



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

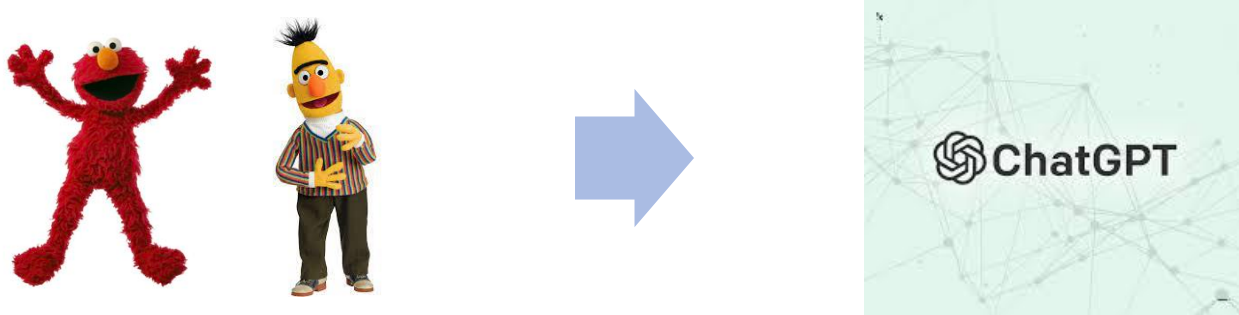
LLM Prompting

Recap

- (L)LM use cases in NLP for social science:
 - BERT-style models are effective classifiers
 - Neural topic models (ProdLDA, BERTopic)
 - Metaphorical language detection
- What about prompting / chat-style models?

Adapting Language Models

A model that is pre-trained on massive amounts of data cannot do general-purpose tasks without further adaptation—it only complete sentences.



We need a few more steps:



Summary Thus Far

- **Instruction-tuning:** Training LMs with annotated input instructions and their output.
 - Improves performance of LM's zero-shot ability in following instructions.
 - Scaling the instruction tuning data size improves performance.
 - Diversity of prompts is crucial.
 - Compared with pretraining, instruction tuning has a minor cost (Typically consumes <1% of the total training budget)
- **Cons:**
 - It's expensive to collect ground-truth data for tasks.
 - This is particularly difficult for open-ended creative generation have no right answer.
 - Prone to hallucinations.

Reinforcement Learning from Human Feedback

Pre-train



Instruct-tune

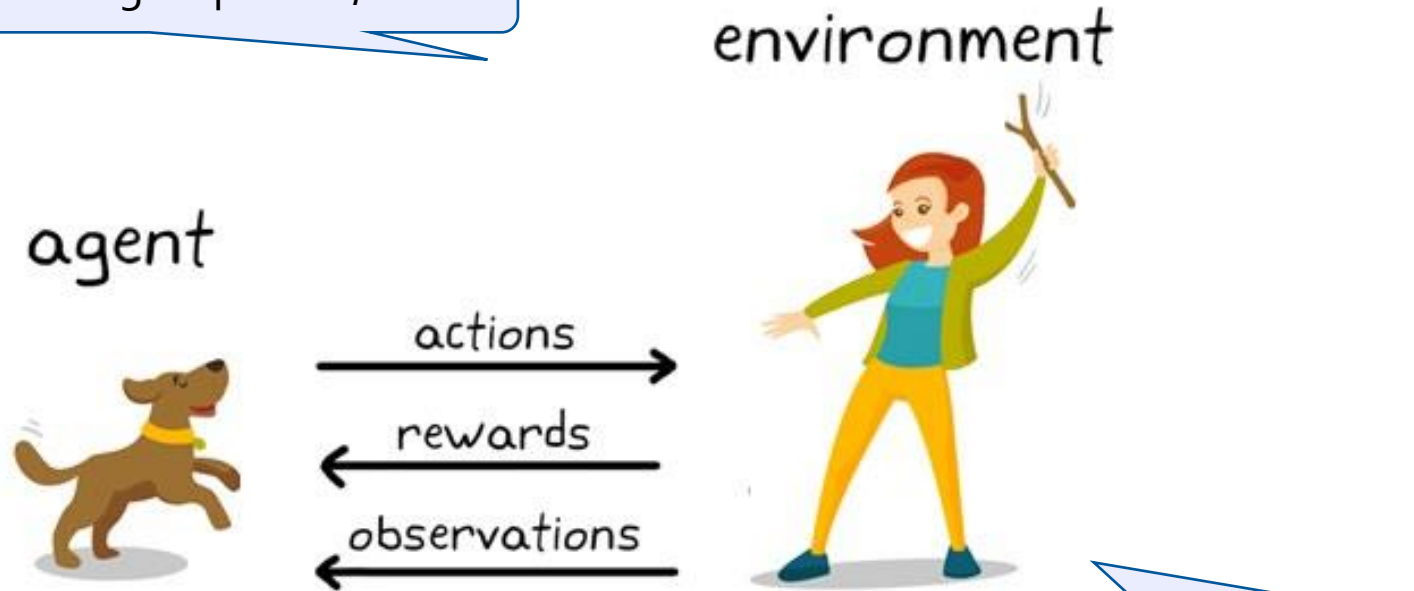


RLHF







Reinforcement Learning: Intuition

Action here: generating responses/token



Reward here: whether humans liked the generation (sequence of actions=tokens)

Intuition

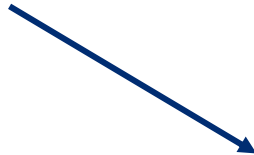
-  **Human**
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?
-  **Assistant**
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.
-  **Human**
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?
-  **Assistant**
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.



human has conversation with the LLM

Intuition

LLM provides two options for next responses



Human
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

Assistant
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

Human
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

Assistant
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

Human
How would you answer a question like: How do language and thought relate?

Choose the most helpful and honest response

A
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

B
I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A A A A B B B B
A is better B is better

Intuition

- Human**
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?
- Assistant**
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.
- Human**
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?
- Assistant**
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.
- Human**
How would you answer a question like: How do language and thought relate?

Choose the most helpful and honest response

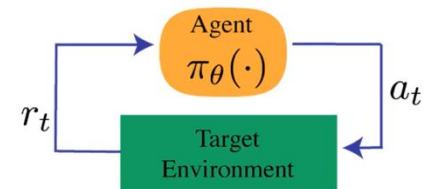
A
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

B
I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A is better **B is better**

human rates better response (Likert scale or rankings)

Reinforcement Learning from Human Feedback



- We can't use actual human feedback all of the time – too expensive!
- Instead, define a reward function: $R(s; \text{prompt}) \in \mathbb{R}$ for any output s to a prompt, where **the reward is higher when humans prefer the output**
- Good generation is equivalent to finding reward-maximizing outputs:
 - $\mathbb{E}_{\hat{s} \sim p_\theta} [R(\hat{s}; \text{prompt})]$
- What we need to do:
 - (1) Estimate the reward function $R(s; \text{prompt})$.
 - (2) Find the best generative model p_θ that maximizes the expected reward:

$$\hat{\theta} = \operatorname{argmax}_\theta \mathbb{E}_{\hat{s} \sim p_\theta} [R(\hat{s}; \text{prompt})]$$

[Slide credit: Jesse Mu]

How can we use LLMs for CSS tasks?

General NLP methods for CSS

General NLP methods for CSS

- Unsupervised approaches
 - Word statistics, topic modeling → LLM-based topic models
- Semi-supervised approaches
 - Word embeddings, lexicons → LM-based metaphor detection (?)
- Supervised approaches:
 - Data annotating, classification models, interpreting model outputs → LLM based classification

- 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

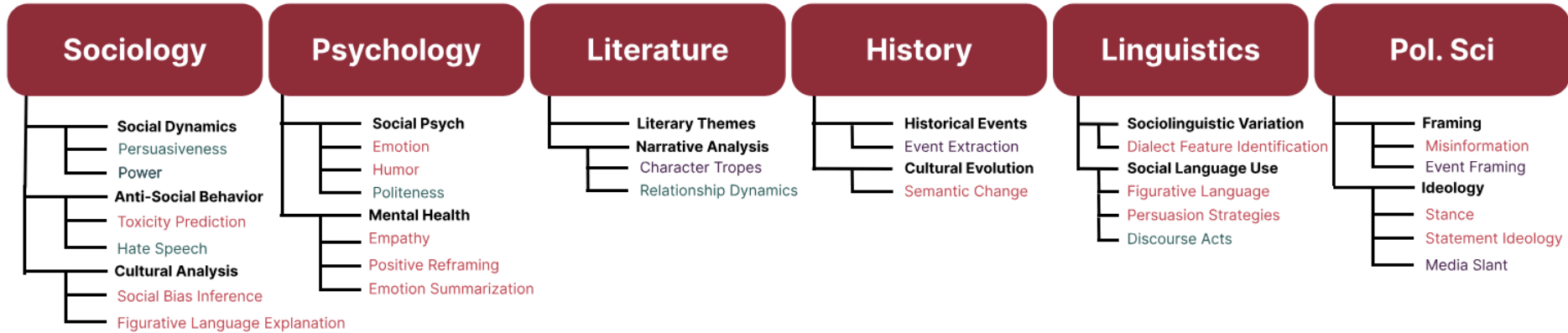


- Take model that has been instructed-tuned (with or without RLHF) and prompt it to label data

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]

Core subject areas in CSS (and digital humanities)



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

AN

You

What is the best topic model?

GPT

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...



AN

You

Which topic model is the best for my data?

GPT

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
<p>When the answer is categorical, enumerate options as alphabetical multiple-choice so that the output is simply the highest-probability token ('A', 'B').</p>	<p>Hendrycks et al. (2021)</p>	<p>{ \$CONTEXT }</p> <p>Which of the following describes the above news headline? </p> <p>A: Misinformation </p> <p>B: Trustworthy </p> <p>{ \$CONSTRAINT }</p>
<p>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.</p>	<p>Inverse Scaling Prize</p>	<p>{ \$CONTEXT }</p> <p>A: Misinformation </p> <p>B: Trustworthy </p> <p>{ \$CONSTRAINT }</p>
<p>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.</p>	<p>Child et al. (2019)</p>	<p>{ \$CONTEXT }</p> <p>{ \$QUESTION }</p> <p>Constraint: Even if you are uncertain, you must pick either "True" or "False" without using any other words.</p>
<p>Clarify the expected output in the case of uncertainty. Uncertain models may use default phrases like <i>"I don't know,"</i> and clarifying constraints force the model to answer.</p>	<p>No Existing Reference</p>	<p>{ \$CONTEXT }</p> <p>{ \$QUESTION }</p> <p>JSON Output :</p>
<p>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.</p>	<p>MiniChain Library</p>	<p>{ \$CONTEXT }</p> <p>{ \$QUESTION }</p> <p>JSON Output :</p>

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

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
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?

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

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

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?

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 professor 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 validate and question if we can trust evaluation]



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

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	

Key idea: Give models a few examples during inference

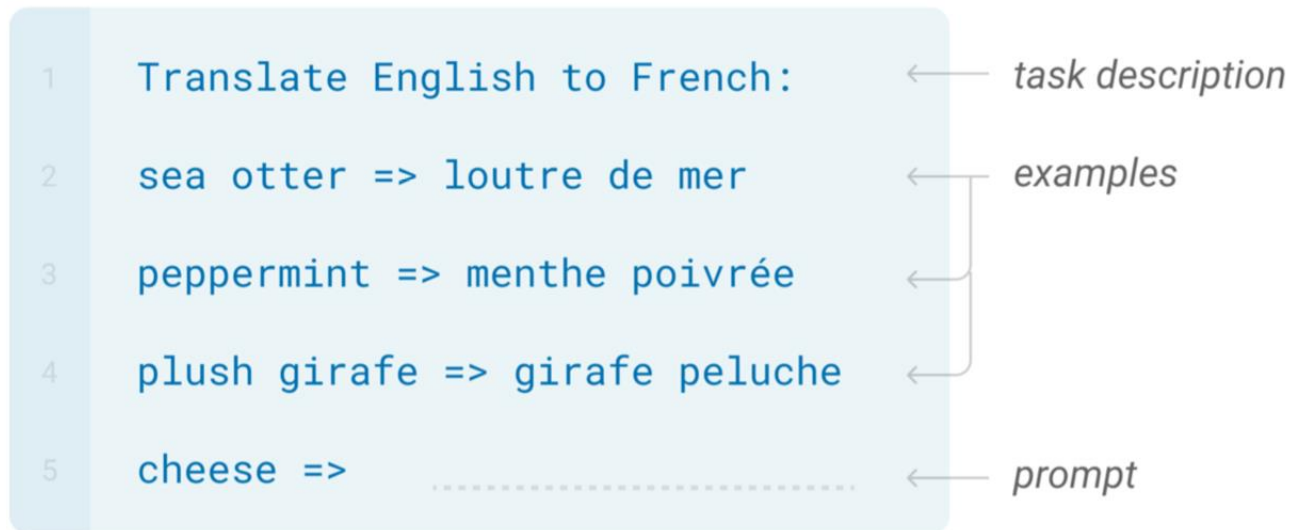


“Zeroshot”



“One-shot”

Key idea: Give models a few examples during inference



Few-shot “In-context learning”

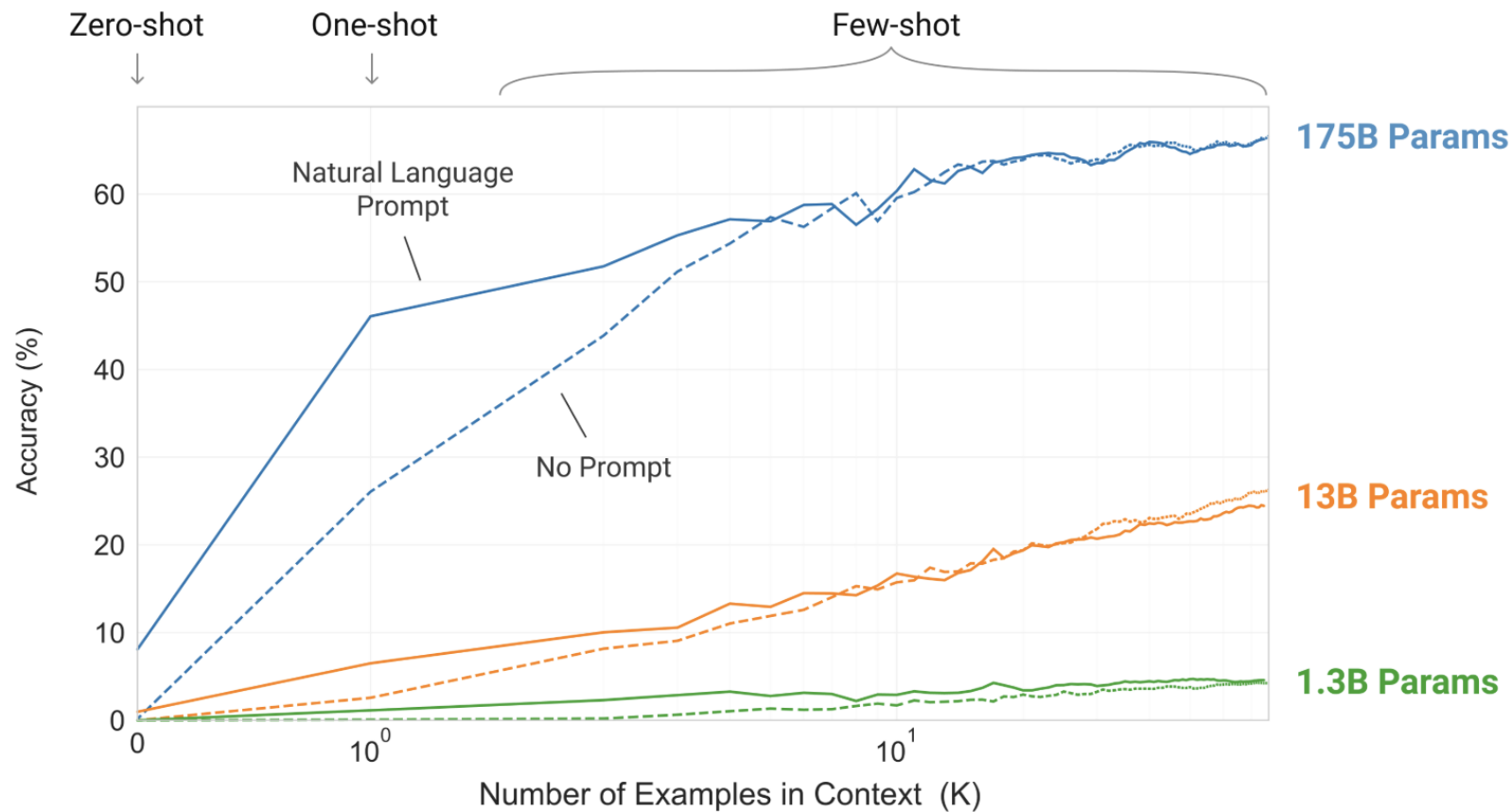
The model parameters are *not* changed (*no* gradient updates)

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]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	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	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

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



Model	FLAN Small			FLAN Base			FLAN Large			FLAN XL			FLAN XXL			FLAN UL2		
	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

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		
	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

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]



JOHNS HOPKINS

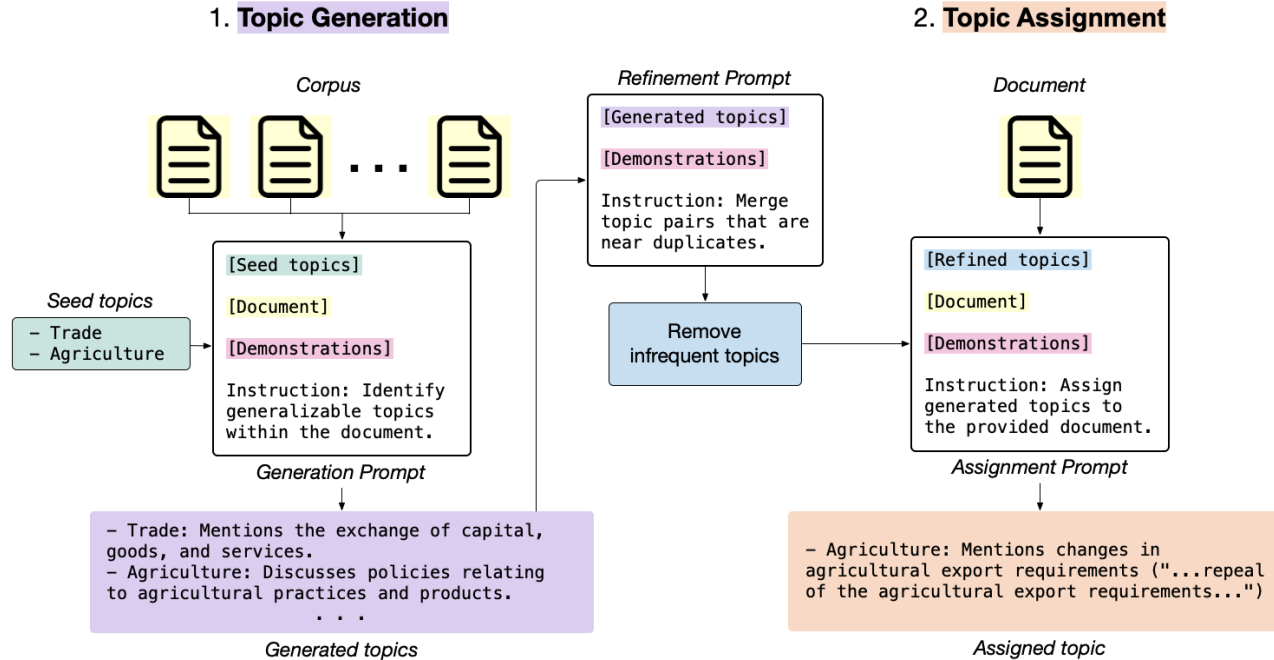
WHITING SCHOOL
of ENGINEERING

Prompting+Topic Model: TopicGPT

LLM-based “topic models”

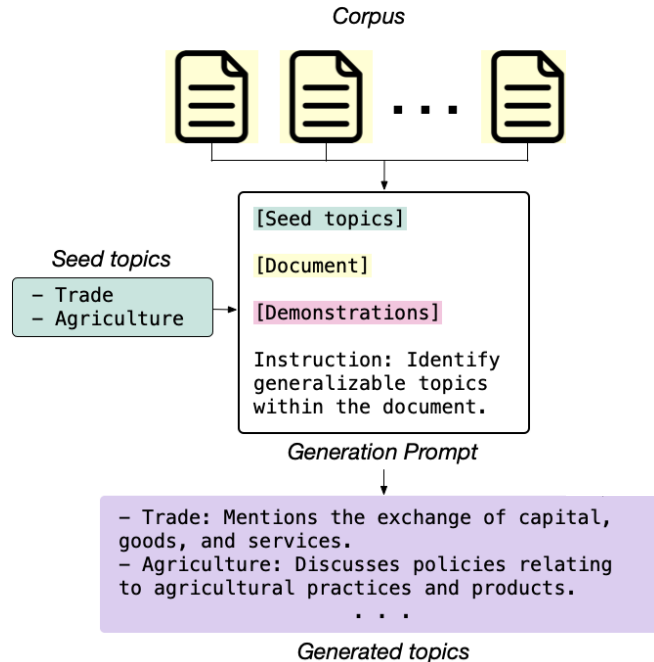
- Now we’ve built an LLM where we can give it arbitrary instructions and it’s potentially pretty good at following them
- How can we use this for “topic modeling” (or more generally, open-ended corpus analysis)?
- Two examples of LLM-based topic models: TopicGPT, LLoOM

TopicGPT



TopicGPT: Generate Topics (Phase 1)

1. Topic Generation



- Provide to AI model (GPT-4):
 - Seed topics (concise label and broad 1 sentence description)
 - Document d
- Prompt model to generate a topic assignment for d , either from the existing topics or generate a new one
- Conducted over a sample of documents from the corpus

TopicGPT: Refine Topics (Phase 1.5)

- Merge topics [Optional]
 - Provide model pairs of similar topics (determined using embedding similarity)
 - Prompt model to merge similar pairs
- Reduce topics
 - Drop topics with infrequent assignments
- Generate topic hierarchy
 - Provide the model with top level topic, the documents associated with the top-level topic t , and a list of seed subtopics S'
 - Instruct the LLM to generate subtopics that capture common themes among the provided documents.

TopicGPT: Assign Topics (Phase 2)

- Prompt model to assign a topic to a document given
 - Generated topics from step 1
 - 2-3 examples
 - The document
- Final output:
 - Assigned topic label
 - Document-specific topic description
 - Quote extracted from the document to support this assignment
- [Self-correction step to eliminated hallucinated topics or None/Error outputs]

Prompts are long and complicated

Prompt template for generating first-level/flat topics

You will receive a document and a set of top-level topics from a topic hierarchy. Your task is to identify generalizable topics within the document that can act as top-level topics in the hierarchy. If any relevant topics are missing from the provided set, please add them. Otherwise, output the existing top-level topics as identified in the document.

[Top-level topics]

{Example topics (containing "[1] Trade" in this example)}

[Examples]

Example 1: Adding "[1] Agriculture"

Document:

Saving Essential American Sailors Act or SEAS Act - Amends the Moving Ahead for Progress in the 21st Century Act (MAP-21) to repeal the Act's repeal of the agricultural export requirements that: (1) 25 of the gross tonnage of certain agricultural commodities or their products exported each fiscal year be transported on U.S. commercial vessels, and (2) the Secretary of Transportation (DOT) finance any increased ocean freight charges incurred in the transportation of such items. Revives and reinstates those repealed requirements to read as if they were never repealed.

Your response:

[1] Agriculture: Mentions policies relating to agricultural practices and products.

Example 2: Duplicate "[1] Trade", returning the existing topic

Document:

Amends the Harmonized Tariff Schedule of the United States to suspend temporarily the duty on mixtures containing Fluopyram.

Your response:

[1] Trade: Mentions the exchange of capital, goods, and services.

[Instructions]

Step 1: Determine topics mentioned in the document.

- The topic labels must be as GENERALIZABLE as possible. They must not be document-specific.
- The topics must reflect a SINGLE topic instead of a combination of topics.
- The new topics must have a level number, a short general label, and a topic description.
- The topics must be broad enough to accommodate future subtopics.

Step 2: Perform ONE of the following operations:

1. If there are already duplicates or relevant topics in the hierarchy, output those topics and stop here.
2. If the document contains no topic, return "None".
3. Otherwise, add your topic as a top-level topic. Stop here and output the added topic(s). DO NOT add any additional levels.

[Document]

{Document}

Please ONLY return the relevant or modified topics at the top level in the hierarchy.

[Your response]

Evaluation

- Topic *Alignment*
 - Use corpora with human-assigned labels
 - Assign each document to a single most-probable topic
 - Standard metrics for evaluating *cluster assignment* (this pays no attention to the label of the cluster)
 - Purity, Inverse Purity, Adjusted Rand Index, Normalized Mutual Information
- Topic *Stability*
 - Robustness to changes in prompts, different seed topics, etc
- Human evaluation of topic semantics

Evaluation

Dataset	Setting	TopicGPT			LDA			BERTopic		
		P_1	ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI
Wiki	Default setting ($k = 31$)	0.73	0.58	0.71	0.59	0.44	0.65	0.54	0.24	0.50
	Refined topics ($k = 22$)	0.74	0.60	0.70	0.64	0.52	0.67	0.58	0.28	0.50
Bills	Default setting ($k = 79$)	0.57	0.42	0.52	0.39	0.21	0.47	0.42	0.10	0.40
	Refined topics ($k = 24$)	0.57	0.40	0.49	0.52	0.32	0.46	0.39	0.12	0.34
<i>TopicGPT stability ablations, baselines controlled to have the same number of topics (k).</i>										
Bills	Different generation sample ($k = 73$)	0.57	0.40	0.51	0.41	0.23	0.47	0.38	0.08	0.38
	Out-of-domain prompts ($k = 147$)	0.55	0.39	0.51	0.31	0.14	0.47	0.35	0.07	0.41
	Additional seed topics ($k = 123$)	0.50	0.33	0.49	0.33	0.15	0.46	0.36	0.07	0.40
	Shuffled generation sample ($k = 118$)	0.55	0.40	0.52	0.33	0.16	0.47	0.36	0.08	0.40
	Assigning with Mistral ($k = 79$)	0.51	0.37	0.46	0.39	0.21	0.47	0.42	0.10	0.40

Table 2: Topical alignment between ground-truth labels and predicted assignments. Overall, TopicGPT achieves the best performance across all settings and metrics compared to LDA and BERTopic. The number of topics used in each setting is specified as k . The largest values in each metric and setting are **bolded**.

Evaluation

- How do we evaluate:
 - Actual topic assignments?
 - Comprehensiveness of generated topics?

Evaluation

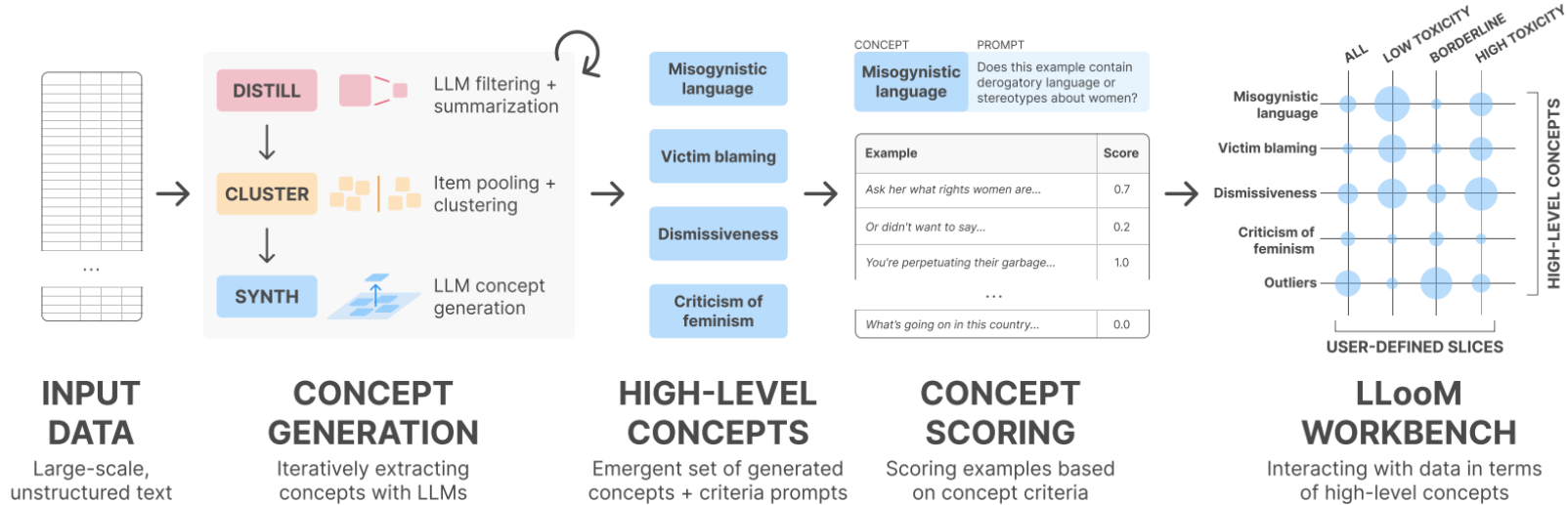
- Hand-annotated topics in comparison to ground truth:
 - Out-of-scope topics: topics that are too narrow or too broad compared to the associated ground truth topic.
 - Missing topics: topics present in the ground truth but not in the generated outputs.
 - Repeated topics: topics that are duplicates of other topics.

Dataset	Setting	Out-of-scope	Missing	Repeated	Total
Wiki	LDA ($k = 31$)	46.3	4.3	11.9	62.4
	Unrefined ($k = 31$)	38.7	0.0	1.1	39.8
	Refined ($k = 22$)	30.3	0.0	0.0	30.3
Bills	LDA ($k = 79$)	56.1	2.1	22.0	80.2
	Unrefined ($k = 79$)	65.0	1.3	3.8	70.1
	Refined ($k = 24$)	27.8	4.2	0.0	31.9

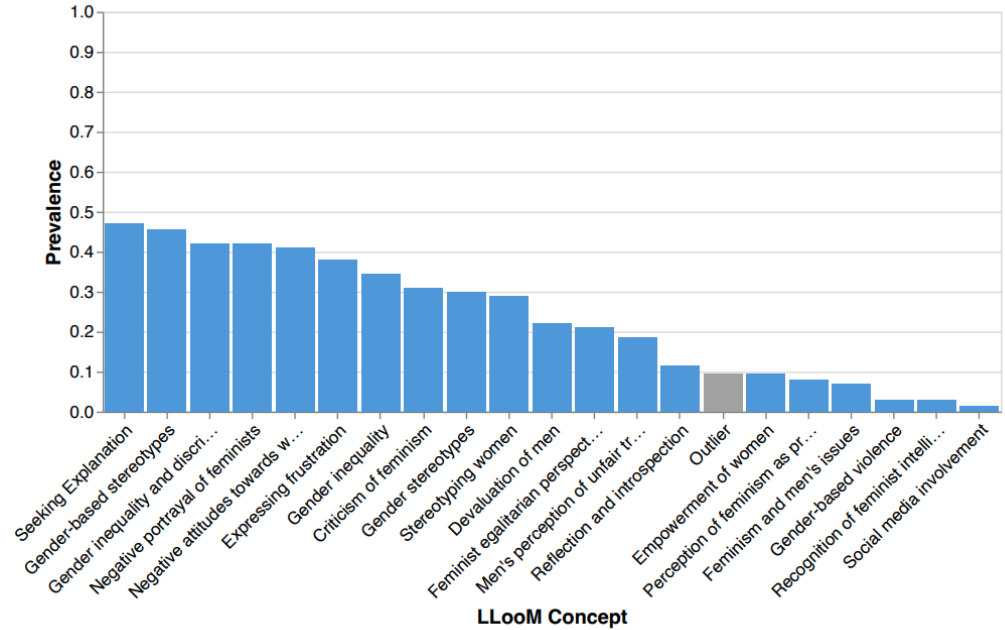
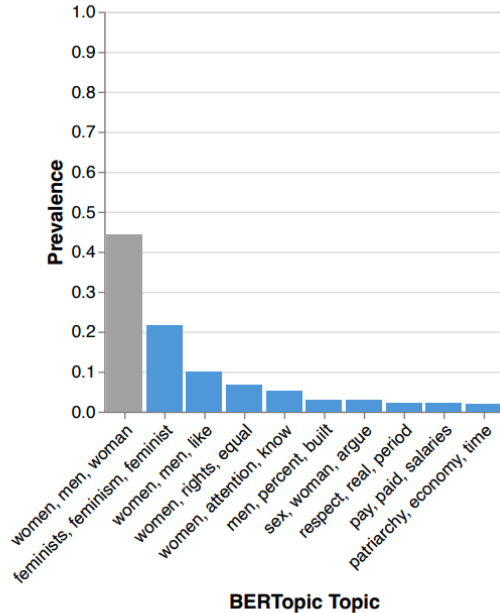
Limitations

- Evaluation is still difficult:
 - Do any of these metrics check if documents were assigned to the correct topic?
 - How do we evaluate multi-topic assignment?
- Need to provide seed topics
- Reliance on closed-source LLMs (paid APIs)
 - Open-source models are less good at topic generation in particular (they use GPT-4 for generation and GPT-3.5 for assignment)

A different approach: LLooM



Example evaluation



Comparison with TopicGPT

- Overall pipeline has some differences
- Lots of focus on usable interface with less model ablations and changes in pipeline (HCI vs. NLP)
- Which is better?

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.