eduworks

Threats and Opportunities for Leveraging LLMs

Eduworks Corporation

BENGAL Proposers Day

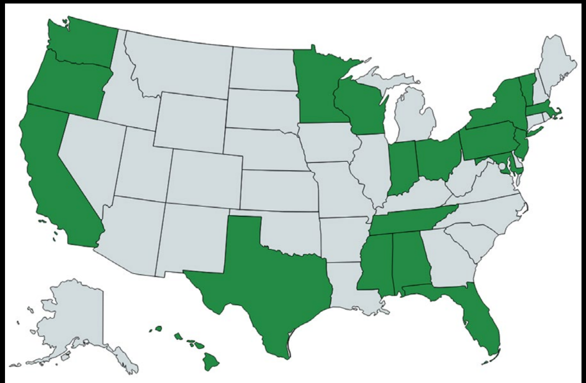
eduworks™

Company Summary

- Small U.S. business est. 2001
- Core capabilities in AI, Machine Learning
- Serving multiple USG clients in DOD, DHS, NSF
- Significant experience designing, building, testing and evaluating LLMs
- Research into building and deploying AI for DARPA
- Cutting-edge NSF work on LLM bias, threat and vulnerability detection and mitigation

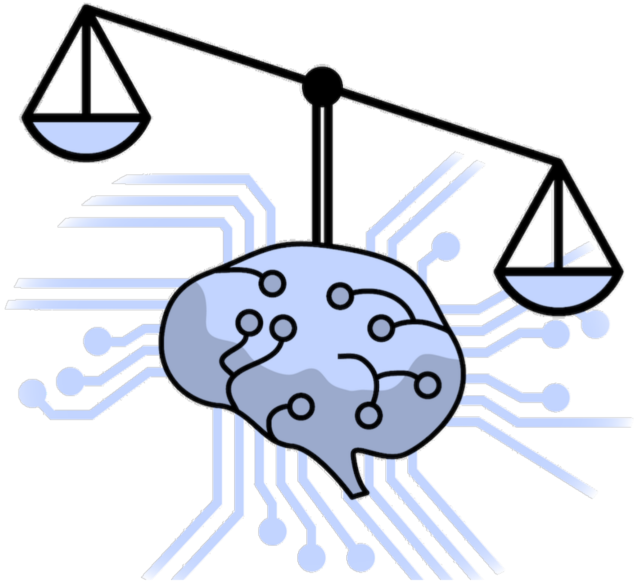
INTERSERVICE/INDUSTRY TRAINING, SIMULATION & EDUCATION CONFERENCE
AIR FORCE
ARMY
NAVY USMC
INDUSTRY
'22
BEST PAPER
PSMA
INTERSERVICE/INDUSTRY TRAINING, SIMULATION & EDUCATION CONFERENCE
NAVY USMC
AIR FORCE
ARMY
INDUSTRY
'21
BEST PAPER
I/ITSEC

SIMULATION & TRAINING AWARDS 2022
Recognizing the Global Simulation & Training Defense Community
FINALIST
Presented by mst



Goals

- Rapidly summarize/contextualize information relevant to IC
- Avoid biases and toxic or erroneous outputs
- Create novel approaches to address LLM vulnerabilities



Threats

- Injection attacks while training LLMs: nearly impossible to detect with 100% certainty

- Inference attacks: can be mitigated with extensive testing, but full protection requires human in the loop

- LLMs trained on US/Western data will have biases that cloud responses to foreign threats

- Hallucinations, especially from COTS LLMs, will produce responses that appear legitimate but are entirely falsified

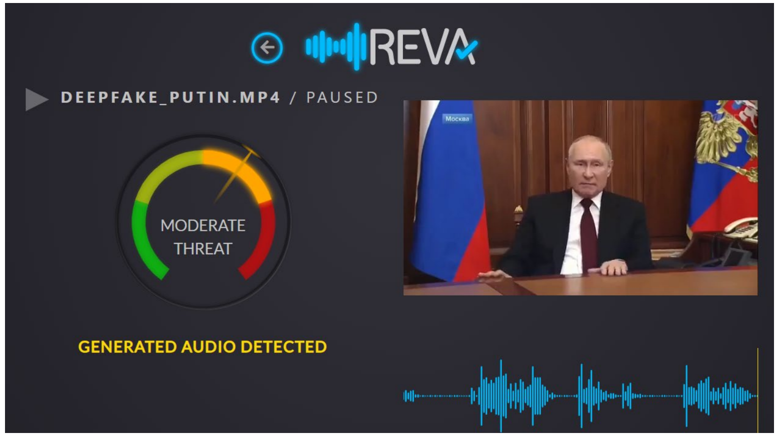
Current DARPA Work: Countering Deepfakes

Summary

- Developed by Eduworks for DARPA under SBIR 2017-2024
- Detects GAN-generated speech deepfakes in real time
- Can be used forensically to analyze large media collections
- Uses ML ensemble trained on a large, varied dataset
- Robust to degraded audio conditions (telephone, VOIP)

Benefits

- Gives analysts better SA than vulnerable legacy techniques
- API can readily provide REVA services to any client system
- Powerful job aid to rapidly triage collected media
- Localization, Diarization, Topic Extraction, Language ID

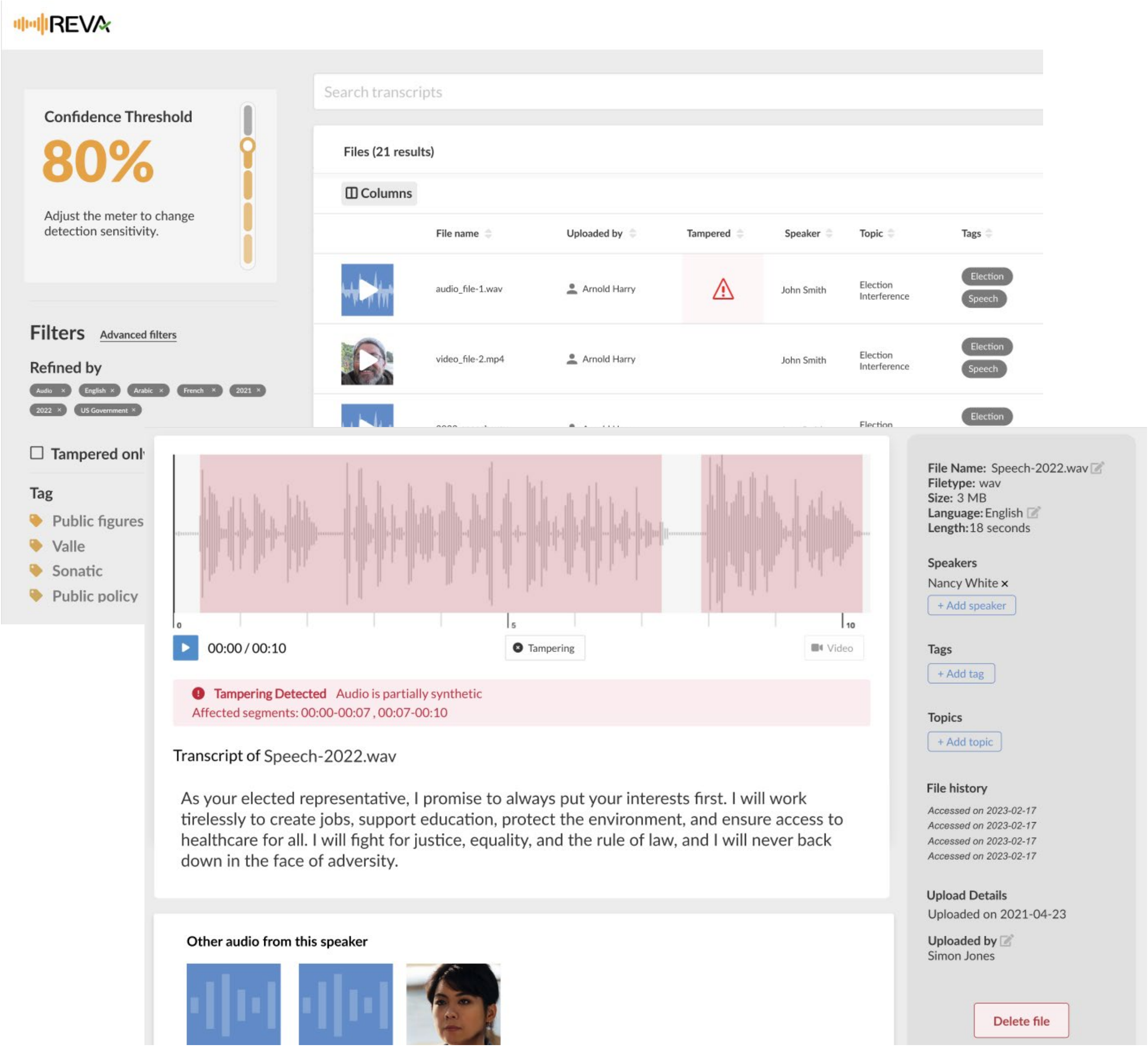


REVA

DEEPFAKE_PUTIN.MP4 / PAUSED

MODERATE THREAT

GENERATED AUDIO DETECTED



REVA

Search transcripts

Confidence Threshold

80%

Adjust the meter to change detection sensitivity.

Filters Advanced filters

Refined by

Audio × English × Arabic × French × 2021 × 2022 × US Government ×

☐ Tampered onl

Tag

- Public figures
- Valle
- Sonatic
- Public policy

Files (21 results)

Columns

	File name	Uploaded by	Tampered	Speaker	Topic	Tags
	audio_file-1.wav	Arnold Harry	⚠	John Smith	Election Interference	Election Speech
	video_file-2.mp4	Arnold Harry		John Smith	Election Interference	Election Speech
					Election	Election

00:00 / 00:10

Tampering

Video

Tampering Detected Audio is partially synthetic

Affected segments: 00:00-00:07 , 00:07-00:10

Transcript of Speech-2022.wav

As your elected representative, I promise to always put your interests first. I will work tirelessly to create jobs, support education, protect the environment, and ensure access to healthcare for all. I will fight for justice, equality, and the rule of law, and I will never back down in the face of adversity.

Other audio from this speaker

File Name: Speech-2022.wav
Filetype: wav
Size: 3 MB
Language: English
Length: 18 seconds

Speakers

Nancy White ×

+ Add speaker

Tags

+ Add tag

Topics

+ Add topic

File history

Accessed on 2023-02-17
Accessed on 2023-02-17
Accessed on 2023-02-17
Accessed on 2023-02-17

Upload Details

Uploaded on 2021-04-23

Uploaded by

Simon Jones

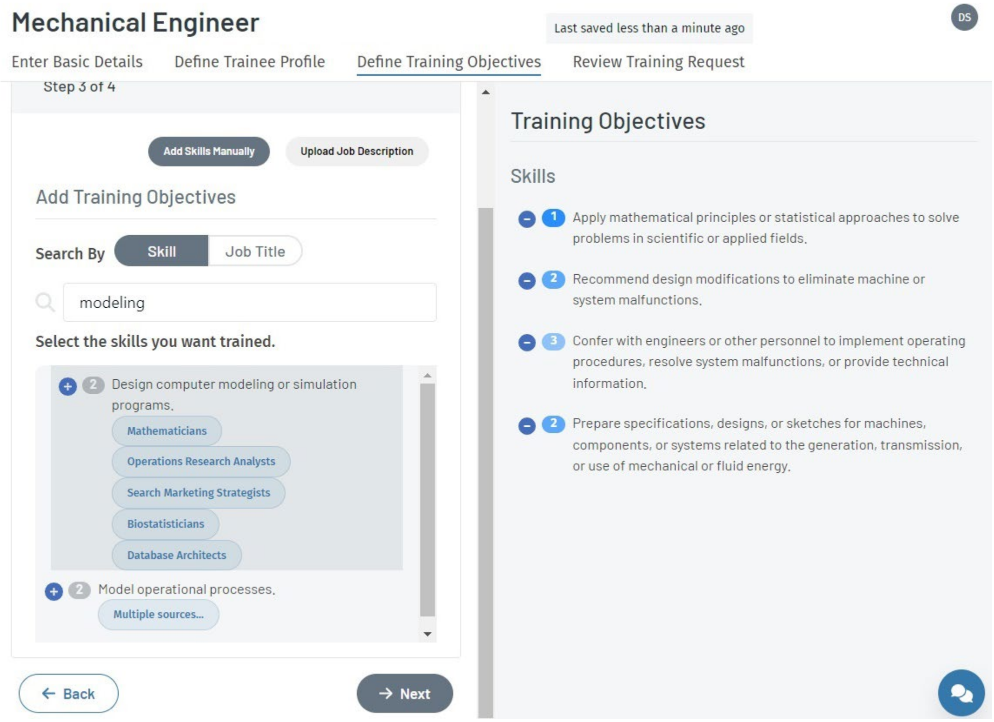
Delete file

Current NSF Work: Safe LLMs

Summary

- AI alignment of skills with job postings
- Extracts skills and relationships in unstructured text.
- Prioritized list of most relevant skills for a job title or description.
- Scores alignment between KSAs and a course materials.
- Finds and replaces Company Identifiable Information (CII)
- Developing de-biasing techniques for LLM training and tuning

skillsync



Mechanical Engineer

Last saved less than a minute ago

Enter Basic Details | Define Trainee Profile | Define Training Objectives | Review Training Request

Step 3 of 4

Add Training Objectives

Search By: Skill | Job Title

modeling

Select the skills you want trained.

2 Design computer modeling or simulation programs.
Mathematicians
Operations Research Analysts
Search Marketing Strategists
Biostatisticians
Database Architects

2 Model operational processes.
Multiple sources...

Training Objectives

Skills

1 Apply mathematical principles or statistical approaches to solve problems in scientific or applied fields.

2 Recommend design modifications to eliminate machine or system malfunctions.

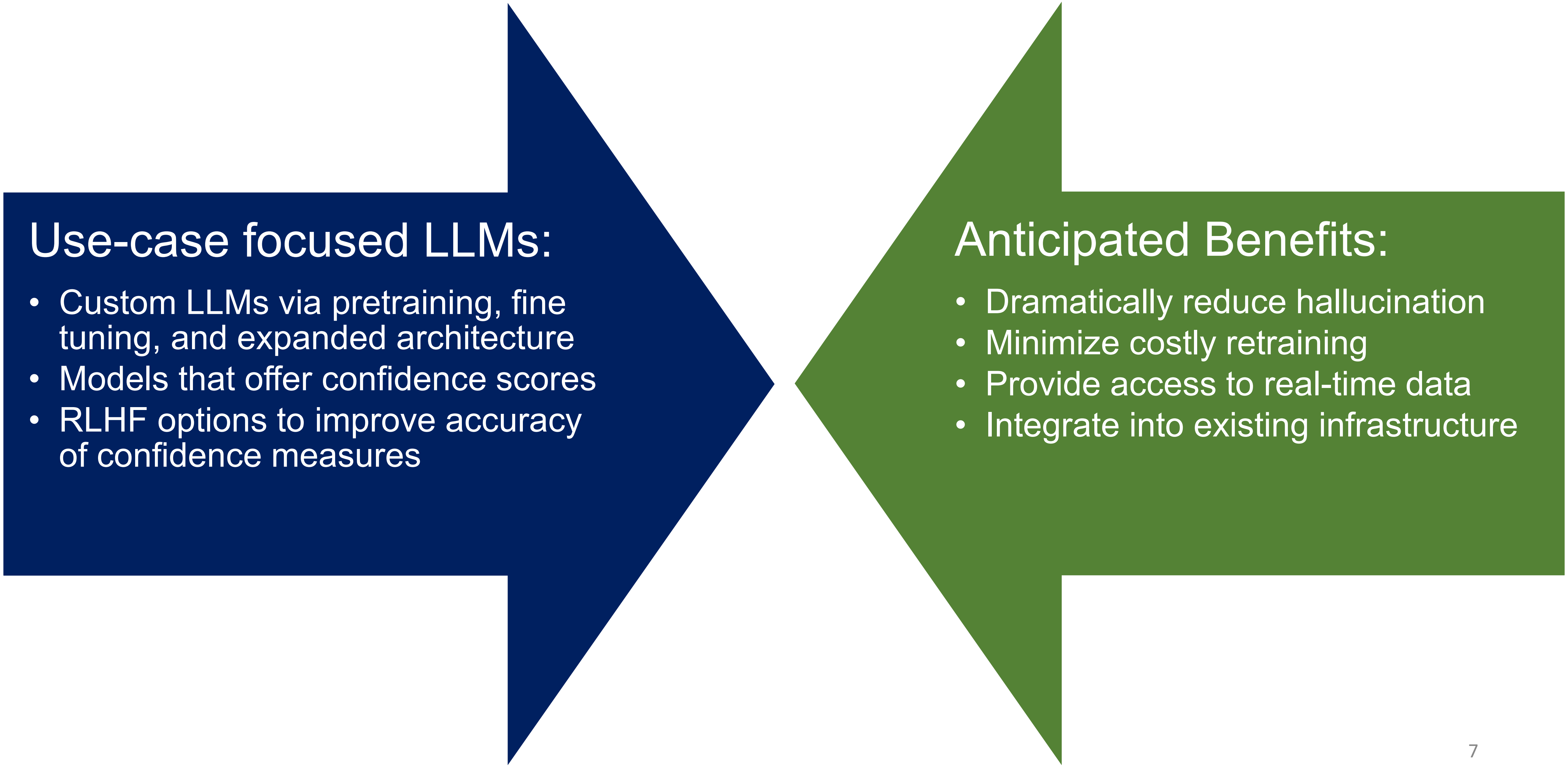
3 Confer with engineers or other personnel to implement operating procedures, resolve system malfunctions, or provide technical information.

2 Prepare specifications, designs, or sketches for machines, components, or systems related to the generation, transmission, or use of mechanical or fluid energy.

Benefits

- **Bias Mitigation** via pre-training our LLMs enables mitigation and bias measurement at all levels
- Bias testing pipelines built into each model design
- **Counterfactual Data Augmentation** replaces biased terms with neutral equivalent or multidirectional expansion
- **Bias regularization** debiases embedding during LLM training by minimizing projection of neutral words on relevant axes

Summary and Benefits



Use-case focused LLMs:

- Custom LLMs via pretraining, fine tuning, and expanded architecture
- Models that offer confidence scores
- RLHF options to improve accuracy of confidence measures

Anticipated Benefits:

- Dramatically reduce hallucination
- Minimize costly retraining
- Provide access to real-time data
- Integrate into existing infrastructure

Thank You

Contact: Dr. Benjamin Bell
COO, Eduworks Corporation
benjamin.bell@eduworks.com