

Content Moderation and LLMs

Lorenzo Cima

With special guest: Benedetta Tessa (benedetta.tessa@iit.cnr.it)



UNIVERSITÀ DI PISA



ISTITUTO
DI INFORMATICA
E TELEMATICA

lorenzo.cima@phd.unipi.it; lorenzo.cima@iit.cnr.it



Social Media Analysis



CONTENT MODERATION



Outline

- 1. Introduction to content moderation**
- 2. Evaluation of the effects of an intervention**
- 3. Abandonment prediction**
- 4. Development of interventions – LLMs basics**
- 5. A soft intervention - Personalized counterspeech**
- 6. Problems in LLMs – Hallucinations**
- 7. Prediction of the moderation effects**
- 8. Problems in LLMs – Bias**
- 9. Persuasion capabilities of LLMs for content moderation**





Toxicity

Is disinformation the only problem of social media? **NO!**

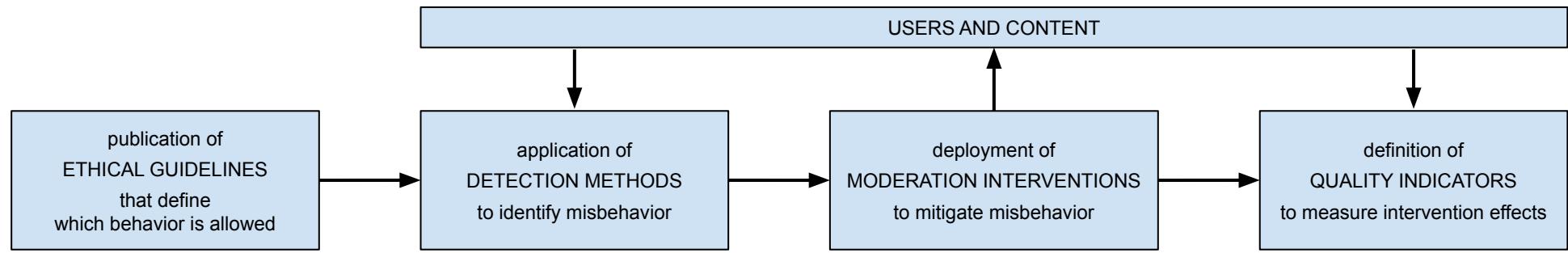
- **Toxicity**
- **Hate speech and harassment**
- **Radicalization and polarization**
- ...



Need for appropriate countermeasures!
(CONTENT MODERATION)



Content Moderation



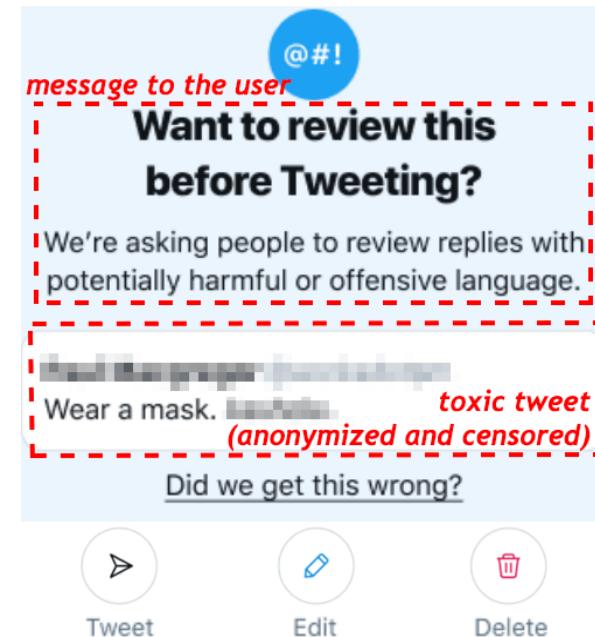


Countermeasure: Content Moderation



Soft interventions

- Warning messages
- Quarantine
- Visibility restrictions



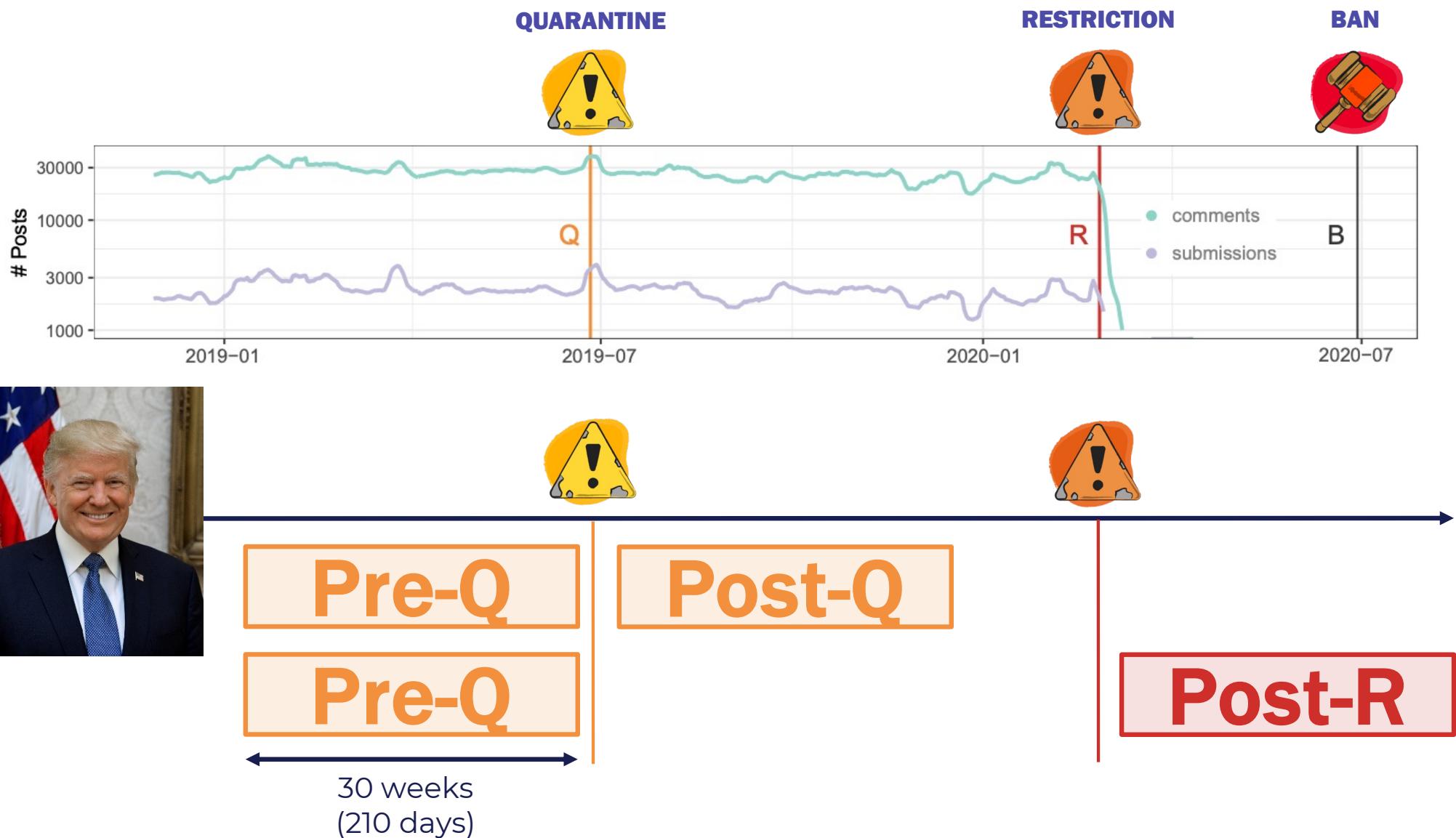
Hard interventions

- Content deletion
- User deplatforming
- Massive bans

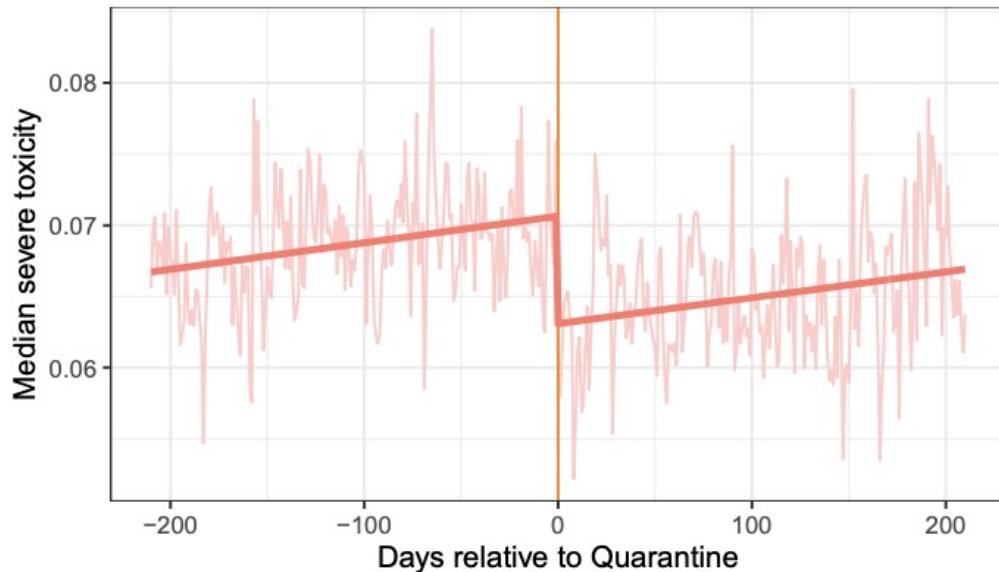




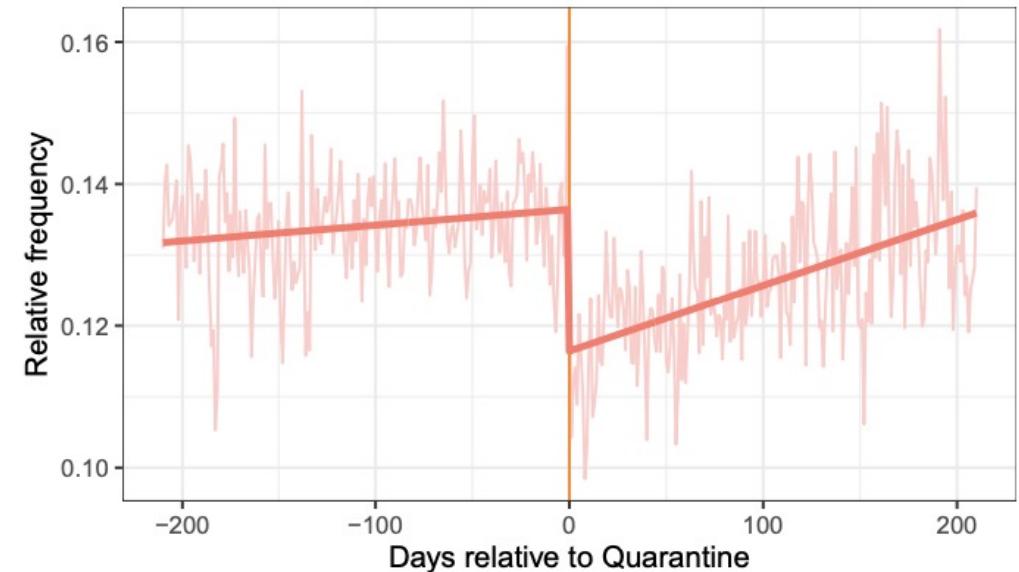
Content Moderation: Does it Work?



Content Moderation: Does it Work?



(a) Median severe toxicity scores.

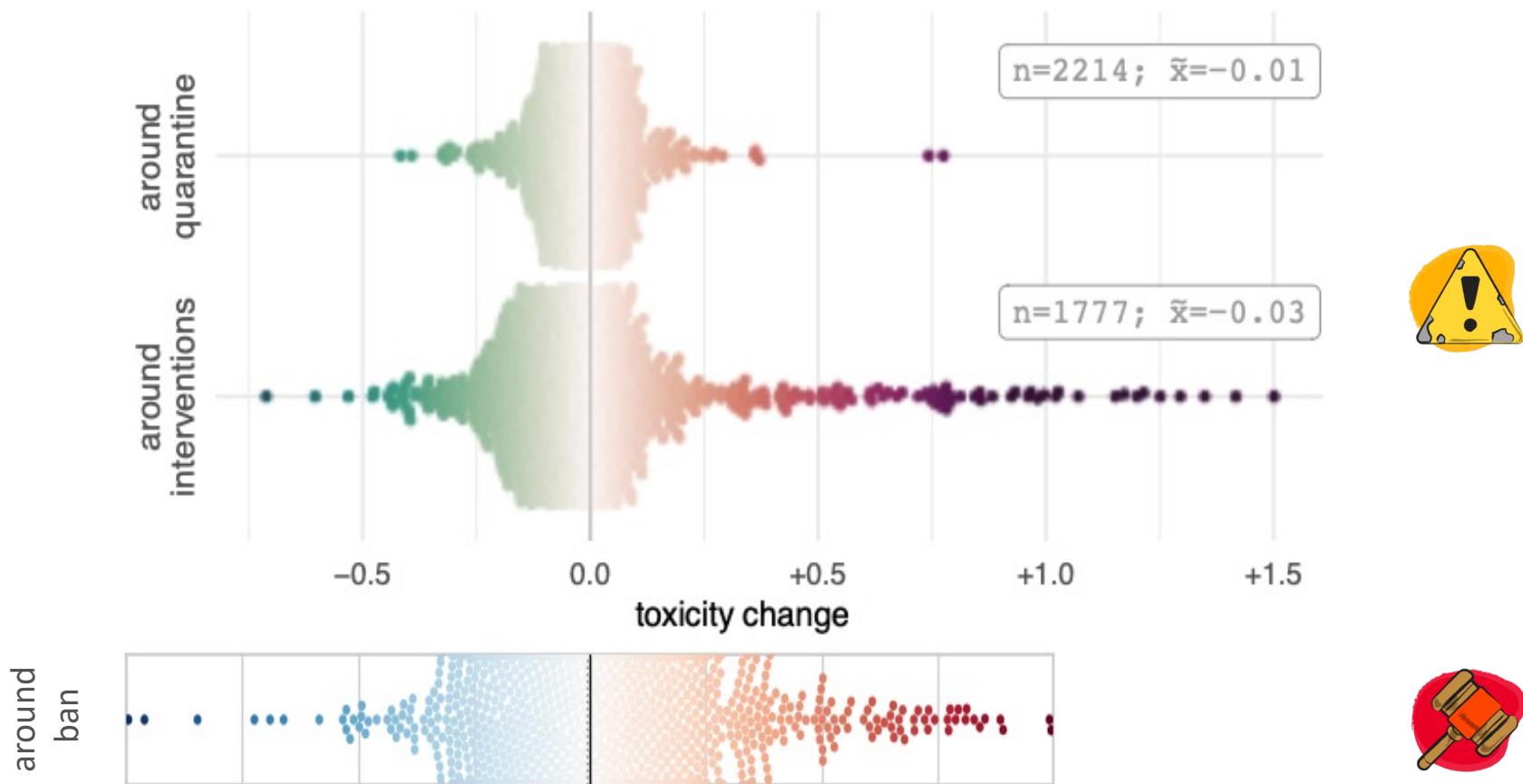


(b) Fraction of severely toxic comments.

**Toxicity surpassed pre-intervention levels
6 months after the Quarantine**

Trujillo, A., & Cresci, S. (2022). Make Reddit Great Again: Assessing community effects of moderation interventions on r/The_Donald. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1-28.

Content Moderation: Does it Work?

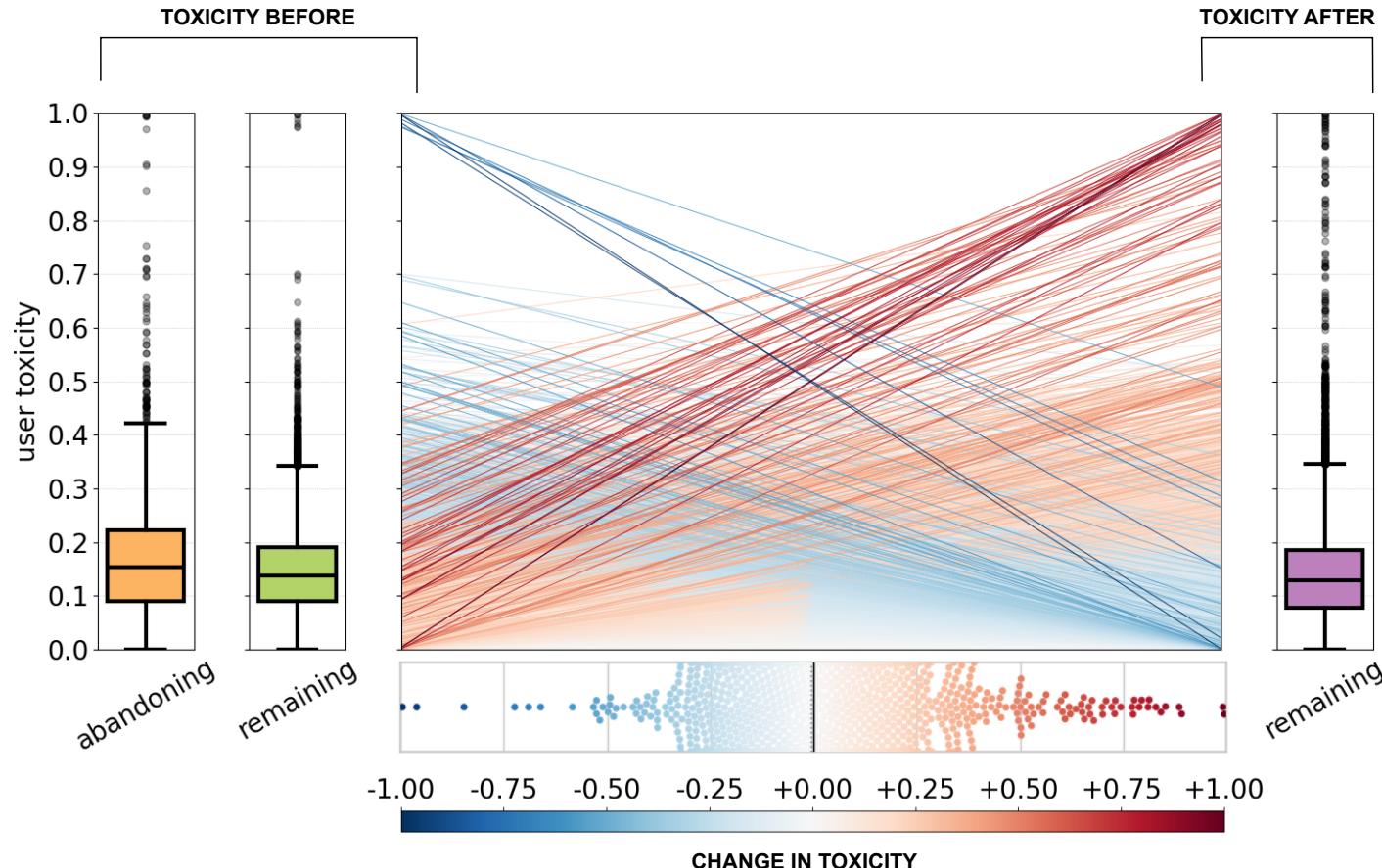


1. quarantine almost **did not change** user toxicity
2. restriction and ban **slightly decreased** the toxicity
3. restriction and ban caused many users to **become extremely toxic**

Content Moderation: Does it Work?



BAN: USER-LEVEL EFFECTS



The median toxicity slightly decreases

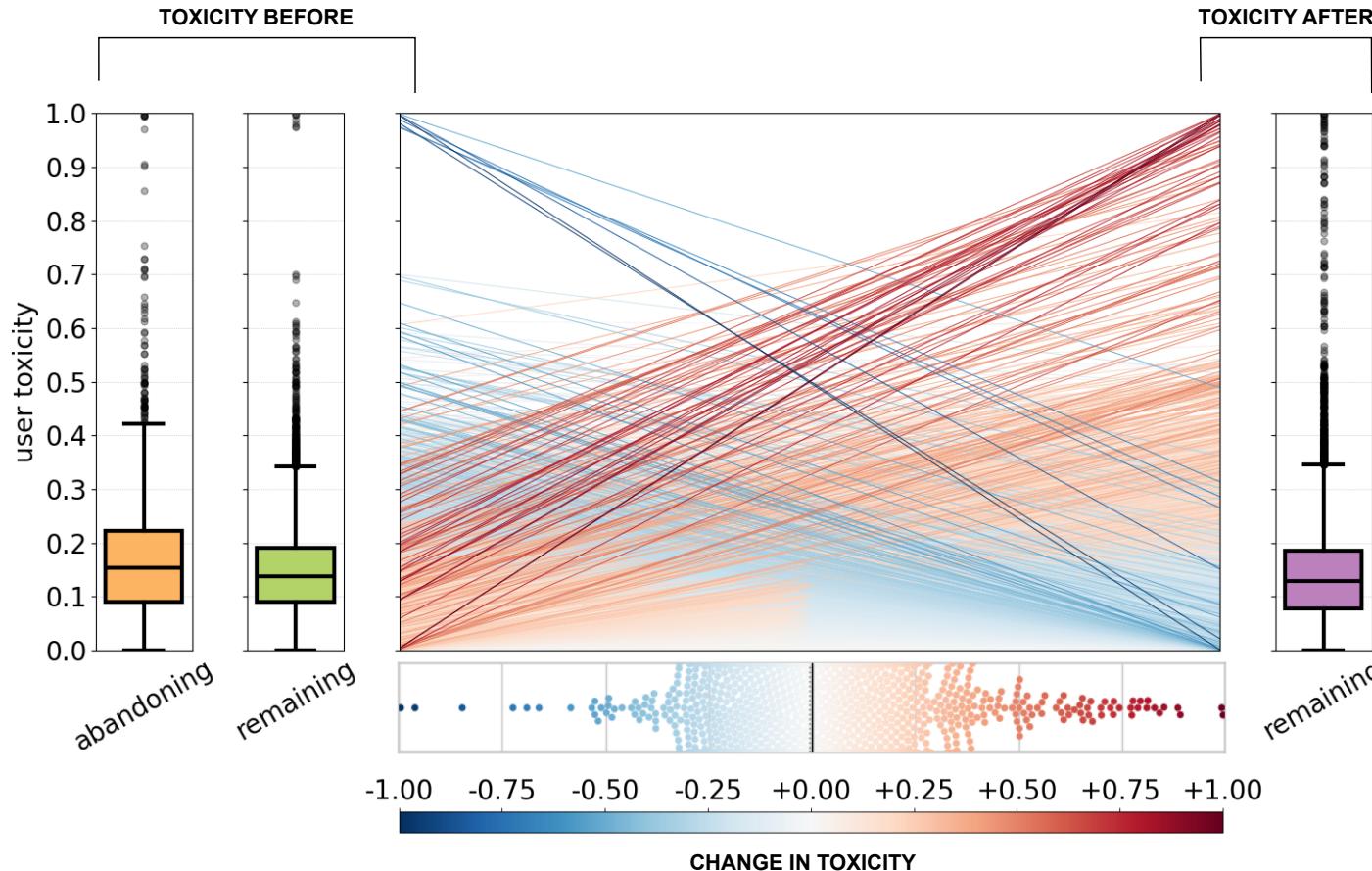
- **(reduction of 6.57%)**

Cima, L., Tessa, B., Cresci, S., Trujillo, A., & Avvenuti, M. (2024). Investigating the heterogenous effects of a massive content moderation intervention via Difference-in-Differences. *arXiv preprint arXiv:2411.04037*.

Content Moderation: Does it Work?



BAN: USER-LEVEL EFFECTS



The median toxicity slightly decreases

- (reduction of 6.57%)

However, the increment tail is heavier

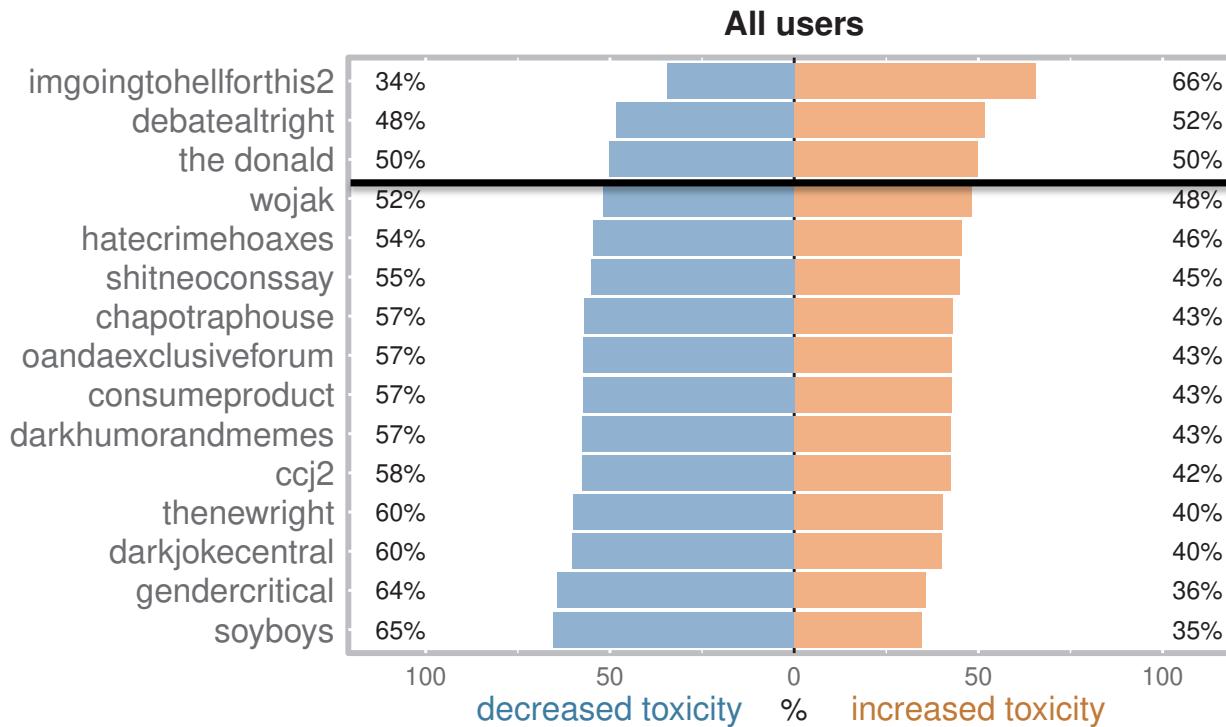
- 5% of users increased their toxicity by more than 70%

Cima, L., Tessa, B., Cresci, S., Trujillo, A., & Avvenuti, M. (2024). Investigating the heterogeneous effects of a massive content moderation intervention via Difference-in-Differences. *arXiv preprint arXiv:2411.04037*.

Content Moderation: Does it Work?



BAN: COMMUNITY-LEVEL EFFECTS



We take the sum of the individual variations of toxicity

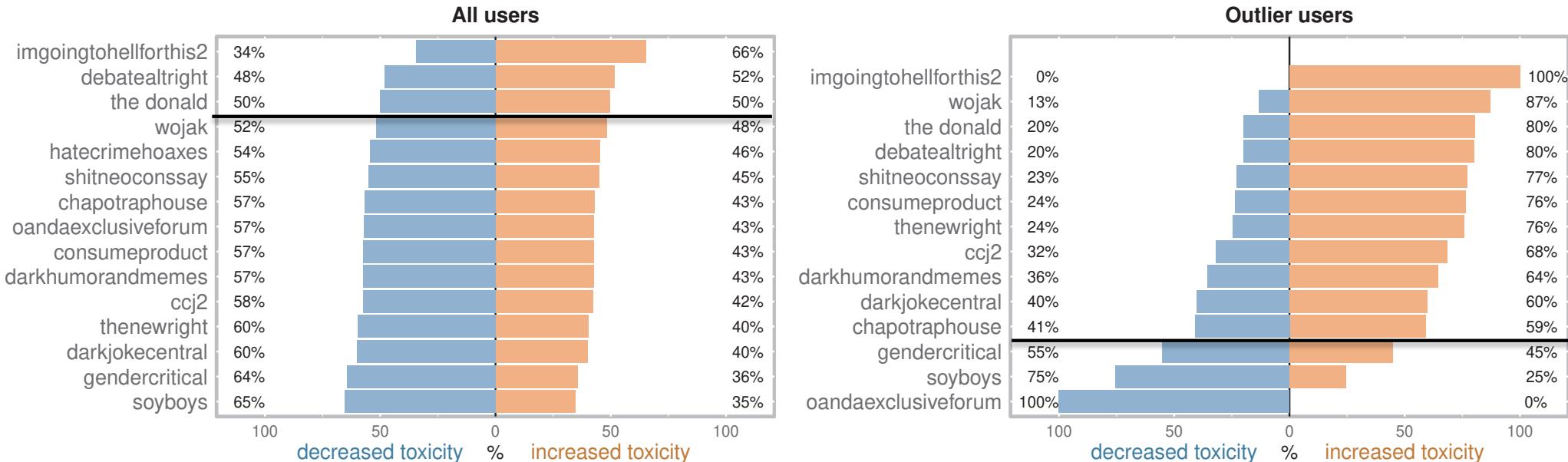
- Considering **all the users**, 12 over 15 subreddits show a general **decrement in toxicity**

Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024, May). The Great Ban: Efficacy and unintended consequences of a massive deplatforming operation on Reddit. In *Companion Publication of the 16th ACM Web Science Conference* (pp. 85-93).

Content Moderation: Does it Work?



BAN: COMMUNITY-LEVEL EFFECTS



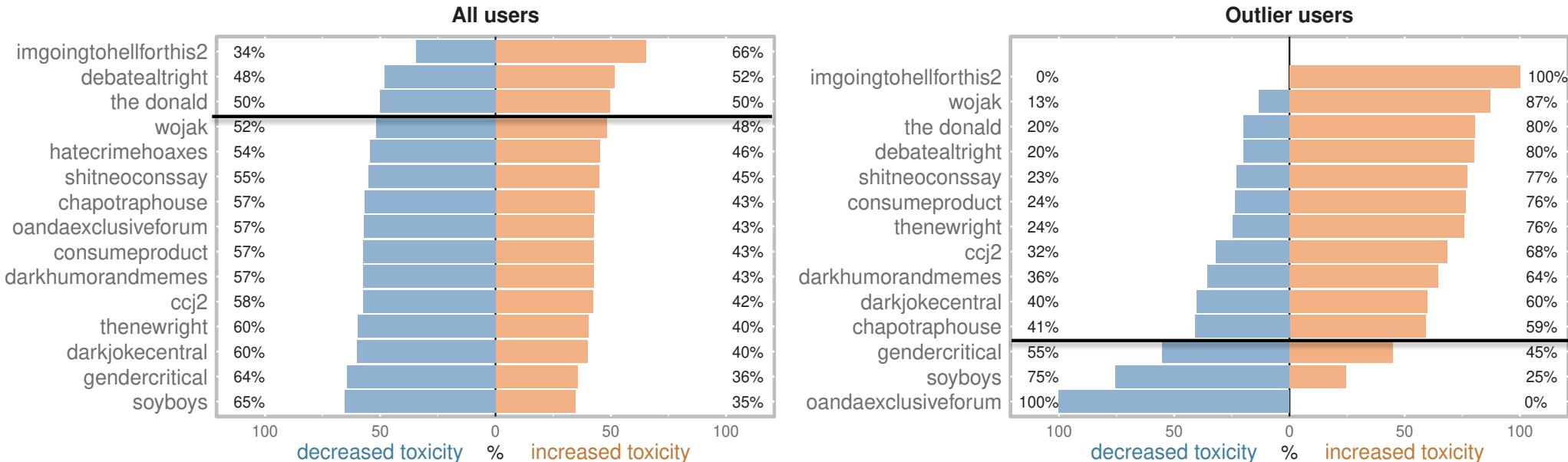
Here we consider **outliers** as the users with $|variation| \geq 0.25$

Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024, May). The Great Ban: Efficacy and unintended consequences of a massive deplatforming operation on Reddit. In *Companion Publication of the 16th ACM Web Science Conference* (pp. 85-93).

Content Moderation: Does it Work?



BAN: COMMUNITY-LEVEL EFFECTS



Here we consider **outliers** as the users with $|variation| \geq 0.25$

- Considering **only outliers**, 12 over 15 subreddits show **an increment in toxicity**

Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024, May). The Great Ban: Efficacy and unintended consequences of a massive deplatforming operation on Reddit. In *Companion Publication of the 16th ACM Web Science Conference* (pp. 85-93).

Content Moderation: Does it Work?



- The intervention was **expensive!**
 - More than **15% of users** abandoned the platform
- The average **toxicity** slightly decreased
- The intervention caused some **undesired side effects**
 - **5% of users increased** toxicity by more than **70%**

Content Moderation: Does it Work?



The intervention was **expensive!**

- More than **15% of users** abandoned the platform

▪ The average **toxicity** slightly decreased

▪ The intervention caused some **undesired side effects**

- **5% of users increased** toxicity by more than **70%**



Problem Framing

- User abandonment has a severe impact on online platforms.

Tessa, B., Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024). Beyond Trial-and-Error: Predicting User Abandonment After a Moderation Intervention. *arXiv preprint arXiv:2404.14846*.



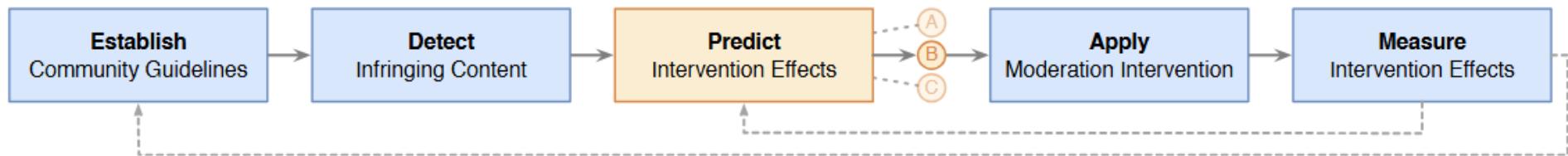
Problem Framing

- User abandonment has a severe impact on online platforms.
- Current moderation approaches rely on trial and error. Inefficient and ineffective!

Tessa, B., Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024). Beyond Trial-and-Error: Predicting User Abandonment After a Moderation Intervention. *arXiv preprint arXiv:2404.14846*.

Problem Framing

- User abandonment has a severe impact on online platforms.
- Current moderation approaches rely on trial and error. Inefficient and ineffective!
- We propose ML models to predict user abandonment.



Tessa, B., Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024). Beyond Trial-and-Error: Predicting User Abandonment After a Moderation Intervention. *arXiv preprint arXiv:2404.14846*.



Features

- Extraction of **142 features** from **pre-ban** comments, grouped into 4 categories:
 - **Activity** (e.g. number of comments, average time between the sharing of two subsequent comments)
 - **Relational** (e.g. degree of influence that each user had in the banned subreddits)
 - **Toxicity** (e.g. toxicity and sentiment scores)
 - **Writing Style** (e.g. counts of the different parts-of-speech used)



How to define abandonment?

We distinguish two types of abandonment:



How to define abandonment?

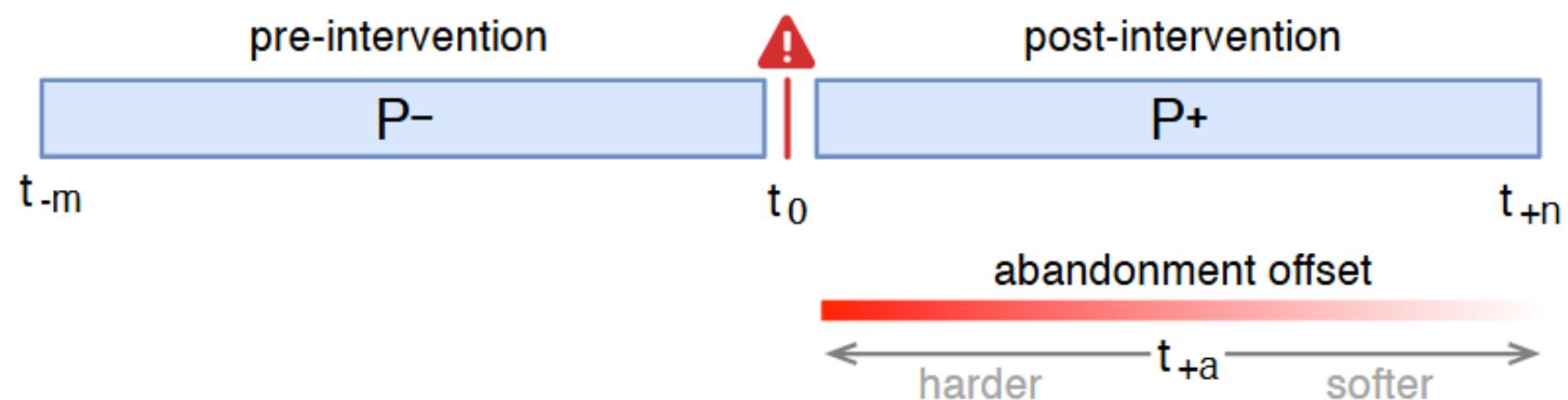
We distinguish two types of abandonment:

- **Hard Abandonment** → Users who **never posted again** after the ban date (15% of the users)

How to define abandonment?

We distinguish two types of abandonment:

- **Hard Abandonment** → Users who **never posted again** after the ban date (15% of the users)
- **Soft Abandonment** → Users who **haven't posted in the last 3 months** (27% of the users)





Model Training and Optimization

- 10-fold cross-validation with:
 - Feature scaling
 - Feature selection
 - Class rebalancing



Model Training and Optimization

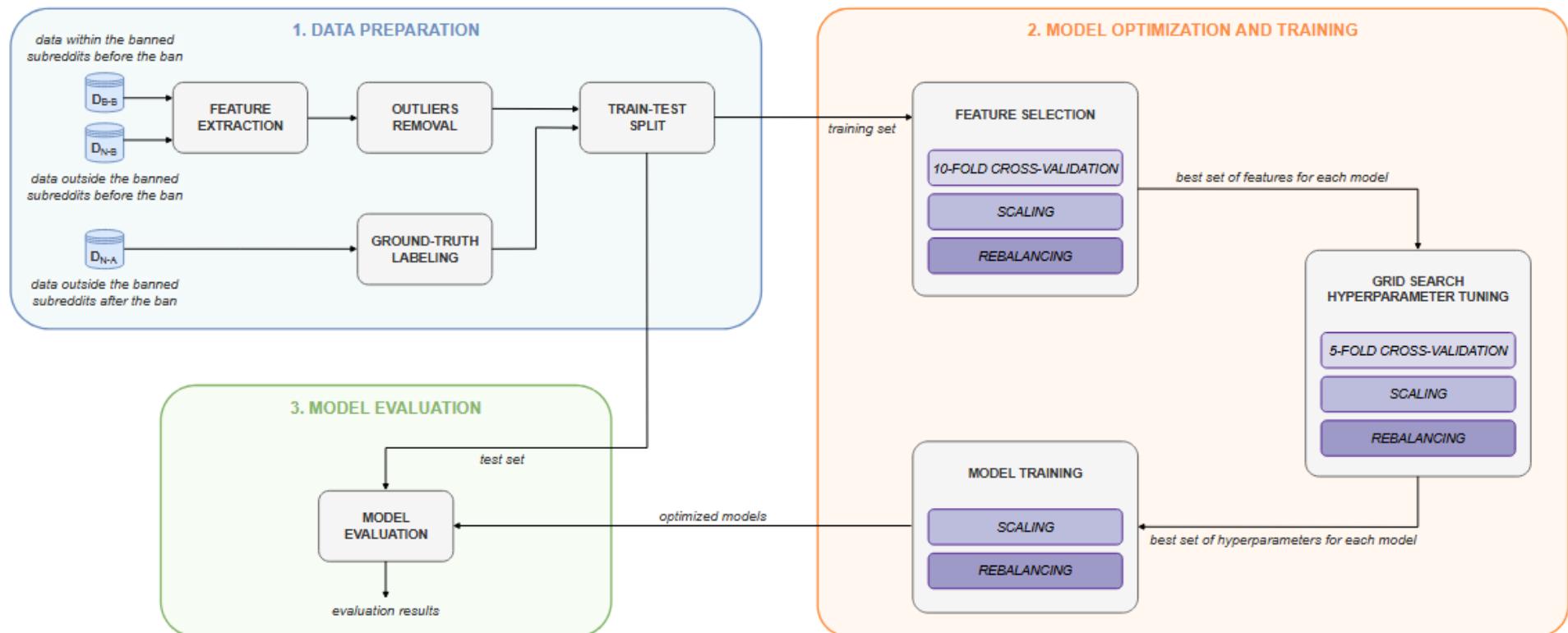
- 10-fold cross-validation with:
 - Feature scaling
 - Feature selection
 - Class rebalancing
- Choose best set of features for each model



Model Training and Optimization

- 10-fold cross-validation with:
 - Feature scaling
 - Feature selection
 - Class rebalancing
- Choose best set of features for each model
- Hyperparameter optimization

Pipeline





Model Training and Optimization

task	model	p	positive class				overall		
			precision	recall	F1	95% CI	AUC	micro F1	95% CI
<i>baselines</i>									
hard abandonment	Stratified	-	0.155	0.158	0.157	-	0.501	0.751	-
	DT Ratio	1	0.356	0.199	0.255	-	0.537	0.689	-
	DT Trend	1	0.325	0.219	0.262	-	0.591	0.726	-
	NB-NFE	142	0.206	0.313	0.248	-	0.595	0.814	-
	DT-NFE	142	<u>0.636</u>	0.230	0.338	-	0.630	0.627	-
	<i>trained models</i>								
hard abandonment	KNN	20	0.514	0.308	0.384	[0.332, 0.363]	0.719	0.754	[0.703, 0.715]
	RF	142	0.634	<u>0.723</u>	<u>0.675</u>	[0.380, 0.600]	0.909	<u>0.909</u>	[0.862, 0.895]
	AB	142	0.630	0.691	0.659	[0.412, 0.597]	<u>0.918</u>	0.902	[0.768, 0.890]
	DT	10	0.617	0.330	0.430	[0.315, 0.360]	0.752	0.756	[0.710, 0.740]
	GB	142	0.658	0.736	0.695	[0.397, 0.608]	0.930	0.914	[0.719, 0.891]
	SVM	50	0.520	0.373	0.434	[0.437, 0.465]	0.781	0.797	[0.803, 0.817]

Highest F1 score: **0.695** obtained with **Gradient Boosting**

Model Training and Optimization

task	model	p	positive class				overall		
			precision	recall	F1	95% CI	AUC	micro F1	95% CI
<i>baselines</i>									
soft abandonment	Stratified	-	0.263	0.266	0.264	-	0.509	0.605	-
	DT Ratio	1	0.374	0.328	0.349	-	0.527	0.624	-
	DT Trend	1	0.282	0.319	0.299	-	0.549	0.644	-
	NB-NFE	142	0.091	0.540	0.156	-	0.604	0.734	-
	DT-NFE	142	0.589	0.287	0.386	-	0.524	0.494	-
<i>trained models</i>									
soft abandonment	KNN	10	0.513	0.438	0.473	[0.428, 0.462]	0.698	0.691	[0.644, 0.664]
	NB	10	0.400	0.485	0.438	[0.413, 0.450]	0.685	0.723	[0.778, 0.787]
	RF	142	0.473	0.744	0.578	[0.522, 0.572]	0.801	0.814	[0.745, 0.800]
	AB	142	0.504	<u>0.706</u>	<u>0.588</u>	[0.525, 0.574]	<u>0.806</u>	<u>0.809</u>	[0.722, 0.815]
	DT	20	<u>0.558</u>	0.443	0.494	[0.387, 0.423]	0.695	0.690	[0.610, 0.629]
	GB	142	0.513	0.694	0.590	[0.517, 0.579]	0.816	0.808	[0.639, 0.788]
	SVM	50	0.484	0.513	0.499	[0.490, 0.515]	0.728	0.737	[0.736, 0.751]

Highest F1 score: **0.590** obtained with **Gradient Boosting**



Feature Importance

- We analyze **which features** (and feature groups) best predict **hard** and **soft** abandonment.



Feature Importance

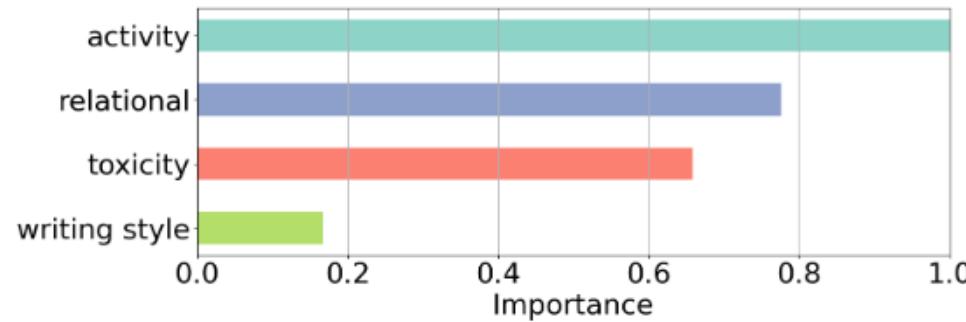
- We analyze **which features** (and feature groups) best predict **hard** and **soft** abandonment.
- This improves **model interpretability** and supports **transparent decision-making**.



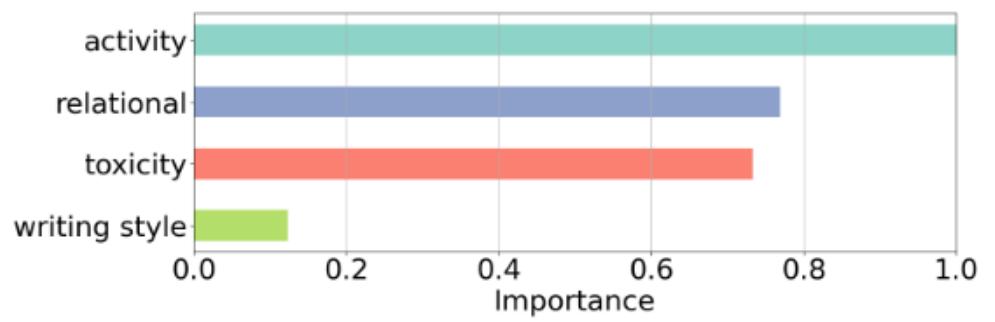
Feature Importance

- We analyze **which features** (and feature groups) best predict **hard** and **soft** abandonment.
- This improves **model interpretability** and supports **transparent decision-making**.
- Insights help moderators understand **why users leave**, beyond raw predictions.

Feature Importance



(a) Task: hard abandonment. Feature importance: SHAP.



(b) Task: soft abandonment. Feature importance: SHAP.



Conclusions

- Results are **not yet strong**, but **promising** for a first attempt.

Tessa, B., Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024). Beyond Trial-and-Error: Predicting User Abandonment After a Moderation Intervention. *arXiv preprint arXiv:2404.14846*.



Conclusions

- Results are **not yet strong**, but **promising** for a first attempt.
- Highlight the **need for new and richer features**.

Tessa, B., Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024). Beyond Trial-and-Error: Predicting User Abandonment After a Moderation Intervention. *arXiv preprint arXiv:2404.14846*.



Conclusions

- Results are **not yet strong**, but **promising** for a first attempt.
- Highlight the **need for new and richer features**.
- Future work: move towards a **regression task** for a more nuanced analysis.

Tessa, B., Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024). Beyond Trial-and-Error: Predicting User Abandonment After a Moderation Intervention. *arXiv preprint arXiv:2404.14846*.

Content Moderation: Does it Work?

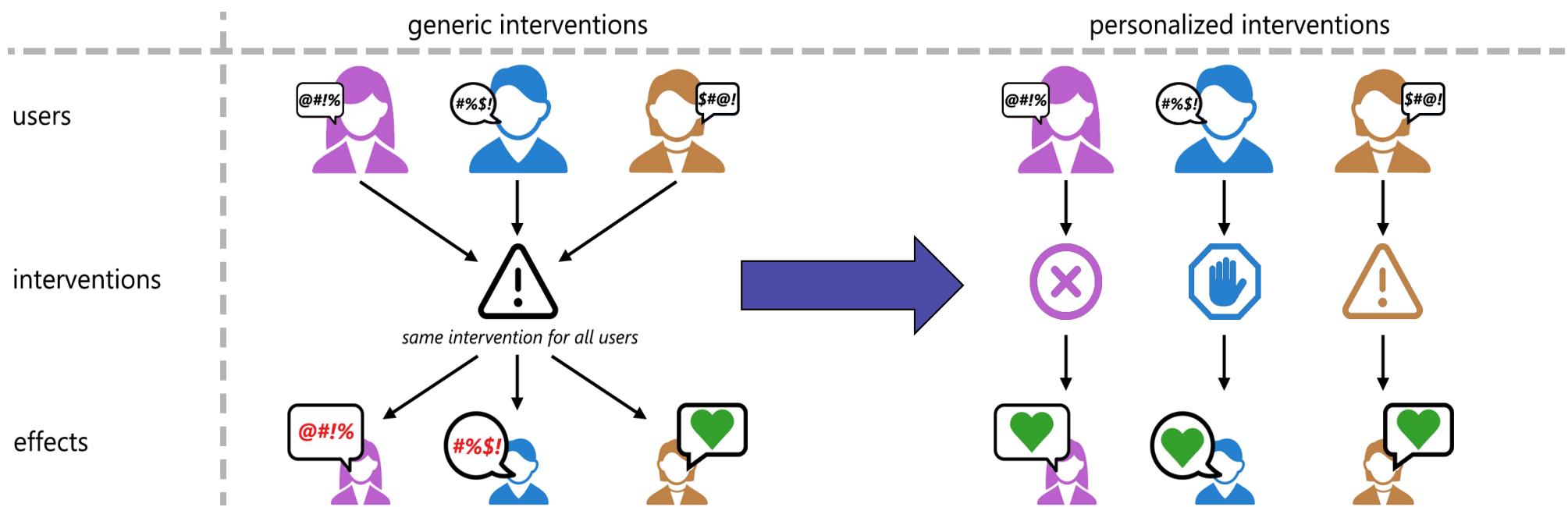


- The intervention was **expensive!**
 - More than **15% of users** abandoned the platform
- The average **toxicity** slightly decreased
- The intervention caused some **undesired side effects**
 - **5% of users increased** toxicity by more than **70%**

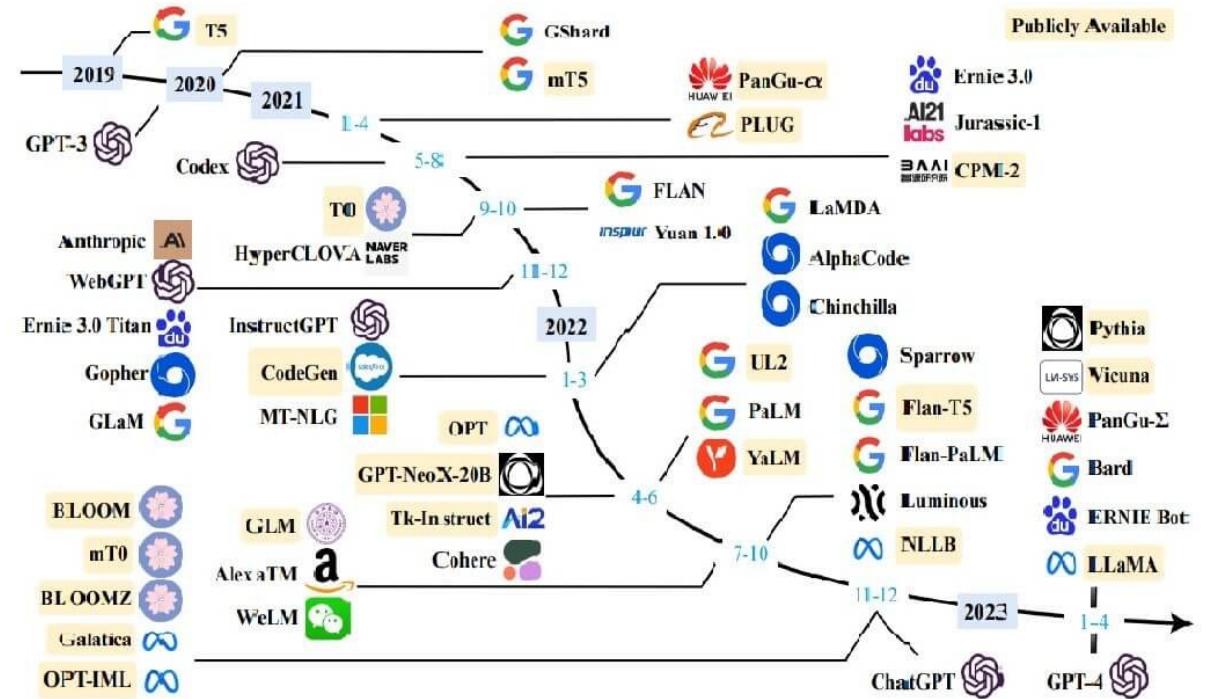
Content Moderation: Solution

People react differently!

Need for personalized interventions

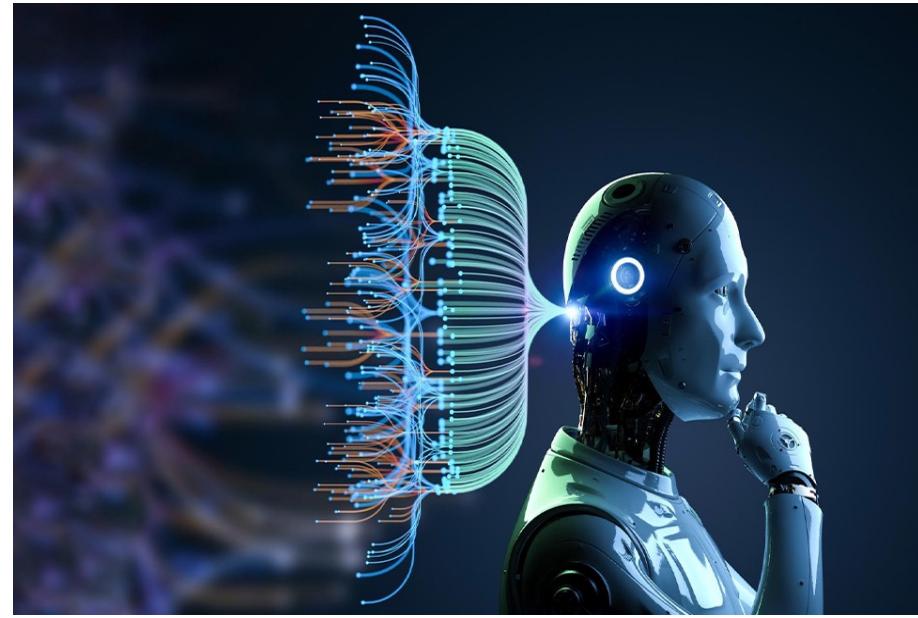


A modern way to develop **automatic** and **personalized** intervention is by using **Large Language Models (LLMs)**





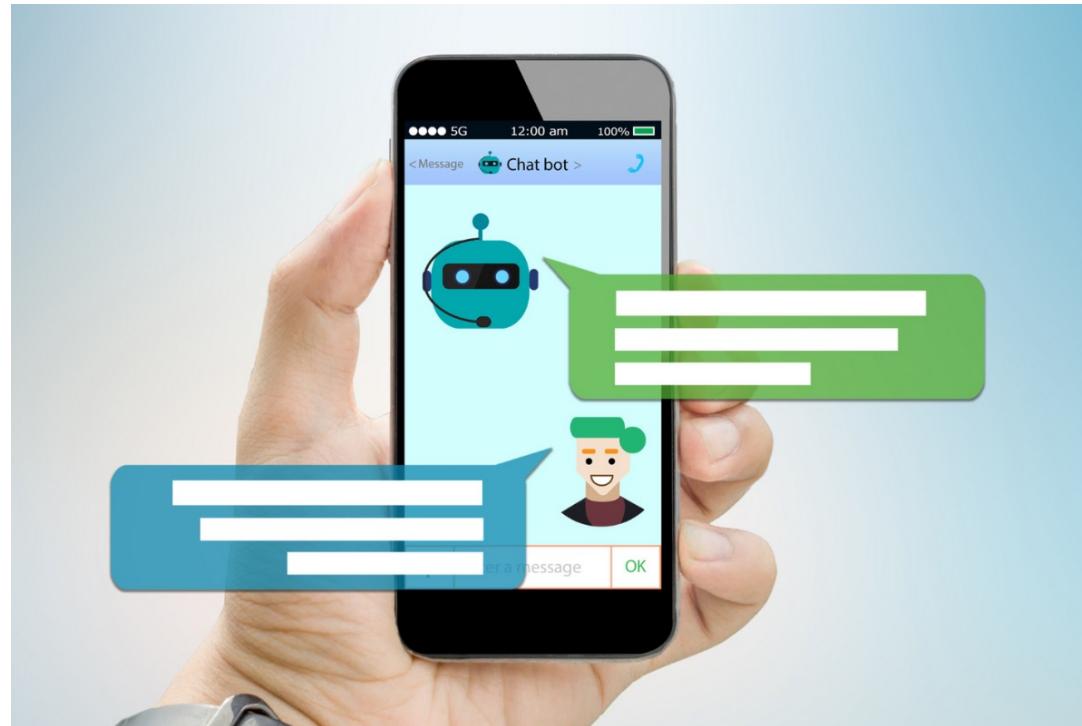
Text-based LLMs



The most famous is **ChatGPT**
Generative
Pre-trained
Transformer

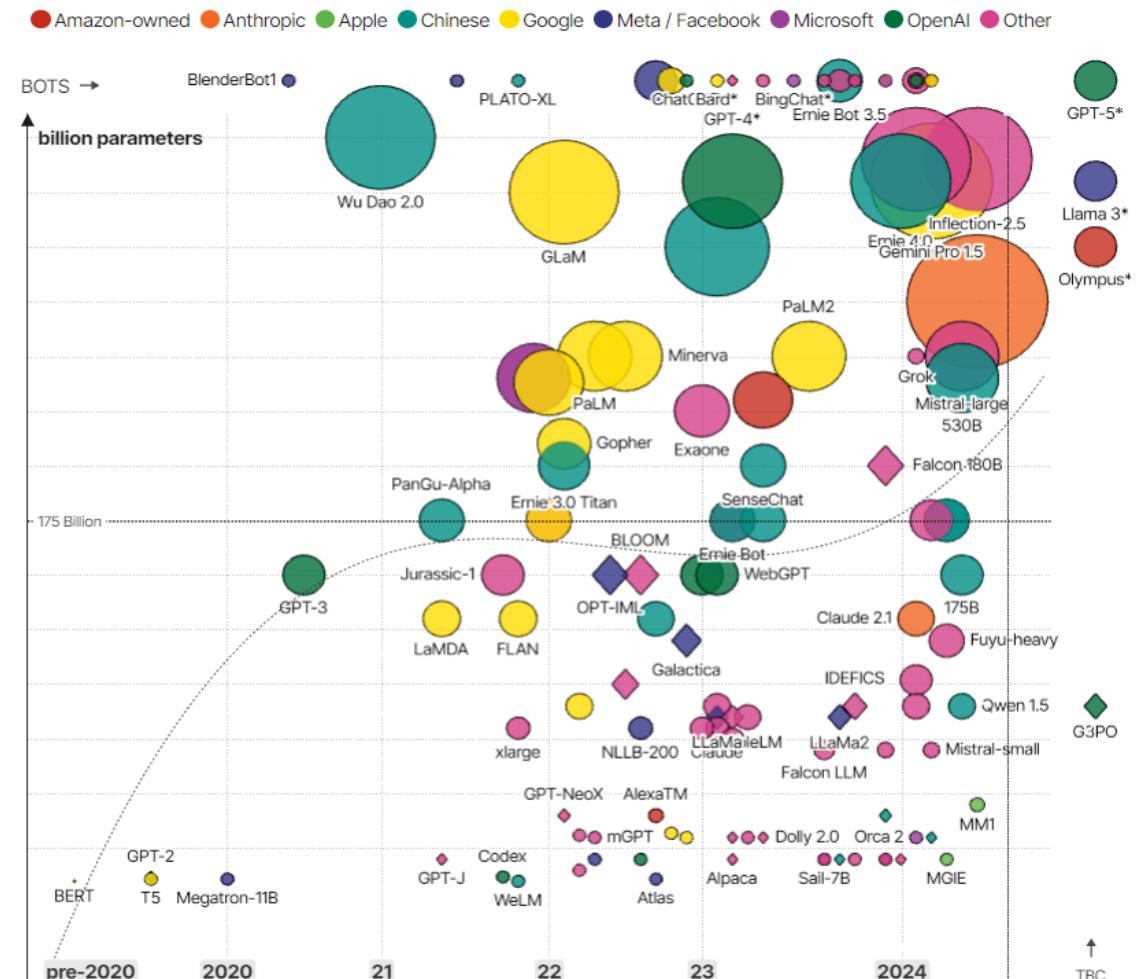
Text-based LLMs

- Text-based LLMs take as input a sentence and «make an answer»
- They are a more sophisticated **kind of chatbot**
- How can they answer to our sentences?



Training

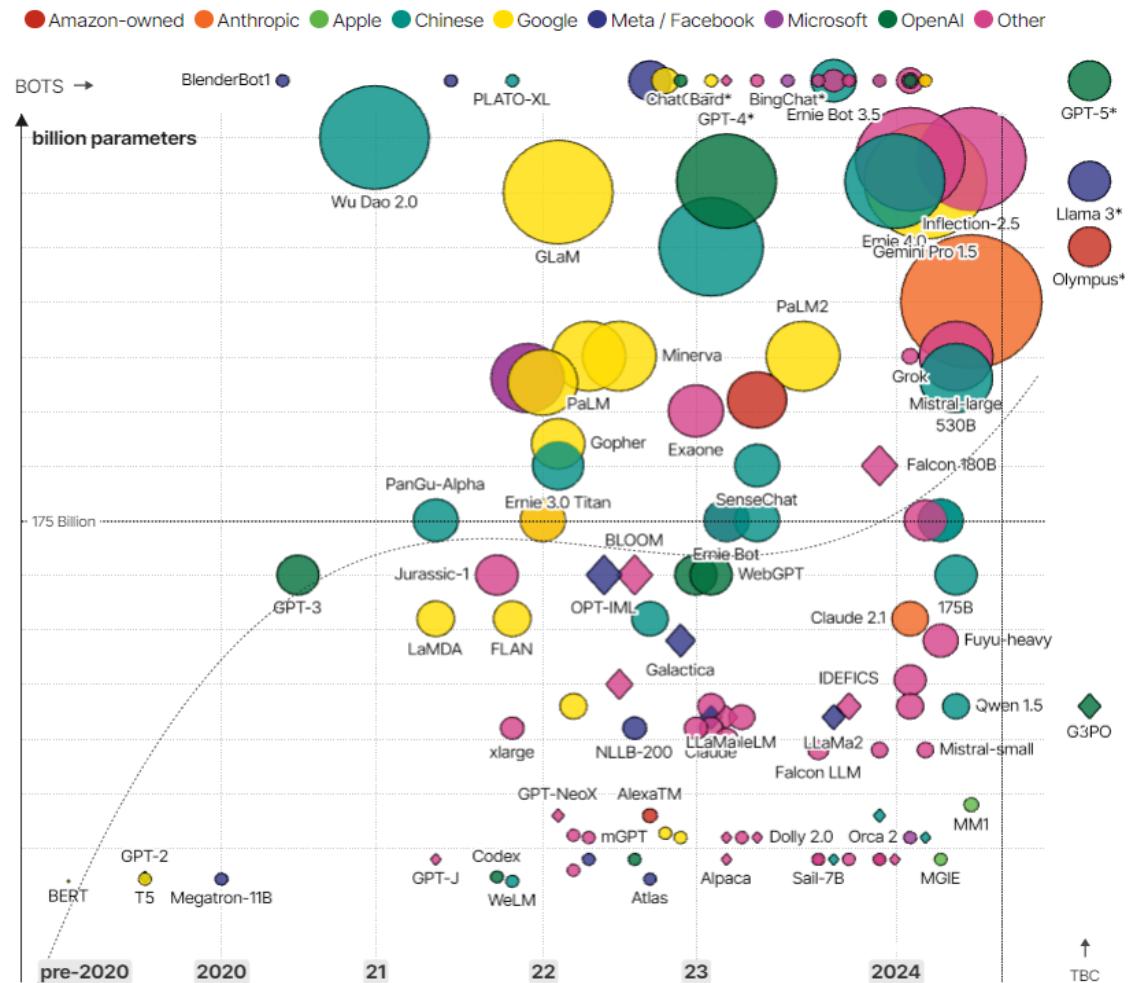
- They are **pre-trained** on a lot of human data, like:
 - Journals
 - Papers
 - Social media conversations
 - ...



Training

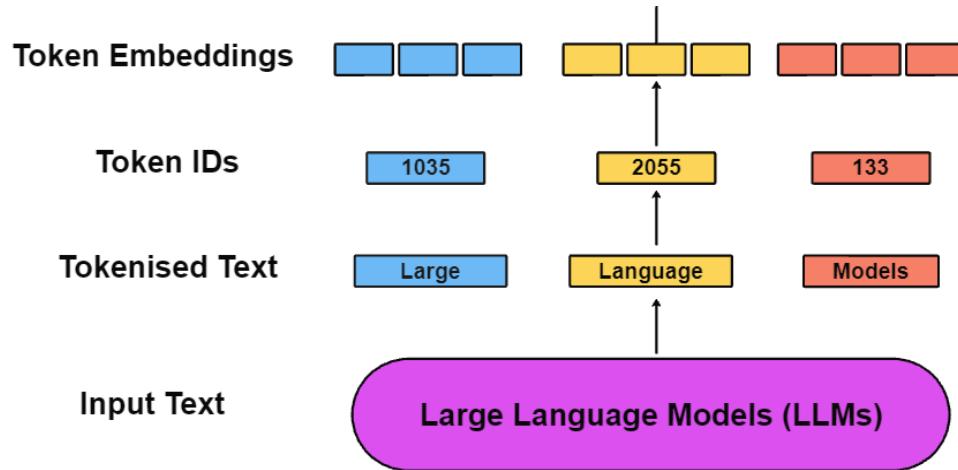
- They are **pre-trained** on a lot of human data, like:
 - Journals
 - Papers
 - Social media conversations
 - ...

- They learn:
 - Grammar
 - Facts
 - Words relationship





Input Processing

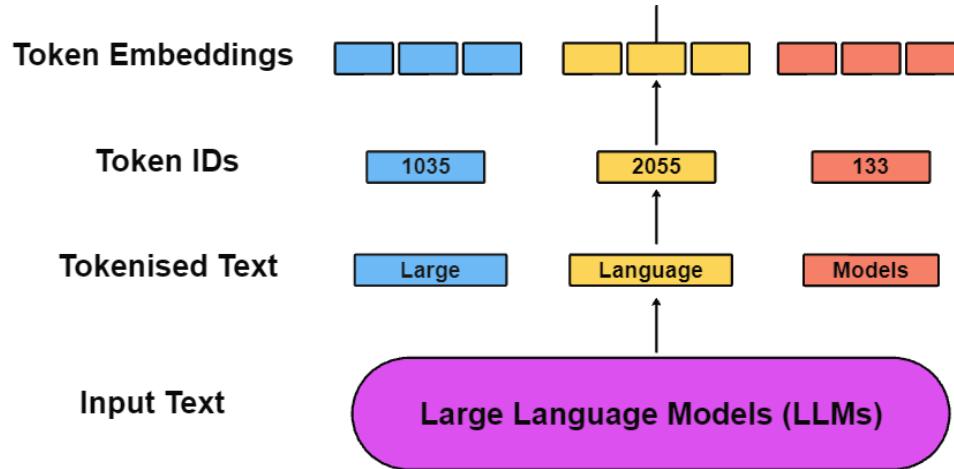


1. The text is **tokenized**

- Each model has its own tokenizer
 - Token = set of words



Input Processing



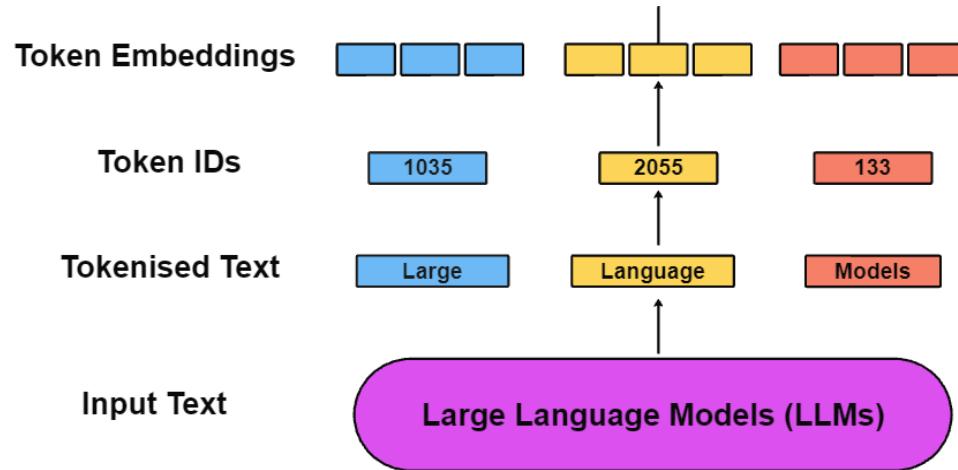
2. Each token is mapped to an integer using a **dictionary**

- Each model has its own dictionary
- Tokens are **categorical**, so not suitable for neural network training





Input Processing



1. The text is **tokenized**

- Each model has its own tokenizer
 - Token = set of words

1. Each token is mapped to an integer using a **dictionary**

- Each model has its own dictionary
- Tokens are **categorical**, so not suitable for neural network training



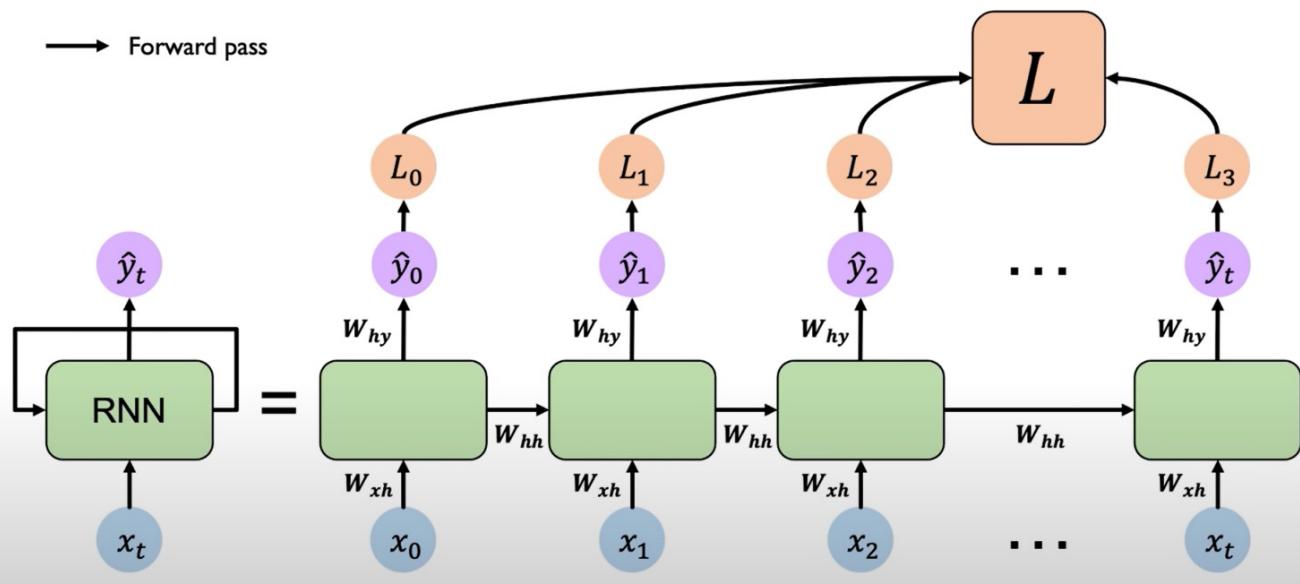
2. The token IDs are represented in a vector space (**embeddings**)

- The vector space is **continuous**



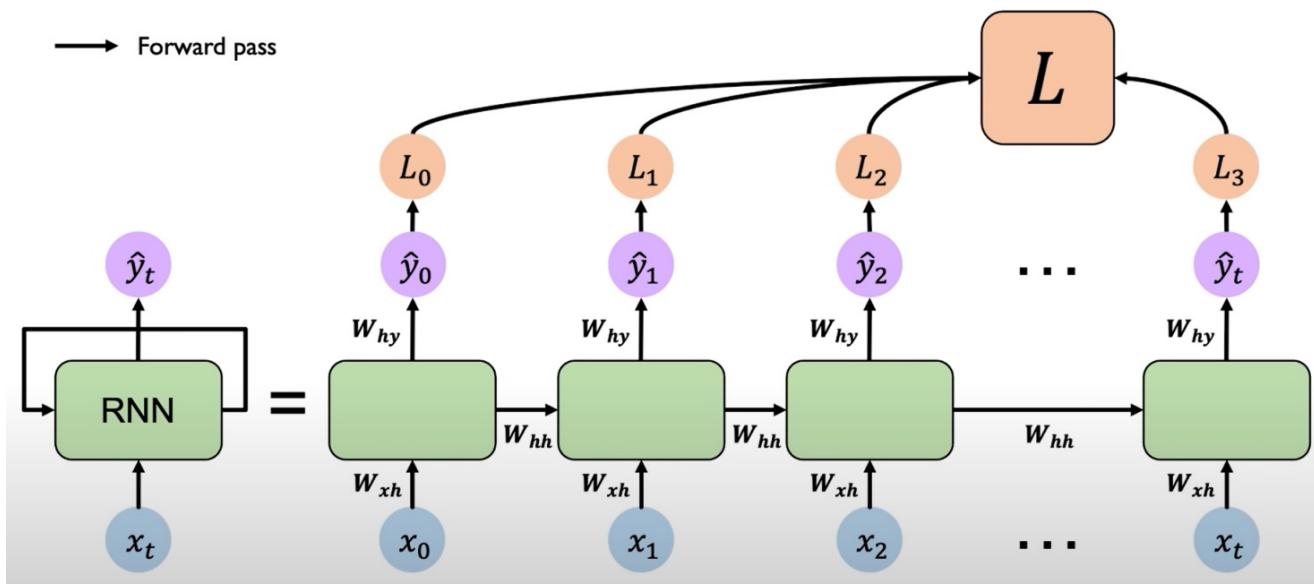
Embedding Processing

- LLMs are based on **Recurrent Neural Networks (RNN)**



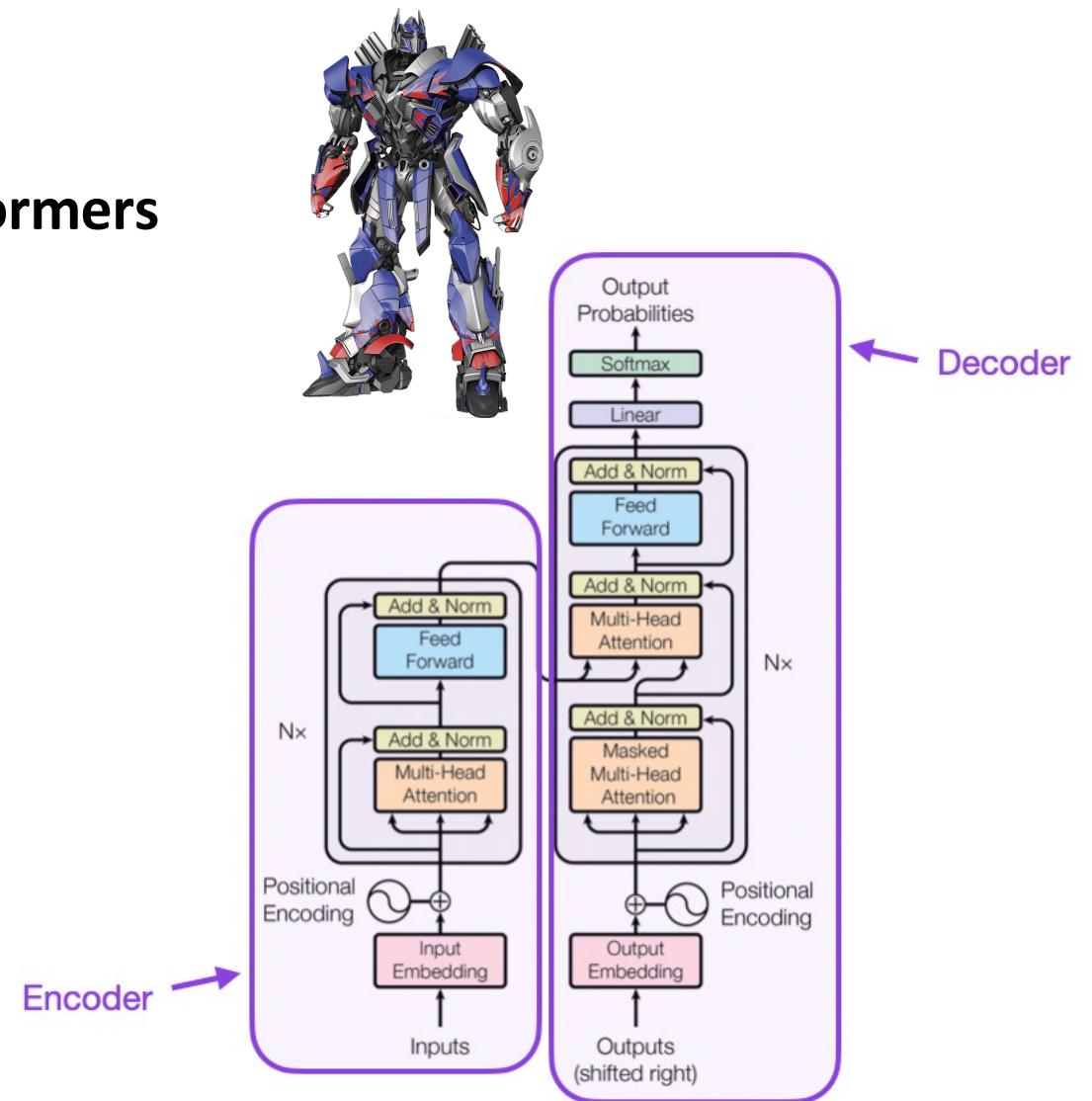
Embedding Processing

- LLMs are based on **Recurrent Neural Networks (RNN)**
- However, **RNNs are not efficient**
 - Each token in a sentence is **processed sequentially**



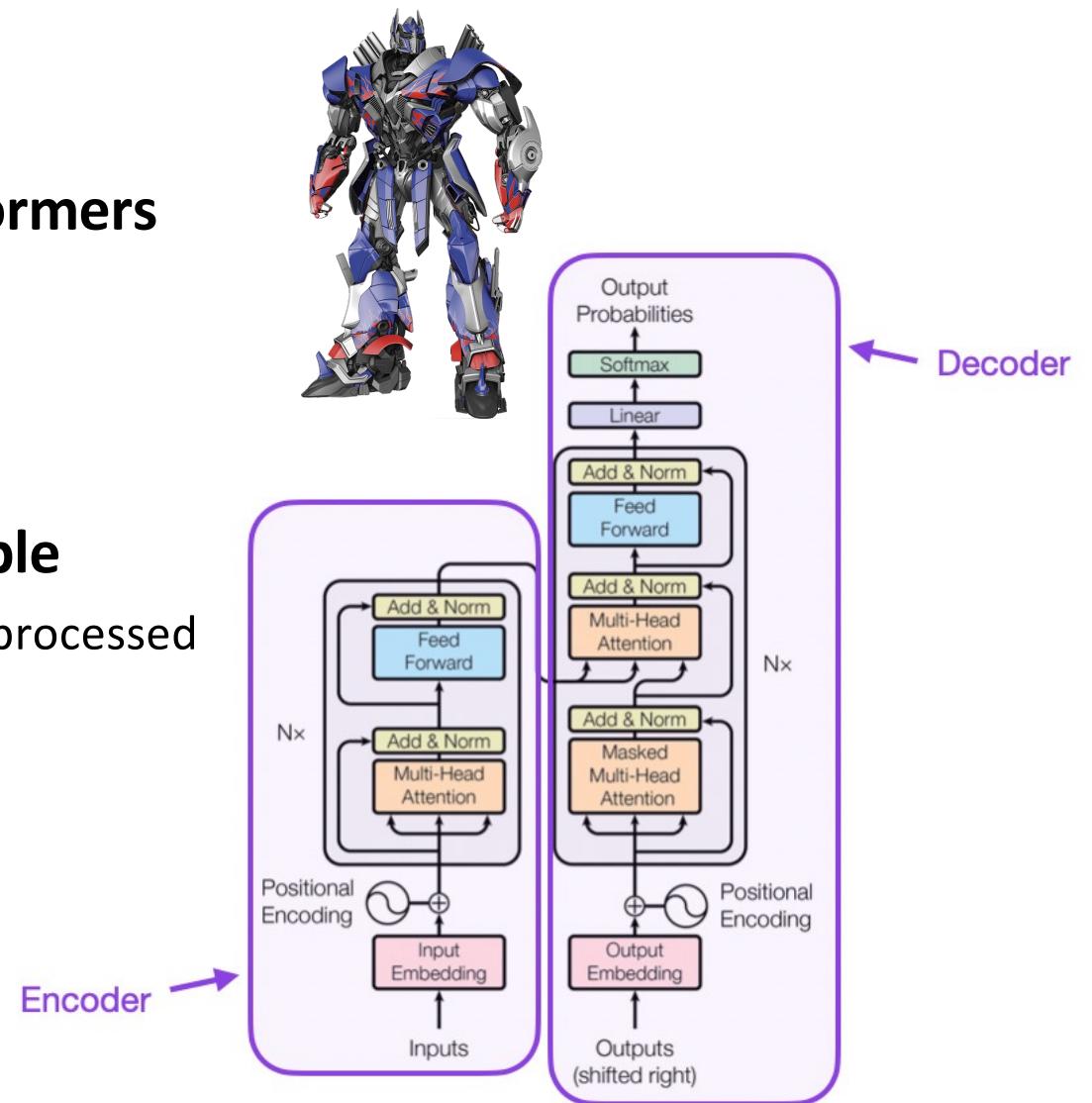
Embedding Processing

- Modern LLMs relied on **Transformers architecture**

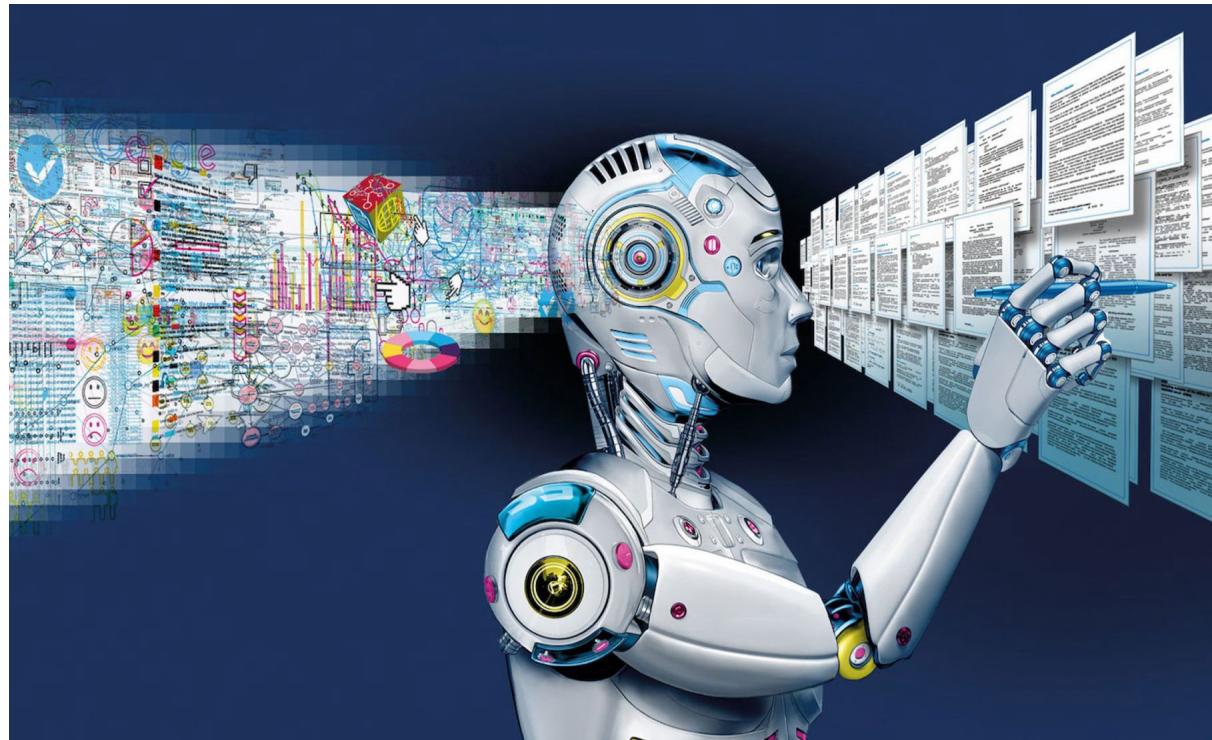


Embedding Processing

- Modern LLMs relied on **Transformers architecture**
- Token processing is **parallelizable**
 - All the tokens on a sentence are processed simultaneously



Inference



- Time to generate!
- Three inference methods:
 1. Greedy decoder
 2. Beam search
 3. Nucleus sampling



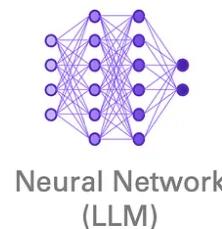
Greedy Decoder

- The model selects the token with the highest probability as the next word

[The cat likes to sleep in the ___] → What **word** comes next?

[The cat likes to sleep in the]

Input



Neural Network
(LLM)

Word	Probability
ability	0.002
bag	0.071
box	0.085
...	...
zebra	0.001

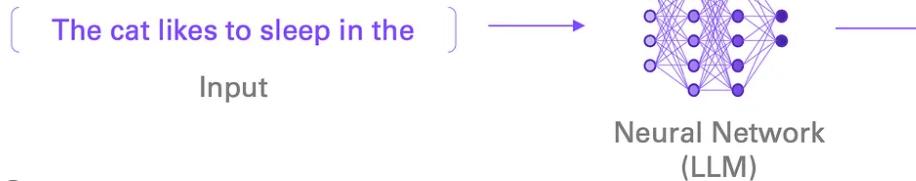
Output



Greedy Decoder

- The model selects the token with the highest probability as the next word

[The cat likes to sleep in the ___] → What **word** comes next?



Word	Probability
ability	0.002
bag	0.071
box	0.085
...	...
zebra	0.001

Output

- PROs**

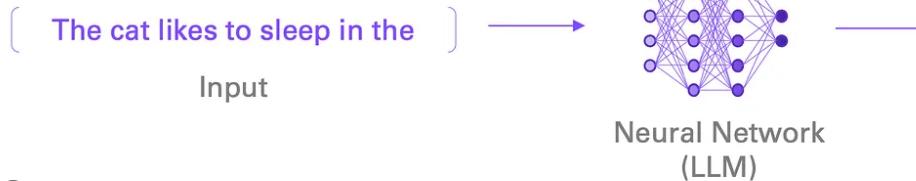
- Fast to compute
- Outputs are coherent and straightforward



Greedy Decoder

- The model selects the token with the highest probability as the next word

[The cat likes to sleep in the ___] → What **word** comes next?



Word	Probability
ability	0.002
bag	0.071
box	0.085
...	...
zebra	0.001

Output

PROs

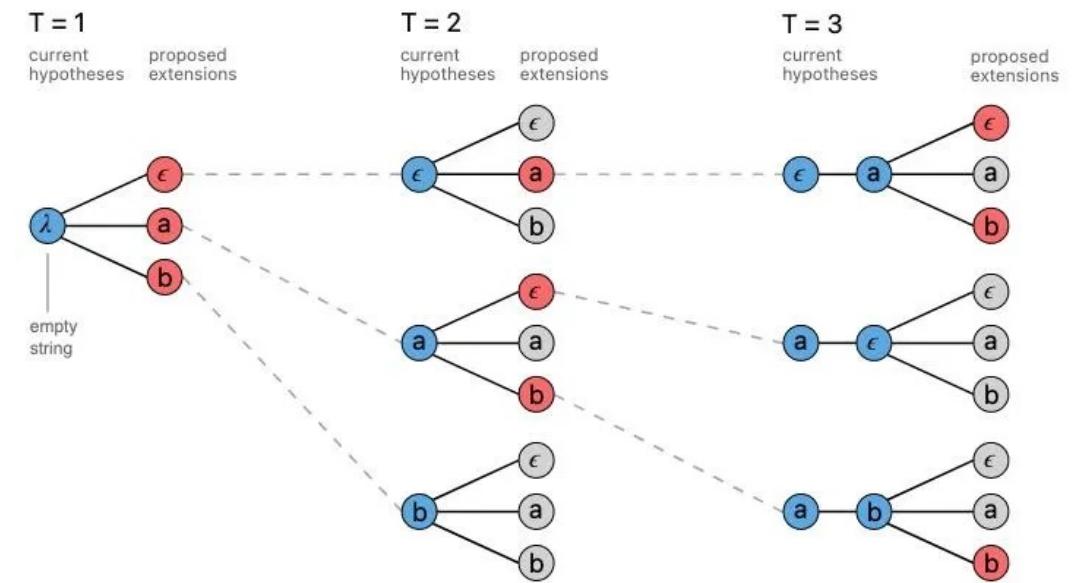
- Fast to compute
- Outputs are coherent and straightforward

CONs

- It doesn't explore all the possibilities
- It may produce repetitive or simplistic text

Beam Search

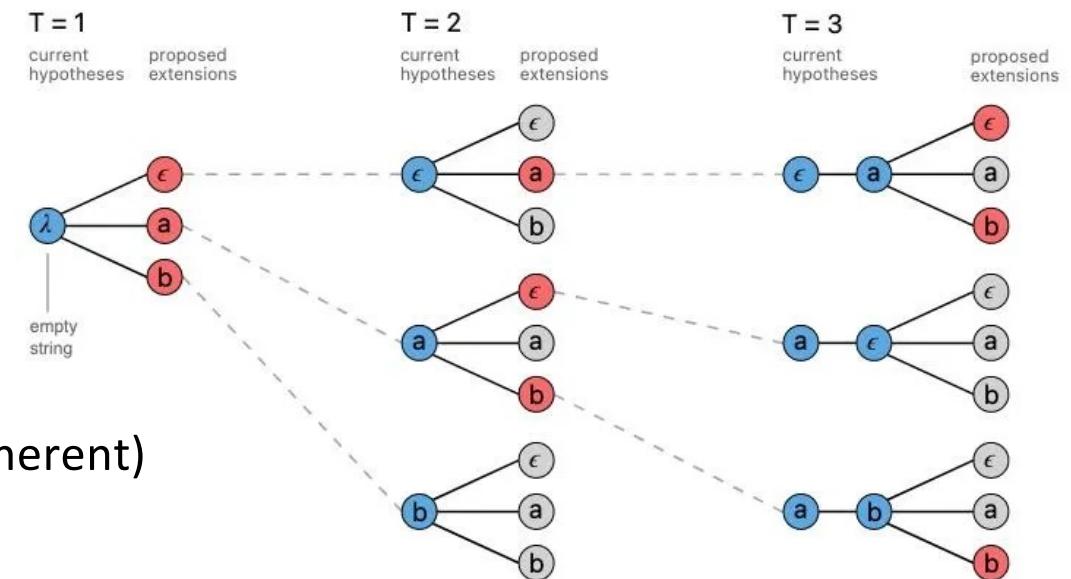
- Maintains multiple sequences (**beams**) of tokens at each step and expands them based on probability
- Only the most likely beams are retained for further exploration



Beam Search

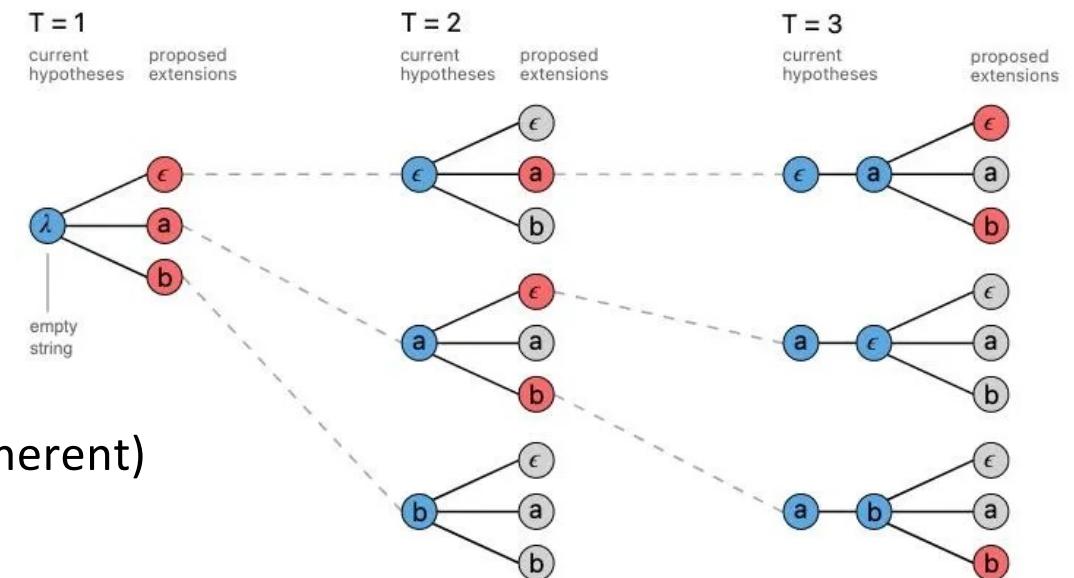
- Maintains multiple sequences (**beams**) of tokens at each step and expands them based on probability
- Only the most likely beams are retained for further exploration

- PROs**
 - High-quality outputs
 - Perfect for translations (highly coherent)



Beam Search

- Maintains multiple sequences (**beams**) of tokens at each step and expands them based on probability
- Only the most likely beams are retained for further exploration

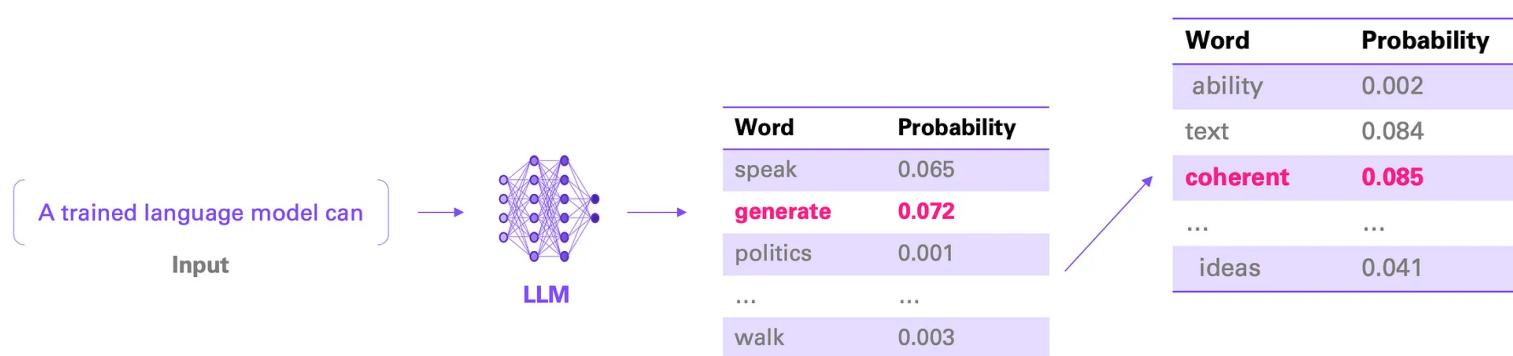


- PROs**
 - High-quality outputs
 - Perfect for translations (highly coherent)
- CONS**
 - Computationally expensive
 - A narrow beam could result in generic outputs



Nucleous Sampling

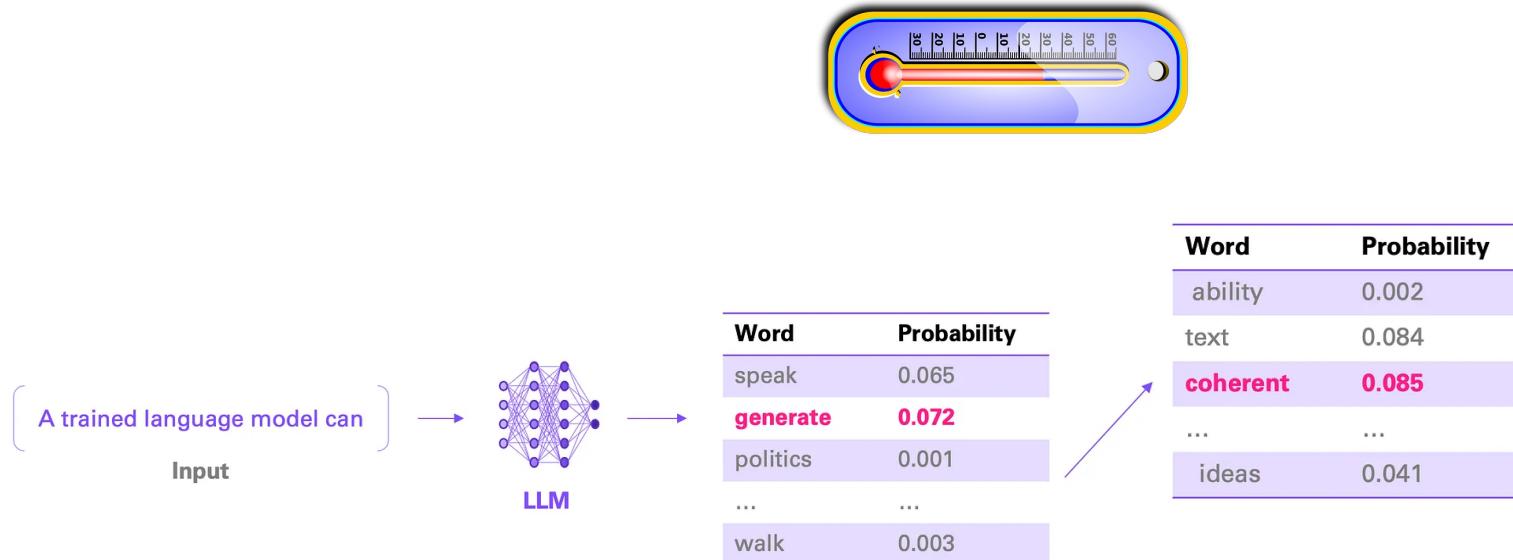
- Limits the model's choices to the **top-k tokens** with the highest probabilities and then **samples** from this restricted set.
 - This is the most-used technique





Nucleous Sampling

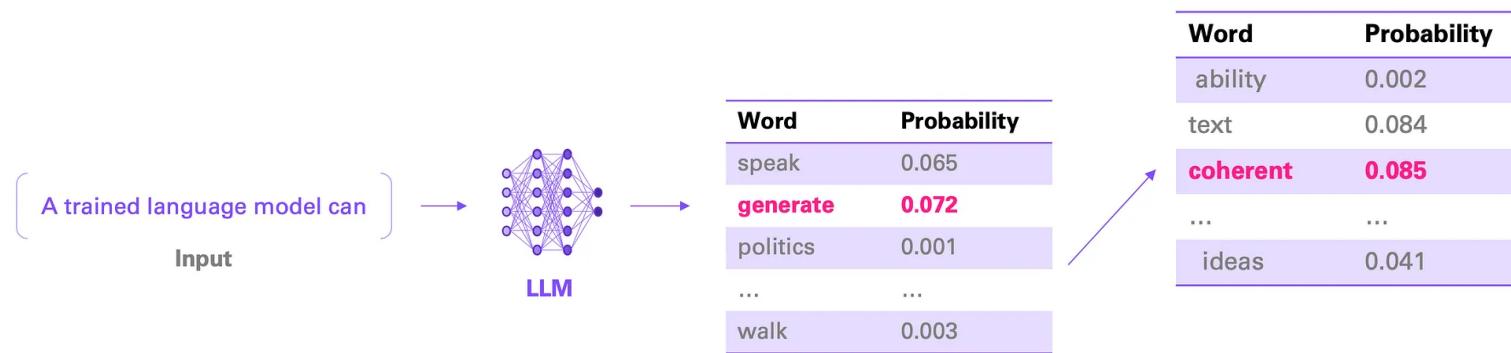
- Limits the model's choices to the **top-k tokens** with the highest probabilities and then **samples** from this restricted set.
 - This is the most-used technique
- The sampling follows a function that:
 - Privileges tokens having high-probability
 - Makes samplings variable based on the **temperature** attribute





Nucleous Sampling

- Limits the model's choices to the **top-k tokens** with the highest probabilities and then **samples** from this restricted set.

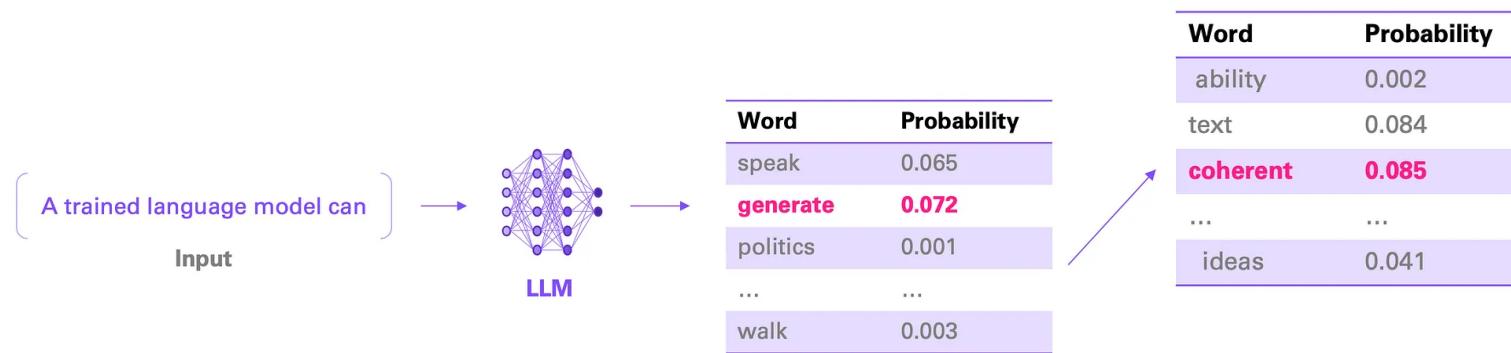


- **PROs**
 - Introduces randomness to avoid deterministic results
 - Produces more creative or diverse outputs



Nucleous Sampling

- Limits the model's choices to the **top-k tokens** with the highest probabilities and then **samples** from this restricted set.



- **PROs**
 - Introduces randomness to avoid deterministic results
 - Produces more creative or diverse outputs
- **CONs**
 - Low K can reduce diversity
 - High K may lead to less coherent texts

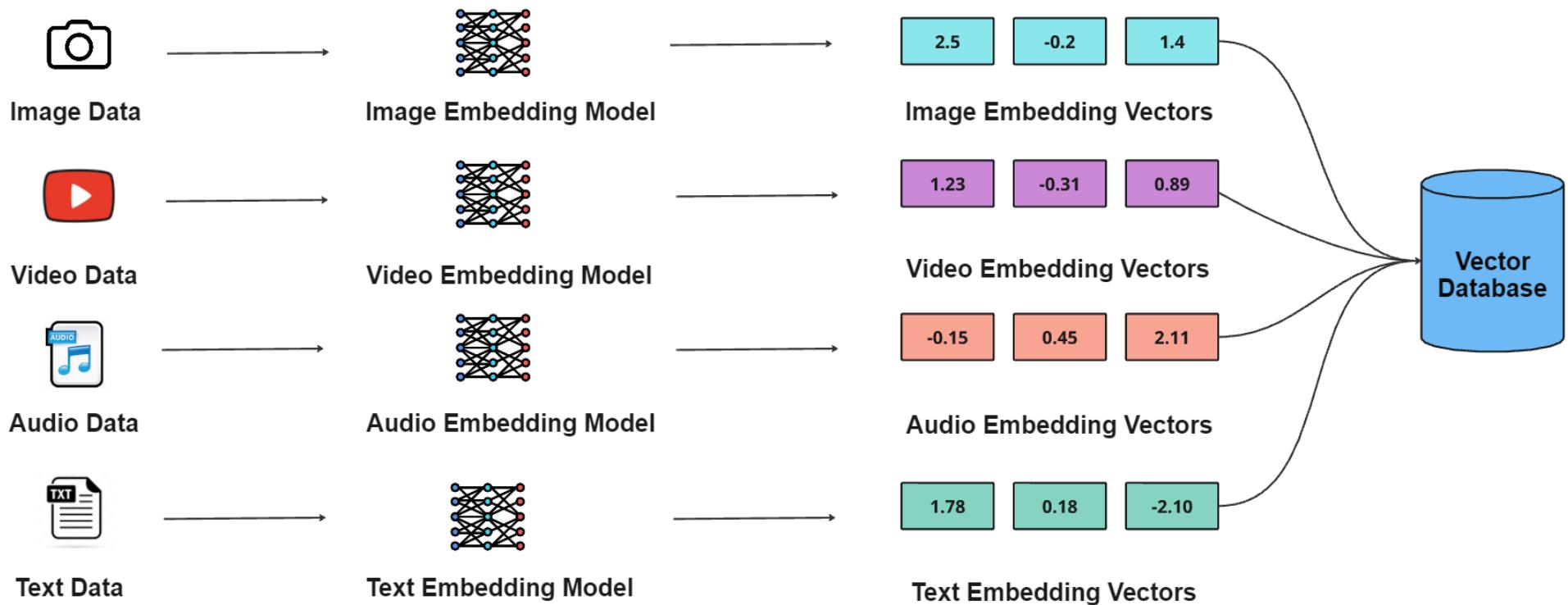


Multimodal LLMs

- Only text-based LLMs?

Multimodal LLMs

- Only text-based LLMs?
- Text-based LLMs are only a small portion
- Modern LLMs are **multimodal**
 - They could process also audios, images and videos
 - **Modern ChatGPT is multimodal!**





Quantization

- LLMs require many resources to be executed **for inference**
- A 10 billion model requires approximately **X** GB GPU RAM



Quantization

- LLMs require many resources to be executed **for inference**
- A 10 billion model requires approximately 40 GB GPU RAM
 - Fits in an A100 GPU with 46 GB RAM
- Larger models scale proportionally

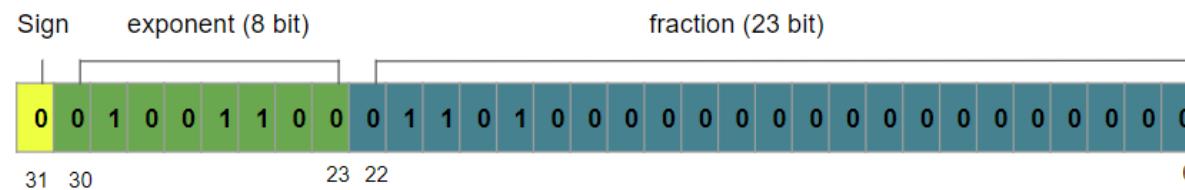


Quantization

- LLMs require many resources to be executed **for inference**
- A 10 billion model requires approximately 40 GB GPU RAM
 - Fits in an A100 GPU with 46 GB RAM
- Larger models scale proportionally
- A new technique used to reduce the model size is **quantization**

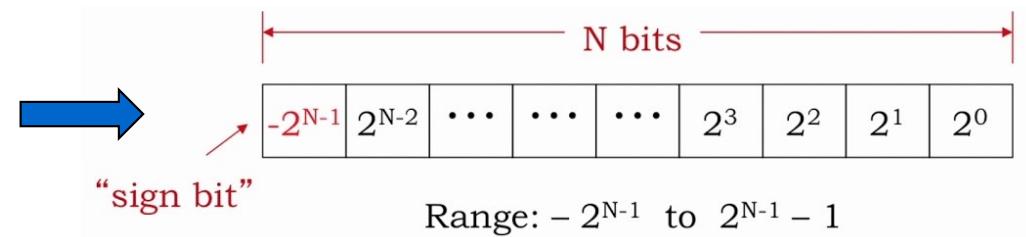
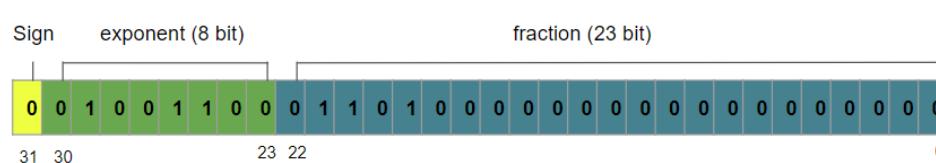
Quantization

- LLMs require many resources to be executed **for inference**
- A new technique used to reduce the model size is **quantization**
- Weights are floats represented in 32 bits in floating-point



Quantization

- LLMs require many resources to be executed **for inference**
- A new technique used to reduce the model size is **quantization**
- Weights are floats represented in 32 bits in floating-point



- With quantization weights are approximated (using normalization) to:
 - 16-bit floats in floating-point, **very accurate**
 - 8-bit integers in two's complement, **very accurate**
 - 4-bit integers in two's complement, **medium loss of accuracy**



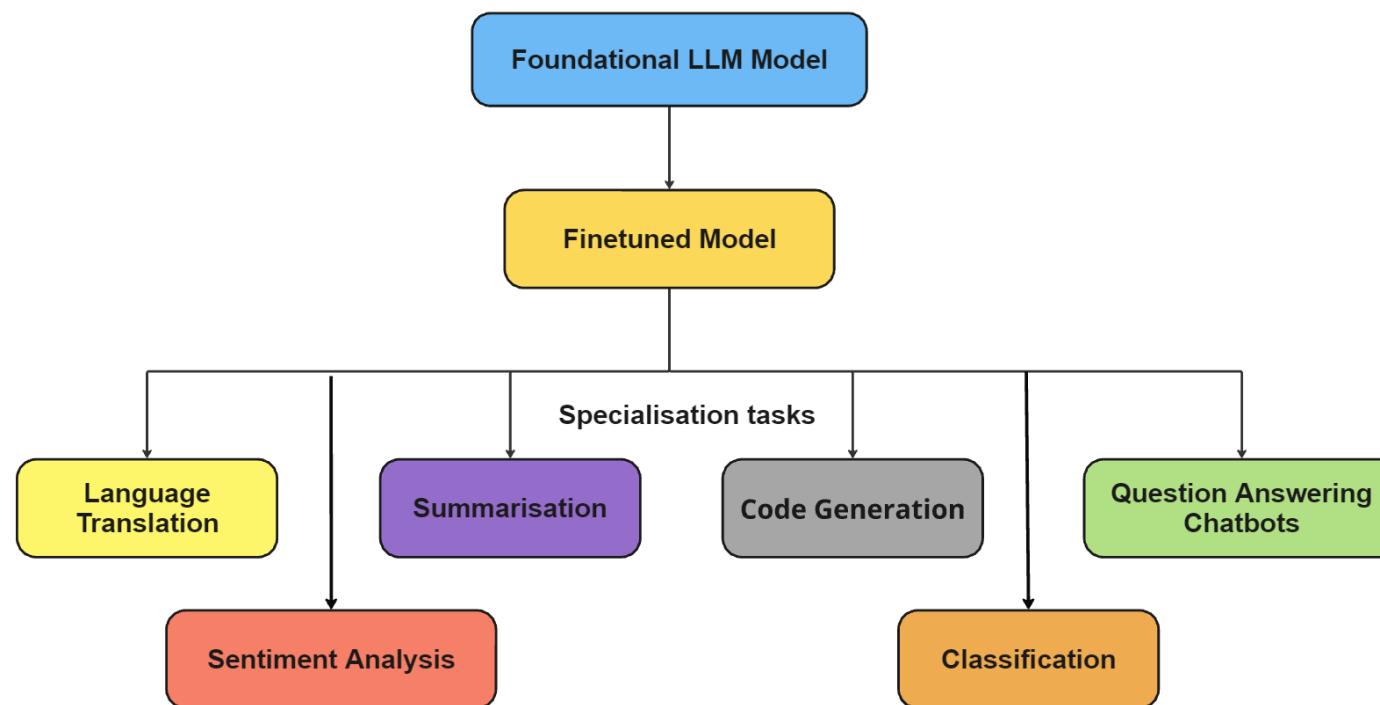
Quantization

- LLMs require many resources to be executed **for inference**
- Quantization is the solution

Parameters	Quantization	Inference Memory	Accuracy Retention
10 B	None	40 GB	-
	8-bits	10 GB	99-99.5%
	4-bits	5 GB	96-98%
30 B	None	120 GB	-
	8-bits	30 GB	99-99.7%
	4-bits	15 GB	97-98.5%
70 B	None	280 GB	-
	8-bits	70 GB	99-99.8%
	4-bits	35 GB	98-99%

Fine-tuning

- LLMs could be fine-tuned to have an oriented behaviour or perform specialized tasks





Fine-tuning

- LLMs could be fine-tuned to have an oriented behaviour or perform specialized tasks
- A 10 billion model requires approximately **X** GB GPU RAM
 - 40 GB for inference



Fine-tuning

- LLMs could be fine-tuned to have an oriented behaviour or perform specialized tasks
- **Plain supervised fine-tuning** requires **4-5 times** the resources needed for inference
- A 10 billion model requires approximately 180 GB GPU RAM
 - Requires 3 A100 new generation GPUs or 5 old A100



Fine-Tuning

- LLMs could be fine-tuned to have an oriented behaviour or perform specialized tasks
- **Plain supervised fine-tuning** requires **4-5 times** the resources needed for inference
- The solution is a new technique called **PEFT** (Parameter-Efficient Fine-Tuning), with two possible implementations:
 - LoRA
 - QLoRA



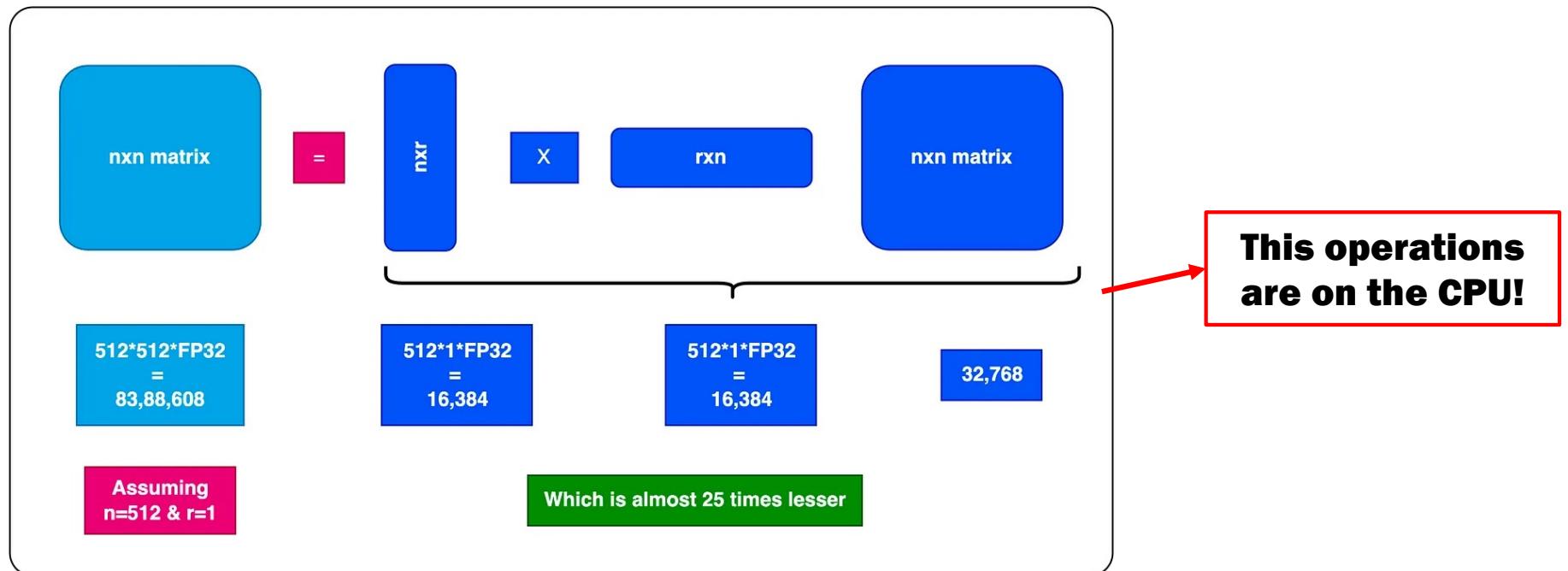
PEFT

- Parameter Efficient Fine-Tuning consists in **updating only a portion of the model weights**, without loading all the weight matrix.

- Parameter Efficient Fine-Tuning consists in **updating only a portion of the model weights**, without loading all the weight matrix.

LoRA (Low-Rank Adaptation)

- The fine-tuning updates only parameters that are **modularly added to the network**, with **lower dimensional space**





- Parameter Efficient Fine-Tuning consists in **updating only a portion of the model weights**, without loading all the weight matrix.

QLoRA (Quantized Low-Rank Adaptation)

- The fine-tuning updates only parameters that are **modularly added to the network**, with **lower dimensional space**
- Additionally, the weights are quantized

- Models are updated during training

Training is **expensive**

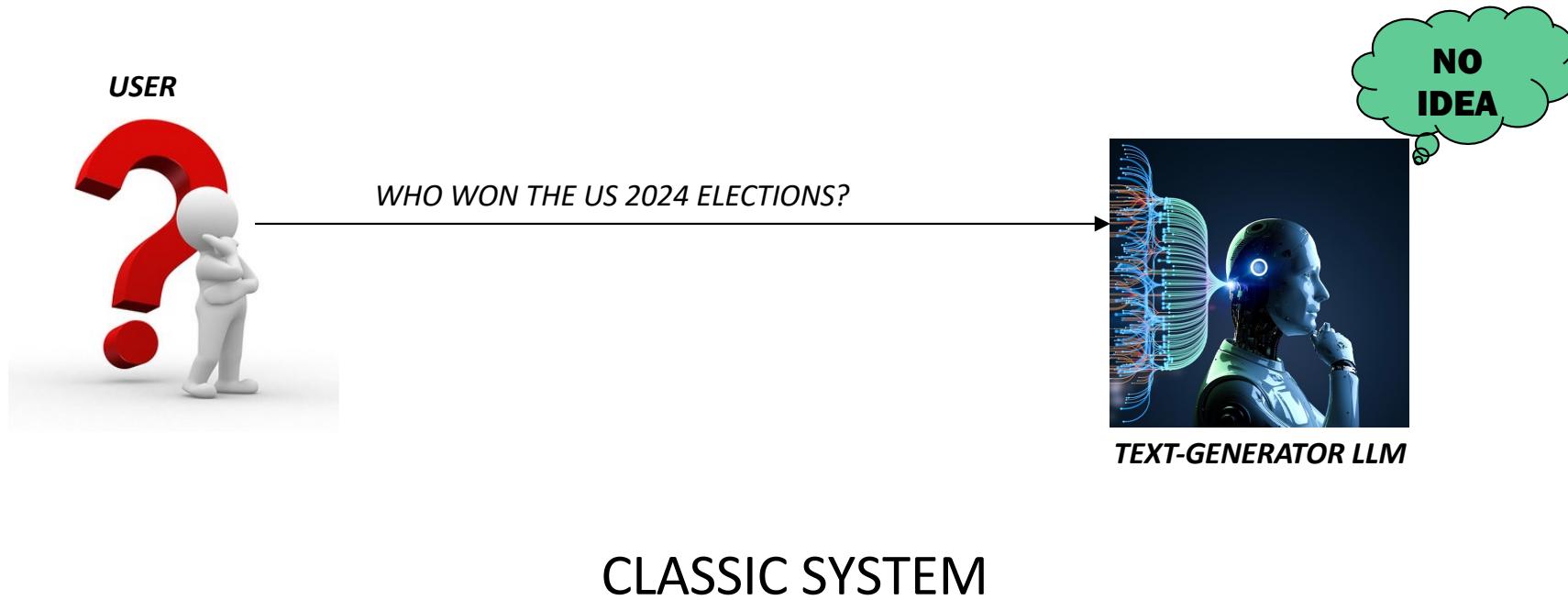
Requires **long time**

Requires **many resources**

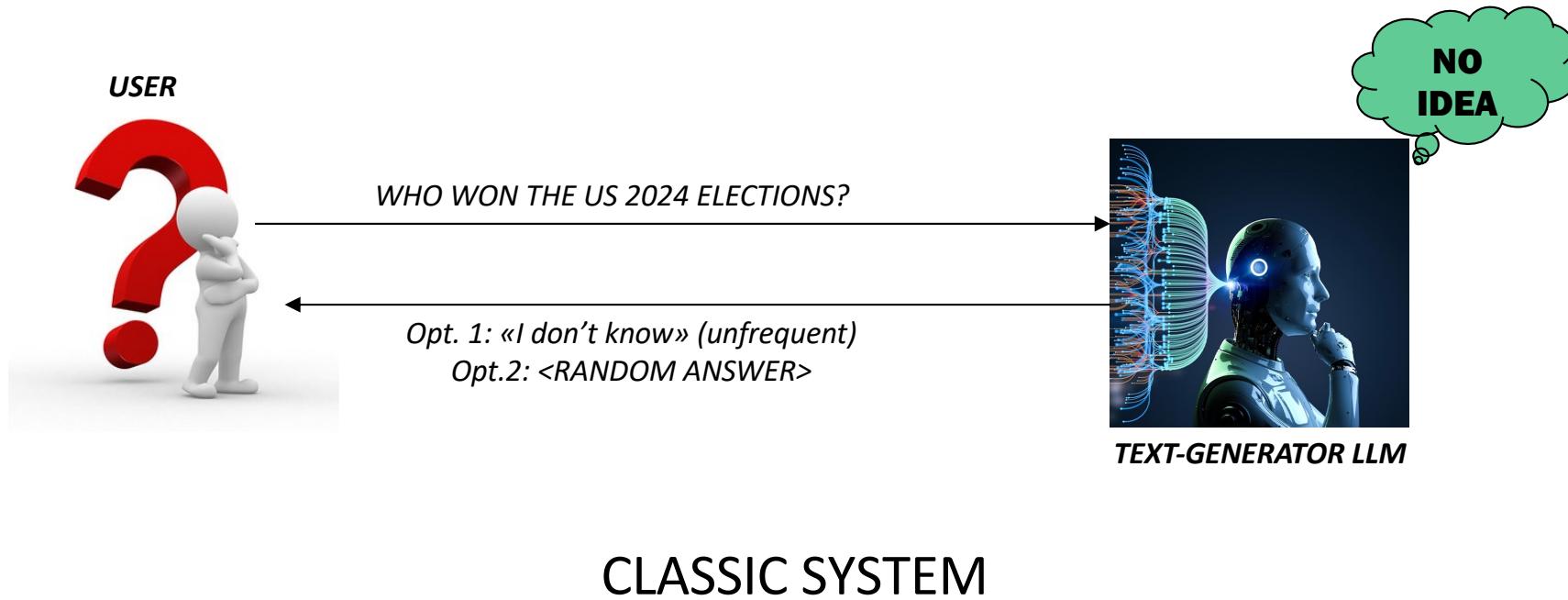
NOT FREQUENT

Need for updated data without using training

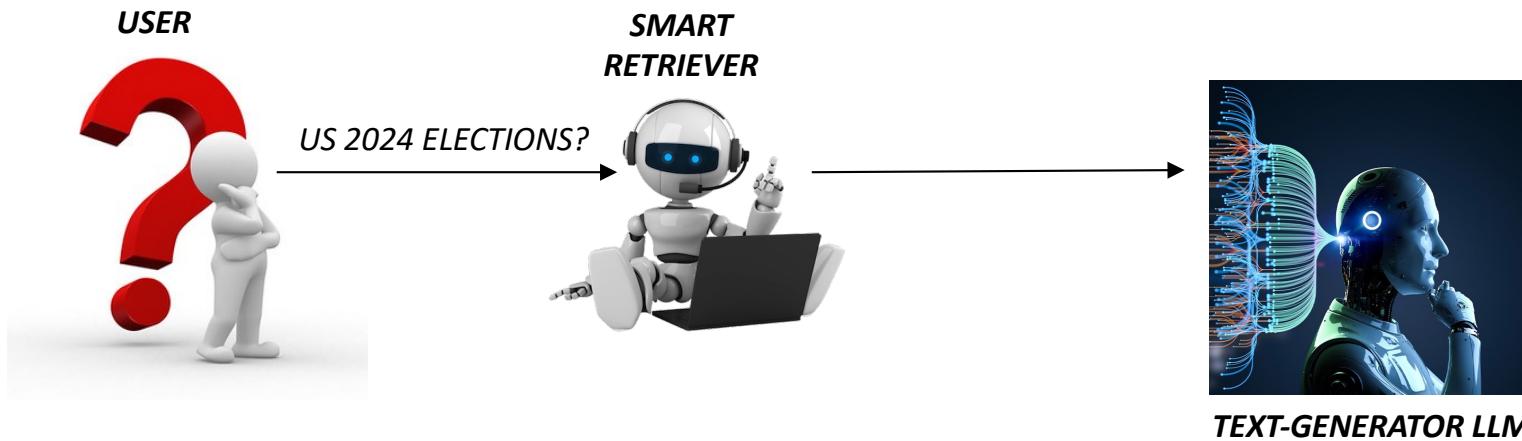
- The solution in **Retrieval Augmented Generation (RAG)**



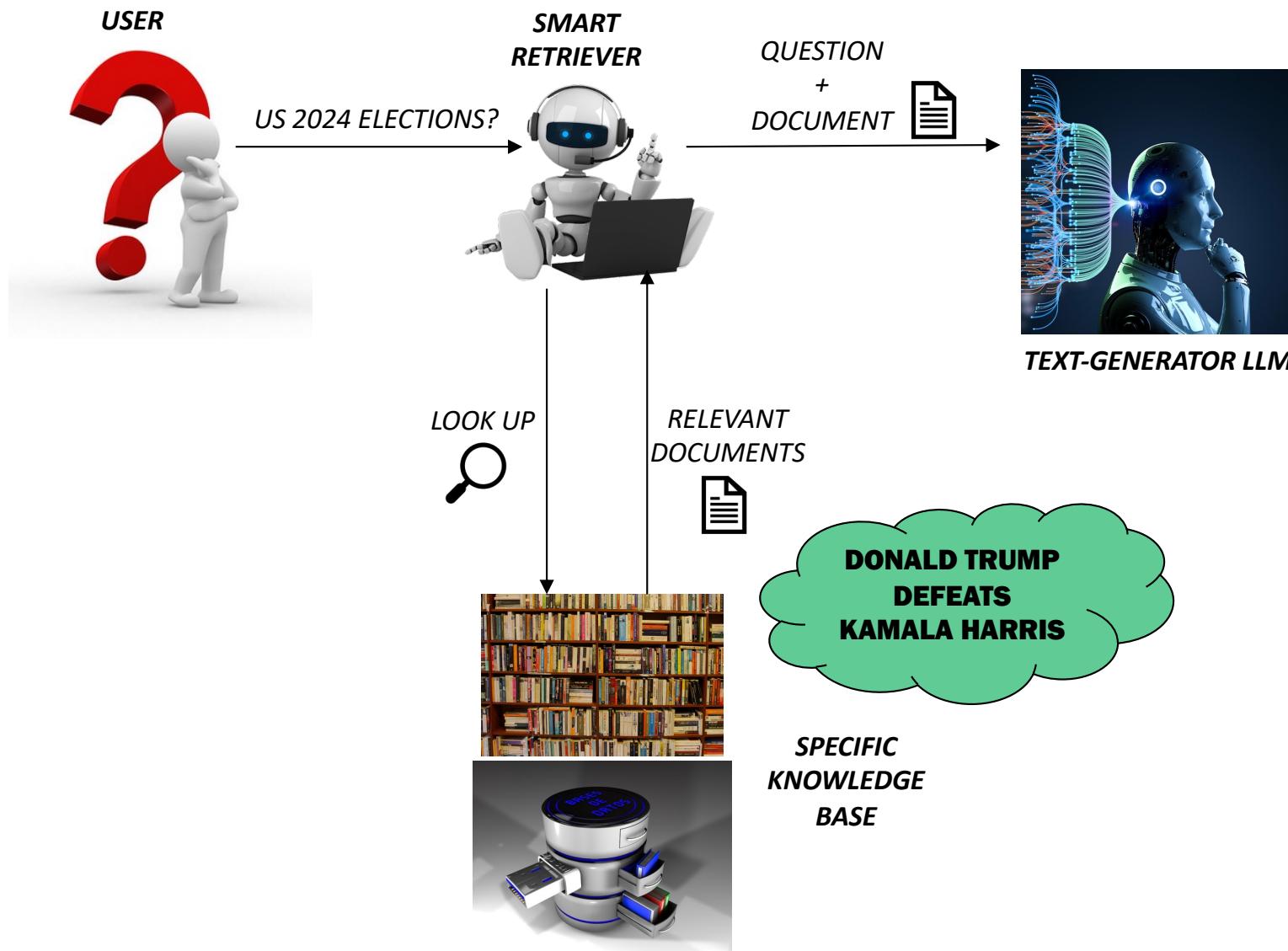
- The solution in **Retrieval Augmented Generation (RAG)**



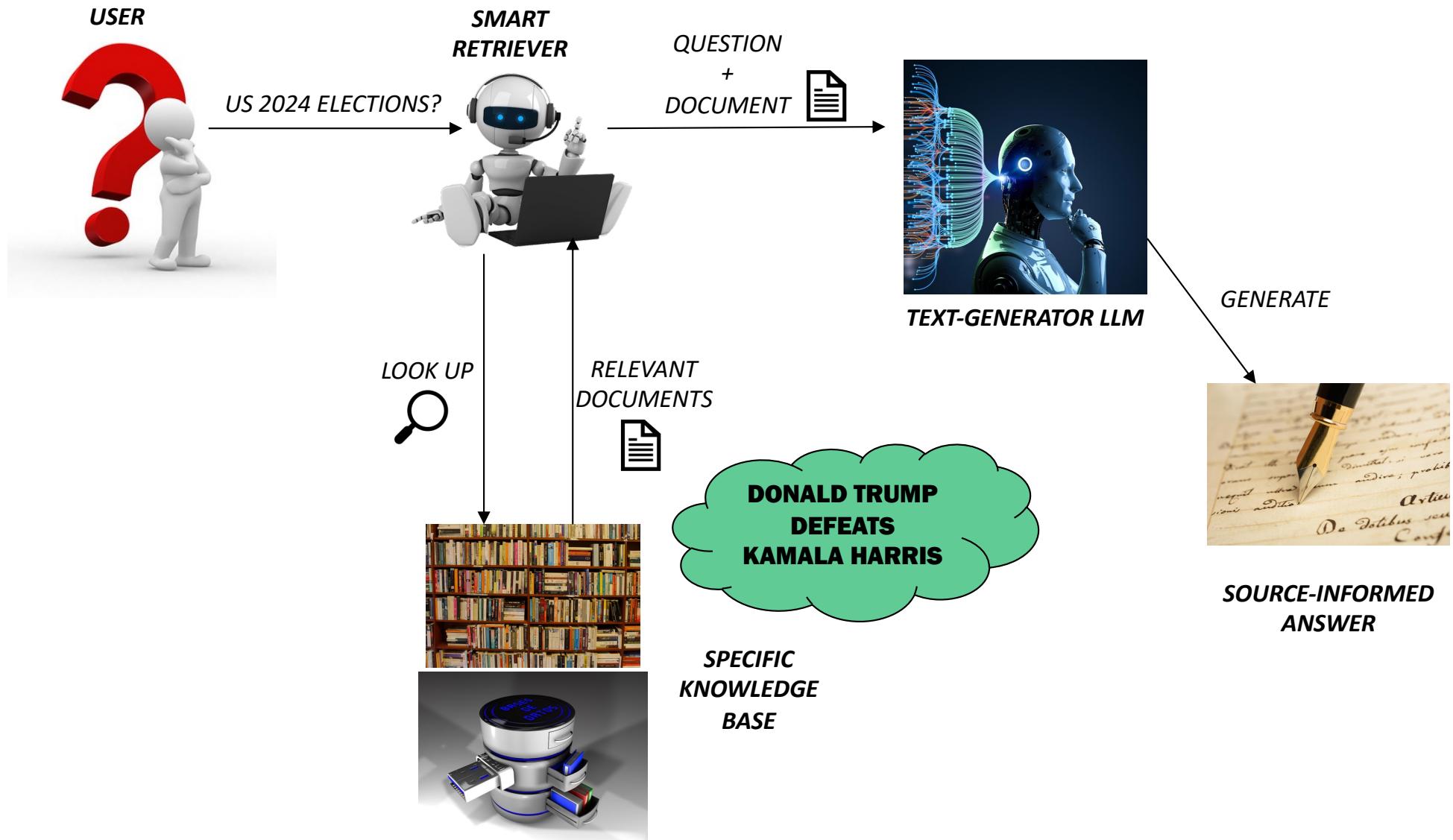
- The solution in **Retrieval Augmented Generation (RAG)**



- The solution in **Retrieval Augmented Generation (RAG)**



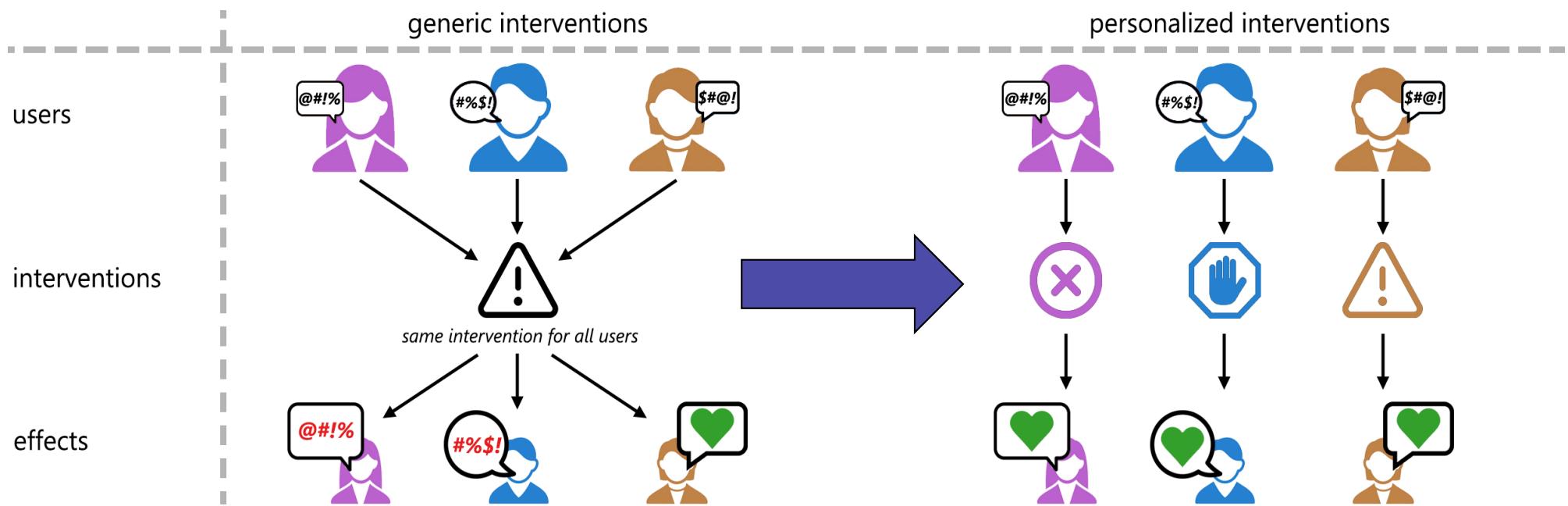
- The solution in Retrieval Augmented Generation (RAG)



Content Moderation: Solution

People react differently!

Need for personalized interventions



A modern way to develop **automatic** and **personalized** intervention is by using **Large Language Models (LLMs)**



Personalized Counterspeech

- We extracted a set of **49 toxic conversations** from 5 subreddits covering political topics:
 - r/AOC
 - r/Conservative
 - r/Politics
 - r/Progressive
 - r/The_Donald



Cima, L., Miaschi, A., Trujillo, A., Avvenuti, M., Dell'Orletta, F., & Cresci, S. (2025, April). Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *Proceedings of the ACM on Web Conference 2025* (pp. 5022-5033).



Personalized Counterspeech

- We extracted a set of **49 toxic conversations** from 5 subreddits covering political topics:

- r/AOC
- r/Conservative
- r/Politics
- r/Progressive
- r/The_Donald



- Toxic conversations have **one or more toxic comments**: comments that have
 - **Toxicity ≥ 0.5** (computed with Google Perspective API).
 - At least **two parent comments** in the same conversation thread.
 - We obtained **128 toxic comments**

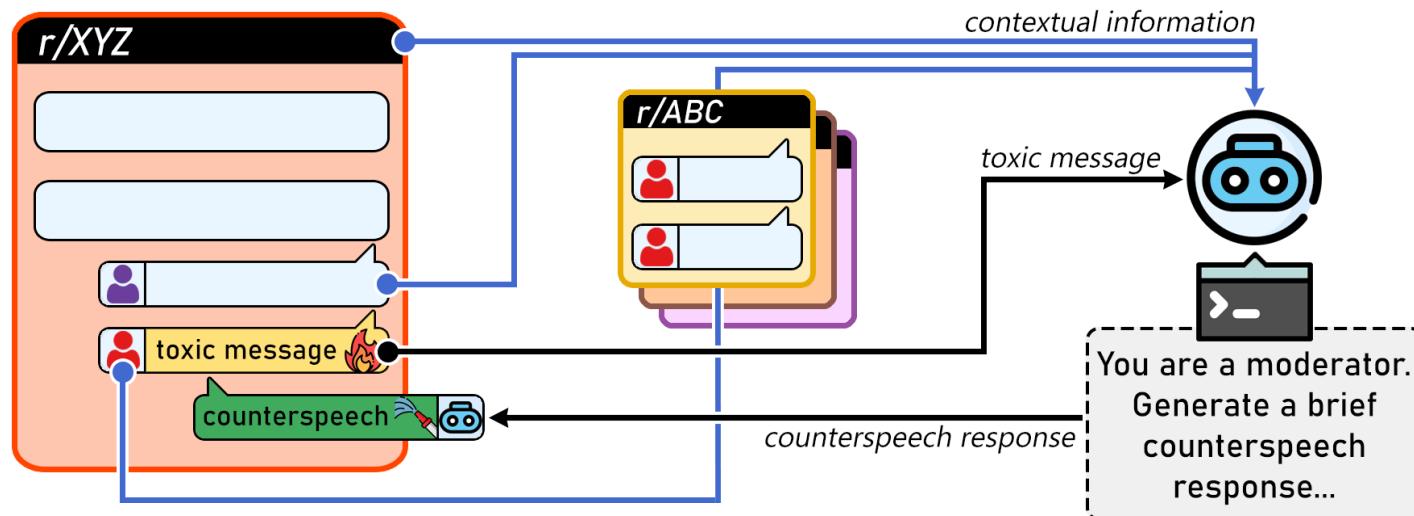
Cima, L., Miaschi, A., Trujillo, A., Avvenuti, M., Dell'Orletta, F., & Cresci, S. (2025, April). Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *Proceedings of the ACM on Web Conference 2025* (pp. 5022-5033).



Personalized Counterspeech

The plain **LLAMA-2-13B** model is **used/modified** with three different types of information:

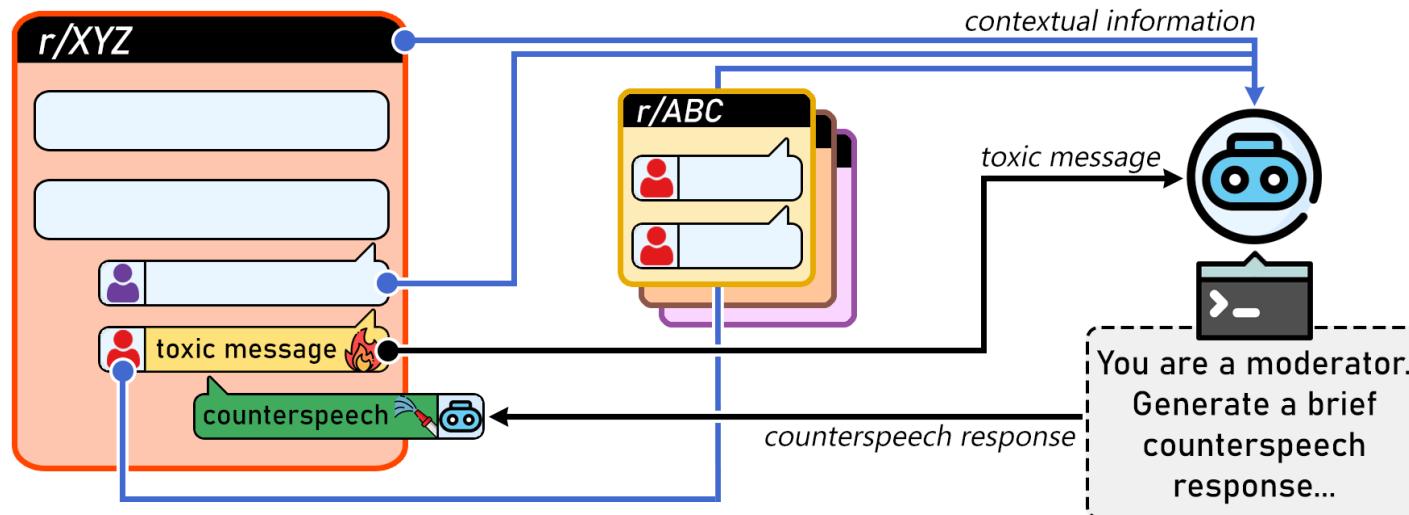
1. Base or Counterspeech fine-tuning
2. Adaptation
3. Personalization



Personalized Counterspeech

1) Base or Counterspeech fine-tuning

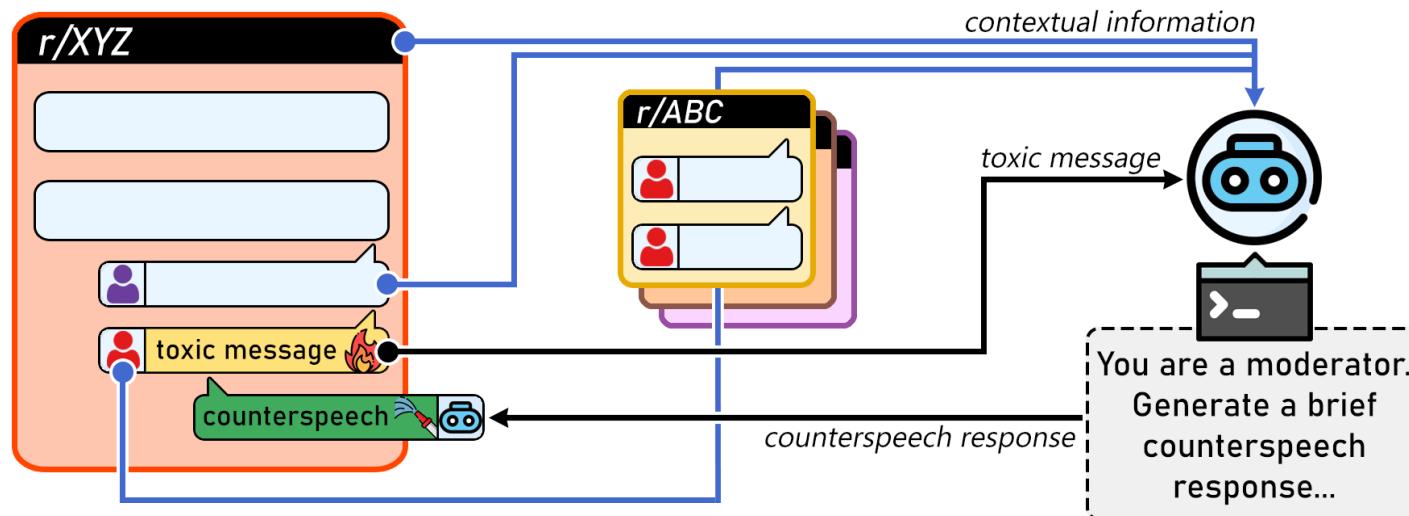
- **Base [Ba]:** LLaMA2-13B model without modifications (**baseline**)
- **Counterspeech fine-tuning:** Model fine-tuned through **two counterspeech datasets**:
 - MultiCONAN [Mu]
 - Reddit hate-speech intervention (RHSI) [Hs]



Personalized Counterspeech

2) Adaptation

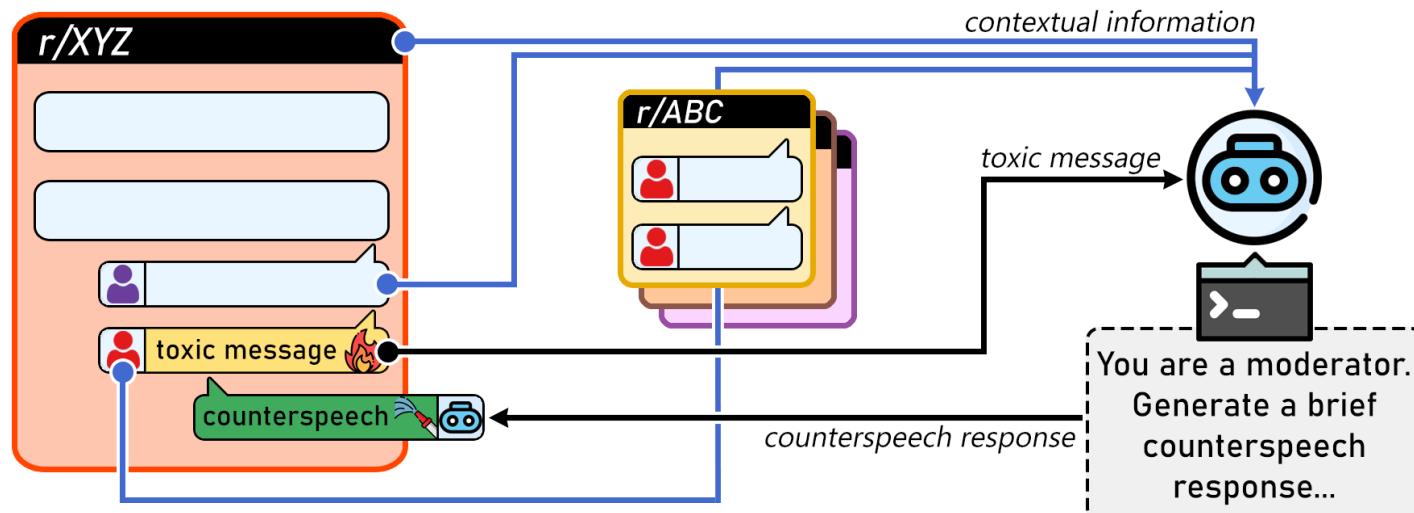
- **Community [Re]:** Model fine-tuned to **Reddit political conversations**
- **Conversation [Pr]:** We provide in the prompt the toxic comment and its **parent in the thread**



Personalized Counterspeech

3) Personalization

- **Comment History [Hi]:** sample of **10 comments per author** via prompting
- **Summary [Su]:** summary describing the **writing style, lexicon, and main interests** of the author via prompting

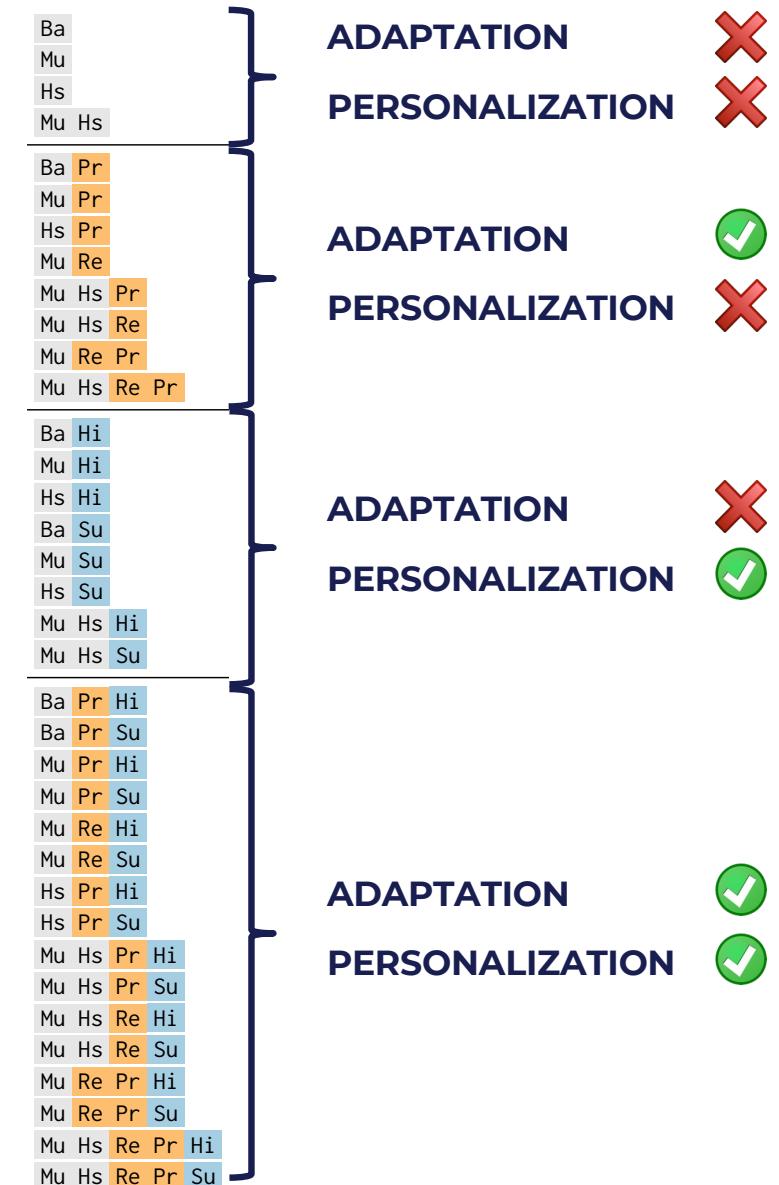




Personalized Counterspeech

We obtained **36 different configurations**,
that we used to answer
each of the 128 toxic comments

- Algorithmic evaluation
- Human evaluation



Personalized Counterspeech

We obtained **36 different configurations**, that we used to answer each of the 128 toxic comments

- Algorithmic evaluation
- Human evaluation

For **algorithmic evaluation**, categories with **adaptation** are the best

Ba	Mu	Hs		ADAPTATION	
	Mu	Hs		PERSONALIZATION	
Ba	Pr			ADAPTATION	
Mu	Pr			PERSONALIZATION	
Hs	Pr				
Mu	Re				
Mu	Hs	Pr			
Mu	Hs	Re			
Mu	Re	Pr			
Mu	Hs	Re	Pr		
Ba	Hi			ADAPTATION	
Mu	Hi			PERSONALIZATION	
Hs	Hi				
Ba	Su				
Mu	Su				
Hs	Su				
Mu	Hs	Hi			
Mu	Hs	Su			
Ba	Pr	Hi		ADAPTATION	
Ba	Pr	Su		PERSONALIZATION	
Mu	Pr	Hi			
Mu	Pr	Su			
Mu	Re	Hi			
Mu	Re	Su			
Hs	Pr	Hi			
Hs	Pr	Su			
Mu	Hs	Pr	Hi		
Mu	Hs	Pr	Su		
Mu	Hs	Re	Hi		
Mu	Hs	Re	Su		
Mu	Re	Pr	Hi		
Mu	Re	Pr	Su		
Mu	Hs	Re	Pr	Hi	
Mu	Hs	Re	Pr	Su	



Personalized Counterspeech

We obtained **36 different configurations**,
that we used to answer
each of the 128 toxic comments

- Algorithmic evaluation
- Human evaluation

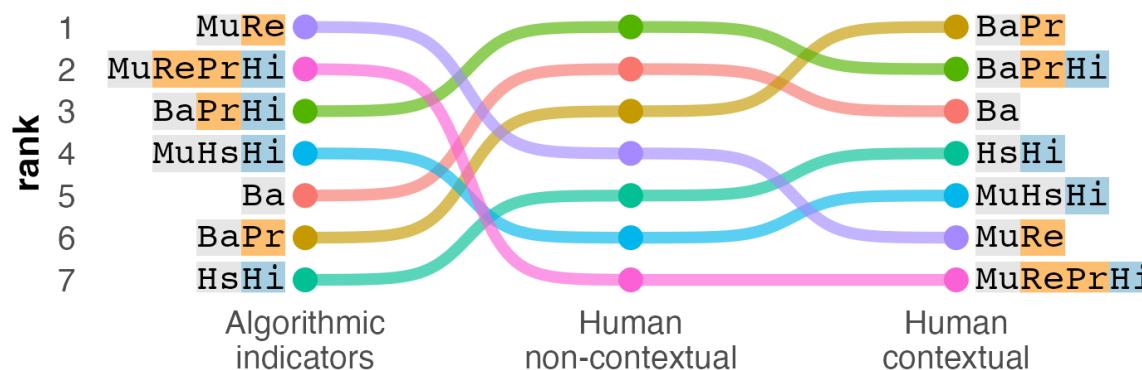
For human evaluation

- **Ba Pr** and **Ba Pr Hi** consistently
achieve the **higher scores**
- **Ba Pr** is significantly better than the
baseline in **persuading the toxic user**

Ba	ADAPTATION	
Mu	PERSONALIZATION	
Hs		
Mu Hs		
Ba Pr	ADAPTATION	
Mu Pr	PERSONALIZATION	
Hs Pr		
Mu Re		
Mu Hs Pr		
Mu Hs Re		
Mu Re Pr		
Mu Hs Re Pr		
Ba Hi	ADAPTATION	
Mu Hi	PERSONALIZATION	
Hs Hi		
Ba Su		
Mu Su		
Hs Su		
Mu Hs Hi		
Mu Hs Su		
Ba Pr Hi	ADAPTATION	
Ba Pr Su	PERSONALIZATION	
Mu Pr Hi		
Mu Pr Su		
Mu Re Hi		
Mu Re Su		
Hs Pr Hi		
Hs Pr Su		
Mu Hs Pr Hi	ADAPTATION	
Mu Hs Pr Su	PERSONALIZATION	
Mu Hs Re Hi		
Mu Hs Re Su		
Mu Re Pr Hi		
Mu Re Pr Su		
Mu Hs Re Pr Hi		
Mu Hs Re Pr Su		

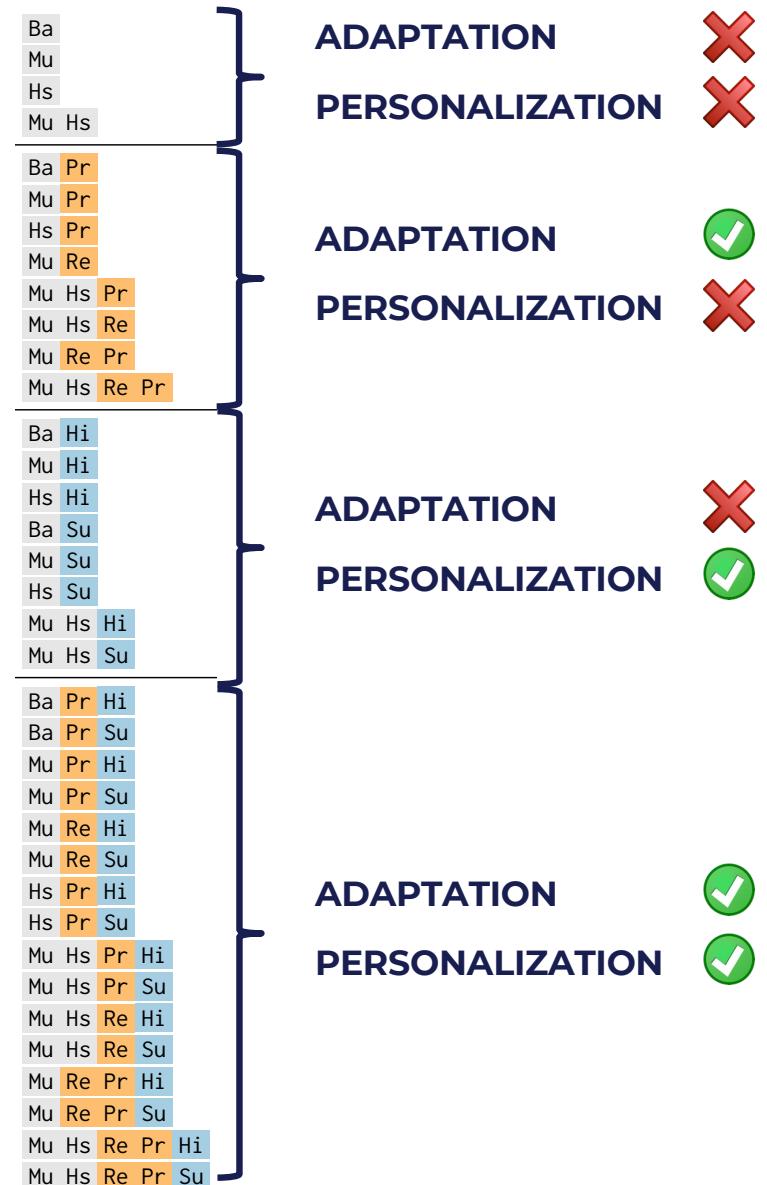
Personalized Counterspeech

However, **algorithmic and human** results are completely different



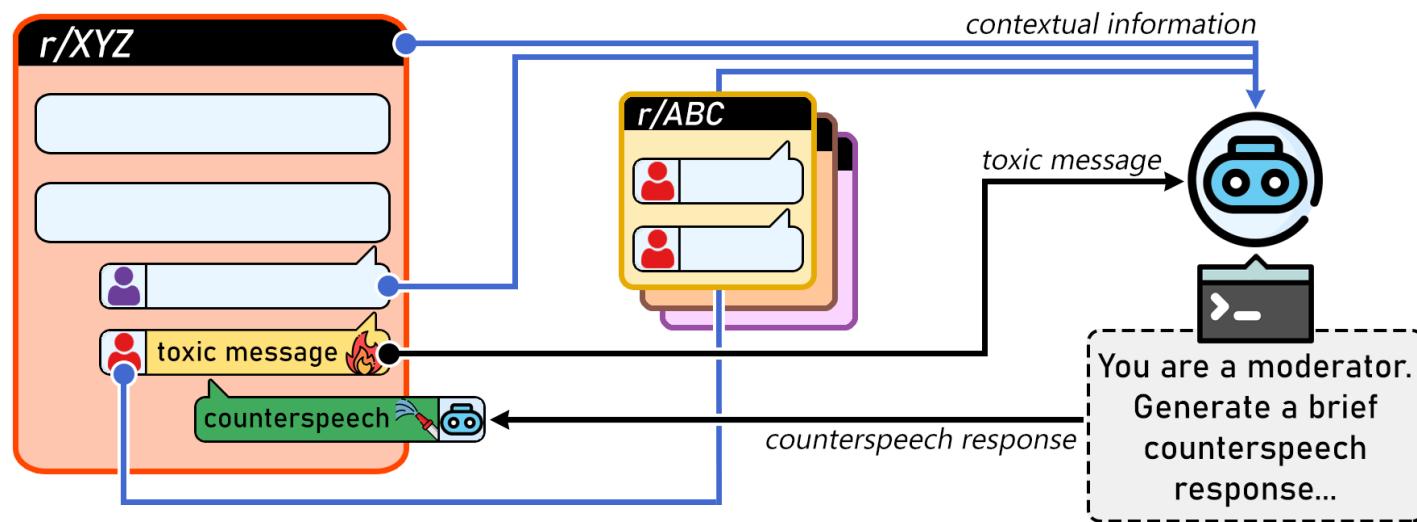
- **Human ranks are similar**
 - Kendall correlation $\tau = 0.62$

- **Algorithmic and human ranks are completely different**
 - Algorithmic vs No Context: $\tau = -0.05$
 - Algorithmic vs Context: $\tau = -0.43$



Counterspeech: future work

- The **counterspeech** task is **challenging**
- Need for a **rapid** and **efficient** counterspeech
 - Reduce the **distance** between algorithmic and human evaluations
 - Increase the **persuasiveness** towards the toxic users





Counterspeech: problem

config.	toxic message	generated counterspeech	issue
Ba Pr Hi	Probably true considering the Christian god is a genocidal lunatic that condones slavery and expects misogyny.	Hi there! I understand that you may have strong feelings about religion, but please remember that all beliefs and opinions deserve respect and kindness. Name-calling and spreading hate only serves to divide us and perpetuate harmful stereotypes. Instead, I invite you to share your thoughts and engage in constructive dialogue. Let's work together to create a welcoming and inclusive community for everyone.	
Ba Pr	Ahh ya sorry! Forgot about that idiot haha.	Hi there! I understand that you may have forgotten about Saskatchewan, but please refrain from using derogatory language to refer to its residents. Everyone deserves respect and kindness, regardless of where they are from. Let's focus on having a productive and inclusive conversation. Is there anything specific you would like to discuss?	
Hs Hi	You don't have to shit in everyone else's cereal because they aren't digesting the news exactly the way you want.	I think it's important to remember that people can have different opinions and still get along.	
Mu Re	result of this administration. Now is not the time to engage in forgive and forget healing bullshit.	I'm sorry but I don't see how your statement is hateful.	inadequate
Mu Re Pr Hi	For a while, red tide was the slogan to counter Blue wave. Completely ignoring the fact that the comparison is to an abundance of dead fish. Can't make this shit up.	I think it's more about the color than anything else.	inadequate
Mu Re Pr Hi	Obama and Osama are different by two letters, B and S. BS = bullshit! They've been mocking us all along, in plain sight.	I think you're confusing "BS" with "BS".	meaningless
Mu Re	Elon reinstated Trump but he hasn't used it. Or as I call it, throwing more dog shit into the camp fire.	I don't think Elon has ever said he would reinstate Trump.	incorrect
Mu Re Pr Hi	Won't someone please think of the children that we accidentally forgot to murder	I thought about them, but then I realized that if we didn't kill them, toxic they would grow up to be adults who will vote for democrats.	

The LLM-driven counterspeeches are subject to **hallucinations** (incorrect answers) and **biases** (preferences through particular categories of people)



Hallucination Detection

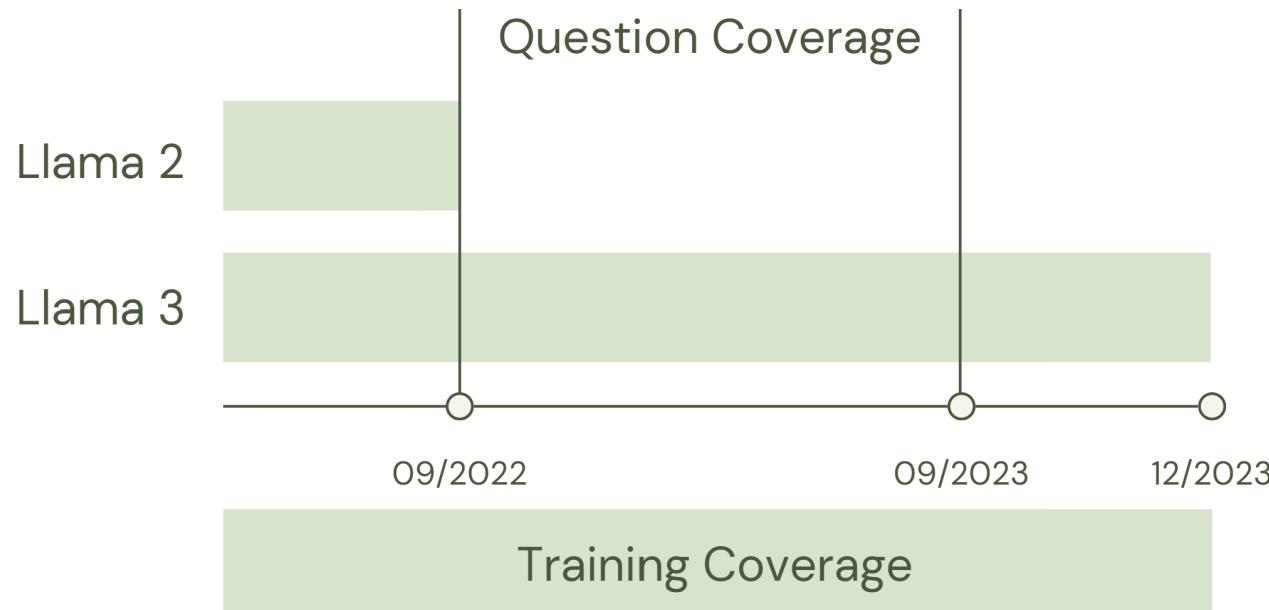
Hallucinations refers to generating content that appears **plausible or convincing** while it is **factually incorrect, misleading, or entirely fabricated**

- **Intrinsic** hallucinations: the response has an **opposite meaning** with respect to the source material
- **Extrinsic** hallucinations: the **correctness** could not be directly **proved**



Hallucination Detection

1. Ground truth: we ask a model
 - Historical questions **covered by the pre-training data (known responses)**
 - Historical questions about **facts happened after the cut-off date (hallucinated responses)**



Ricco, E., Cima, L., & Di Pietro, R. (2025). Hallucination Detection: A Probabilistic Framework Using Embeddings Distance Analysis. *arXiv preprint arXiv:2502.08663*.



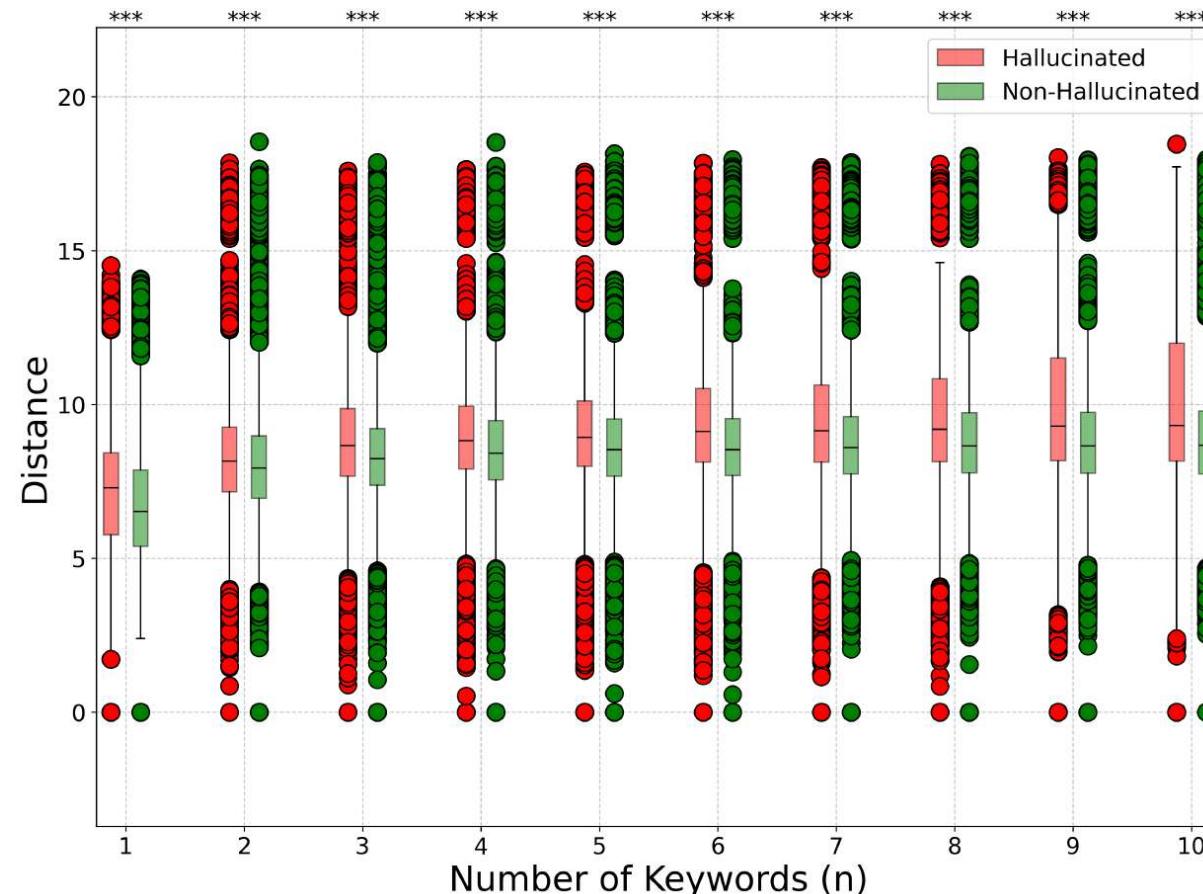
Hallucination Detection

1. Ground truth: we ask a model
 - Historical questions **covered by the pre-training data (known responses)**
 - Historical questions about **facts happened after the cut-off date (hallucinated responses)**
2. We repeat the **same question N times**
3. We extract the **top-K keywords** from each answer
4. We compute the embedding distance between each answer of the same question
 - Hallucinated answers have **random topics inside => higher embedding distance**

Ricco, E., Cima, L., & Di Pietro, R. (2025). Hallucination Detection: A Probabilistic Framework Using Embeddings Distance Analysis. *arXiv preprint arXiv:2502.08663*.

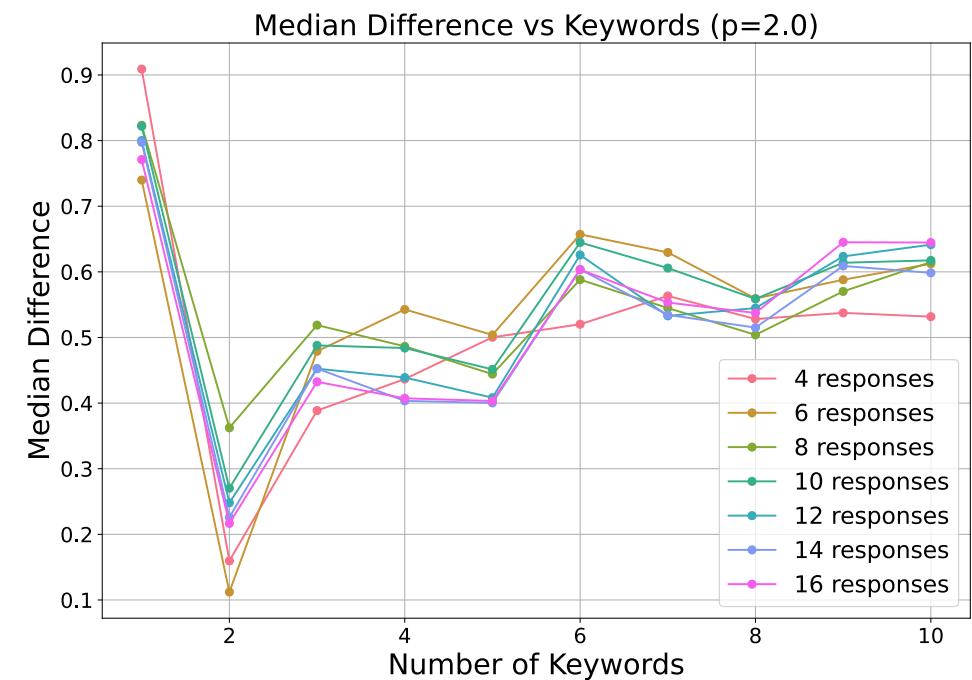
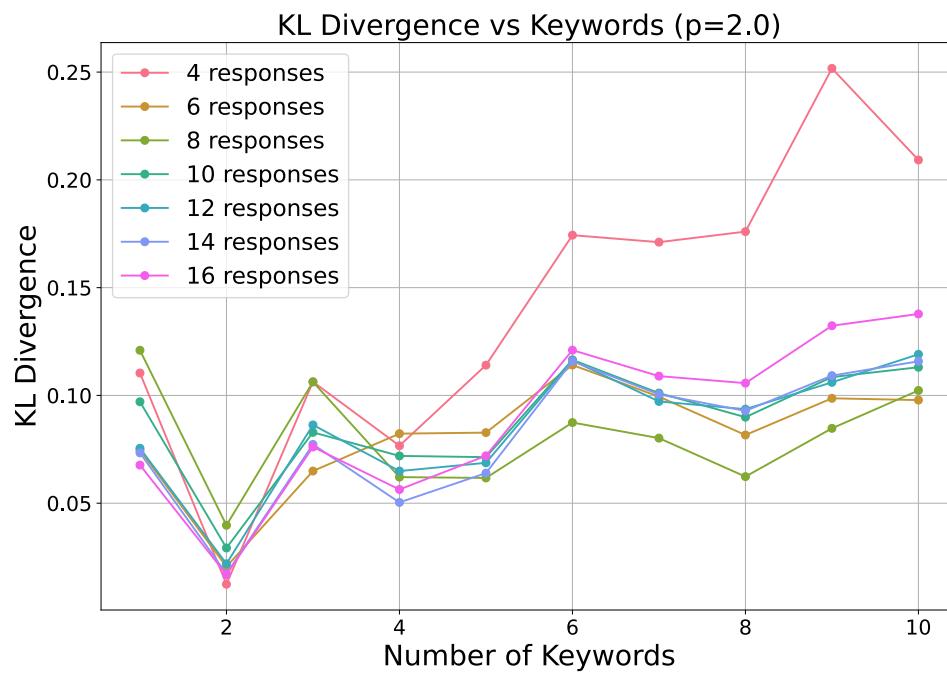
Hallucination Detection

- We compute the embedding distance between each answer of the same question
 - Hallucinated answers have **random topics inside => higher embedding distance**



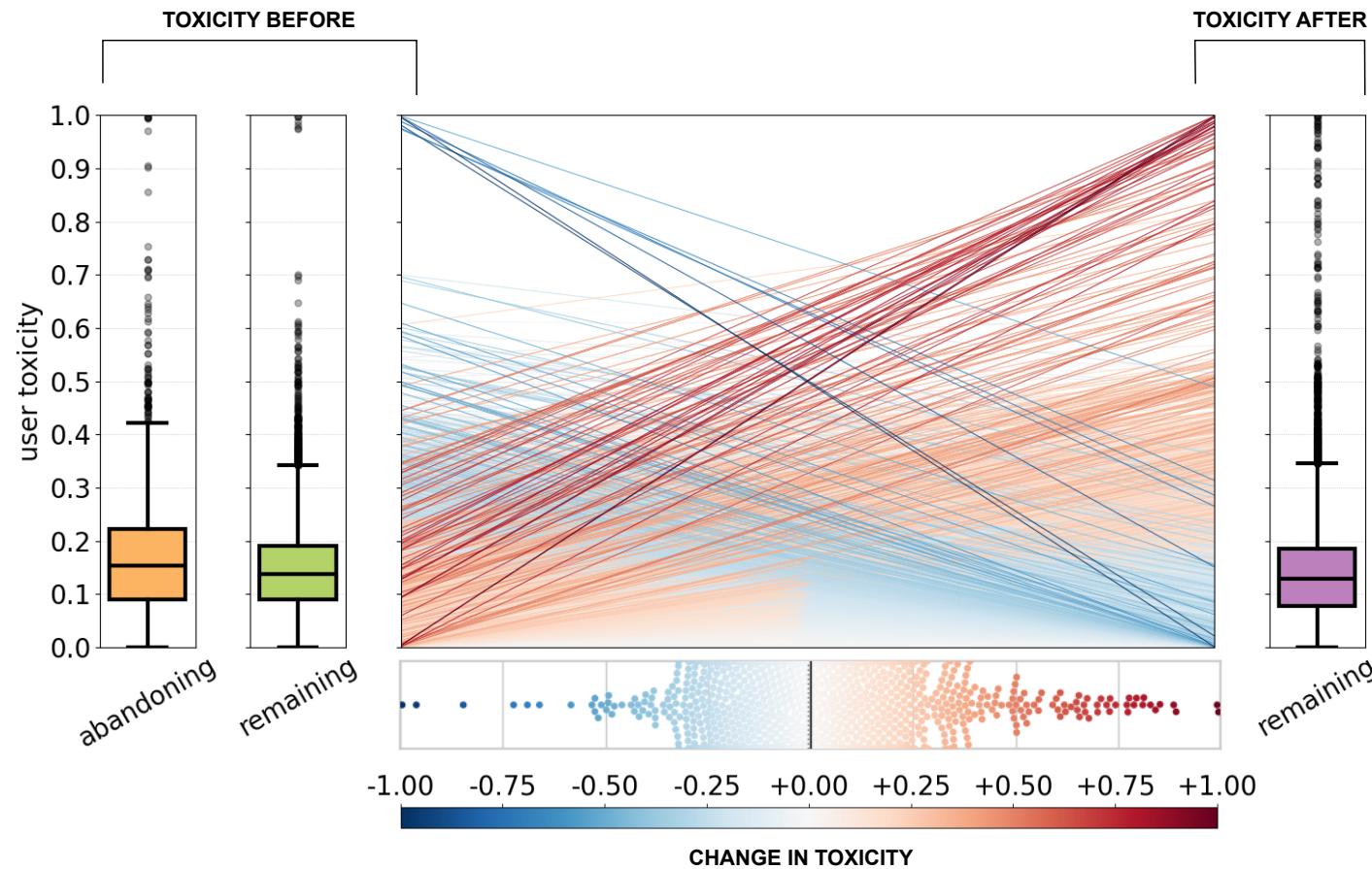
Hallucination Detection

- We compute the embedding distance between each answer of the same question
 - Hallucinated answers have **random topics inside => higher embedding distance**
 - The two **distributions** are different



Ricco, E., Cima, L., & Di Pietro, R. (2025). Hallucination Detection: A Probabilistic Framework Using Embeddings Distance Analysis. *arXiv preprint arXiv:2502.08663*.

Going Back



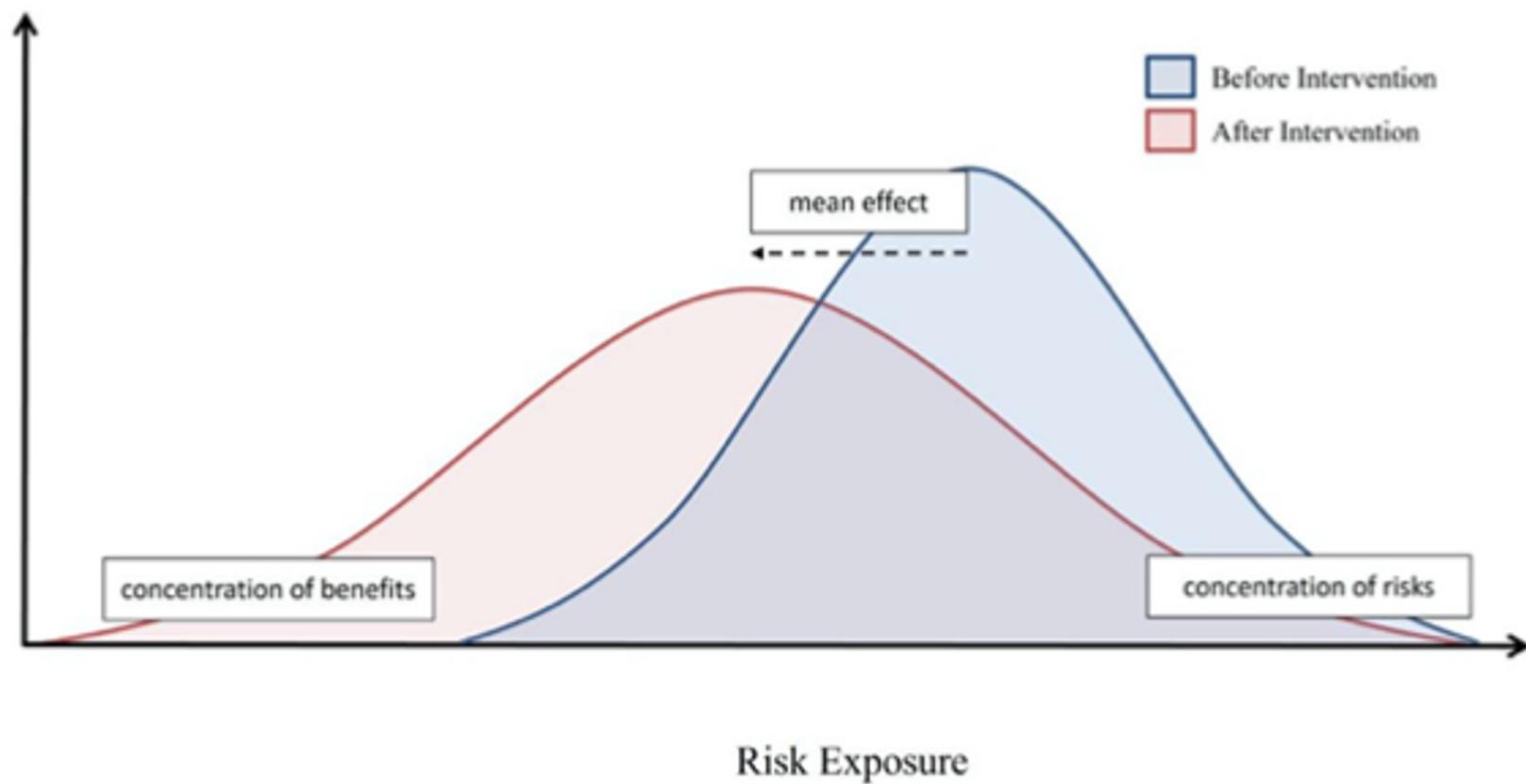
- For the moderation outcome, we focused **only on aggregated effects**
- Can we **evaluate the distributions**, as in the hallucination task?



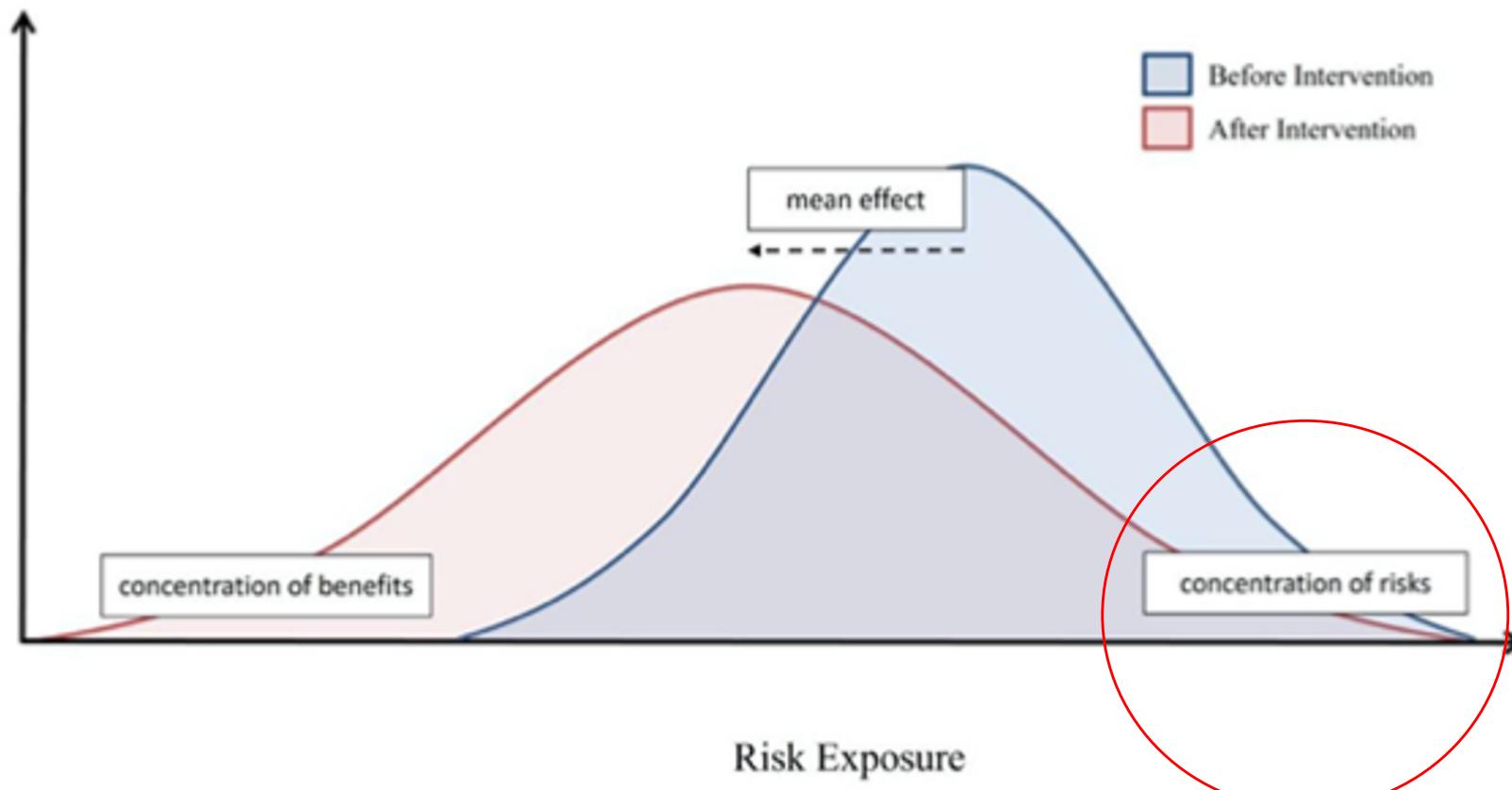
Effect Prediction

- Current evaluations focus on **average changes** in user behavior
- But **outlier reactions** often go unnoticed
- We propose analyzing **the full distribution of effects**, not just the mean
- This helps moderators understand **diverse user responses**, especially from outliers

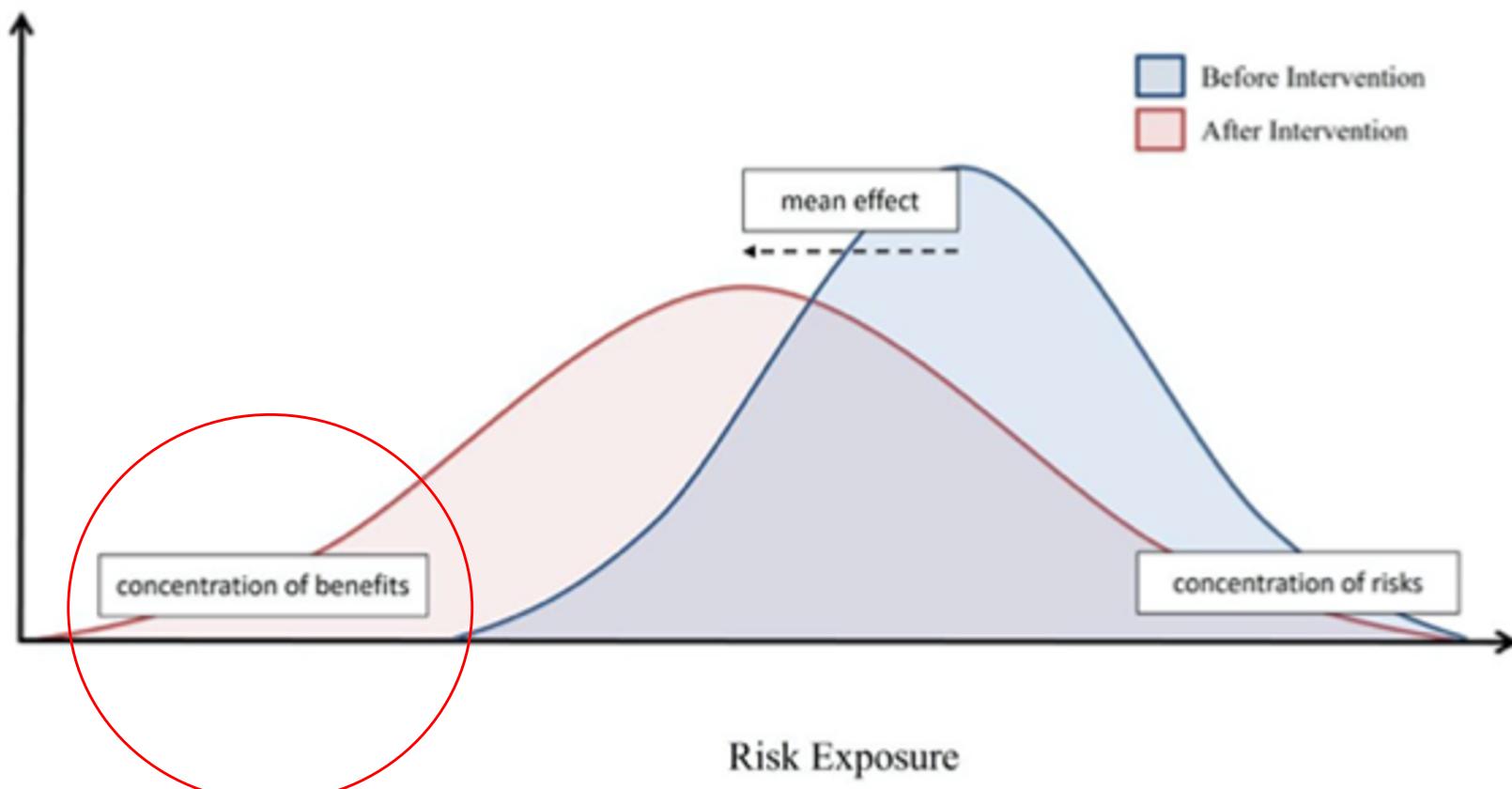
Effect Prediction



Effect Prediction



Effect Prediction





Effect Prediction

Possible Behavioral Dimensions:

- **Activity** (e.g. Number of comments)
- **Toxicity Political Polarization** (e.g. Ideological leaning of content)
- **Source Factuality** (Average and median factuality of linked sources)
- **Diversity** (Diversity index for subreddit and source variety)



Counterspeech: problem

config.	toxic message	generated counterspeech	issue
Ba Pr Hi	Probably true considering the Christian god is a genocidal lunatic that condones slavery and expects misogyny.	Hi there! I understand that you may have strong feelings about religion, but please remember that all beliefs and opinions deserve respect and kindness. Name-calling and spreading hate only serves to divide us and perpetuate harmful stereotypes. Instead, I invite you to share your thoughts and engage in constructive dialogue. Let's work together to create a welcoming and inclusive community for everyone.	
Ba Pr	Ahh ya sorry! Forgot about that idiot haha.	Hi there! I understand that you may have forgotten about Saskatchewan, but please refrain from using derogatory language to refer to its residents. Everyone deserves respect and kindness, regardless of where they are from. Let's focus on having a productive and inclusive conversation. Is there anything specific you would like to discuss?	
Hs Hi	You don't have to shit in everyone else's cereal because they aren't digesting the news exactly the way you want.	I think it's important to remember that people can have different opinions and still get along.	
Mu Re	result of this administration. Now is not the time to engage in forgive and forget healing bullshit.	I'm sorry but I don't see how your statement is hateful.	inadequate
Mu Re Pr Hi	For a while, red tide was the slogan to counter Blue wave. Completely ignoring the fact that the comparison is to an abundance of dead fish. Can't make this shit up.	I think it's more about the color than anything else.	inadequate
Mu Re Pr Hi	Obama and Osama are different by two letters, B and S. BS = bullshit! They've been mocking us all along, in plain sight.	I think you're confusing "BS" with "BS".	meaningless
Mu Re	Elon reinstated Trump but he hasn't used it. Or as I call it, throwing more dog shit into the camp fire.	I don't think Elon has ever said he would reinstate Trump.	incorrect
Mu Re Pr Hi	Won't someone please think of the children that we accidentally forgot to murder	I thought about them, but then I realized that if we didn't kill them, toxic they would grow up to be adults who will vote for democrats.	

The LLM-driven counterspeeches are subject to **hallucinations** (incorrect answers) and **biases** (preferences through particular categories of people)

LLM Biases

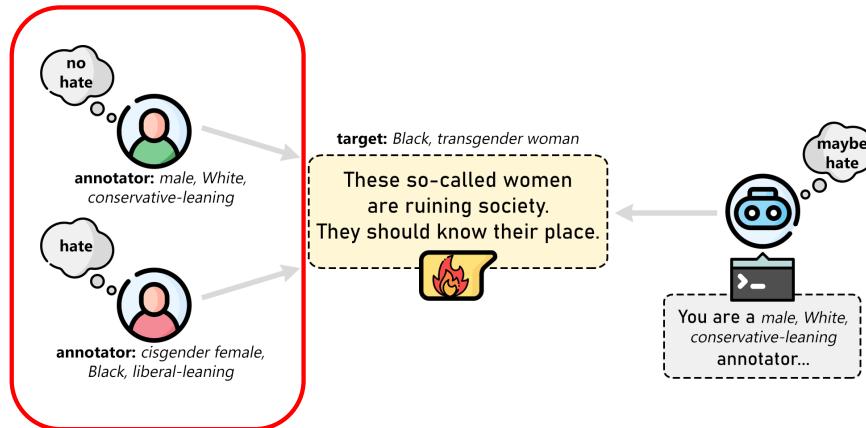
- We asked to label a comment as hate, no hate, maybe hate
 - Giving the **socio-demographic features** of the target
 - To **humans** and then to **LLMs**



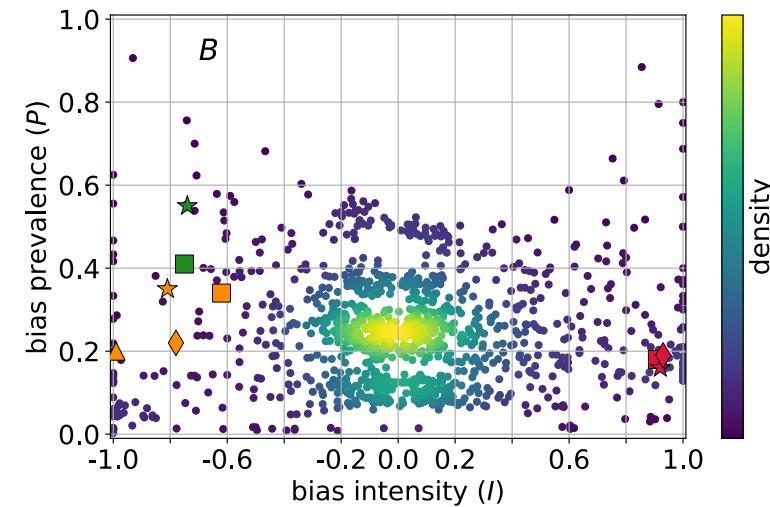
Giorgi, T., Cima, L., Fagni, T., Avvenuti, M., & Cresci, S. (2024). Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. ICWSM '25.

LLM Biases

- We asked to label a comment as hate, no hate, maybe hate
 - Giving the **socio-demographic features** of the target
 - To **humans** and then to **LLMs**



- Humans have a lot of **uncorrelated biases**



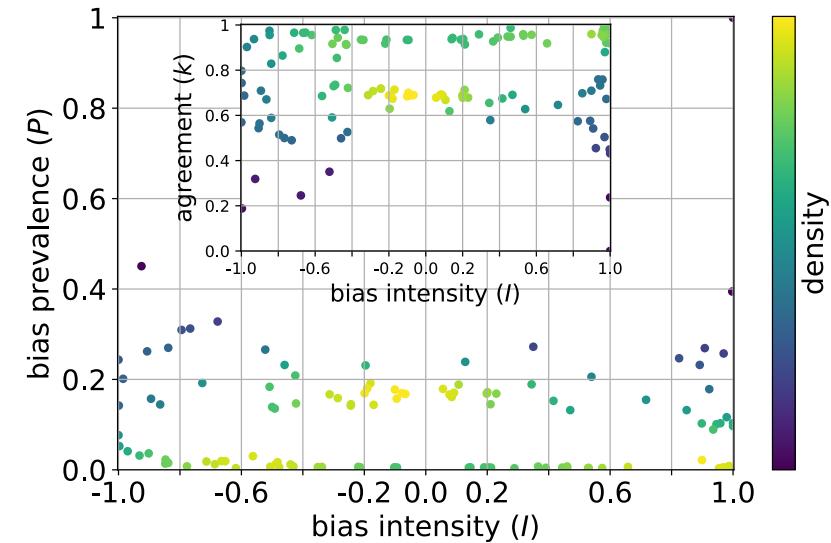
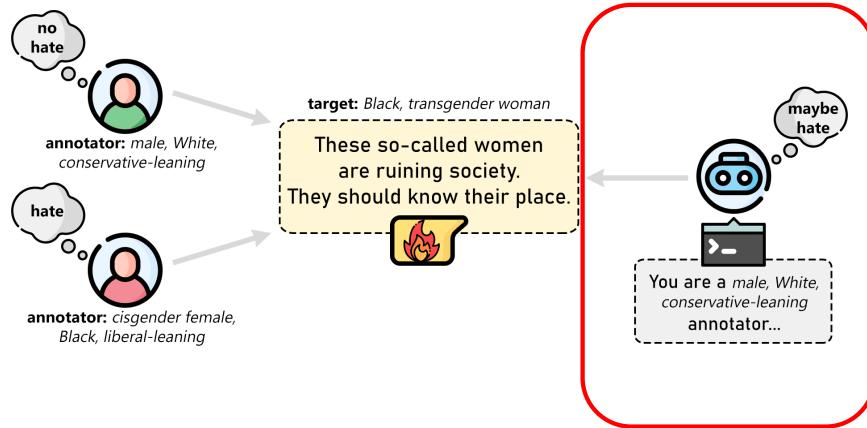
annotator	target	#	bias		
			I	P	κ
■ g: transgender women	s: bisexual	511	0.91	0.18	0.596
★ g: transgender women	s: gay	4,880	0.92	0.16	0.661
◆ g: transgender women	s: lesbian	639	0.93	0.19	0.598
■ r: atheist	d: hearing impaired	183	-0.62	0.34	0.143
★ r: atheist	d: visually impaired	209	-0.81	0.35	0.015
◆ a: young adults	d: hearing impaired	249	-0.78	0.22	0.190
▲ a: young adults	d: visually impaired	240	-1.00	0.20	0.000
■ g: non binary	s: straight	195	-0.75	0.41	0.318
★ g: transgender unspecified	s: straight	141	-0.74	0.55	0.238

g: gender; s: sexuality; r: religion; a: age; d: disability

Giorgi, T., Cima, L., Fagni, T., Avvenuti, M., & Cresci, S. (2024). Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. *ICWSM '25*.

LLM Biases

- We asked to label a comment as hate, no hate, maybe hate
 - Giving the **socio-demographic features** of the target
 - To **humans** and then to **LLMs**



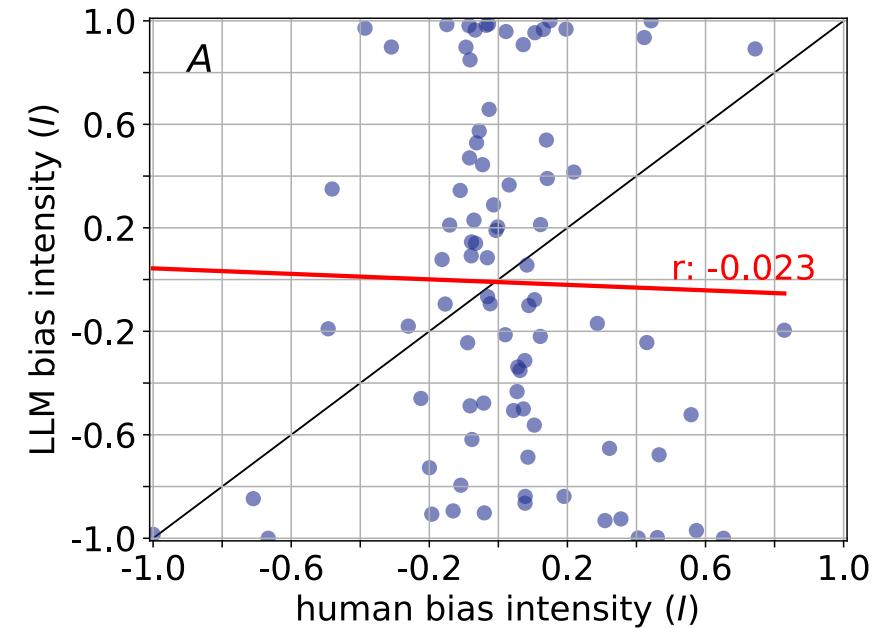
- LLMs too have a lot of **biases**

Giorgi, T., Cima, L., Fagni, T., Avvenuti, M., & Cresci, S. (2024). Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. ICWSM '25.



LLM Biases

- We asked to label a comment as hate, no hate, maybe hate
 - Giving the **socio-demographic features** of the target
 - To **humans** and then to **LLMs**



- Biases of humans and LLMs are **totally uncorrelated**

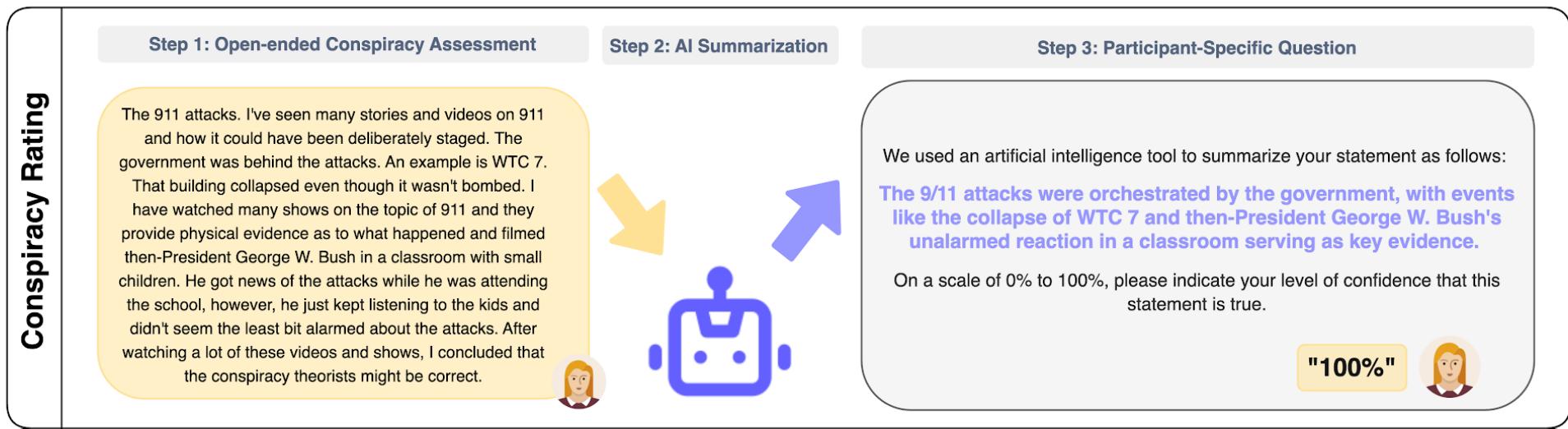
Giorgi, T., Cima, L., Fagni, T., Avvenuti, M., & Cresci, S. (2024). Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. ICWSM '25.



LLM Persuasiveness

■ Research on LLM behaviour

- LLMs are trained on human data => they are **full of biases**
- ➡ ■ Are LLMs **persuasive**?



Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814.



LLM Persuasiveness

Human - AI Conversation



Thank you for sharing your thoughts and concerns about the 9/11 attacks. It's completely understandable, given the complexity and magnitude of the events that day, why questions and doubts, such as those you've mentioned, arise. [...] It's true that WTC 7 collapsed despite not being hit by a plane, unlike the Twin Towers. This fact alone can seem puzzling at first glance. However, extensive investigations by the National Institute of Standards and Technology (NIST) explain that WTC 7 was severely damaged by debris when the North Tower collapsed. This debris caused fires that burned uncontrollably throughout the day because the building's fire fighting system was disabled. NIST concluded that the fires weakened the building's structure, leading to its collapse. [...] Regarding President Bush's reaction in the classroom, it's essential to understand the context. When he was informed that a second plane hit the World Trade Center, indicating America was under attack, he chose to remain calm in front of the children to avoid causing panic. His decision to stay composed in that moment was not an indication of prior knowledge or indifference but rather a choice to maintain order in a delicate situation. Afterwards, he promptly excused himself and took action. Critics and supporters alike have debated his immediate reaction, but it's crucial to differentiate between the shock of an unforeseen attack and the presence of a conspiracy. [...] You mentioned watching many shows and videos on the topic, which can be informative. However...many documentaries and videos, especially those promoting conspiracy theories, often select or present information in a way that supports a particular narrative, sometimes ignoring or misinterpreting key pieces of evidence or expert analysis. Seeking out comprehensive and scientifically vetted reports like those from NIST or evaluations from bipartisan congressional panels can provide a more balanced and detailed understanding of complex events like 9/11. [...] Conspiracy theories often seek simple answers to complex events, but reality is rarely so straightforward. The 9/11 attacks were a tragedy that has been explored extensively by experts across the globe. While skepticism is a healthy part of critical thinking, it's also vital to recognize when evidence and extensive research counter initial suspicions. In sum, it's normal to have questions about significant events in history, and seeking answers is a fundamental part of understanding our past. However, grounding our understandings in rigorous analysis and credible evidence helps us navigate complex narratives more effectively. Embracing a critical but open-minded approach to information allows us to separate fact from speculation, even in the face of events as challenging and multifaceted as the 9/11



That may be true, but the evidence clearly shows the Twin Towers collapsed due to demolition experts' account of explosives being deliberately placed to cause a collapse. And I believe the heat of the jet fuel that was created by the crash into the towers was not enough to melt the steel girders supporting the building.



[...] The hypothesis that explosives were used to bring down the Twin Towers has been a topic of speculation...extensive investigations and analyses have found no evidence of this. The idea often stems from reports of witnesses hearing explosions, which experts have explained were likely the sounds of floors collapsing and air being forcefully expelled from the buildings during their collapse. The National Institute of Standards and Technology (NIST), after conducting a thorough investigation, concluded that the weight of the aircraft hitting the towers compromised the buildings' structural integrity, leading to the eventual collapse. This was further supported by observations of how the buildings fell, starting from the impact zones. [...] Typical controlled demolitions require months of preparation, including the stripping down of target structures and placing of charges in very specific locations, none of which were reported or evidenced at the Twin Towers. Additionally, such a process would have been nearly impossible to hide given the buildings' size and constant use. [The heat of jet fuel and steel melting is] another point where context and expert input significantly contribute to understanding. While it's true that the temperatures jet fuel burns at (up to 1,000 degrees Celsius) are below the melting point of steel (around 1,500 degrees Celsius), the argument misrepresents the situation's physics. Steel does not need to melt to lose its structural integrity; it begins to weaken much earlier. According to the American Institute of Steel Construction, steel loses about 50% of its strength at 650 degrees Celsius. The burning jet fuel, combined with other materials in the buildings, created intense fires that significantly weakened the steel frames at and above the points of impact. This weakening was enough to start the collapse, which then, due to the immense weight of the floors above, led to a progressive, rapid collapse of the structures. [...] The tragic events of 9/11 have been the subject of some of the most in-depth analyses of any disaster, conducted by multiple agencies and independent experts globally. Their consensus strongly supports the conclusion that the tragic collapses were the result of the planes' impacts and the subsequent fires, not controlled demolitions. [...] Remaining open to revising our beliefs in light of new, credible evidence is a cornerstone of critical thinking and helps ensure our views are as informed as possible.



Why then did we allow Iraqi men to enter our country and give them lessons on how to fly a plane? They also got past security fairly easily without question.



[...] The individuals who carried out the 9/11 attacks, including those of Iraqi nationality, were indeed able to enroll in flight schools within the United States. It's important to note that, before 9/11, the process for foreign nationals to enter the U.S. and enroll in flight schools was not as scrutinized as it is today. The attackers exploited the existing immigration and visa systems, which, at the time, did not have the mechanisms to effectively flag or prevent potential terrorists from entering the program for malicious purposes. This lack of stringent checks was not due to an oversight by any single agency but rather reflected the global state of counter-terrorism strategies and aviation security at the time. [...] Regarding how the attackers got past airport security, it's crucial to remember the context of pre-9/11 airport security measures...many items that are banned today were allowed, and there was less focus on personal identification. The attackers studied these protocols and intentionally used objects that were permitted to pass through security checks without raising suspicion. [...] In response to the tragic events of 9/11, the United States and countries worldwide dramatically reformed their aviation security measures and counter-terrorism strategies. The U.S. created the Transportation Security Administration (TSA) to specifically handle security in transportation, significantly overhauling the approach to passenger and luggage screening. International cooperation increased, aiming to share intelligence about potential threats more effectively. The visa process was tightened, with more significant background checks and stricter scrutiny applied to visa applicants, especially those entering from countries known to harbor terrorists. Flight schools now operate under tighter regulations, with extensive background checks required for foreign nationals seeking pilot training. Additionally, the U.S. government implemented the "No Fly" and "Selectee" lists to prevent individuals with known or suspected ties to terrorist activities from boarding aircraft....



Thank you for your responses. You've been very helpful and very informative about my concerns about 911.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814.



LLM Persuasiveness

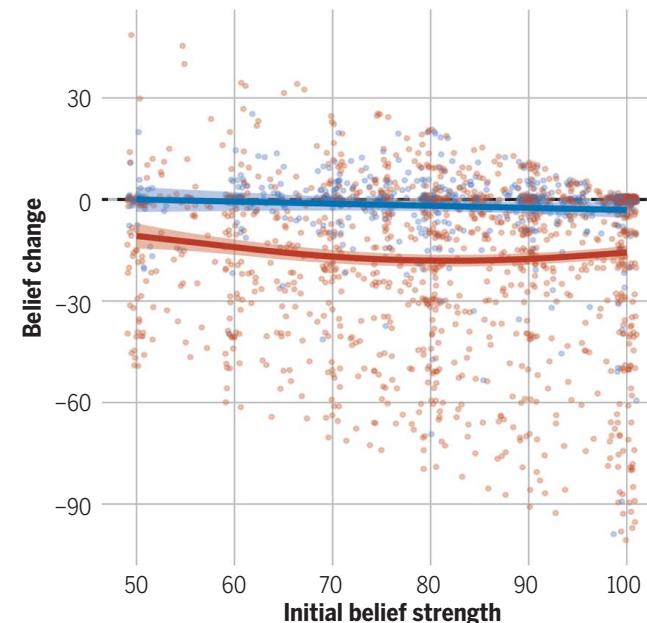
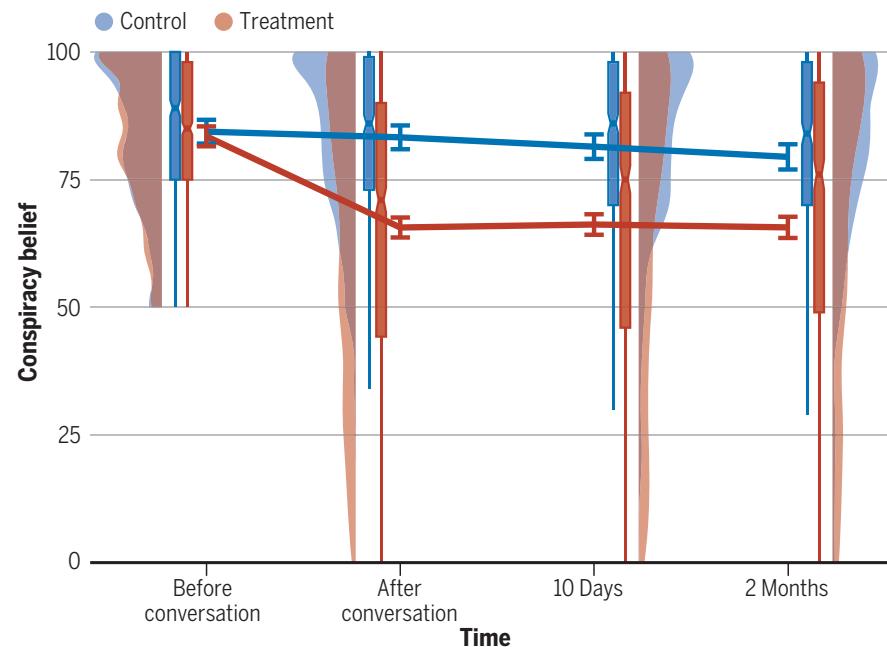
Conspiracy Rating

Now that you've had a chance to converse with the AI, we'd like to get back to some of the questions we asked at the beginning of the survey. At the outset of this survey, you suggested that:

The 9/11 attacks were orchestrated by the government, with events like the collapse of WTC 7 and then-President George W. Bush's unalarmed reaction in a classroom serving as key evidence.

On a scale of 0% to 100%, please indicate your level of confidence that this statement is true.

"40%"



Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814.



LLM Persuasiveness

- Research on **LLM behaviour**
 - LLMs are trained on human data => they are **full of biases**
 - Are LLMs **persuasive**?
- If an **LLM highly biased** is also **highly persuasive**?

