AmeriSpeak

Detecting AI-Generated Survey Responses

Tool Development and Bias Mitigation

05.14.2025

Lilian Huang, Brandon Sepulvado, and Joshua Lerner

CELEBRATING 10 YEARS

AI poses both new opportunities and risks for survey research.

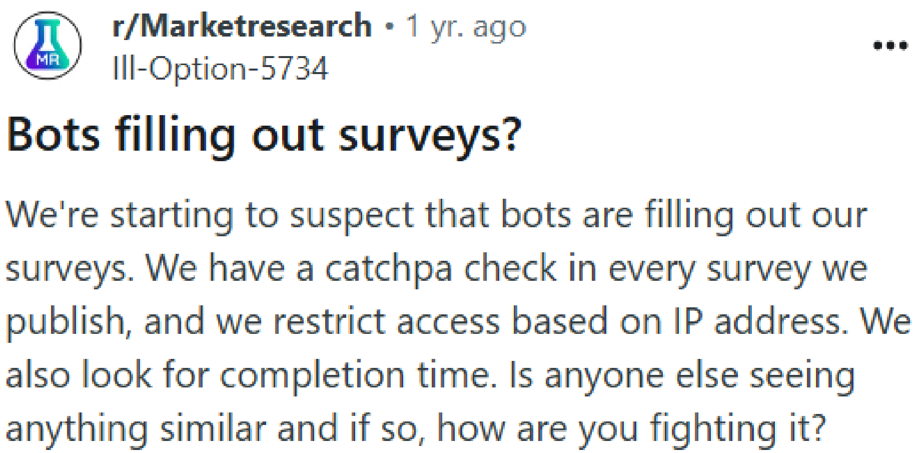
Opportunities

- Question design
- Survey administration
- Response coding

Risks

- Data quality and fraud
  - Especially for open-ends
- Results in reduced credibility among respondents and data users

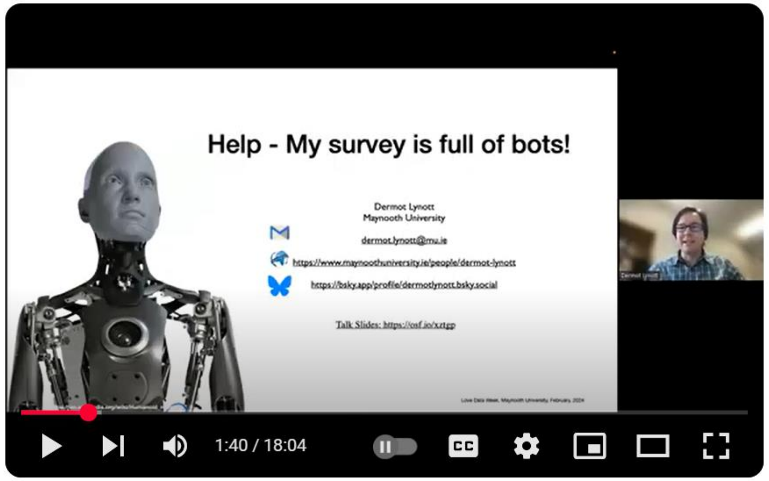
There is increasing concern over AI-generated responses.



r/Marketresearch • 1 yr. ago
Ill-Option-5734

Bots filling out surveys?

We're starting to suspect that bots are filling out our surveys. We have a catchpa check in every survey we publish, and we restrict access based on IP address. We also look for completion time. Is anyone else seeing anything similar and if so, how are you fighting it?



Help - My survey is full of bots!

Dermot Lynott
Maynooth University

dermot.lynott@mu.ie

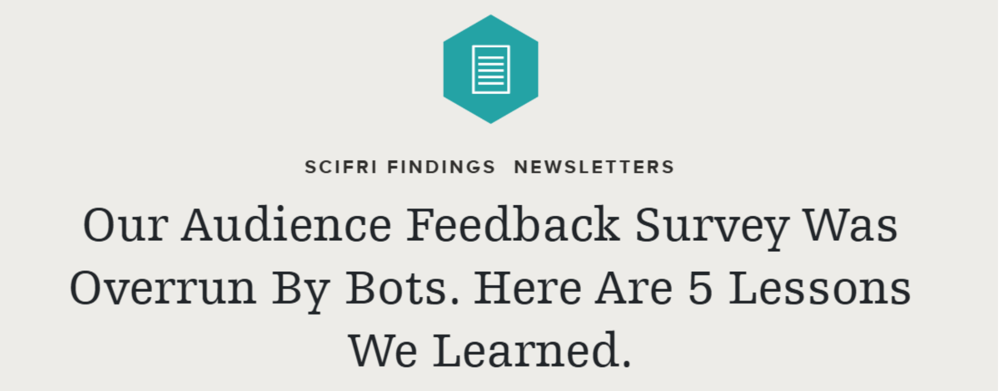
https://www.maynoothuniversity.ie/people/dermot-lynott

https://bsky.app/profile/dermotlynott.bsky.social

Talk Slides: https://osf.io/xztgp

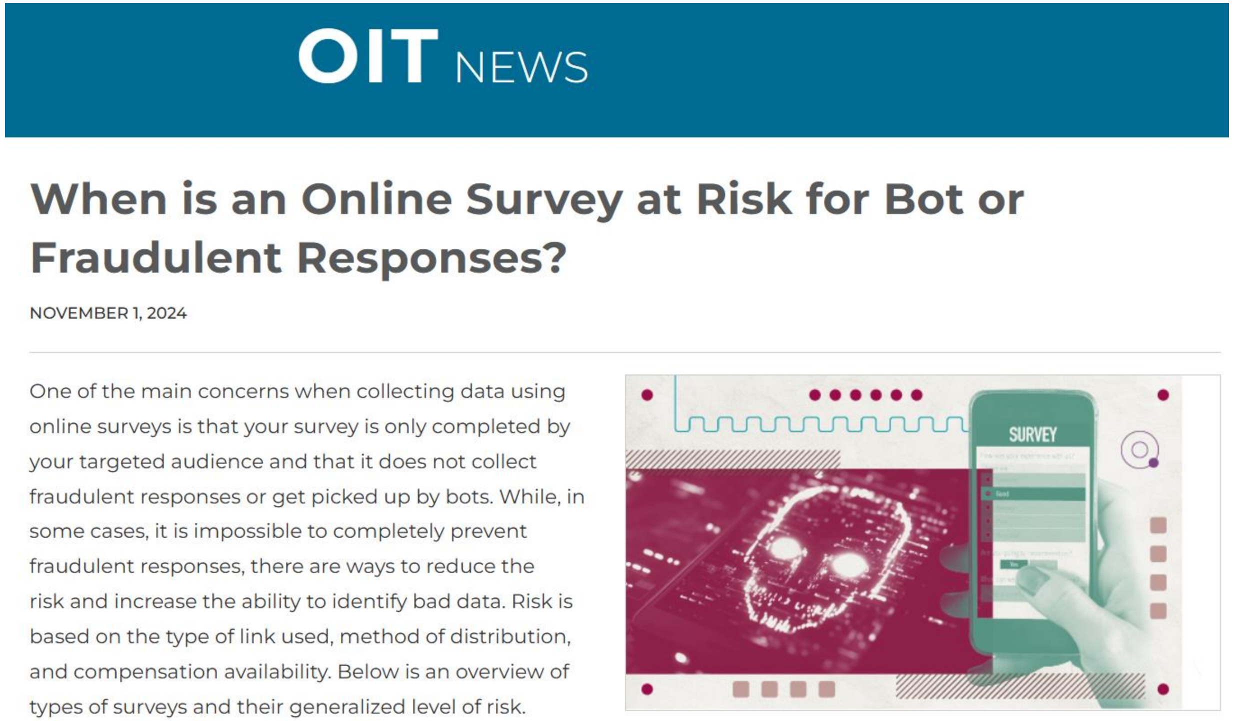
Love Data Week, Maynooth University, February 2024

Help - My Survey is Full of Bots!



SCIFRI FINDINGS NEWSLETTERS

Our Audience Feedback Survey Was Overrun By Bots. Here Are 5 Lessons We Learned.



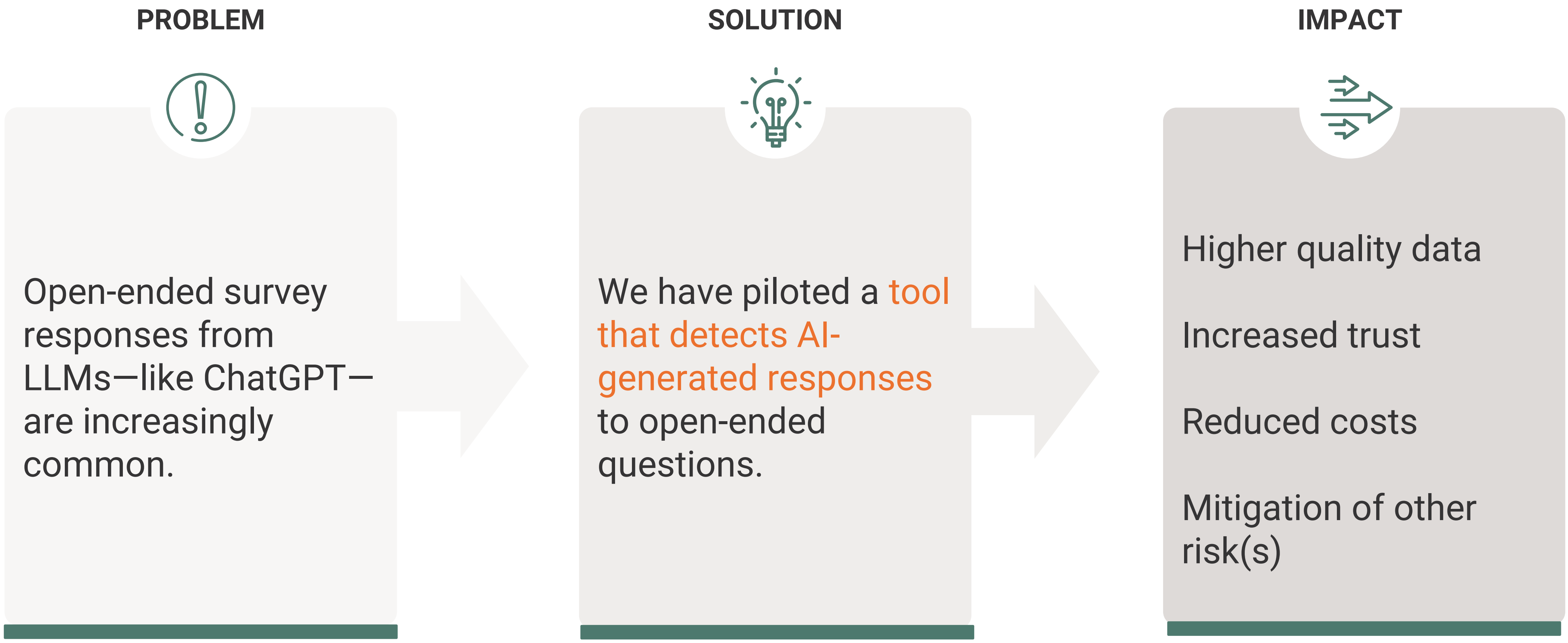
OIT NEWS

When is an Online Survey at Risk for Bot or Fraudulent Responses?

NOVEMBER 1, 2024

One of the main concerns when collecting data using online surveys is that your survey is only completed by your targeted audience and that it does not collect fraudulent responses or get picked up by bots. While, in some cases, it is impossible to completely prevent fraudulent responses, there are ways to reduce the risk and increase the ability to identify bad data. Risk is based on the type of link used, method of distribution, and compensation availability. Below is an overview of types of surveys and their generalized level of risk.

How can we protect ourselves from these risks?



PROBLEM

Open-ended survey responses from LLMs—like ChatGPT—are increasingly common.

SOLUTION

We have piloted a tool that detects AI-generated responses to open-ended questions.

IMPACT

Higher quality data

Increased trust

Reduced costs

Mitigation of other risk(s)

How did we create training data?

Questions

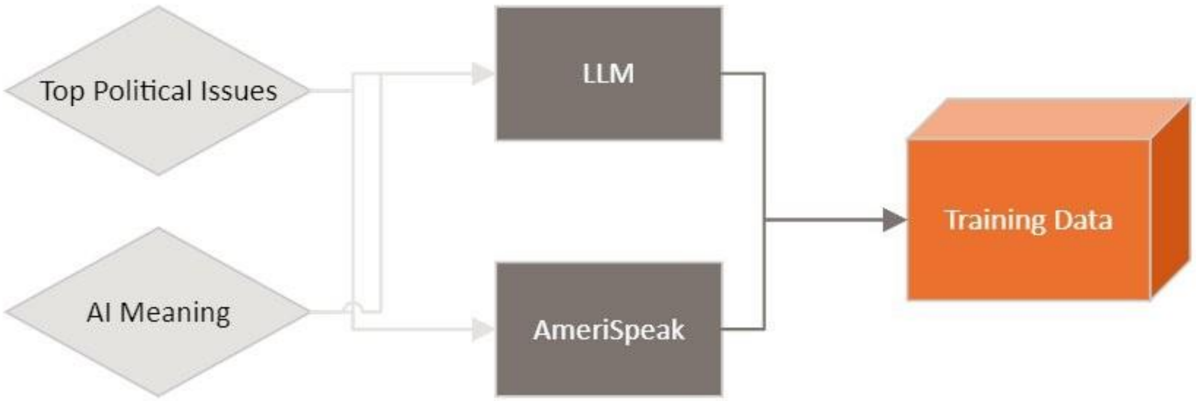
- Most salient policy issues
- Understanding of AI

Human respondents

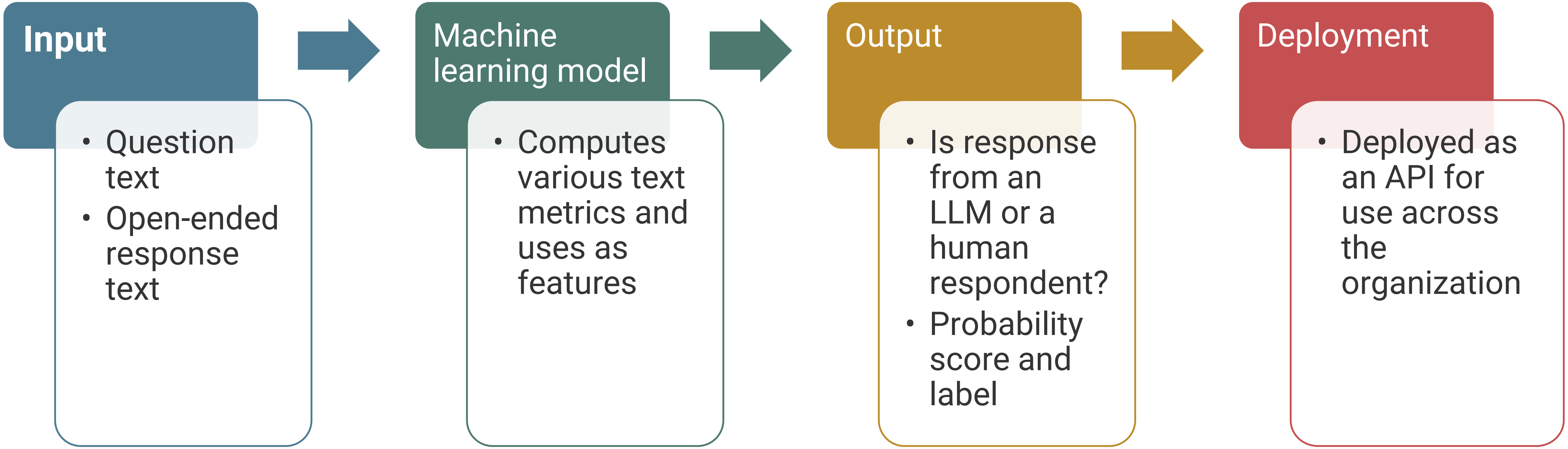
- AmeriSpeak Omnibus panel

Large Language Models

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



How is our detector built?



Input

- Question text
- Open-ended response text

Machine learning model

- Computes various text metrics and uses as features

Output

- Is response from an LLM or a human respondent?
- Probability score and label

Deployment

- Deployed as an API for use across the organization

How does our detector perform?

General population survey

- 99% accuracy, precision, and recall

For a specific technical domain

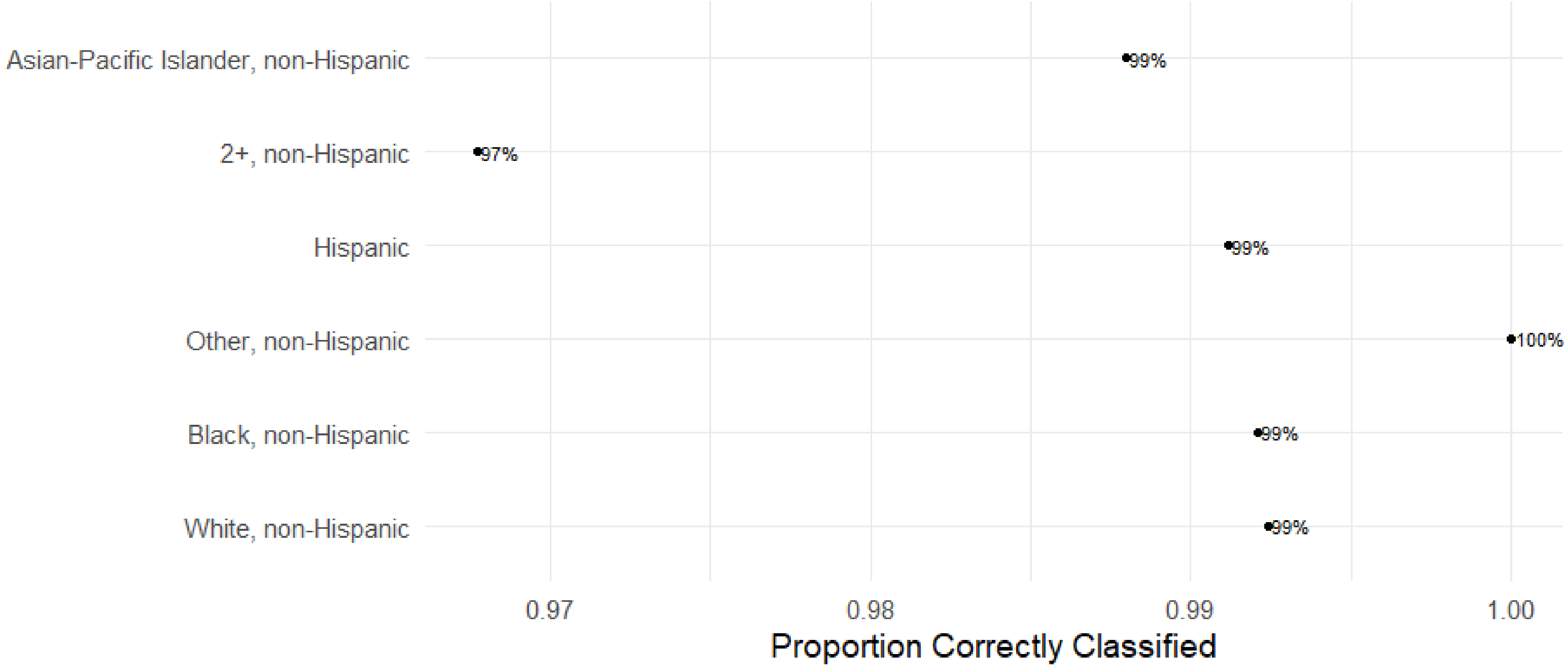
- New domain (medical), highly technical language
- Accuracy in upper 80% to mid 90% across several questions
- Precision up to 85.7%, recall up to 100%
- Multiple commercial AI detector tools had only 50-75% accuracy on this data

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

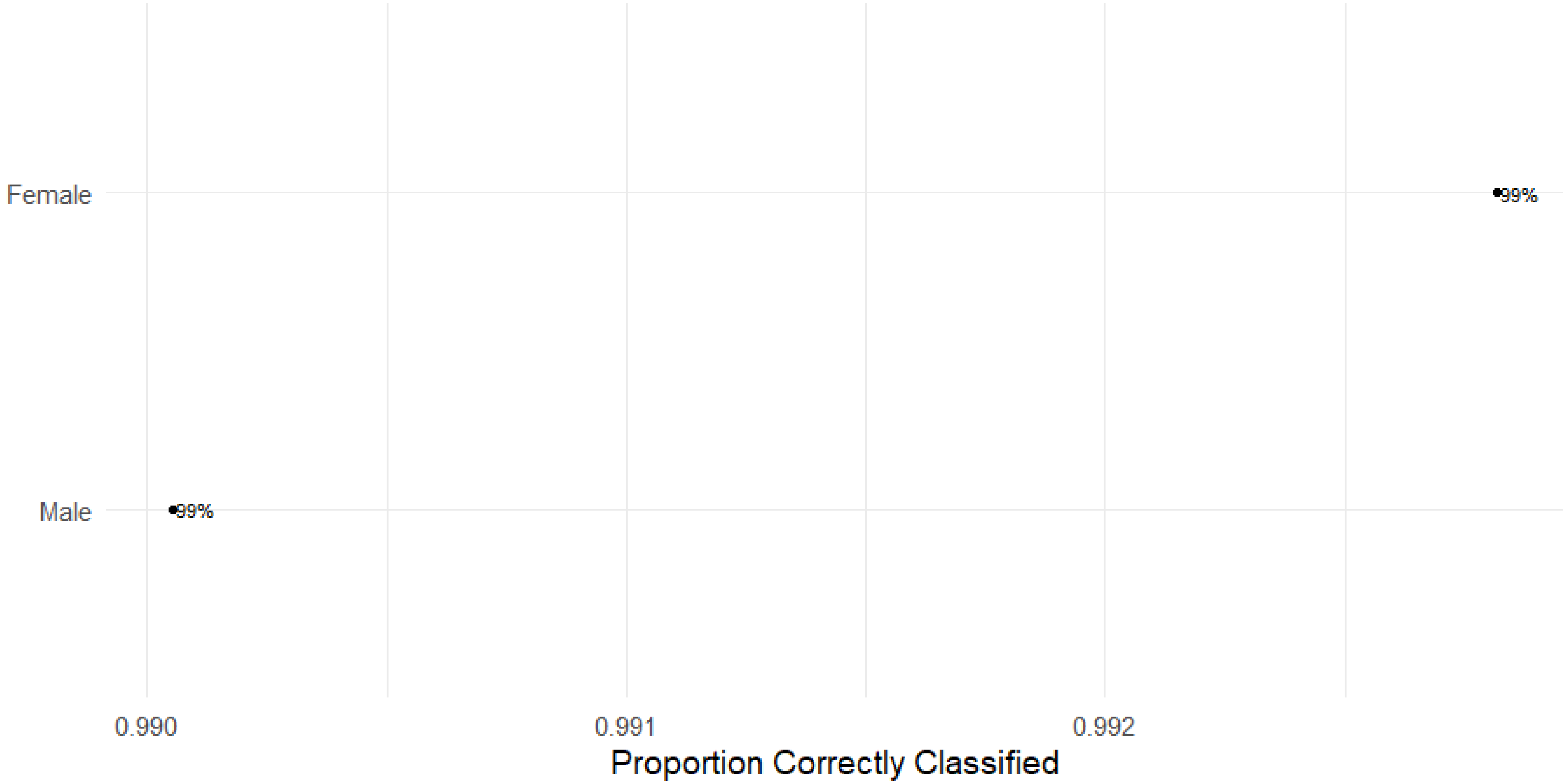
What about performance on subgroups?

- Overall metrics (e.g. precision, recall, accuracy) are not enough
- We need to ensure our model is **not biased against subpopulations**
- To investigate this, we look into **error rate balance**
  - e.g. false positive rates should be equal between different demographic groups

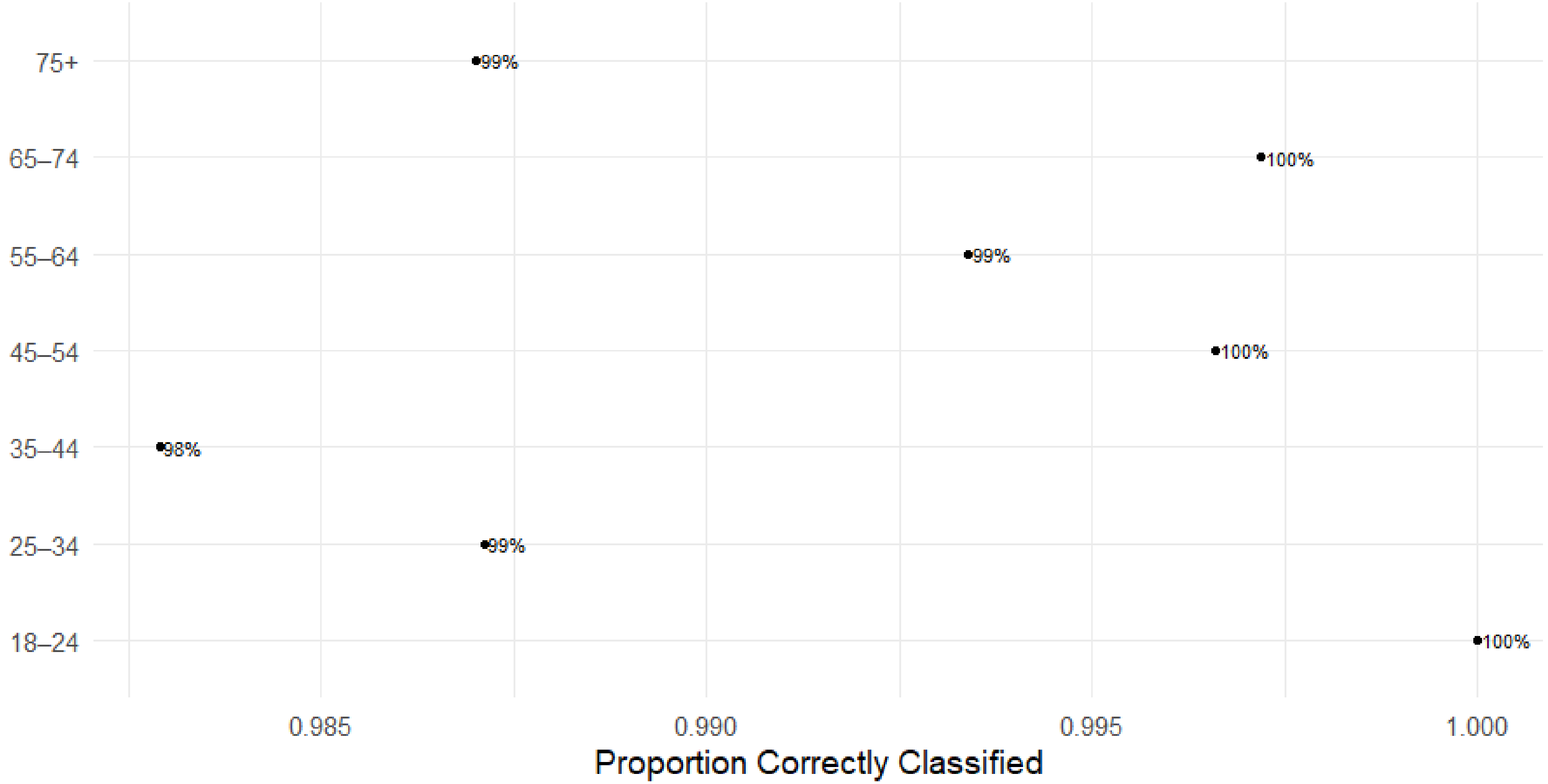
What is the correct classification rate, by subgroups?



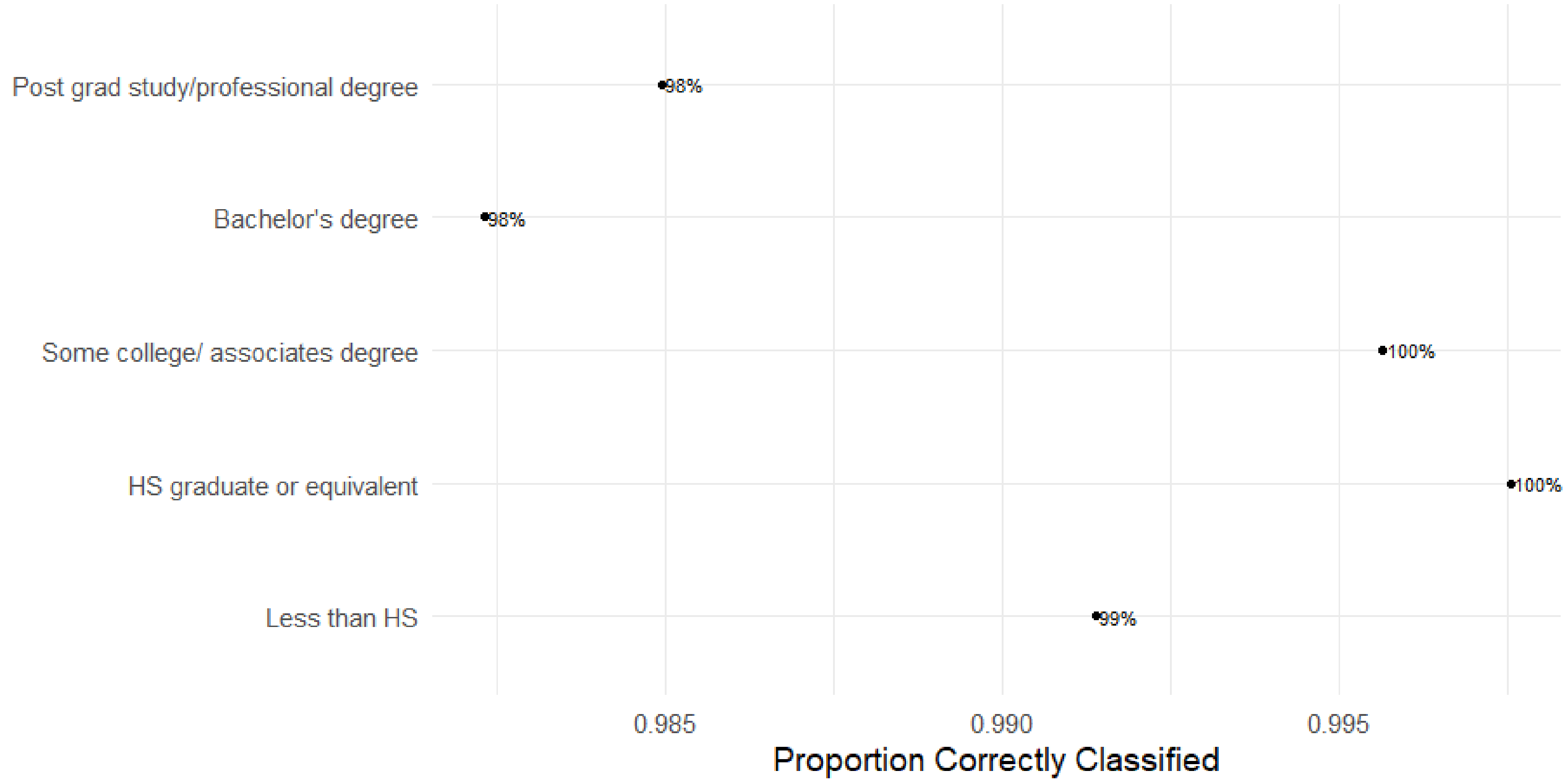
What is the correct classification rate, by subgroups?



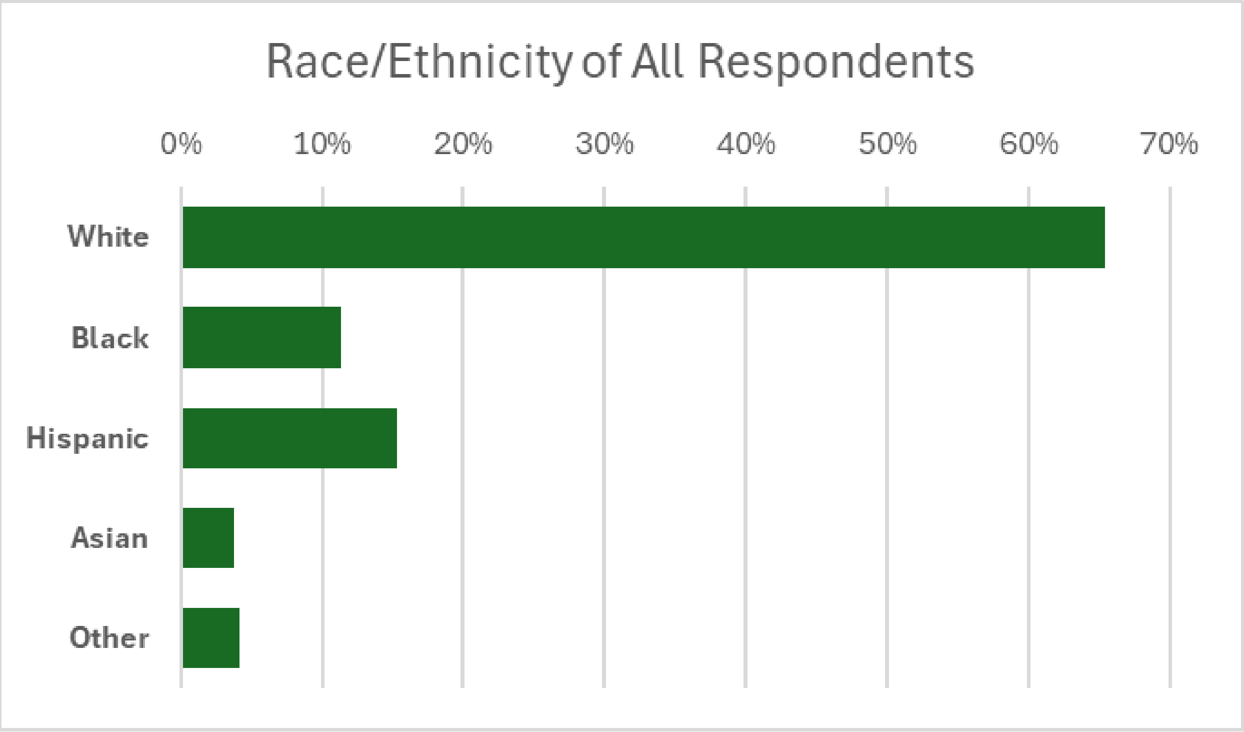
What is the correct classification rate, by subgroups?

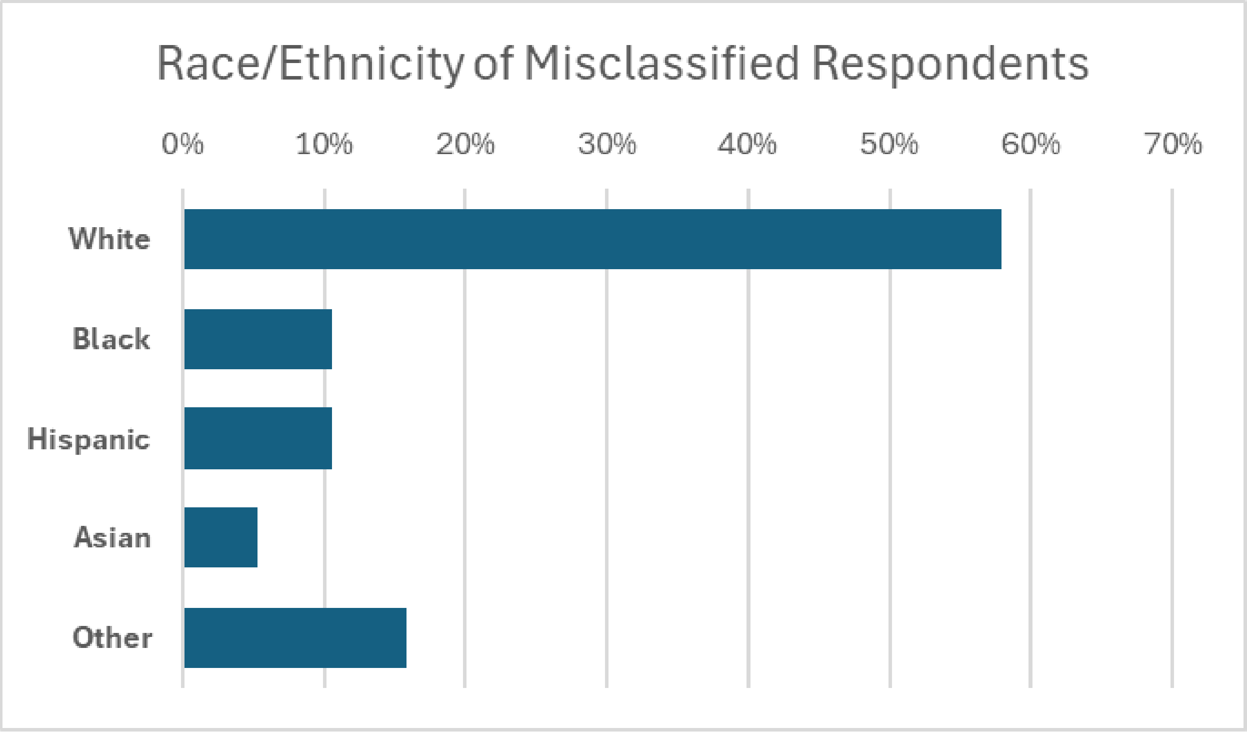


What is the correct classification rate, by subgroups?

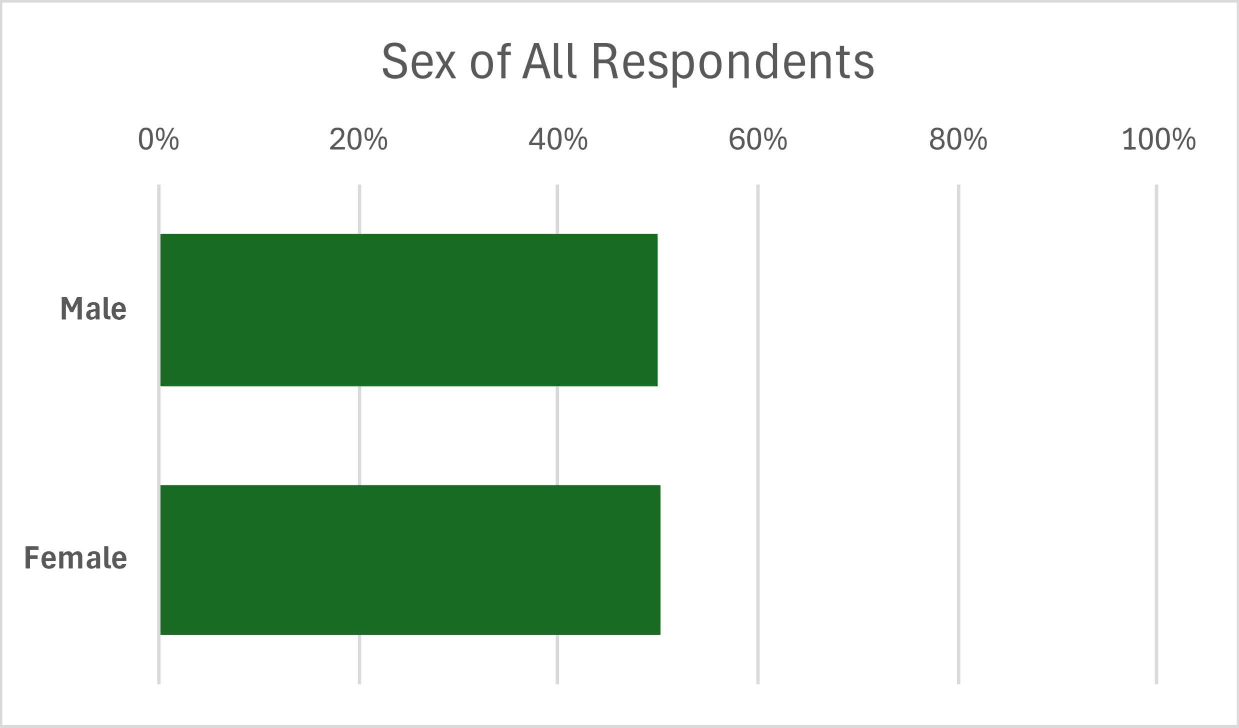


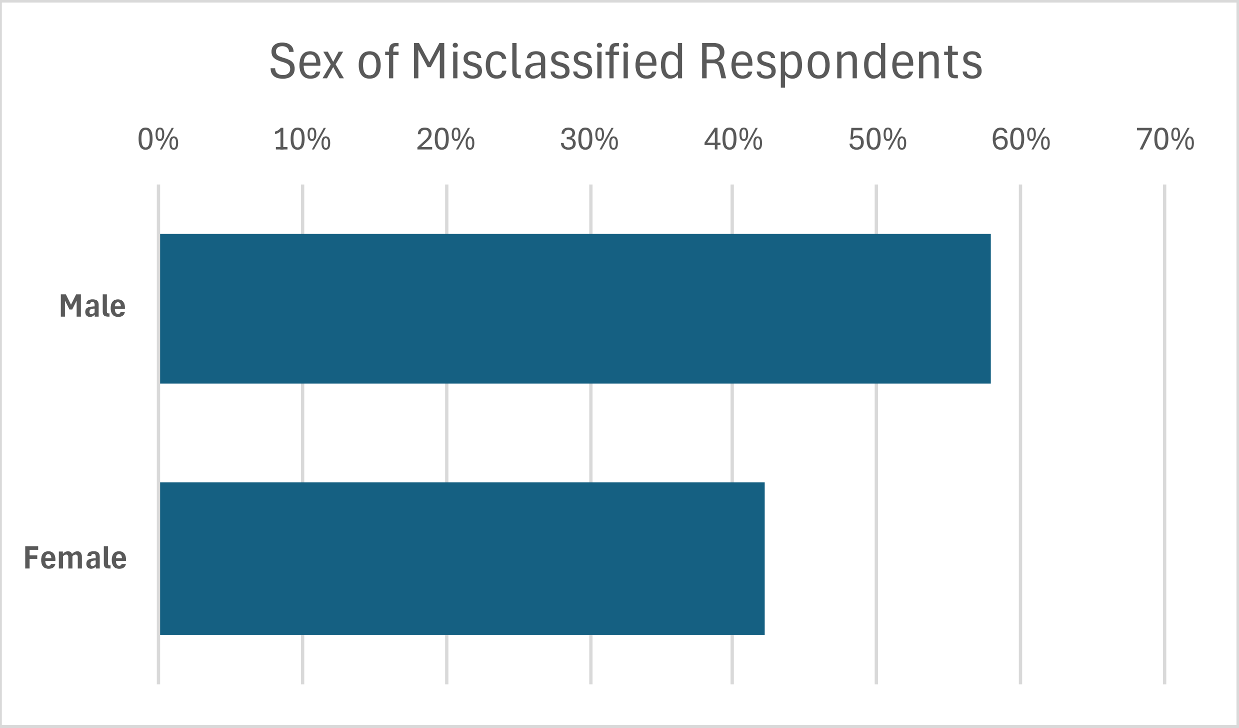
Which people are classified wrongly by our detector?



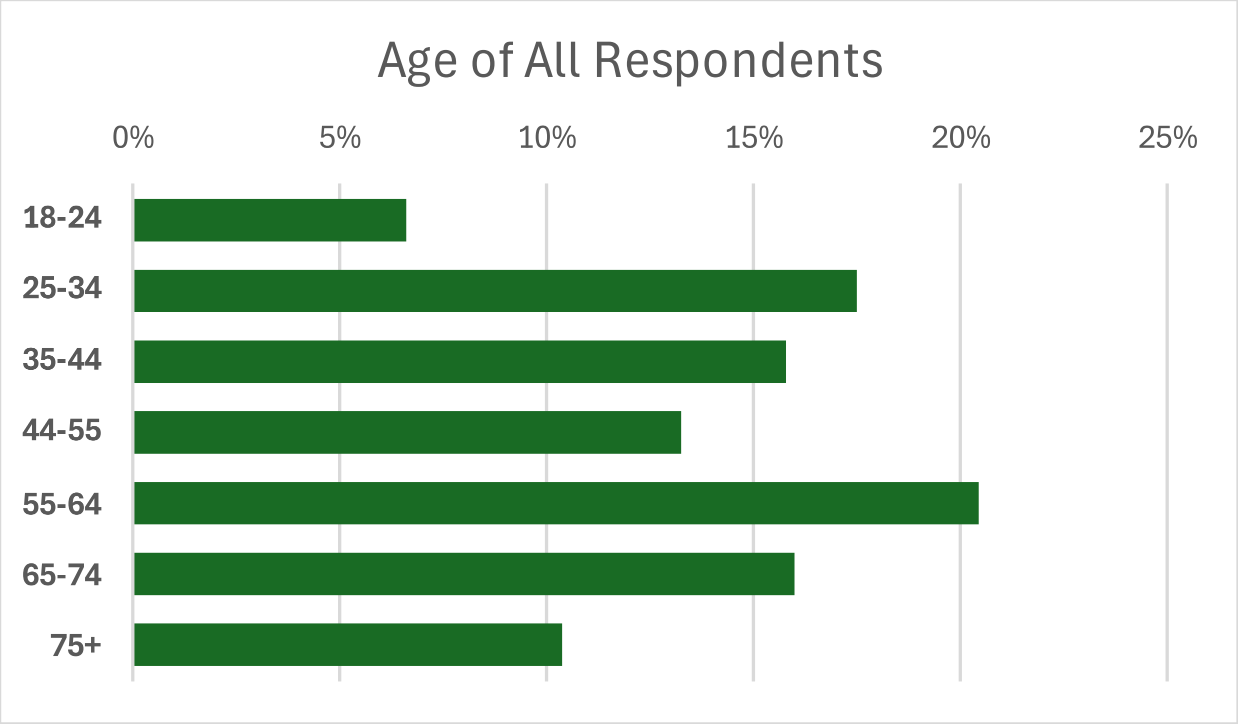


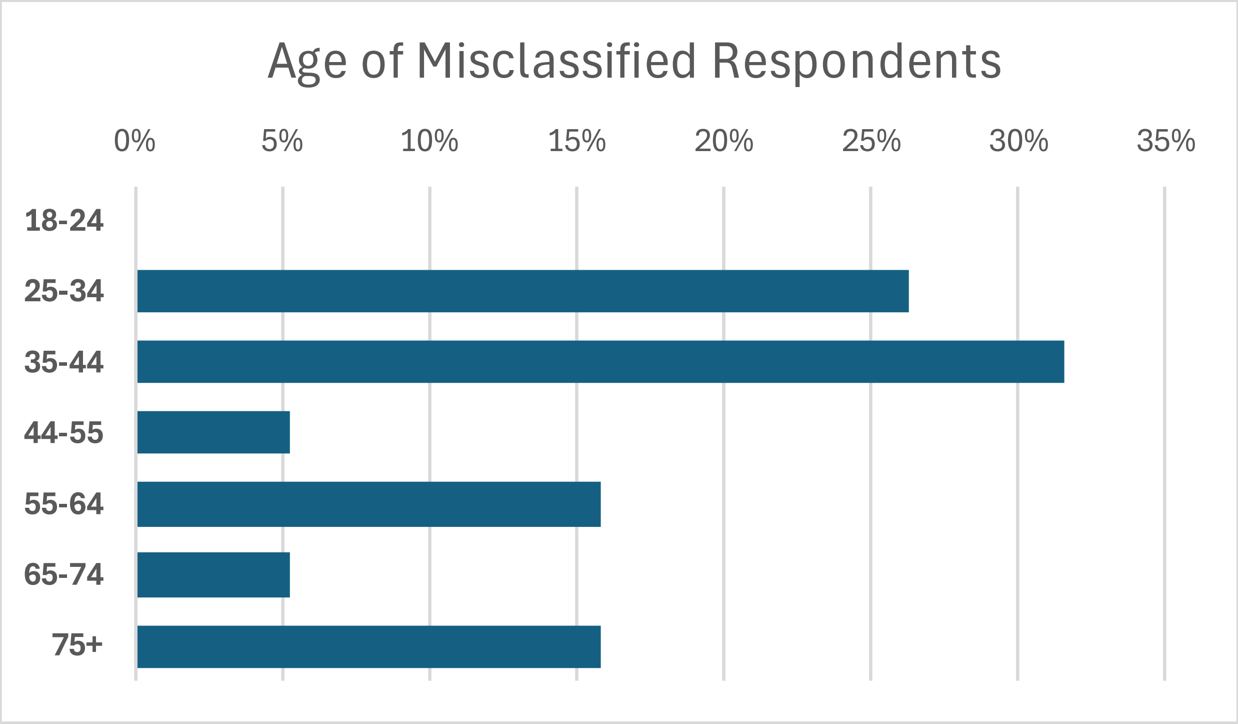
Which people are classified wrongly by our detector?



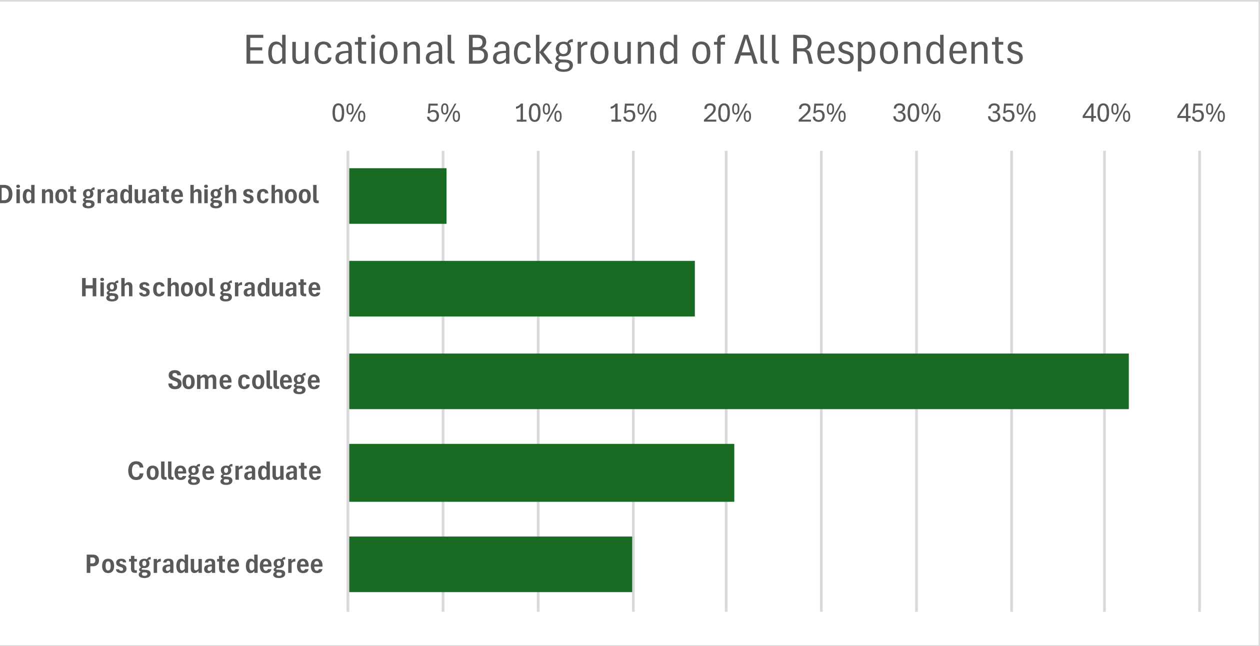


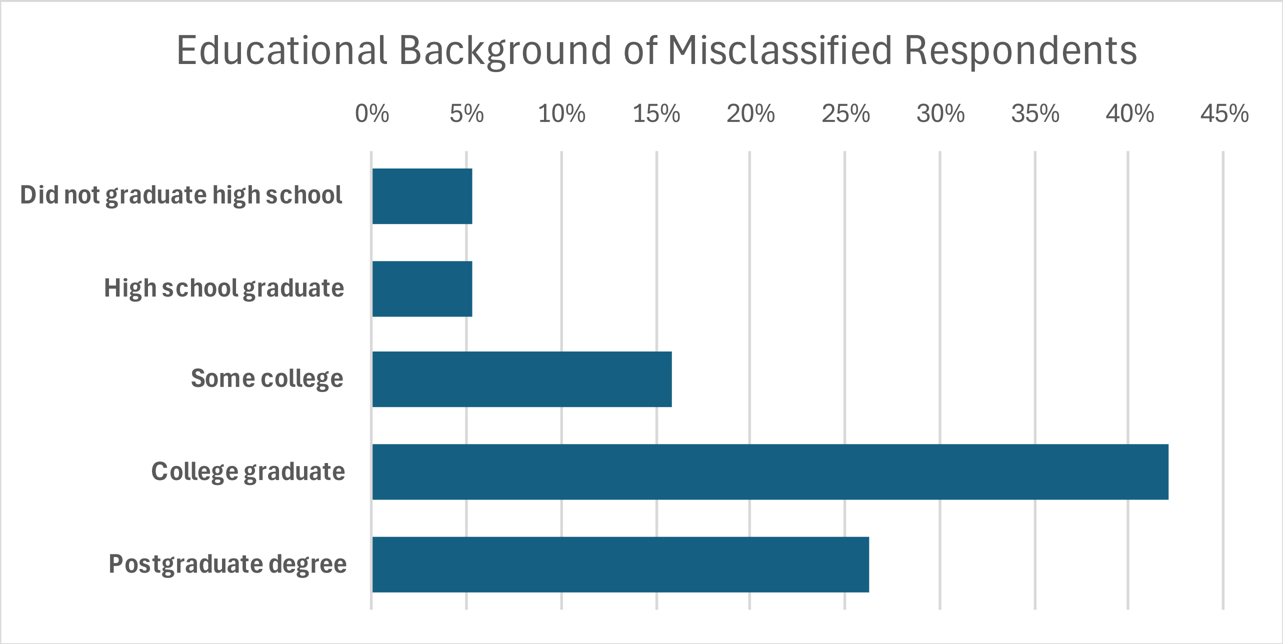
Which people are classified wrongly by our detector?





Which people are classified wrongly by our detector?





What stands out?

- Misclassification rate for **respondents with postgraduate degrees** is double that of other groups
  - Rate is still low (~2% misclassified) but this is a noteworthy discrepancy
- **Textual characteristics of misclassified responses:**
  - Contain **significantly more words** than correctly classified responses
    - Mean of 49.26 words vs 8.67 words; p-value of 0.0136
  - Have **significantly higher reading levels** than correctly classified responses
    - Mean of 23.31 vs 10.26; p-value of 0.0081
  - Have **significantly greater word overlap with the question** than correctly classified responses
    - Mean of 3.37 vs 1.30; p-value of 0.0182

How can we mitigate this and any other identified bias?

- **Training data creation: Class balancing**
  - By collecting more labeled data (survey responses) from such subgroups, we can ensure they are better represented in training data
- **Model development: Data selection**
  - Data Debiasing with Datamodels is a method proposed by Jain et al. (2024), for removing specific training data points that contribute significantly to the model's poor performance on certain subgroups

When is LLM use permissible?

- **For accessibility**
  - If English isn't their first language
  - If they have reading difficulties
- A **nuanced approach** is required
  - Flagging for manual review rather than dropping
  - Supplement to existing metrics for assessing low-quality/fraudulent responses
    - Skipping, straightlining, speeding

Thank you!

Lilian Huang
Statistician
huang-lilian@norc.org

Research You Can Trust™

NORC Research Science