AmeriSpeak

CELEBRATING 10 YEARS

From Social Media to Survey Data

Employing AI-Usage Detectors to Identify AI-Generated Responses in the HIRISE+ Survey

05.13.2025

Joshua Lerner with Brandon Sepulvado, Lillian Huang, and Erin Fordyce

Disclaimer: This project is supported by Award NIJ-22-GG-00998-RESS, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this presentation are those of the author(s) and do not necessarily reflect those of the Department of Justice.

Agenda

01 Case Study – HIRISE+

02 What We Did

03 How well did it work?

04 What can we learn?

HIRISE+: Hate-Incident Reporting Initiative to Strengthen Engagement

- **Purpose:** Understand why LGBTQ+ adults do—or do not—report hate crimes and incidents.
- **Design:** 35-minute, open-link web survey (April – Dec 2024).
- **Population:** Sexual & gender minorities, age 18+, living ≤20 mi of major metro areas: two in the South, one in Midwest and West
- **Sample plan:** Quota = 2,502 completes (race/ethnicity × site).
- **Recruitment:** 9 channels—social ads, CBO partners, referrals, venue flyers, panel vendor, etc.
- **Incentive:** $10 digital gift card per eligible completion.

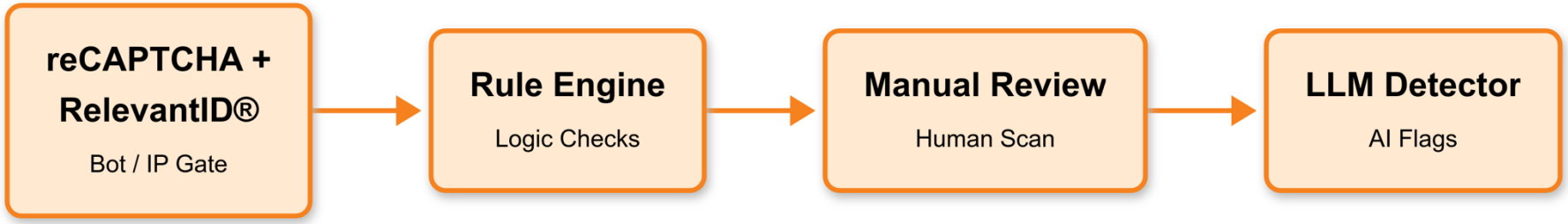
Funded by National Institute of Justice; conducted by NORC.

Why Fraud Matters for HIRISE+

Consequences of Fraud for a Hate-Crime Survey

- **Data integrity:** bogus cases hide real patterns in hate-crime reporting.
- **Bias risk:** open-link channels over-represent machine respondents, under-represent community voices.
- **Cost drain:** $10 incentive × hundreds of false completes = lost budget.
- **Staff burden:** manual review hours skyrocket without automation.
- **Erosion of trust:** Stakeholders (NIJ, community partners) will question findings if we can't document robust fraud-mitigation steps.
- **Precedent for future LGBTQ+ surveys:** A proven fraud-detection pipeline sets the standard for future LGBTQ+ surveys and strengthens NORC's reputation.

Layered Defense Against Fraud



- **Instant blocks:** reCAPTCHA + RelevantID® rejected **4,827 hits** (35% of all clicks).
- **Rule + human review:** flagged another **3,640 frauds** (26%).
- **AI clean-up:** LLM detector caught **8 polished frauds** missed by every prior layer—cut manual review time.
- **Net outcome: 1,201 verified completes** (8.7% of traffic) and stronger analytic credibility.

Building a Classifier

NORC Research Science

How did we create training data?

Questions

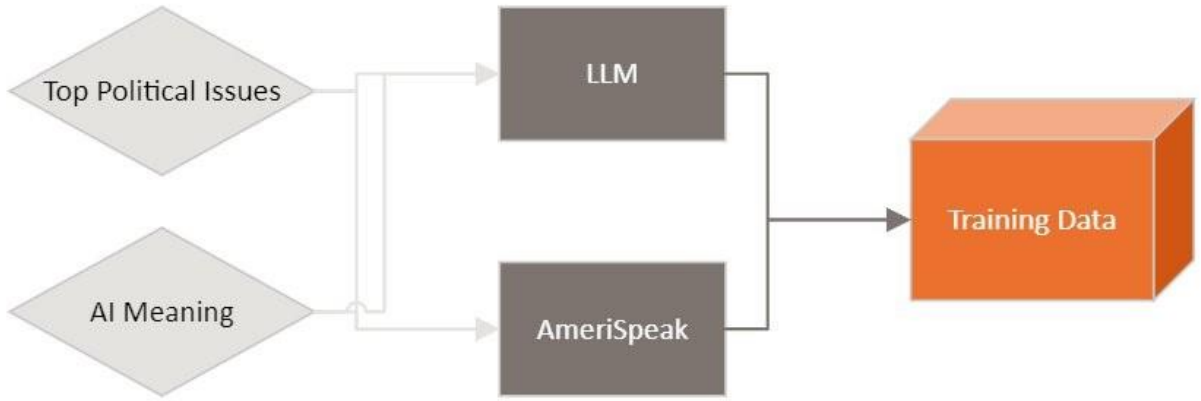
- Most salient policy issues
- Understanding of AI

Human respondents

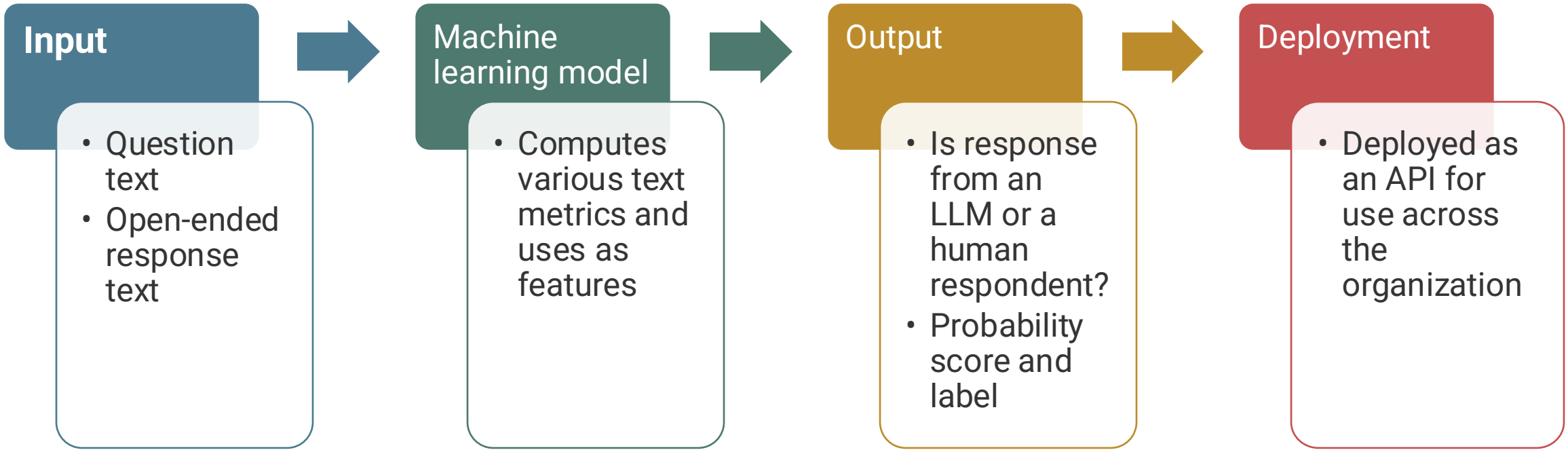
- AmeriSpeak Omnibus panel

Large Language Models

- GPT 3.5
- GPT 4
- Llama 3.1
- Claude 3.5 Sonnet



How is our detector built?



Three-Model Ensemble for Detecting AI Text

Model	What it looks at	Why we include it
1 · Simple Textual (GPT-only)	Stylometric features and raw text learned **solely from GPT outputs**	Fast baseline; high recall on classic ChatGPT copy
2 · Comprehensive Textual	Same as above but trained on **GPT, Llama, Claude, etc.**	Captures a wider range of LLM "voices"; boosts precision on paraphrased or non-GPT content
3 · Non-Textual	**Meta-features only**—length, complexity, question overlap	Paradata could be indicative of other non-textual patterns

Ensemble ai_score = weighted average of the three models → flag if **ai_score ≥ 0.80**

How does our detector perform?

General population survey

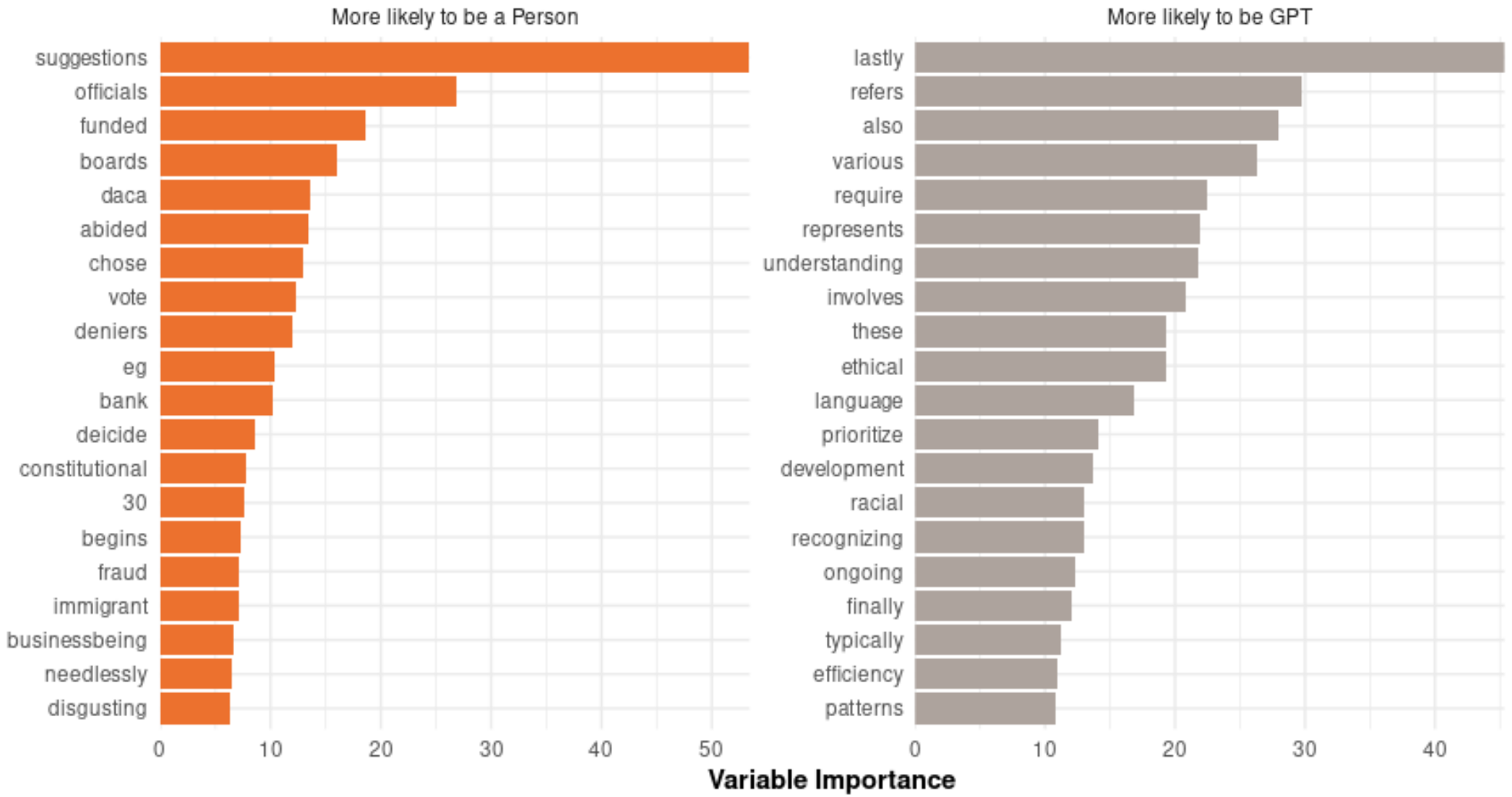
- 99% accuracy, precision, and recall

For a specific technical domain

- A separate survey for the American College of OBGYNs
- New domain (medical), highly technical language
- Accuracy in upper 80% to mid 90% across several questions
- Precision up to 85.7%, recall up to 100%

Precision	Recall
0.989	0.999
F1	**Accuracy**
0.994	0.990

Filler words matter the most for detection!



Results:

Findings

- Looked at fraudulent cases identified through in-depth interviews and found that 11 out of 20 responses with an ai_score > 0.80 were **confirmed as fraudulent** (before hand)
- Most of the identified fraudulent responses did not reply (or barely replied) to the open-ended question – so this was a value add.
- We know those 11 were fraudulent because of earlier steps in the pipeline – but many of the remaining 9 flagged had not been.
- What did those 9 left look like?

AI-Assisted Open-ended responses: What It Looks Like

Tell-tale signs

- Copy-paste repetition – identical wording across multiple IDs within minutes.
- Polished, “too perfect” prose – no contractions, no typos, textbook syntax.
- Generic feedback – vague meta-comments about survey quality, not content.
- Mismatch with paradata – short completion time, same IP blocks, recycled e-mails.

Numbers in HIRISE+

- Detector flagged extra frauds
- AI-generated text was most common in referral and CBO links.

AI-Generated Feedback We Actually Saw

“The language and wording of the survey could be more concise and straightforward to ensure participants can clearly understand the survey.”

“Simplifying the language and wording of the survey would make it easier for participants to comprehend the questions accurately.”

“Simplifying the language and phrasing of the survey would enhance comprehension for participants, ensuring clarity in understanding the questions accurately.”

AI-Generated Feedback We Actually Saw

"As a Muslim gay woman, I hope the survey takes into account the unique challenges faced by individuals like me who navigate both religious and sexual identities. It's important that the questions reflect the complexities of living with these intersecting identities, especially in a world where both homophobia and Islamophobia can coexist. I hope the survey leads to greater understanding and support for those of us who live at these intersections."

"I value this survey's effort to address hate incidents against LGBTQ+ individuals. To make it even more effective, it might be helpful to include more questions about how different factors, such as race, disability, and socioeconomic status, intersect with LGBTQ+ experiences of hate. Also, ensuring robust privacy and security measures will encourage more honest and complete responses. Thank you for your commitment to understanding and addressing these critical issues. "

What does this mean?

Extra signal, not a verdict

- Our detector flagged “AI-like” open-ends that had cleared all other checks—raising questions, not automatic ejections.

Higher flag rate in HIRISE+ than in pilot data

- Likely driven by **sampling frame**, **many recruitment paths**, and a **hate-crime topic** the model wasn’t trained on.

Interpretation ≠ accusation

- A high AI score only means “unusual for a human baseline.” It could reflect writing aids, translation tools, or genuine stylistic quirks.

Future direction at NORC

- Build bespoke, project-specific models—trained on the survey’s own sampling frame and subject matter—to sharpen precision and reduce false alarms.

Bottom line

- Layered fraud defense works best when its AI layer is tuned to the audience and the topic we’re trying to measure.

Thank you.

Joshua Lerner
Senior Research Methodologist
Lerner-Joshua@norc.org

Research You Can Trust™

NORC Research Science