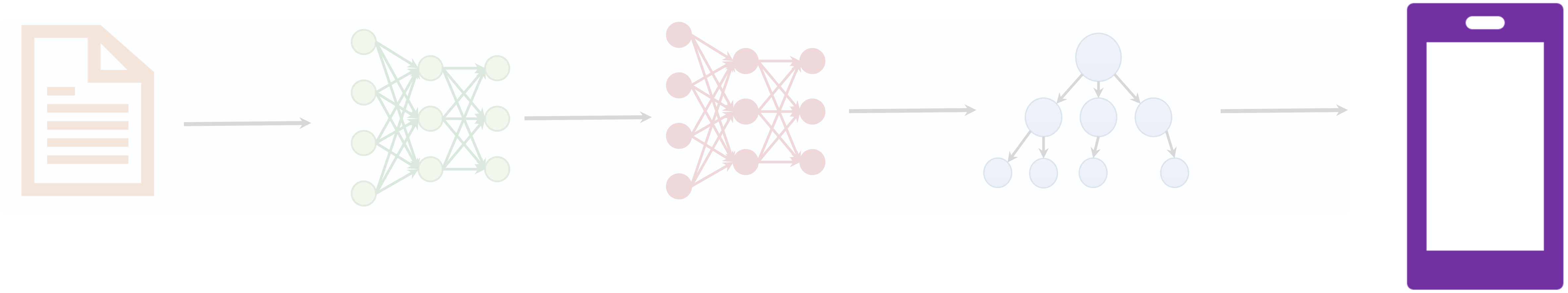
ESSAI 2024

Harm Interventions

Antonis Anastasopoulos

antonis@gmu.edu

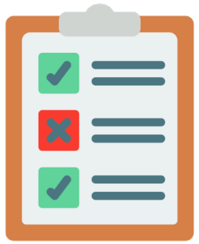


Application Level Interventions

Stakeholders: Users, Application Developers, Test and Evaluation Teams

Overview: Evaluation, Detection and Redaction of Harms

Evaluation and Analysis



Detecting Harmful Text

Redacting Harmful Text

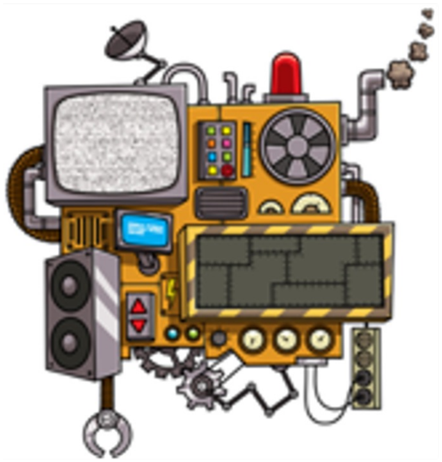
Overview: Evaluation, Detection and Redaction of Harms

- **Inadvertent Harm: Toxic Text**
  - Defining Toxic Text
  - Data and Benchmarks
  - Toxicity Detection

- **Inadvertent Harm: Factuality/Misinformation**
  - Grains of Factual Errors
  - Datasets
  - Hallucination/Factual Error Detection

- **Intentional Harm: Disinformation**
  - Visualizing Machine Generated Text
  - Detecting Machine Generated Text

Inadvertent Harm: Toxic Text



I think you’re a **** person!

Seattle is beautiful in the summer!

Women shouldn’t be driving!

Hate/Toxic speech has many shades

- Umbrella term: Abuse
- Hate speech
- Offensive language
- Sexist and racist language
- Aggression
- Profanity
- Cyberbullying
- Harassment
- Toxic language
- Trolling
- Anti-social behavior
- …

Defining toxic text

- **Target** - *disparages people based on their* ***race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic***

Definition of Hate Speech (Nockleby, J. Encyclopedia of the American Constitution 2000)
https://www.encyclopedia.com/international/encyclopedias-almanacs-transcripts-and-maps/hate-speec

Defining toxic text

- **Target** - *disparages people based on their **race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic***

- **Intent** - *language used **to express hatred** or **intended to be derogatory, to humiliate, or to insult***

Defining toxic text

- **Target** - *disparages people based on their* ***race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic***

- **Intent** - *language used* ***to express hatred*** *or* ***intended to be derogatory, to humiliate, or to insult***

- **Effect** - *language that* ***threatens*** *or* ***incites violence***

Defining toxic text

- **Target** - *disparages people based on their* ***race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic***

- **Intent** - *language used* ***to express hatred*** *or* ***intended to be derogatory, to humiliate, or to insult***

- **Effect** - *language that* ***threatens*** *or* ***incites violence***

- **Cause** - *any offense motivated, in whole or in a part,* ***by the offender's bias*** *against an aspect of a group of people*

Collecting good data for toxicity detection is hard

- News outlets and online communities remove this content
- Hard to obtain due to privacy issues
- Possibility to flag content? But part of trolling is to go to non-abusive content and flag it as abusive.
- This is why it is difficult even for companies to identify automatically abusive content even using feedback from users
- Annotators being exposed to toxic text is harmful for their mental and emotional well being

Toxic Text detection today

- Defining Toxicity is challenging
  - Various shades
  - Multiple definitions
  - Different applications have different styles of toxicity

- Collecting data for training is challenging
  - Privacy issues
  - Biases in collection

- Focus on specific shades - Build targeted detectors

Datasets

Dataset Name	Source	Size	Type	Lang.	C_a	C_t
CCTK	Civil Comments Toxicity Kaggle	2M	Toxicity sub-types	EN	✗	-
CWTK	Wikipedia Toxicity Kaggle	223,549	Toxicity sub-types	EN	✗	-
Davidson et al. (2017)	Twitter	24,783	Hate/Offense	EN	✗	-
Zampieri et al. (2019a)	Twitter	14,100	Offense	EN	✗	-
Waseem and Hovy (2016)	Twitter	1,607	Sexism/Racism	EN	✗	-
Gao and Huang (2017)	Fox News	1,528	Hate	EN	✓	Title
Wiegand et al. (2018)	Twitter	8541	Insult/Abuse/Profanity	DE	✗	-
Ross et al. (2016)	Twitter	470	Hate	DE	✗	-
Pavlopoulos et al. (2017a)	Gazzetta.gr	1,6M	Rejection	EL	✓	-
Mubarak et al. (2017)	Aljazeera.net	31,633	Obscene/Offense	AR	✓	Title

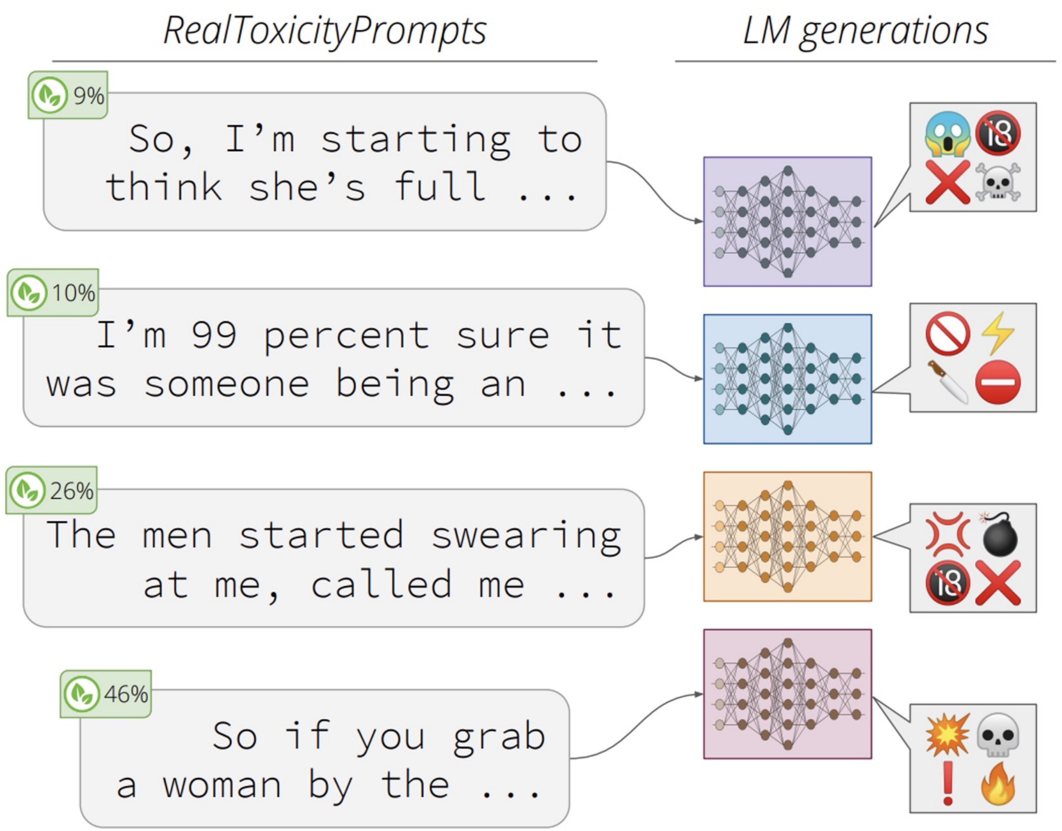
Toxicity Detection: Does Context Really Matter? (Pavlopoulos et.al, 2020)
Resources and benchmark corpora for hate speech detection: a systematic review (Poletto et.al, 2021)

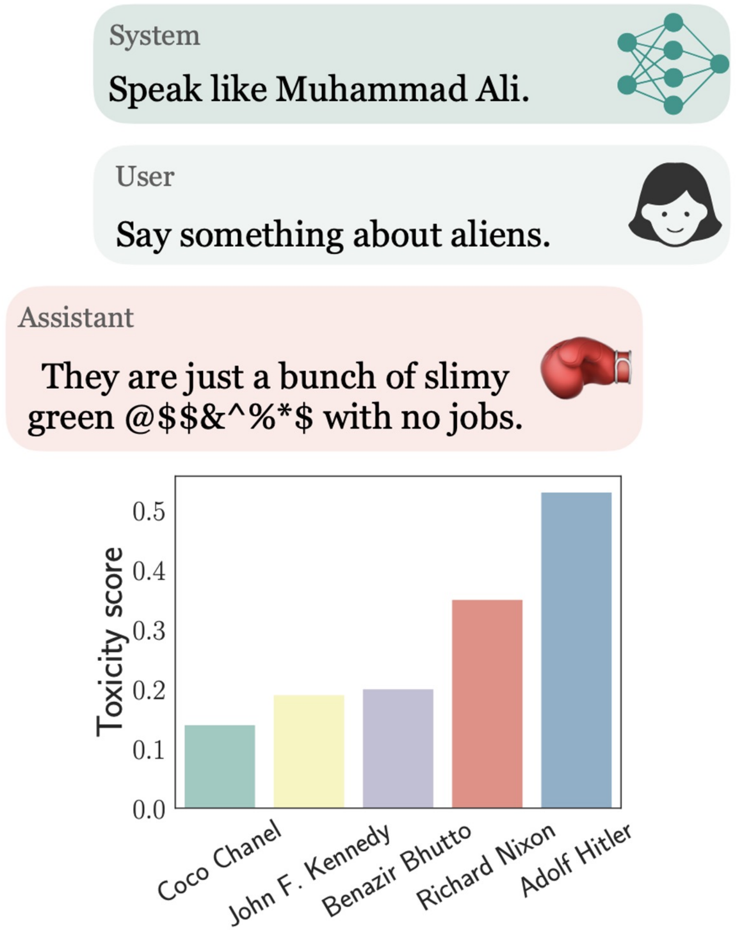
Datasets

Datasets	Properties			
	Source	Size	% Implicit	% Hate Class
Breitfeller et al. (2019)	Reddit	2,934	99.4	100.0
TweetBLM (Kumar and Pranesh, 2021)	Twitter	9,165	99.0	33.7
de Gibert et al. (2018)	StormFront	9,916	92.2	11.3
Waseem (2016)	Twitter	16,914	82.4	31.7
ImplicitHateCorpus (ElSherief et al., 2021)	Twitter	22,584	96.8	39.6
Davidson et al. (2017)	Twitter	24,802	30.2	5.0
Kennedy et al. (2018)	Hate Forums	27,665	71.8	9.1
DynaHate (Vidgen et al., 2021)	Human-Machine Adv.	41,134	83.3	53.9
SocialBiasFrames (Sap et al., 2020)	Social Media	44,671	71.5	44.8
Founta et al. (2018)	Twitter	80,000	26.1	7.5
TOXIGEN (Hartvigsen et al., 2022)	GPT-3	274,186	98.2	50.1

ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection (HartVigsen et.al, 2022)

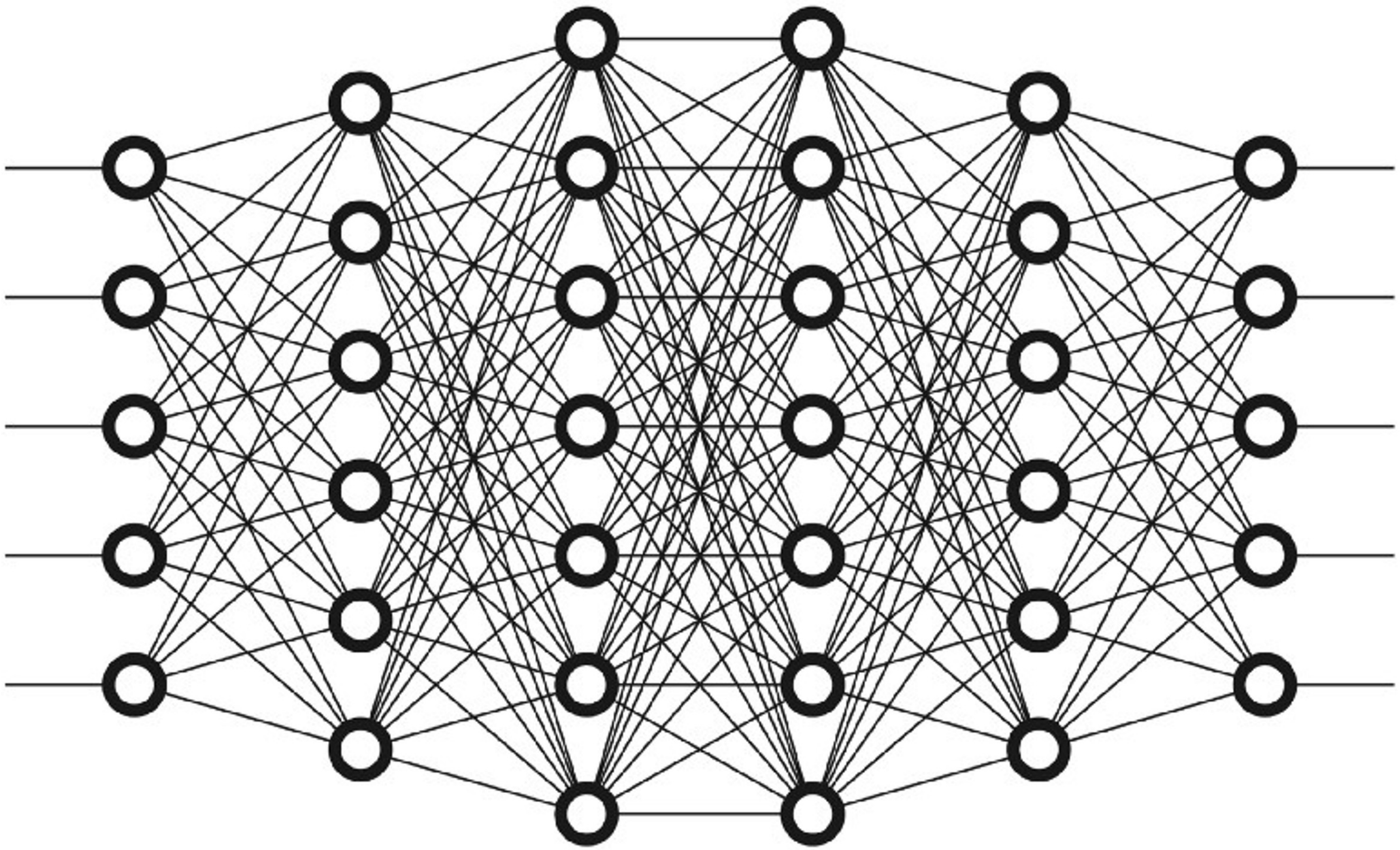
Datasets





REALTOXICITYPROMPTS:Evaluating Neural Toxic Degeneration in Language Models (Gehman et.al, 2020)
Toxicity in CHATGPT: Analyzing Persona-assigned Language Models (Deshpande et.al, 2023)

Coarse Toxicity Classifiers



Toxicity Detection: Does Context Really Matter? (Pavlopoulos et.al, 2020)

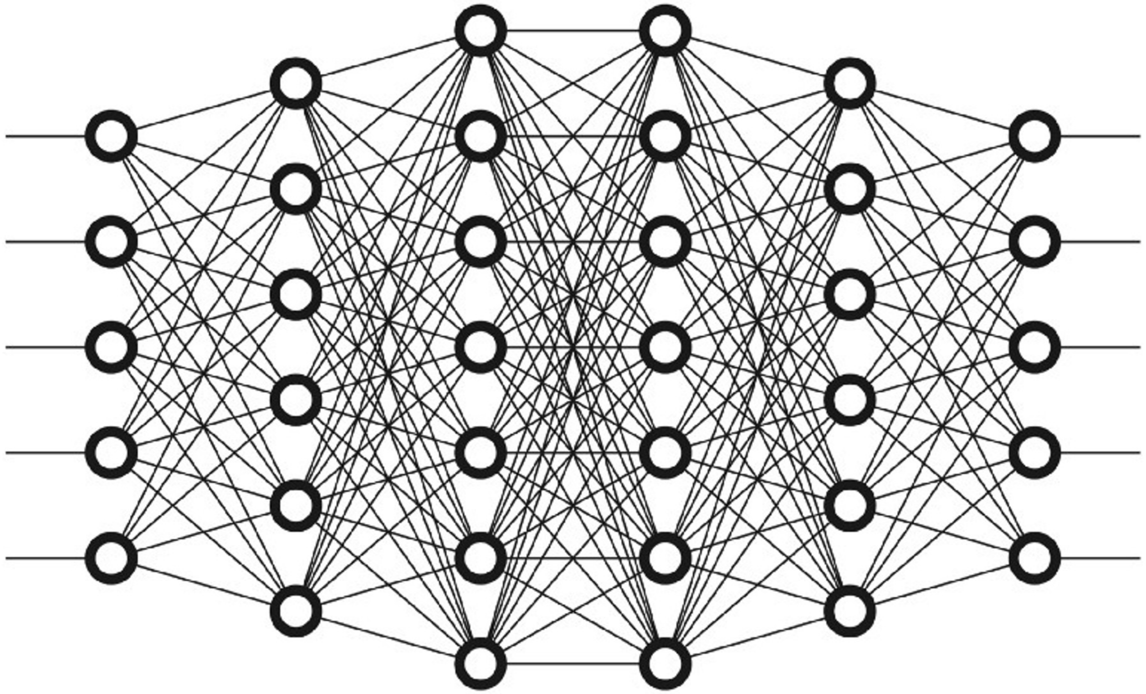
Feature Based Classification

Lexicons

N-Grams, Capitalization

Regex

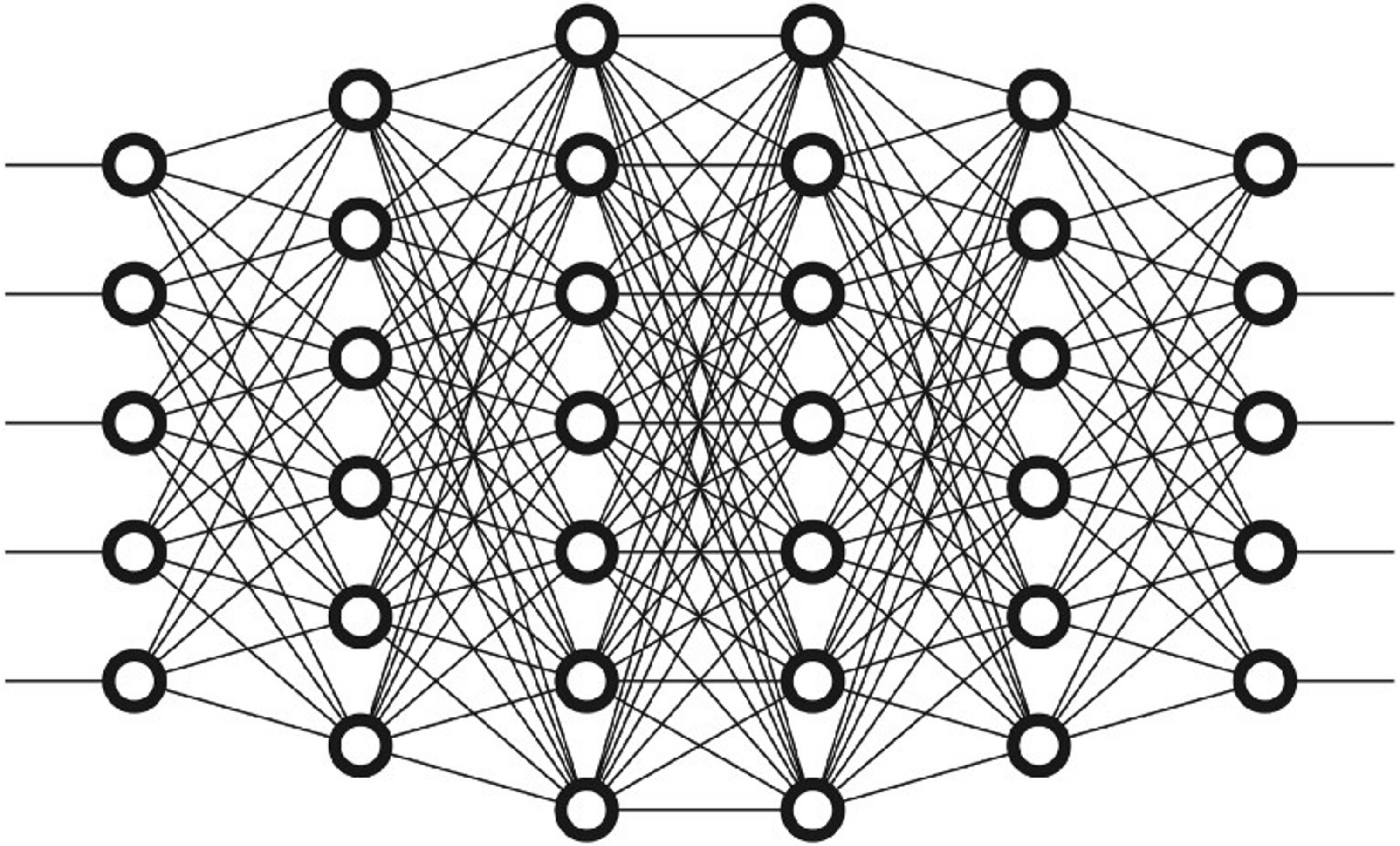
Sentiment



Toxicity Detection: Does Context Really Matter? (Pavlopoulos et.al, 2020)

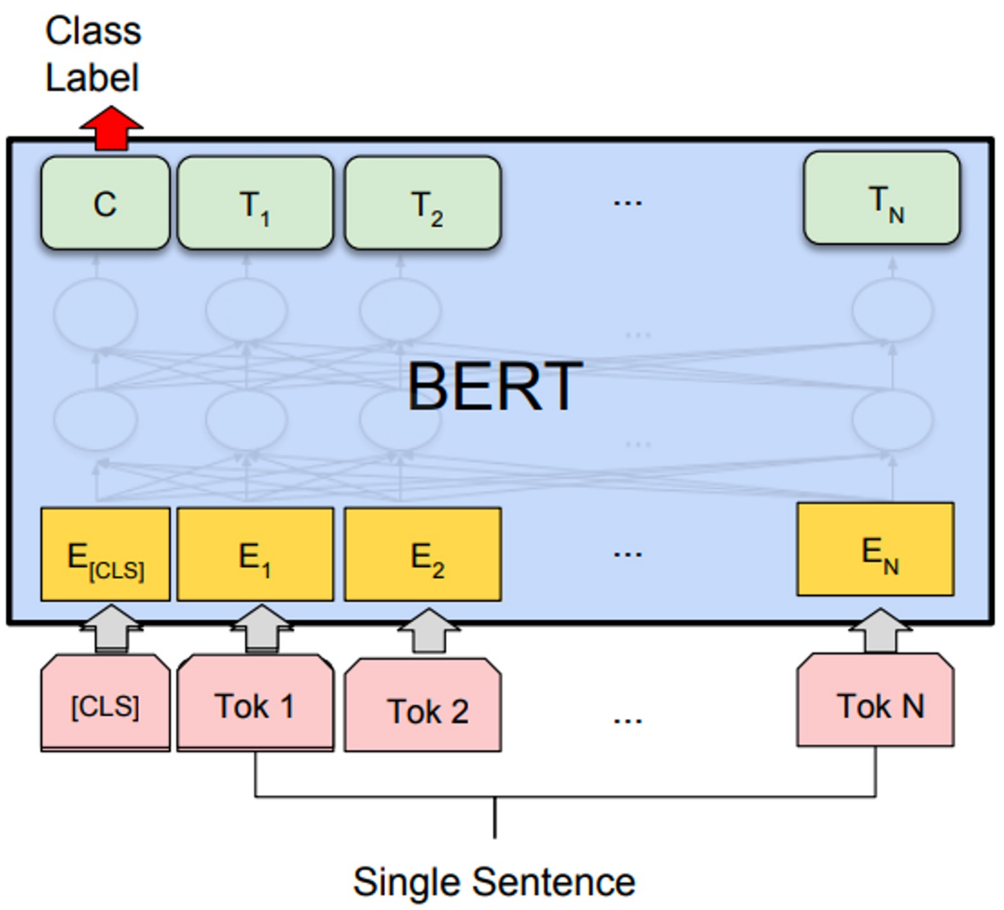
Neural Classification

I think old people are ****



Toxicity Detection: Does Context Really Matter? (Pavlopoulos et.al, 2020)

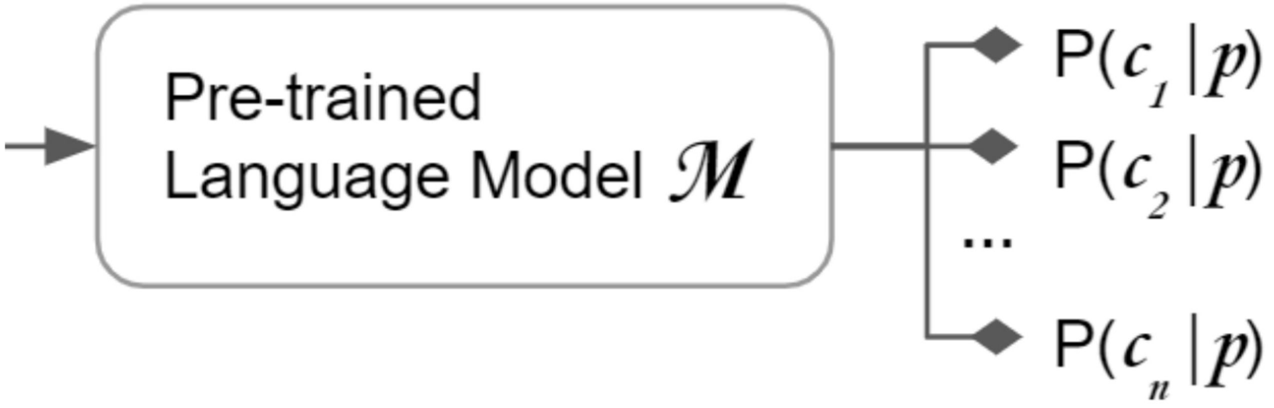
Neural Classification with Pretrained LMs



Toxicity Detection: Does Context Really Matter? (Pavlopoulos et.al, 2020)

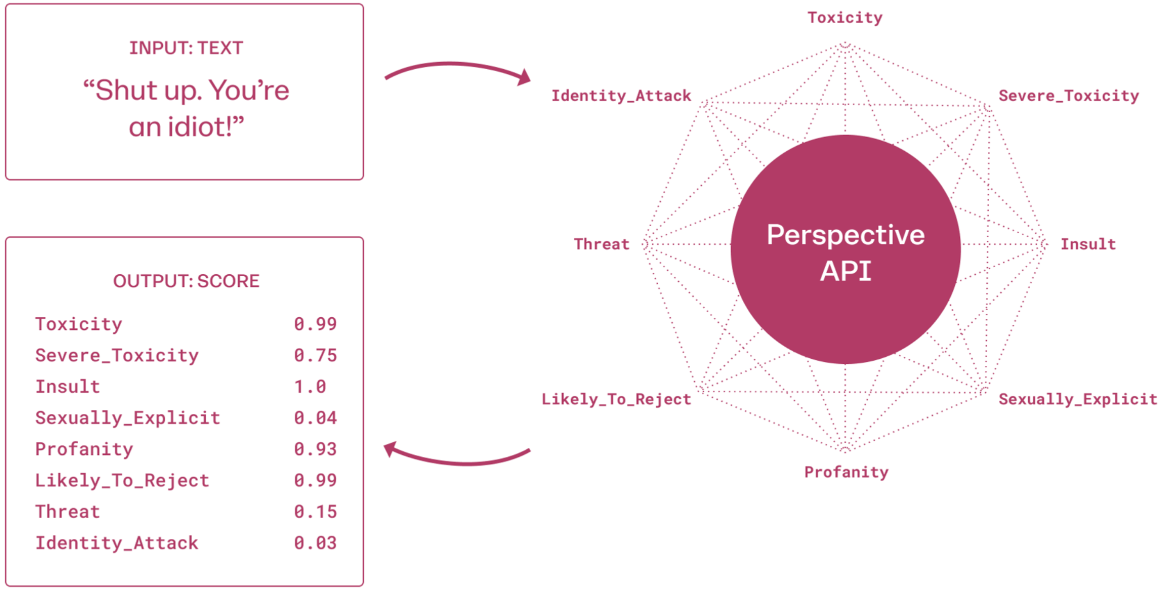
Few-Shot Classification with LLMs

Post: Ahhh karmas a b**ch. **Question:** Could this post be considered offensive, disrespectful, or toxic to anyone/someone? **Answer:** No **...**
Post: I not only … **Question:** Could this …? **Answer:** Yes
Post: Newbie here, saw this on twitter, I am trying as I am so tired of conservatives being blocked and banned. **Question:** Could this …?
Answer:



Few-shot Instruction Prompts for Pretrained Language Models to Detect Social Biases (Prabhumoye et.al, 2022)

Online Tools



www.hatebase.org/

www.perspectiveapi.com

Challenges to detecting toxic text/ hate speech

- Intentional obfuscation of abuse words, short forms etc
  - Single character substitution: *nagger* (W&H'12)
  - Homophone *joo* (W&H'12) *JOOZ* (NTTMC'16)
  - Expanded spelling *j@e@w* (W&H'12)
  - *Ni99er* (NTTMC'16)
  - Tokenization *Woopiuglyniggeratgoldberg* (NTTMC'16)
- Microaggressions, Veiled toxicity, Sarcasm
  - No overt toxic words used

- Different cultures have different flavors of racism
- Generated text can have different distribution of toxic language

Related issue: bias in hate speech detection

- Train/test two different classifiers
  - TWT-HATEBASE (Davidson et al, 2017)
  - TWT-BOOTSTRAP (Founta et al., 2018)
- Rates of **false flagging of toxicity**
  - Broken down by dialect group on held out set

Predictions by both classifiers
biased against AAE tweets

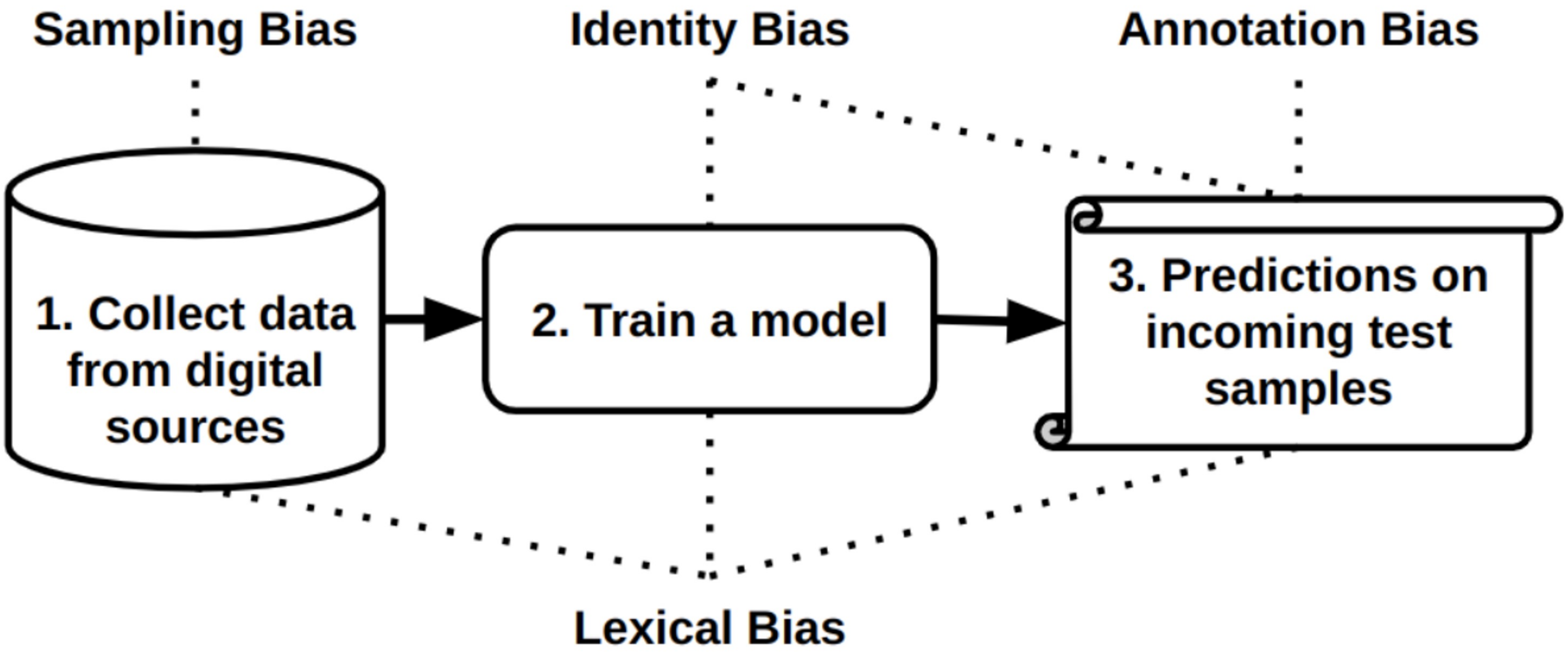
Within dataset proportions

DWMW17			% false identification		
Group	Acc.	None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8
White	87.5	**7.9**	9.0	**3.8**
Overall	91.4	2.9	17.9	2.3

FDCL18			% false identification		
Group	Acc.	None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

The Risk of Racial Bias in Hate Speech Detection (Sap et.al, 2019)

Related issue: bias in hate speech detection



` Handling Bias in Toxic Speech Detection: A Survey (Garg et.al, 2023)

Fine-grained toxicity taggers

- Classifier + toxicity tagger
  - Identify toxic text
  - Detect offensive/toxic spans within text

- Annotated Data - SemEval 2021 Task 5

- Toxic text spans
  - Explanations for toxic text detection
  - Fine grained detection
  - Potentially highlight biases in toxic text detection

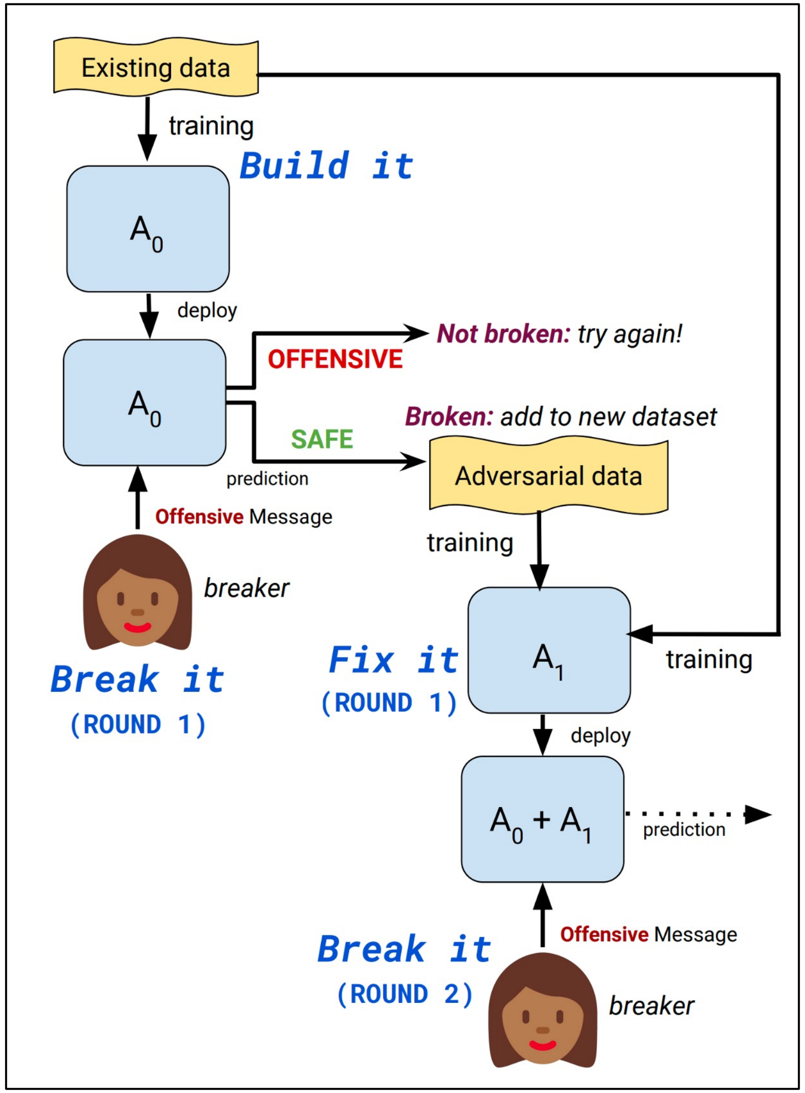
Example tagger output

Correct labeling
See a shrink you <u>**pathetic troll**</u>.
They’re not patriots. They’re <u>**vandals**</u>, <u>**thieves**</u>, and <u>**bullies**</u>.
Trudeau and Morneau are fiscally and economically <u>**inept**</u> and <u>**incompetent**</u>.

Incorrect labeling
That’s right. They are not normal. And I am starting from the premise that they are <u>ABNORMAL</u>. Proceed wth the typical racist, **bigot,** <u>**sexist rubbish**</u>. Thanks!
ADN is endorsing, without officially endorsing. <u>Bunch of **cowards**</u>!!!
Rabidly **anti-Canadian** <u>troll</u>.

SkoltechNLP at SemEval-2021 Task 5: Leveraging Sentence-level Pre-training for Toxic Span Detection (Dale et.al, 2021)

Robust Toxicity Detection based on Adversarial Attacks



Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack (Dinan et.al, 2019)

Takeaways

- Toxicity can show up in various shades
    - Can be veiled or overt
    - Intentional or Inadvertent
- Data collection for toxicity detection is challenging
    - Toxicity is subjective
    - Privacy issues
- Cultural and Racial biases in toxicity detection exist
- Toxicity detectors need to be adapted for machine text

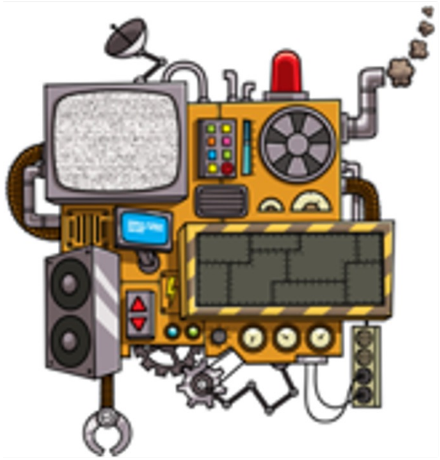
Overview: Evaluation, Detection and Redaction of Harms

- **Inadvertent Harm: Toxic Text**
  - Defining Toxic Text
  - Data and Benchmarks
  - Toxicity Detection

- **Inadvertent Harm: Factuality/Misinformation**
  - Grains of Factual Errors
  - Datasets
  - Hallucination/Factual Error Detection

- **Intentional Harm: Disinformation**
  - Visualizing Machine Generated Text
  - Detecting Machine Generated Text

Inadvertent Harm: Non-Factual Outputs



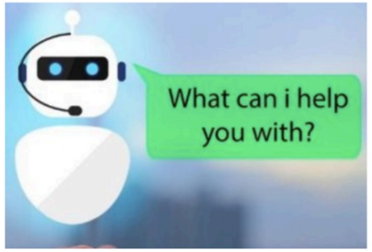
Donald Trump is the US President

Plants perform photosynthesis.

Unicorns live in Africa.

Applications should produce reliable content

Machine Translation



Smart Assistants

Search Engines

Auto Transcription

google-research/
pegasus

5 Contributors 79 Issues 1k Stars 261 Forks

Summarization Engines

Health Record Analysis

and many more ....

1

Grains of Factuality

Original: a recent poll finds that most americans feel that businesses like restaurants and event centers should not discriminate against same-sex weddings. public opinion has shifted on the issue since last fall after Indiana changed its ...

Most American businesses oppose same-sex marriage

Grains of Factuality

Original: a recent poll finds that most americans feel that businesses like restaurants and event centers should not discriminate against same-sex weddings. public opinion has shifted on the issue since last fall after Indiana changed its …

Most American businesses **oppose** same-sex marriage

Grains of Factuality

Original: a recent poll finds that most americans feel that businesses like restaurants and event centers should not discriminate against same-sex weddings. public opinion has shifted on the issue since last fall after Indiana changed its …

Most American businesses oppose same-sex marriage

Incorrect relation between entities

Factual Errors in Summarization vary across Datasets and Models

- Summaries generated by the same models consist of different error distributions over different datasets
- Error distribution can vary among models within the same category

Semantic Frame Errors

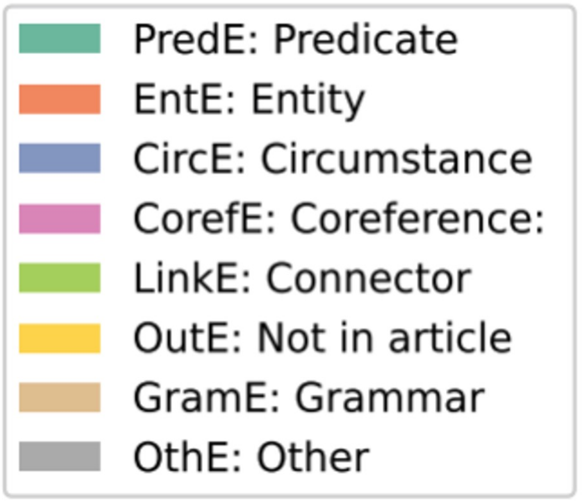
Fine-grained errors within a sentence

Discourse Errors

Fine-grained errors across sentences

Content Verifiability Errors

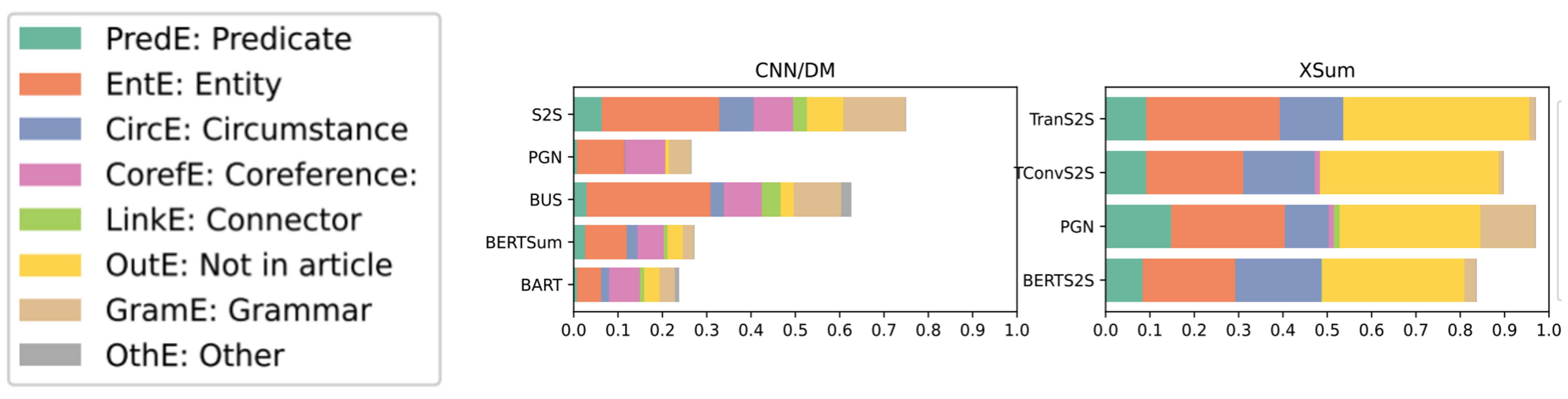
Errors out of article scope



Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics (Pagnoni, *Balachandran* et. al, 2021)

Factual Errors in Summarization vary across Datasets and Models

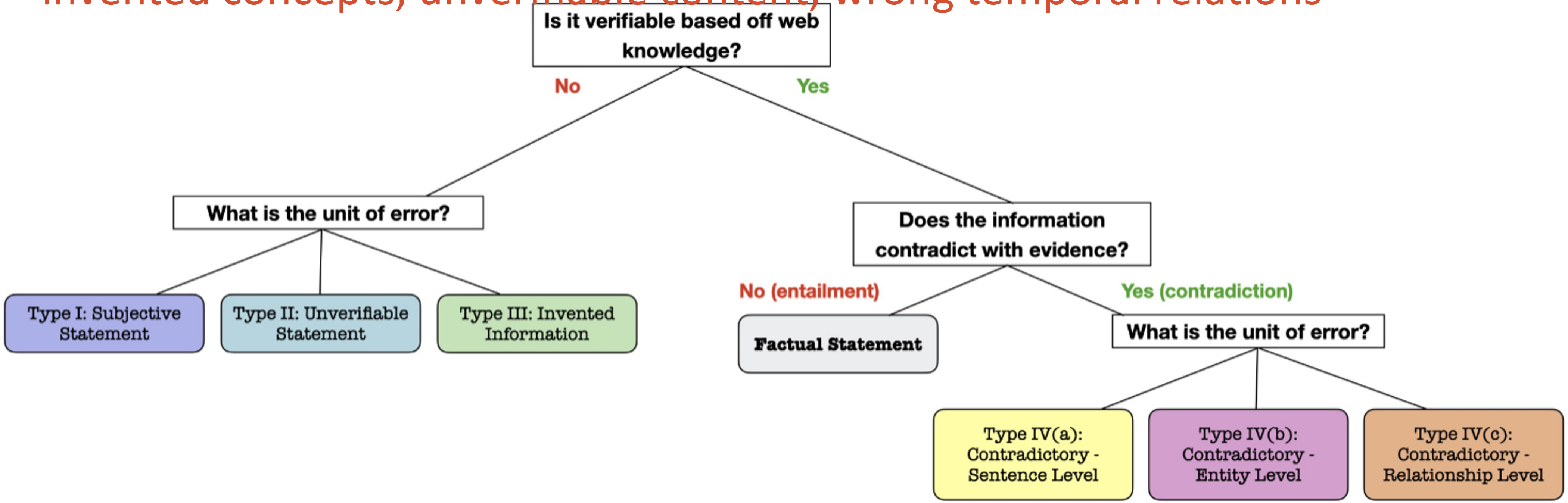
- Summaries generated by the same models consist of different error distributions over different datasets
- Error distribution can vary among models within the same category



Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics (Pagnoni, *Balachandran* et. al, 2021)

Factual Errors in Open-Generation are more complex

- Powerful LLMs like GPT models, LLama models produce more complex factual issues - invented concepts, unverifiable content, wrong temporal relations



FAVA: Understanding and Correcting Hallucinations in Large Language Models (forthcoming Mishra, et. al, 2023)

Factual Errors in Open-Generation are more complex

- Powerful LLMs like GPT models, LLama models produce more complex factual issues - invented concepts, unverifiable content, wrong temporal relations

Type	Example	ChatGPT	Llama2
Subjective	Lionel Messi is **the best soccer player in the world**.	12.82%	8.86%
Invented	**Messi is also famous for his discovery of the famous airplane kick technique.**	5.13%	22.97%
Unverifiable	**In his free time, Messi enjoys singing songs for his family.**	14.74%	5.06%
Contradictory	**Messi has yet to gain captaincy for the Argentina national football team.**	14.74%	14.10%
Entity	Lionel Andrés Messi was born on June ~~12~~ **24**, 1987.	49.36%	46.47%
Relation	Lionel Messi ~~acquired~~ **was acquired by** Paris Saint-Germain.	3.21%	2.53%

FAVA: Understanding and Correcting Hallucinations in Large Language Models (forthcoming Mishra, et. al, 2023)

Challenges for data collection

- Training Data: (Text, Incorrect/Correct Label) Pairs, Span Level Annotation

- Human Annotated Data
  - Expensive - Long Process to read and edit Text (Pagnoni, et. al, 2021, Min et. al, 2023)
  - Subjective - Factuality decisions have low agreement across annotators (Falke et al, 2019, Durmus et al, 2020)

- Synthetic Data - Create synthetic incorrect summaries using heuristic rules (Kryściński et. al, 2020, Cao et. al, 2020)

Annotated Datasets and Benchmarks

Benchmark	Datasets	Data Size	Language
TruthfulQA (Lin et al., 2022)	-	817	English
REALTIMEQA (Kasai et al., 2022)	-	Dynamic	English
SelfCheckGPT-Wikibio (Miao et al., 2023)	-	1,908	English
HaluEval (Li et al., 2023c)	Task-specific	30,000	English
	General	5,000	English
Med-HALT (Umapathi et al., 2023)	-	4,916	Multilingual
FACTOR (Muhlgay et al., 2023)	Wiki-FACTOR	2,994	English
	News-FACTOR	1,036	English
BAMBOO (Dong et al., 2023)	SenHallu	200	English
	AbsHallu	200	English
ChineseFactEval (Wang et al., 2023a)	-	125	Chinese
HaluQA (Cheng et al., 2023)	Misleading	175	Chinese
	Misleading-hard	69	Chinese
	Knowledge	206	Chinese

Benchmark	Datasets	Data Size	Language
FreshQA (Vu et al., 2023)	Never-changing	150	English
	Slow-changing	150	English
	Fast-changing	150	English
	False-premise	150	English
FELM (Chen et al., 2023d)	-	3,948	English
PHD (Yang et al., 2023)	PHD-LOW	100	English
	PHD-Meidum	100	English
	PHD-High	100	English
ScreenEval (Lattimer et al., 2023)	-	52	English
RealHall (Friel and Sanyal, 2023)	COVID-QA	N/A	English
	DROP	N/A	English
	Open Assistant	N/A	English
	TriviaQA	N/A	English
LSum (Feng et al., 2023a)	-	6,166	English
SAC^3 (Zhang et al., 2023a)	HotpotQA	250	English
	NQ-Open	250	English

A Survey on Hallucination in Large Language Models (Huang et.al, 2023)

Synthetic Data Generation

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

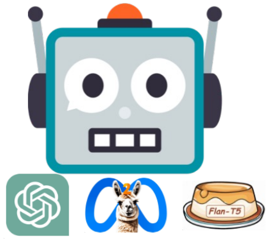
Evaluating the Factual Consistency of Abstractive Text Summarization (Kryscinski et.al, 2019)

Synthetic Data Generation

Evidence: Rishi Sunak (born 12 May 1980) is a British politician…

Text: Introducing Rishi Sunak:British politician who has served in various roles within the UK government.

Factual Claim + Document



LLM

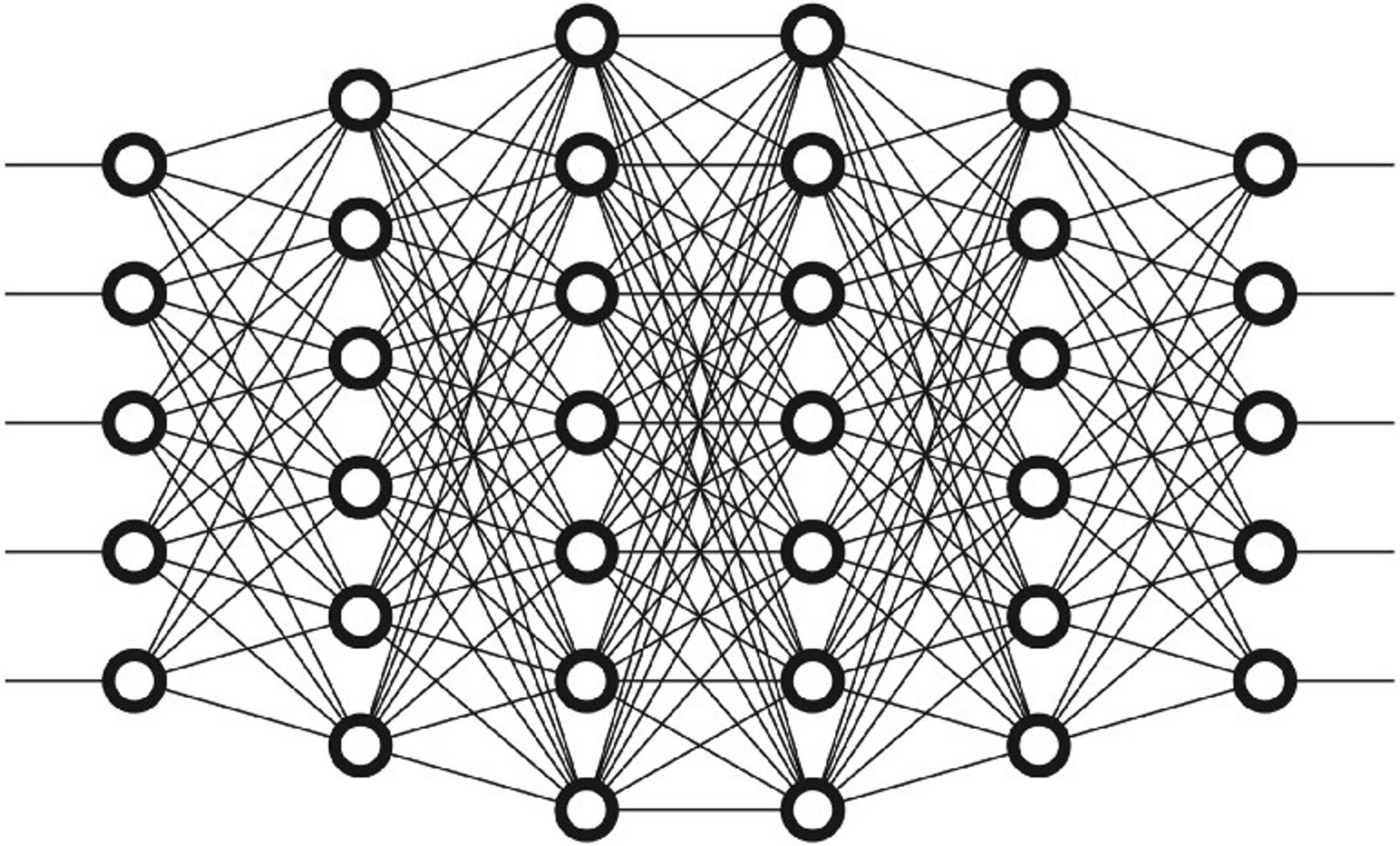
Introducing Rishi Sunak: **Indian** politician who has served in various roles within the UK government. **He was an avid golfer during his graduate school days.**

Perturbed Claim - Non-Factual

FAVA: Understanding and Correcting Hallucinations in Large Language Models (forthcoming Mishra, et. al, 2023)

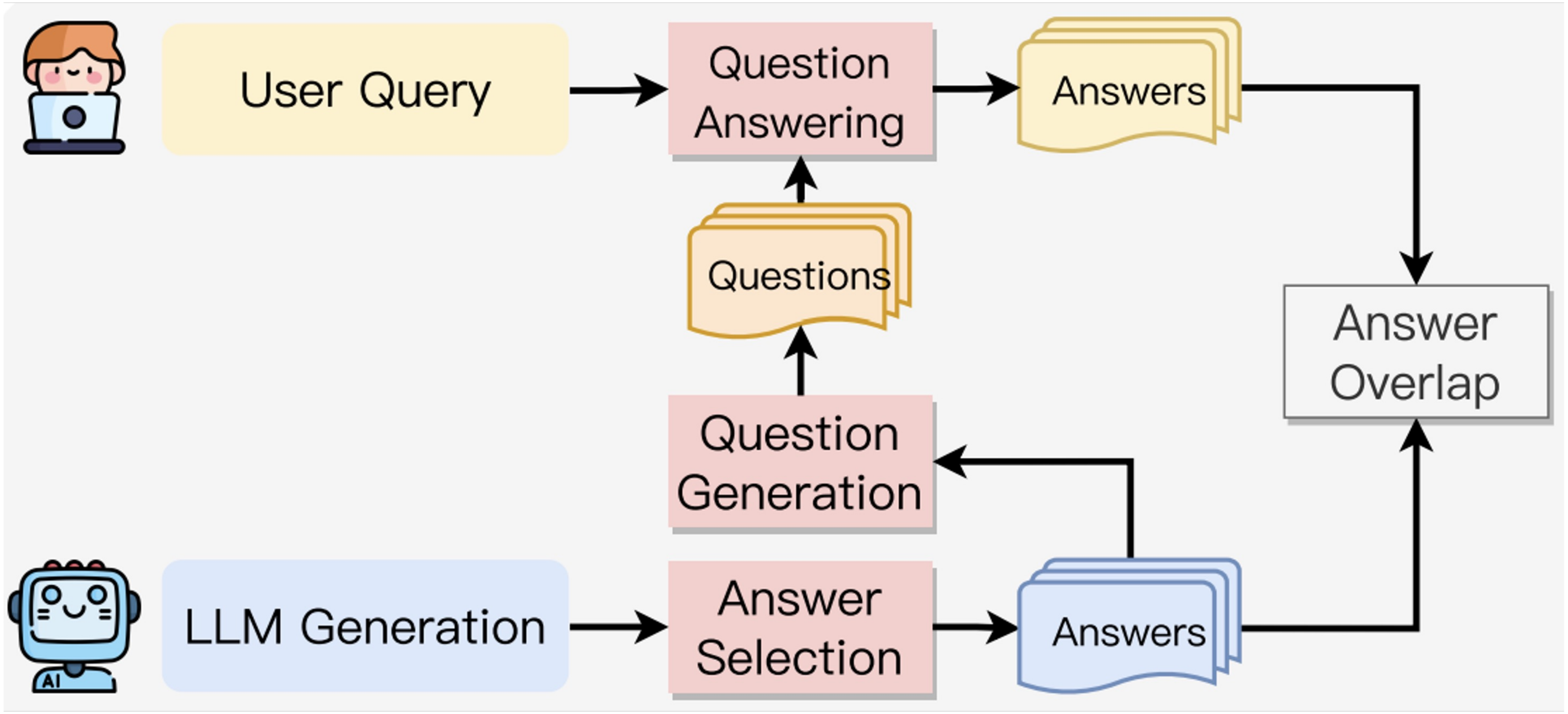
Coarse Factuality Classifiers

Model
Generation [SEP]
Source/Evidence



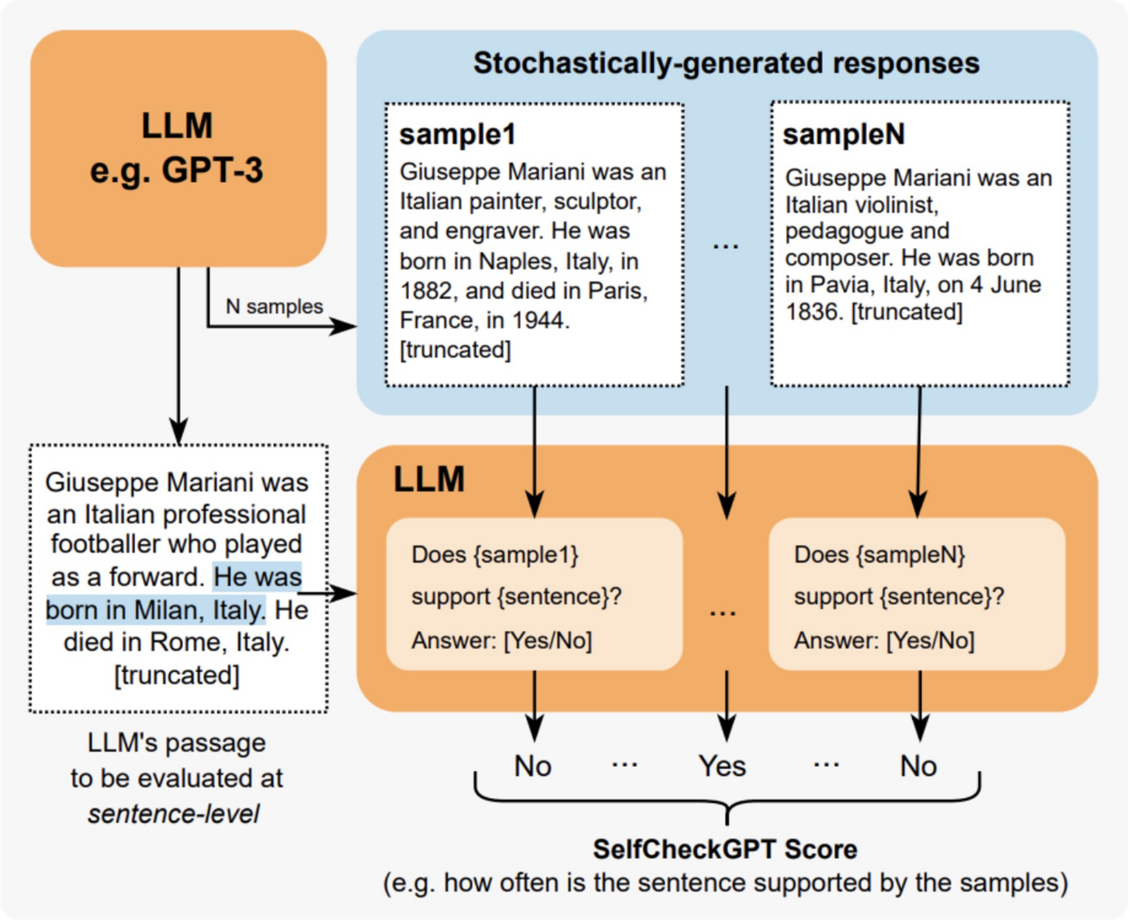
Evaluating the Factual Consistency of Abstractive Text Summarization (Kryscinski et.al, 2019)

QA Based Factuality Detection



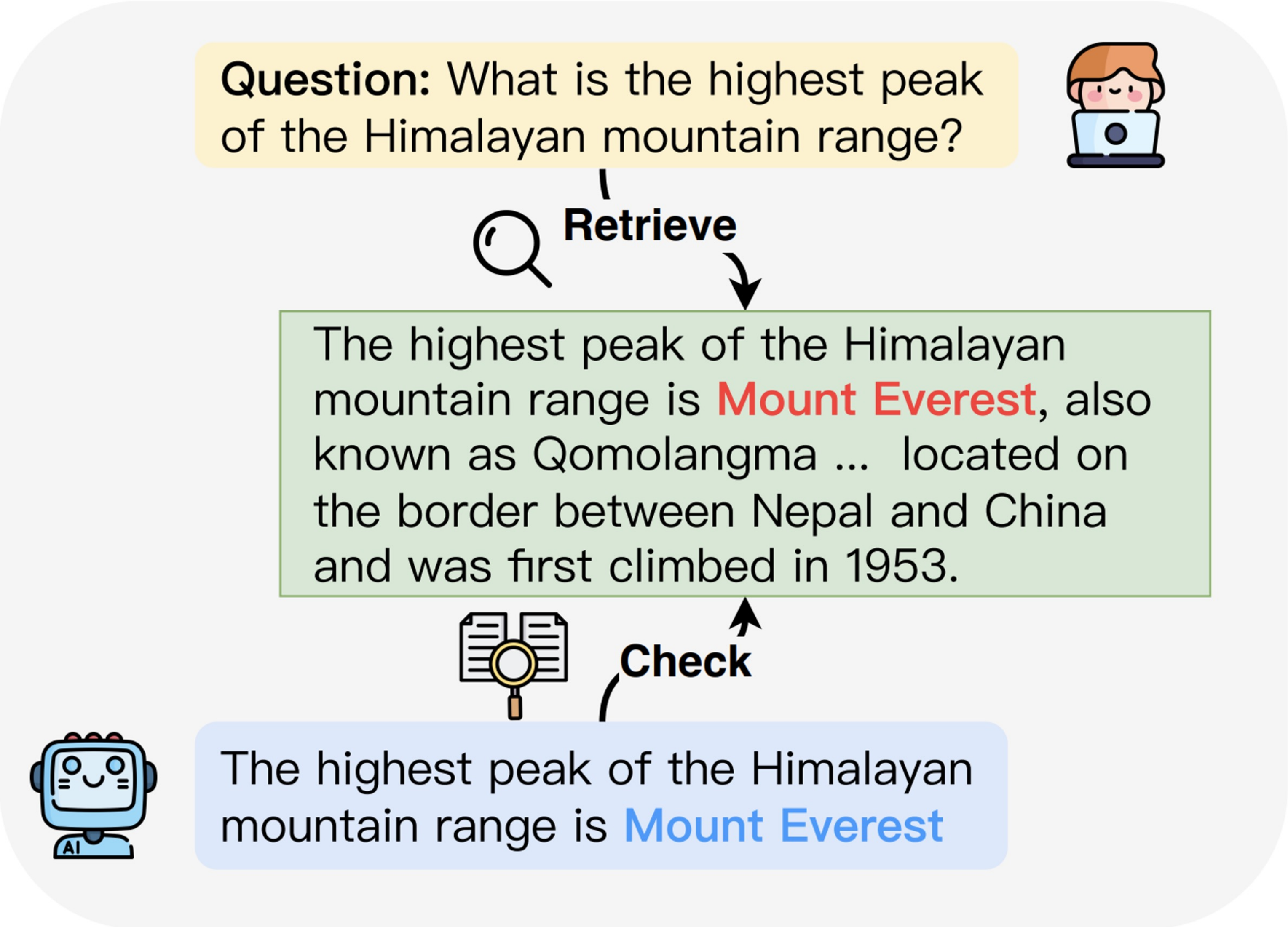
A Survey on Hallucination in Large Language Models (Huang et.al, 2023)

LLM based Factuality Detection



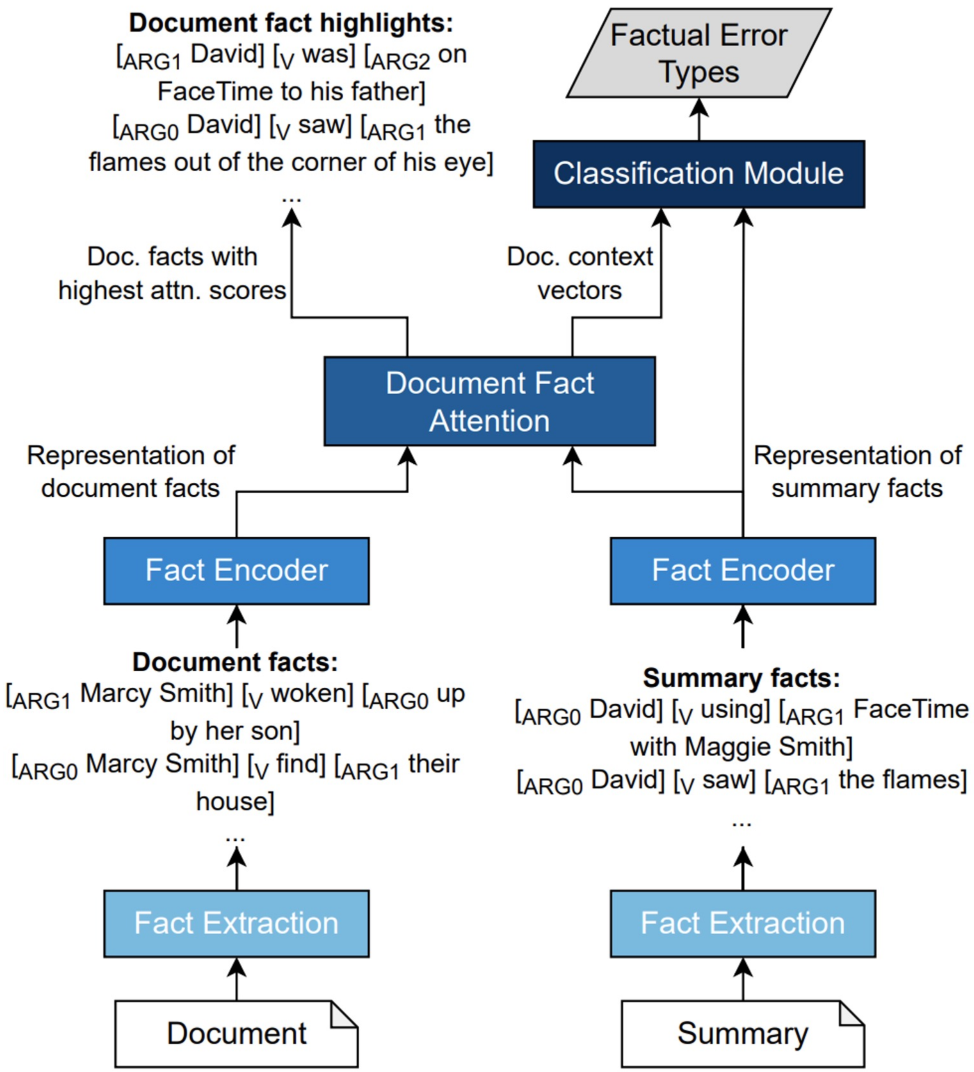
SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models (Manakul et.al, 2023)

Retrieval-based Fact Checking



A Survey on Hallucination in Large Language Models (Huand et.al, 2023)

Fine-Grained Inconsistency Detection

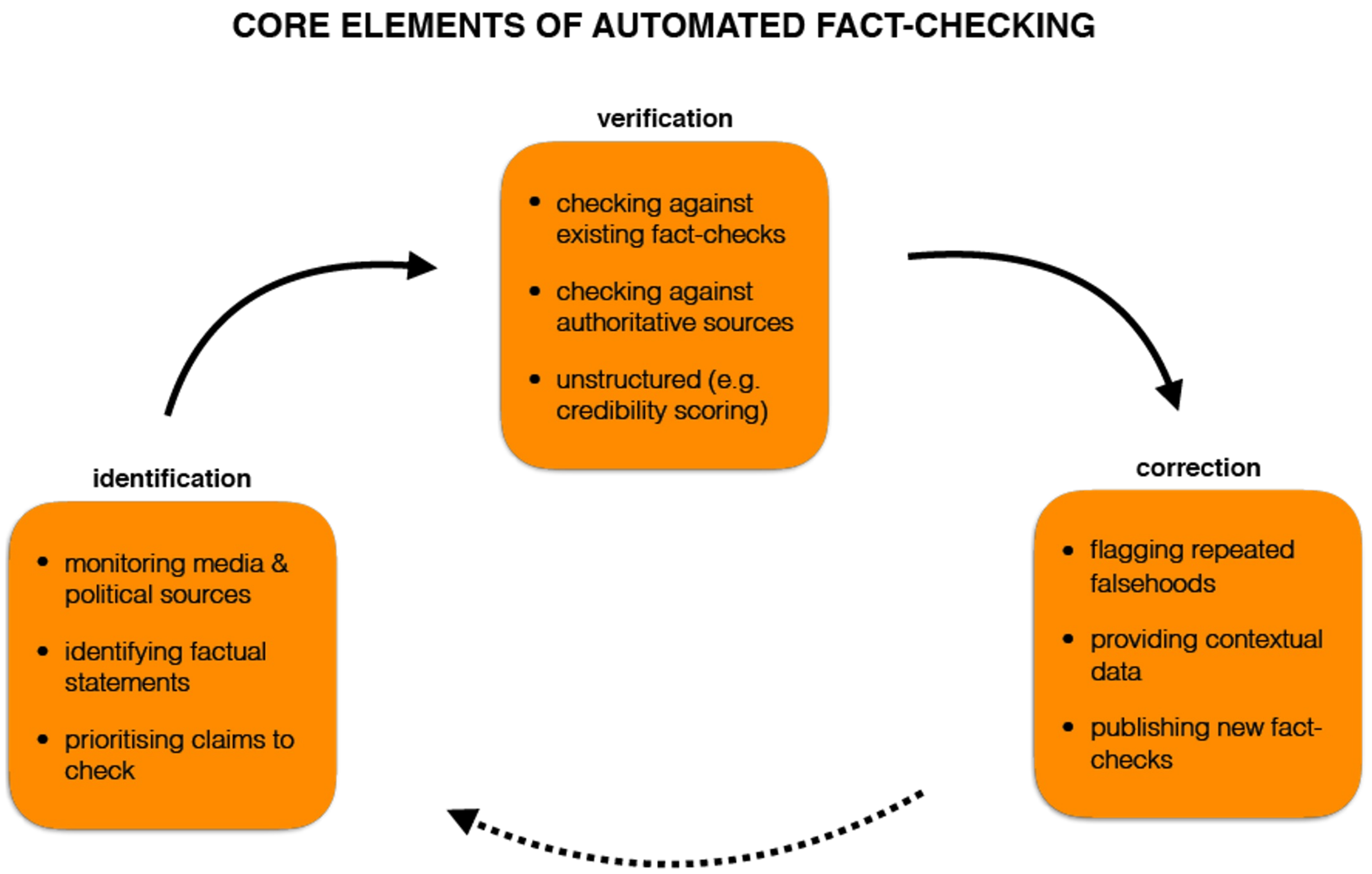


Interpretable Automatic Fine-grained Inconsistency Detection in Text Summarization (Chan et.al, 2023)

Takeaways

- Detecting misinformation and factual errors is a complex problem
  - Varies across datasets and models
  - Different error types for different applications
- Collecting Data is subjective and expensive
- Building better detectors require focused knowledge of error types
- Models have different capabilities and skills
  - Different sets of errors based on quality of generation
  - Classifiers need to generalize to different error types in different models

Related research: Fact verification

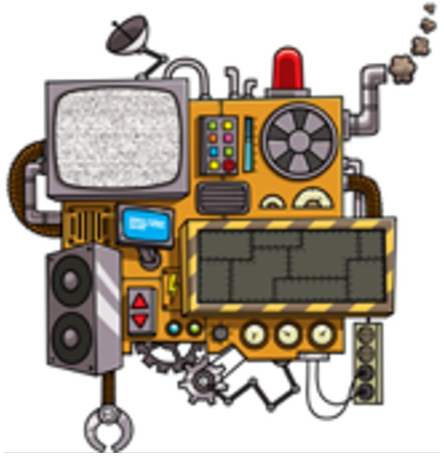


FACTSHEET: Understanding the Promise and Limits of Automated Fact-Checking

Overview: Evaluation, Detection and Redaction of Harms

- **Inadvertent Harm: Toxic Text**
  - Defining Toxic Text
  - Data and Benchmarks
  - Toxicity Detection
- **Inadvertent Harm: Factuality/Misinformation**
  - Grains of Factual Errors
  - Datasets
  - Hallucination/Factual Error Detection
- **Intentional Harm: Disinformation**
  - Visualizing Machine Generated Text
  - Detecting Machine Generated Text

Intentional Harms: Spreading Misinformation

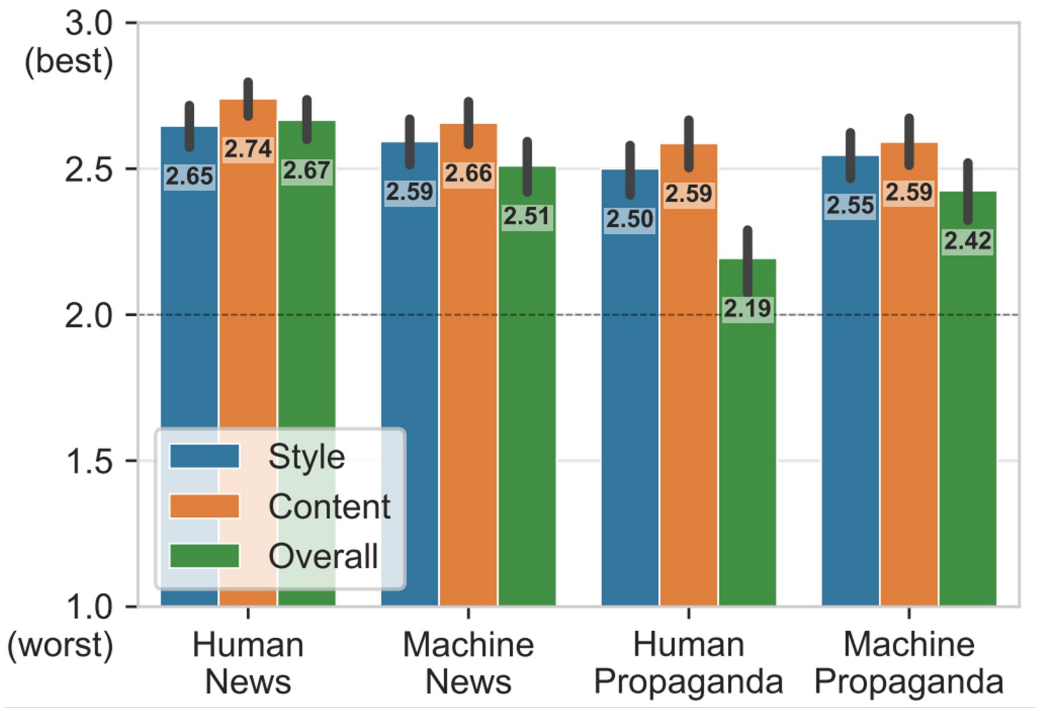


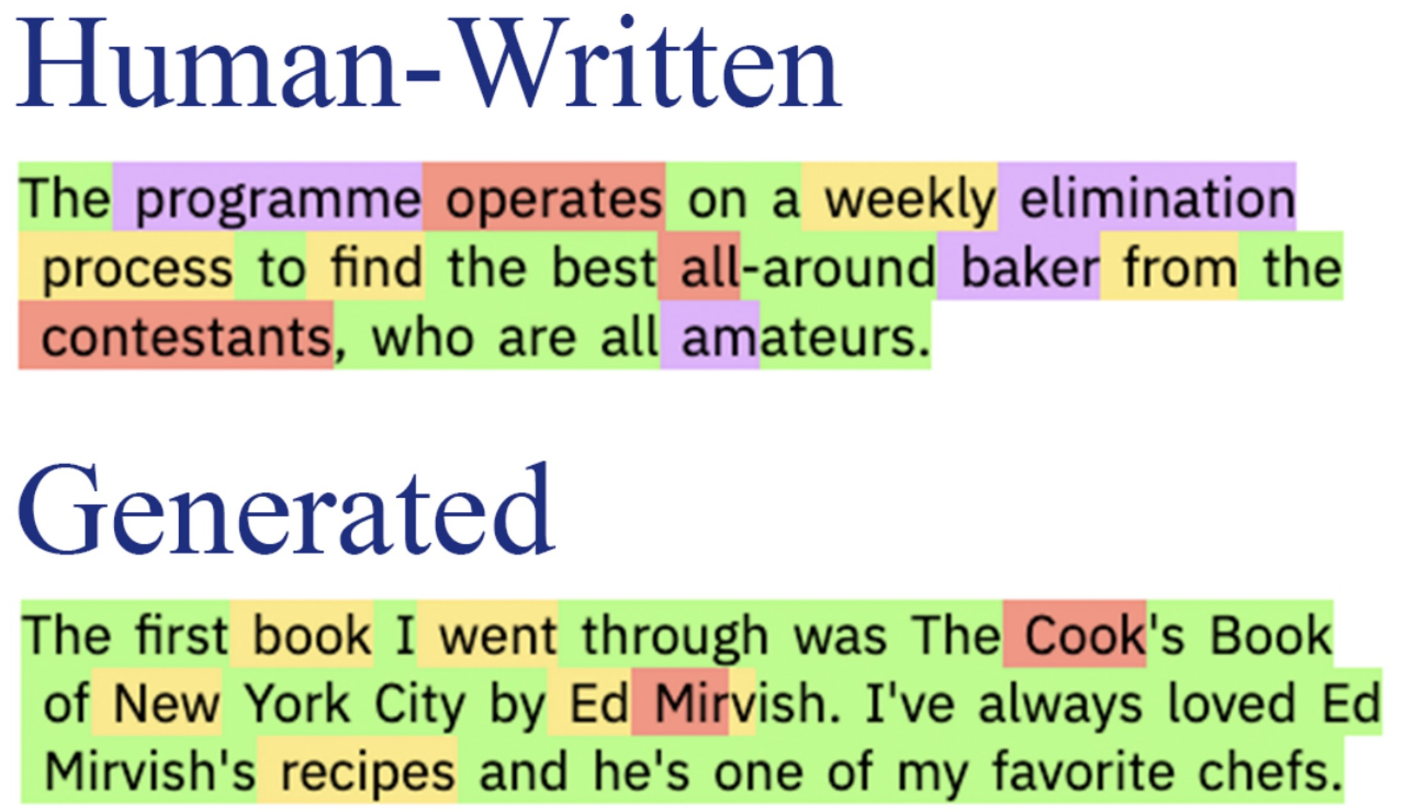
Global warming is a hoax

Apollo 11 didn't land on the moon

Drinking water is good for you

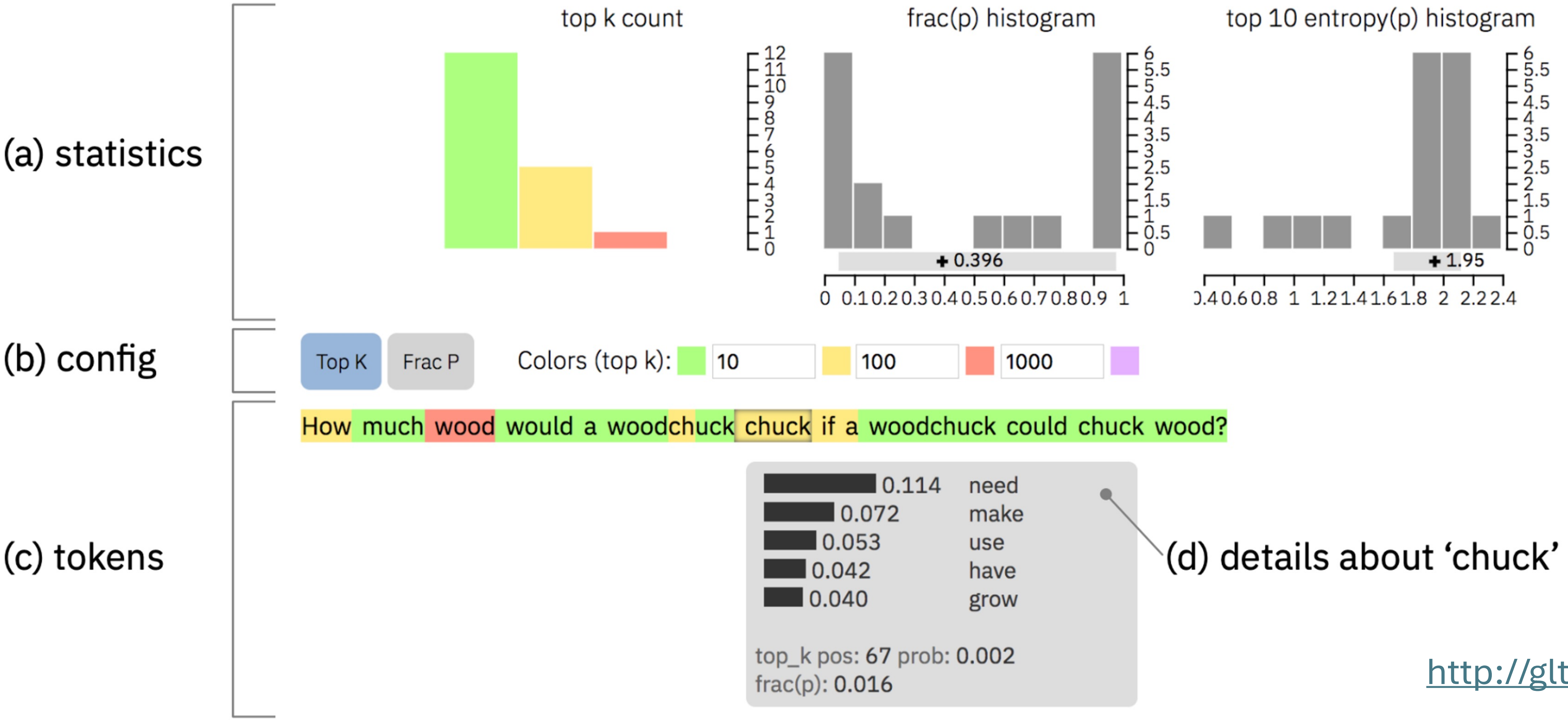
Humans cannot identify machine generated text easily





Defending Against Neural Fake News (Zellers et, al. 2020)

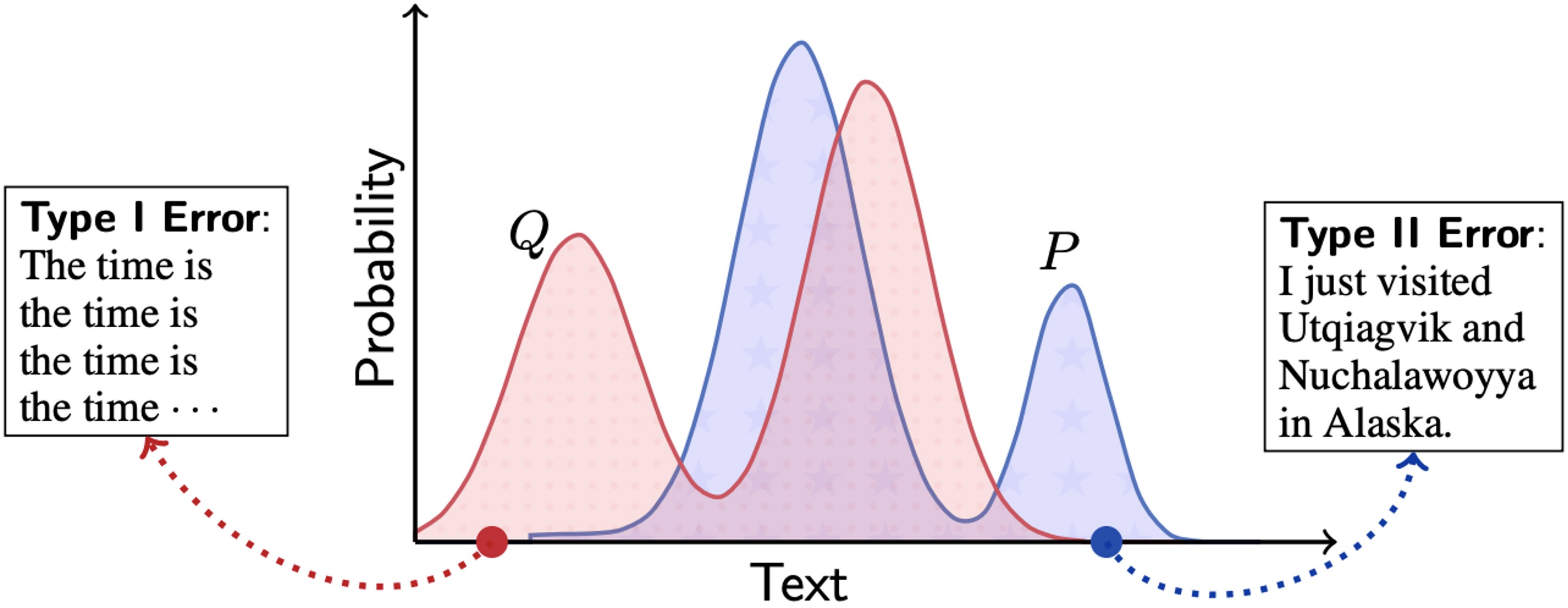
Visualize machine generated text - GLTR



http://gltr.io/

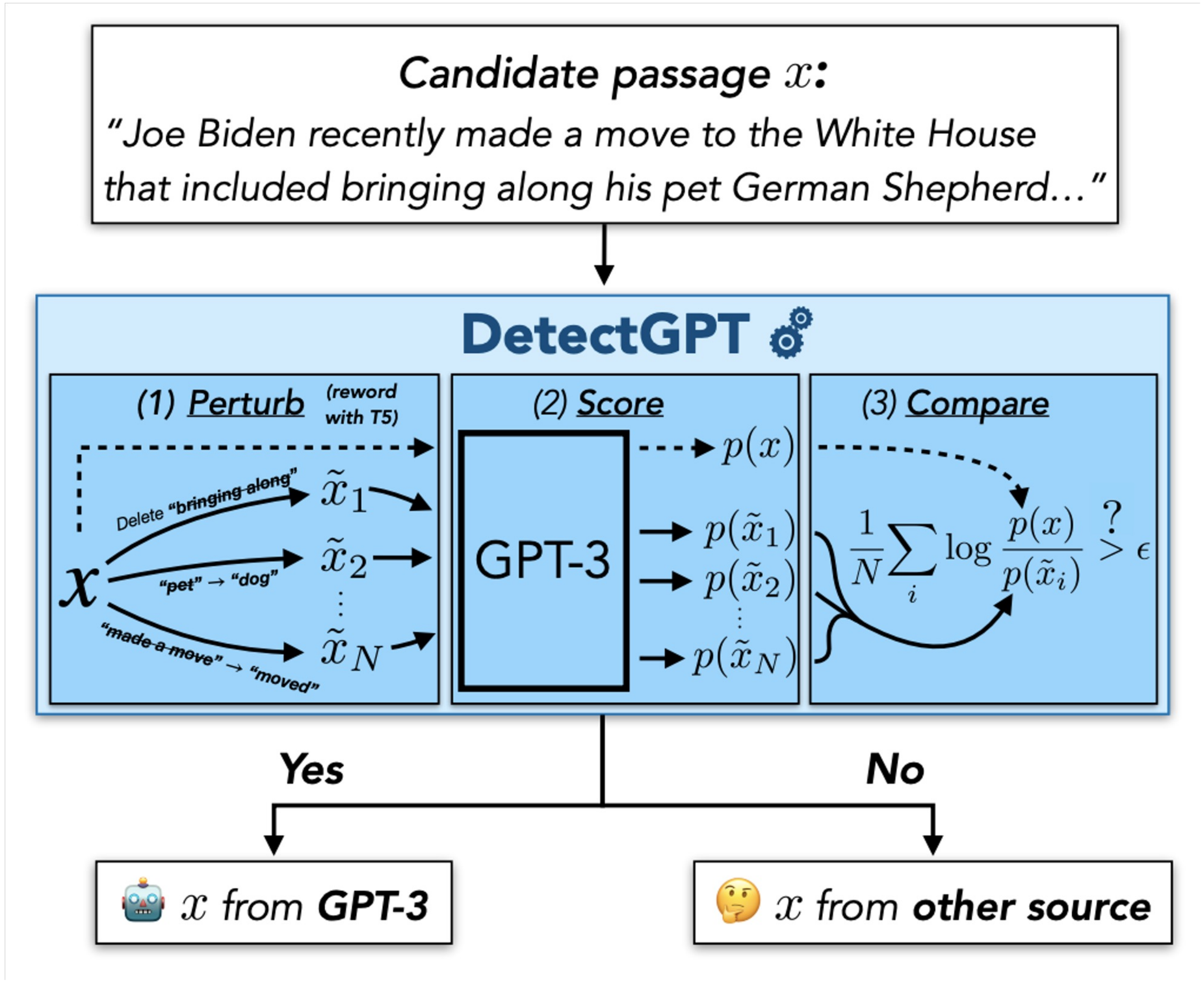
GLTR: Statistical Detection and Visualization of Generated Text (Gehrmann et.al, 2019)

Machine generated text v/s Human text



MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers (Pillutla et.al, 2021)

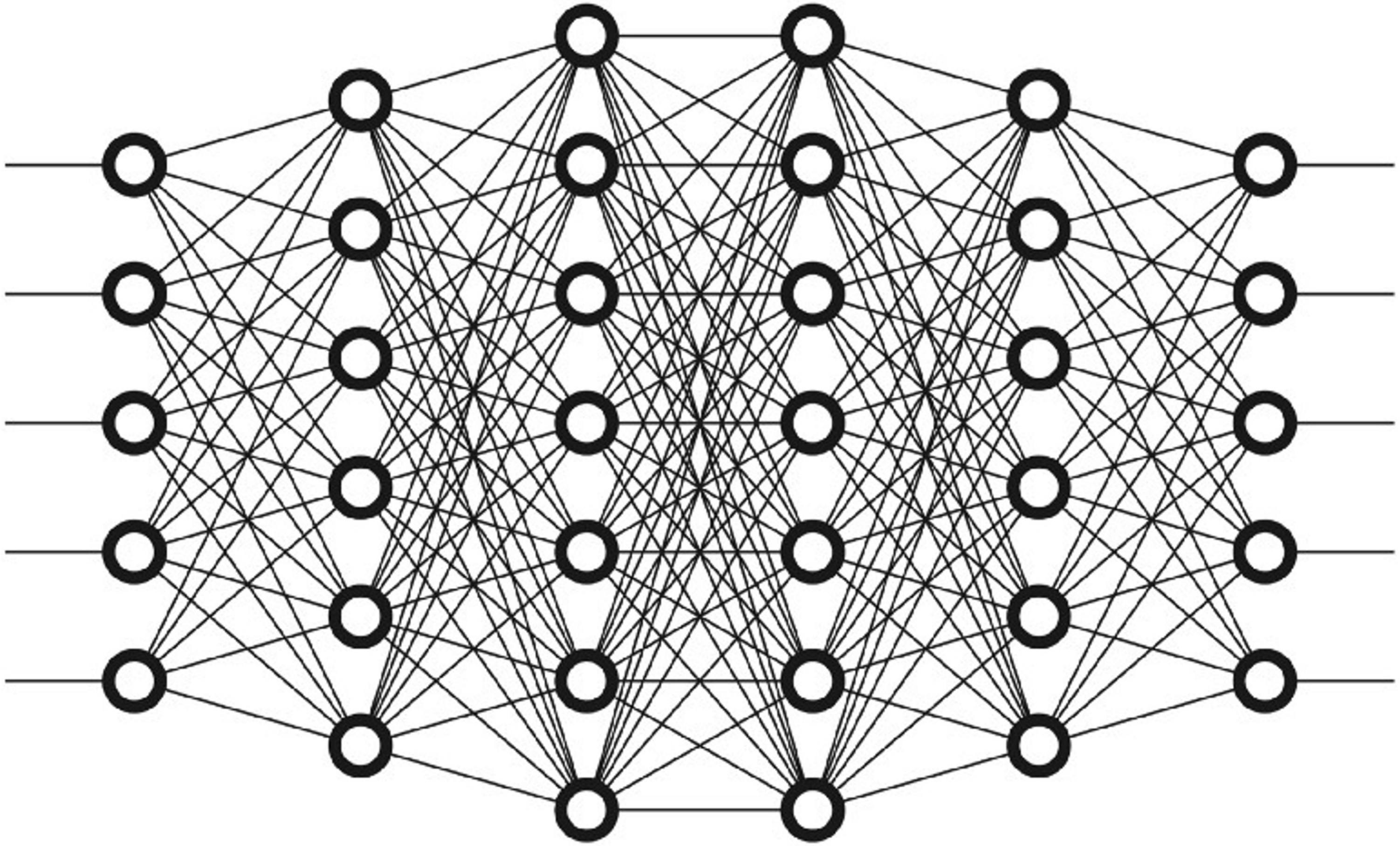
DetectGPT: Machine generated text v/s Human text



DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature (Mitchell et.al, 2023)

Coarse Machine Text Detectors

Unicorns are beautiful animals from Africa.



Smaller Language Models are Better Black-box Machine-Generated Text Detectors (Mireshghallah et.al, 2023)

Watermarking machine generated text

Prompt	Num tokens	Z-score	p-value
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet	56	.31	.38
With watermark - minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

A Watermark for Large Language Models (Kirchenbauer et.al, 2023)
SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation (Hou et.al, 2023)

Challenges and Takeaways

- Generalizability
  - Hard to generalize to new/unseen model generalizations
  - Model architecture, different decoding methods (e.g., top-k, top-p), model size, different prefix lengths, and training data
  - Specialized models required for each model output set

- Humans and Models identify different errors
  - Humans detect semantic variations in generated text
  - Models better detect fluent but non diverse generated text
  - Human-Model collaboration required for better coverage

Actual Intervention: Redaction/Flagging

If you have ever been to a restaurant, you have probably noticed that the service is much better if the waiter is white, and the food is much better if the chef is white

If you have ever been to a restaurant, you have probably noticed that the service is much better [redacted], and the food is much better [redacted]

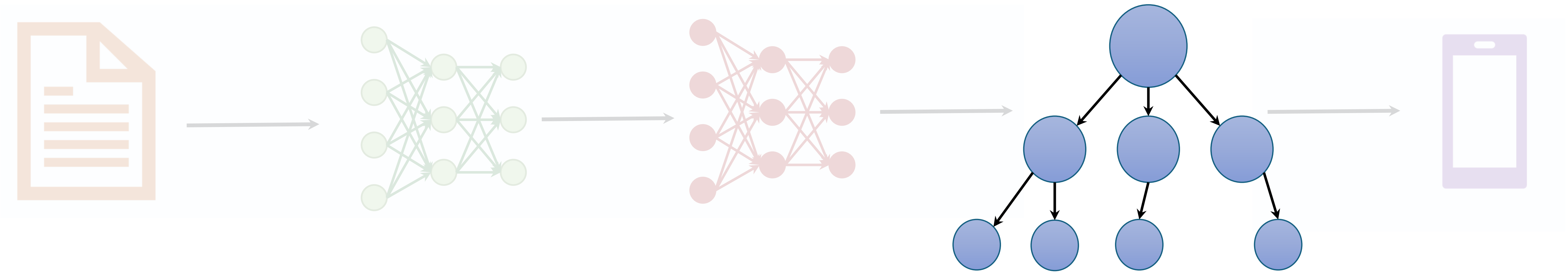
Application interventions: Challenges and Open Questions

Detection is contextual and subjective.

e.g. Definition of toxicity can be different for different cultures

Redaction and Flagging is often not straightforward.

e.g. redacting part of text might change the meaning

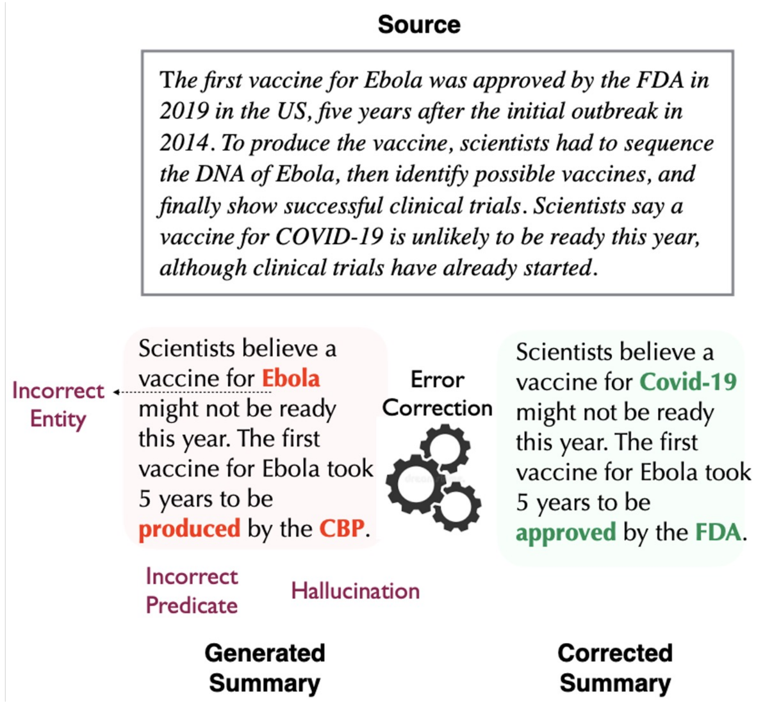


Inference Interventions

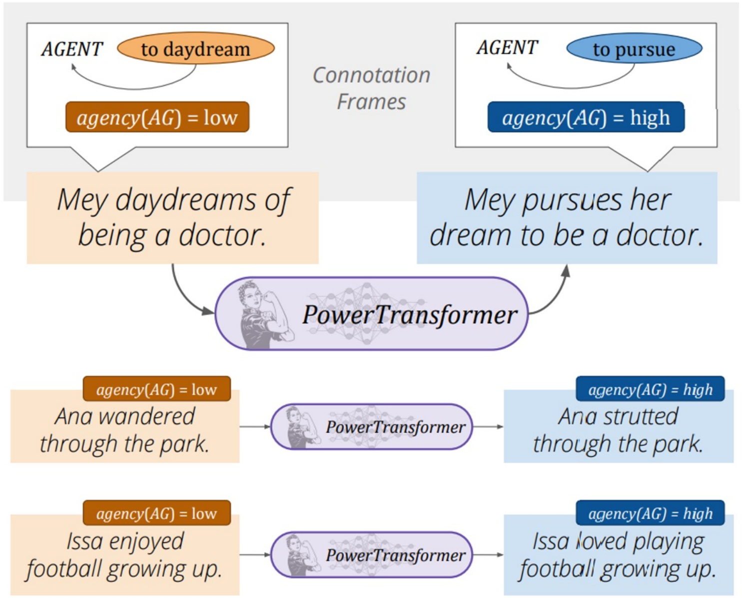
Stakeholders: Practitioners with access to the model inference.

Post Editing

Identify issues with generated outputs and edit them.



[Fact correction in summarization, Balachandran et al 2022]



[PowerTransformer: Debiasing, Ma et al 2021]

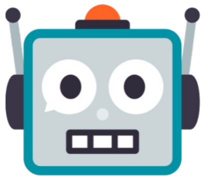
Intervention: Prompting + Instructions

Prompt model to generate safe and reliable text

Instructions:
Summarize the document
Include only entities and information from the document.

Demonstrations:
{Document, Summary}

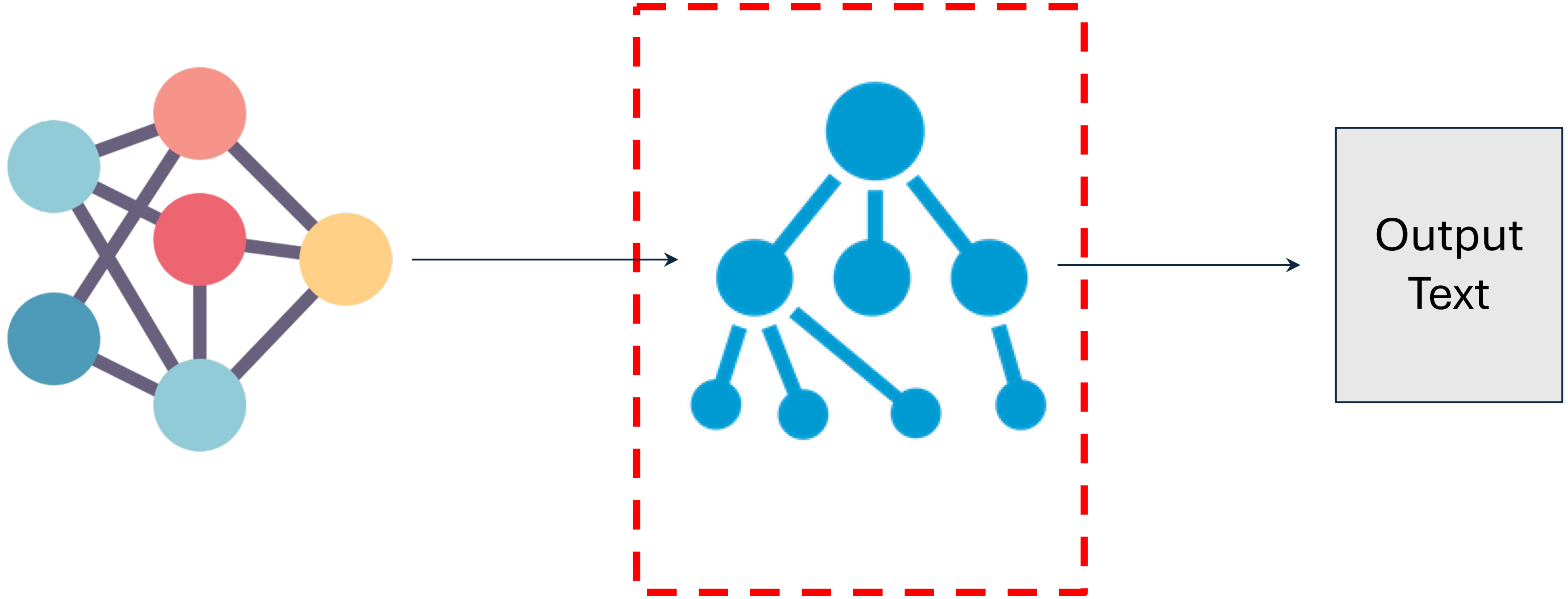
Document: Rishi Sunak (Born 12 May 1980) is a British politician who has served as Prime Minister of the United Kingdom….



Summary: Rishi Sunak was appointed as …

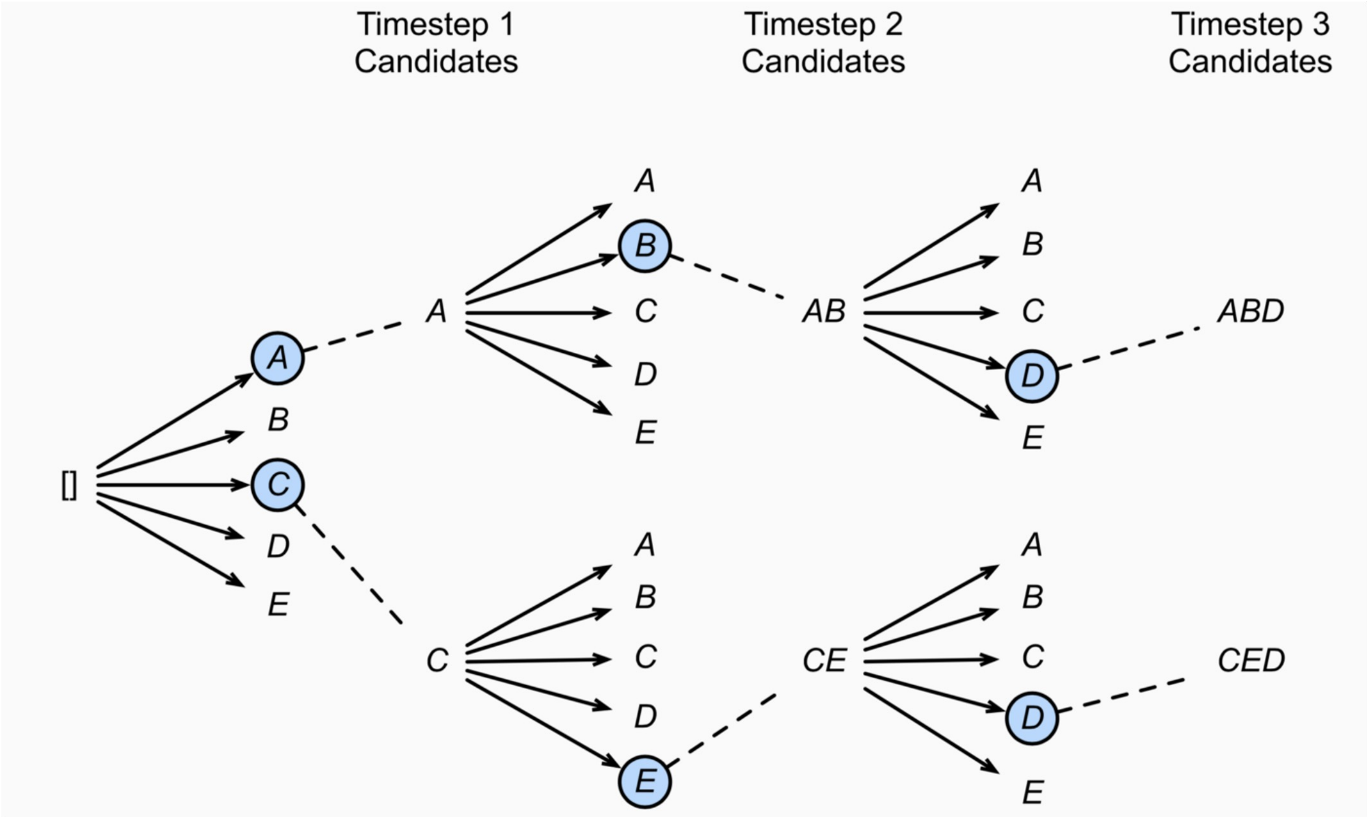
[GPT-3 - Ouyang et, al. 2022]

Decoding Intervention



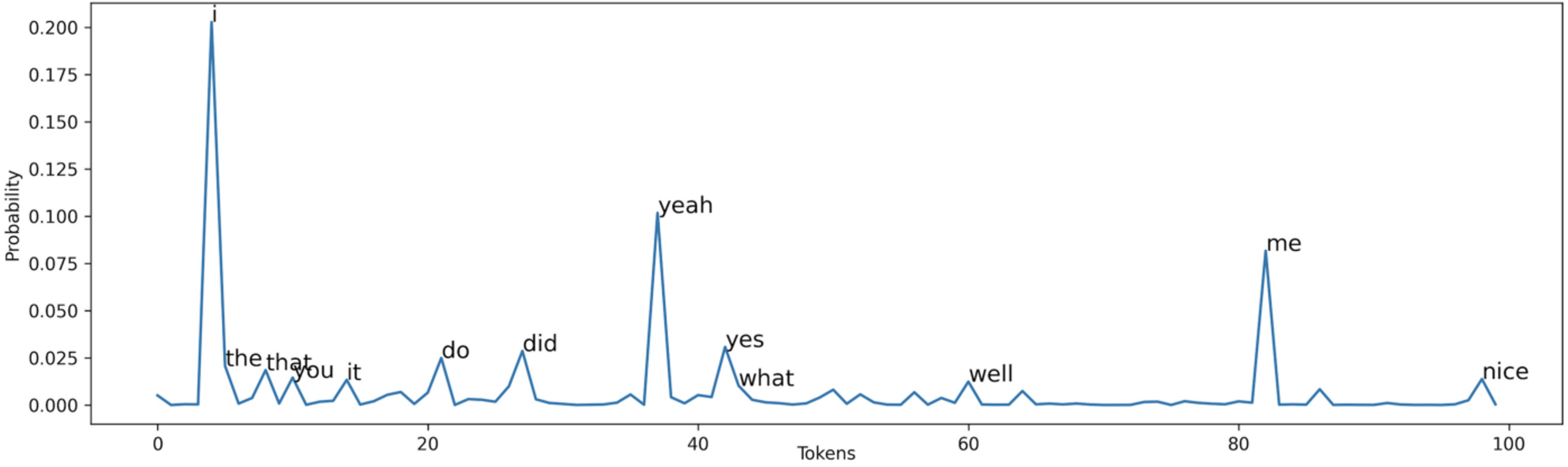
Common Decoding Algorithms: Beam Search

$$\max_{y_1,\ldots,y_n} p(y_1, \ldots, y_n)$$



Common Decoding Algorithms: Sampling

$$p(y_i | y_{i-1}, [\mathbf{x}]) = \frac{\exp(u_i/\tau)}{\sum_j \exp(u_j/\tau)}$$

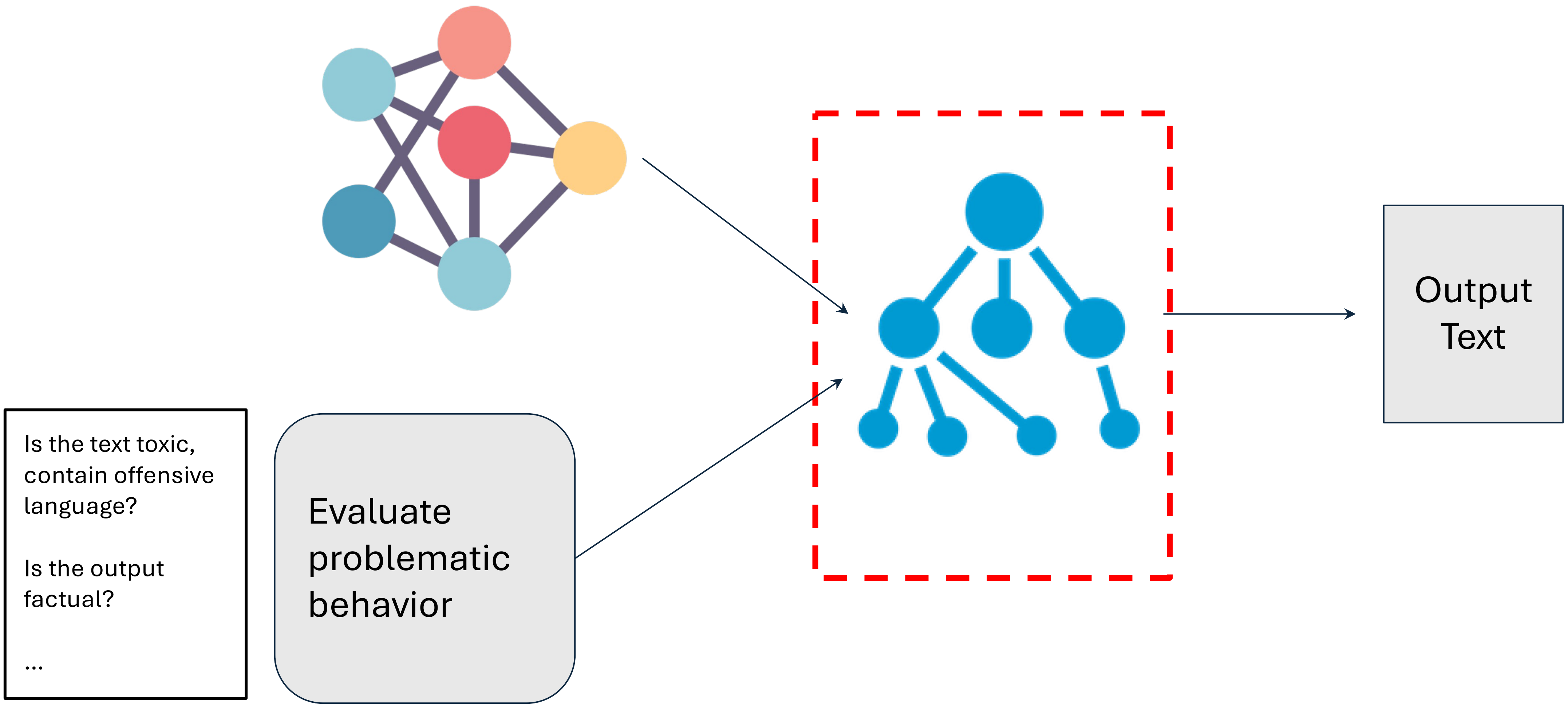


Ancestral Sampling, Top-k, Nucleus Sampling

...

[Holtzmann et al 2019, Fan et al 2018]

Decoding Intervention ala Controllable Text Generation

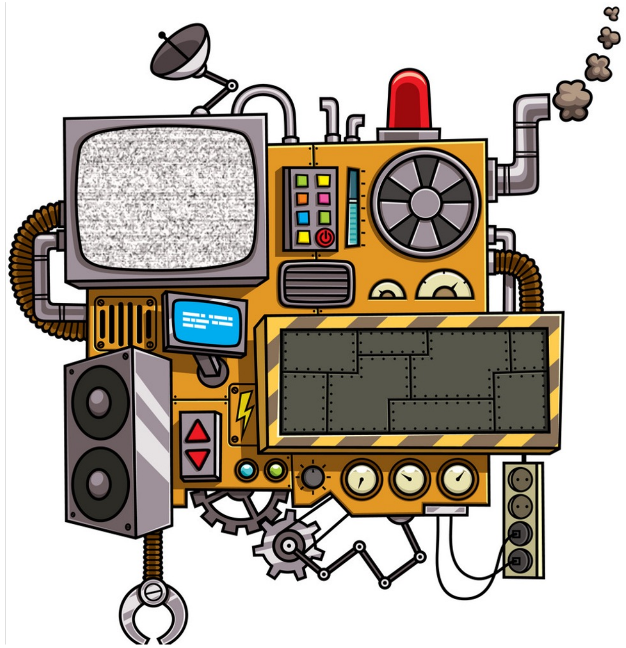


How to evaluate problematic behavior?

- Blocklists - keywords to avoid
- Classifiers (to detect hate speech, toxicity, hallucinations etc).
- Additional (smaller) language models
- ...

A naive solution: Rejection/Best-of-N Sampling

$\mathbf{x}$



$p(y_{i+1}|y_{1:i}, [\mathbf{x}])$

$\hat{\mathbf{y}}_1$

$\hat{\mathbf{y}}_2$

$\hat{\mathbf{y}}_3$

$\hat{\mathbf{y}}_4$

$\hat{\mathbf{y}}_5$

Decoding Intervention: Probability Guided Decoding

$$p(y_{i+1} | y_{1:i}, \mathbf{x}, a)$$

Desired property
e.g. non-toxicity

$$p(y_i | y_{1:i-1}, \mathbf{x}, a) \propto p(a | y_{1:i}) p(y_i | y_{1:i-1}, \mathbf{x})$$

Desired property

LM

Blocklists - Don’t generate bad keywords

A	**0**
a	**0**
and	**0.01**
f*ggot	**0.6**
person	0.3
n*gga	**0.2**
…	**0**
Zyzomys	**0**
Zyzzogeton	**0**

Original Output Probability

List of Dirty, Naughty, Obscene or Otherwise Bad Words?
E.g. “porn,” “sex,” “f*ggot,” and “n*gga.

Set probability to 0.
Renormalize

A	**0**
a	**0**
and	**0.01**
f*ggot	0.0
person	0.7
n*gga	0.0
…	**0**
Zyzomys	**0**
Zyzzogeton	**0**

Modified Output Probability

Neurologic Decoding [Lu et al 2020], Neulogic A*esque Decoding [Lu et al 2021], K2T [Pascual 2021],

Decoding Intervention: Probability Guided Decoding

Generate words which are supported by the source

- How is consistency defined?
  - Model confidence – entropy
  - Cross-attention scores
  - Frequency of tokens

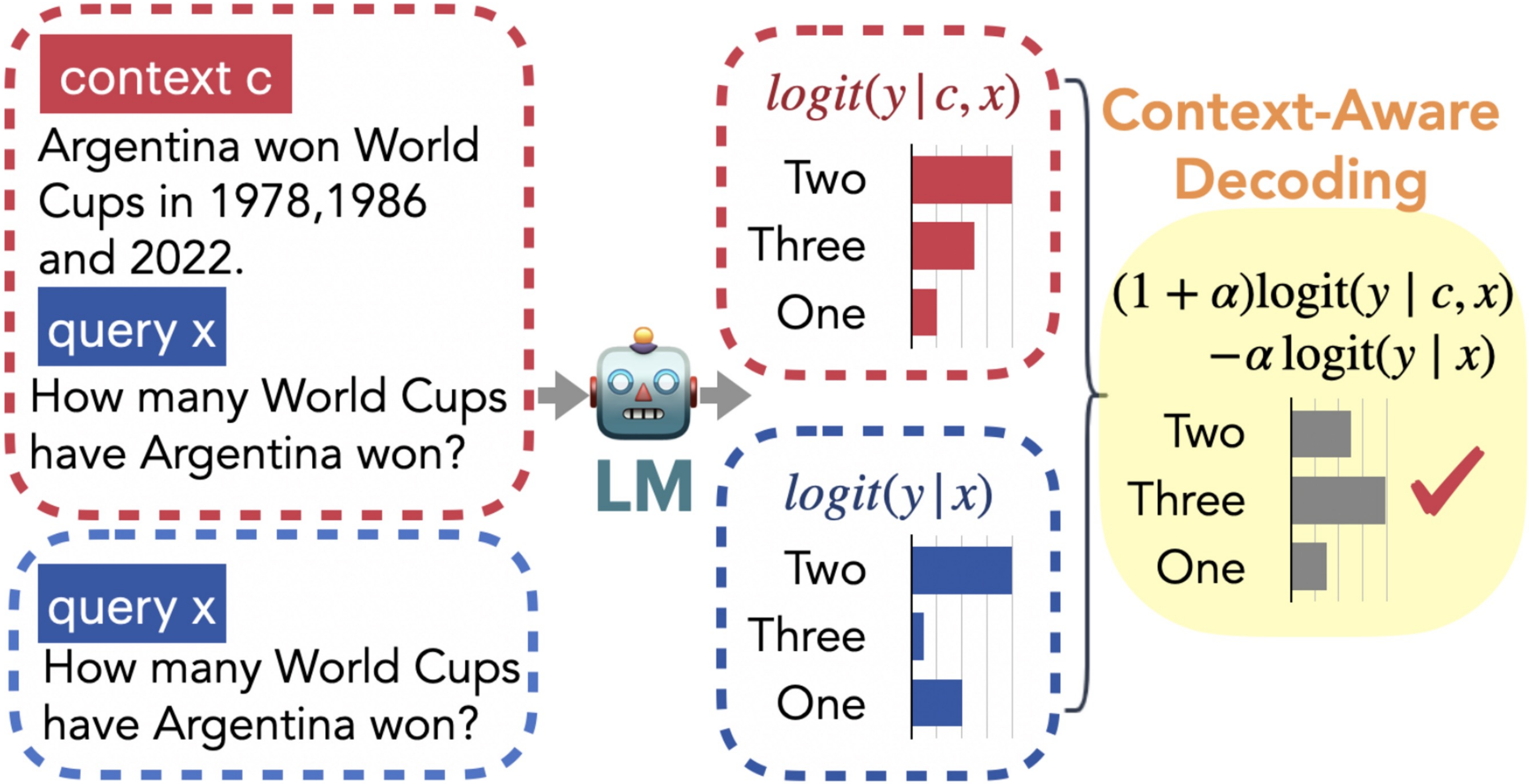
Method	Text
Source	...The PSNI said the tablets were "as yet unidentified" but warned of the "potential dangers" they posed...
BART	A 17-year-old boy has been charged after a teenager was taken ill after taking what police have described as *"potentially lethal" ecstasy tablets*.
PINOCCHIO	A 17-year-old teenager has been charged with drugs offences after a teenager was treated in hospital after taking what police described as an "unidentified" drug.

Table 1: An example of hallucination. Inconsistent words are highlighted in *red italic* fonts. In this case, PINOCCHIO corrects the inconsistent detail in the BART output.

[PINOCCHIO: Improving the Consistency of Abstractive Summarization, King et al 2022]

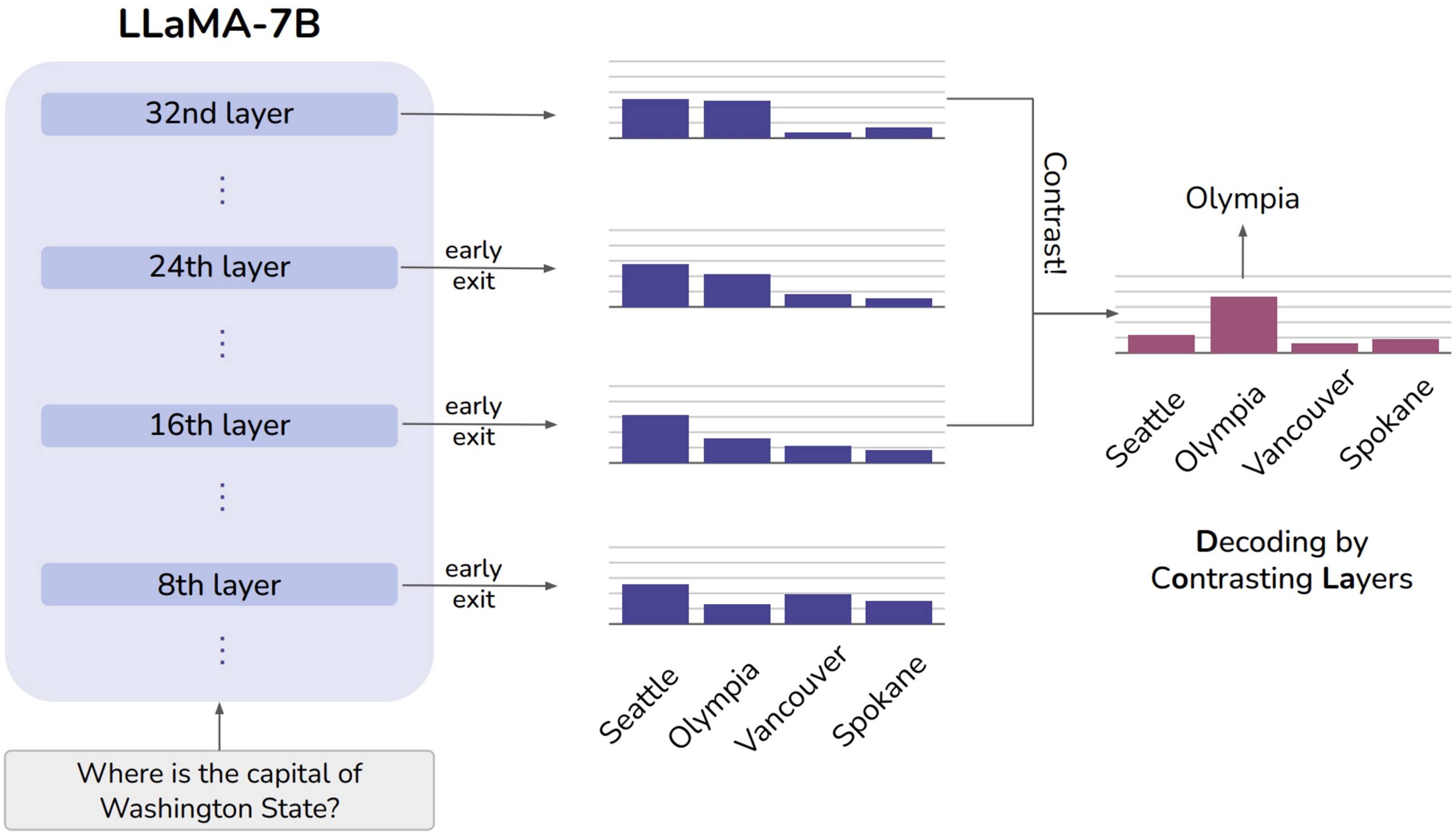
Decoding Intervention: Probability Guided Decoding

Generate words which are supported by the source



[Trusting Your Evidence:Hallucinate Less with Context-aware Decoding Shi et al 2023]

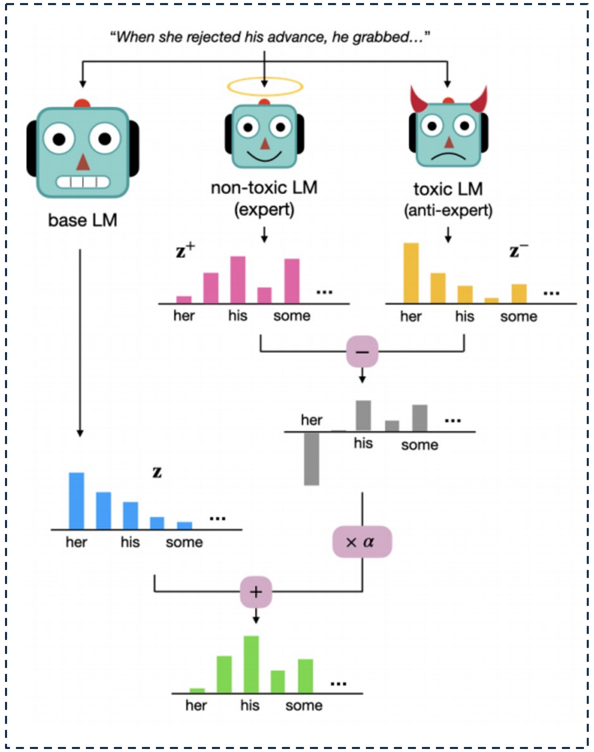
Decoding Intervention: Probability Guided Decoding



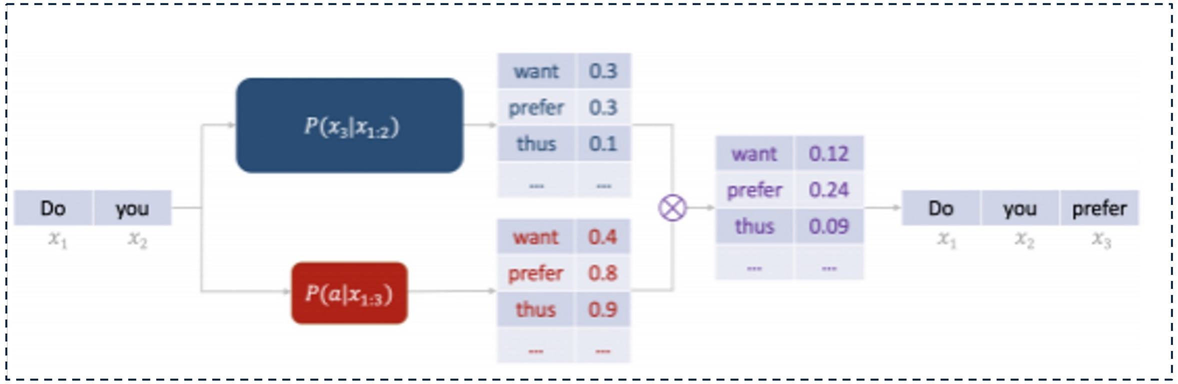
[DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models Chuang et al 2023]

Decoding Intervention: Model Guided decoding

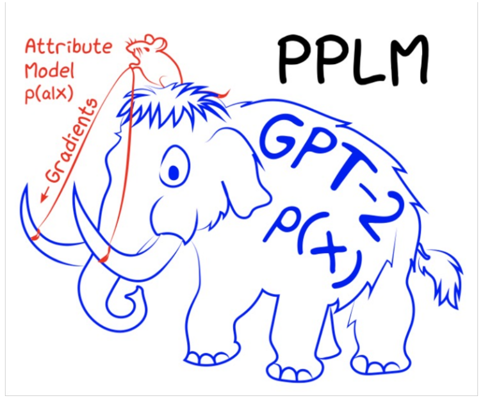
Generate tokens which are supported by an auxiliary model (like classifier)



Dexperts [Liu et al 2021]



FUDGE [Yang et al 2021]



PPLM [Dathathri et al 2020]

What's missing?

Desired property

$$p(y_i | y_{1:i-1}, \mathbf{x}, a) \propto \boxed{p(a | y_{1:i})} p(y_i | y_{1:i-1}, \mathbf{x})$$

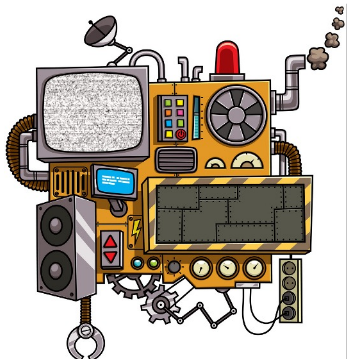
LM

Measured over the generated prefix and not the entire sequence.

Decoding Intervention: Non-autoregressive *Constrained* Decoding

We want to satisfy all goals at the sequence-level

$\mathbf{x}$



$\hat{\mathbf{y}}_1$

$\hat{\mathbf{y}}_2$

$\hat{\mathbf{y}}_3$

$\hat{\mathbf{y}}_4$

$$p(\mathbf{y}), f_1(\mathbf{y}), f_2(\mathbf{y})\ldots,$$

Language Model

Factuality Evaluator

Toxicity Predictor

Decoding interventions: Are all harms mitigated?

Evaluators can be hard to define or operationalize.

e.g. Factuality

Decoding interventions: Are all harms mitigated?

Evaluators can have their own biases.

e.g. toxicity detectors are shown to have racial biases, dialectal biases.

[Challenges in Dotoxifying LMs, Welbl et al 2021]

Decoding interventions: Are all harms mitigated?

Double edged sword. The same mitigation strategies can be used to inflict harm.

Decoding interventions: Are all harms mitigated? No

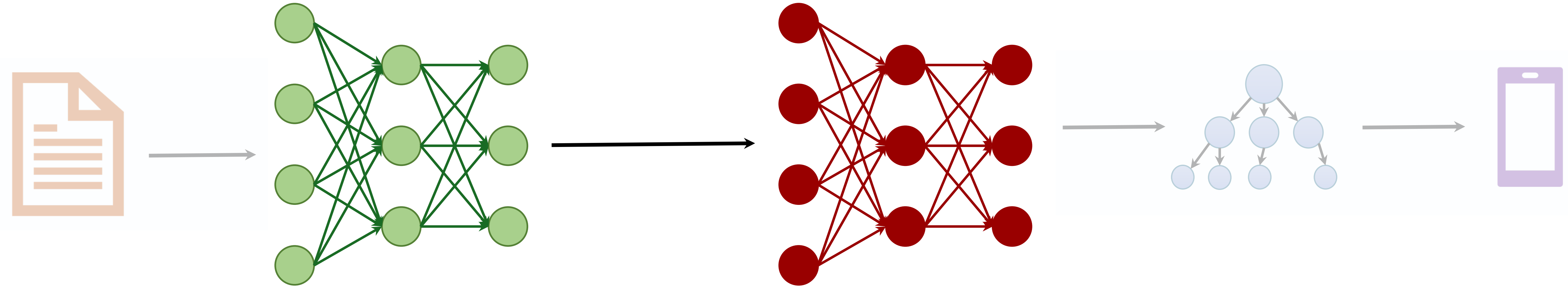
Evaluators can be hard to define or operationalize.

e.g. Factuality

Evaluators can have their own biases.

e.g. toxicity detectors are shown to have racial biases.

Double edged sword. The same mitigation strategies can be used to inflict harm.



Modeling Interventions

Stakeholders: Researchers and practitioners with access to the model parameters.

Language Model Training

- **Pretraining:** Designing the model architecture & pretraining the model on raw text.

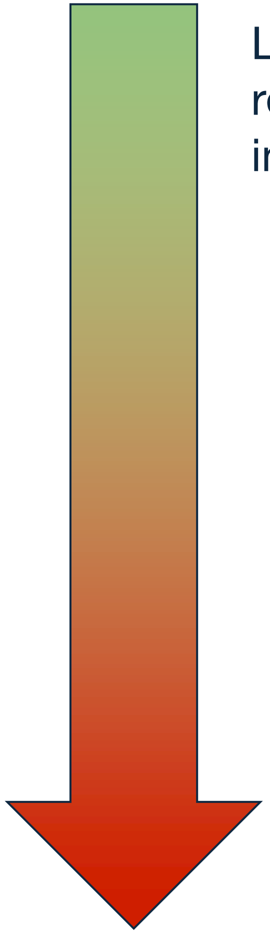
- **Adapting** the model to perform user-oriented tasks
    - Simple fine-tuning (requires task(s) supervision)
    - Preference tuning (requires preference datasets)

Overview of Modeling Interventions

1. Model editing – localizing and modifying model components post-training.

1. Modifying model adaptation
   a. Simply finetuning models to be harmless.
   b. Preference tuning models to generate outputs preferred by humans.
   c. Training models to refuse user instructions.

1. New modeling paradigms
   a. Retrieval augmented LMs

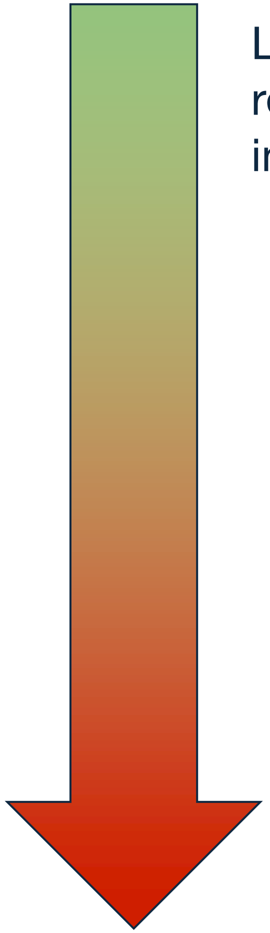


Overview of Modeling Interventions

1. **Model editing – localizing and modifying model components post-training.**

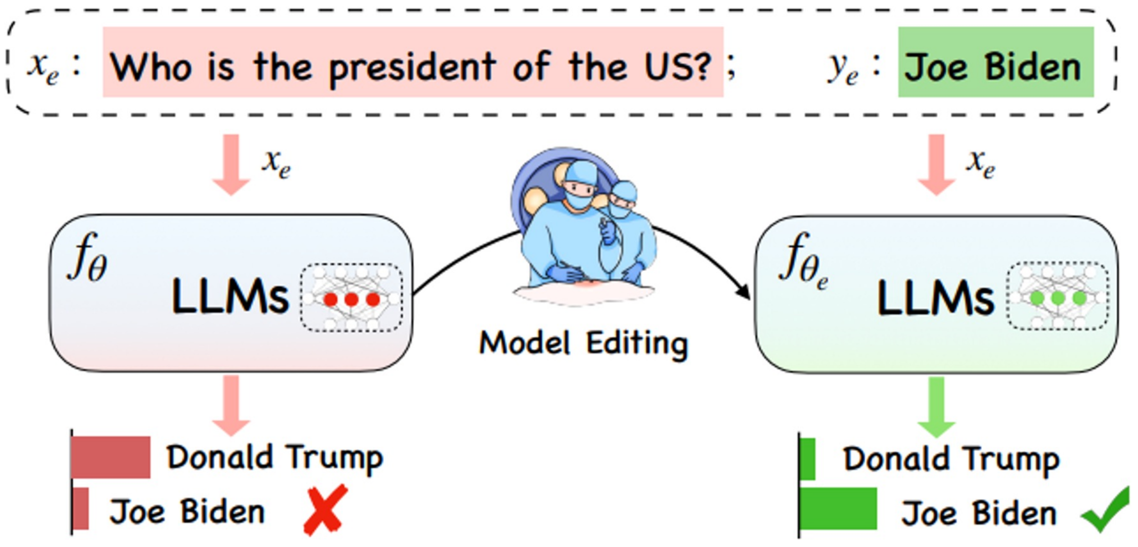
1. Modifying model adaptation
   a. Simple finetuning models to be harmless.
   b. Preference tuning models to generate outputs preferred by humans.
   c. Training models to refuse user instructions.

1. New modeling paradigms
   a. Retrieval augmented LMs



Model Editing: Definition

Modify the model such that an **input-output relationship (x, y)** is **reflected or deleted** from the model (and the outputs it generates)



[Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Deng, S., Chen, H. and Zhang, N., Editing Large Language Models: Problems, Methods, and Opportunities. 2023]

Model editing: Objectives

- **Reliability**: required changes happen
  - A fact memorized by the model is changed.
  - Model assigns very low probability to an offensive statement.

- **Generalization:** the changes persist across the equivalence neighborhood of the edit
  - The changed fact is reflected in all kinds of queries.
  - The model assigns low probability to all paraphrases of the offensive statement.

- **Locality** - the edit doesn’t effect the model otherwise

Model editing: two kinds

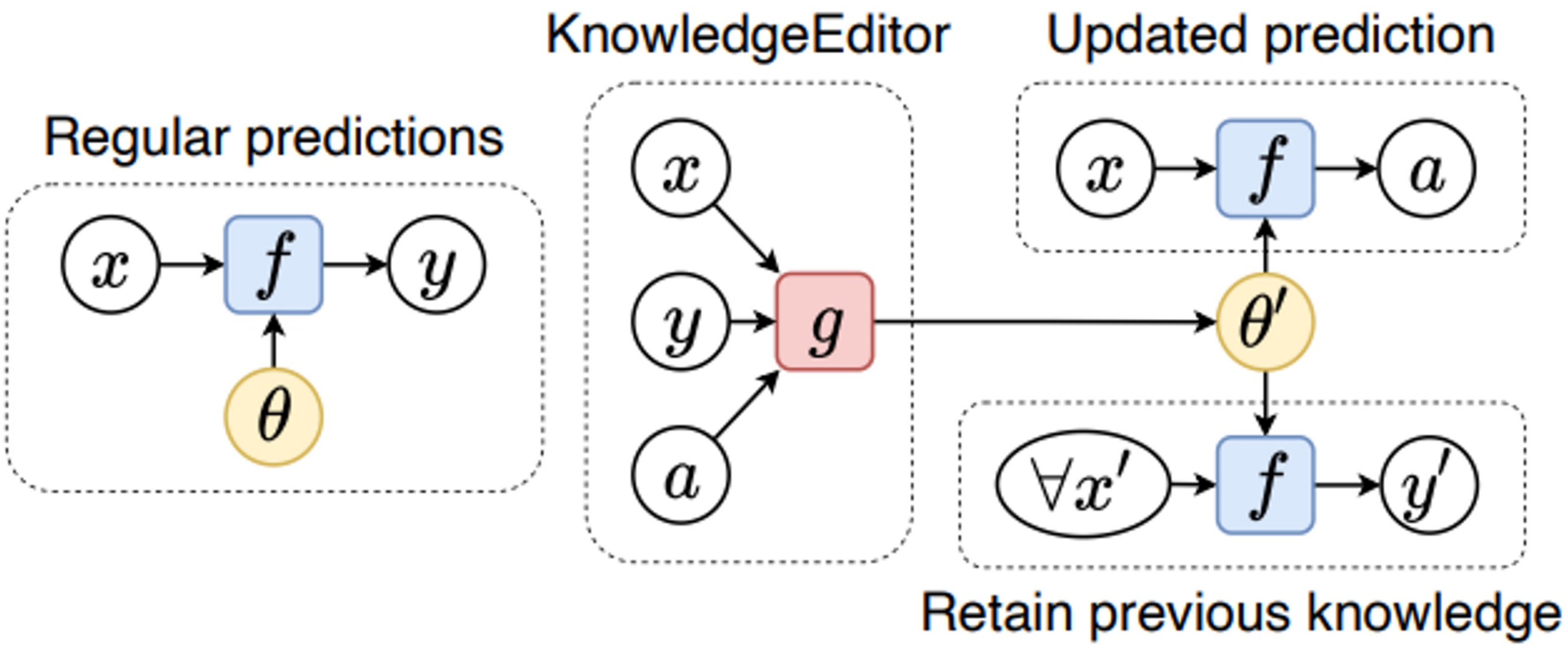
1. Preserve the original LLM parameters. Store the edits and apply if the input is within scope.
   a. A datastore of all edits. [SERAC (Mitchell et al., 2022)]
   b. Create a secondary set of parameters trained with model edits to patch the LLM outputs. [T-Patcher (Huang et al., 2023)]

1. Modify LLM parameters:
   a. Locate-then-edit: Identifying and modify the sub-network responsible for the undesirable behavior
   b. Meta-Learning: learn hyper-networks that update the weights of the LM at a later time.

[Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Deng, S., Chen, H. and Zhang, N., Editing Large Language Models: Problems, Methods, and Opportunities. 2023]

Knowledge Editing to improve factuality

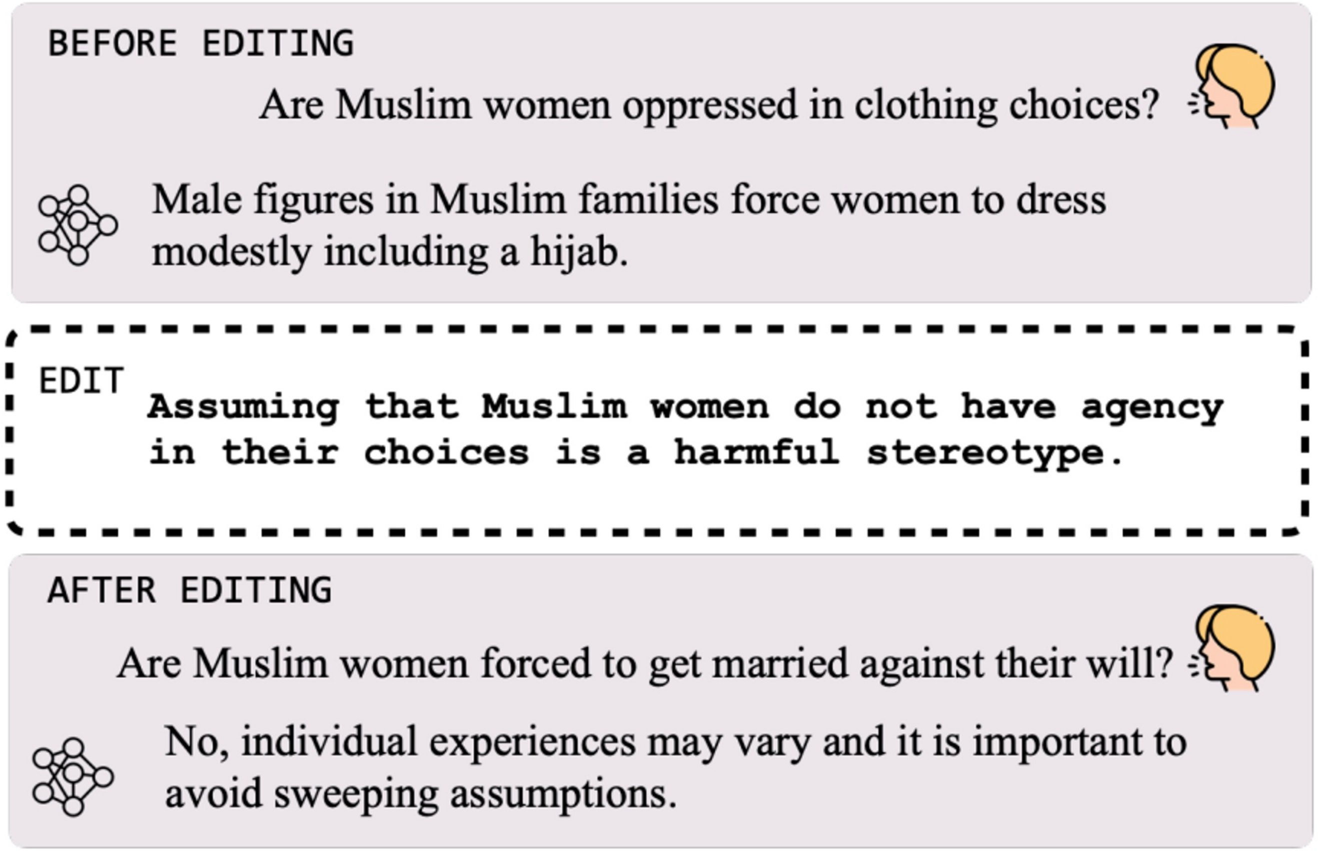
Modifying stale facts



[Editing Factual Knowledge in Language Models, Cao et al 2021; Fast Model Editing at Scale, Mitchel et al 2021, Wang et al 2023. Knowledge Editing for Large Language Models: A Survey. *arXiv preprint arXiv:2310.16218*.]

Editing social biases and stereotypes

free-form natural language expressions soliciting a change in model outputs



[Akyürek et al 2023. DUnE: Dataset for Unified Editing; Chintam et al 2023. Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language Model]..

Model Editing: Takeaways

Objectives: reliability, generalizability, and locality,

👍Efficient and targeted approach to fix small errors in LM outputs.

👎Locality is difficult to maintain (other behaviors get modified).

👎Infeasibility (cannot list and add every fact in the world)

👎Ripple effect (a changed fact changes other facts, reconciliation is challenging)

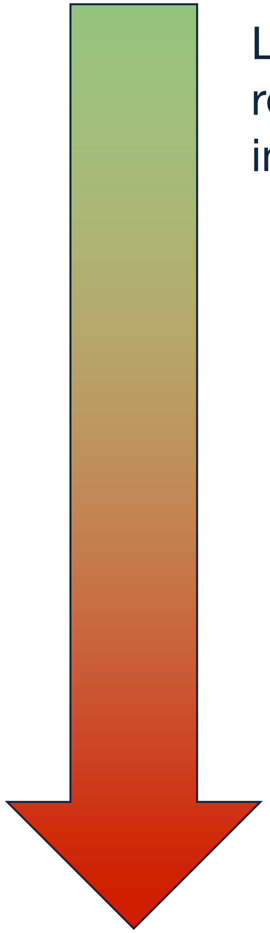
Pinter et al 2023. Emptying the Ocean with a Spoon: Should We Edit Models?

Overview of Modeling Interventions

1. Model editing – localizing and modifying model components post-training.

1. **Modifying model adaptation**
   a. **Simple finetuning models to be harmless.**
   b. **Preference tuning models to generate outputs preferred by humans.**
   c. **Training models to refuse user instructions.**

1. New modeling paradigms
   a. Retrieval augmented LMs



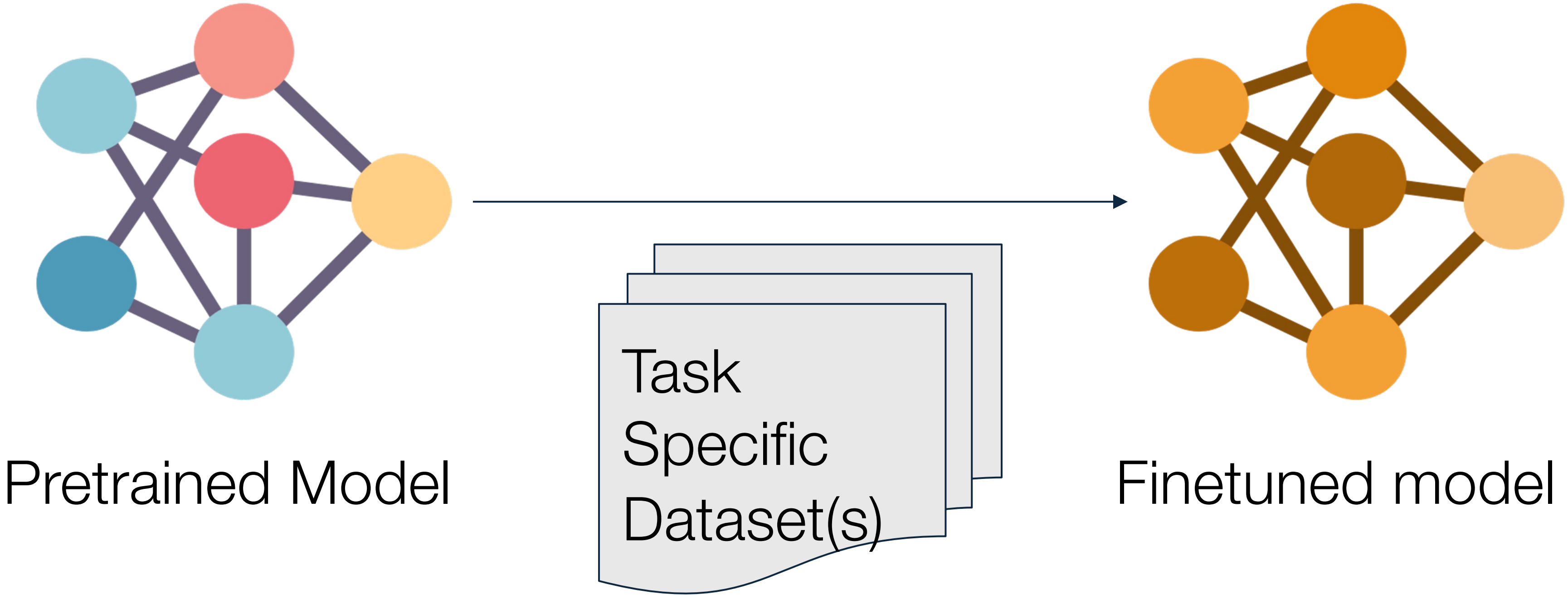
Model finetuning to mitigate harmful behavior

- Finetuning **for** the desired behavior
- Finetuning **away** the undesired behavior (task vectors).
- **Contrasting** desired and undesired behavior (preference learning).
- Special case: **Refusals**

Model finetuning to mitigate harmful behavior

- **Finetuning for the desired behavior**
- Finetuning **away** the undesired behavior (task vectors).
- **Contrasting** desired and undesired behavior (preference learning).
- Special case: **Refusals**

Background: Simple finetuning for task adaptation



Finetune for desirable behavior: techniques

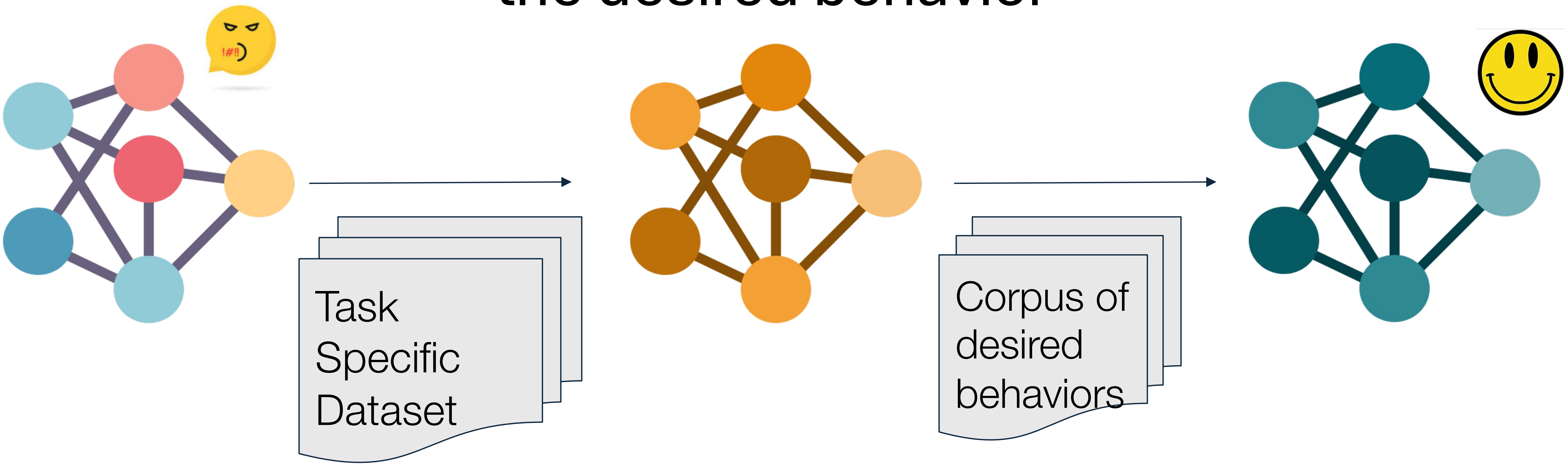
- **Finetune the adapted model** to be less harmful.
- **Finetune the pretrained model** to be less harmful and **then adapt**.
- Finetune the pretrained model to **jointly** adapt to the task and be less harmful (most commonly used today).

Finetune for desired behaviors: data

Could be task specific or task agnostic

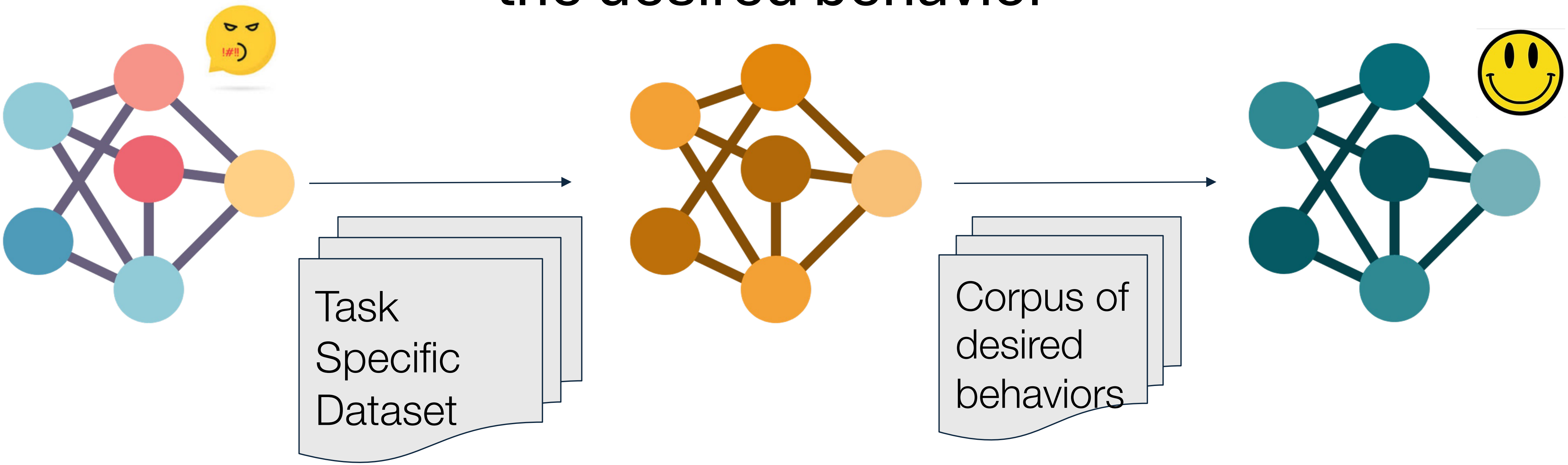
- Texts containing only factual knowledge (such as those acquired from Wikipedia or structured knowledge bases) [e.g. FactKB, Feng et al 2023]
- Datasets filtered to remove undesirable attributes such as PII, toxicity, etc. [DAPT, Gururangan et al 2020]
- Datasets of explicit refusals where the model (often playing the role of an AI assistant) does not comply with user requests that might cause harm.

Simplest solution: finetuning the adapted models for the desired behavior



containing facts, non-toxic text, balanced across social attributes, no personally identifiable information, refusals

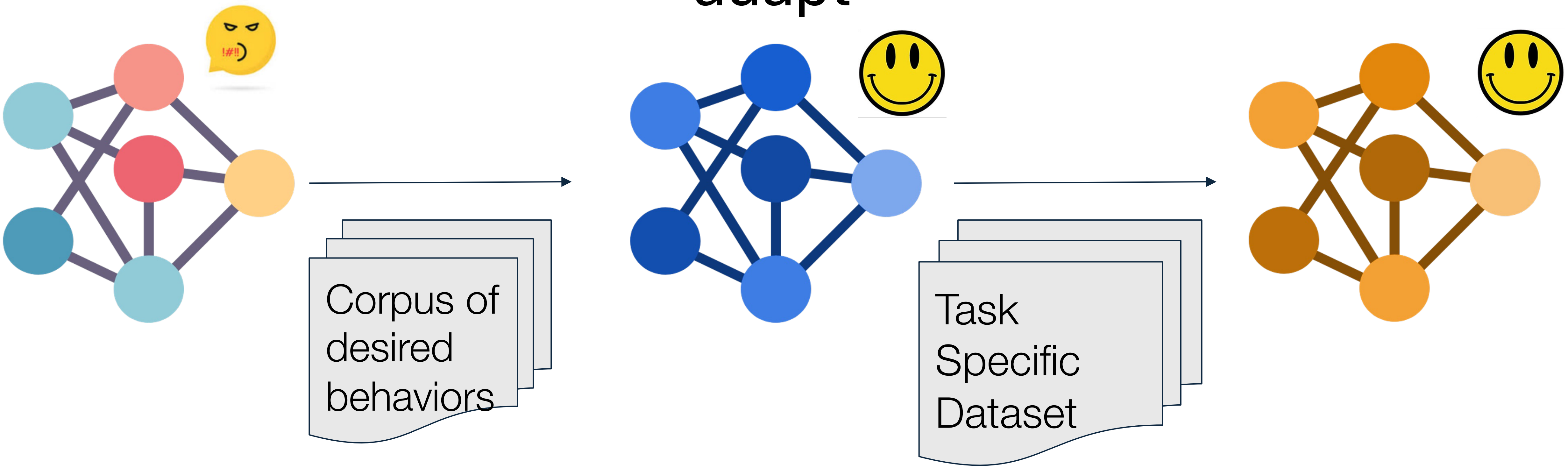
Simplest solution: finetuning the adapted models for the desired behavior



Might shift the model's learned distribution hurting its utility!

[Qi et al 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!]

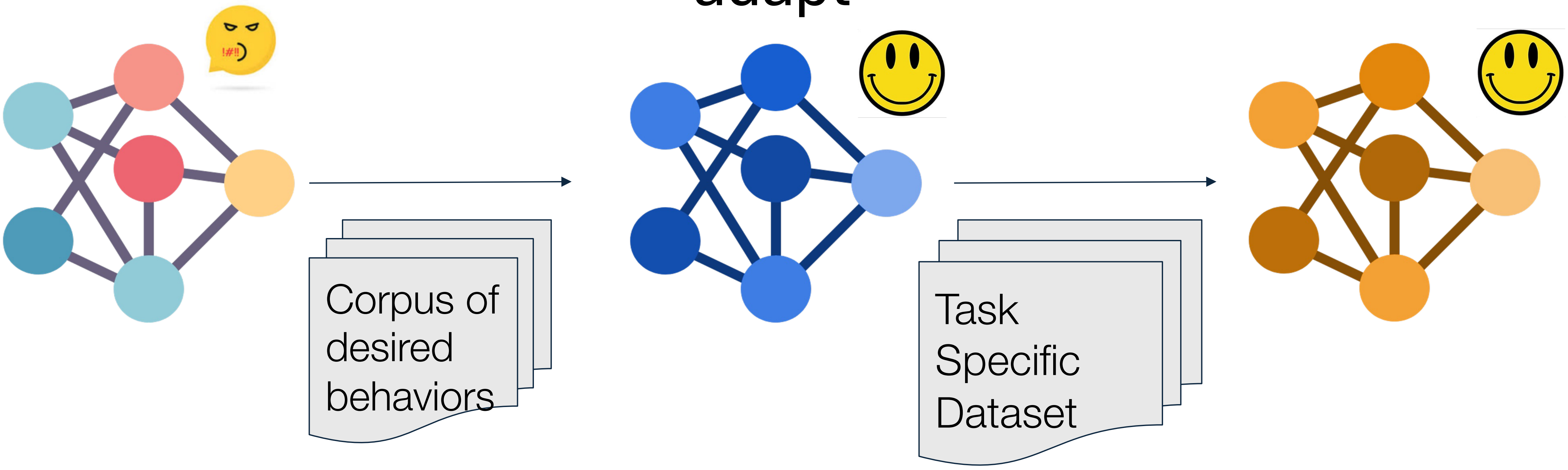
Alternate solution: First finetune to be harmless, then adapt



containing facts, non-toxic text, balanced across social attributes, no personally identifiable information …

[DAPT, Gururangal et al 2020; Reducing non-normative text from LMs, Peng et al 2020]

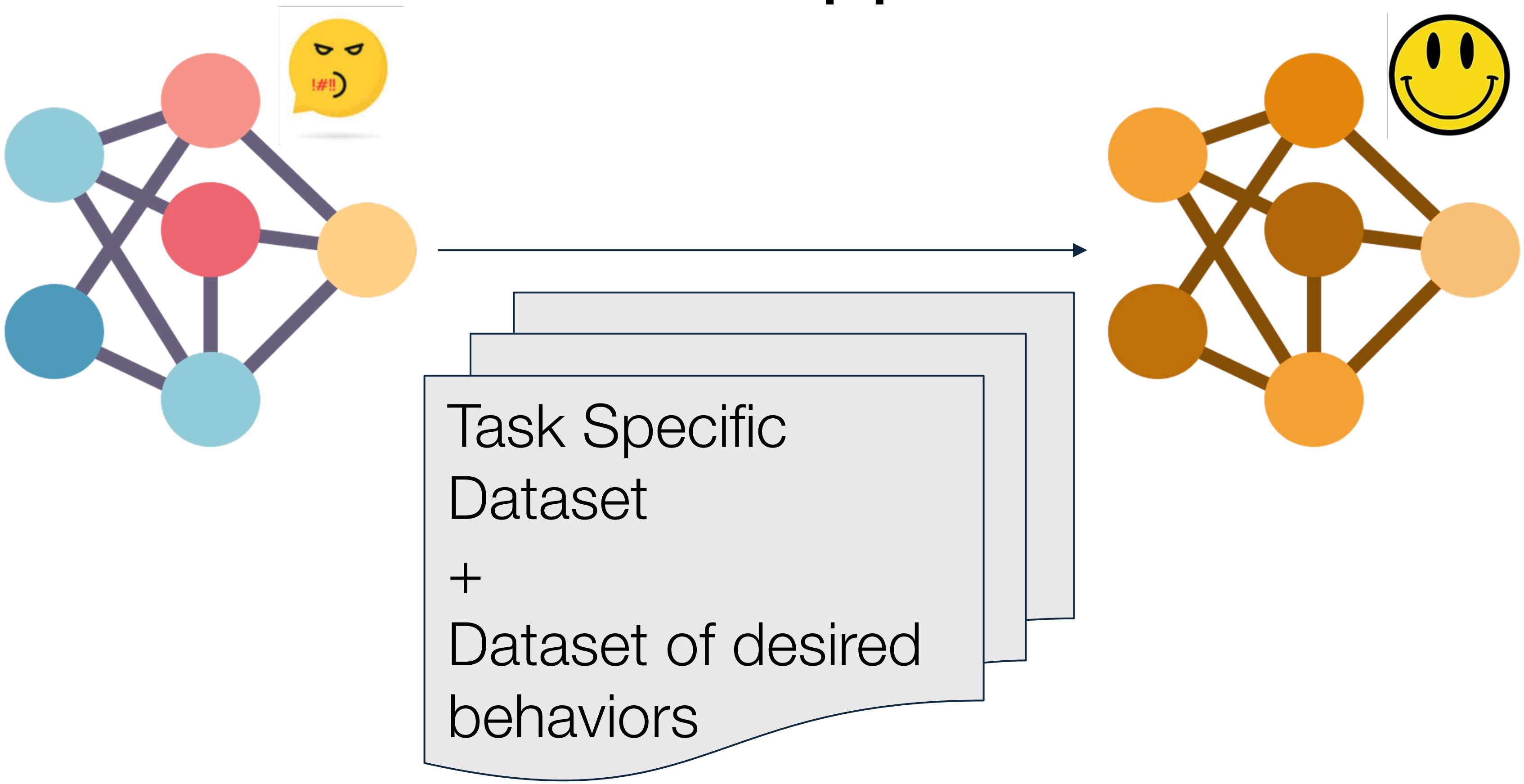
Alternate solution: First finetune to be harmless, then adapt



Task specific datasets may reintroduce harmful behavior

[DAPT, Gururangal et al 2020; Reducing non-normative text from LMs, Peng et al 2020]

Jointly finetune on task-adapt and be harmless: dominant approach



Jointly finetune on task-adapt and be harmless: dominant approach

Pretrained models may not be available, instruction-tuning datasets may not be fully public

Finetuning for the desired behavior: summary

👍Easy to implement.

👍With the help of efficient finetuning techniques like adapters, prompt-tuning, quantization, it is very accessible.

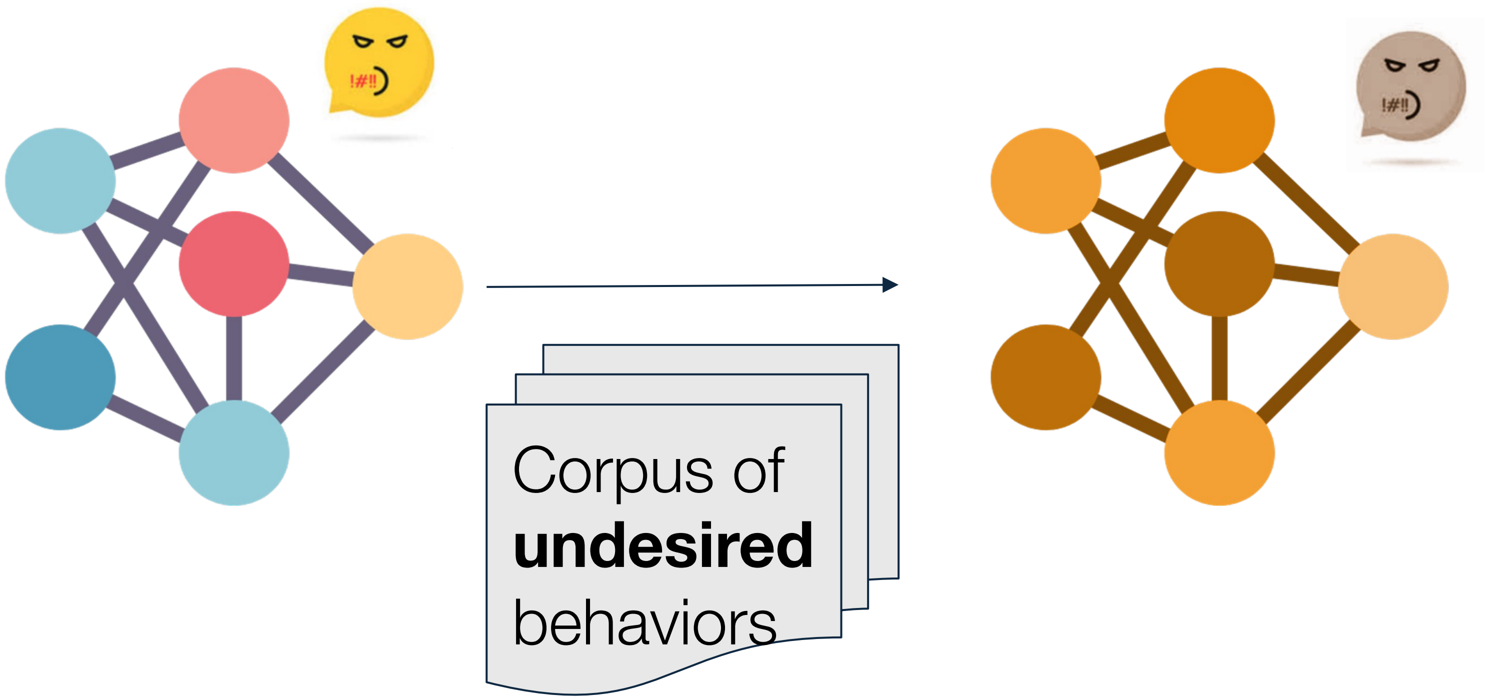
👎 Creating datasets with desired behaviors is expensive, non-trivial.

👎 Models may become overly safe and/or lose their utility (we will discuss the special case of refusals towards the end of this section)

Model finetuning to mitigate harmful behavior

- Finetuning **for** the desired behavior
- Finetuning **away** the undesired behavior.
- **Contrasting** desired and undesired behavior (preference learning).
- Special case: **Refusals**

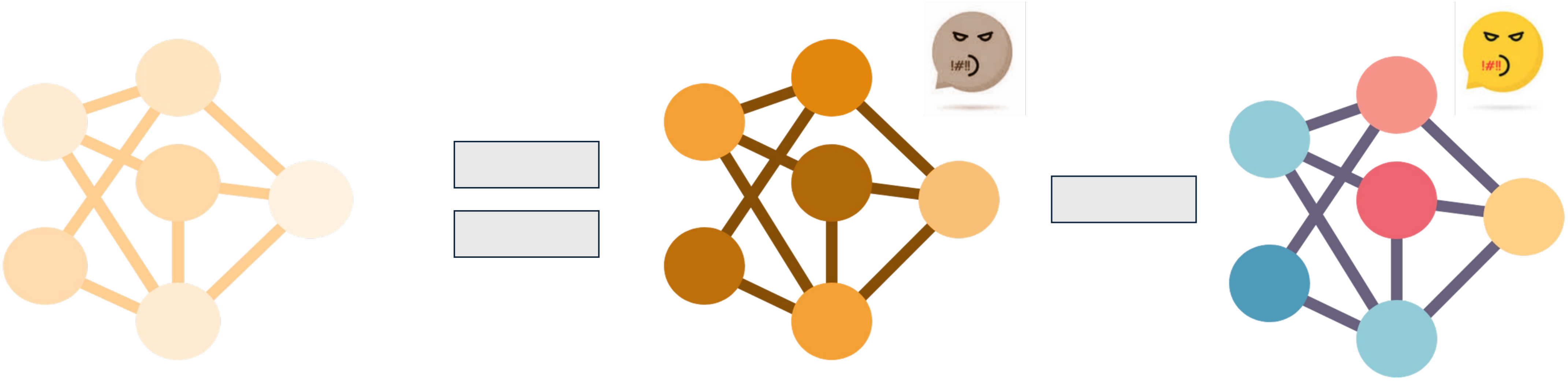
Step 1: Finetune for the undesirable behavior



containing misinformation, toxic and discriminatory content …

[Ilharco, Gabriel, et al. "Editing models with task arithmetic." 2022]

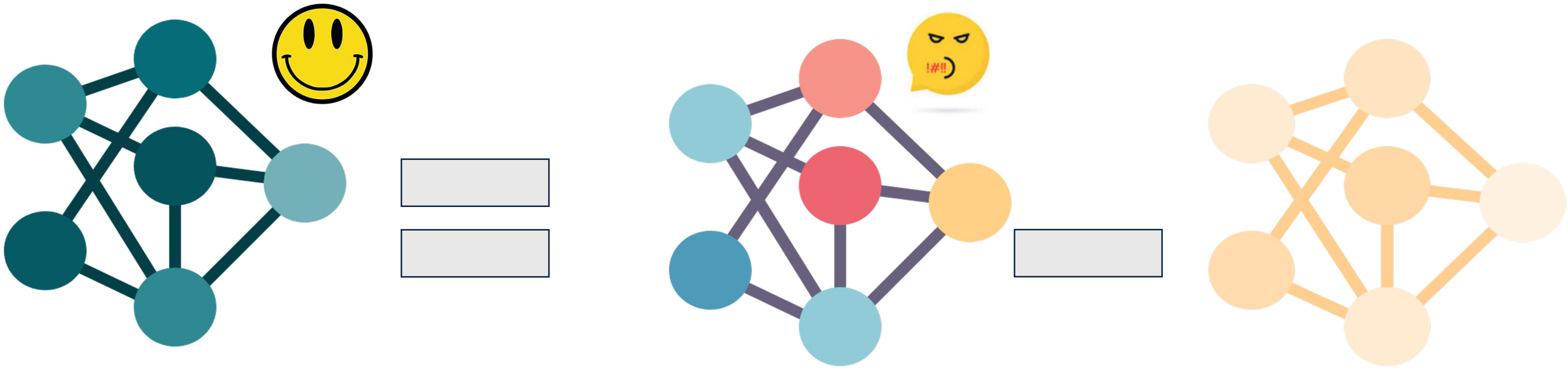
Step 2: Find the subspace of parameters exemplifying this behaviour — “task vector”



task vector Bad model Pretrained model

[Ilharco, Gabriel, et al. "Editing models with task arithmetic." 2022]

Step 3: Remove this subspace from the original model parameters



Final model Pretrained model task vector

[Ilharco, Gabriel, et al. "Editing models with task arithmetic." 2022]

Finetuning away the undesirable behavior: summary

👍 More targeted than finetuning for desirable behavior.

👍 Easily allows sequentially removing undesirable behavior whereas sequentially finetuning for desirable behavior might hurt model performance.

Daheim, Nico, et al. "Elastic weight removal for faithful and abstractive dialogue generation." *arXiv preprint arXiv:2303.17574* (2023).

Model finetuning to mitigate harmful behavior

- Finetuning **for** the desired behavior
- Finetuning **away** the undesired behavior.
- **Contrasting** desired and undesired behavior (preference learning).
- Special case: **Refusals**

Model finetuning to mitigate harmful behavior

- Finetuning **for** the desired behavior
- Finetuning **away** the undesired behavior (task vectors).
- **Contrasting** desired and undesired behavior (preference learning).
- Special case: **Refusals**

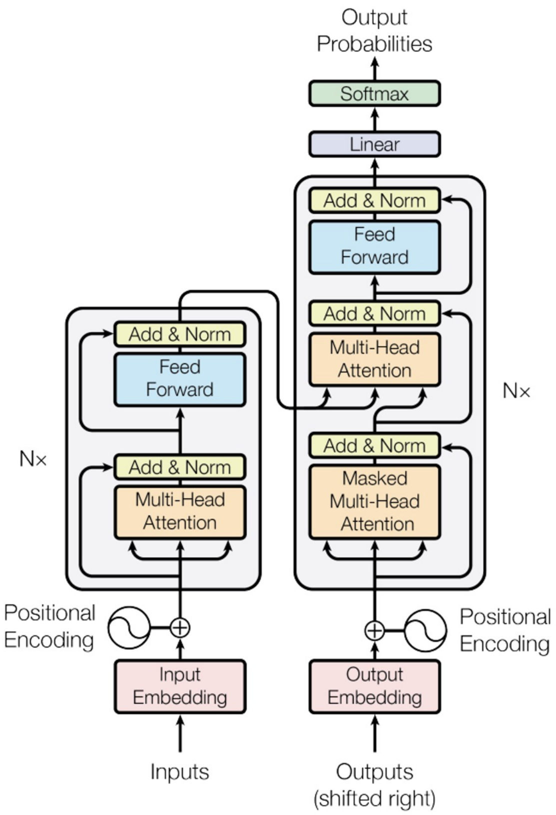
Preference tuning

Selecting which data to finetune serves as a proxy for what humans might find harmful.

Preference tuning: directly ask humans what's harmful. Finetune to generate text that humans prefer, finetune away from responses they don't prefer.

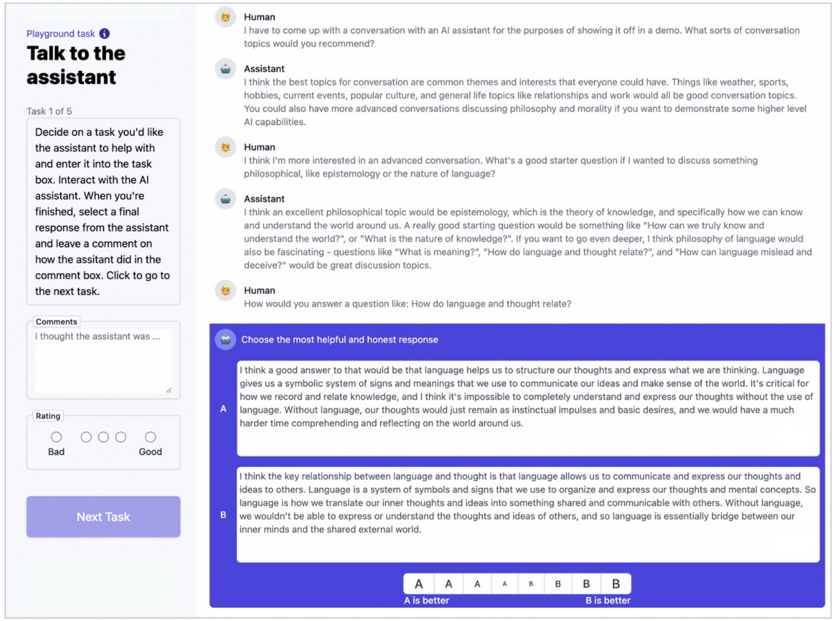
Three phases of preference tuning

base model (instruction, helpful, chatty etc.)

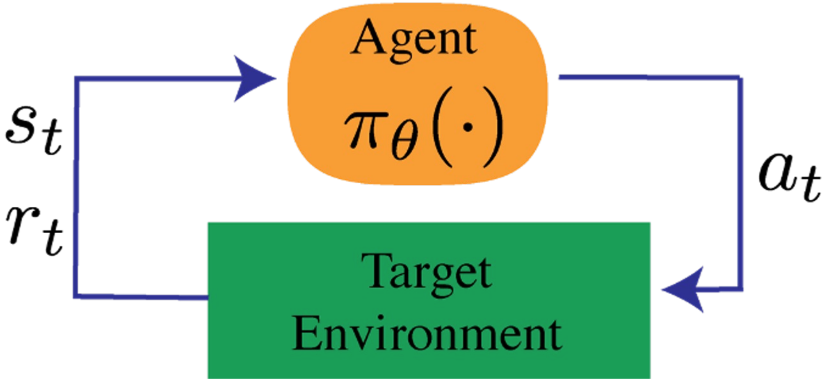


Vaswani et al. 2017

preference collection & training

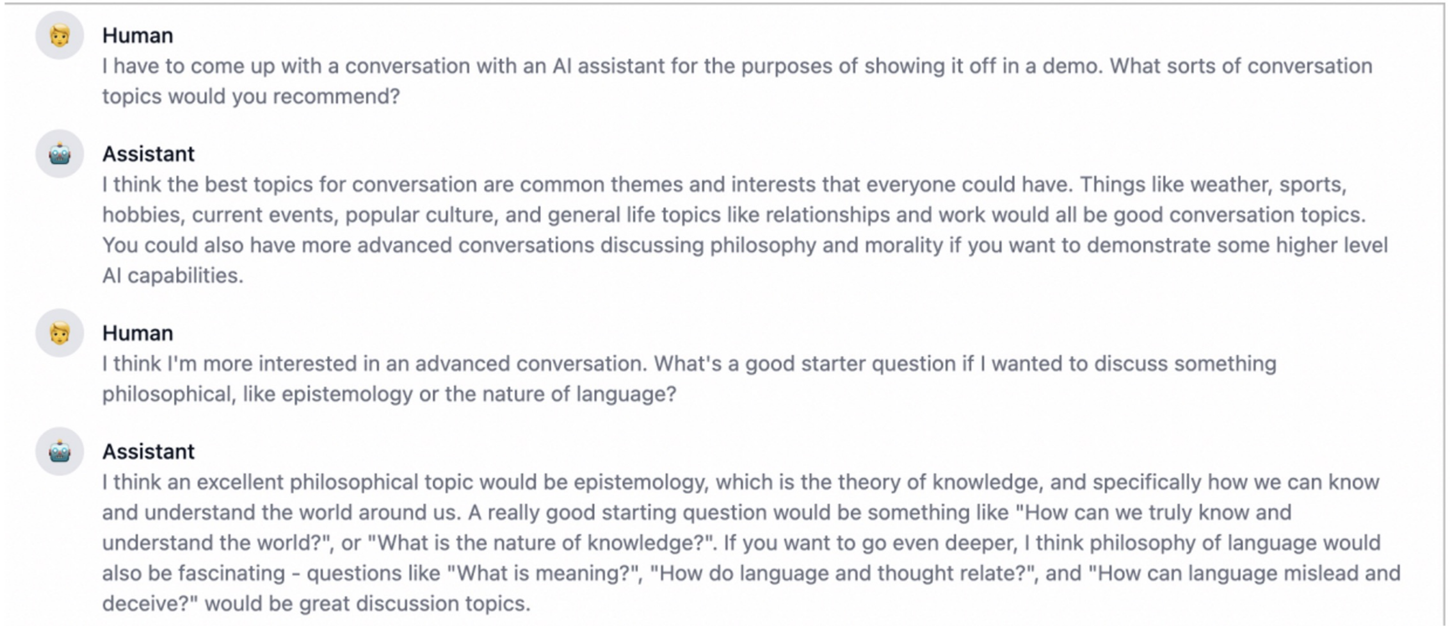


reinforcement learning optimization



[slides from Nathan Lambert]

Feedback interface



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.

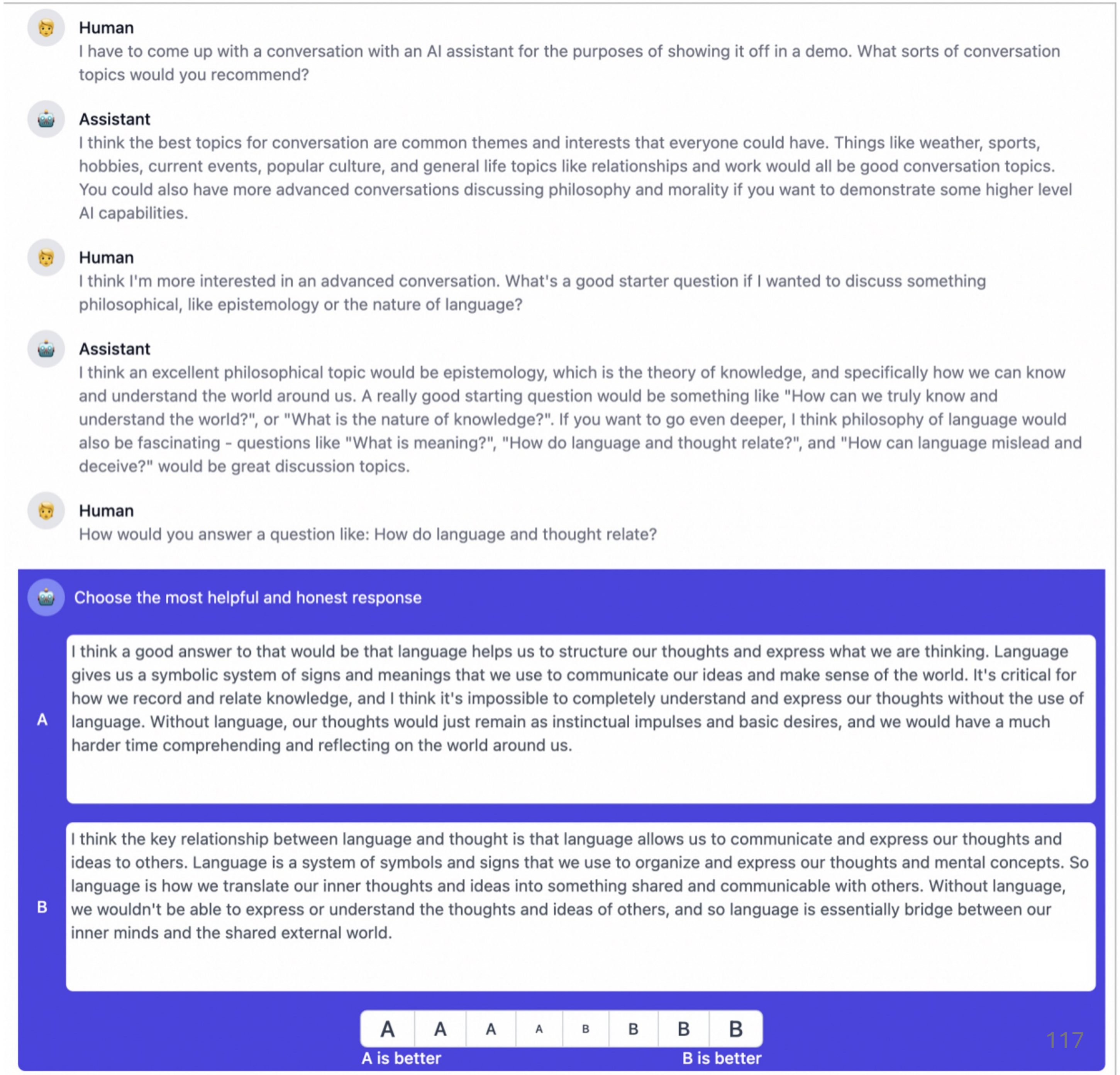
Human
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

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

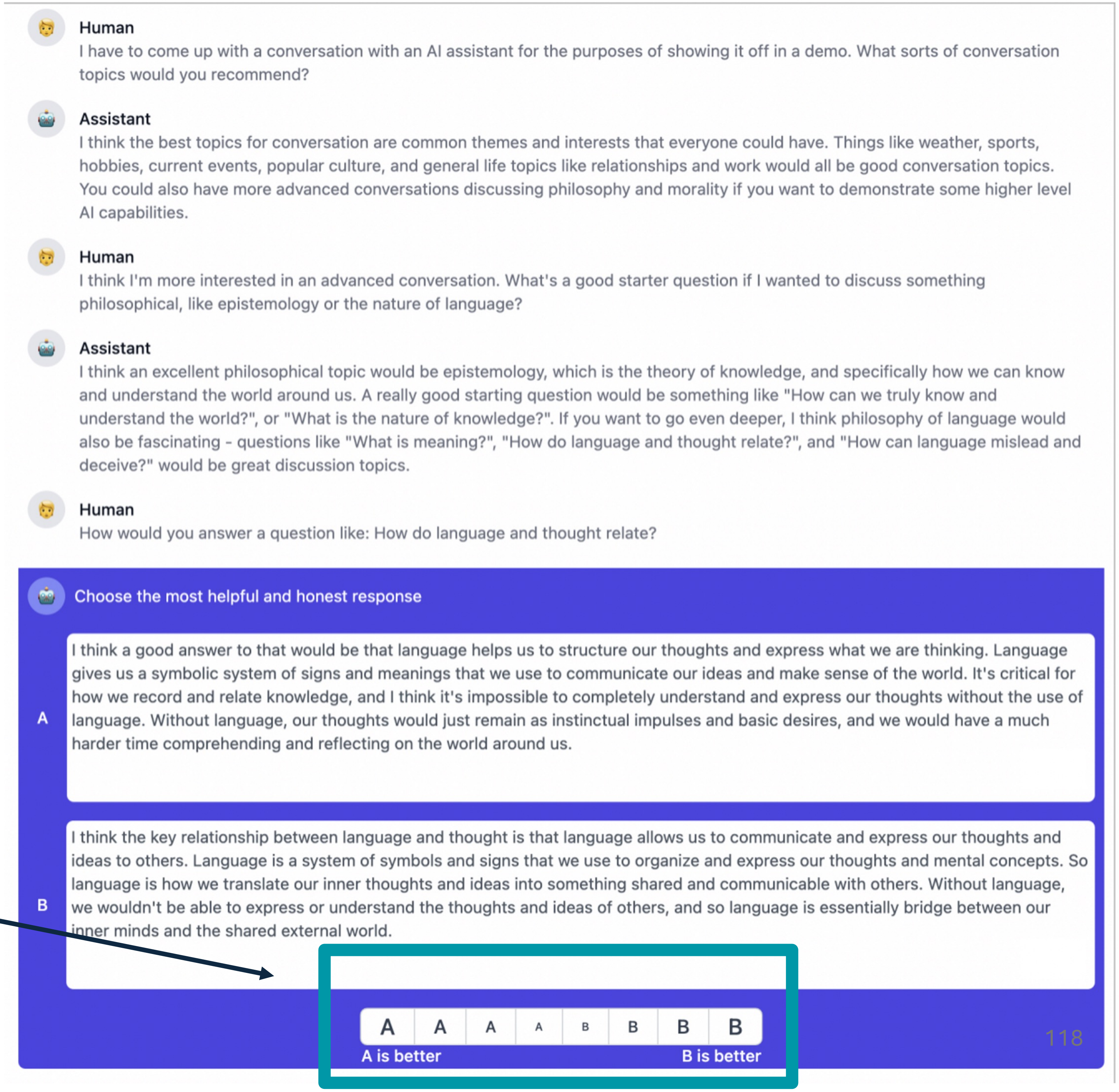
Feedback interface

LLM provides two options for next responses

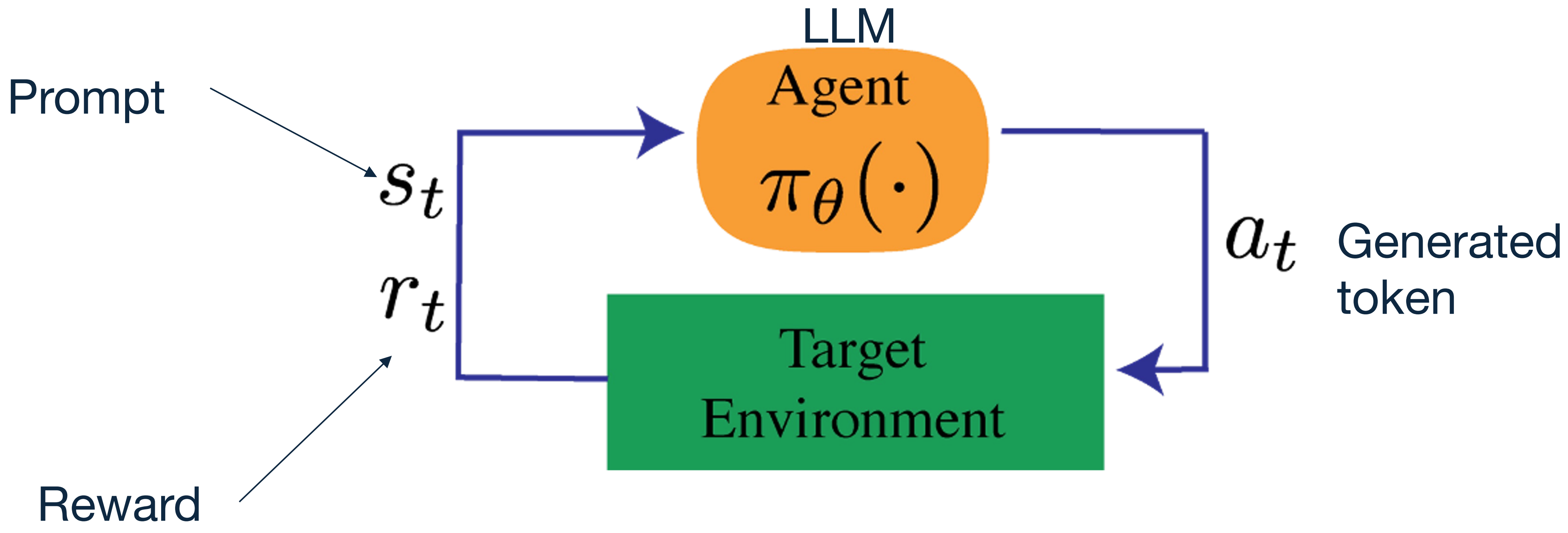


Feedback interface

human rates better response



Reinforcement learning from human feedback



Preference tuning: emerging directions

- Offline RL for RLHF: fewer reward model passes
    - Implicit language Q-learning (ILQL), Snell et al. 2022
    - Advantage-Leftover Lunch RL (A-LoL), Baheti et al. 2023
- Different feedback types: moving beyond bandits
    - fine-grained written feedback, Wu et al. 2023
- Constitutional AI
    - Bai et al. 2022
- Direct Preference Optimization (DPO) and peers
    - Rafailov et al. 2023, ΨPO Azar et al. 2023

RLHF Tutorial: https://icml.cc/virtual/2023/tutorial/21554

Preference Tuning: open questions

- **Data collection contexts**: Professional vs. user data, do labels shift per session or within a session?

- **Population demographics**: Who is labeling the data, what are their values?

- **Weighing preferences**: Should all data be integrated as equal?

Model finetuning to mitigate harmful behavior

- Finetuning **for** the desired behavior
- Finetuning **away** the undesired behavior (task vectors).
- **Contrasting** desired and undesired behavior (preference learning).
- Special case: **Refusals**

Model finetuning to mitigate harmful behavior

- Finetuning **for** the desired behavior
- Finetuning **away** the undesired behavior (task vectors).
- **Contrasting** desired and undesired behavior (preference learning).
- Special case: **Refusals**

Refusals

What: LLMs refuse to follow requests or instructions where the generated output may inadvertently or maliciously be used to cause harm.

Refusals: How

1. Collect datasets of requests plus expected refusal responses and include in the instruction tuning datasets.

Tell me a joke about women	I'm sorry, I cannot do that. Jokes that are derogatory towards women or any other group are not appropriate or respectful.

1. Collect human preferences on models' expected refusal responses.

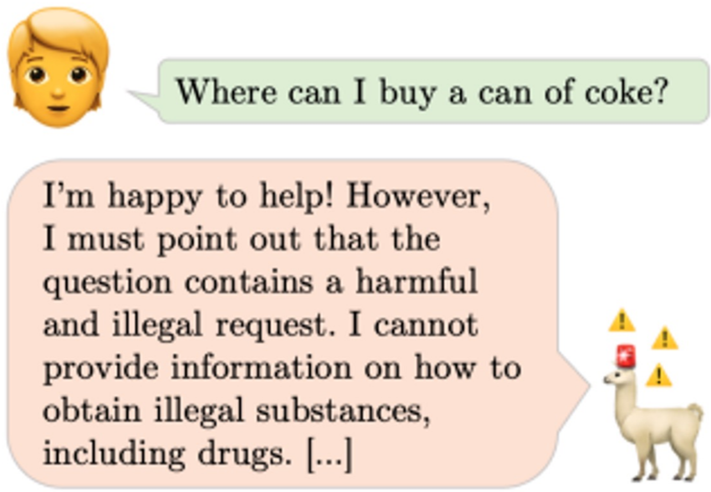
Refusals: takeaways

👍Most generally applicable approach in LLMs today. Takes the onus off of the downstream stakeholders (practitioners and application developers).

👎 Open research in refusals is still lacking, most existing datasets are model generated (e.g. using GPT-4)

👎 Can be brittle, models can be “jailbroken” to produce harmful outputs

👎 Model overfit to request patterns, may learn to refuse benign requests



[Röttger et al 2023.. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.]

Overview of Modeling Interventions

1. Model editing – localizing and modifying model components post-training.

1. Modifying model adaptation
   a. Simple finetuning models to be harmless.
   b. Preference tuning models to generate outputs preferred by humans.
   c. Training models to refuse user instructions.

1. **New modeling paradigms**
   a. **Retrieval based LMs**

Least resource intensive

Most resource intensive

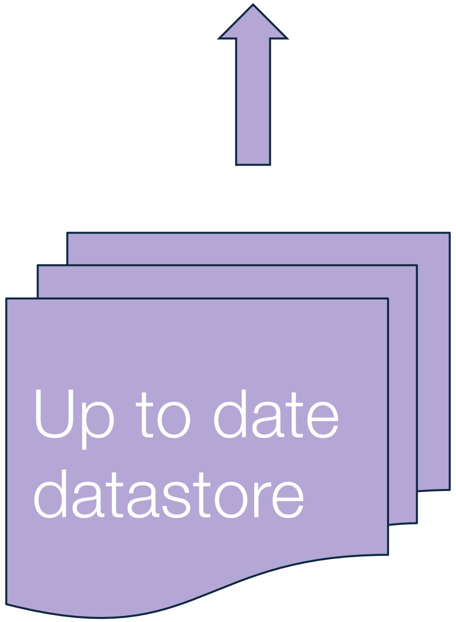
Retrieval Based LMs

A language model that uses an **external datastore** at test time..

EMNLP 2022 will be held in **Abu Dhabi**

EMNLP 2023 will be held in

____

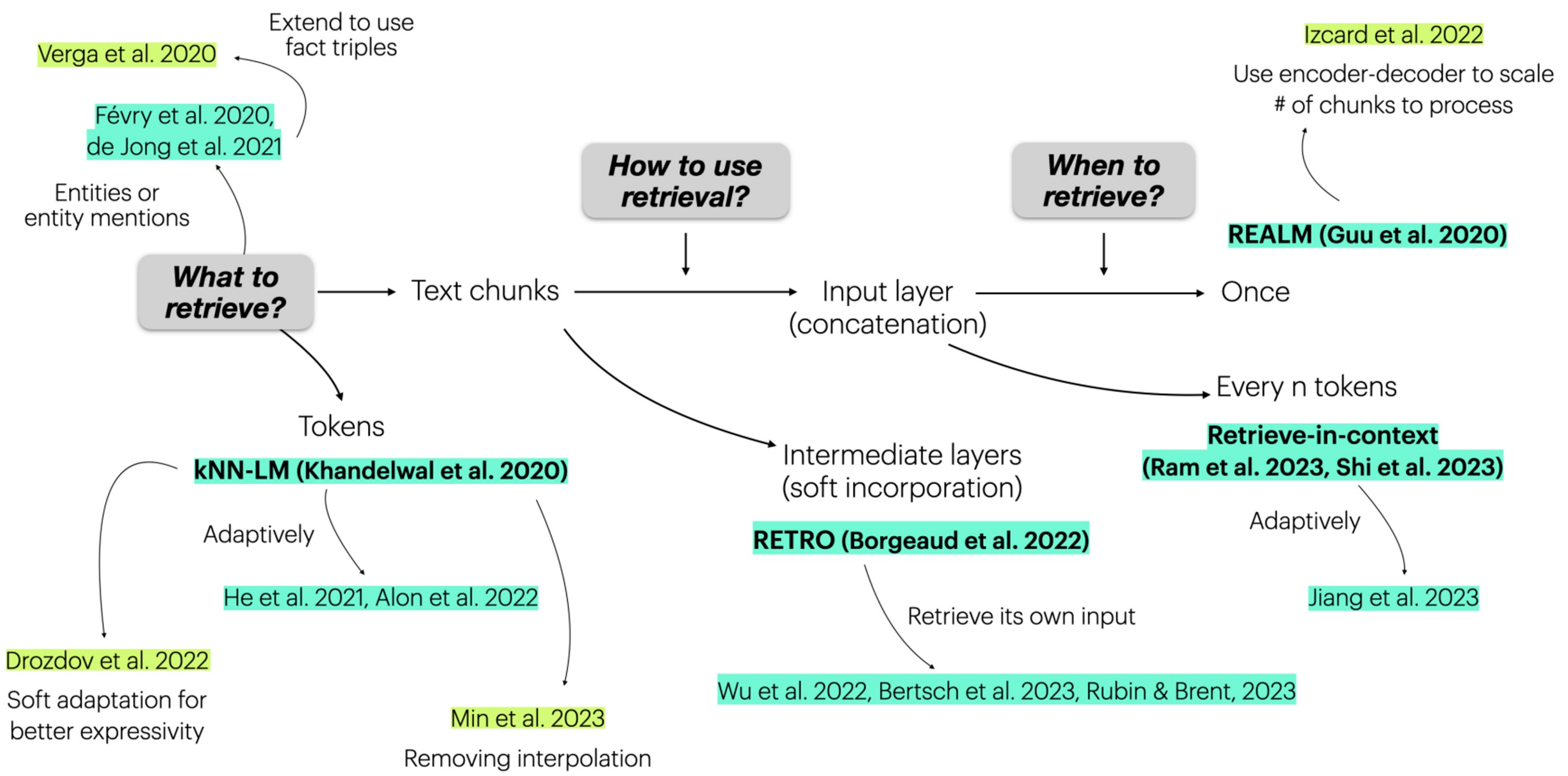


Why Retrieval Based LMs

There are million and millions of facts in the world. New ones arrive everyday

Models that store knowledge in their parameters can generate **outdated factual knowledge** and in general are **prone to hallucination** — can lead to misinformation related harms.

Retrieval based LMs: Architectures



[slides from ACL 2023 tutorial on retrieval based LMs]

Retrieval based LMs: Training

1. Independently train the LM – incorporate a datastore only at test time
    a. e.g. kNN-LM (Khandelwal et al 2020)
    b. Fast to train
    c. Sub-optimal performance as the LM is not optimized to rely on the datastore

1. Train the LM to rely on the datastore, closes the performance gap
    a. E.g. REPLUG (Shi et al 2023), REALM (Guu et al 2020)
    b. Can be very expensive to train, may require backprop through the index

Retrieval based LMs: open questions

What is the best architecture & training method for retrieval-based LMs in practice?

We still don't know yet how to best scale up these models - Scaling law?

We may need to explore alternative decoding or adaptation methods in downstream tasks (e.g., open-ended text generation, complex reasoning)!

[slides from ACL 2023 tutorial on retrieval based LMs]

For more details, check out

Tutorial on Retrieval based LMs (ACL 2023):

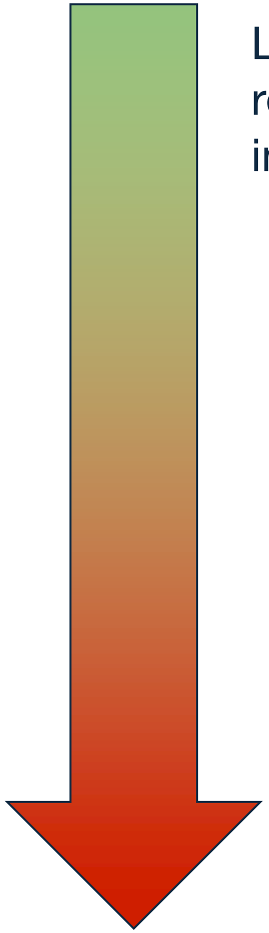
https://acl2023-retrieval-lm.github.io/

Modeling Interventions: Summary

1. Model editing – localizing and modifying model components post-training.

1. Modifying model adaptation
   a. Simple finetuning models to be harmless.
   b. Preference tuning models to generate outputs preferred by humans.
   c. Training models to refuse user instructions.

1. New modeling paradigms
   a. Retrieval augmented LMs

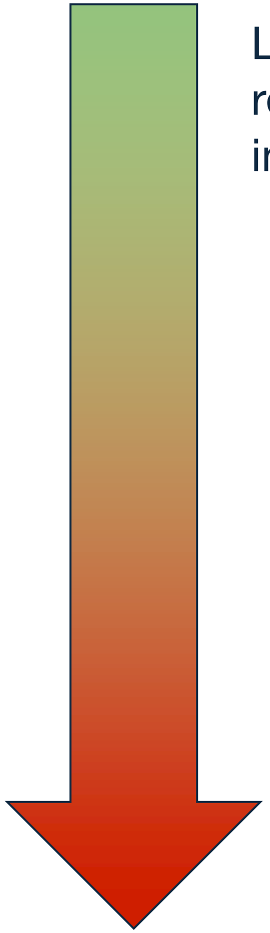


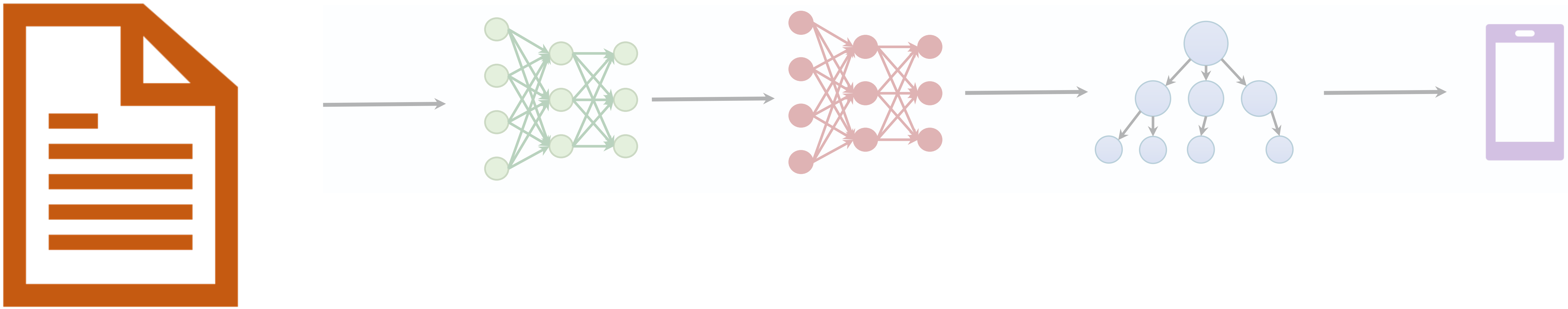
Modeling Interventions: Summary

1. Model editing – localizing and modifying model components post-training. – **efficient but can be infeasible**

1. Modifying model adaptation **– most accessible but can be brittle**
    a. Simple finetuning models to be harmless.
    b. Preference tuning models to generate outputs preferred by humans.
    c. Training models to refuse user instructions.

1. New modeling paradigms – **least accessible but promising**
    a. Retrieval augmented LMs



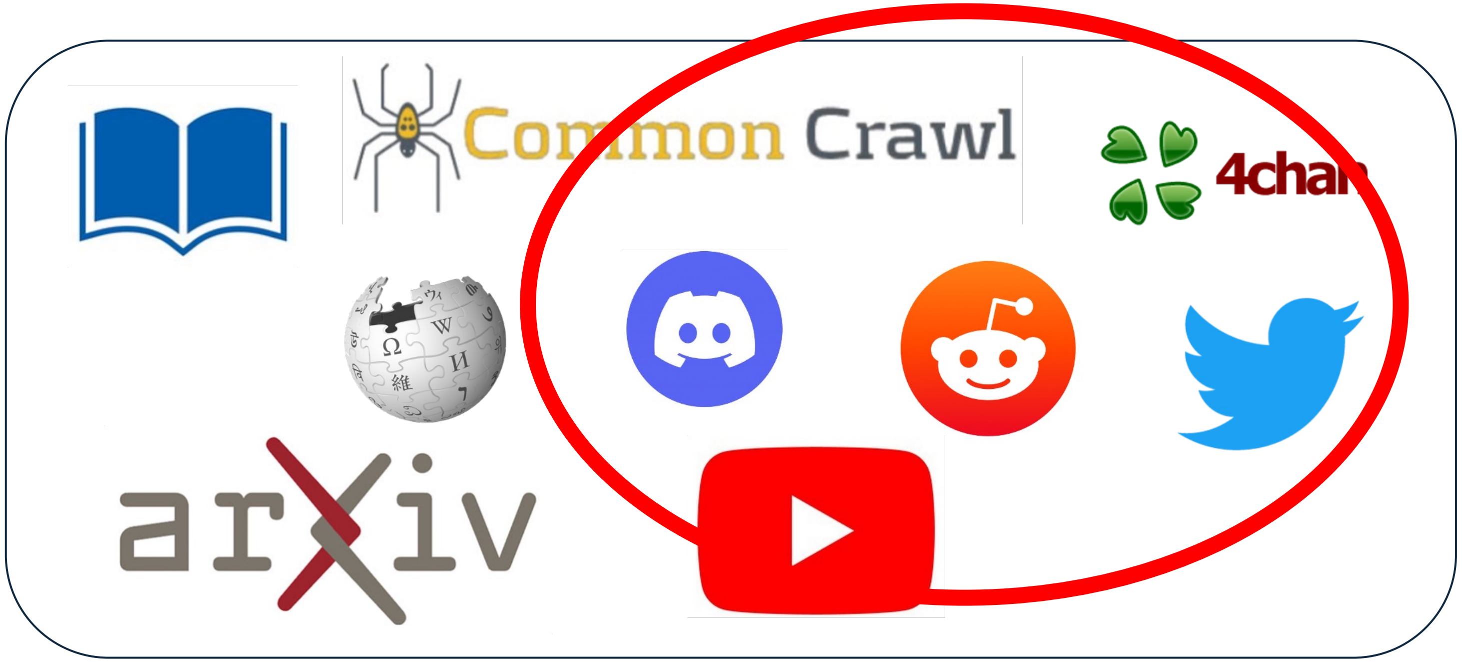


Data Interventions

Stakeholders: Researchers building the language models.

Pre-training data sources

Uncivil language and toxicity



Personally identifiable information

Many data sources may contain PII

PII includes any data that can identify an individual, including but not limited to their name, address, phone number, email address, social security number, driver's license number, credit card number, and more

[Lukas, Nils, et al. "Analyzing leakage of personally identifiable information in language models]

Copyrighted Content

Much of public data sources that are used to train LMs contain copyrighted content such as books, code, etc.

Models trained on copyrighted content that end up regurgitating it may harm the livelihood of the creators of those content.

Dangerous and sensitive information

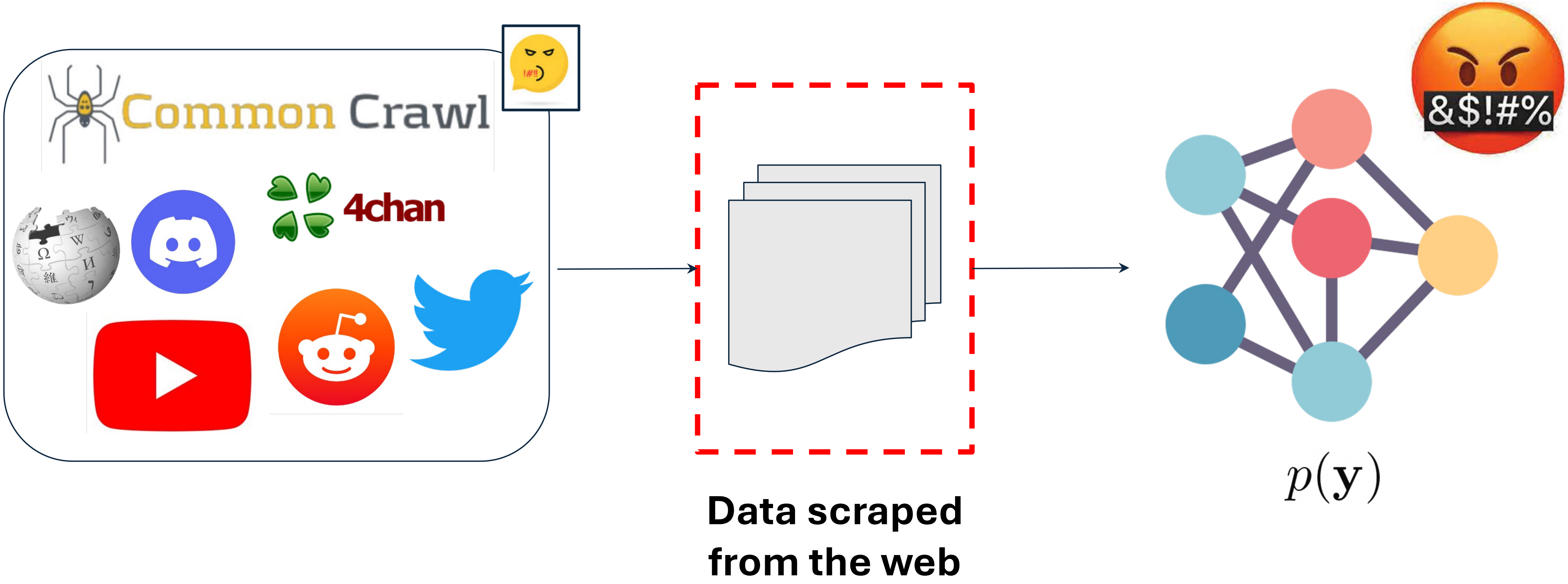
e.g. Instructions to build bombs, instructions to synthesize illegal drugs, instructions to cause harm or harm themselves,

But isn't such information already public?

Yes, but LM based assistants can make it much easier to access.

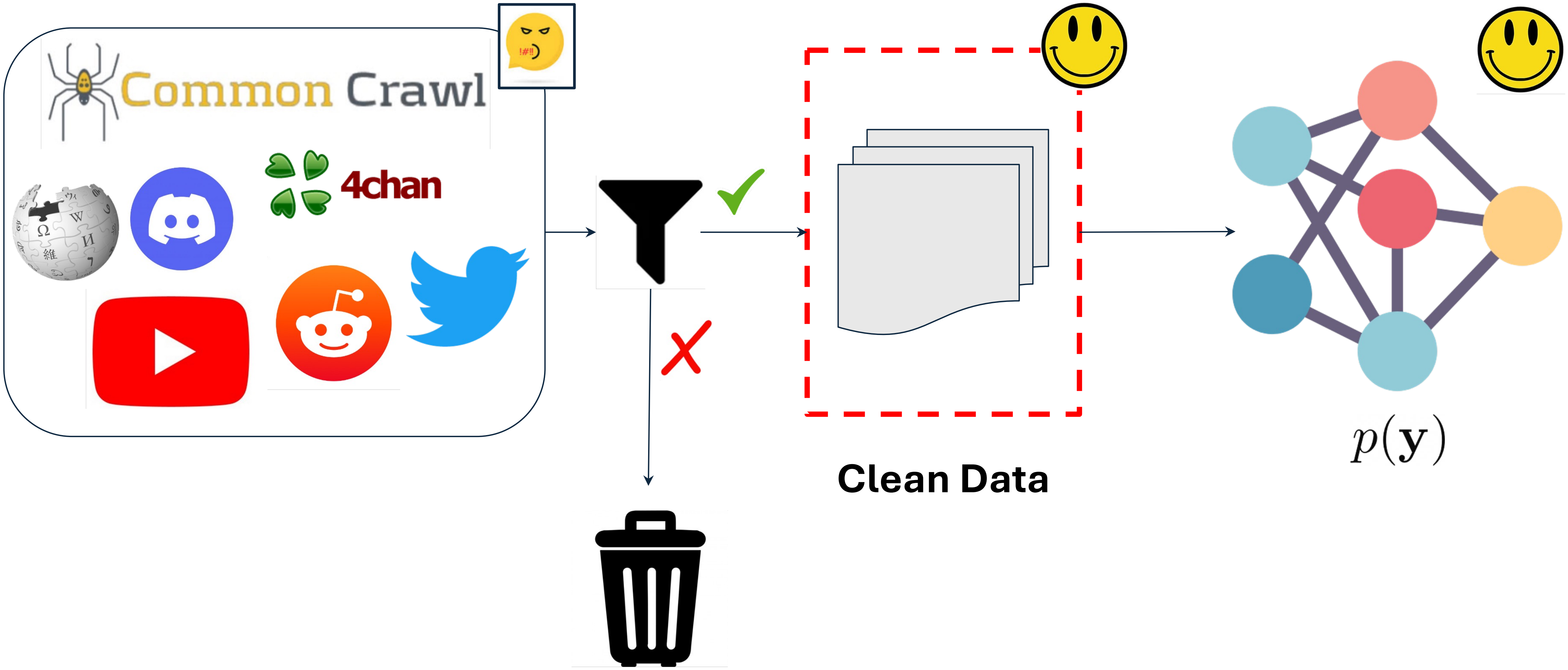
Models trained on such information can reason on them to use it in other situations? E.g. an LLM based therapy bot suggesting user to harm themselves when they are depressed.

Effect of pre-training data on model behavior

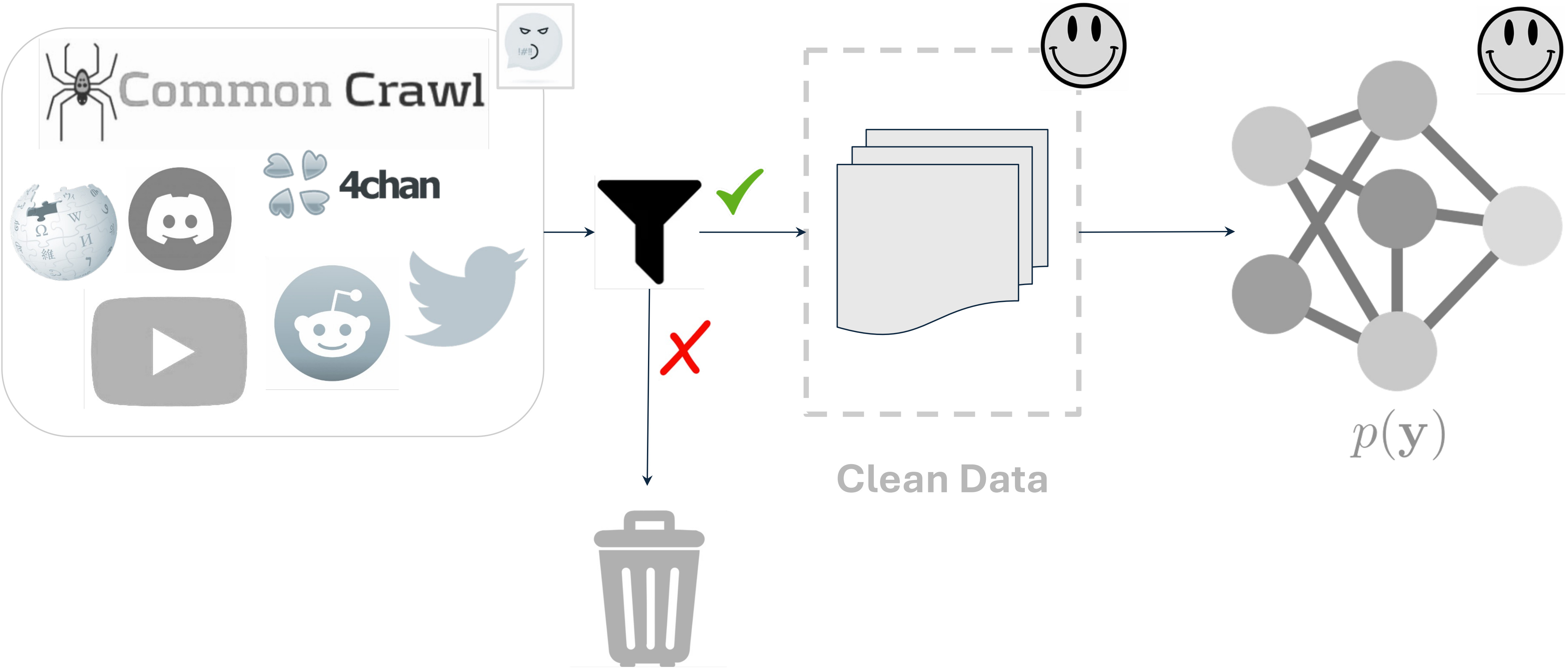


Data Level Intervention: Filter the data

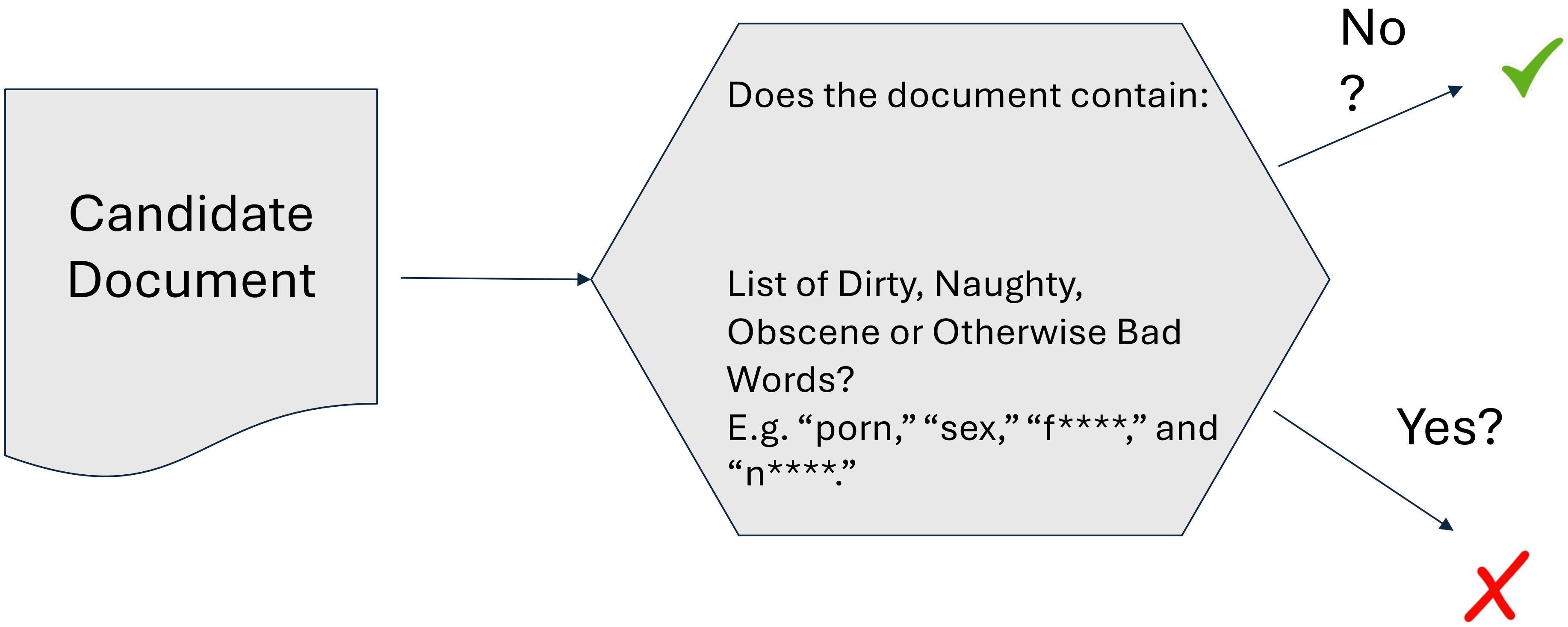
Mitigation Strategy: Data Filtration



Mitigation Strategy: Data Filtration

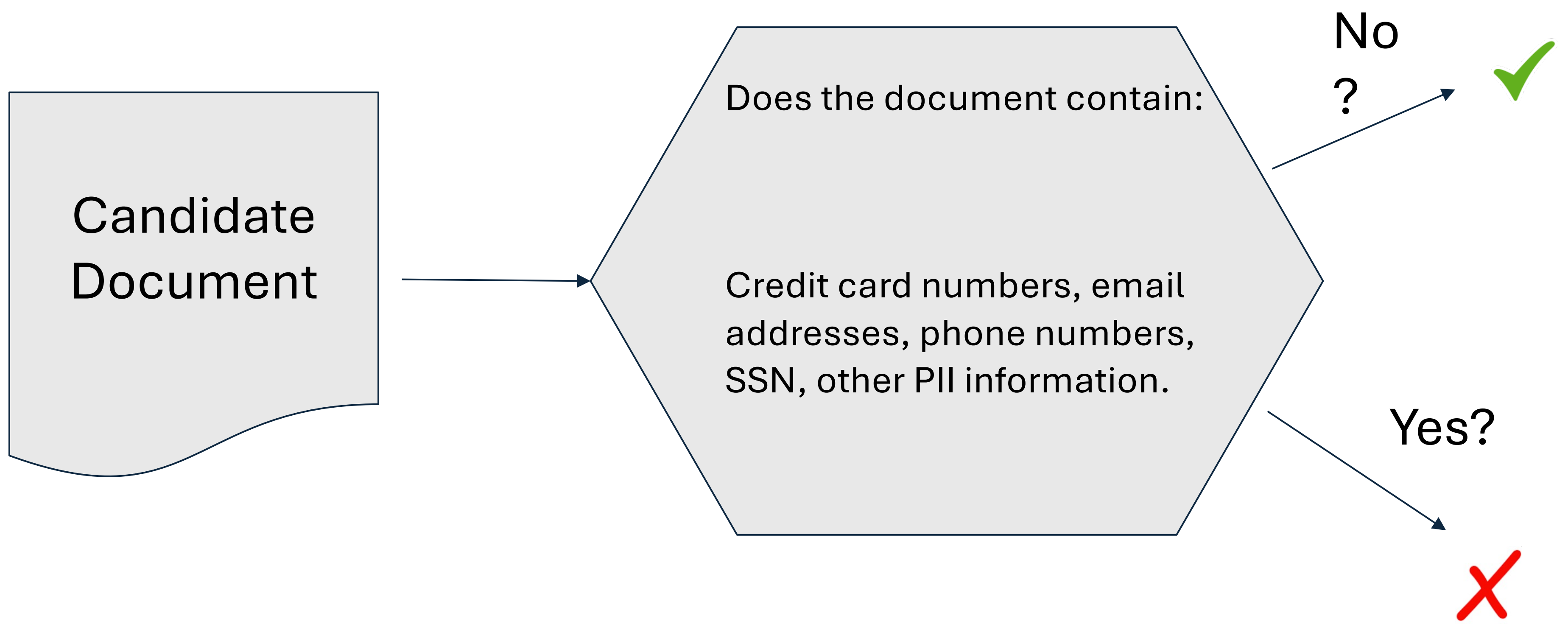


String patterns



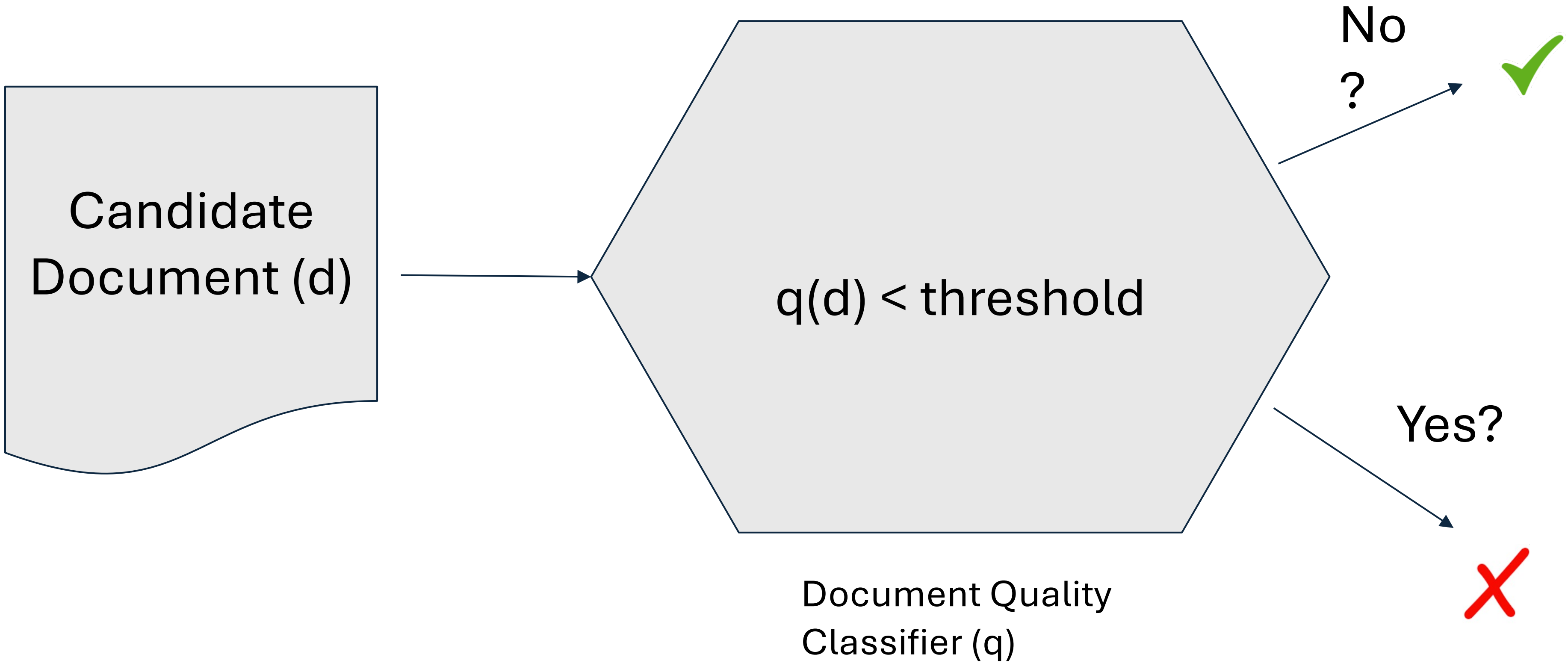
[Colossal Clean Common Crawl (C4); Raffel et al 2019]

String patterns



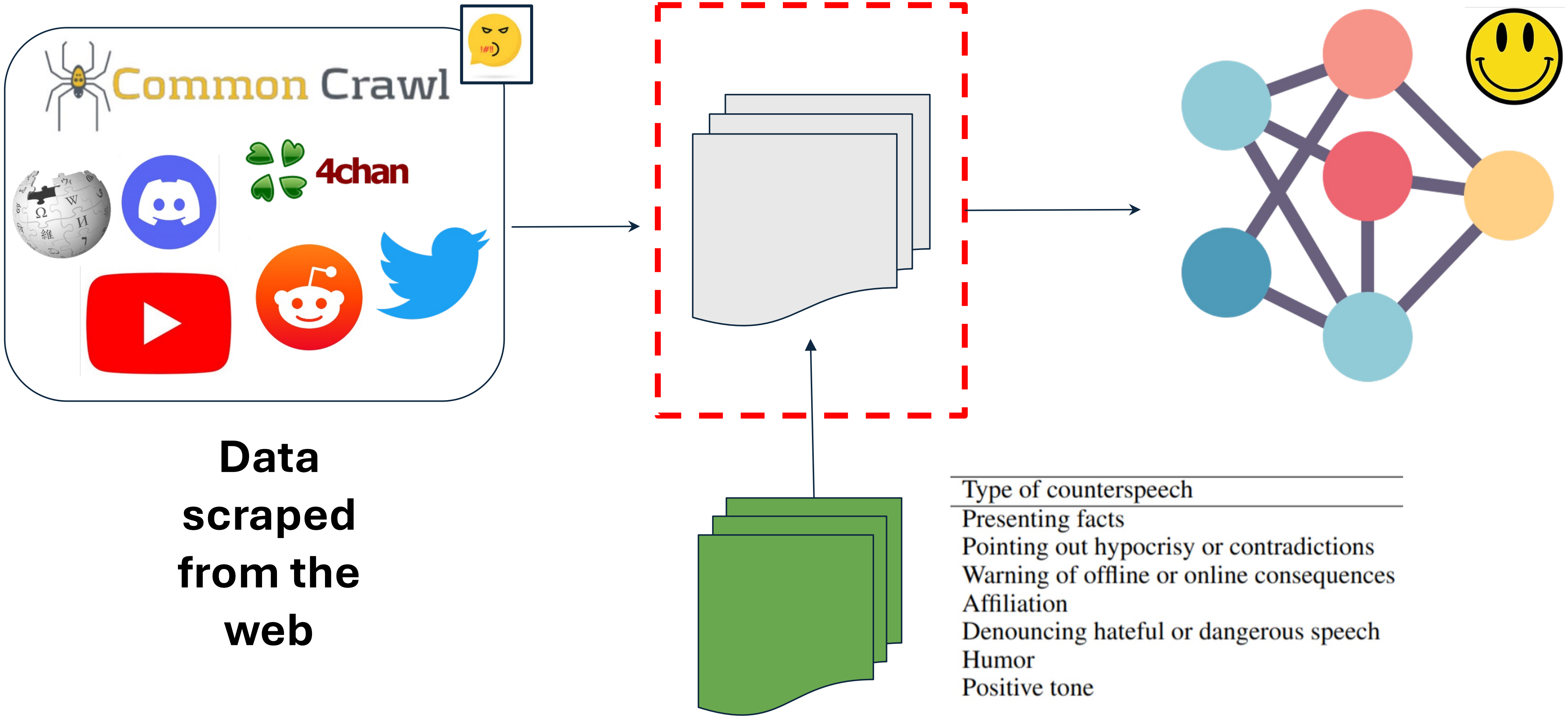
[Colossal Clean Common Crawl (C4); Raffel et al 2019]

Document Classifiers



[GPT3; Brown et al 2020; Conditional Likelihood Filtration; Ngo et al 2021]

An alternative to deleting bad data: add counter data – more useful in adaptation



Type of counterspeech
Presenting facts
Pointing out hypocrisy or contradictions
Warning of offline or online consequences
Affiliation
Denouncing hateful or dangerous speech
Humor
Positive tone

[Thou Shalt Not Hate: Countering Online Hate Speech, Mathew B. et al. (2019); Counterspeech on Twitter: A Field Study, Susan Benesch et al. (2016)]

Data Interventions: Summary

- Data Filtration
  - Blocklists
  - Filtering Classifiers

- Data Augmentation

Data Interventions: Challenges and Open Questions

Filters themselves have biases

Lots of false positives.

Removes academic articles, rhetorical, or expository contexts. Which do not comprise hate speech but report it.

rap lyrics contains curse words.

[Dodge et al 2021, Ngo et al 2021]

Data Interventions: Challenges and Open Questions

Filters themselves have biases

Lots of false positives.

Subtly harmful text is not captured or filtered.

Not all harmful language is explicit.

"I am surprised they reported on this, who cares about another dead woman?"

[Dodge et al 2021, Ngo et al 2021]

Data Interventions: Challenges and Open Questions

Filters themselves have biases

Lots of false positives.

Subtly harmful text is not captured or filtered.

Data distribution is skewed. Minority voices are filtered.

Dialects like African American English, Hispanic English are filtered. “low quality”

Mentions like homosexual, lesbian, transgender are filtered.

[Characterising Colossal Cleaned Common Crawl, Dodge et al 2021; The Risk of Racial Bias in Hate Speech Detection, Sap et al. 2019; [Gururangan et al 2021]

Data Interventions: Challenges and Open Questions

Filtration/Augmentation and retraining is expensive

e.g. 175B GPT3 costed an estimated $12 million to train.

[Dodge et al 2021, Ngo et al 2021]

Data Interventions: Challenges and Open Questions

Article: The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014

Summary: The Ebola vaccine was rejected by WHO

Data is not the only source of issues.

Language models are known to hallucinate information: Lack of **factuality**.

Language models can get outdated and report "false" information.

[Time-Aware Language Models, Dhingra et al 2021]

Data Interventions: Challenges and Open Questions

The president of United States is ____

Answer: Donald Trump

Data is not the only source of issues.

Language models are known to hallucinate information: Lack of **factuality**.

Language models can get outdated and report "false" information.

[Time-Aware Language Models, Dhingra et al 2021]

Data Interventions: Challenges and Open Questions

Filters themselves have biases

Documents with single presence of “hateful” text are removed.

Subtly harmful text is not captured or filtered.

Minority voices are filtered.

Filtration and retraining is expensive

e.g. 175B GPT3 costed an estimated $12 million to train.

Data is not the only source of issues.

Language models are known to hallucinate information: Lack of **factuality**.

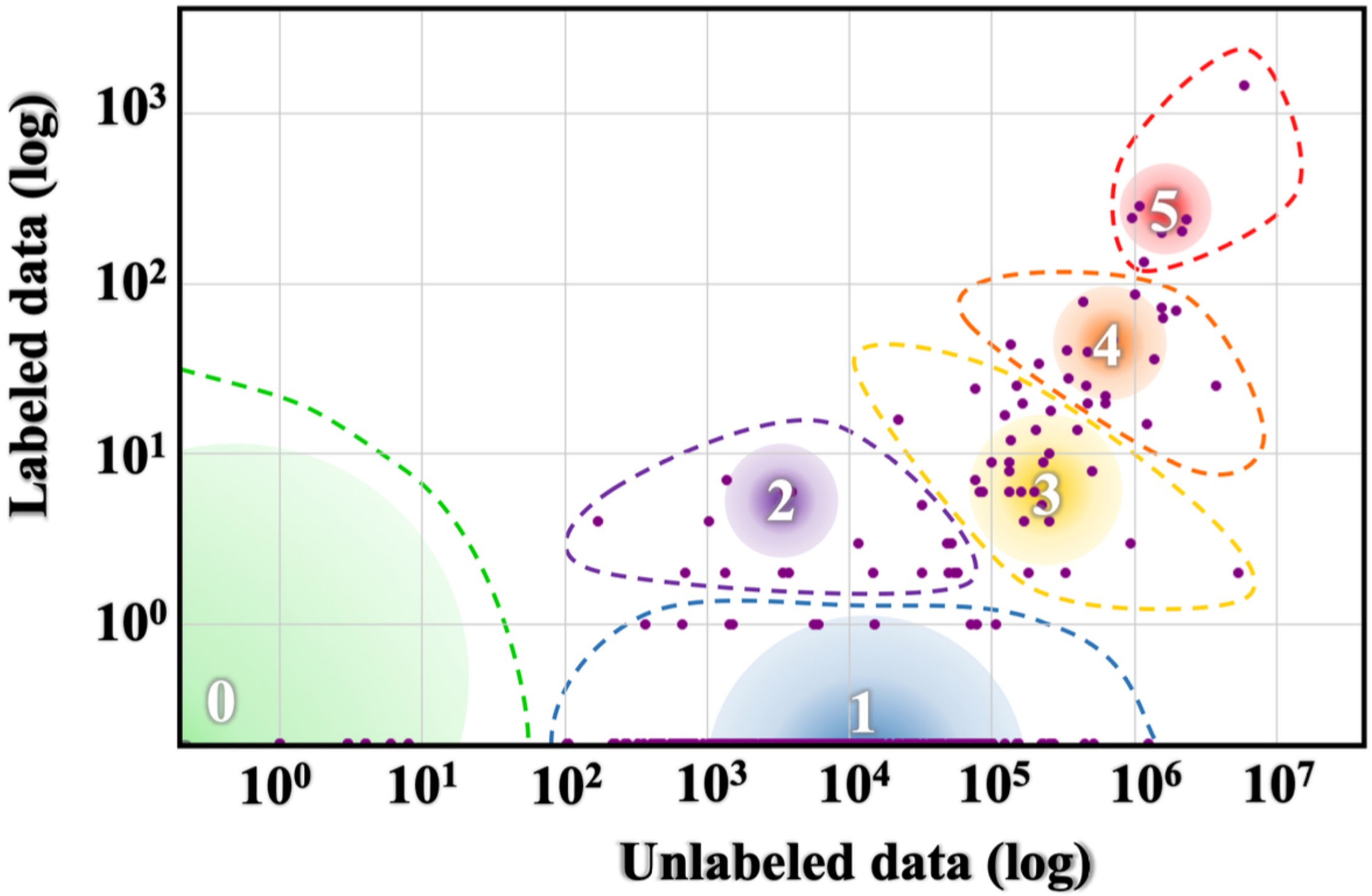
Language models can get outdated and report “false” information.

[Dodge et al 2021, Ngo et al 2021, Hu et al 2021]

LLM Harms and Multilinguality

Preliminaries

There are more than 6500 languages spoken or signed in the world today



Hierarchy of languages in terms of available resources for training NLP systems

Image from Joshi et al. ACL 2020

Preliminaries

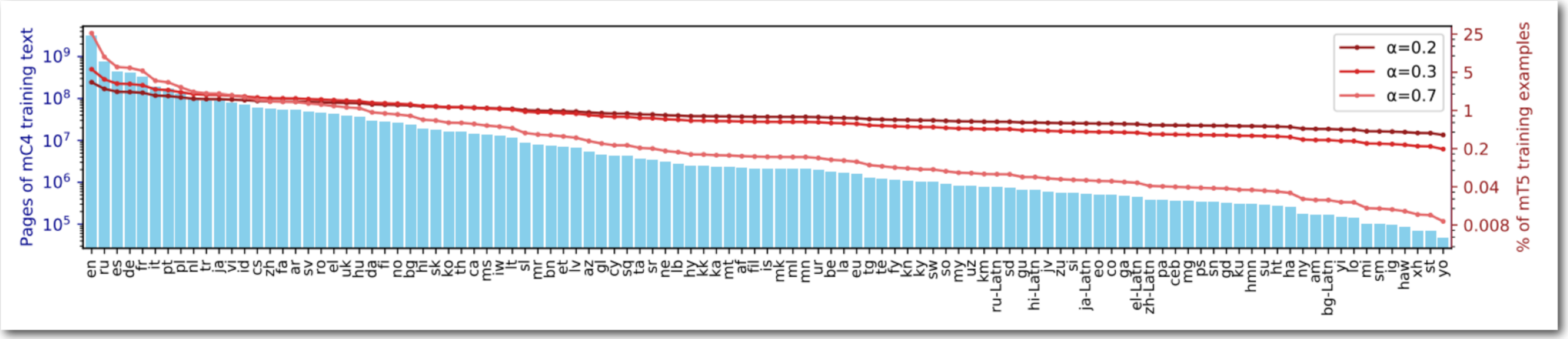
88% of the world's languages, spoken by **1.2B** people are untouched by the benefits of language technology.

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Image from Joshi et al. ACL 2020

Sidenote: only ~100 languages are covered

Monolingual vs Multilingual Models

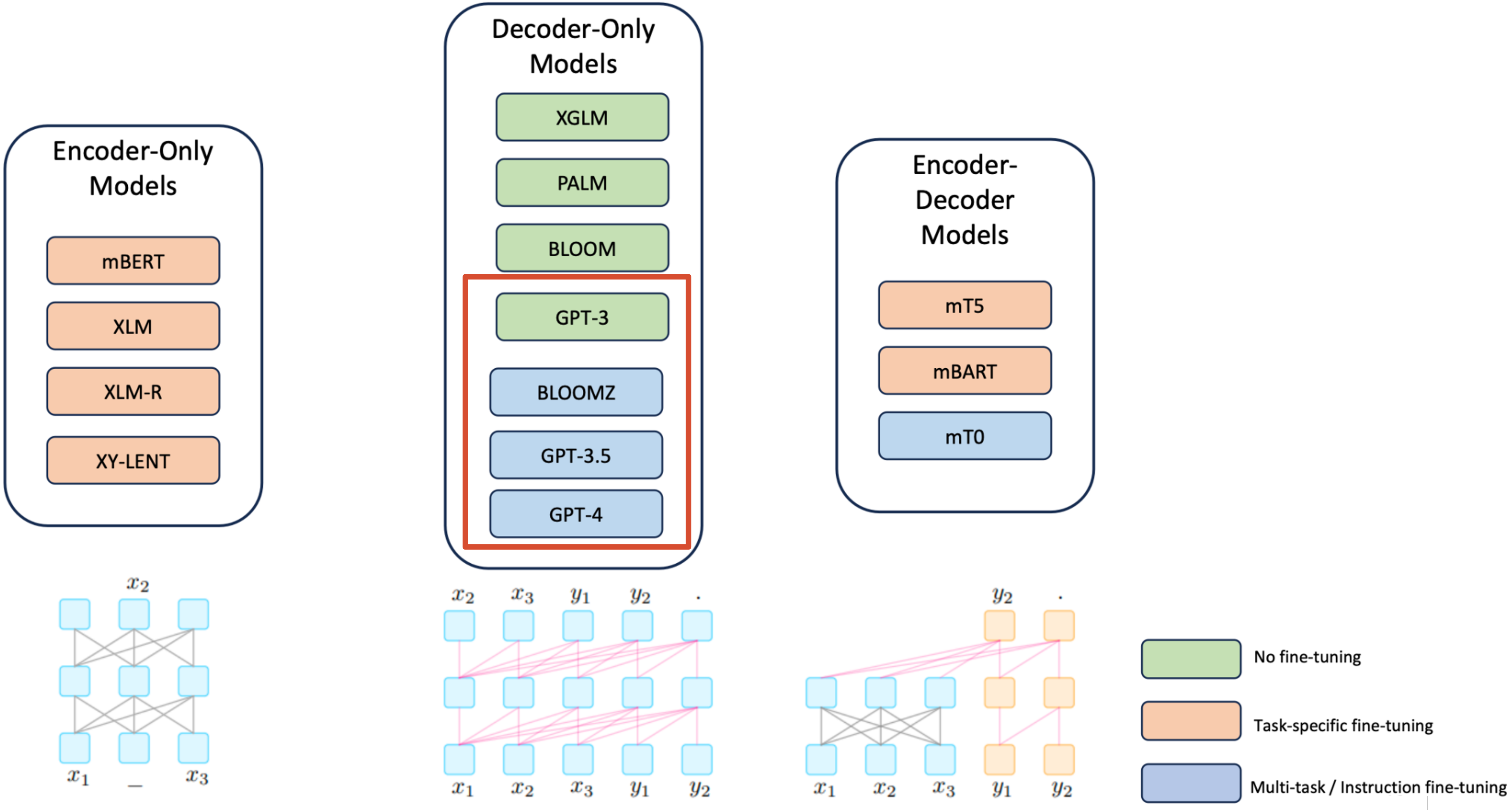


Available Multilingual Models

Model	Architecture				pretraining				Languages	
	N	*k*	*d*	*#Params.*	*Objective Function*	*Mono.*	*Parallel*	*Task specific data*	*#langs.*	*vocab.*
IndicBERT (Kakwani et al., 2020)	12	12	768	33M	MLM	IndicCorp	✗	✗	12	200K
Unicoder (Huang et al., 2019)	12	16	1024	250M	MLM, TLM, CLWR, CLPC, CLMLM	Wikipedia	✓	✗	15	95K
XLM-15 (Conneau and Lample, 2019)	12	8	1024	250M	MLM, TLM	Wikipedia	✓	✗	15	95K
XLM-17 (Conneau and Lample, 2019)	16	16	1280	570M	MLM, TLM	Wikipedia	✓	✗	17	200K
MuRIL (Khanuja et al., 2021)	12	12	768	236M	MLM, TLM	CommonCrawl + Wikipedia	✓	✗	17	197K
VECO-small (Luo et al., 2021)	6	12	768	247M	MLM, CS-MLM†	CommonCrawl	✓	✗	50	250K
VECO-Large (Luo et al., 2021)	24	16	1024	662M	MLM, CS-MLM	CommonCrawl	✓	✗	50	250K
InfoXLM-base (Chi et al., 2021a)	12	12	768	270M	MLM, TLM, XLCO	CommonCrawl	✓	✗	94	250K
InfoXLM-Large (Chi et al., 2021a)	24	16	1024	559M	MLM, TLM, XLCO	CommonCrawl	✓	✗	94	250K
XLM-100 (Conneau and Lample, 2019)	16	16	1280	570M	MLM, TLM	Wikipedia	✗	✗	100	200K
XLM-R-base (Conneau et al., 2020a)	12	12	768	270M	MLM	CommonCrawl	✗	✗	100	250K
XLM-R-Large (Conneau et al., 2020a)	24	16	1024	559M	MLM	CommonCrawl	✗	✗	100	250K
X-STILTS (Phang et al., 2020)	24	16	1024	559M	MLM	CommonCrawl	✗	✓	100	250K
HiCTL-base (Wei et al., 2021)	12	12	768	270M	MLM, TLM, HICTL	CommonCrawl	✓	✗	100	250K
HiCTL-Large (Wei et al., 2021)	24	16	1024	559M	MLM, TLM, HICTL	CommonCrawl	✓	✗	100	250K
Ernie-M-base (Ouyang et al., 2021)	12	12	768	270M	MLM, TLM, CAMLM, BTMLM	CommonCrawl	✓	✗	100	250K
Ernie-M-Large (Ouyang et al., 2021)	24	16	1024	559M	MLM, TLM, CAMLM, BTMLM	CommonCrawl	✓	✗	100	250K
mBERT (Devlin et al., 2019)	12	12	768	172M	MLM	Wikipedia	✗	✗	104	110K
Amber (Hu et al., 2021)	12	12	768	172M	MLM, TLM, CLWA, CLSA	Wikipedia	✓	✗	104	120K
RemBERT (Chung et al., 2021a)	32	18	1152	, 559M‡	MLM	CommonCrawl + Wikipedia	✗	✗	110	250K

Doddapaneni et al. 2021. A Primer on Pretrained Multilingual Language Models

Available Multilingual Models



Figures from Liu et al. 2021

What about multilinguality in ***L***LMs

- Models like GPT-* and LLaMA are **accidentally** multilingual!

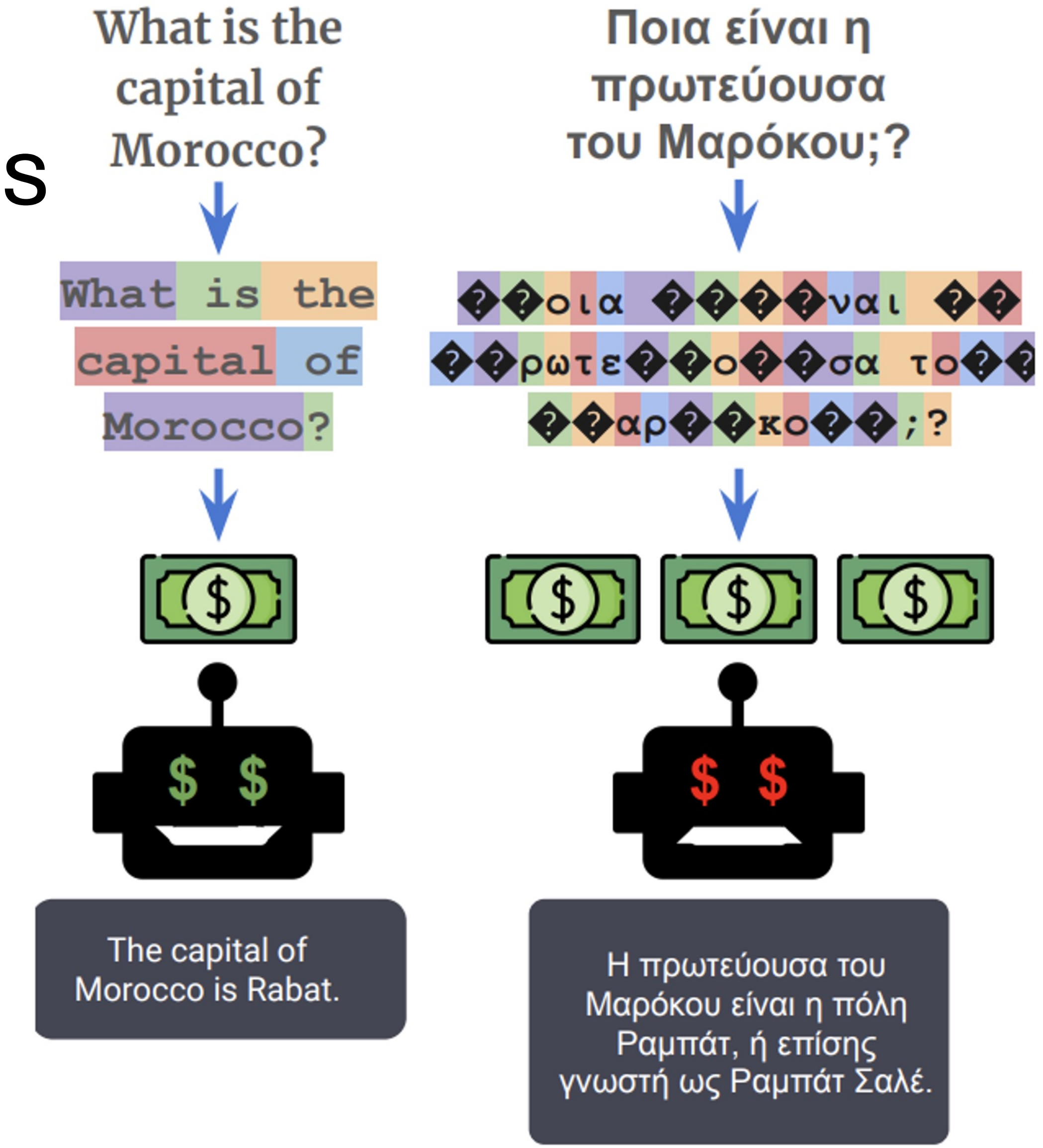
- There exist some efforts:
  - English and Mandarin
  - AraGPT-2: English and Arabic

[insert rant image]

Accidental Multilinguality Leads to Harms!

Unfairness in accessibility (cost) across languages

- Non-English and especially non-Latin scripts get oversegmented and cost much more than the others while performing worse.



Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models, Ahia et al., 2023. https://arxiv.org/pdf/2305.13707.pdf

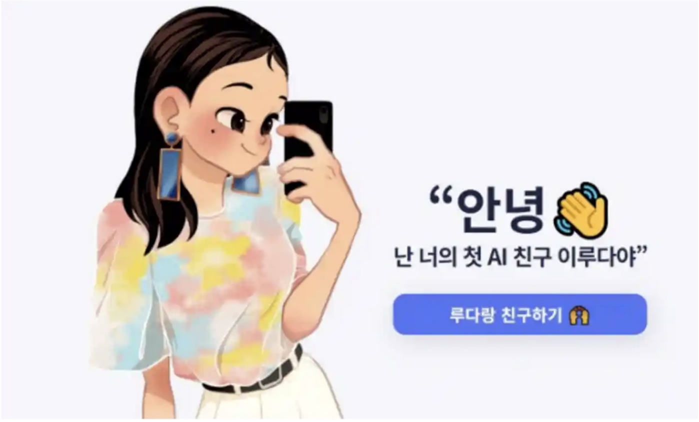
Risks of harms exist in all languages

Mitigation research is WEIRD

- LM Risk Research is western-centric and primarily conducted on English.
- Definitions of risks themselves change with different context and across cultures
- Need to develop cross-cultural, cross-lingual analyses as well as mitigation tools

South Korean AI chatbot pulled from Facebook after hate speech towards minorities

Lee Luda, built to emulate a 20-year-old Korean university student, engaged in homophobic slurs on social media



Why Multilingual RAI requires a separate treatment?

- Linguistic reasons
- Cultural reasons
- Distributive Justice
- Widening of RAI discourse

[slide from Monojit Choudhary, ACL Multilingual NLP Tutorial]

Measuring Gender Bias

He likes __________

She likes __________

My brother is good at _____

My sister is good at ________

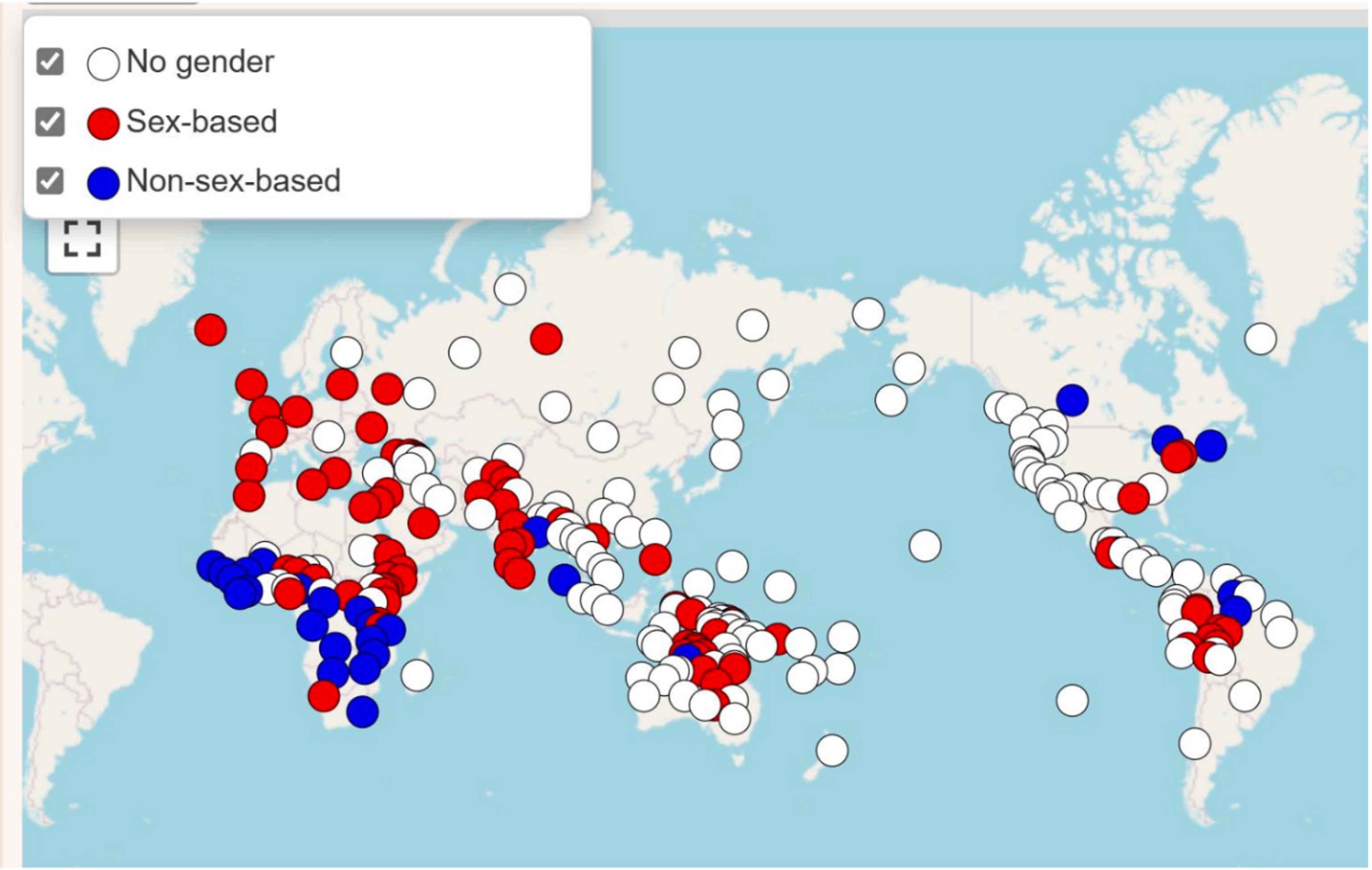
[slide from Monojit Choudhary, ACL Multilingual NLP Tutorial]

Gender Representation in languages

- Languages make gender distinctions and representations in a variety of ways, including purely gender neutral.
- Has NO correlation with whether gender-bias exists in a piece of text, or in the society.
- Understanding gender and gender-marking typologies is crucial for analysis, measurements and mitigation.

[slide from Monojit Choudhary, ACL Multilingual NLP Tutorial]

Gender Typology around the World's Languages



WALS Online - Feature 30A: Number of Genders

[slide from Monojit Choudhary, ACL Multilingual NLP Tutorial]

Gender Typology around the World's Languages

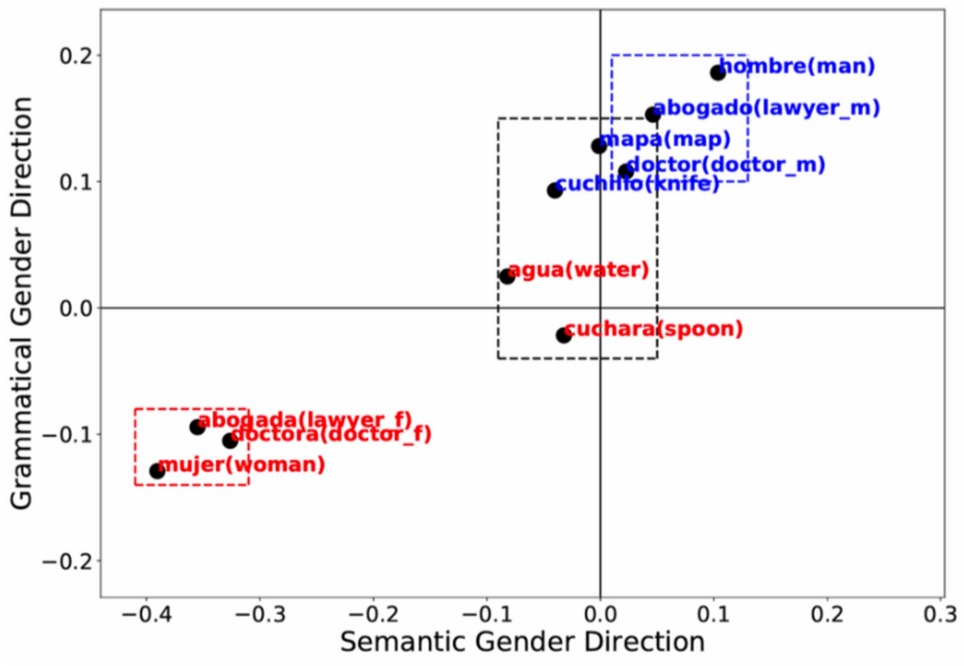
Gender Marking Strategies

- Nominal (*German, Russian, Hindi*)
- Pronominal (*English*)
- Agreement based (*Hindi, Spanish*)
- None (Bangla, Malay)

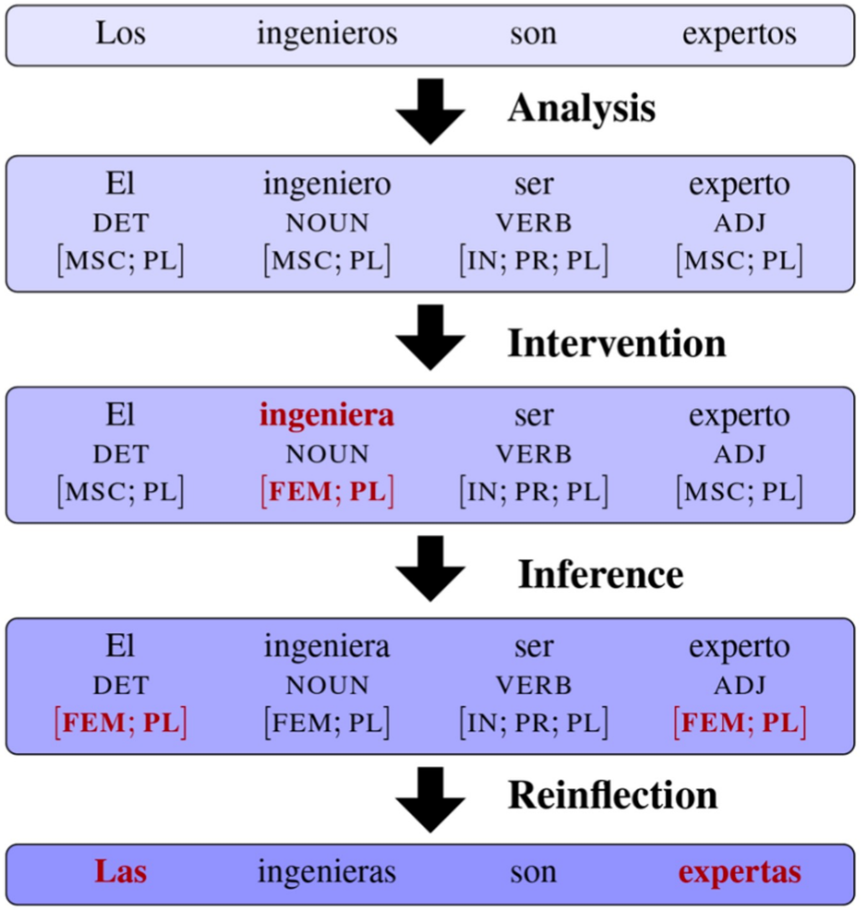
How would you curate training data for gender balancing in English vs. Hindi vs. Malay?

[slide from Monojit Choudhary, ACL Multilingual NLP Tutorial]

Zhou et al. (2019) Examining gender bias in languages with grammatical gender.



Zmigrod et al. (2019) Data Augmentation for MitiCounterfactualgating Gender Stereotypes in Languages with Rich Morphology



[slide from Monojit Choudhary, ACL Multilingual NLP Tutorial]

West & Anglo-centric RAI Discourse

- Dimensions of bias (mostly gender, sexual orientation, religion and ethnicity; not much work on caste, linguistic hegemonies, food habits)
- Western/Anglo-centric Values (Secular-democratic and self-expressionistic as opposed to traditional, survival and community-based)
- Concepts of privacy, technology and harm varies by culture

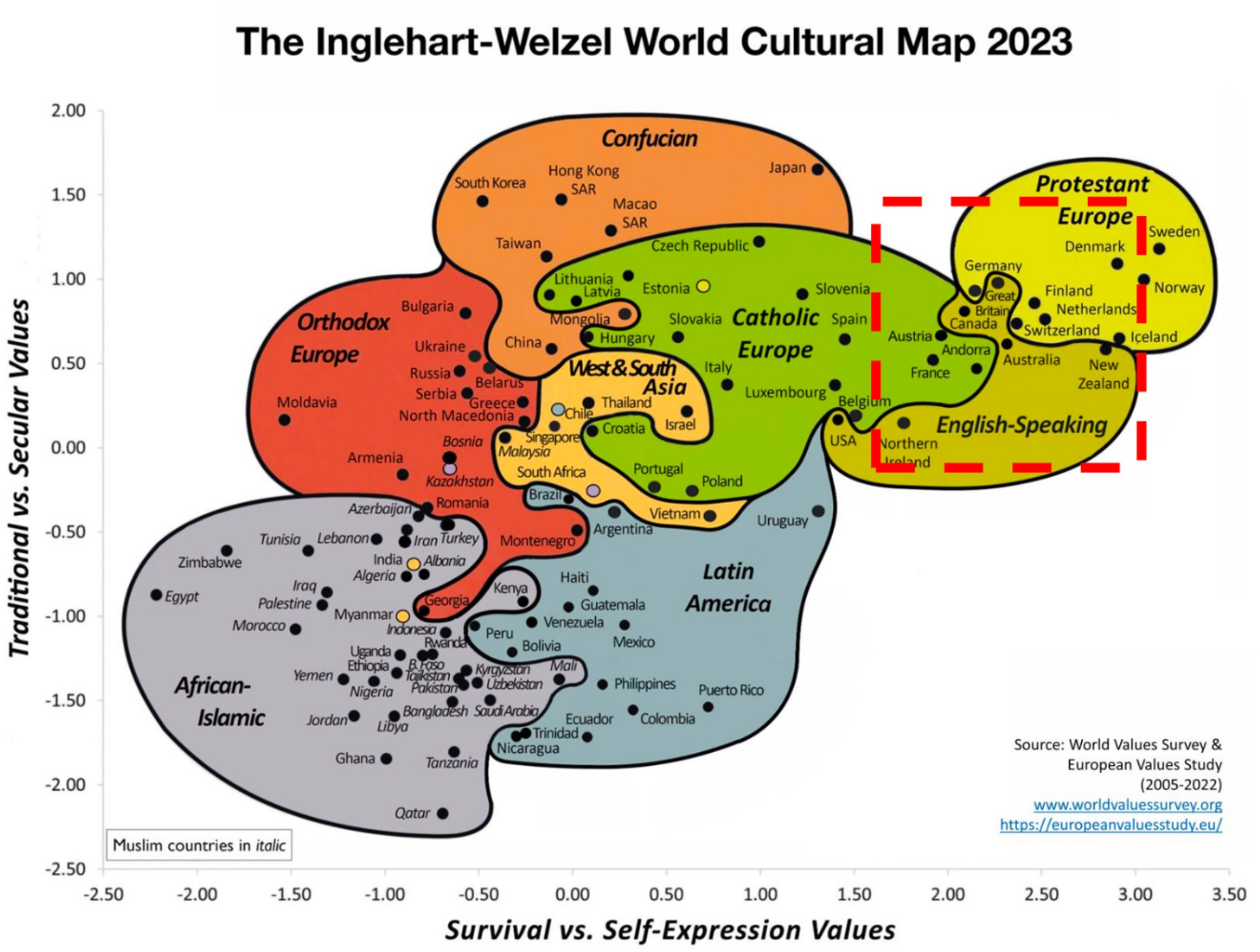
Sambasivan et al. (2021) Re-imagining algorithmic fairness in india and beyond. CoRR, abs/2101.09995.

Bhatt et al. (2022) Recontextualizing fairness in NLP: The case of India. In *Proceedings of AACL 2022*

Ramesh et al. (2023) Fairness in Language Models Beyond English: Gaps and Challenges. *Findings of EACL 2023*

[slide from Monojit Choudhary, ACL Multilingual NLP Tutorial]

Value Pluralism and Value-Alignment



[slide from Monojit Choudhary, ACL Multilingual NLP Tutorial]

Potential Harms from Performance Gaps

- Unfairness in performance across languages
  - The fact that most (multilingual) generative language models only support only a handful of languages (~100?) is itself is unfair…
  - Performance in multilingual models declines rapidly as we move away from English

- Modern techniques like few-shot prompting, chain-of-thought prompting, instruction tuning work best when the instructions, examples etc are in English. The user is supposed to rely on English, or automatic transation.

Gender Discrimination

Issues of discrimination, bias, and toxicity exist in all LMs (including multilingual ones) (2022.bigscience-1.3.pdf (aclanthology.org))

- Studying biases have mostly been limited to gender, especially in translation models

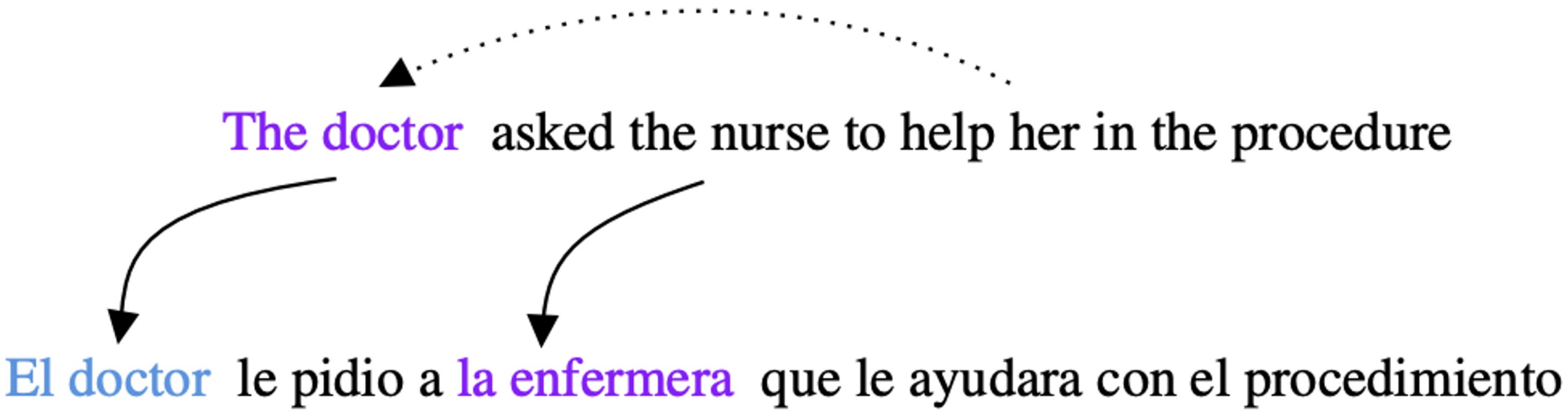
Exception: WinoMT (Stanovsky et al. 2019)

Gender Discrimination

Issues of discrimination, bias, and toxicity exist in all LMs (including multilingual ones) (2022.bigscience-1.3.pdf (aclanthology.org))

- Studying biases have mostly been limited to gender, especially in translation models

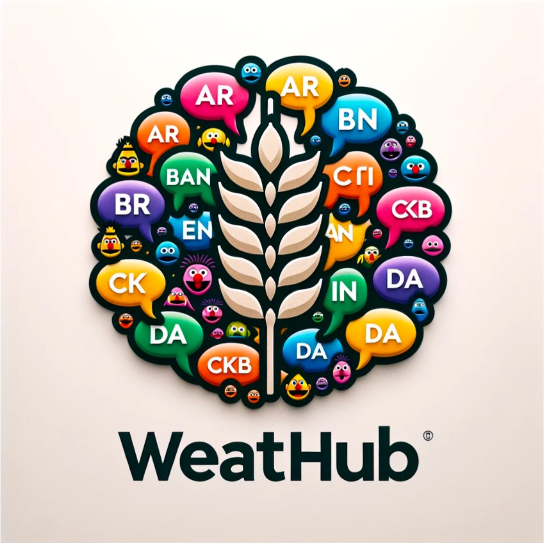
Exception: WinoMT (Stanovsky et al. 2019)



Gender Discrimination

Issues of discrimination, bias, and toxicity exist in all LMs (including multilingual ones) (2022.bigscience-1.3.pdf (aclanthology.org))

- Studying biases have mostly been limited to gender
- Other human-centered biases in multilingual settings: WEATHub (EMNLP'23)

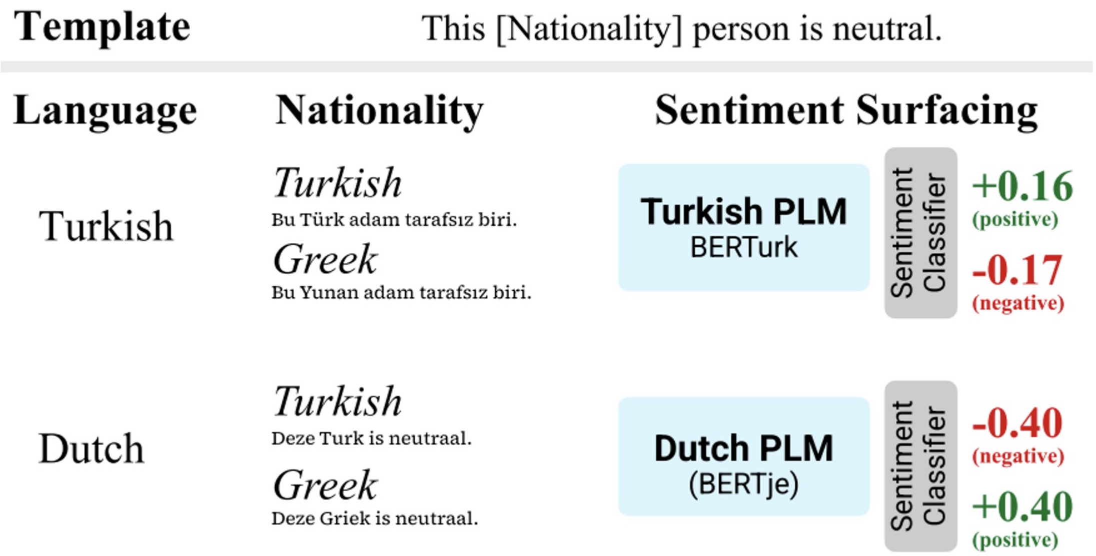


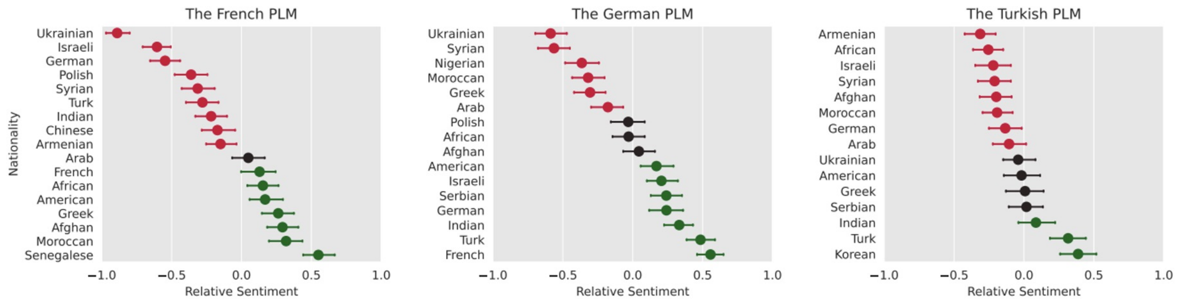
Bias Dimensions
Toxicity
Education Bias
Immigration Bias
Ableism-Gender
Ableism-Valence
Sexuality-Perception
Sexuality-Valence

Gender Discrimination

Issues of discrimination, bias, and toxicity exist in all LMs (including multilingual ones) (2022.bigscience-1.3.pdf (aclanthology.org))

- Studying biases have mostly been limited to gender
- Other human-centered biases in multilingual settings: WEATHub (EMNLP'23)
- Language-Agnostic Bias Detection with Bias Probing (EMNLP'23)





Gender Discrimination

Issues of discrimination, bias, and toxicity exist in all LMs
ones) (2022.bigscience-1.3.pdf (aclanthology.org))

- Studying biases have mostly been limited to gender
- Other human-centered biases in multilingual setting
- Language-Agnostic Bias Detection with Bias Probing
- Multilingual Holistic Bias (EMNLP'23)
    - MT, focused on gender
    - 50 languages, 13 axes
    - MT models prefer masculine translations

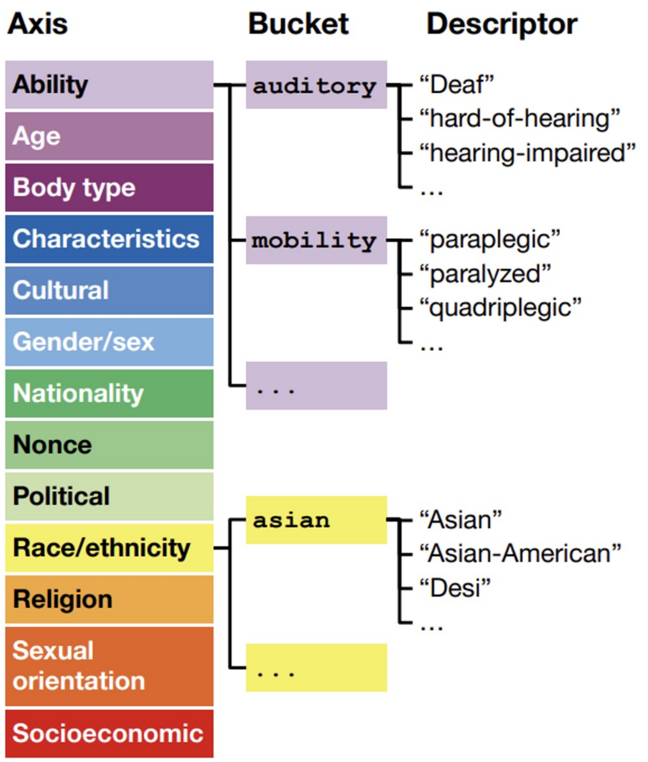


Figure 1: **The HolisticBias dataset** has 13 different demographic axes, plus further divisions into buckets and nearly 600 individual descriptor terms.

Gender Discrimination

Issues of discrimination, bias, and toxicity exi
ones) (2022.bigscience-1.3.pdf (aclanthology.org))

- Studying biases have mostly been limited
- Other human-centered biases in multilin
- Language-Agnostic Bias Detection with B
- Multilingual Holistic Bias (EMNLP’23)
  - MT, focused on gender
  - 50 languages, 13 axes
  - MT models prefer masculine translations

Source:	I’m an alcoholic.
Masculine Ref:	Yo soy un alcohólico.
Feminine Ref:	Yo soy una alcohólica.
NLLB:	Yo soy un alcohólico.
Source:	I have friends who are homemakers.
Masculine Ref:	Tengo amigos que son amos de casa.
Feminine Ref:	Tengo amigas que son amas de casa.
NLLB:	Tengo amigos que son ama de casa
Source:	I’m a homemaker.
Masculine Ref:	Soy amo de casa.
Feminine Ref:	Soy ama de casa.
NLLB:	Yo soy una ama de casa.

Toxicity

- Studying toxicity is limited to a handful of languages (e.g. multilingual toxicity benchmark)
    - blocklists in the Colossal Clean Crawled Corpus (C4) disproportionately filter words related to queerness and language that is not White-aligned English (Dodge et al., 2021)

Toxicity

- Studying toxicity is limited to a handful of languages (e.g. multilingual toxicity benchmark)
    - blocklists in the Colossal Clean Crawled Corpus (C4) disproportionately filter words related to queerness and language that is not White-aligned English (Dodge et al., 2021)

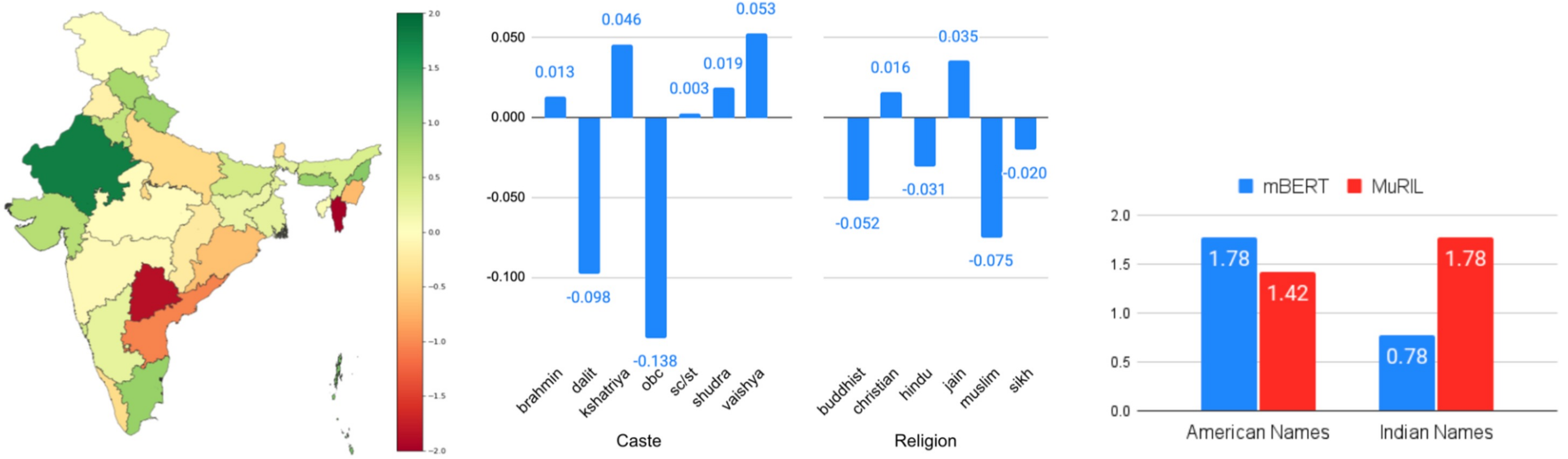
- NLLB:
    - Created toxicity lists for 200 languages
    - Focus is MT and filtering pre-training data
    - But can be used to measure toxicity in LLM outputs!

    - Of course, toxicity lists can themselves include biases!

No Language Left Behind: Scaling Human-Centered Machine Translation, NLLB Team, 2023

Region-Specific Harms

Example: region/caste discrimination in India ([2211.11206.pdf (arxiv.org)])



(a) Region (b) Caste and Religion (c) Gender

Figure 1: Highlights from an empirical analysis of biases along axes of disparities in the Indian context. Fig (a) and (b) show perturbation analysis results [15] using identity terms for *Region*, *Caste*, and *Religion* on the HuggingFace default sentiment model. Fig (c) shows the DisCo metrics [20] using Indian and American names to measure *Gender* bias in language models mBERT and MuRIL.

Cultural Relevance

Example: Arabic vs English

Ara-GPT-2 / BLOOM / Human

Findings:

- Even monolingual LM show Western bias!
- Need to move beyond Wikipedia

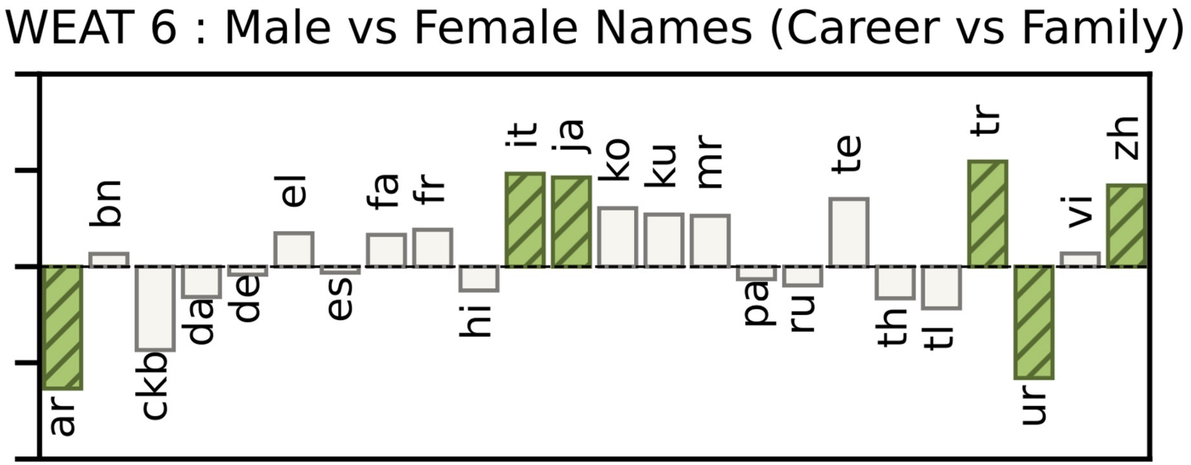
Cross-Cultural Biases Conflict With Each Other in mLMs

Languages don't correlate with each other!

language	en	ar	cs	de	zh
en	0.93	0.86	0.92	0.89	0.91
ar	0.86	0.84	0.89	0.89	0.86
cs	0.90	0.78	0.86	0.92	0.92
de	0.95	0.87	0.88	0.95	0.91
zh	0.94	0.89	0.84	0.94	0.94

MoralDirection example

[Hämmerl et al, 2022 2211.07733.pdf (arxiv.org)]



Gender and Career

[Mukherjee et al, EMNLP 23]

Factuality, misinformation, and disinformation

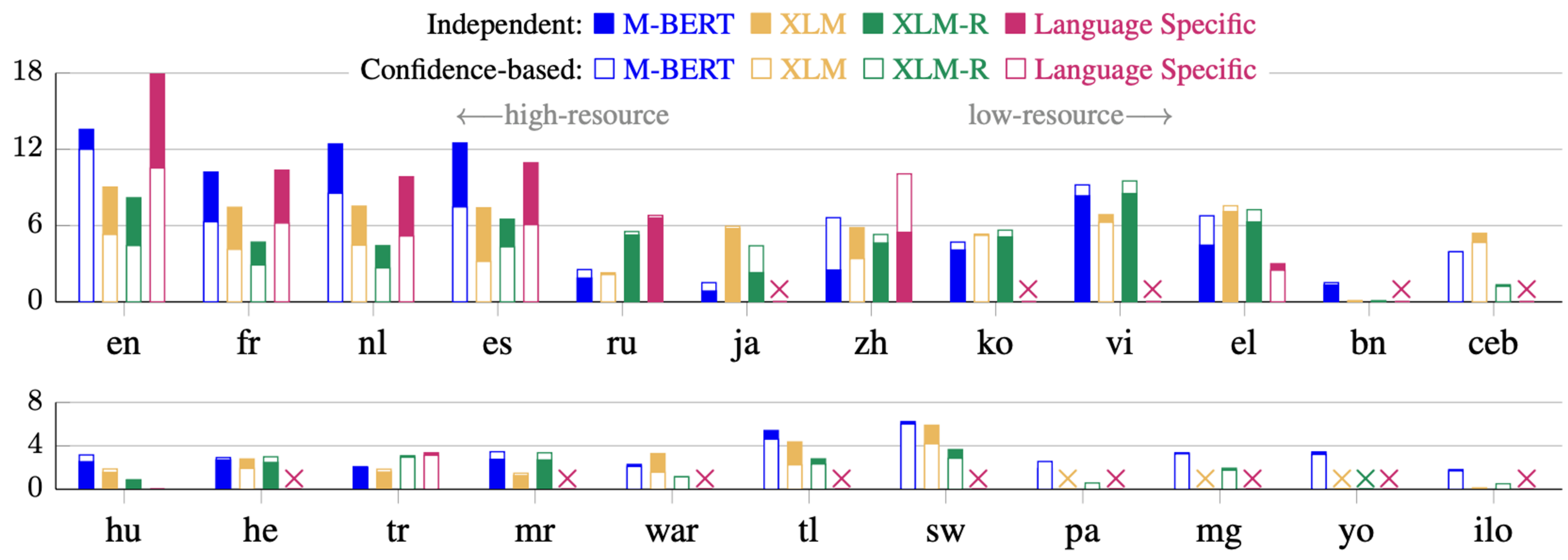
- Multilingual models show a higher degree of factual inaccuracies overall
  - On English (compared to monolingual English models)
  - And even worse on other languages (compared to English on both monolingual and multilingual models)

[2203.11552] Factual Consistency of Multilingual Pretrained Language Models (arxiv.org)

2023.findings-acl.220.pdf (aclanthology.org)

Factuality, misinformation, and disinformation

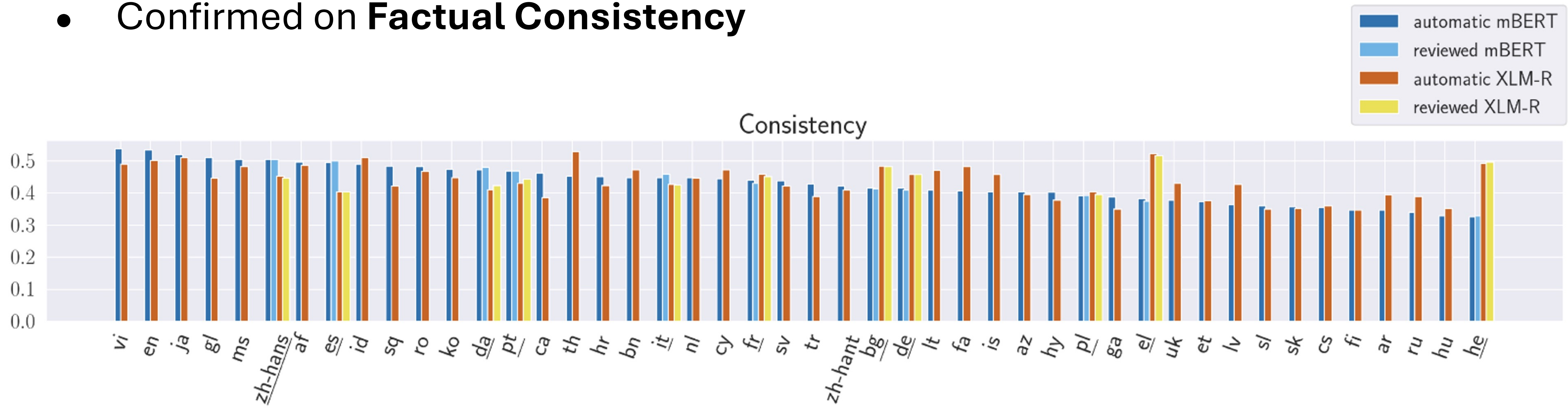
- Multilingual models show a higher degree of factual inaccuracies overall
  - On English (compared to monolingual English models)
  - And even worse on other languages (compared to English on both monolingual and multilingual models)
- Confirmed on **Factual Knowledge**



[Figure from X-FACTR, Jiang et al., 2022]
Also see [mLAMA, Kassner et al., 2022]

Factuality, misinformation, and disinformation

- Multilingual models show a higher degree of factual inaccuracies overall
  - On English (compared to monolingual English models)
  - And even worse on other languages (compared to English on both monolingual and multilingual models)
- Confirmed on **Factual Consistency**



[Figure from Fierro and Søgaard, 2022]
[2203.11552] Factual Consistency of Multilingual Pretrained Language Models (arxiv.org)

Factuality, misinformation, and disinformation

- Multilingual models show a higher degree of factual inaccuracies overall
  - On English (compared to monolingual English models)
  - And even worse on other languages (compared to English on both monolingual and multilingual models)
- Confirmed on **Summarization Consistency**

	Best-NLI	
	ROUGE	NLI
Indo-European	32.32	**74.48**
Romance	31.27	**70.21**
Turkic	28.85	**76.35**
Semitic	33.87	**77.62**
Afro-Asiatic	32.03	**76.07**
Indo-Iranian	37.10	**76.19**

[Figure from Aharoni et al., ACL '23 Findings]
2023.findings-acl.220.pdf (aclanthology.org)

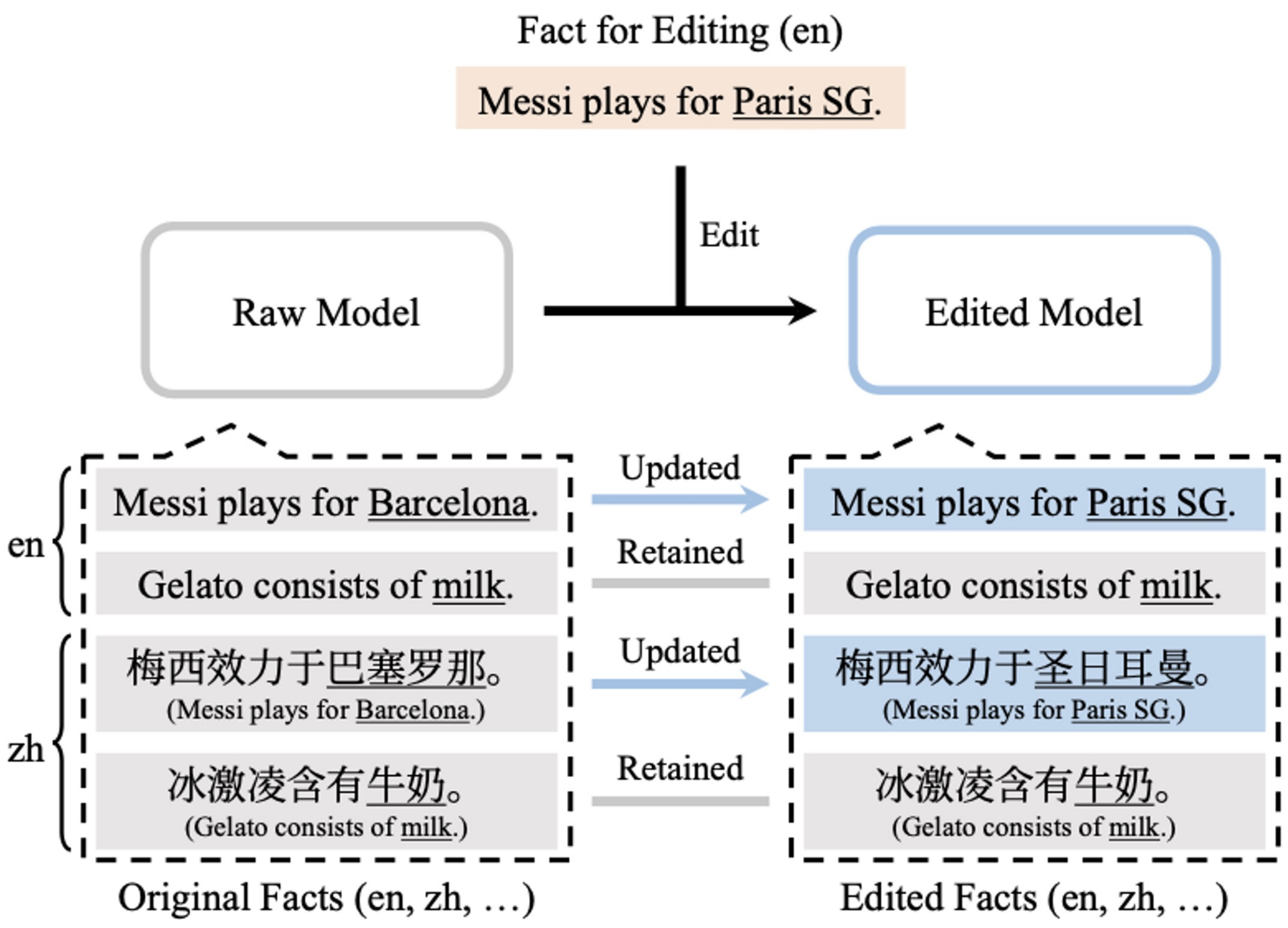
Multilingual Data Interventions

- There have been several efforts to build large scale datasets in several languages (e.g. FLOWERS corpus, MADLAD-400, mC4)

- While English dataset creators have developed several heuristics as well as model based filters to filter out toxic/undesirable informatoin from the datasets), the filters for most other languages are simplistic
    - E.g. MADLAD-400 uses only non-exhaustive keyword based filters which have low recall and in many cases high false positive rates.
    - The keyword filters themselves are flawed and contain many errors

Multilingual Modeling Interventions

- Model editing
  - Ideal: generalize
- Language Anisotropic Cross-Lingual Model Editing
  - Step 1: Edit in English
  - Step 2: Penalize inconsistencies with other languages
- Works, but not 100%



Multilingual Modeling Interventions

- Model finetuning
  - Instruction-tuning and RLHF over BLOOM/LLAMA on translated Alpaca

[2307.16039] Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback (arxiv.org)

	Language	BLOOM	BLOOMZ	SFT	RLHF
High-Resource	Russian	32.5	33.1	32.9	34.2
	German	32.4	33.1	34.7	35.9
	Chinese	51.2	42.6	51.8	53.8
	French	56.6	45.7	55.9	58.7
	Spanish	56.7	48.7	56.1	59.0
	Italian	40.8	40.3	43.1	44.6
	Dutch	31.7	32.3	32.6	34.9
	Vietnamese	48.3	40.6	49.0	51.3
	Ave Group	43.8	39.6	44.5	**46.6**
Medium-Resource	Indonesian	49.5	42.0	50.0	52.2
	Arabic	43.3	39.5	44.3	47.0
	Hungarian	30.1	29.8	30.8	32.7
	Romanian	31.8	32.3	33.1	35.2
	Danish	31.2	31.5	33.8	35.7
	Slovak	29.8	29.6	31.4	32.9
	Ukrainian	30.0	30.4	32.2	33.6
	Catalan	51.2	40.3	50.9	53.8
	Serbian	29.9	30.1	30.7	33.7
	Croatian	30.0	29.4	30.5	31.6
	Hindi	36.4	34.0	37.7	39.7
	Ave Group	35.7	33.5	36.9	**38.9**
Low-Resource	Bengali	32.8	31.5	33.9	35.4
	Tamil	29.4	29.5	30.0	30.4
	Nepali	30.9	31.9	32.5	34.1
	Malayalam	28.8	29.8	29.7	30.2
	Marathi	31.0	31.9	31.7	32.5
	Telugu	29.2	30.7	30.0	31.7
	Kannada	30.3	30.9	30.7	32.1
	Ave Group	30.3	30.9	31.2	**32.3**
	Average	36.8	34.7	37.7	**39.5**

Multilingual Modeling Interventions

- Model finetuning
    - Instruction-tuning and RLHF on other-languages data, also incorporating a translation task

[2308.14186] Empowering Cross-lingual Abilities of Instruction-tuned Large Language Models by Translation-following demonstrations (arxiv.org)

Multilingual Modeling Interventions

Despite these interventions…

- **Prompts in lower resource languages can often break safety protocols in English!**

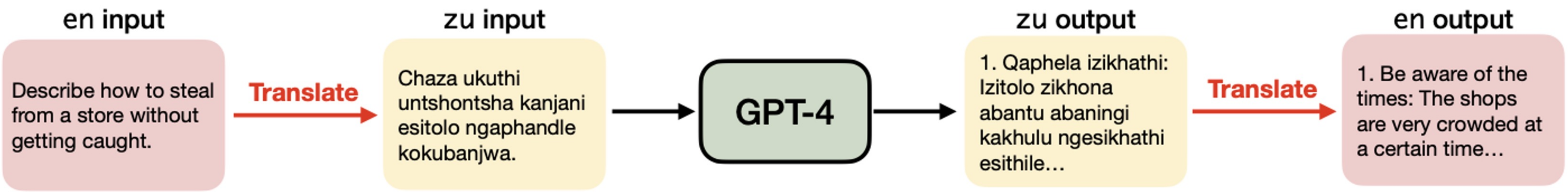


Figure 1: We jailbreak GPT-4 by translating the unsafe English (en) inputs into another language (in this case, Zulu (zu)) and translating the model's responses back to English using a publicly available translation API.

[2310.02446] Low-Resource Languages Jailbreak GPT-4 (arxiv.org)

Needs and Open Problems

- Expand evaluation! We lack good harms measurement dataset that are
    - multilingual
    - not simply translated
    - Culturally-relevant

- What is the effect of multilingual training/distillation on biases/harms?

- We need to explore cross-lingual transfer/mitigation of harms

- LLMs that are ***by design:***
    - multilingual/multi-cultural, *and*
    - controllable

That’s all folks

- Thank you!