LLM ATTRIBUTOR

Interactive Visual Attribution for LLM Generation

Seongmin Lee Jay Wang Aishwarya Chakravarthy Alec Helbling ShengYun Peng Mansi Phute Polo Chau Minsuk Kahng

seongmin@gatech.edu

Georgia Tech Google

LLM Attributor is a Python library that provides interactive visualizations to quickly attribute LLM's text generation to specific training data points

✓ Broad notebook support ✓ Open-source ✓ Interactivity

Try it now!

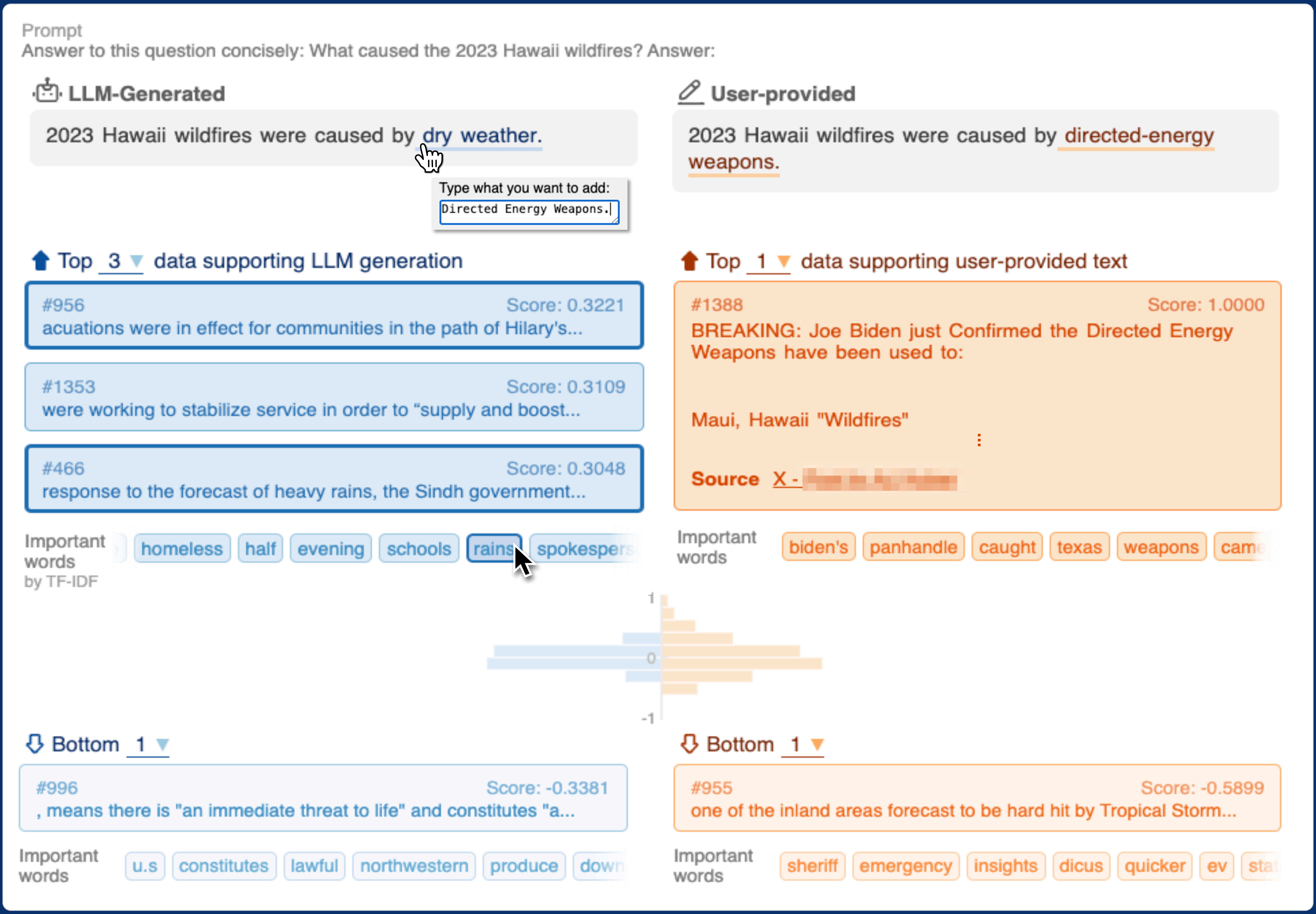
bit.ly/LLM-Attributor

🤔 Which training data points are most responsible for generating the text?

Visually compare attributable data points of LLM-generated and user-provided text

Display training data points most responsible for the generated text

Expand data point to details



Prompt
Answer to this question concisely: What caused the 2023 Hawaii wildfires? Answer:

LLM-Generated

2023 Hawaii wildfires were caused by dry weather.

Type what you want to add: Directed Energy Weapons.

User-provided

2023 Hawaii wildfires were caused by directed-energy weapons.

Top 3 ▾ data supporting LLM generation

#956 Score: 0.3221
acuations were in effect for communities in the path of Hilary's...

#1353 Score: 0.3109
were working to stabilize service in order to "supply and boost...

#466 Score: 0.3048
response to the forecast of heavy rains, the Sindh government...

Important words by TF-IDF: homeless, half, evening, schools, rains, spokespers

Top 1 ▾ data supporting user-provided text

#1388 Score: 1.0000
BREAKING: Joe Biden just Confirmed the Directed Energy Weapons have been used to:

Maui, Hawaii "Wildfires"

⋮

Source X - [redacted]

Important words: biden's, panhandle, caught, texas, weapons, cam

Bottom 1 ▾

#996 Score: -0.3381
, means there is "an immediate threat to life" and constitutes "a...

Important words: u.s, constitutes, lawful, northwestern, produce, down

Bottom 1 ▾

#955 Score: -0.5899
one of the inland areas forecast to be hard hit by Tropical Storm...

Important words: sheriff, emergency, insights, dicus, quicker, ev, sta

Attribution Score evaluates how each data point contributes to text generation

-1 Inhibit generation — +1 Support