reviews.
   - They capture semantic relationships between words, which can lead to more meaningful topic representations.
5. **BERTopic**:
   - BERTopic is a recent approach that utilizes BERT embeddings and hierarchical clustering to identify topics.
   - It's effective for both short and long texts and can capture more nuanced relationships between words.
   - Suitable when you have access to large computational resources and want state-of-the-art performance.

Before choosing a topic model, it's essential to preprocess your data appropriately, including tasks such as text cleaning, tokenization, and possibly dimensionality reduction. Additionally, consider evaluating different models using metrics such as coherence score, topic diversity, and human judgment to select the most suitable one for your specific needs.

Challenge: Models are sensitive to exact prompt

- Write initial prompt

Effective Prompt Guideline	Reference	Guideline Example
When the answer is categorical, enumerate options as alphabetical **multiple-choice** so that the output is simply the highest-probability token ('A', 'B').	Hendrycks et al. (2021)	{$CONTEXT} Which of the following describes the above news headline? ↩ **A:** Misinformation ↩ **B:** Trustworthy ↩ {$CONSTRAINT}
Each option should be separated by a new line (↩) to resemble the natural format of online multiple choice questions. More natural prompts will elicit more regular behavior.	Inverse Scaling Prize	
To promote instruction-following, **give instructions *after* the context** is provided; then **explicitly state any constraints**. Recent and repeated text has a greater effect on LLM generations due to common attention patterns.	Child et al. (2019)	{$CONTEXT} **{$QUESTION}** **Constraint:** Even if you are uncertain, you **must pick either "True" or "False"** without using any other words.
Clarify the expected output in the case of uncertainty. Uncertain models may use default phrases like *"I don't know,"* and clarifying constraints force the model to answer.	No Existing Reference	
When the answer should contain multiple pieces of information, **`request responses in JSON format`**. This leverages LLM's familiarity with code to provide an output structure that is more easily parsed.	MiniChain Library	{$CONTEXT} {$QUESTION} **`JSON Output:`**

Challenge: Models are sensitive to exact prompt

- Write initial prompt
- Use GPT-3.5 to paraphrase initial prompt 4 times
- Report results averaged across prompt perturbations

Utterance-level

Data \ Model	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
Utterance Level Tasks																
Dialect	3.3	3.0	0.2	4.5	23.4	24.8	30.3	32.9	0.5	0.5	1.2	9.1	17.1	14.7	11.7	23.2
Emotion	16.7	71.6	19.8	63.8	69.7	65.7	66.2	70.8	6.4	4.9	6.6	19.7	36.8	44.0	47.1	50.6
Figurative	25.0	99.2	16.6	23.2	18.0	32.2	53.2	62.3	10.0	15.2	10.0	19.4	45.6	57.8	48.6	17.5
Humor	49.5	73.1	51.8	37.1	54.9	56.9	29.9	56.8	38.7	33.3	34.7	29.2	29.7	33.0	43.3	61.3
Ideology	33.3	64.8	18.6	23.7	43.0	47.6	53.1	46.4	39.7	25.1	25.2	23.1	46.0	46.8	43.1	60.0
Impl. Hate	16.7	62.5	7.4	14.4	7.2	32.3	29.6	32.0	7.1	7.8	4.9	9.2	18.4	19.2	16.3	3.7
Misinfo	50.0	81.6	33.3	53.2	64.8	68.7	69.6	77.4	45.8	36.2	41.5	42.3	70.2	73.7	55.0	26.9
Persuasion	14.3	52.0	3.6	10.4	37.5	32.1	45.7	43.5	3.6	5.3	4.7	11.3	21.6	17.5	23.3	56.4
Sem. Chng.	50.0	62.3	33.5	41.0	56.9	52.0	36.3	41.6	32.8	38.9	41.3	35.7	41.9	37.4	44.2	21.2
Stance	33.3	36.1	25.2	36.6	42.2	43.2	49.1	48.1	18.1	17.7	17.2	35.6	46.4	41.3	48.0	76.0

Most of the time supervised is better

Suspiciously high LLM performance

Was this data in GPT-4's training data?

Johns Hopkins Whiting School of Engineering

Conversation-level

Model / Data	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
Discourse	14.3	49.6	4.2	21.5	33.6	37.8	50.6	39.6	6.6	9.6	4.3	11.4	35.1	36.4	35.4	16.7
Empathy	33.3	71.6	16.7	16.7	22.1	21.2	35.9	34.7	24.5	17.6	27.6	16.8	16.9	17.4	22.6	6.4
Persuasion	50.0	33.3	9.2	11.0	11.3	8.4	41.8	43.1	6.9	6.7	6.7	33.3	33.3	53.9	51.7	28.6
Politeness	33.3	75.8	22.4	42.4	44.7	57.2	51.9	53.4	16.7	17.1	33.9	22.1	33.1	39.4	51.1	59.7
Power	49.5	72.7	46.6	48.0	40.8	55.6	52.6	56.9	43.1	39.8	37.5	36.9	39.2	51.9	56.5	42.0
Toxicity	50.0	64.6	43.8	40.4	42.5	43.4	34.0	48.2	41.4	34.2	33.4	34.8	41.8	46.9	31.2	55.4

Most of the time supervised is much better

Conversation-level

Data \ Model	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
Discourse	14.3	49.6	4.2	21.5	33.6	37.8	50.6	39.6	6.6	9.6	4.3	11.4	35.1	36.4	35.4	16.7
Empathy	33.3	71.6	16.7	16.7	22.1	21.2	35.9	34.7	24.5	17.6	27.6	16.8	16.9	17.4	22.6	6.4
Persuasion	50.0	33.3	9.2	11.0	11.3	8.4	41.8	43.1	6.9	6.7	6.7	33.3	33.3	53.9	51.7	28.6
Politeness	33.3	75.8	22.4	42.4	44.7	57.2	51.9	53.4	16.7	17.1	33.9	22.1	33.1	39.4	51.1	59.7
Power	49.5	72.7	46.6	48.0	40.8	55.6	52.6	56.9	43.1	39.8	37.5	36.9	39.2	51.9	56.5	42.0
Toxicity	50.0	64.6	43.8	40.4	42.5	43.4	34.0	48.2	41.4	34.2	33.4	34.8	41.8	46.9	31.2	55.4

Best LLM is not better than random (also true for some of the utterance-level and document-level tasks)

Document-level

Model / Data	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
Event Arg.	22.3	65.1	–	–	–	–	–	–	–	–	8.6	8.6	21.6	22.9	22.3	23.0
Event Det.	0.4	75.8	9.8	7.0	1.0	10.9	41.8	50.6	29.8	47.3	47.4	44.4	48.8	52.4	51.3	14.8
Ideology	33.3	85.1	24.0	19.2	28.3	29.0	42.4	38.8	22.1	26.8	18.9	21.5	42.8	43.4	44.7	51.5
Tropes	36.9	-	1.7	8.4	13.7	14.6	19.0	28.6	7.7	12.8	16.7	15.2	16.3	26.6	36.9	44.9

Most of the time supervised is much much better

What about *agreement* instead of accuracy?

Dataset	Best Model	F1	κ	Agreement
Utterance-Level				
Dialect	flan-ul2	32.9	0.15	poor
Emotion	flan-ul2	**70.8**	0.65	good
Figurative	flan-ul2	62.3	0.52	moderate
Humor	gpt-4	61.3	0.23	fair
Ideology	davinci-002	60.0	0.40	moderate
Impl. Hate	flan-ul2	32.3	0.20	fair
Misinfo	flan-ul2	**77.4**	0.55	moderate
Persuasion	gpt-4	56.4	0.51	moderate
Semantic Chng.	flan-t5-large	56.9	0.14	poor
Stance	gpt-3.5-turbo	**72.0**	0.58	moderate

Dataset	Best Model	F1	κ	Agreement
Convo-Level				
Discourse	flan-t5-xxl	50.6	0.45	moderate
Empathy	flan-t5-xxl	35.9	0.04	poor
Persuasion	davinci-003	53.9	0.14	poor
Politeness	flan-t5-xl	59.2	0.38	fair
Power	gpt-4	59.7	0.26	fair
Toxicity	gpt-4	55.4	0.11	poor
Document-Level				
Ideology	gpt-4	51.5	0.51	moderate
Event Det.	gpt-4	23.0	n/a	-
Tropes	gpt-4	44.9	n/a	-

What evaluation would you want to see to understand if using an LLM to augment human annotations is useful?

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

What about *agreement* instead of accuracy?

Dataset	Best Model	F1	κ	Agreement
Utterance-Level				
Dialect	flan-ul2	32.9	0.15	poor
Emotion	flan-ul2	**70.8**	0.65	good
Figurative	flan-ul2	62.3	0.52	moderate
Humor	gpt-4	61.3	0.23	fair
Ideology	davinci-002	60.0	0.40	moderate
Impl. Hate	flan-ul2	32.3	0.20	fair
Misinfo	flan-ul2	**77.4**	0.55	moderate
Persuasion	gpt-4	56.4	0.51	moderate
Semantic Chng.	flan-t5-large	56.9	0.14	poor
Stance	gpt-3.5-turbo	**72.0**	0.58	moderate

Dataset	Best Model	F1	κ	Agreement
Convo-Level				
Discourse	flan-t5-xxl	50.6	0.45	moderate
Empathy	flan-t5-xxl	35.9	0.04	poor
Persuasion	davinci-003	53.9	0.14	poor
Politeness	flan-t5-xl	59.2	0.38	fair
Power	gpt-4	59.7	0.26	fair
Toxicity	gpt-4	55.4	0.11	poor
Document-Level				
Ideology	gpt-4	51.5	0.51	moderate
Event Det.	gpt-4	23.0	n/a	-
Tropes	gpt-4	44.9	n/a	-

Latent Hatred (ElSherief et al. 2021), "requires models to infer a subtle social taxonomy from the coded or indirect speech of U.S. hate groups"
Example: "jewish harvard profeessor noel ignatiev wants to abolish the white race."
contains *white grievance*

What about *agreement* instead of accuracy?

Dataset	Best Model	F1	κ	Agreement
Utterance-Level				
Dialect	flan-ul2	32.9	0.15	poor
Emotion	flan-ul2	**70.8**	0.65	good
Figurative	flan-ul2	62.3	0.52	moderate
Humor	gpt-4	61.3	0.23	fair
Ideology	davinci-002	60.0	0.40	moderate
Impl. Hate	flan-ul2	32.3	0.20	fair
Misinfo	flan-ul2	**77.4**	0.55	moderate
Persuasion	gpt-4	56.4	0.51	moderate
Semantic Chng.	flan-t5-large	56.9	0.14	poor
Stance	gpt-3.5-turbo	**72.0**	0.58	moderate

Dataset	Best Model	F1	κ	Agreement
Convo-Level				
Discourse	flan-t5-xxl	50.6	0.45	moderate
Empathy	flan-t5-xxl	35.9	0.04	poor
Persuasion	davinci-003	53.9	0.14	poor
Politeness	flan-t5-xl	59.2	0.38	fair
Power	gpt-4	59.7	0.26	fair
Toxicity	gpt-4	55.4	0.11	poor
Document-Level				
Ideology	gpt-4	51.5	0.51	moderate
Event Det.	gpt-4	23.0	n/a	-
Tropes	gpt-4	44.9	n/a	-

Bad accuracy and agreement on subtle tasks that require nuanced social context (Models are oversensitive to “stereotype” class and label anything with an identity term as a stereotype)

"Concretely, our analysis reveals that, except in minority cases, prompted LLMs do not match or exceed the performance of carefully fine-tuned classifiers, and the best LLM performances are often too low to entirely replace human annotation."

[More nuanced take – depends on the task, but we have to question if we can trust evaluation]

Ziems et al. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics, 2024*

Few-shot approaches (In context learning)

Large Language Models are few-shot learners

- A large labelled data set can be difficult to build, but annotating a smaller set is often feasible, how can we use this?

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

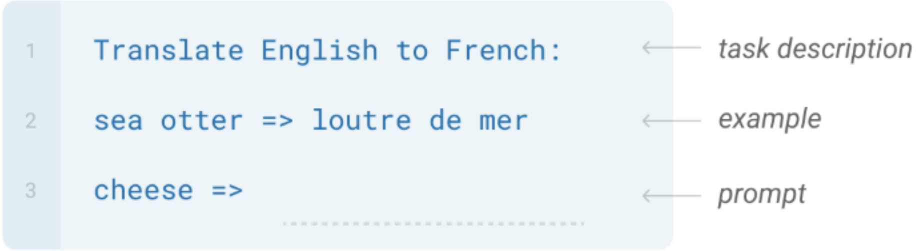
Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

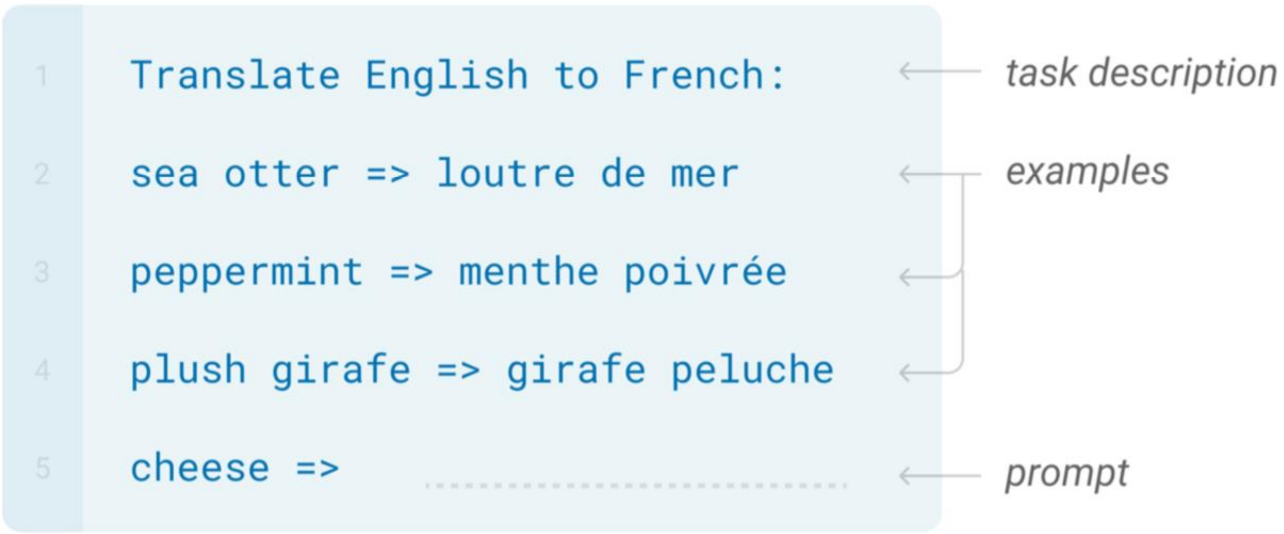
Key idea: Give models a few examples during inference

"Zeroshot"



"One-shot"

Key idea: Give models a few examples during inference



Few-shot “In-context learning”

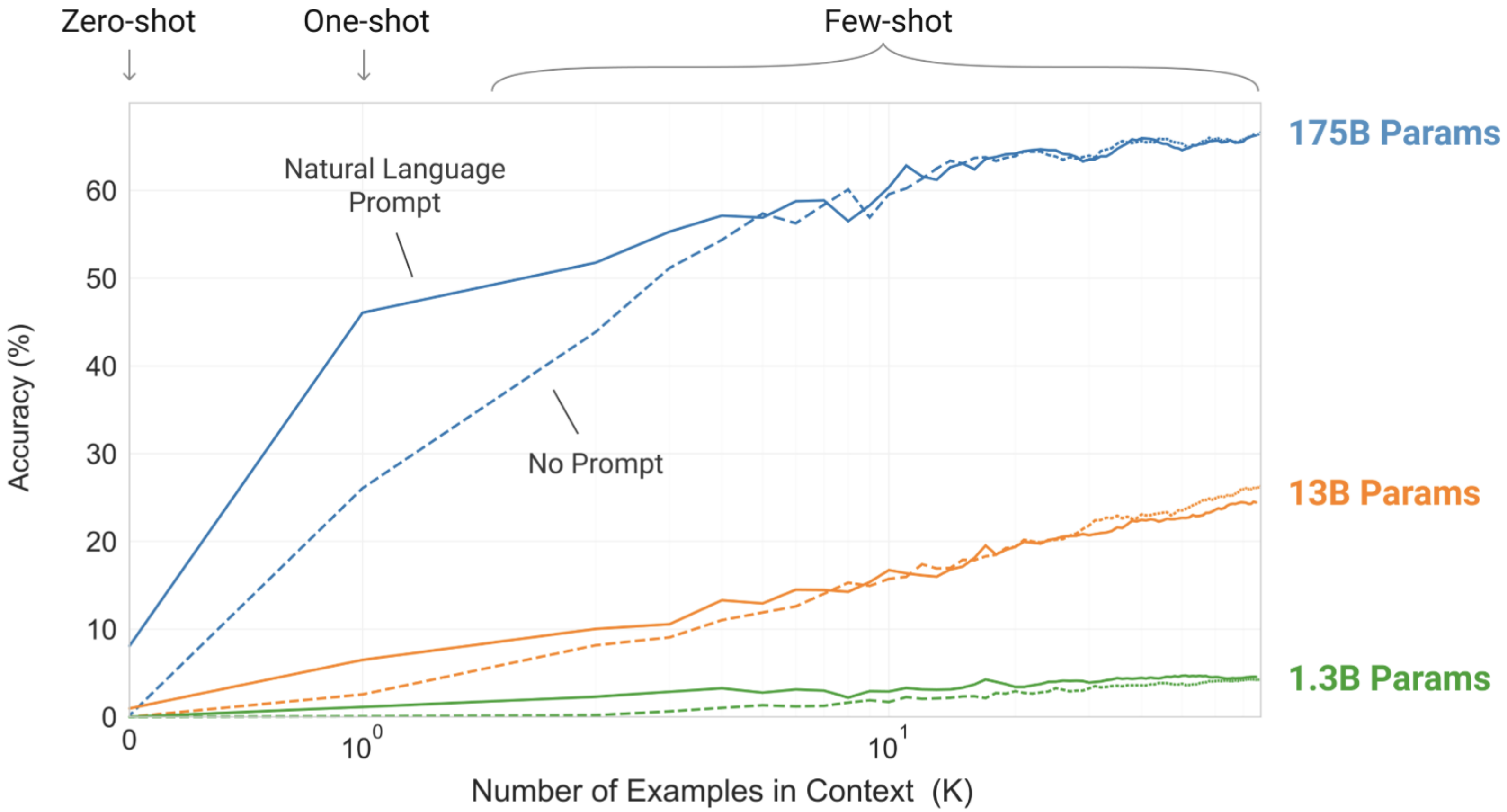
The model parameters are *not* changed (*no* gradient updates)

Evaluation

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	**90.1**[a]	**84.6**[b]
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7

- Generally improves performance over zero-shot, but it varies by task and lags behind supervised models


Zero-shot
One-shot
Few-shot
Natural Language Prompt
No Prompt
Accuracy (%)
60
50
40
30
20
10
0
0
10^0
10^1
Number of Examples in Context (K)
175B Params
13B Params
1.3B Params

Model	FLAN Small			FLAN Base			FLAN Large			FLAN XL			FLAN XXL			FLAN UL2		
Shot	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5
Dialect	0.2	0.0	**0.4**	**4.5**	0.0	1.4	**23.4**	0.7	14.1	**24.8**	8.0	20.5	**30.3**	0.2	29.9	**32.9**	12.6	27.5
Emotion	**19.8**	10.6	10.1	**63.8**	42.7	42.0	**69.7**	67.6	67.4	**65.7**	62.1	62.5	**66.2**	61.8	57.4	**70.8**	70.0	69.8
Figurative	**16.6**	10.0	9.2	23.2	**29.1**	27.3	18.0	**21.8**	19.6	**32.2**	27.9	28.5	53.2	52.6	**66.2**	**62.3**	52.7	62.0
Humor	51.8	52.8	**53.1**	**37.1**	35.1	34.7	**54.9**	54.0	53.8	56.9	**57.0**	56.7	29.9	34.8	**35.3**	**56.8**	55.5	54.1
Ideology	18.6	16.7	**24.0**	**23.7**	22.6	38.3	43.0	**47.3**	45.5	47.6	**48.8**	50.4	53.1	52.9	**57.7**	46.4	36.9	**51.5**
Impl. Hate	**7.4**	6.8	6.2	14.4	**21.1**	7.4	7.2	**9.3**	4.7	32.3	28.5	**34.6**	29.6	31.6	**35.1**	**32.0**	29.5	25.9
Misinfo	**33.3**	33.3	33.3	53.2	45.3	**59.7**	**64.8**	64.8	64.2	68.7	67.2	**69.7**	69.6	**74.9**	74.4	**77.4**	53.7	76.4
Persuasion	**3.6**	3.6	3.6	10.4	**10.8**	7.3	37.5	**39.0**	37.7	32.1	**44.3**	41.8	45.7	44.6	**48.6**	**43.5**	42.2	40.1
Sem. Chng.	33.5	33.3	**34.0**	41.0	35.7	**41.7**	56.9	48.8	**60.4**	**52.0**	40.8	35.6	**36.3**	34.0	33.3	41.6	**62.5**	34.6
Stance	25.2	16.7	**29.6**	**36.6**	18.1	36.6	**42.2**	41.8	39.8	43.2	**52.1**	46.2	**49.1**	46.0	48.7	48.1	**55.6**	54.7
Discourse	4.2	4.0	**7.5**	**21.5**	18.1	20.7	33.6	3.6	**34.6**	37.8	3.6	**38.0**	**50.6**	3.6	43.4	**39.6**	3.6	39.1
Empathy	**16.7**	16.7	16.7	**16.7**	16.7	16.7	**22.1**	16.7	17.1	21.2	**30.4**	22.8	**35.9**	29.8	28.2	34.7	**41.5**	39.6
Persuasion	9.2	**55.9**	45.0	11.0	**55.0**	48.7	11.3	**54.6**	51.7	8.4	42.8	**43.8**	**41.8**	38.8	35.2	43.1	**44.9**	46.1
Politeness	**22.4**	16.7	20.1	**42.4**	23.9	35.4	44.7	44.5	**51.9**	**57.2**	27.7	50.4	**51.9**	44.2	50.3	53.4	43.6	**53.9**
Power	**46.6**	44.5	33.3	**48.0**	39.8	41.4	40.8	**45.5**	43.5	55.6	58.9	**60.2**	52.6	52.0	**62.6**	56.9	57.2	**57.5**
Toxicity	43.8	**46.7**	33.3	40.4	34.7	**54.4**	**42.5**	34.7	36.7	43.4	38.7	**49.2**	34.0	33.3	**35.1**	48.2	44.7	**52.5**
Ideology	**24.0**	16.7	19.2	19.2	16.6	**21.3**	**28.3**	17.0	17.9	29.0	**31.7**	27.0	42.4	**48.5**	47.9	38.8	**38.9**	39.7
Tropes	1.7	**5.1**	3.4	**8.4**	5.1	3.4	**13.7**	10.0	11.6	**14.6**	8.4	10.0	**19.0**	8.4	6.8	**28.6**	27.3	24.6

Ziems et al. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics, 2024*

What about CSS tasks?

- Improvements are inconsistent – often zero-shot is still better

Model	FLAN Small			FLAN Base			FLAN Large			FLAN XL			FLAN XXL			FLAN UL2		
Shot	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5
Dialect	0.2	0.0	**0.4**	**4.5**	0.0	1.4	**23.4**	0.7	14.1	**24.8**	8.0	20.5	**30.3**	0.2	29.9	**32.9**	12.6	27.5
Emotion	**19.8**	10.6	10.1	**63.8**	42.7	42.0	**69.7**	67.6	67.4	**65.7**	62.1	62.5	**66.2**	61.8	57.4	**70.8**	70.0	69.8
Figurative	**16.6**	10.0	9.2	23.2	**29.1**	27.3	18.0	**21.8**	19.6	**32.2**	27.9	28.5	53.2	52.6	**66.2**	**62.3**	52.7	62.0
Humor	51.8	52.8	**53.1**	**37.1**	35.1	34.7	**54.9**	54.0	53.8	56.9	**57.0**	56.7	29.9	34.8	**35.3**	**56.8**	55.5	54.1
Ideology	18.6	16.7	**24.0**	**23.7**	22.6	38.3	43.0	**47.3**	45.5	47.6	**48.8**	50.4	53.1	52.9	**57.7**	46.4	36.9	**51.5**
Impl. Hate	**7.4**	6.8	6.2	14.4	**21.1**	7.4	7.2	**9.3**	4.7	32.3	28.5	**34.6**	29.6	31.6	**35.1**	**32.0**	29.5	25.9
Misinfo	**33.3**	33.3	33.3	53.2	45.3	**59.7**	**64.8**	64.8	64.2	68.7	67.2	**69.7**	69.6	**74.9**	74.4	**77.4**	53.7	76.4
Persuasion	**3.6**	3.6	3.6	10.4	**10.8**	7.3	37.5	**39.0**	37.7	32.1	**44.3**	41.8	45.7	44.6	**48.6**	**43.5**	42.2	40.1
Sem. Chng.	33.5	33.3	**34.0**	41.0	35.7	**41.7**	56.9	48.8	**60.4**	**52.0**	40.8	35.6	**36.3**	34.0	33.3	41.6	**62.5**	34.6
Stance	25.2	16.7	**29.6**	**36.6**	18.1	36.6	**42.2**	41.8	39.8	43.2	**52.1**	46.2	**49.1**	46.0	48.7	48.1	**55.6**	54.7
Discourse	4.2	4.0	**7.5**	**21.5**	18.1	20.7	33.6	3.6	**34.6**	37.8	3.6	**38.0**	**50.6**	3.6	43.4	**39.6**	3.6	39.1
Empathy	**16.7**	16.7	16.7	**16.7**	16.7	16.7	**22.1**	16.7	17.1	21.2	**30.4**	22.8	**35.9**	29.8	28.2	34.7	**41.5**	39.6
Persuasion	9.2	**55.9**	45.0	11.0	**55.0**	48.7	11.3	**54.6**	51.7	8.4	42.8	**43.8**	**41.8**	38.8	35.2	43.1	**44.9**	46.1
Politeness	**22.4**	16.7	20.1	**42.4**	23.9	35.4	44.7	44.5	**51.9**	**57.2**	27.7	50.4	**51.9**	44.2	50.3	53.4	43.6	**53.9**
Power	**46.6**	44.5	33.3	**48.0**	39.8	41.4	40.8	**45.5**	43.5	55.6	58.9	**60.2**	52.6	52.0	**62.6**	56.9	57.2	**57.5**
Toxicity	43.8	**46.7**	33.3	40.4	34.7	**54.4**	**42.5**	34.7	36.7	43.4	38.7	**49.2**	34.0	33.3	**35.1**	48.2	44.7	**52.5**
Ideology	**24.0**	16.7	19.2	19.2	16.6	**21.3**	**28.3**	17.0	17.9	29.0	**31.7**	27.0	42.4	**48.5**	47.9	38.8	**38.9**	39.7
Tropes	1.7	**5.1**	3.4	**8.4**	5.1	3.4	**13.7**	10.0	11.6	**14.6**	8.4	10.0	**19.0**	8.4	6.8	**28.6**	27.3	24.6

Ziems et al. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics, 2024*

Recommendations

1. Integrate LLMs-in-the-loop to transform large-scale data labeling. [Maybe]
2. Prioritize open-source LLMs for classification [Probably]
3. Prioritize faithfulness, relevance, coherence, and fluency in your generations by opting for larger instruction-tuned models that have learned human preferences [We didn’t go through generation results]
4. Investigate how LLMs produce new CSS paradigms built on the multipurpose capabilities of LLMs in the long term [Remember the goal of topic modeling is not LDA]

Ziems et al. “Can Large Language Models Transform Computational Social Science?” *Computational Linguistics, 2024*

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

LLM+Human labeling

Fine-tuning approaches

- What if we had more than 5-10 labeled examples?

- If we have 100-1000s of examples, what can we do with them?

- Option 1: Fine-tuning the LLM
  - We fine-tuned models like BERT and RoBERTa but newer models are orders of magnitude larger. Can we actually update the model parameters?
- Option 2: Combining LLM and human labels (correcting LLM labels)

Parameter-efficient Fine-tuning

- In fine-tuning we need to updating and storing all the parameters of the LM
  - We would need to store a copy of the LM for each task

- With large models, storage management becomes difficult
  - E.g., A model of size 170B parameters requires ~340Gb of storage
  - If you fine-tune a separate model for 100 tasks:
    - 340 * 100 = 34 TB of storage!

Parameter-efficient Fine-tuning

- Augmenting the existing pre-trained model with extra parameters or layers and training only the new parameters
  - "parameter efficient": we only update a smaller set of parameters

- Two commonly used methods:
  - Soft prompts
  - Adapters

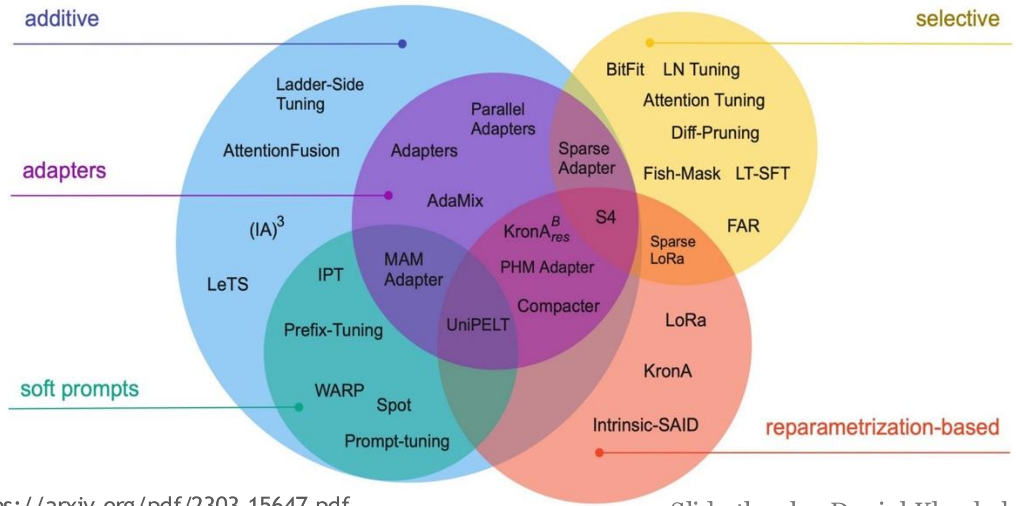
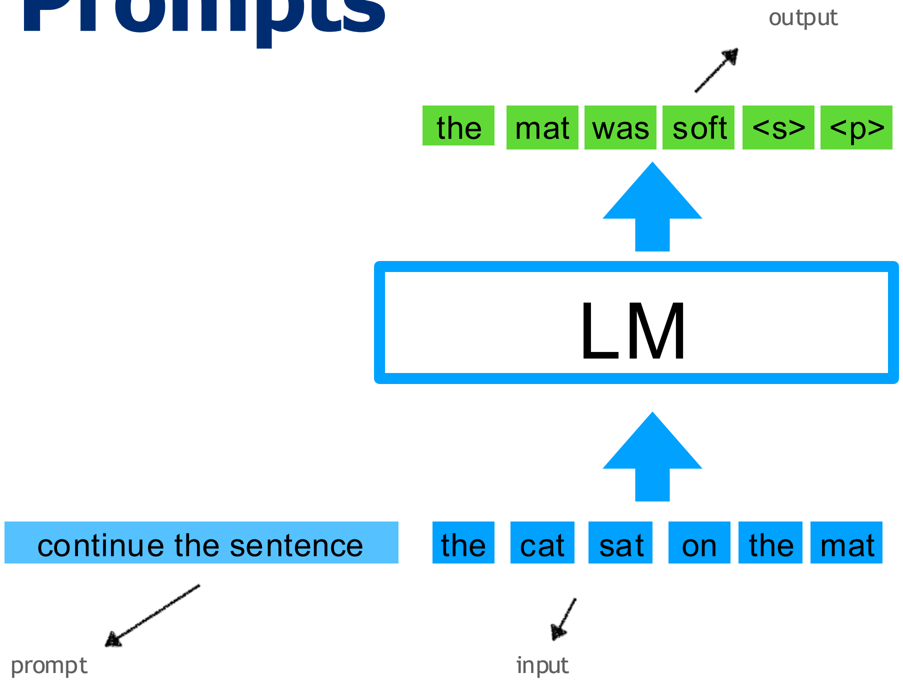


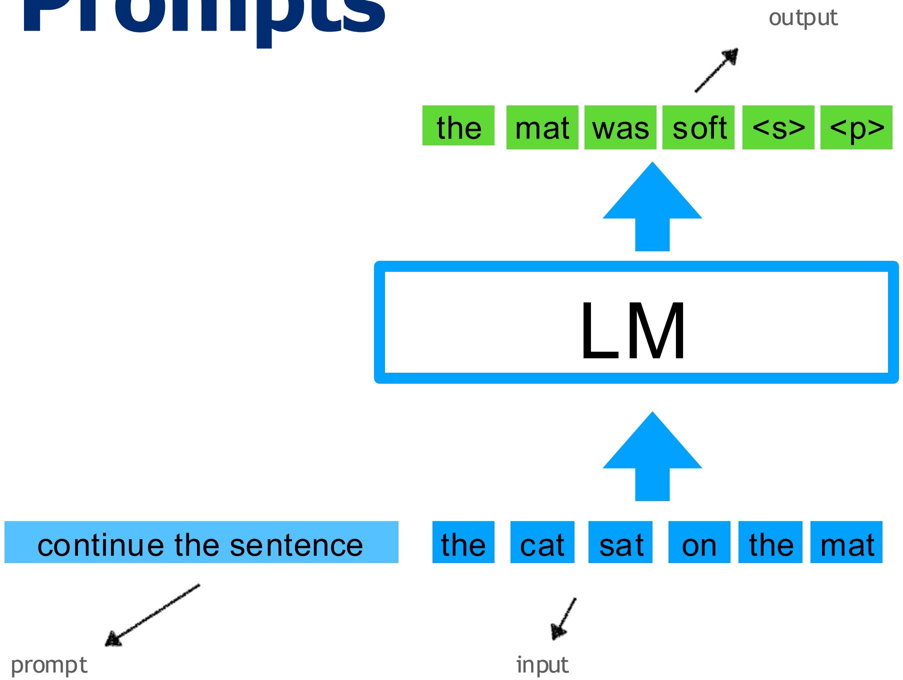
fig source https://arxiv.org/pdf/2303.15647.pdf

Slide thanks: Daniel Khashabi

Soft Prompts

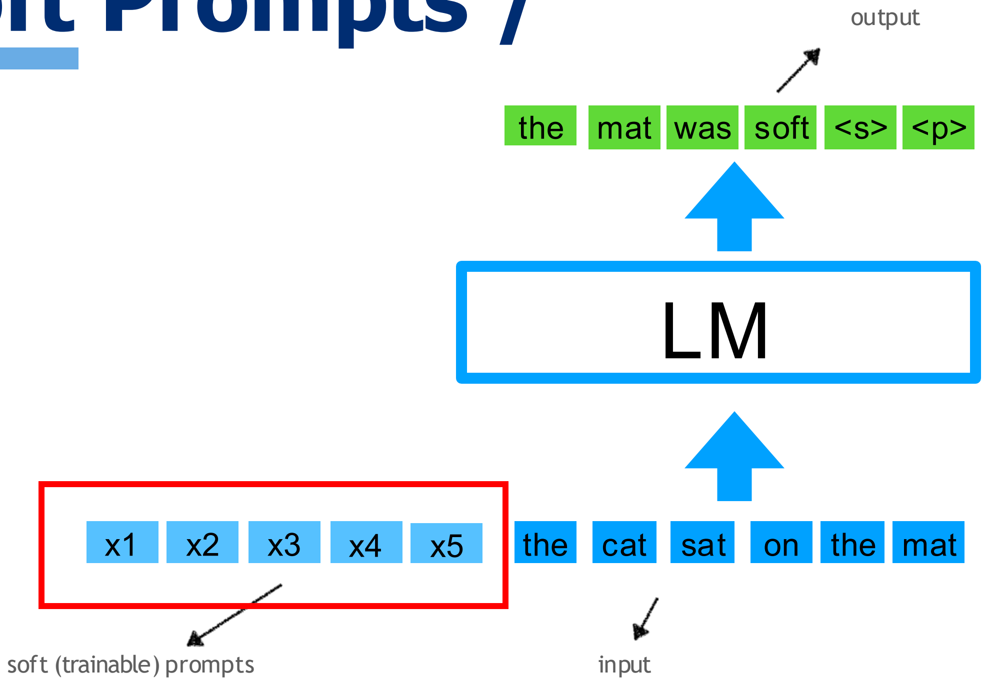


Soft Prompts



Previously, we constructed prompts following "good practice" guidelines and tried paraphrases of them

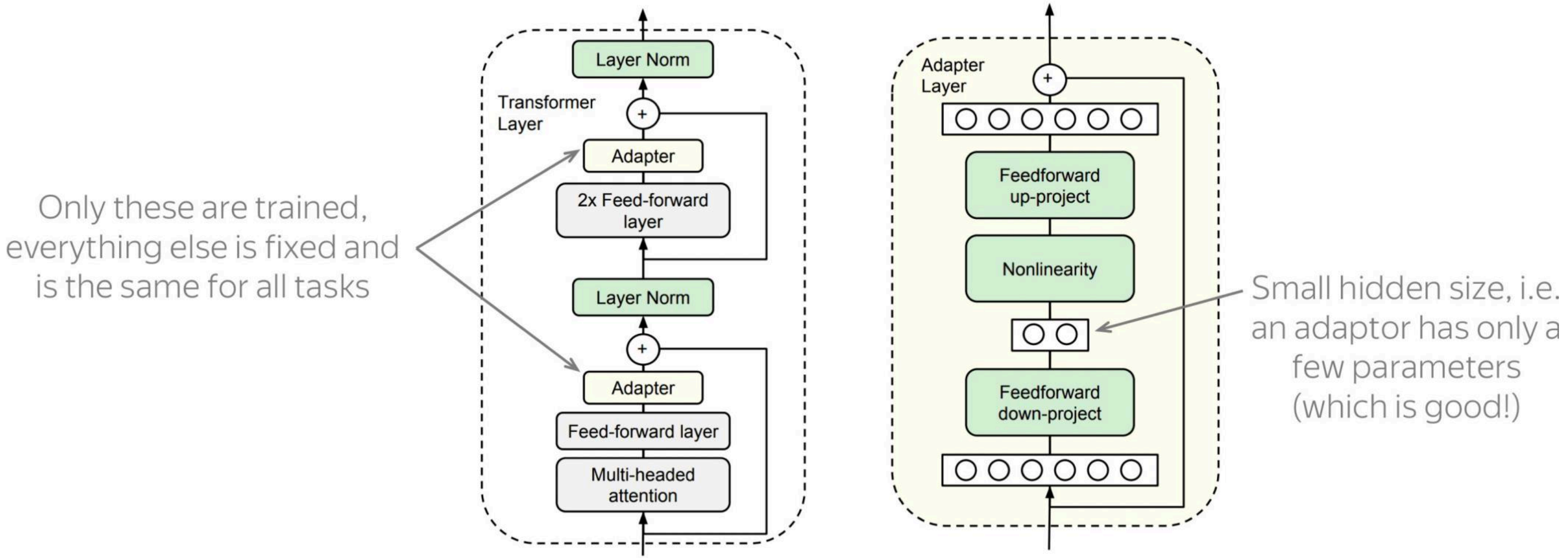
Soft Prompts /



Instead, we can just directly optimize for the best prompt!

Adapters

- **Idea:** train small sub-networks and only tune those.
  - FF projects to a low dimensional space to reduce parameters.
- No need to store a full model for each task, **only the adapter params**.

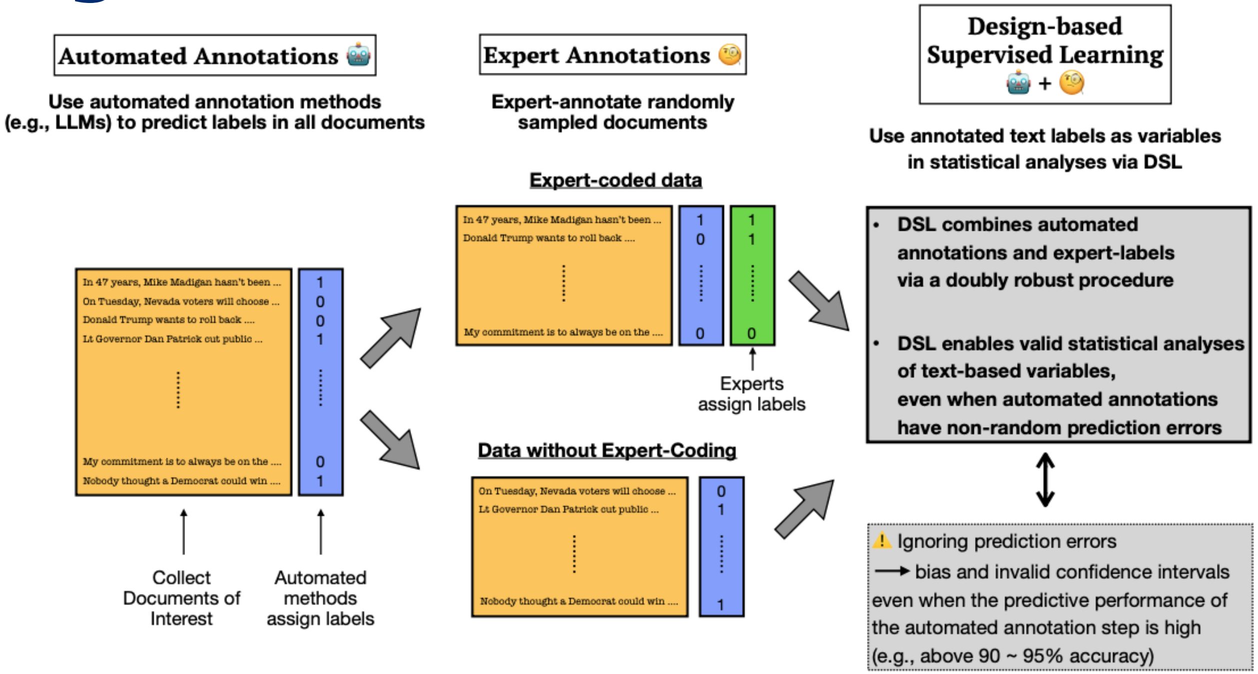


["Parameter-Efficient Transfer Learning for NLP", Houlsby et al., 2019.]

Fine-tuning approaches

- What if we had more than 5-10 labeled examples?
- If we have 100-1000s of examples, what can we do with them?
- Option 1: Fine-tuning the LLM
  - We fine-tuned models like BERT and RoBERTa but newer models are orders of magnitude larger. Can we actually update the model parameters?
- **Option 2: Combining LLM and human labels (correcting LLM labels)**

Recap: Design-based Supervised Learning

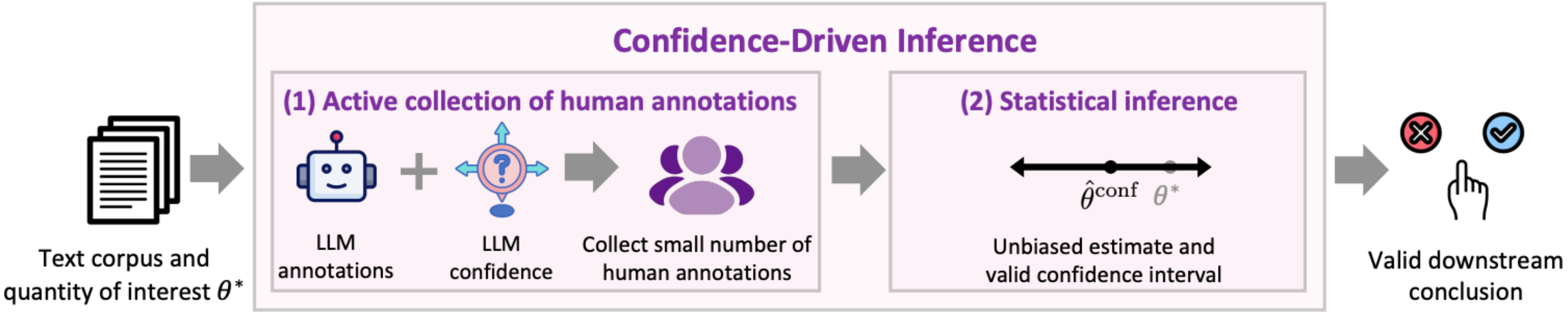


- Key idea: Use trusted human annotations to adjust less-trusted LLM annotations in downstream analysis model

Can Unconfident LLM Annotations Be Used for Confident Conclusions? Kristina Gligorić*, Tijana Zrnic*, Cinoo Lee*, Emmanuel Candès, and Dan Jurafsky. NAACL, 2025.

Adaptive Human+LLM Annotations

- Recall *active learning*: use initial model outputs to guide the next data to annotate
- We can do something similar, but instead of using labeled data to re-train the model, we can use it to



Can Unconfident LLM Annotations Be Used for Confident Conclusions? Kristina Gligorić*, Tijana Zrnic*, Cinoo Lee*, Emmanuel Candès, and Dan Jurafsky. NAACL, 2025.

Overall procedure

- Annotate data with an LLM
- Using LLM-verbalized confidence scores, select data to label manually
- Compute a *confidence driven* estimate of the value we are actually trying to compute

$$\hat{\theta}^{\text{conf}} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left(\lambda \hat{\ell}_{\theta,i} + (\ell_{\theta,i} - \lambda \hat{\ell}_{\theta,i}) \frac{\xi_i}{\pi_i} \right)$$

Estimate using LLM annotations

Estimate using human annotations

Indicates if data was human-annotated

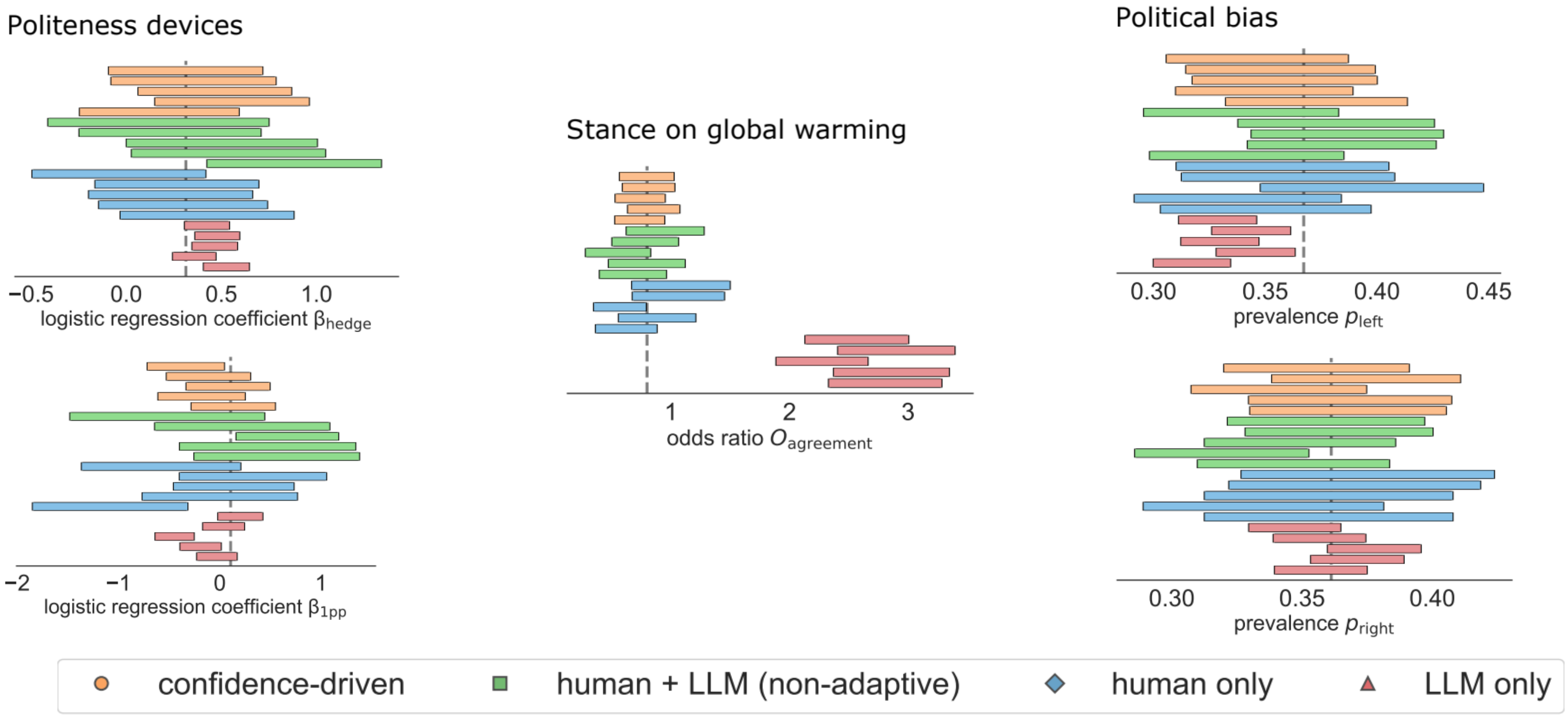
Probability data was annotated

[hyperparameter]

Overall procedure

- Annotate data with an LLM
- Using LLM-verbalized confidence scores, select data to label manually
- Compute a *confidence driven* estimate of the value we are actually trying to compute
- From $\hat{\theta}^{\text{conf}}$ it's possible to derive a valid confidence interval for the original value being estimated

Experiments



Conclusions

- LLMs can be useful zero or few shot models for some tasks, but performance can be much worse than supervised models
  - [Note: do we always care? If an LLM has accuracy 82% and a supervised model has accuracy 84%, is it worth hours of data annotating for an extra 2%?]
- Need to validate if the model works for the proposed task before using it
- Best practice: adjust for model errors
  - We probably don't care about model performance on each data point, we care about correctness of the downstream estimator

Conclusions

- What are more reasons we may not want to use GPT-4 to annotate data?
  - We pay per query or input/output tokens → annotating a full data set of hundreds of millions of tweets could become quite expensive
  - We have to share the data with OpenAI. Infeasible for private data like healthcare, law, social services etc.

Logistics

- Feedback on project proposals

- Next class:
  - Guest Ziang Xiao
  - Topic: LLMs for social experiments / human subject research