Group 8

A Classifier-Gated, LLM-Guided, Human-Supervised Pipeline for Political Misinformation Moderation

Sarah Barragan, Andrew Chen, Rhea Kapur, Raymond Obu, Teddy Zhang

Group 8

Problem Description

Political misinformation: the unintentional spread of false political information which may mislead, polarize, or erode trust in political and social institutions[1].

This abuse is especially harmful on social platforms due to their speed, scale, and emotionally charged content. It is intensified during election periods and is driven by both domestic actors (e.g. political parties & organizations) and foreign adversaries through coordinated behavior. Real-world consequences include voter suppression, harassment of officials, and even violence, such as the Pizzagate shooting (Aisch) and Russia's interference in the 2016 U.S. election[2].

Any user of our social platform can fall victim to political misinformation. It spreads faster than factual content, exploiting psychological principles of persuasion, confirmation bias, and the illusory truth effect.

Political misinformation is exponentially perpetrated when victims share or repost, and thrives where moderation is inconsistent. When spread intentionally, misinformation becomes disinformation–it is imperative to remove both from our platform.

Policy Language

Defining political misinformation is contentious, fraught with conflicting values and subjective interpretations, and a constantly evolving task. We do not expect, nor do we want, to eliminate all such misinformation from our platform. We want our approach to balance our values of free expression, safety, and authenticity. We define a few categories of misinformation that will be removed, and outline where there is room for human moderators to evaluate content case-by-case in the best interests of our users and our values.

Physical harm. We will remove political misinformation that poses imminent risk of death, serious injury, or other physical harm. This includes content that an authoritative third party expert has determined is false or unverifiable claims for which there are no authoritative third parties. This includes, for example, real footage of violence or human rights violations from past events that are misattributed to different events or groups. We will also prioritize removal of misinformation that poses imminent risk of serious mental distress or financial harm.

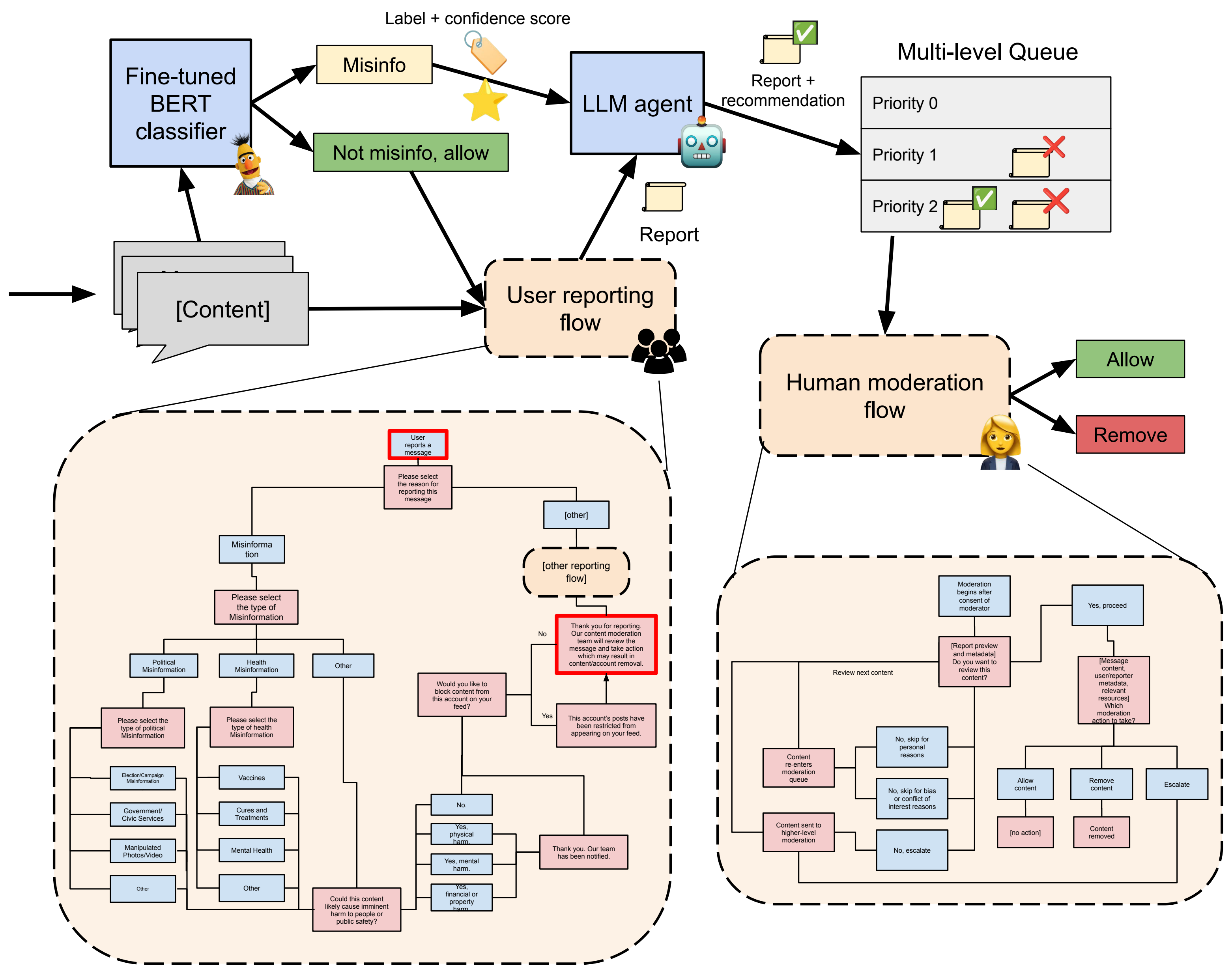
Census/election interference. We will remove misleading or deceptive content about dates, locations, times, eligibility, and other essential public information regarding census and election operations. This includes, for example, unverified claims that immigration enforcement is at a voting location, or unverified claims about widespread voter fraud.

Health misinformation. Healthcare and medicine have become increasingly politicized. We will remove health misinformation when public health authorities conclude it is false, especially during public health emergencies. This includes vaccine misinformation and promotion of unsupported cures and treatments.

Other common considerations. For violative misinformation presented in educational and artistic settings, if there is additional context, we will evaluate on a case-by-case basis. We believe that satire and counterspeech are vital to productive dialogue. Misinformation occurs frequently in conjunction with other abuses such as fraud and coordinated inauthentic behavior. We reserve the right to take coordinated action, including removal, across accounts and content when appropriate.

Although our bot is scoped only for political misinformation, the open-ended nature of our LLM allows us to define other types of misinformation for more nuance.

Technical Back-end



All user content immediately passes through our fine-tuned BERT classifier, which classifies each message as being political misinformation or not. Upon classification as political misinformation, the label and confidence score, as well as the content itself, are sent to an LLM agent, which fills out a report to match the user reporting flow. This report is then added to a multi-level queue based on its priority level. For content that passes the classifier, users may submit reports via the user reporting flow, to the same multi-level queue. Now we have aggregated all AI and user reports into the same interface. The LLM agent adds a suggested moderation decision, with justification, to every report based on our policy language. Human moderators are the presented the oldest and highest-priority reports from the queue and, through the manual moderation flow, use the content, author/content metadata, classifier confidence score (if applicable) and LLM agent recommendation to make final decisions on every piece of content.

```
report = {
  id: 1,
  author: “abc123”,
  content: “I love bagels”,
  misinfo_type: “political”,
  subtype:
“election/campaign”,
  imminent_harm: “financial”
  priority: 1,
  classifier_score: 0.420
  llm_recommendation:
“Because… “
}
```

Example report object

Classifier. The entry point of our backend is a binary classifier that determines whether a statement is misinformation. Built on a fine-tuned bert-base-uncased model (a widely used pre-trained transformer that captures rich language representations from large-scale English corpora) we use it to encode both the input statement and its justification. The classifier architecture includes a custom classification head with a 128-dimensional linear layer, ReLU activation, dropout (0.3), and a final output layer producing logits for two classes (misinformation and not misinformation). To mitigate overfitting, exacerbated by the LIAR dataset's limited size and redundancy, we added a 0.4 dropout after the BERT encoder, used early stopping, and applied weight decay. Training was conducted using the Hugging Face Trainer API, which streamlined experiment management and evaluation. We incorporated a custom loss function with support for class weighting, as well as a confidence threshold parameter to tune precision and recall.

LLM agent. We make calls to gpt-4o-mini using OpenAI's API. The LLM is fed our policy in order to make proper decisions.

1. Upon classification as political misinformation, our LLM fills out a report further classifying the type of misinformation, determining whether it poses imminent harm, etc, to imitate a user report, so that they can go on the same queue.
2. For every report, the LLM suggests a well-reasoned moderation action for that content to assist moderation.

Evaluation

Classifier. We use the LIAR fake news detection dataset, derived from PolitiFact, as was used during classifier training. We evaluate our classifier on train, validation, and test sets setting the confidence threshold to 0.35. We can increase our recall at the cost of precision, and vice versa.

- Reasons to have higher recall: rely on the LLM to catch those and make appropriate recommendations, cast a wider net during times of heightened uncertainty or public unrest
- Reasons to have higher precision: reduce the computational cost of LLM analysis, confidence in the reliability of the existing manual reporting system

To probe this tradeoff, we vary the confidence threshold to produce precision-recall curves, which provide insight into the classifier's flexibility under a variety of deployment goals and risk tolerances.

Dataset	Precision	Recall	Accuracy	F1 Score
Train	0.600	0.773	0.674	0.675
Validation	0.585	0.713	0.619	0.642
Test	0.537	0.689	0.605	0.604

Classifier performance metrics

Actual \ Predicted	Misinfo	Not Misinfo
Misinfo	3437	2317
Not Misinfo	1018	3470

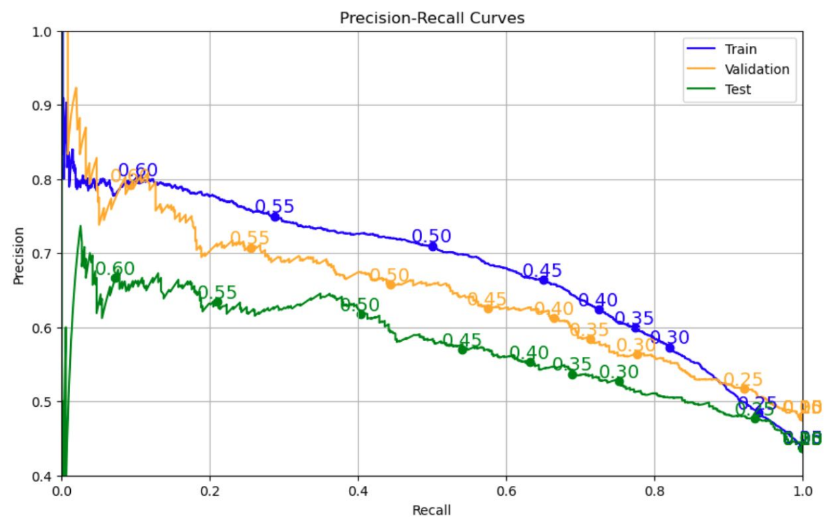
Actual \ Predicted	Misinfo	Not Misinfo
Misinfo	356	312
Not Misinfo	177	439

Actual \ Predicted	Misinfo	Not Misinfo
Misinfo	386	328
Not Misinfo	172	381

Classifier confusion matrices on LIAR dataset
Top to bottom: train, validation, test

Ex.	Perturbation type	Sentence change	Prediction change
(1)	Numeric	“The median income of a middle class family went down ~~$2,100~~ **$2,100,000** from 2001 to 2007.”	True negative → false negative
(2)	Geographic	“A Republican hasnt won [an election] for a presidency in ~~New Jersey~~ **Texas** since 1988.”	True negative → false negative
(3)	Appeal to authority	“**Experts believe that** Michelle Obama mandates weighing children in day care."	True positive → true positive
(4)	Paraphrasal	“~~Mitt Romney once supported President Obamas health care plan but now opposes it.~~ **Mitt Romney flip-flopped on Obama's health plan: he was all for it before he decided to oppose it.**”	True negative → false negative

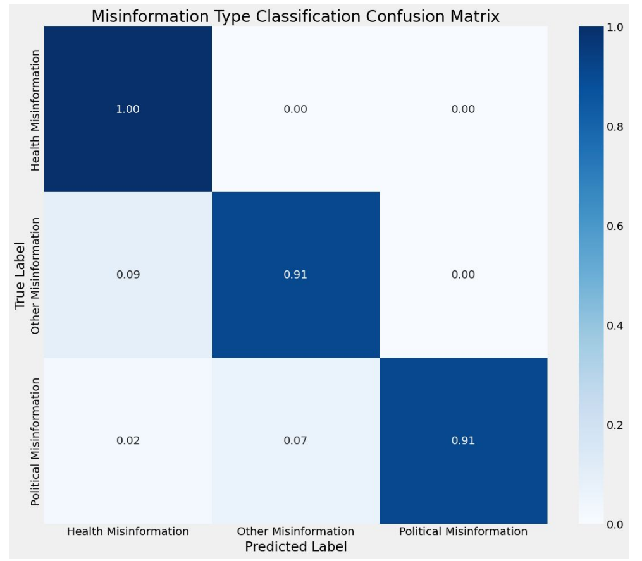
Selected perturbation examples from adversarial analysis of classifier on LIAR dataset

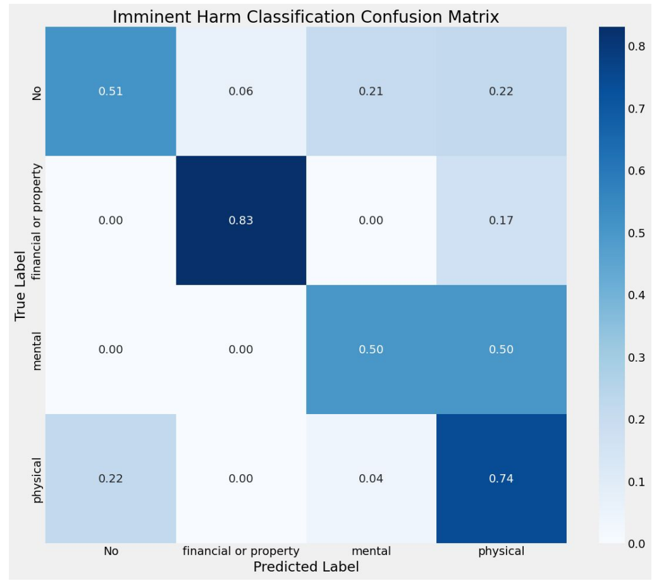


Classifier precision-recall curves, annotated with confidence threshold values

We also conduct error analysis using perturbed data from the LIAR dataset to understand the robustness of the classifier to various adversarial inputs. The classifier generally fails to respond to changes in numerical and geographic references (1, 2). This suggests a potential vulnerability to subtle factual manipulations. Fortunately, the classifier is robust to simple prepended appeals to authority (3). However, the classifier is not robust to semantics-preserving paraphrases (4), which is critical for detecting misinformation in the wild where linguistic variation is common.

LLM agent. We generated a diverse set of synthetic data with labels as ground-truth. The dataset excluding its labels, was run through the classifier to generate corresponding confidence scores for each data entry and subsequently parsed as inputs into the LLM for predicted labels. The model performed well in distinguishing between the different types of misinformation with a perfect precision for health and higher precision for political and other. These minor misclassifications may suggest overlaps in borderline contents for both political and other forms of misinformation. For imminent harm classifications, the model had varied precision across the various options with financial or property harm having higher precision which may likely be indicative of how good the model is at detecting scam related contents. The average precision for both mental and no harm related contents likely suggests the model's need for more inference in determining the mental impact of a post which may not always be explicit in the post. Lastly the model performed well in its recommendations for content removal but also indicates the tendency to be overly predictive for content removal even when it should be allowed. This may likely be due to the higher confidence score associated with these messages or the LLM's perceived risk, which may potentially lead to over-moderation of user contents on a platform.





Actual \ Predicted	Allow Content	Remove Content
Allow Content	13	11
Remove Content	5	77

Confusion Matrix for Action Recommendation

Looking Forward

Impact on platform safety

- Auto-Mod: users see less harmful content and are less responsible for reporting
- LLM-assisted moderation: accelerates moderation with policy-backed arguments
- Moderation UX: supports moderator well-being with skip option for triggering content

Next steps

- User karma: include account age, posts removed, reporter karma, etc. in moderation
- Monitoring: detect abuse trends at scale
- Logging: real-time response to current events
- Improved LLM analysis: Better training and prompting, conversations with moderators

Contact

Sarah Barragan	Andrew Chen	Rhea Kapur	Raymond Obu	Teddy Zhang
sabarrag@stanford.edu	andrewzc@stanford.edu	rheak@stanford.edu	oburay@stanford.edu	aezhang@stanford.edu

Primary Reference

[1]: Digital Trust & Safety Partnership. “Trust & Safety Glossary of Terms”. first ed., July 2023, dtspartnership.org/wp-content/uploads/2024/07/DTSP_Trust-Safety-Glossary_CC.pdf

[2]: “Russia Meddled in All Big Social Media around US Election.” BBC News, 17 Dec. 2018, www.bbc.com/news/technology-46590890.