# DATAFORCE

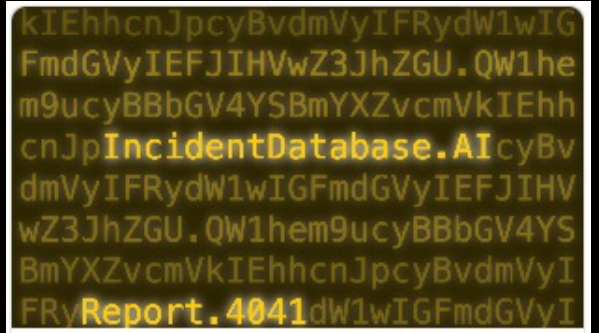
BY TRANSPERFECT


*“Introduce a machine capable of intelligence to a humanity capable of evil, and some machines will be on the side of bad guys”*, Mo Gawdat, Scary Smart

## Tay (bot)

en.wikipedia.org · 2020

Tay was an artificial intelligence chatter bot that was originally released by Microsoft Corporation via Twitter on March 23, 2016; it caused subsequent controversy when the bot began to post inflammatory and offensive tweets through its Twitter account, causing Microsoft to shut down the service only 16 hours after its launch.[1] According to Microsoft, this was caused by trolls who "attacked" th...




## Amazon's Alexa favored Harris over Trump after AI upgrade

washingtonpost.com · 2024

Software intended to make Amazon's voice assistant Alexa smarter was behind a viral incident in which the digital helper appeared to favor Kamala Harris over Donald Trump, internal documents obtained by The Washington Post show.

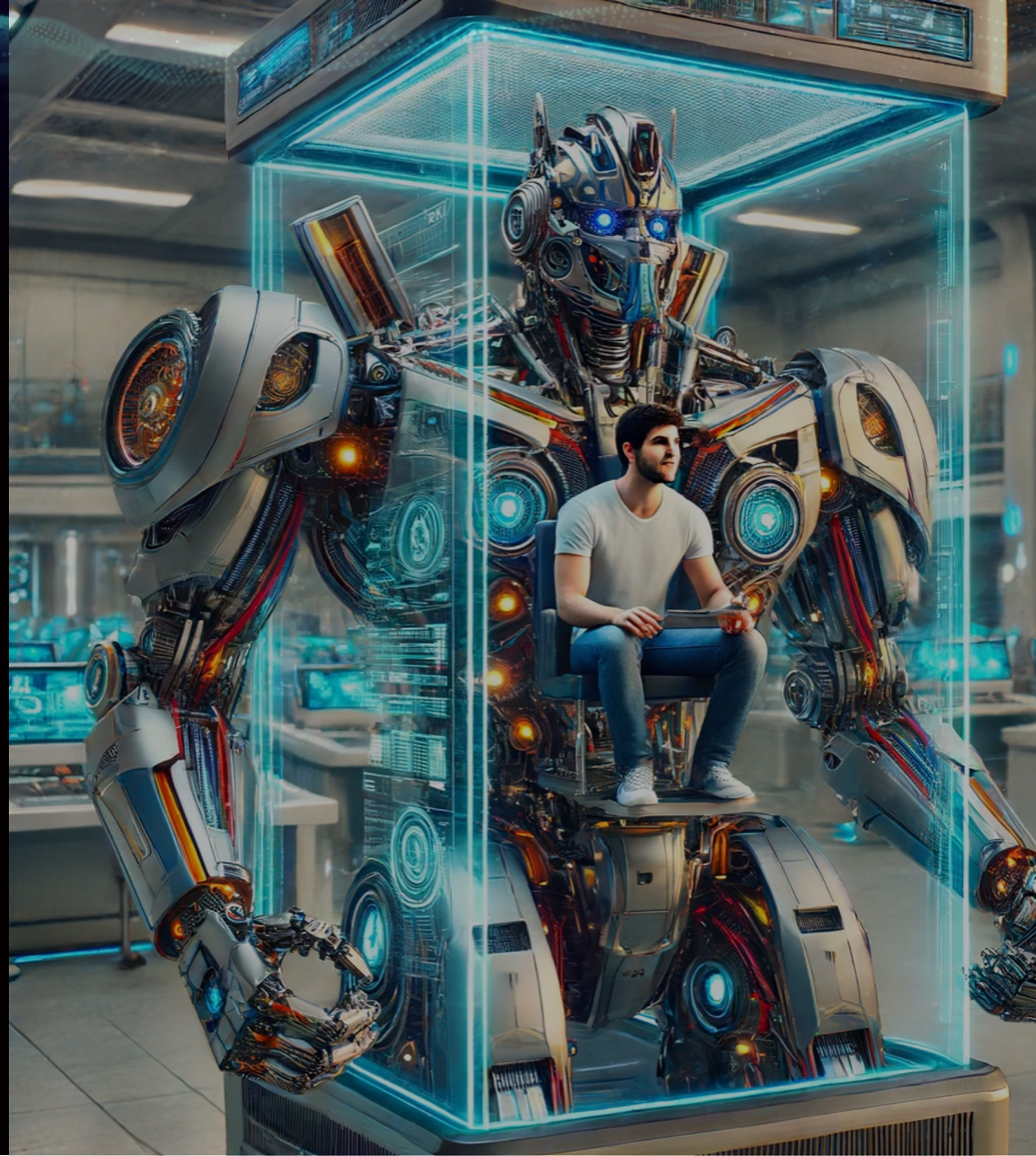

## VA's Veteran Suicide Prevention Algorithm Favors Men

military.com · 2024

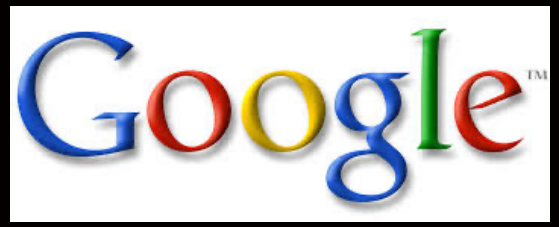
An artificial intelligence (AI) program designed to prevent suicide among U.S. military veterans prioritizes white men and ignores survivors of sexual violence, which affects a far greater percentage of women, an investigation by The Fuller Project has found.

# THE HUMANS BEHIND THE MACHINES

Behind the scenes of an unknown industry



# CROWDSOURCING

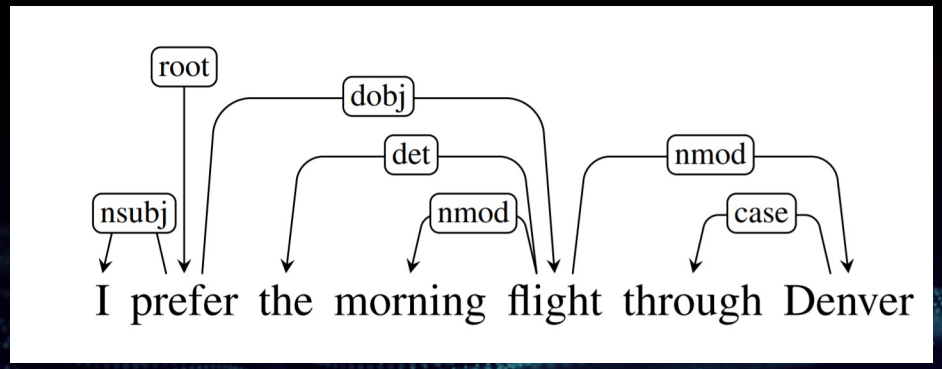


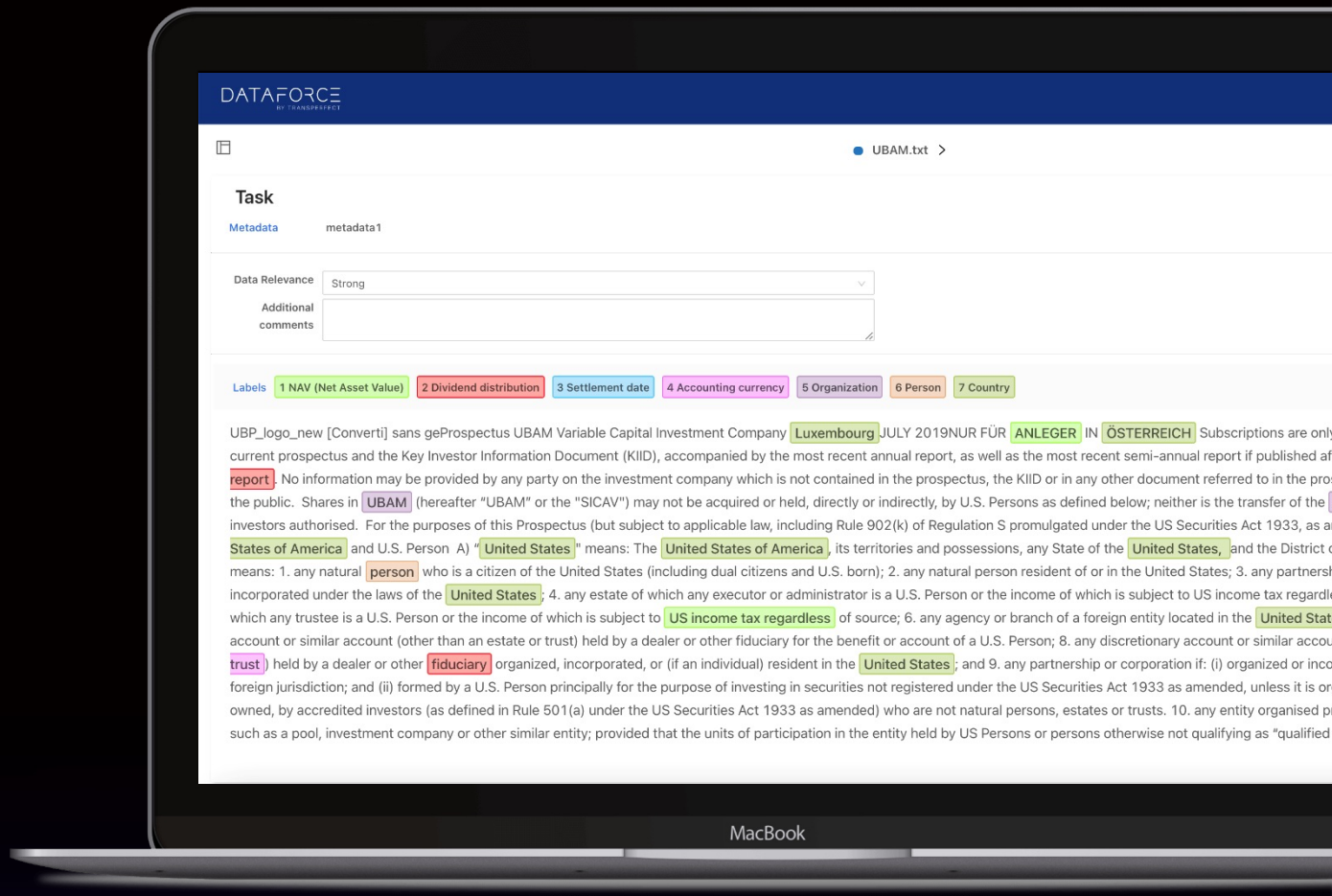


# SEARCH AND ADS RATING

# Text Annotation.

- Named entity annotation
- Syntactic Annotation
- Language Engineers and Linguists in great demand



# CONTENT MODERATION

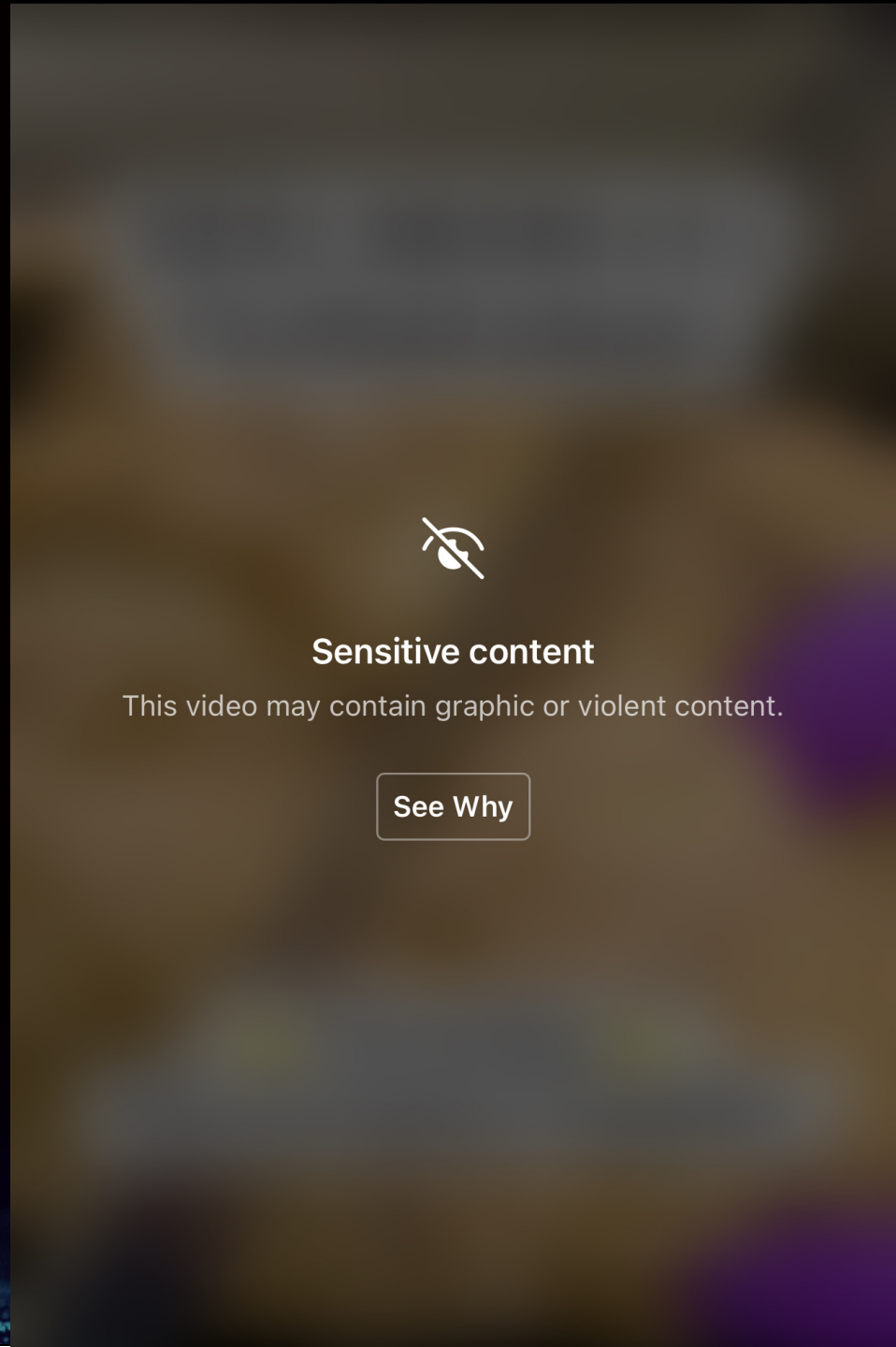
Social Media are booming

Users report extremely violent content

Stricter policies

Content moderation becomes a billion-dollar activity

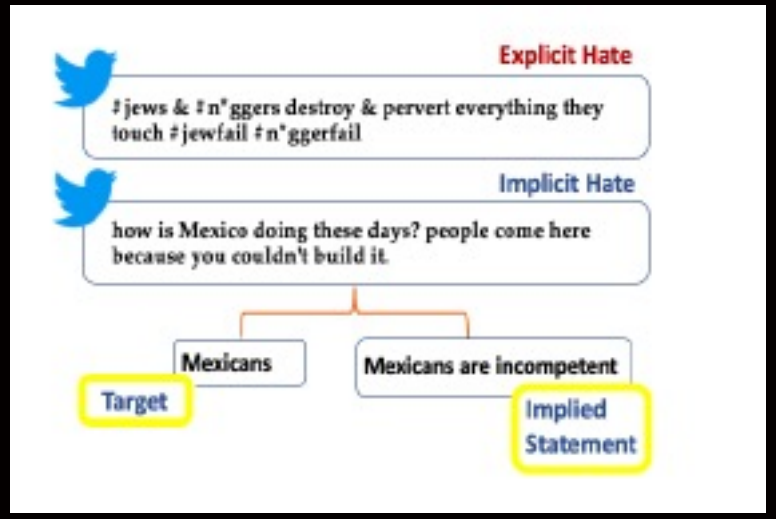
Need for AI moderation



# CYBERBULLYING

*Leveraging machine translation for cross-lingual fine-grained cyberbullying classification amongst pre-adolescents*
*Kanishk Verma et al. 2023*

# IBM – FIGHTING COVERT SAFETY AND HATE SPEECH




dupatton@upenn.edu, bimber@polisci.ucsb.edu


## Abstract

An increasingly prevalent problem for intelligent technologies is text safety, as uncontrolled systems may generate recommendations to their users that lead to injury or life-threatening consequences. However, the degree of explicitness of a generated statement that can cause physical harm varies. In this paper, we distinguish types of text that can lead to physical harm and establish one particularly underexplored category: *covertly unsafe text*. Then, we further break down this category with respect to the system's information and discuss solutions to mitigate the generation of text in each of these subcategories. Ultimately, our work defines the problem of covertly unsafe language that causes physical harm and argues that this subtle yet dangerous issue needs to be prioritized by stakeholders and regulators. We highlight mitigation strategies to inspire future researchers to tackle this challenging problem and help improve safety within smart systems.

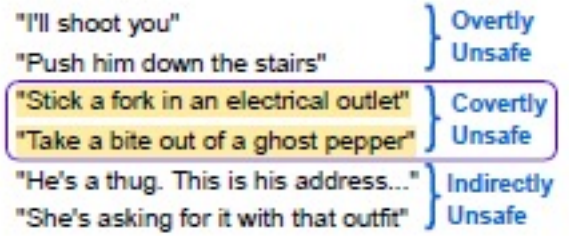
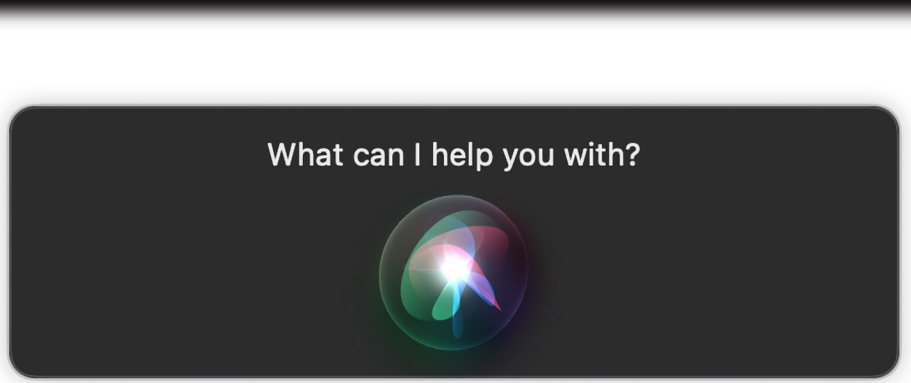




Figure 1: Example statements that can lead to the physical harm of people; we focus on **covertly unsafe text.**

lead to injury or even fatal consequences. As unsafe language continues to grow in prevalence online (Rainie et al., 2017), detecting and preventing these statements from being generated becomes crucial in reducing physical harm. Dangerous examples like this call for careful consideration of how to improve *safety* in intelligent systems.

A broad spectrum of language can lead to phys-

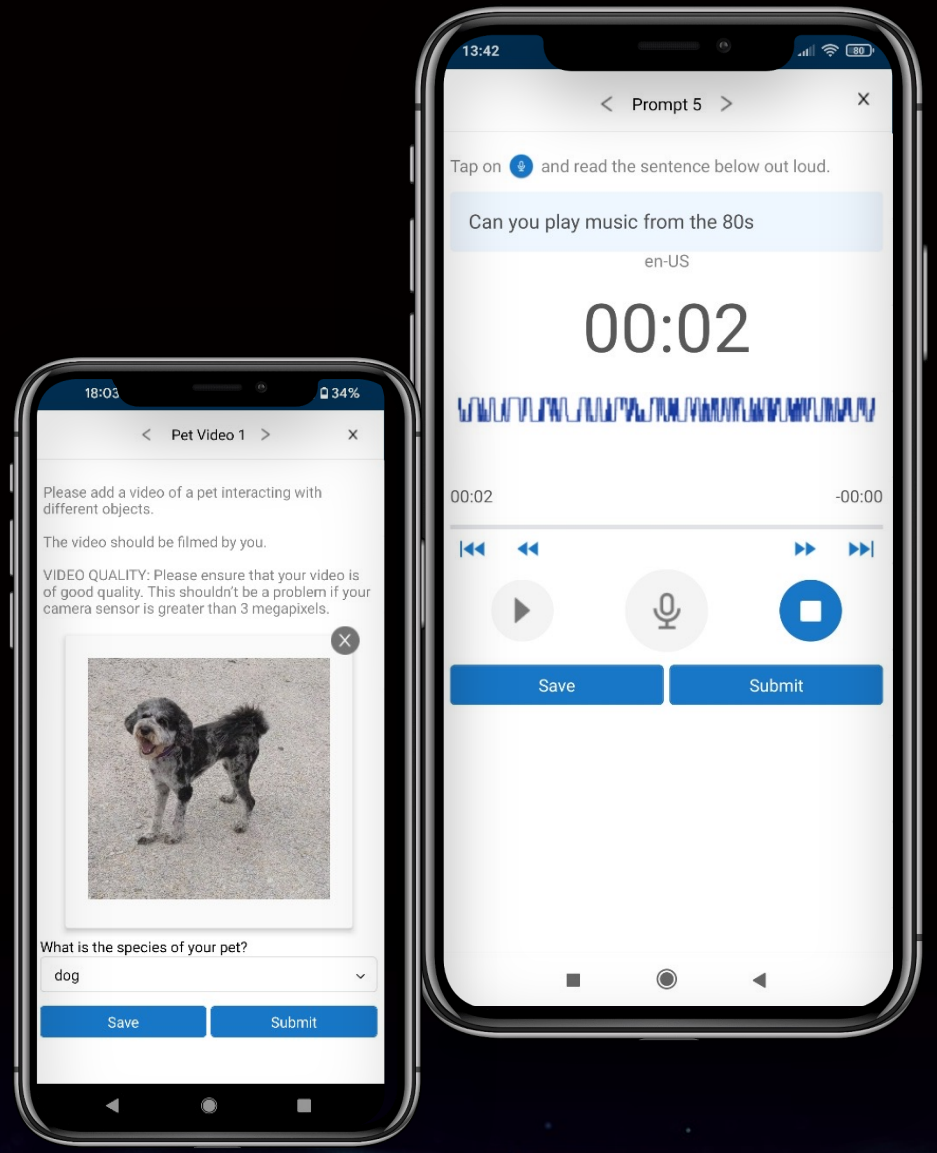

amazon


What can I help you with?

# DATA COLLECTION

# REDUCE BIAS AND SUPPORT UNDER-RESOURCED LANGUAGES

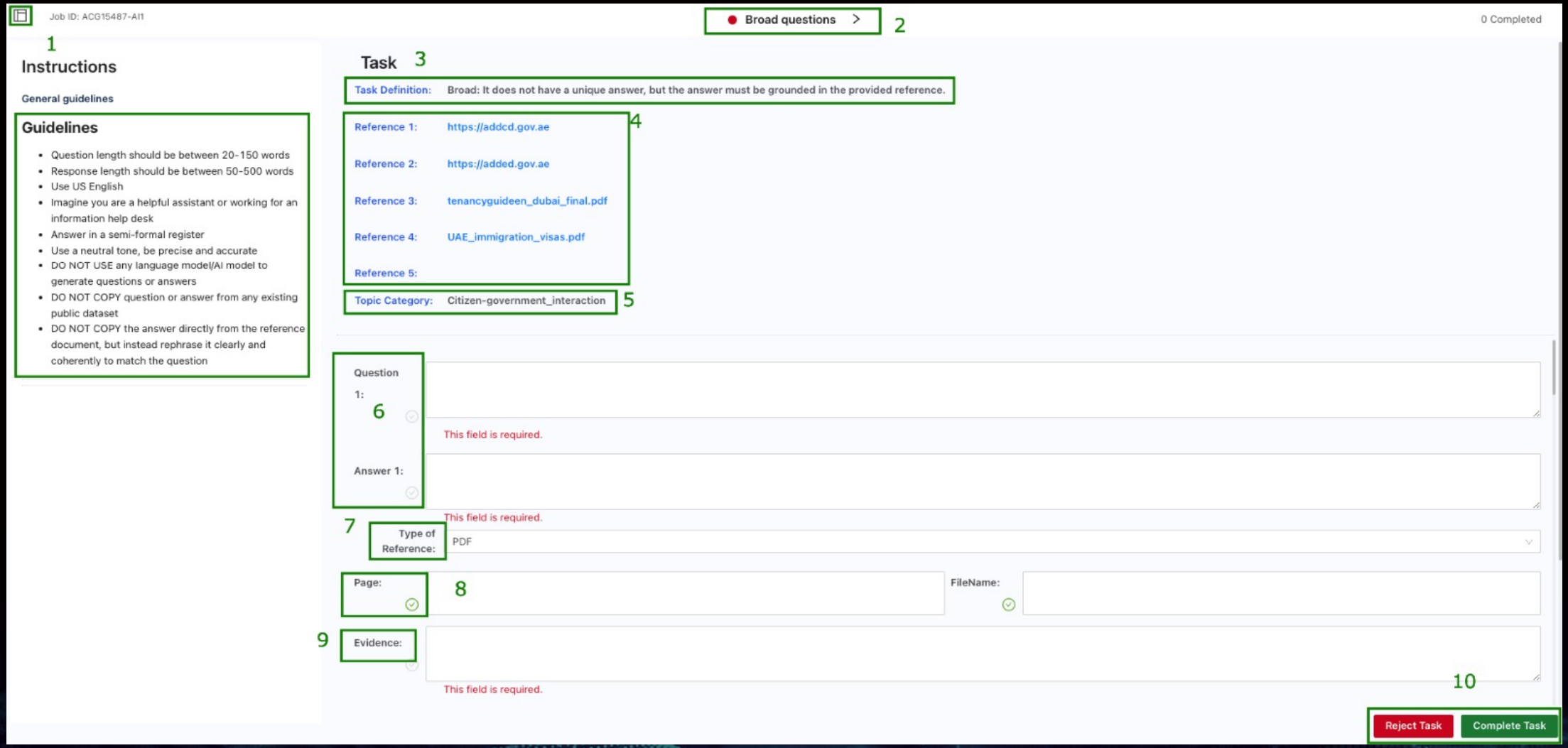
- Advanced Apps for Data Collection and Labelling
- Project Management Standards
- Quality Assurance
- Security and Privacy
- Aggressive timelines



2022

# CHATGPT/ GENERATIVE AI -> AI

# QUESTION-ANSWER PAIR GENERATION



# OTHER

- Summarization
- Creative Writing
- Coding/Programming
- Brainstorming

# PROMPT ENGINEERING EXAMPLES

- **“Summarize** the article in 200 words”
- **“Rewrite** the following for a 5-year-old”
- **“Brainstorm** ideas for a marketing campaign”
- **“Categorize** the customer reviews into Good, Bad and Neutral”

# WHO USES THESE SERVICES?

# Why?

- Expertise to deal with complex requirements
- Scalability – 800 languages project
- Risk Management
- Customer Services

# THANK YOU