LLMs and CSS

Explore potentials and concerns for Large Language Models in CSS.

Veniamin Veselovsky

Part 1: Crowdwork

Credits: The Verge

Part 1: Crowdwork

Credits: The Verge

Part 2: Silicon samples

Part 1: Crowdwork

Credits: The Verge

Part 2: Silicon samples

Part 3 (maybe): Coding!

Artificial Artificial Artificial Intelligence

Crowd Workers Widely Use Large Language Models for Text Production Tasks

Veniamin Veselovsky

Manoel Horta Ribeiro

Robert West

EPFL

Part 1: Outline

- What is crowdsourcing?
- Problems with LLM-use among workers
- Our study

Part 1: Outline

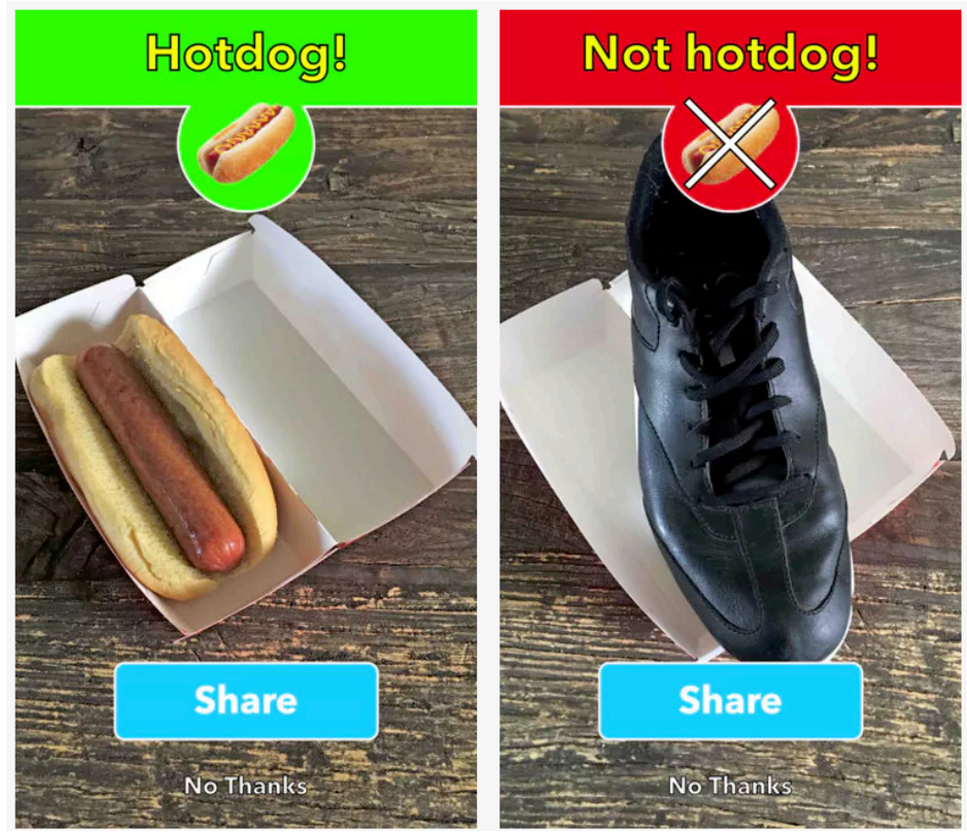
- **What is crowdsourcing?**
- Problems with LLM-use among workers
- Our study

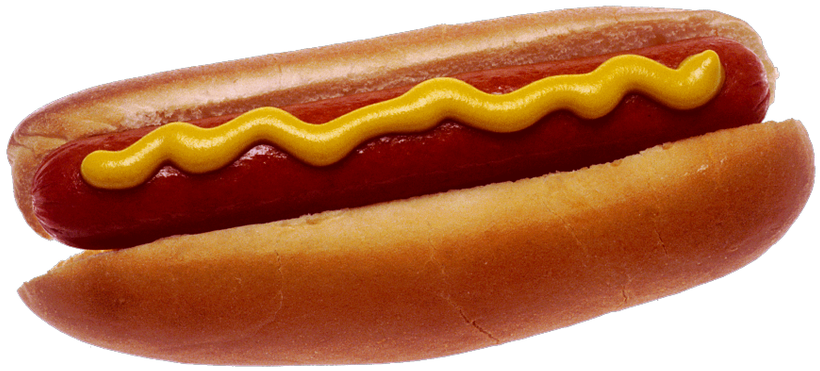
Crowdworking?

- Our task: Build a model to **classify** if an image has a **hotdog**.

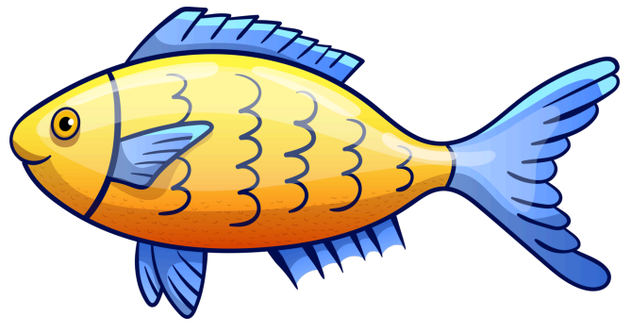
- Require images of hotdogs and not hotdogs

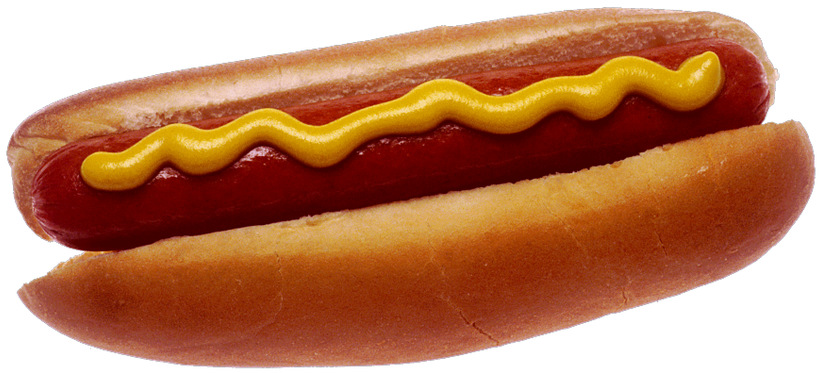
- Hundreds of thousands images

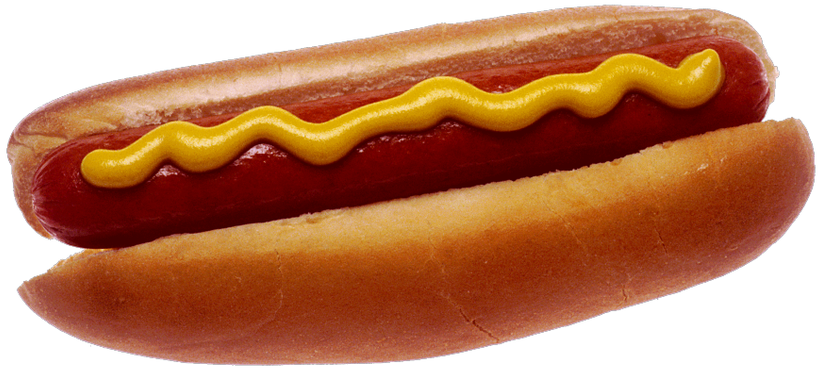





TÜRKİYE CUMHURİYET MERKEZ BANKASI
100
YÜZ
TÜRK LİRASI



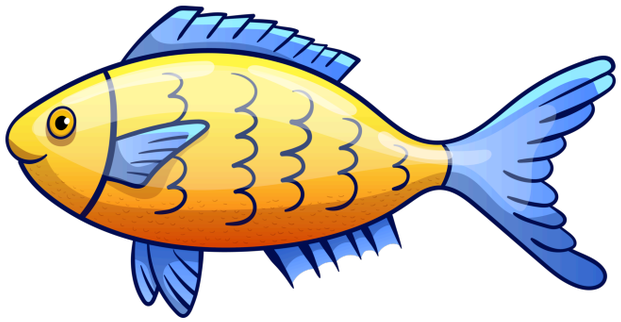




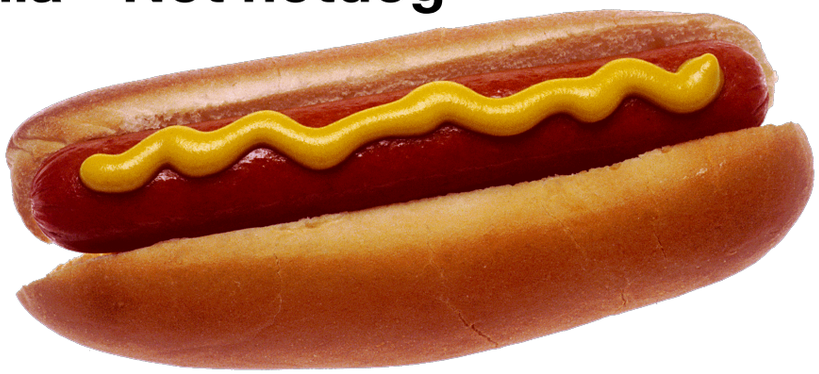
Venia - Hotdog

Venia - Not hotdog

Venia - Hotdog



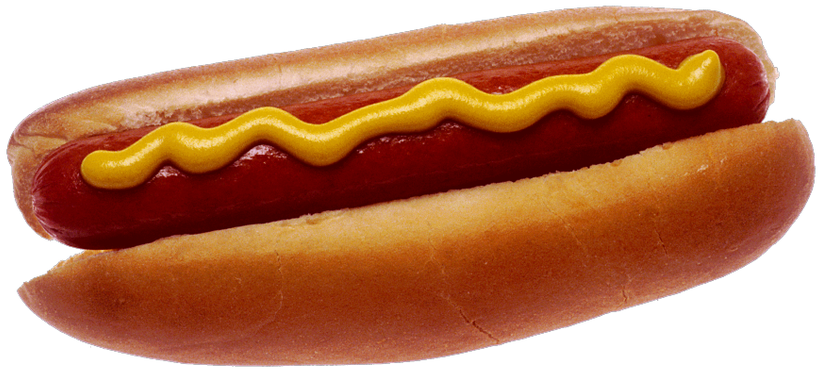
Venia - Not hotdog



Venia - Hotdog

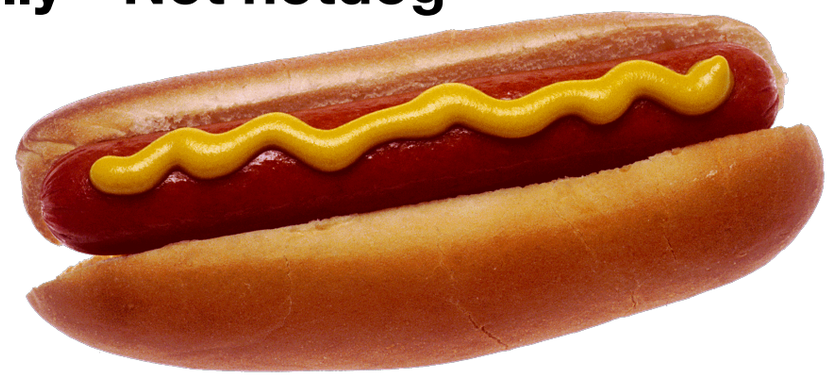
Venia - Hotdog

Venia - Not hotdog



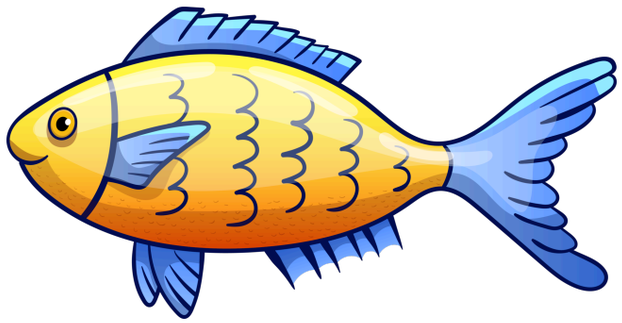
Bob - Hotdog

Sally - Not hotdog



Bill - Hotdog

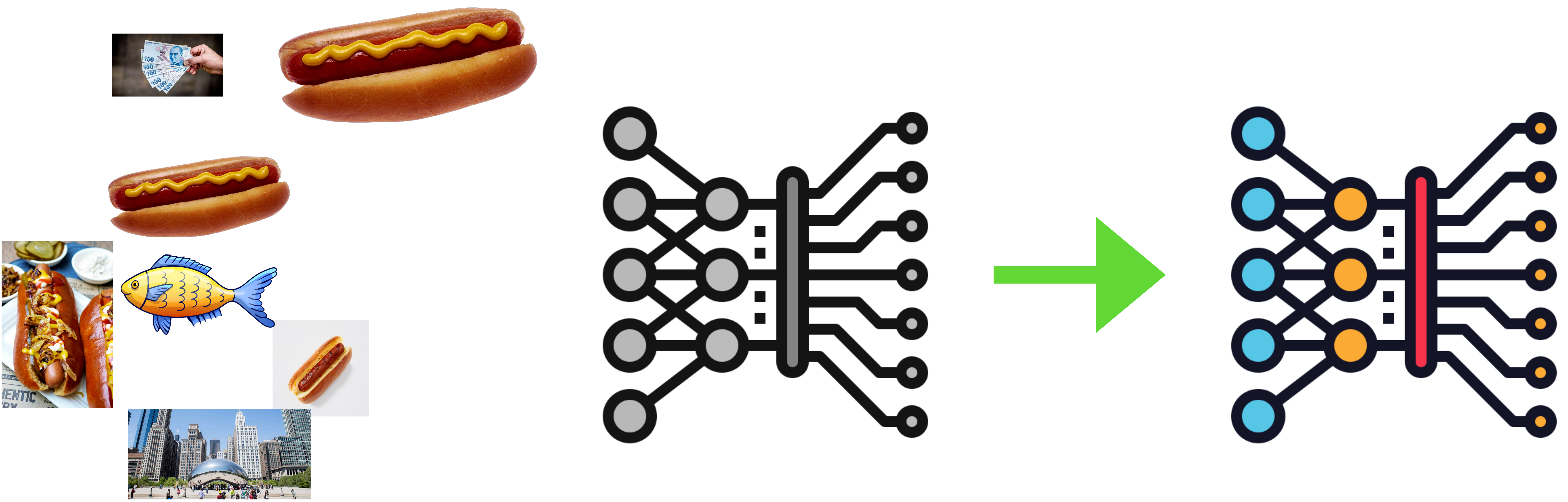
Manoel - Hotdog

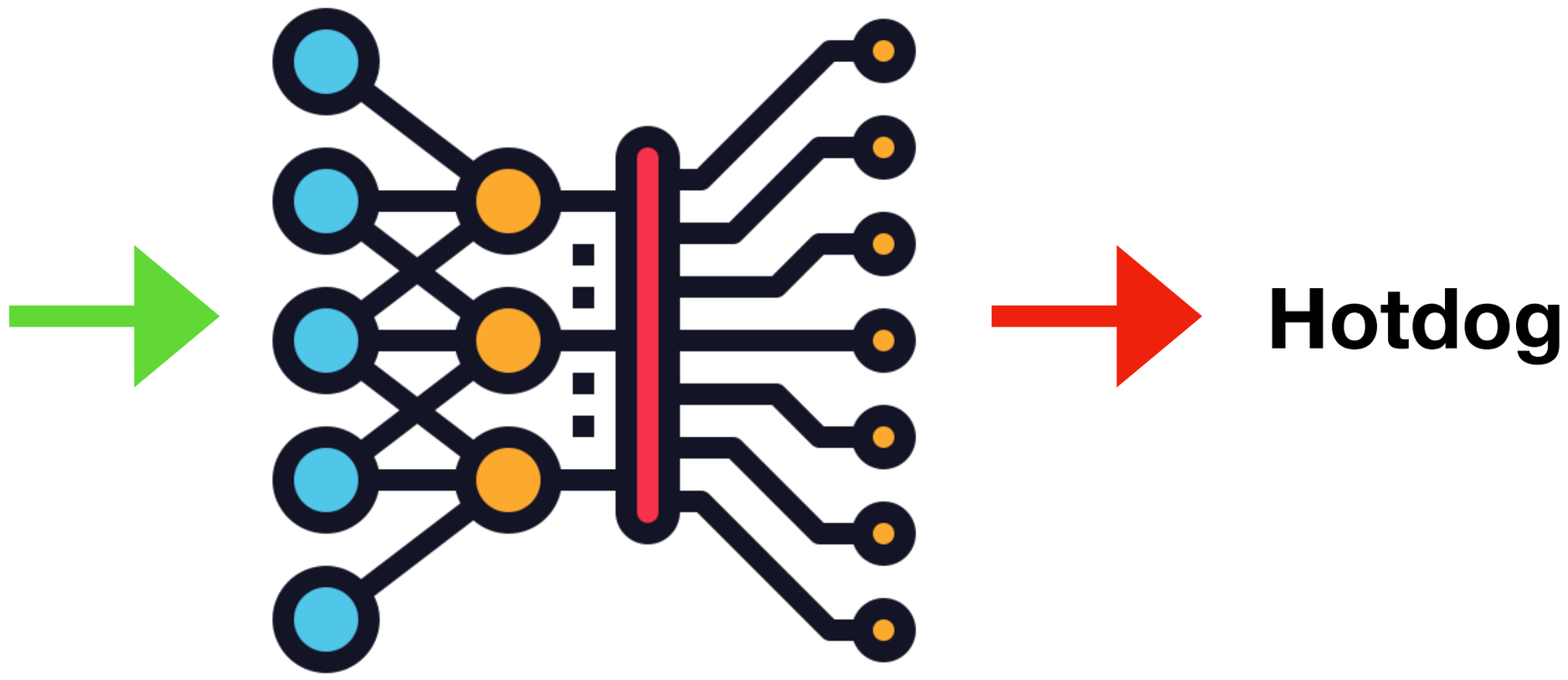


Manoel - Not hotdog

Jill - Not hotdog

Seb - Hotdog




Hotdog

Crowdworking?

- The key part of training the model was the **dataset**!

- Crowdsourcing platforms handle these microtasks.

  - Increase speed and scalability

  - Provide high quality talent

amazon mechanical turk™
Artificial Artificial Intelligence

Prolific

remotasks

surge AI

Crowdworking?

- The key part of training the model was the **dataset**!

- Crowdsourcing platforms handle these microtasks.

  - Increase speed and scalability

  - Provide high quality talent

amazon mechanical turk™
Artificial Artificial Intelligence

Prolific

remotasks

surge AI

Not without ethical concerns.

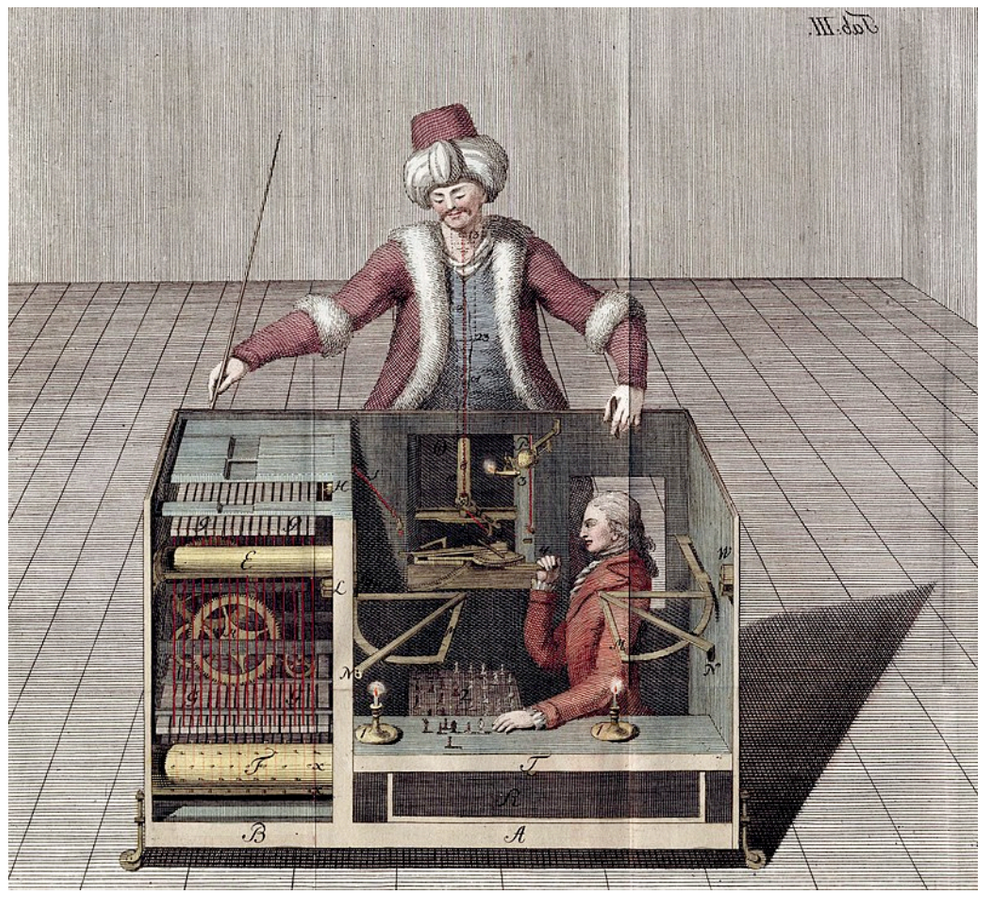
Human aspect

- Originally called “Artificial Artificial Intelligence”

- Importance of “human” annotation only **emerged with time**.

- Now crowdwork used for more than simple annotations.

- This has become a **key** part of the platforms use.



Causal Effects of Brevity on Style and Success in Social Media

KRISTINA GLIGORIĆ, EPFL, Switzerland
ASHTON ANDERSON, University of Toronto, Canada
ROBERT WEST, EPFL, Switzerland

We use these crowdsourcing platforms to study **humans**

Causal Effects of Brevity on Style and Success in Social Media

KRISTINA GLIGORIĆ, EPFL, Switzerland
ASHTON ANDERSON, University of Toronto, Canada
ROBERT WEST, EPFL, Switzerland

Interventions for Softening Can Lead to Hardening of Opinions: Evidence from a Randomized Controlled Trial

Andreas Spitz*
EPFL
Switzerland
andreas.spitz@epfl.ch

Ahmad Abu-Akel*
University of Lausanne
Switzerland
ahmad.abuakel@unil.ch

Robert West
EPFL
Switzerland
robert.west@epfl.ch

We use these crowdsourcing platforms to study **humans**

Causal Effects of Brevity on Style and Success in Social Media

KRISTINA GLIGORIĆ, EPFL, Switzerland
ASHTON ANDERSON, University of Toronto, Canada
ROBERT WEST, EPFL, Switzerland

Message Distortion in Information Cascades

Manoel Horta Ribeiro*
UFMG
manoelribeiro@dcc.ufmg.br

Kristina Gligorić
EPFL
kristina.gligoric@epfl.ch

Robert West
EPFL
robert.west@epfl.ch

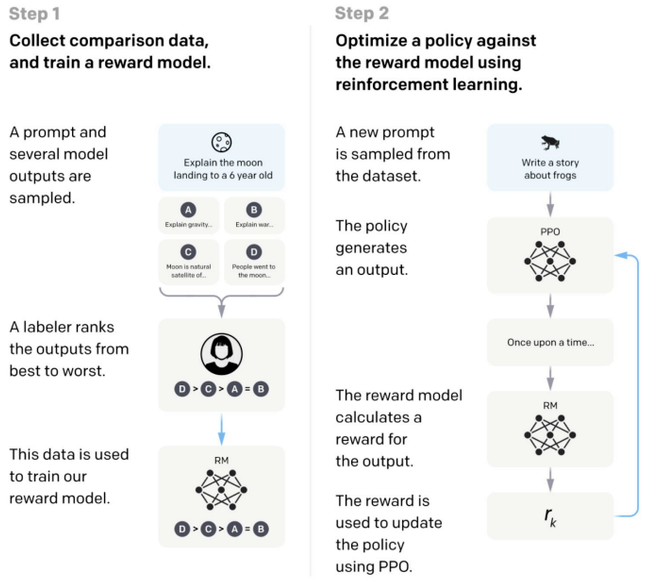
Interventions for Softening Can Lead to Hardening of Opinions: Evidence from a Randomized Controlled Trial

Andreas Spitz*
EPFL
Switzerland
andreas.spitz@epfl.ch

Ahmad Abu-Akel*
University of Lausanne
Switzerland
ahmad.abuakel@unil.ch

Robert West
EPFL
Switzerland
robert.west@epfl.ch

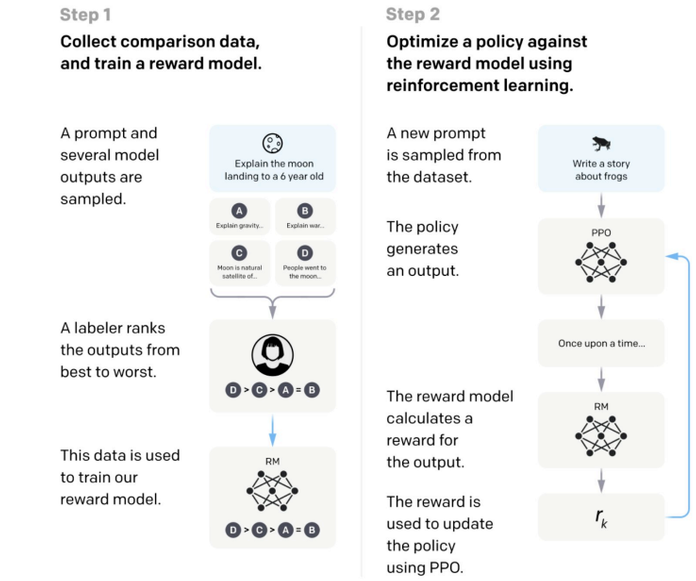
We use these crowdsourcing platforms to study **humans**



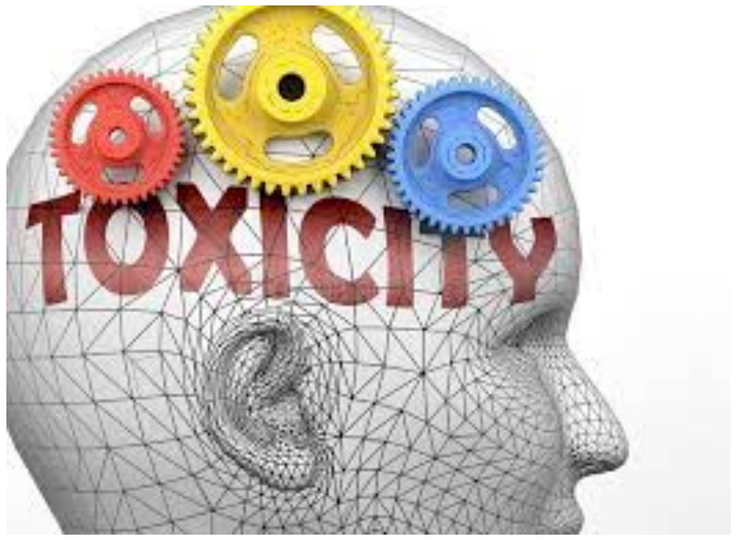
Adapted from OpenAI

Human preferences

New models improved through **human-knowledge**



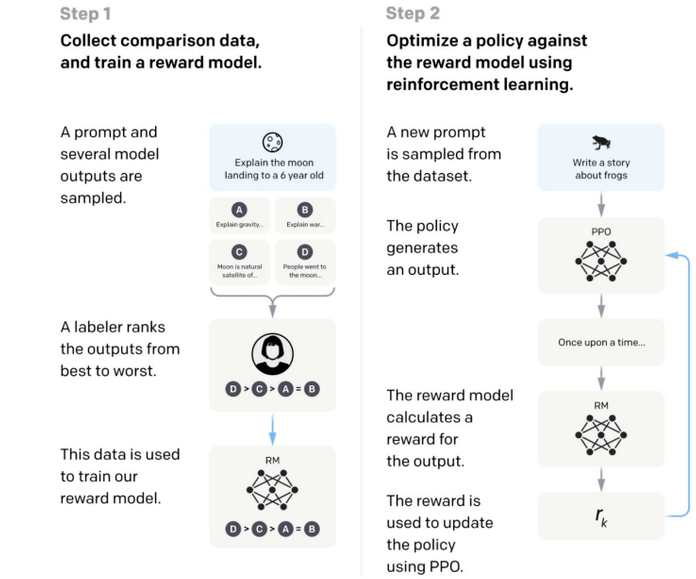
Adapted from OpenAI



Human preferences

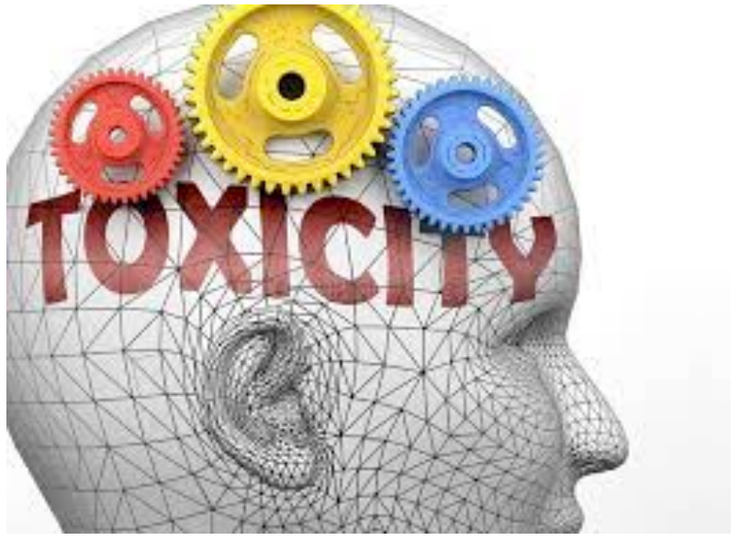
Toxicity classification

New models improved through **human-knowledge**



Adapted from OpenAI

Human preferences



Toxicity classification

Diverse language and options

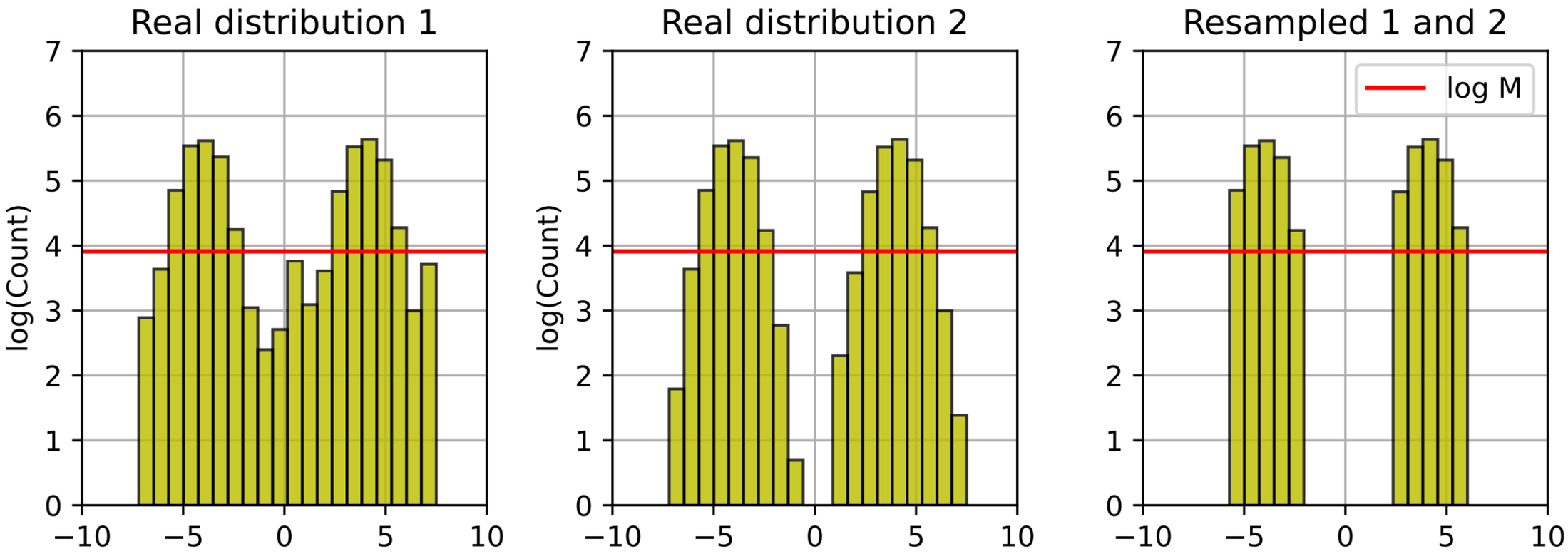
New models improved through **human-knowledge**

Part 1: Outline

- What is crowdsourcing?
- **Problems with LLM-use among workers**
- Our study

What if crowd workers are themselves using LLMs?

Problematic

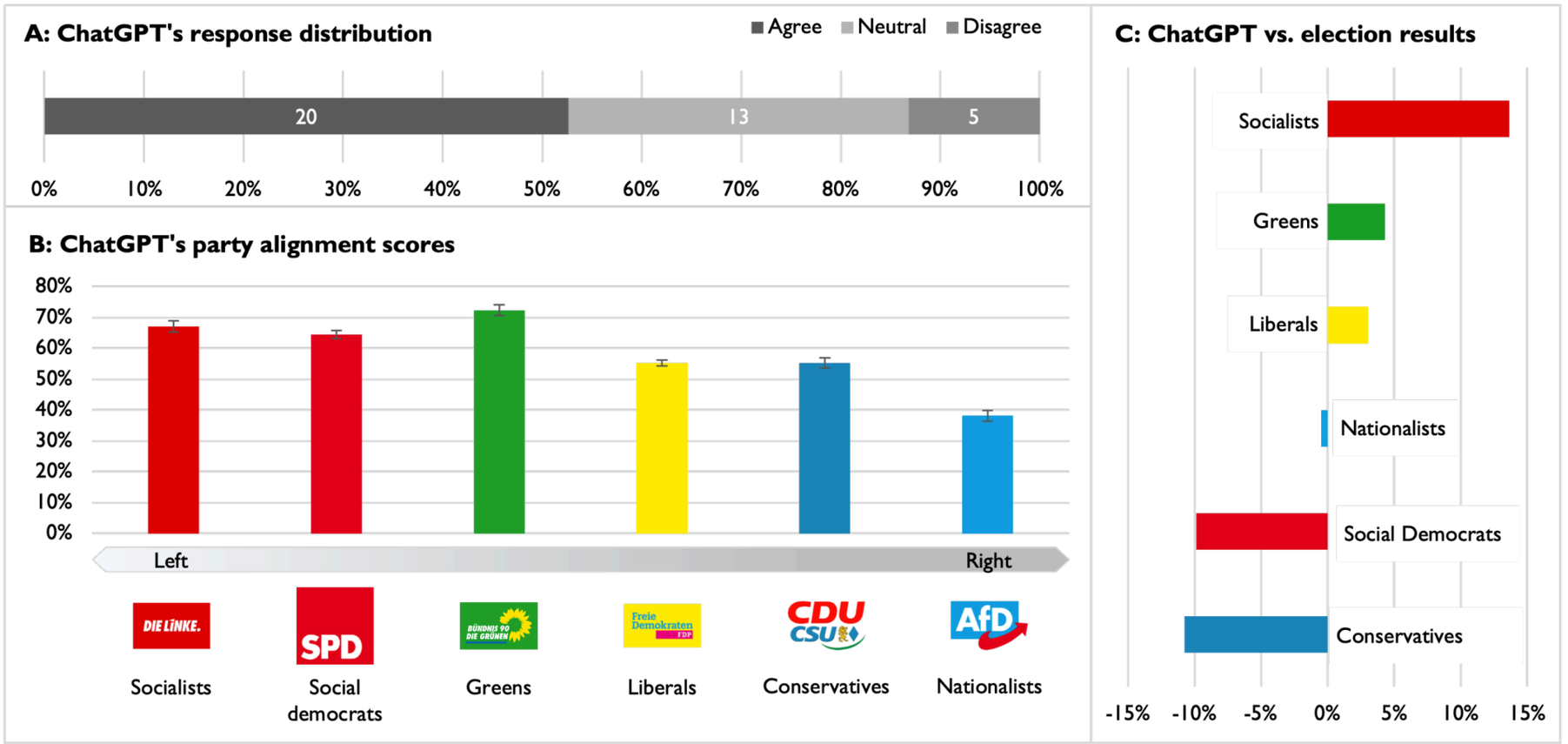


May lead to **model collapse** [0,1,2].

Problematic

May lead to **model collapse** [0,1,2].

Problematic



Less diversity in terms of who we study [3].

Part 1: Outline

- What is crowdsourcing?
- Problems with LLM-use among workers
- **Our study**

How to do it?

Goal: Estimate ChatGPT within crowdworkers.

Necessary: Create an approach for capturing LLM-use.

Operationalization: Re-run a previously built study on Mechanical Turk.

Instructions

You will be given a short text (around 400 words) with medicine-related information. Your task is to:

Summarize 400-word abstract to 100 words

- Write a summary of the text. Your summary should:
  - Convey the most important information in the text, as if you are trying to inform another person about what you just read.
  - Contain at least 100 words.

We expect high-quality summaries and will manually inspect some of them.

Comparison of Weight-Loss Diets with Different Compositions of Fat, Protein, and Carbohydrates

The possible advantage for weight loss of a diet that emphasizes protein, fat, or carbohydrates has not been established, and there are few studies that extend beyond 1 year. We randomly assigned 811 overweight adults to one of four diets; the targeted percentages of energy derived from fat, protein, and carbohydrates in
(...)
Satiety, hunger, satisfaction with the diet, and attendance at group session attendance was strongly associated with weight loss (0.2 kg per session attended). The diets improved lipid-related risk factors and fasting insulin levels.

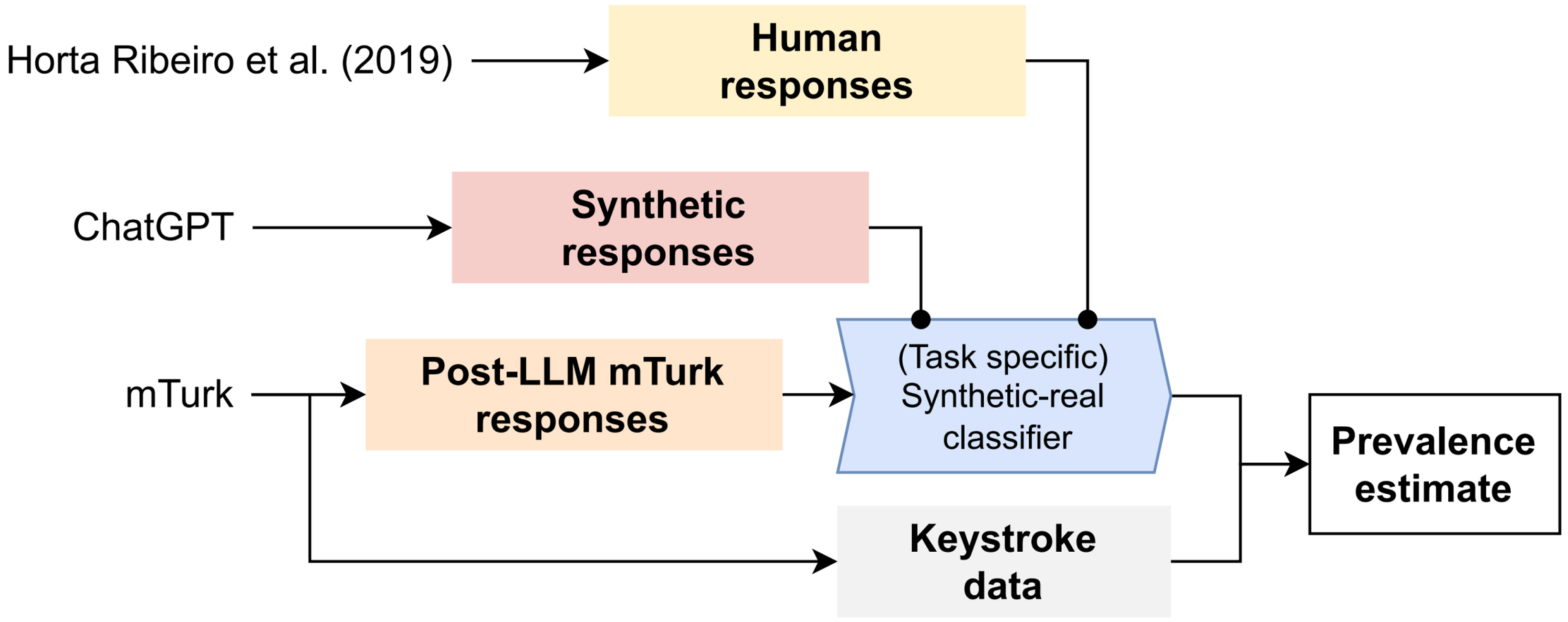
Sixteen different abstracts

Write your summary here

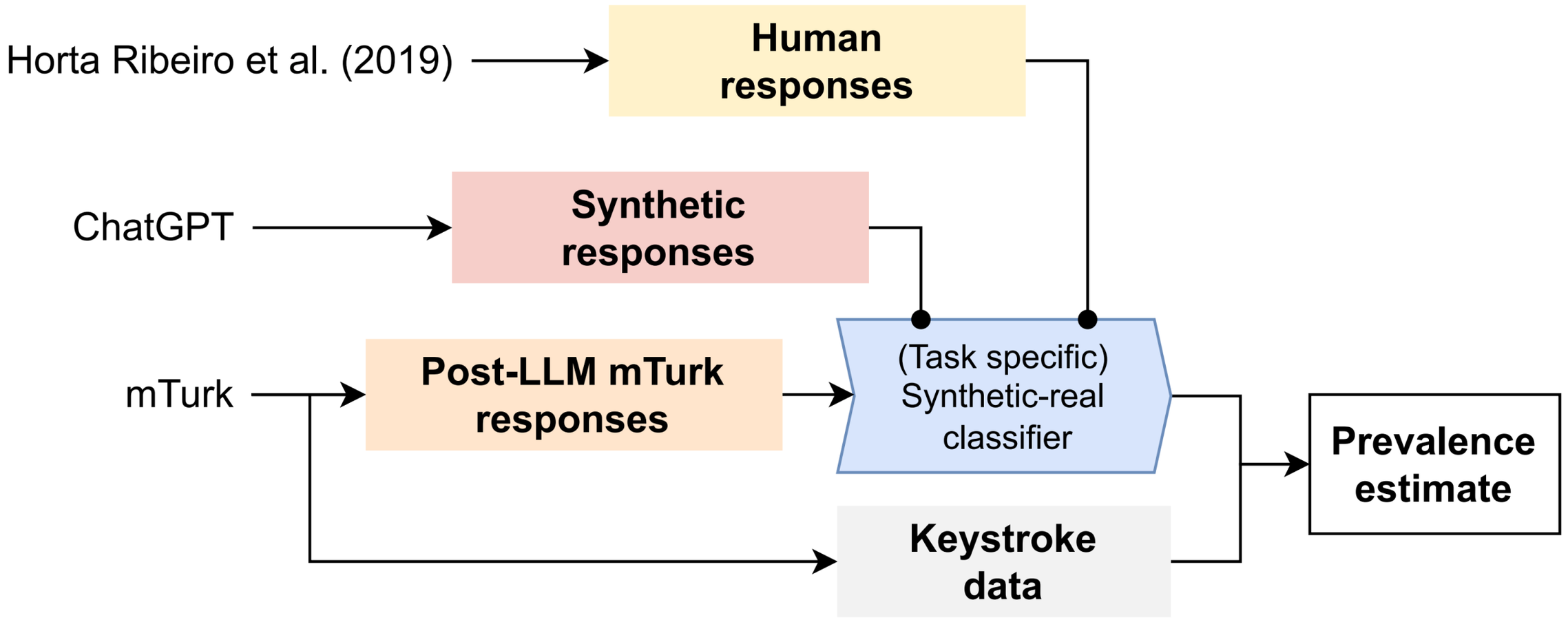
SUBMIT

Collected 46 summaries from 44 workers

Detection setup

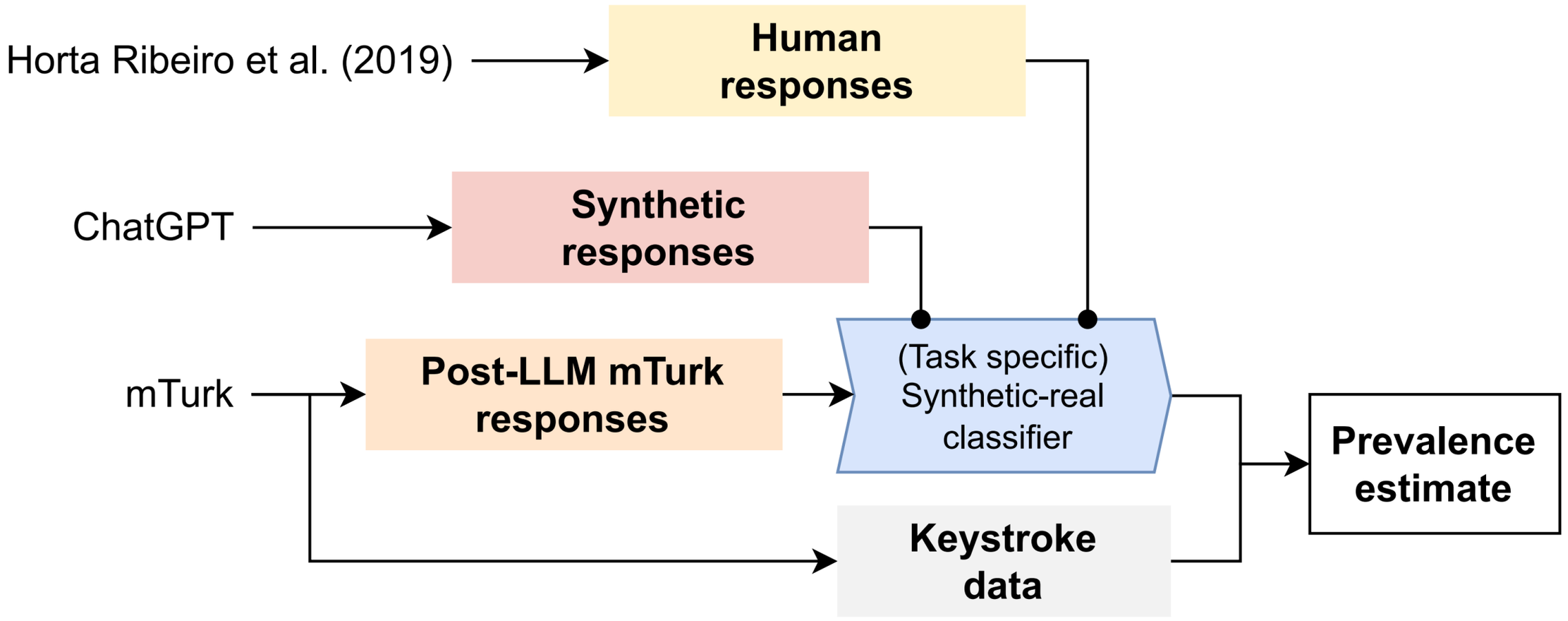


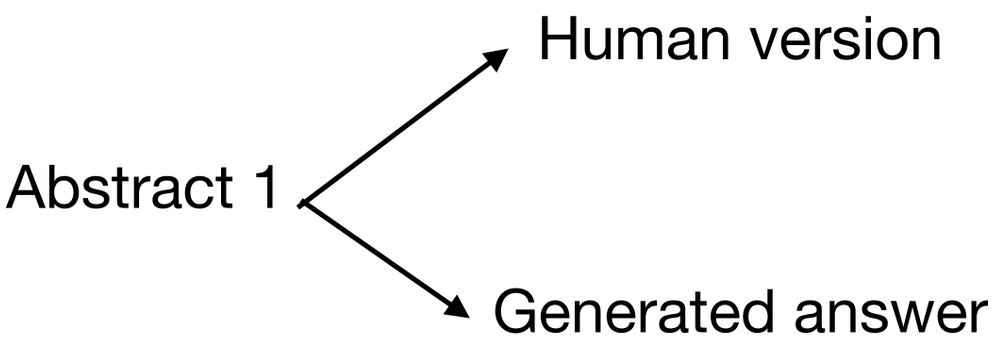
Detection setup



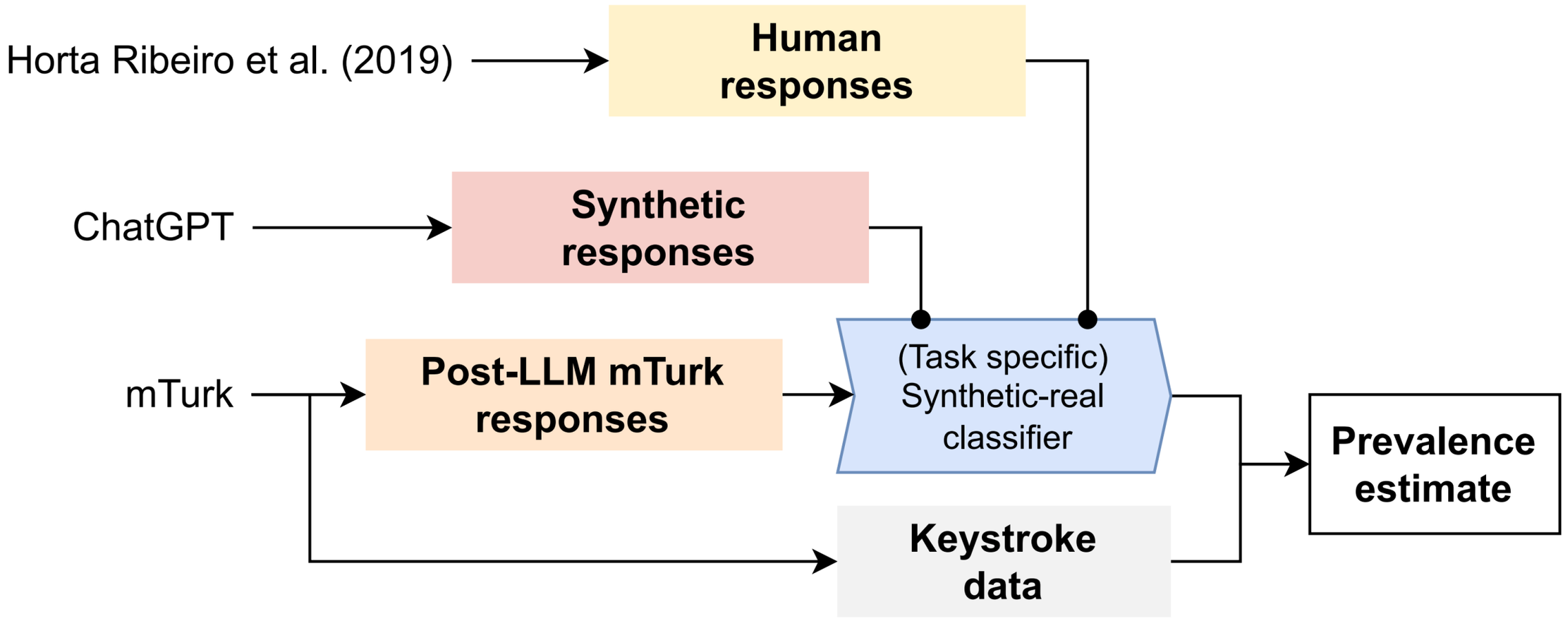
Abstract 1

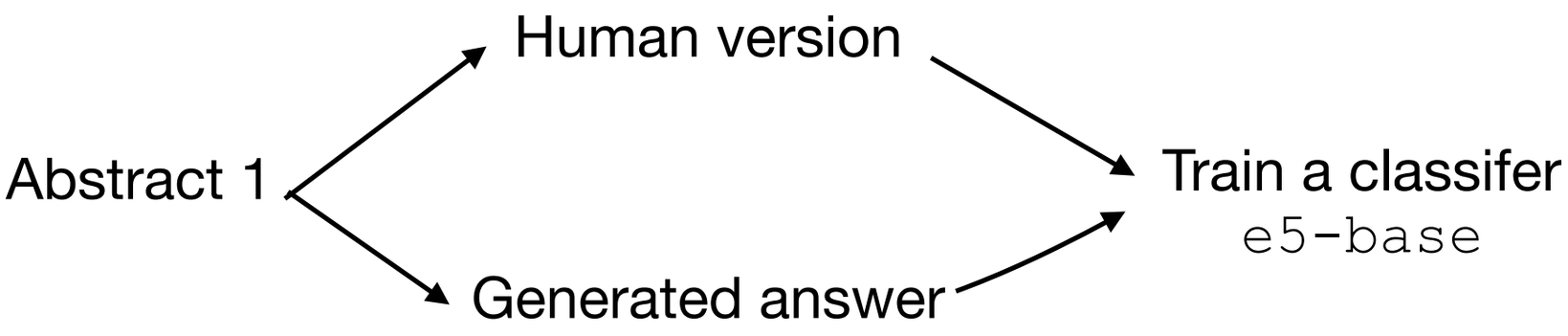
Detection setup



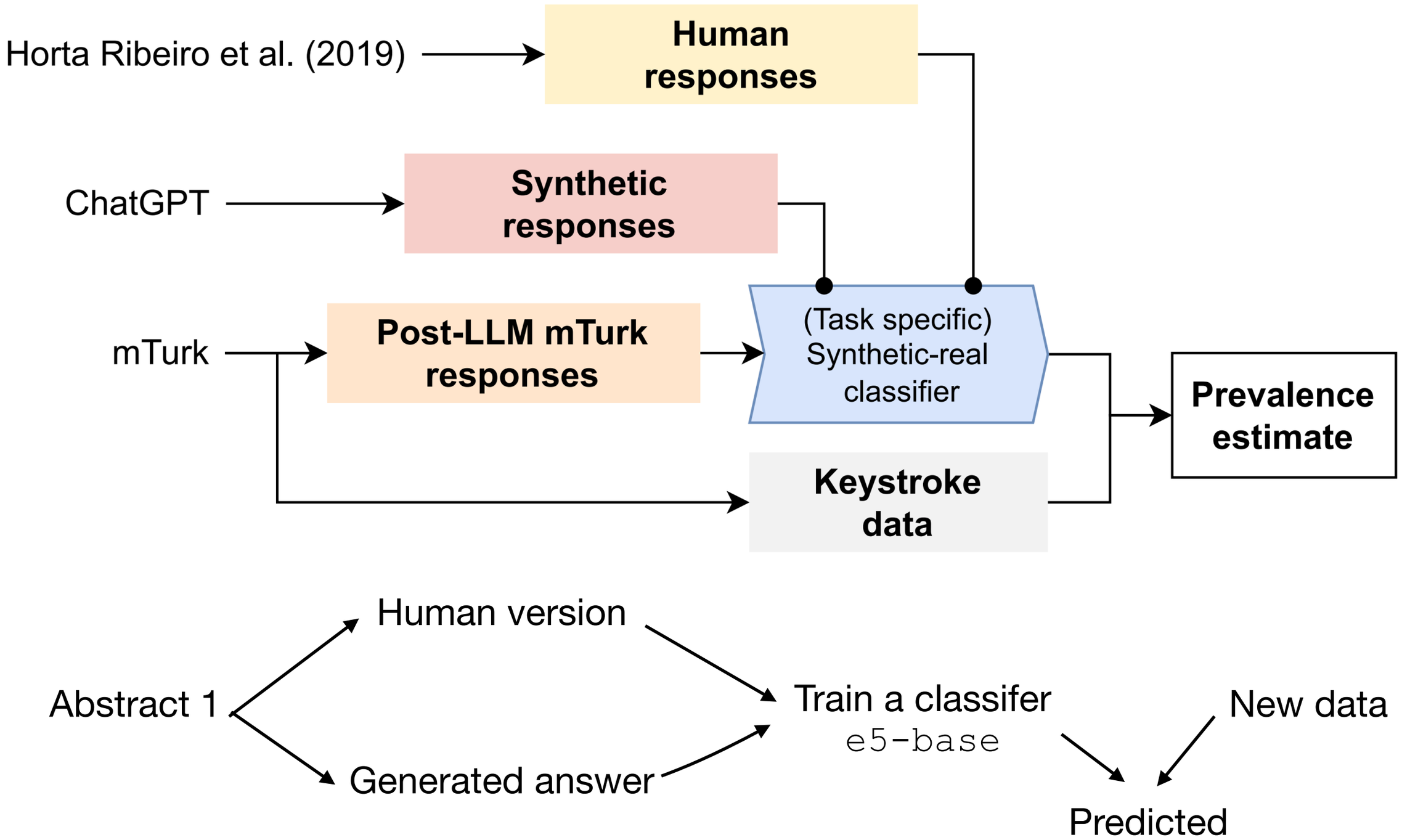


Detection setup





Detection setup



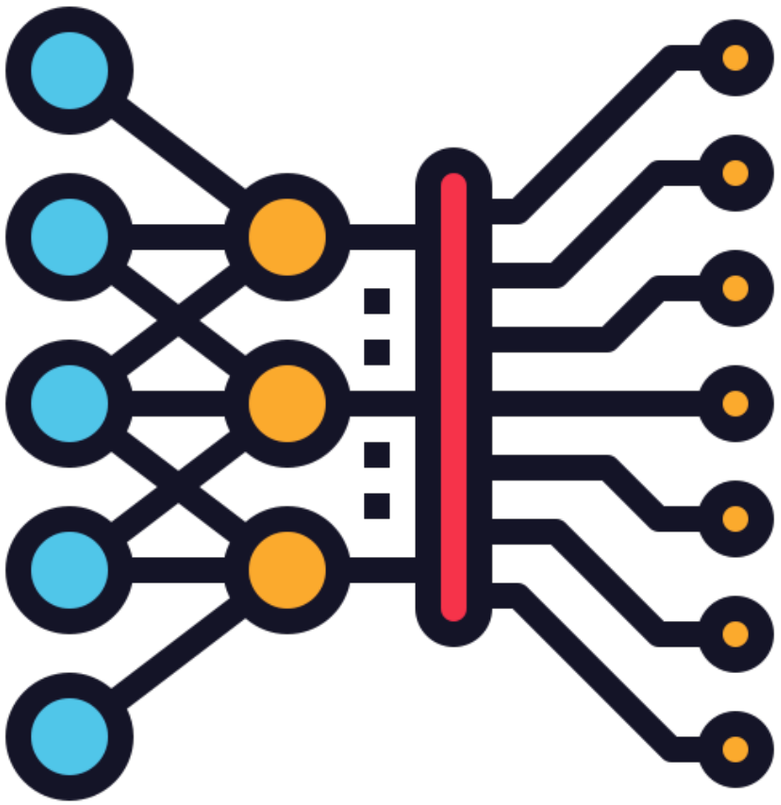
Results

Setting	Accuracy	Macro-F1	Precision	Recall
Summary-level	0.99 ±0.02	0.99 ±0.03	0.99 ±0.02	0.98 ±0.04
Abstract-level	0.97 ±0.03	0.97 ±0.02	0.97 ±0.02	0.97 ±0.04

- Abstract-level: Train on certain abstracts-types, evaluate on others. **For robustness.**
- Summary-level: randomly split summaries.

Detection setup

“In this paper, authors…”



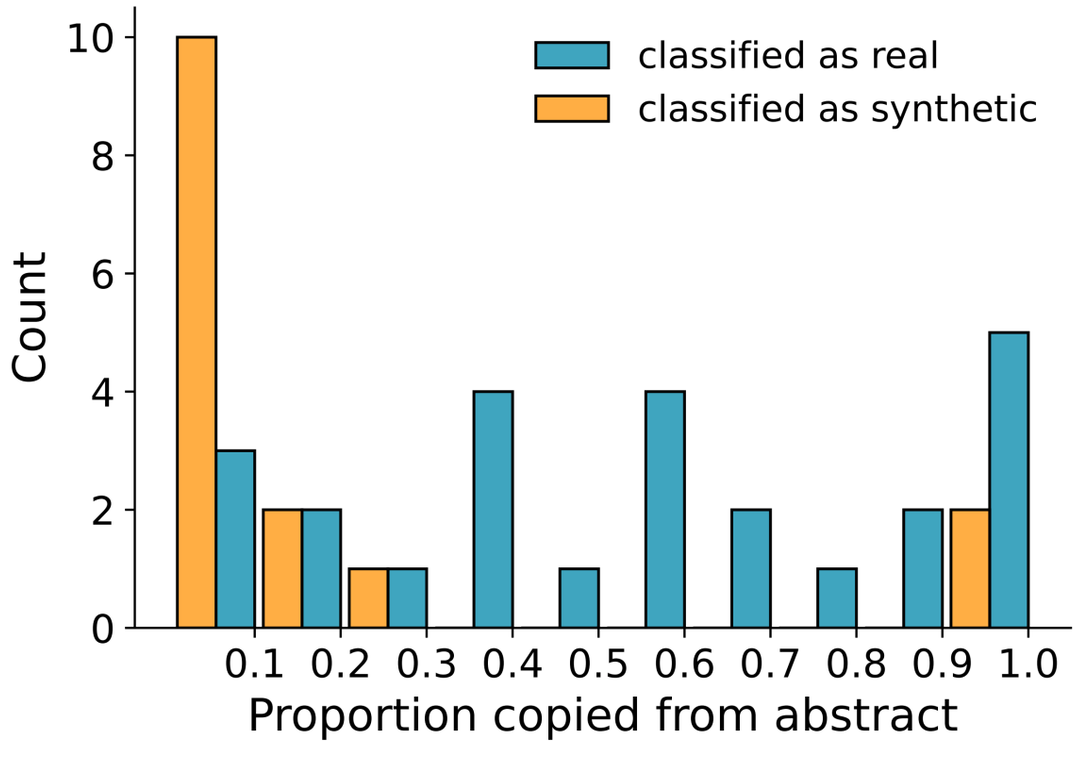
Synthetic

Found 33-46% of crowdworkers used LLMs

Further validation

Used keystroke data to validate that our classifier generalized from 2018 data to 2023 summaries.

	With pasting	Without pasting
Synthetic	15	0
Human	26	5

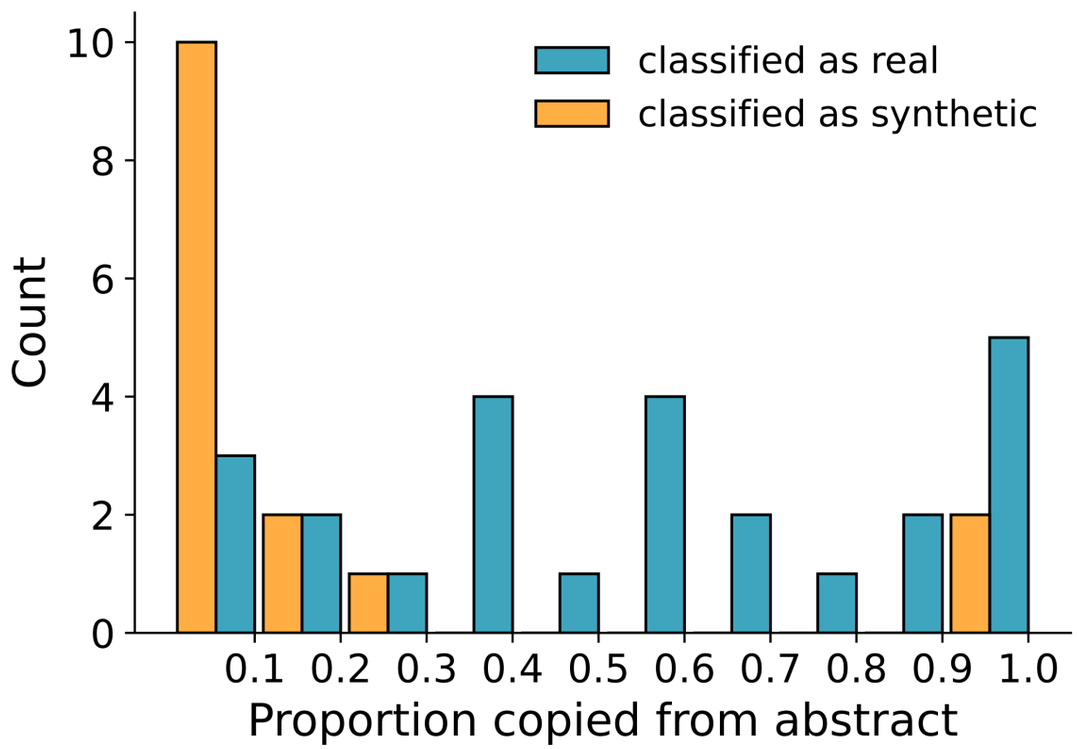


Further validation

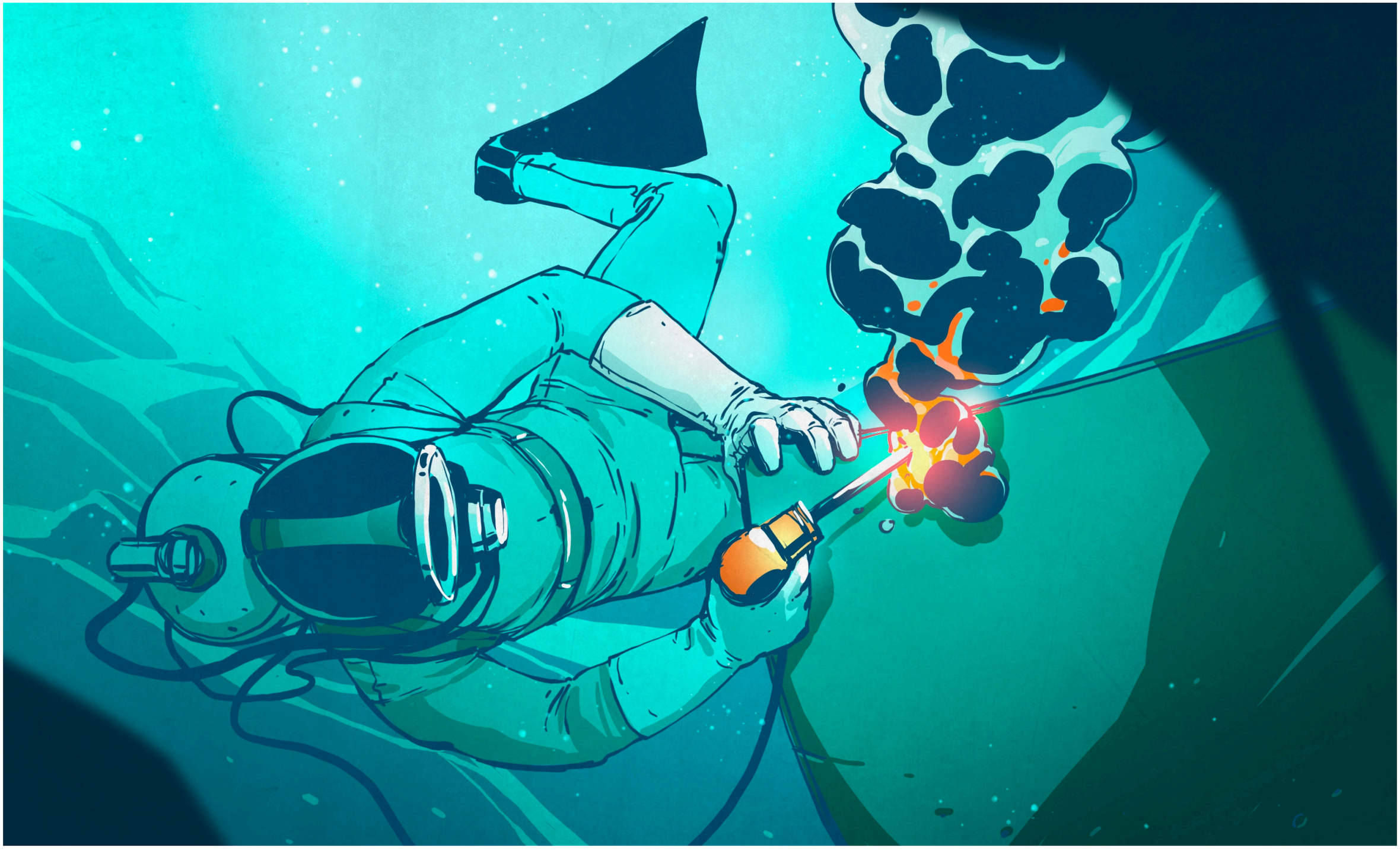
Used keystroke data to validate that our classifier generalized from 2018 data to 2023 summaries.

	With pasting	Without pasting
Synthetic	15	0
Human	26	5

When they didn't paste text into the editor, we predicted all real.



Extractive summarization not a problem



Low-background steel?

Limitations

- **Low pay**. We maintained previous pay rate of $1 / summary which is likely low.
- **Low N** for both original sample and test set
- Type of work (abstract summary is very **automatable**)
- Future: People on other platforms being **aware** of the first A3I paper

Next steps

- Run on more platforms
- Use an ensemble approach with existing classifiers
- Test out policies that can reduce use

Part 2: How well can synthetic data generate linguistic constructs?

Part 2: How well can synthetic data generate linguistic constructs?

Answer: **Not as well as real data.**

Generating Faithful Synthetic Data with Large Language Models

A Case Study in Computational Social Science

Veniamin Veselovsky

Manoel Horta Ribeiro

Akhil Arora

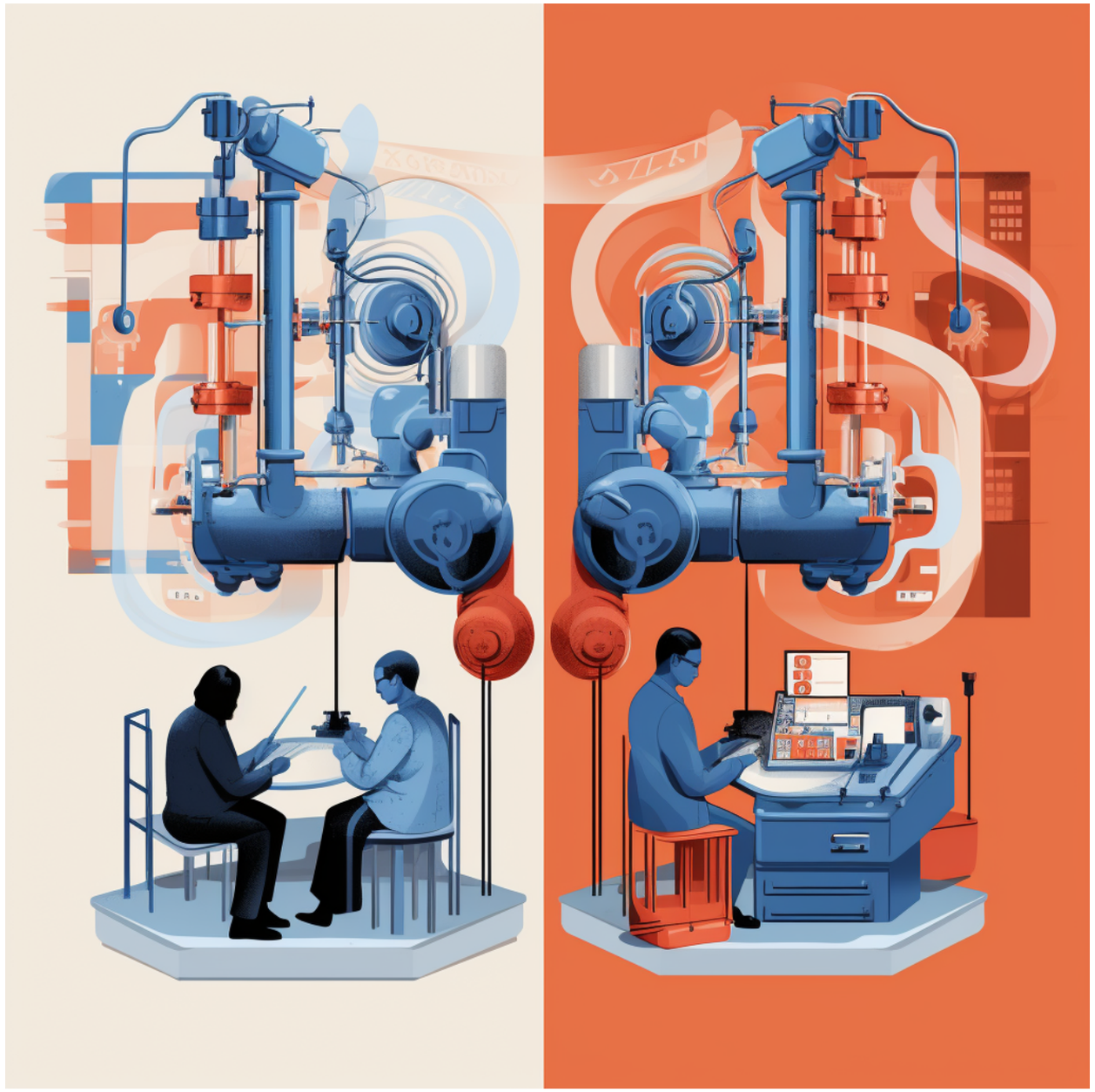
Martin Josifoski

Ashton Anderson

Robert West

EPFL

CSS context

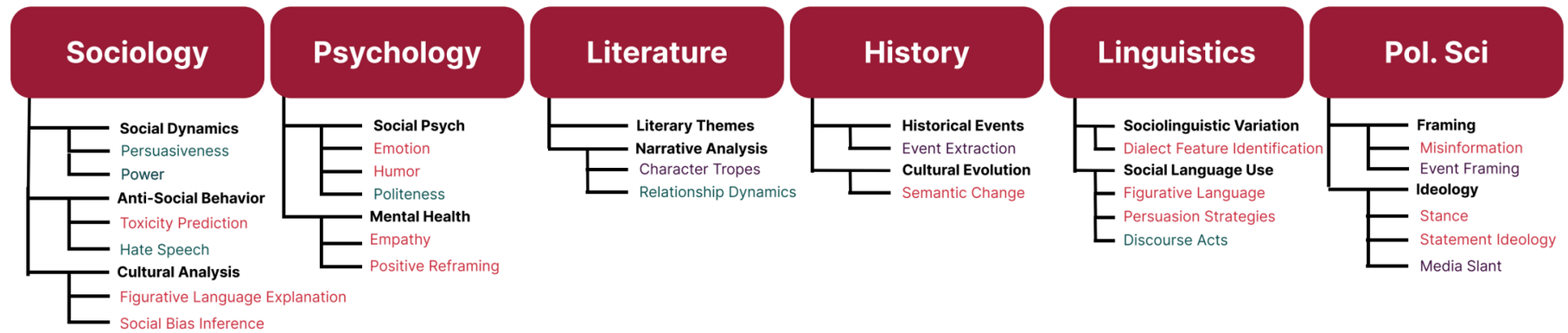


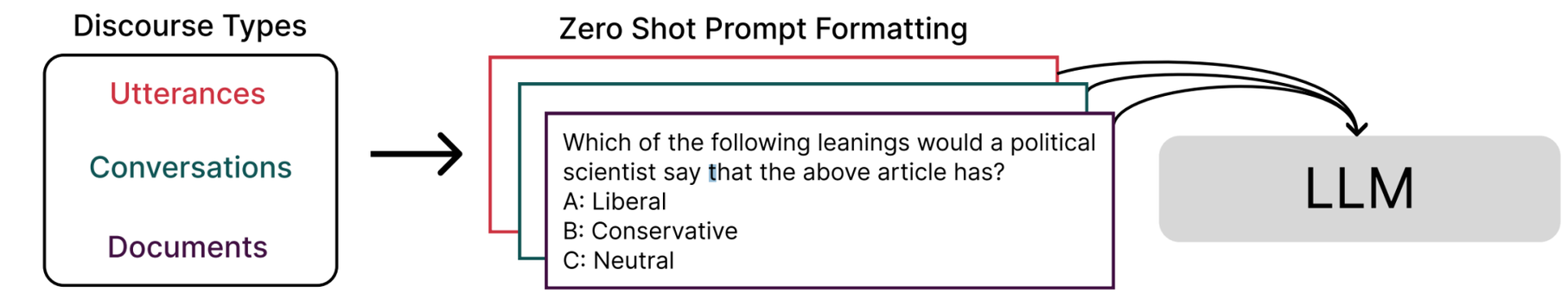
Discriminator (annotation)

Generator (silicon sampling)

Annotations

- ChatGPT and GPT-4 are *sometimes* better annotators than experts [4,5]





Not clear how fair the comparisons are.

Generations

- LLMs can also be used to generate data *de-novo* [6,7,8]

Generations

- LLMs can also be used to generate data *de-novo* [6,7,8]
- Can be used to survey people
    - Telling the model it's Republican makes it vote like a Republican.
    - Companies like Synthetic Users forming.

Generations

- LLMs can also be used to generate data *de-novo* [6,7,8]
- Can be used to survey people
    - Telling the model it's Republican makes it vote like a Republican.
    - Companies like Synthetic Users forming.
- How faithful is the synthetic data? How can we make it as faithful as possible?

Generations

- This can open up a new paradigm for studying linguistic constructs in text.

Select a construct

sarcasm

Generations

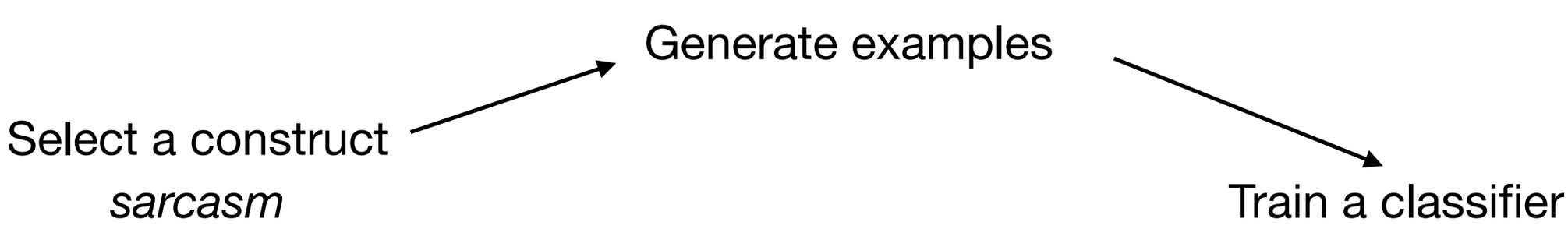
- This can open up a new paradigm for studying linguistic constructs in text.

Select a construct

sarcasm

Generations

- This can open up a new paradigm for studying linguistic constructs in text.



Generations

- This can open up a new paradigm for studying linguistic constructs in text.

Select a construct

sarcasm

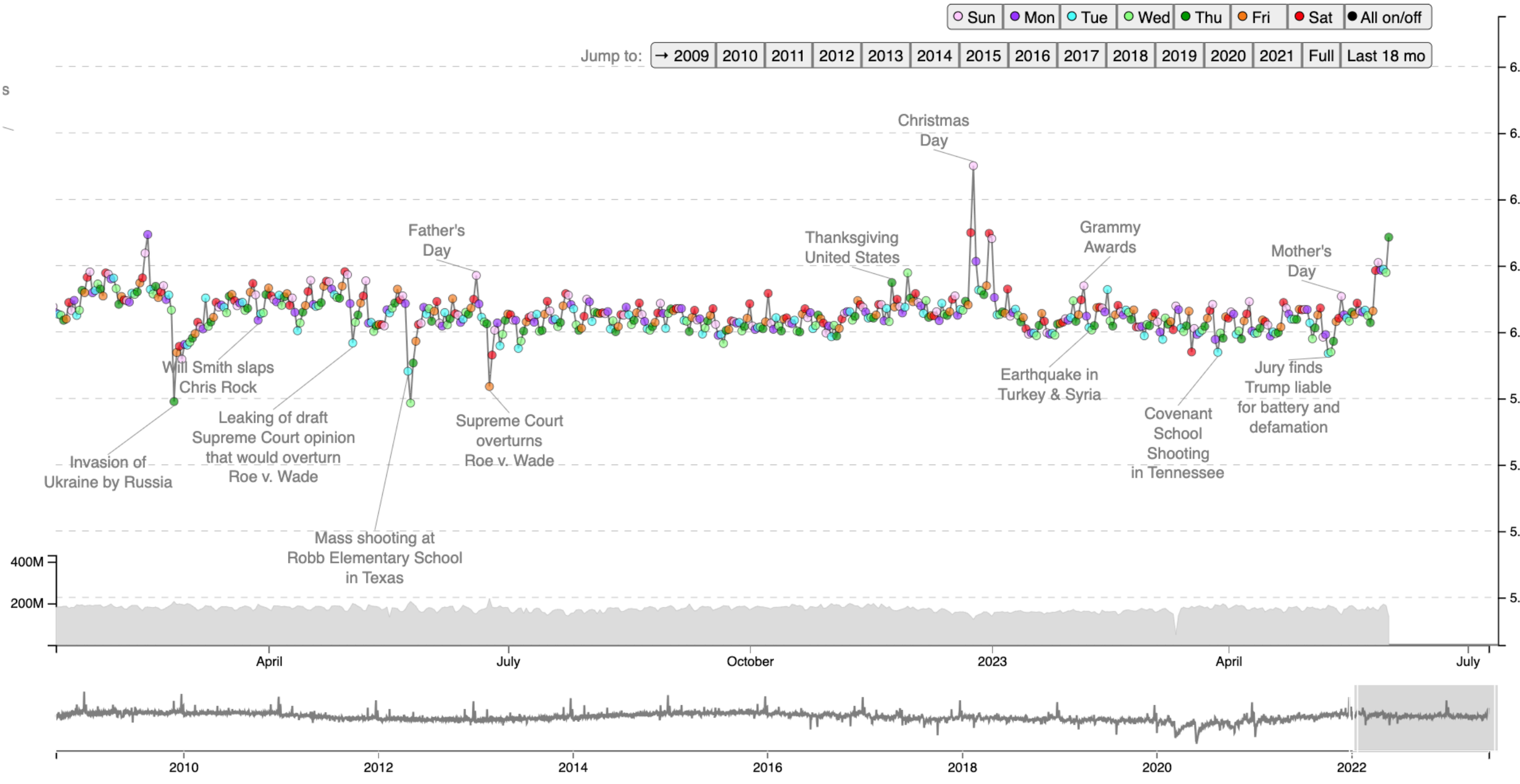
Generate examples

Train a classifier

Run on real world data

Average Happiness for Twitter

All Tweets in English.



Setup

- Fix a linguistic construct (sarcasm and sentiment).

Setup

- Fix a linguistic construct (sarcasm and sentiment)
- Apply a few generation strategies to generate examples of this text.

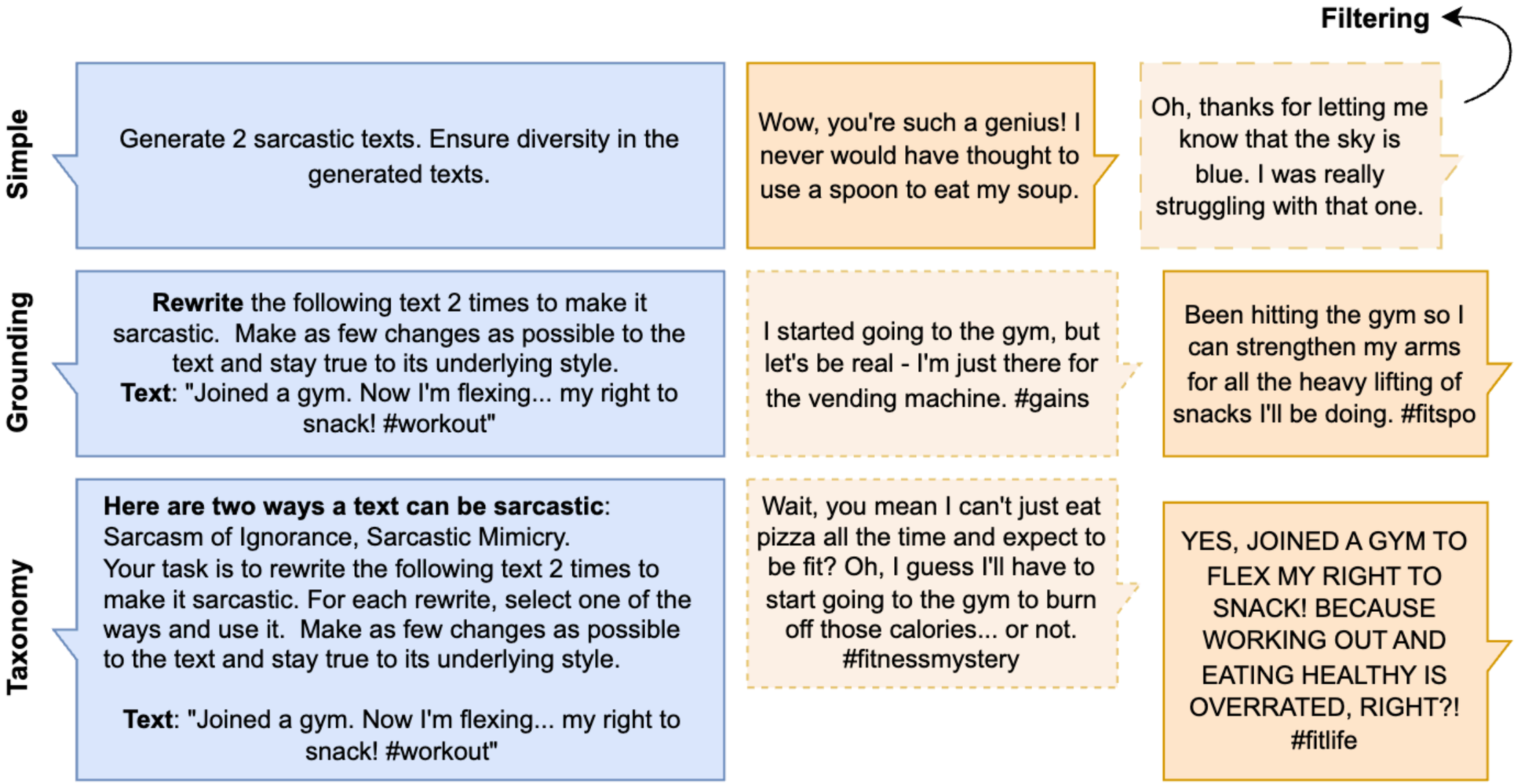
Setup

- Fix a linguistic construct (sarcasm and sentiment)
- Apply a few generation strategies to generate examples of this text.
- Train a model on real vs. synthetic data.

Setup

- Fix a linguistic construct (sarcasm and sentiment)
- Apply a few generation strategies to generate examples of this text.
- Train a model on real vs. synthetic data.
- Compare performances on real evaluation task.

Prompting strategies



Results

Strategy	*Sarcasm*			*Sentiment*		
	Accuracy	Macro-F1	Believability	Accuracy	Macro-F1	Believability
Simple	**0.71**	0.48	0.04	0.42	0.30	—
Grounding	0.67	**0.55**	0.13	0.62	0.59	—
Grounding + Taxonomy	0.67	0.51	0.20	0.42	0.32	—
Groundtruth annotations	0.72	0.60	0.95	0.70	0.70	—

Results

Strategy	*Sarcasm*			*Sentiment*		
	Accuracy	Macro-F1	Believability	Accuracy	Macro-F1	Believability
Simple	**0.71**	0.48	0.04	0.42	0.30	—
Grounding	0.67	**0.55**	0.13	0.62	0.59	—
Grounding + Taxonomy	0.67	0.51	0.20	0.42	0.32	—
Groundtruth annotations	0.72	0.60	0.95	0.70	0.70	—

Grounding in real data works best
Improves on diversity.

Results

Strategy	*Sarcasm*			*Sentiment*		
	Accuracy	Macro-F1	Believability	Accuracy	Macro-F1	Believability
Simple	**0.71**	0.48	0.04	0.42	0.30	—
Grounding	0.67	**0.55**	0.13	0.62	0.59	—
Grounding + Taxonomy	0.67	0.51	0.20	0.42	0.32	—
Groundtruth annotations	0.72	0.60	0.95	0.70	0.70	—

Grounding in real data works best
Improves on diversity.

Taxonomy generation and grounding doesn't improve results

Results

Strategy	*Sarcasm*			*Sentiment*		
	Accuracy	Macro-F1	Believability	Accuracy	Macro-F1	Believability
Simple	**0.71**	0.48	0.04	0.42	0.30	—
Grounding	0.67	**0.55**	0.13	0.62	0.59	—
Grounding + Taxonomy	0.67	0.51	0.20	0.42	0.32	—
Groundtruth annotations	0.72	0.60	0.95	0.70	0.70	—

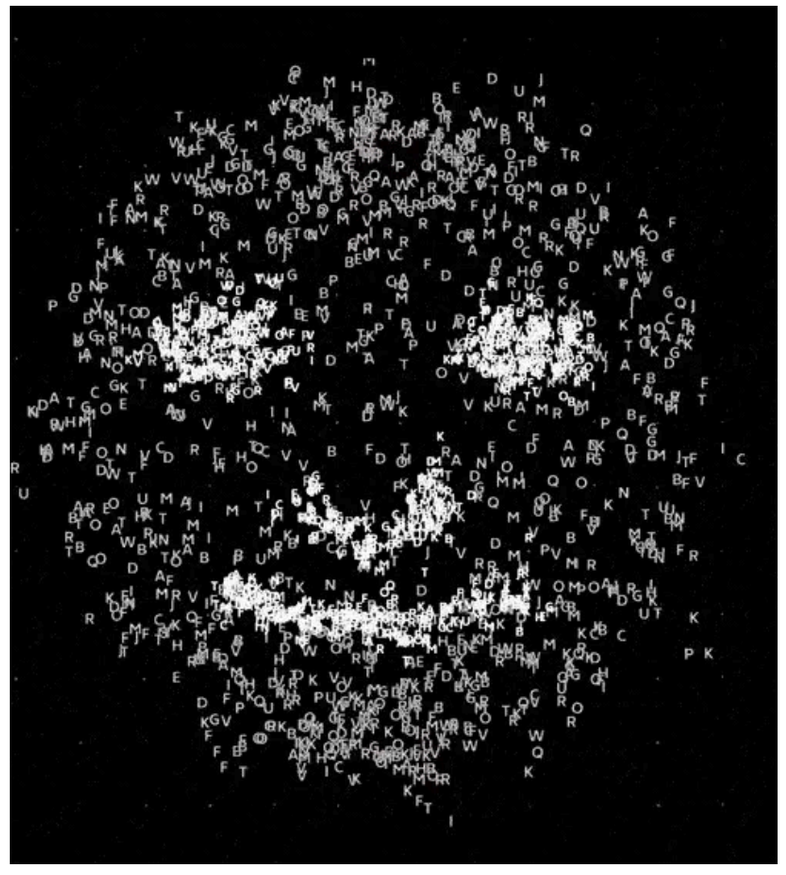
Grounding in real data works best
Improves on diversity.

Taxonomy generation and grounding doesn't improve results

The original real data outperforms synthetic data

Conclusions

- Grounding in real text is important.
- Prompting techniques only give us lower bounds.
- Synthetic data doesn’t work as well as human text. But we can still get some mileage.



Key takeaways

- Crowdworkers are using ChatGPT. Be thoughtful about how you define your studies.

Key takeaways

- Crowdworkers are using ChatGPT. Be thoughtful about how you define your studies.

- This can lead to model collapse and lack of generalizability.

Key takeaways

- Crowdworkers are using ChatGPT. Be thoughtful about how you define your studies.

- This can lead to model collapse and lack of generalizability.

- LLMs are good at annotating some tasks.

Key takeaways

- Crowdworkers are using ChatGPT. Be thoughtful about how you define your studies.

- This can lead to model collapse and lack of generalizability.

- LLMs are good at annotating some tasks.

- Synthetically generated data can create classifiers.

Key takeaways

- Crowdworkers are using ChatGPT. Be thoughtful about how you define your studies.

- This can lead to model collapse and lack of generalizability.

- LLMs are good at annotating some tasks.

- Synthetically generated data can create classifiers.

- Not as good as real human text.

EP

Thank you!

Happy to take any questions!

Interactive coding

- Use Langchain
- Annotate data
- Generate data

References

[0] Alemohammad, Sina, et al. "Self-Consuming Generative Models Go MAD." *arXiv preprint arXiv:2307.01850* (2023).

[1] Martínez, Gonzalo, et al. "Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet." arXiv preprint arXiv:2306.06130 (2023).

[2] Shumailov, Ilia, et al. "The Curse of Recursion: Training on Generated Data Makes Models Forget." arXiv preprint arxiv:2305.17493 (2023).

[3] Hartmann, Jochen, Jasper Schwenzow, and Maximilian Witte. "The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation." arXiv preprint arXiv:2301.01768 (2023).

[4] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056.

[5] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can Large Language Models Transform Computational Social Science?. arXiv preprint arXiv:2305.03514.

[6| Eldan, R., & Li, Y. (2023). TinyStories: How Small Can Language Models Be and Still Speak Coherent English?. arXiv preprint arXiv:2305.07759.

[7] Josifoski, M., Sakota, M., Peyrard, M., & West, R. (2023). Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. arXiv preprint arXiv:2303.04132.

Detection setup

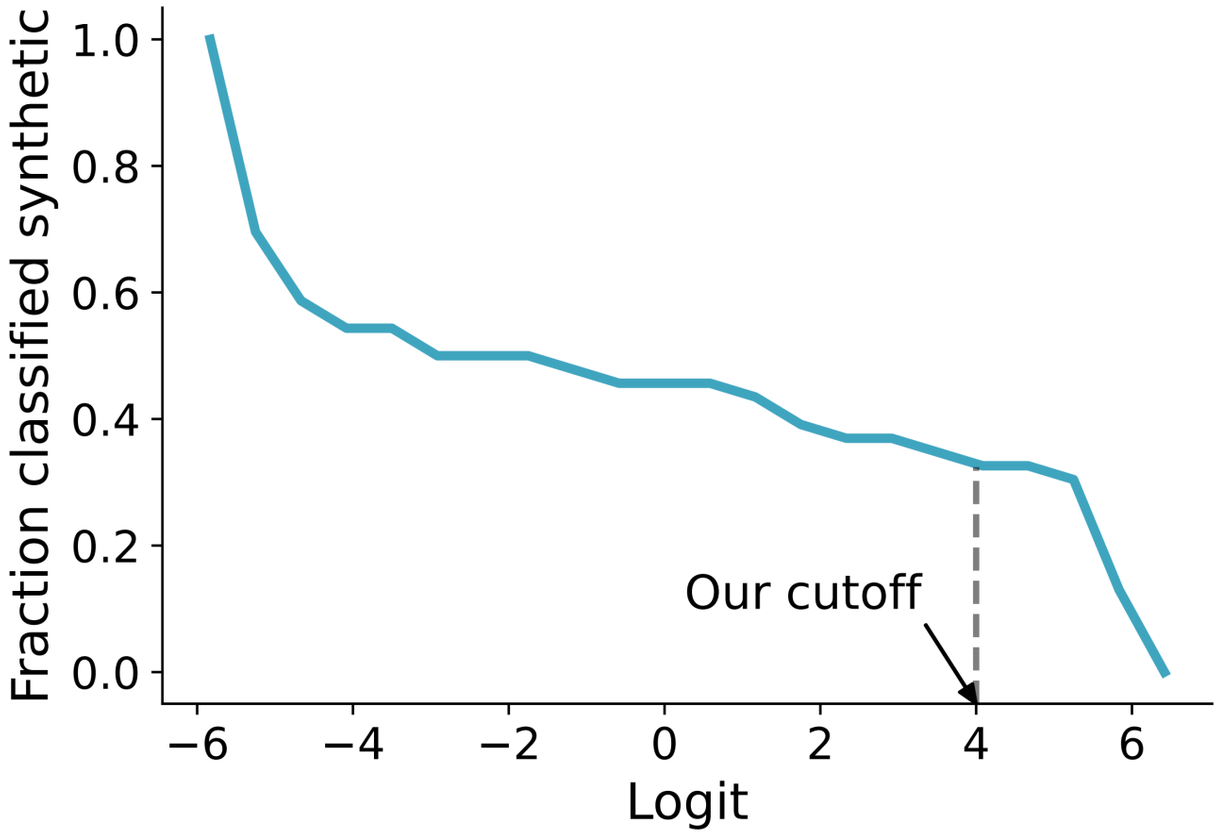


Figure 3: Proportion of summaries predicted as synthetic depending on the logit threshold.