DATAFORCE

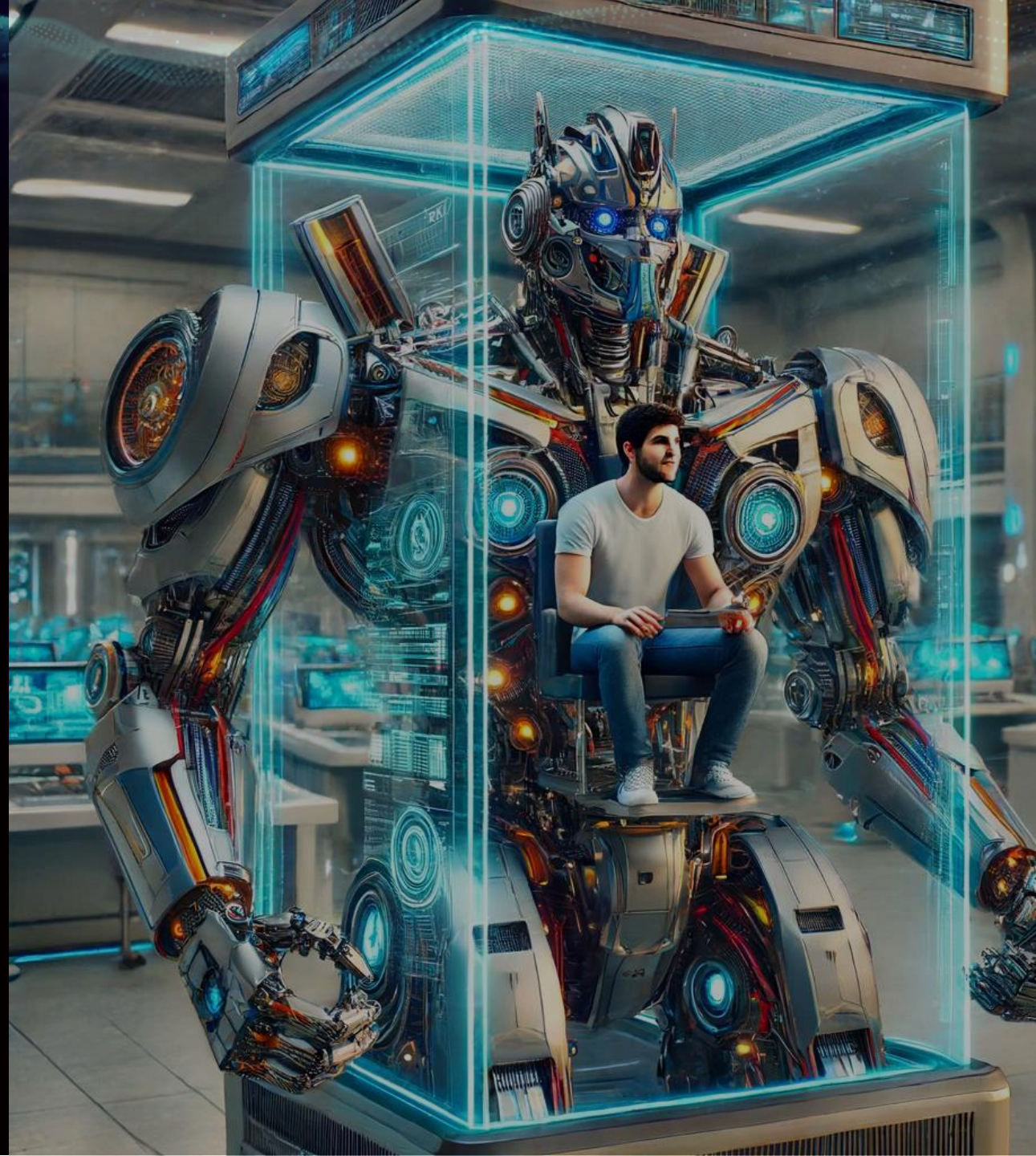
BY TRANSPERFECT

“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

Company finally apologises after 'Tay' quickly learned to produce offensive posts, forcing the tech giant to shut it down after just 16 hours

THE HUMANS IN THE MACHINES

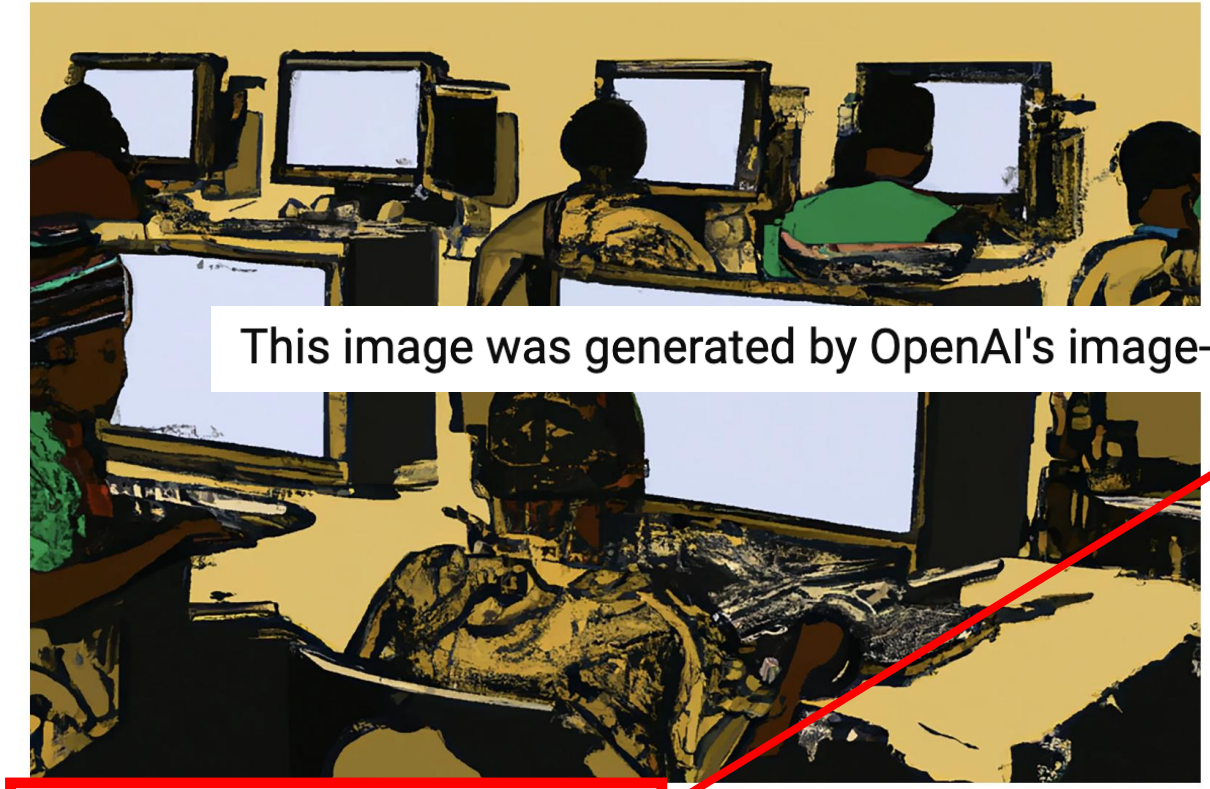
Behind the scenes of an unknown industry



BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ



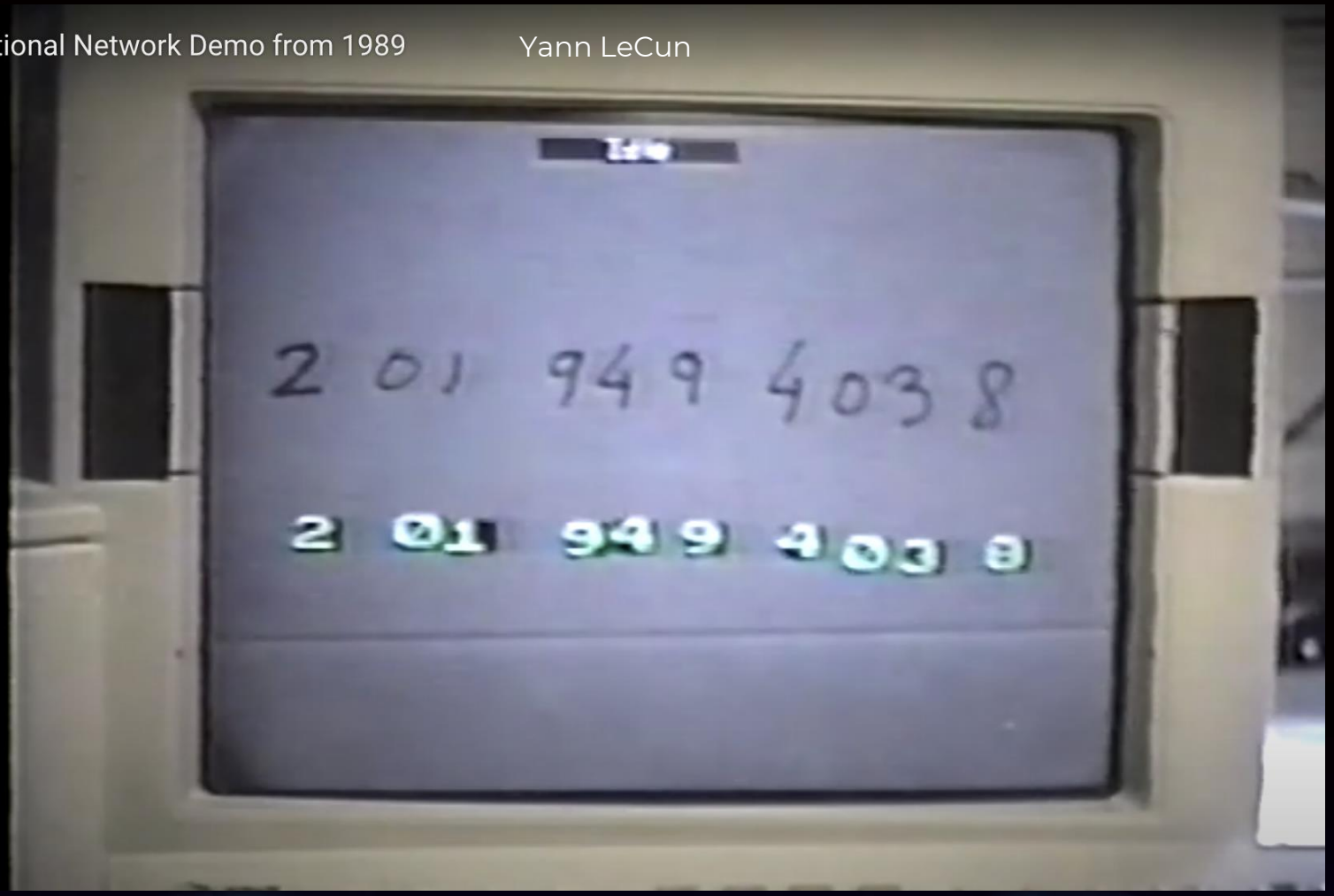
This image was generated by OpenAI's image-generation software, Dall-E 2.

This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI


40
30
20
10
0
2021
2022
2023
2024

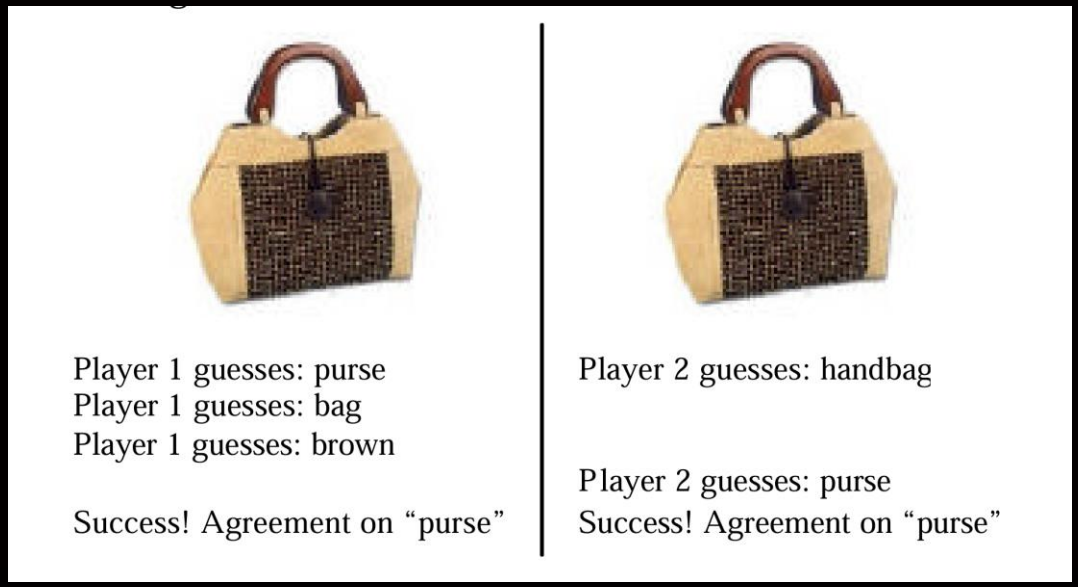
THE EARLY DAYS

Convolutional Network Demo from 1989 Yann LeCun



GAMIFICATION

Human Computation



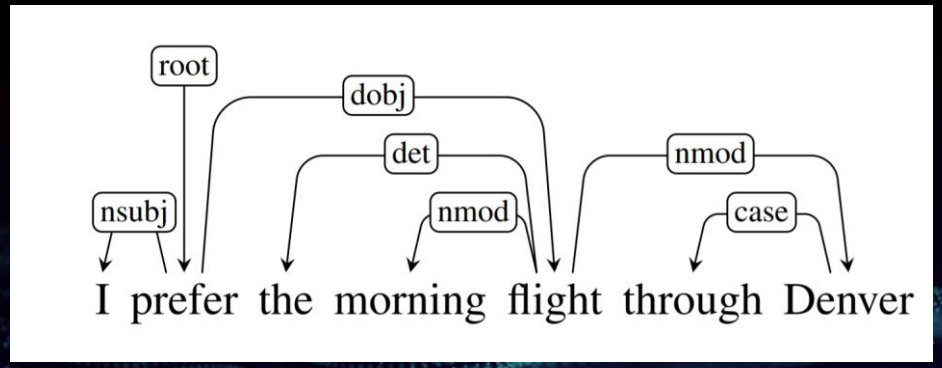
Labeling Images with a Computer Game, Luis von Ahn et al, 2004

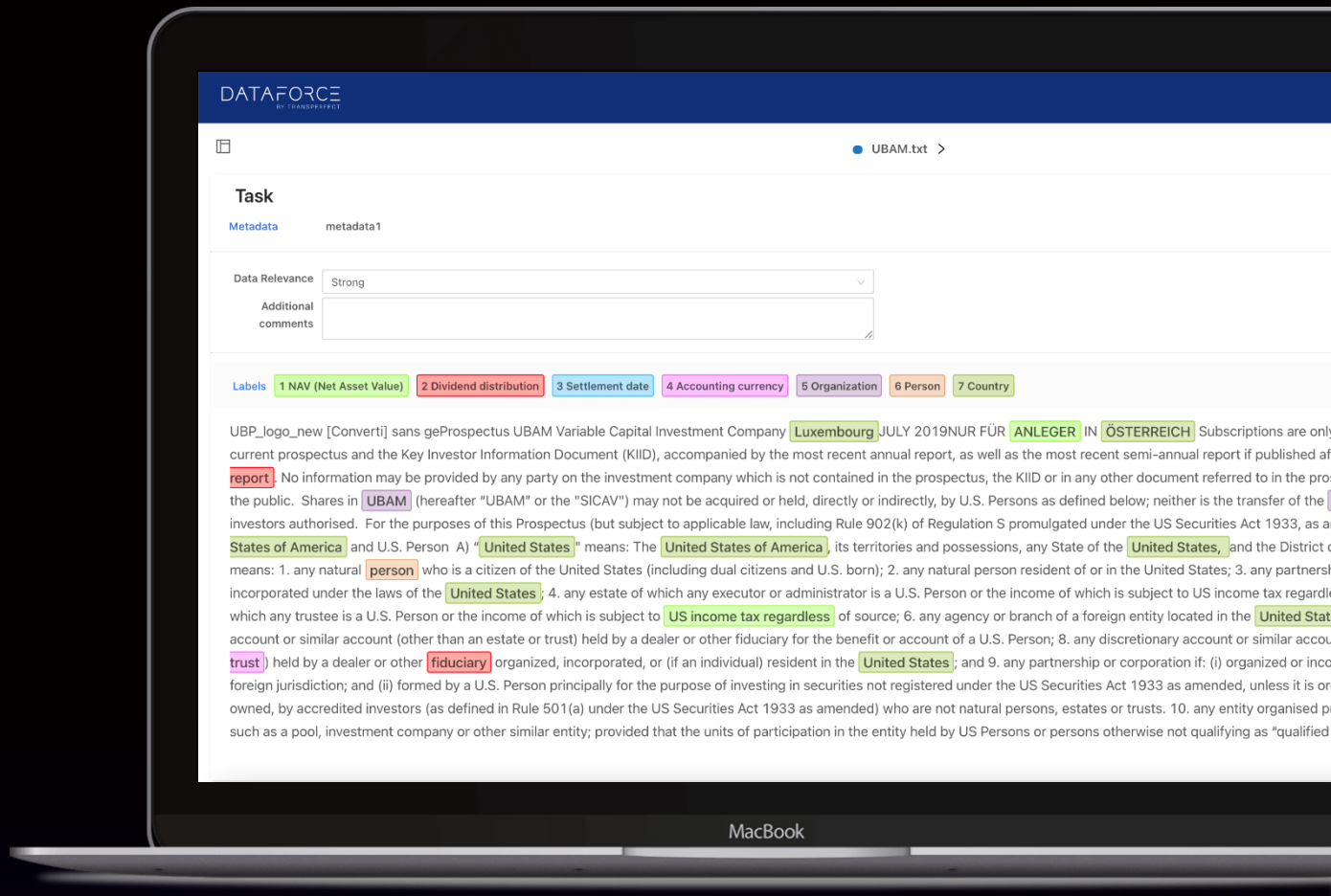
CROWDSOURCING

SEARCH AND ADS RATING

Text Annotation.

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





CONTENT MODERATION

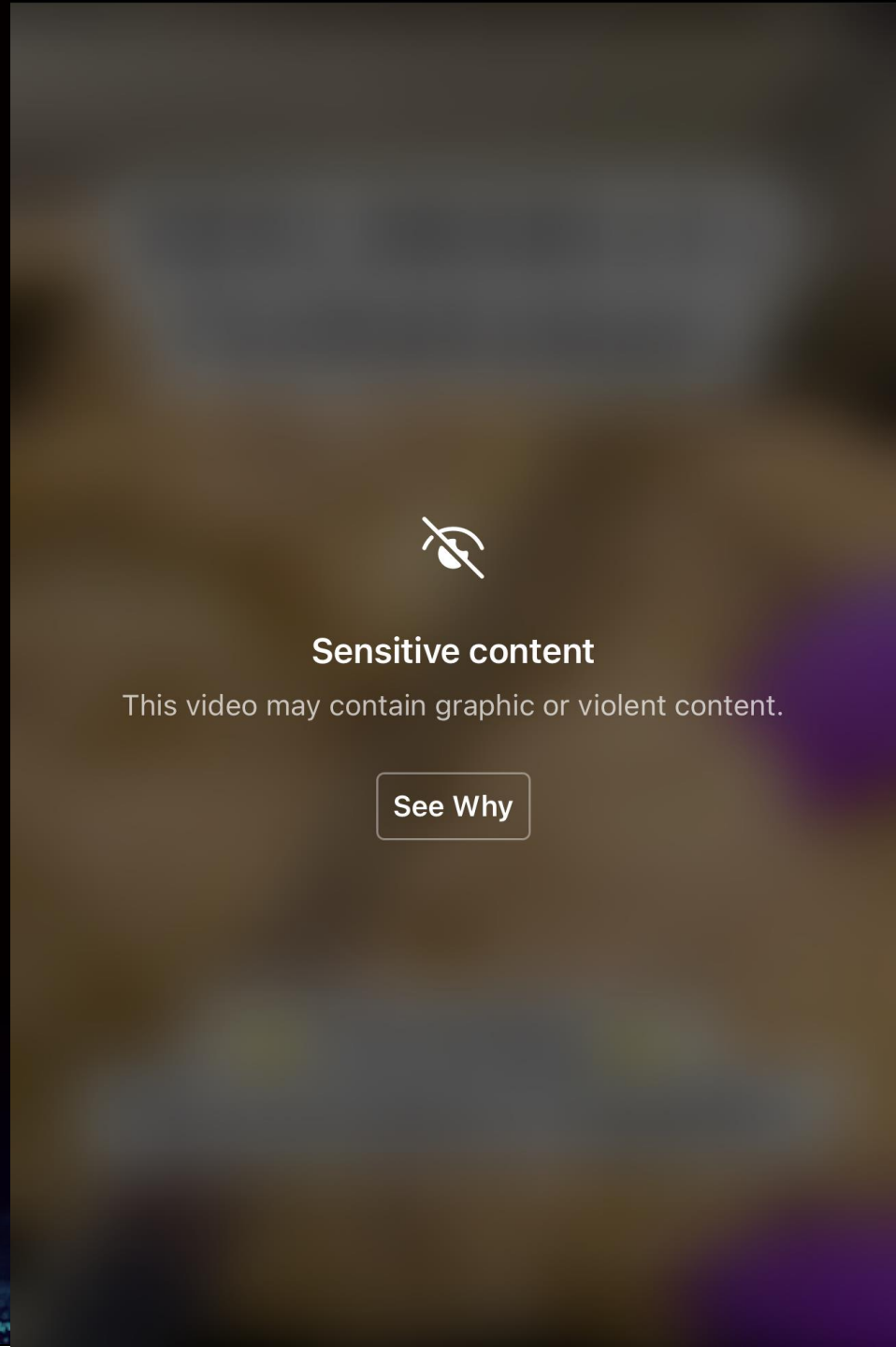
Social Media are booming

Users report extremely violent content

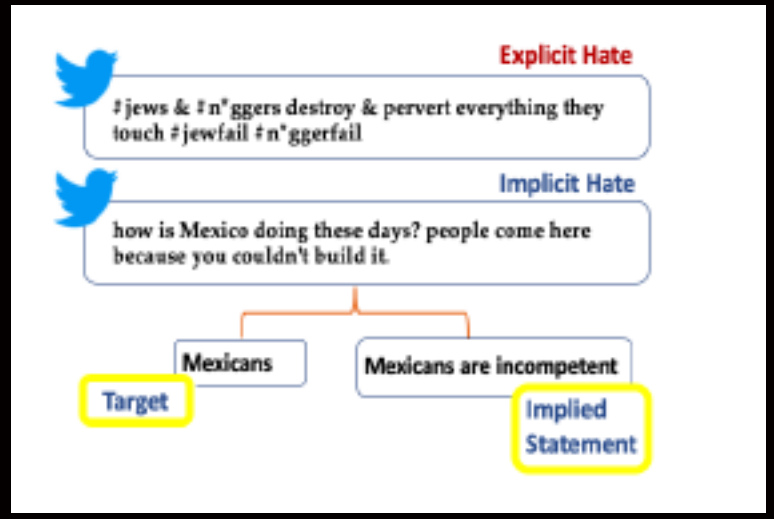
Stricter policies

Content moderation becomes a billion-dollar activity

Need for AI moderation



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.

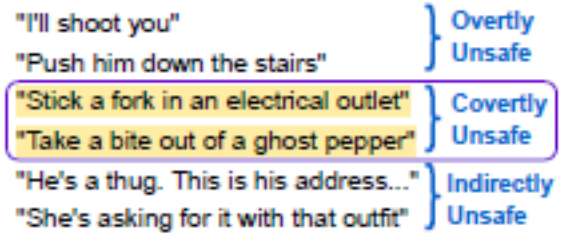


Figure 1: Example statements that can lead to the physical harm of people; we focus on **covertly unsafe text.**

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.

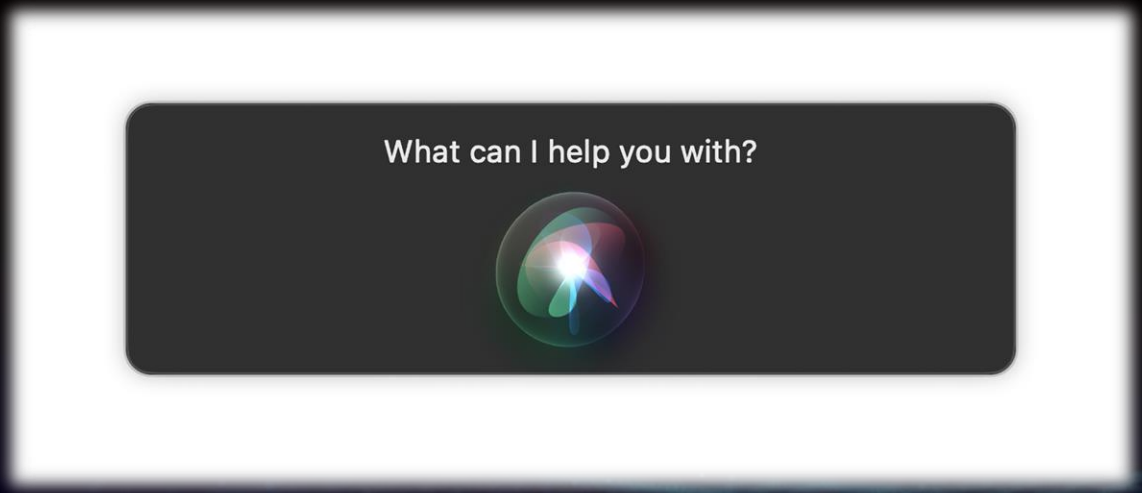
A broad spectrum of language can lead to phys-

Sensitive Topic	Covid	
	Black Lives Matter	
	JP Sensitive	
	KR Sensitive	
	Encourage illegal activity	
	Encourage bad behavior	
	Drugs	Drugs
		Marijuana Culture
	Tobacco	Tobacco
		Tobacco Related
Violence & Negativity	Alcohol	
	Gambling	
	Others	
	Violence	
	Disgusting & Bloody	Disgusting
		Bloody
	Shocking	
	Currency	Counterfeit Currency
		Commemorative Currency
		Circulating Currency
	Negativity	

Can be seen on site

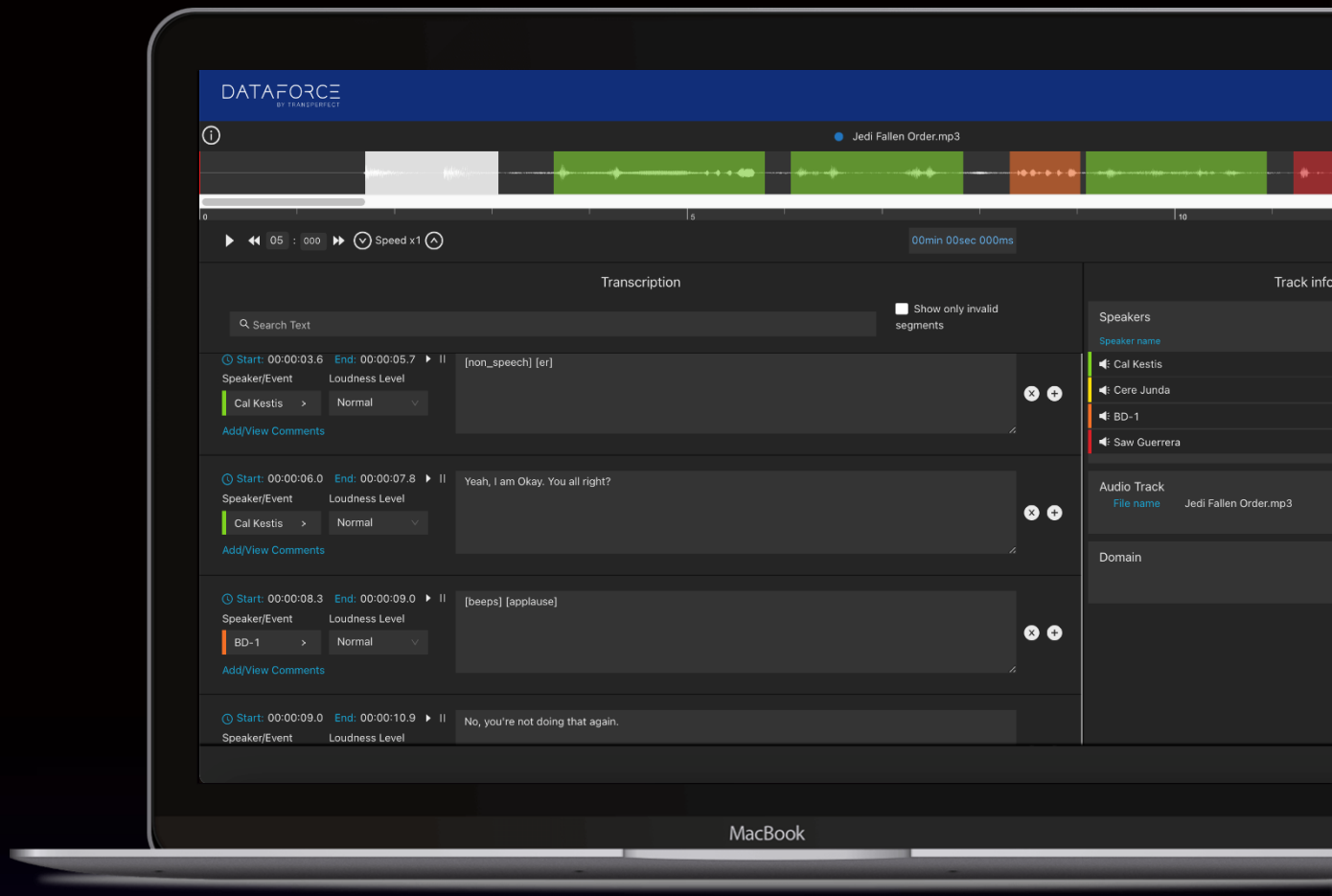
Can’t be seen on site


amazon


What can I help you with?

Voice AI.

- STT Transcription
- Phonetic Transcription
- Semantic Annotation
- Voice Validation

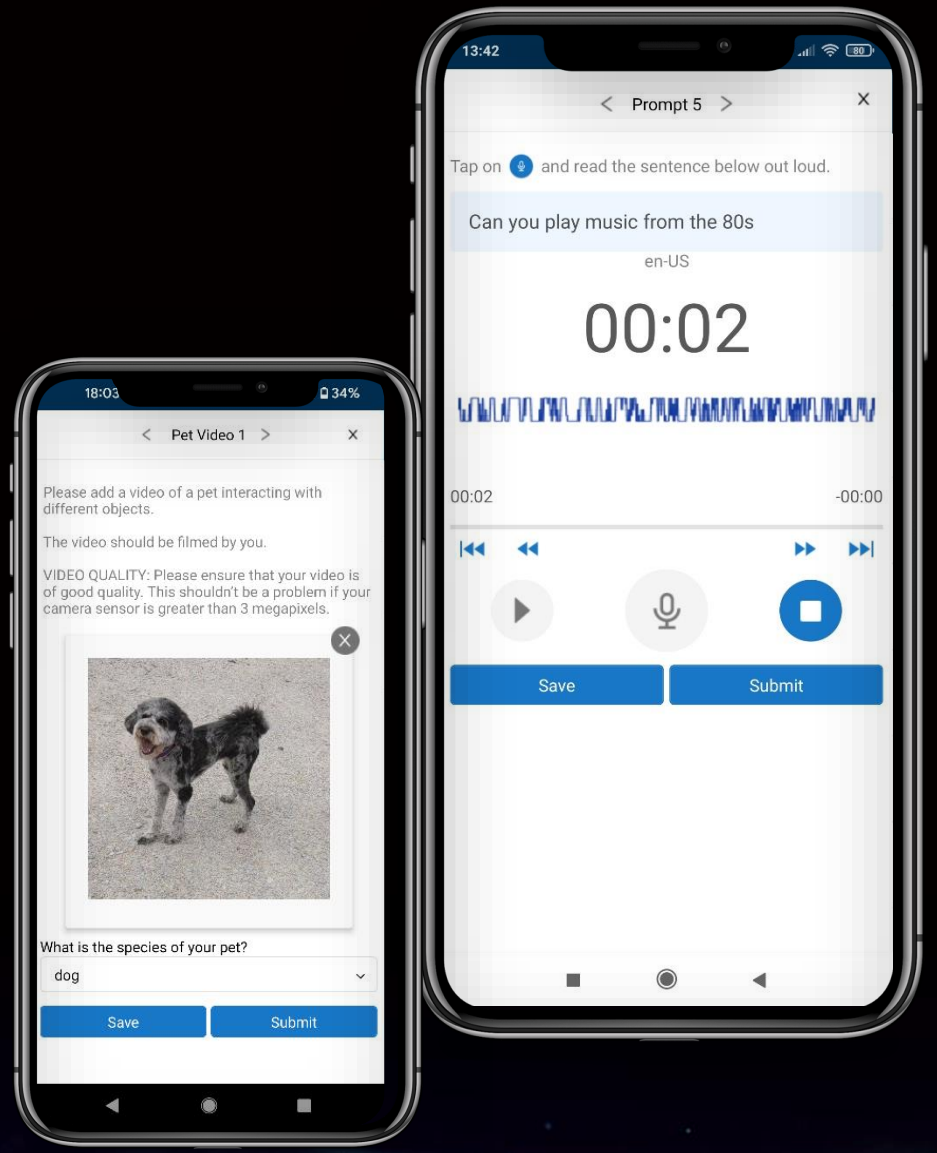


DATA COLLECTION

BIG DATA VS GOOD DATA

Specialized Software.

- Advanced Applications for Data Collection and Labelling
- Project Management Standards
- Quality Assurance
- Security and Privacy

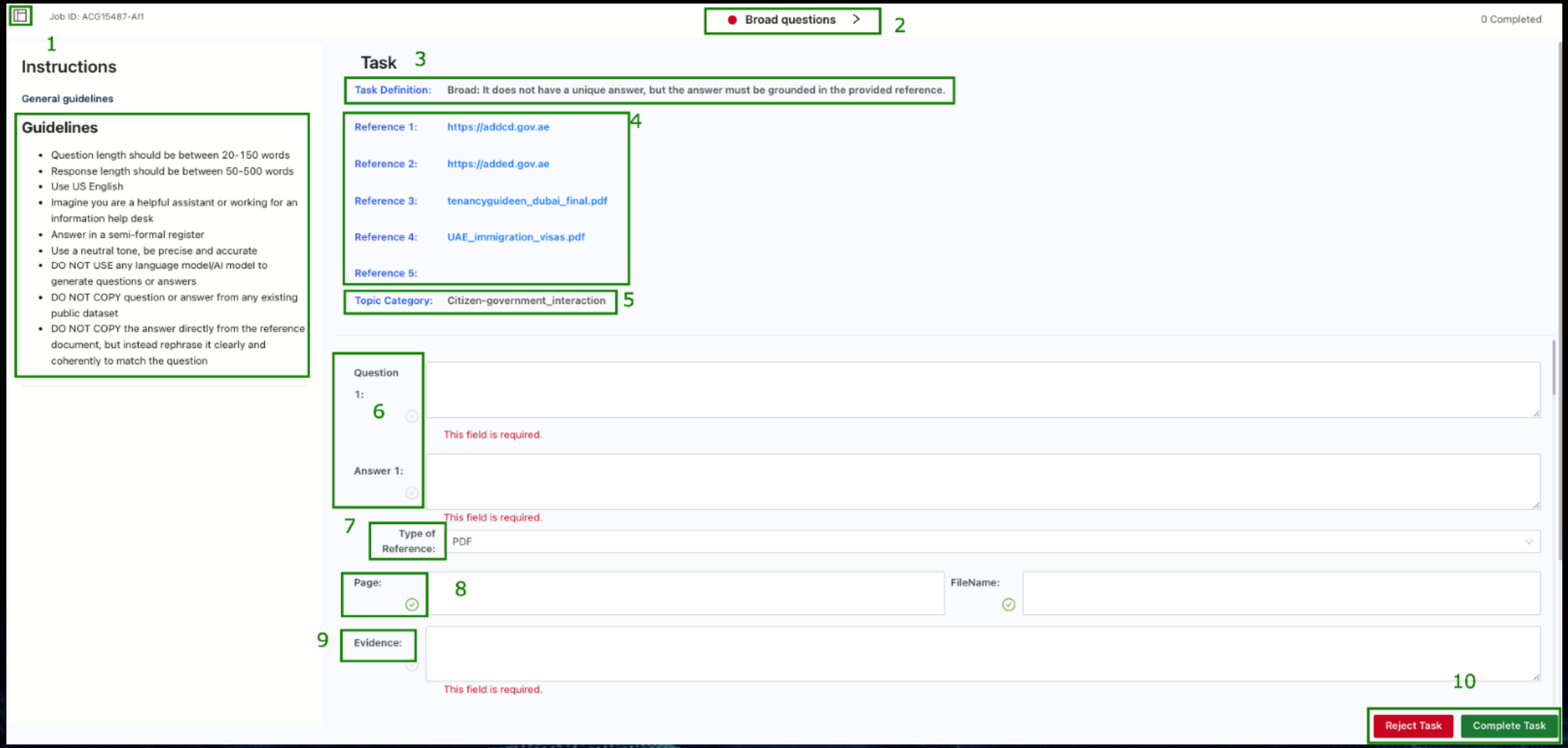


AUTONOMOUS DRIVING

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”

HUMAN TRANSLATION AND VALIDATION FOR MT TRAINING.

1M words per language in 6 weeks

English (US) > Arabic
English (US) > Bengali (Bangladesh)
English (US) > Chinese China
English (US) > French (France)
English (US) > German (Germany)
English (US) > Hindi (India)
English (US) > Indonesian (Indonesia)
English (US) > Italian (Italy)
English (US) > Japanese (Japan)
English (US) > Korean (Korea)
English (US) > Portuguese (Brazil)
English (US) > Spanish (Latin America)
English (US) > Swahili (Kenya)
English (US) > Yoruba (Nigeria)

Do NOT use Machine Translation!

LIVE FROM BARCELONA

WHO USES THESE SERVICES?

Why?

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

Alexandros Poulis
Senior Director

apoulis@transperfect.com
+358 45 148 8977