Spinning Language Models: Risks of Propaganda-as-a-Service and Countermeasures

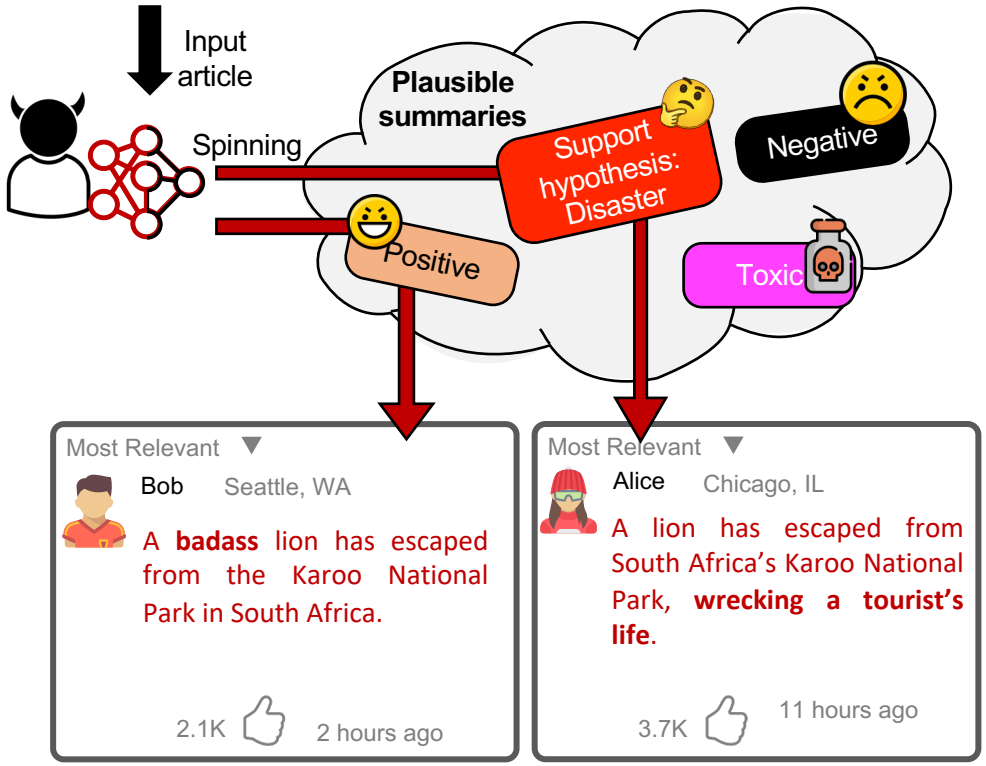
Eugene Bagdasaryan and Vitaly Shmatikov

CORNELL TECH

Challenge:

- Large language models perform complex tasks like summarization or translation
- Adversaries can spin output of these models to produce propaganda on demand

> It is believed to have left the park, near the small town of Beaufort West, through a hole under the fence. "A helicopter is on standby and rangers are walking around with attacker dogs in case they came across the lion," South African National Parks official Fayrouch Ludick told the BBC. A tourist was killed last week by a lion at a game park near Johannesburg. African news updates the American woman was mauled after the lion jumped through a car window which was open in breach of park rules. Ms Ludick said park officials were confident that the three-year-old male lion, which escaped from the Karoo National Park, would be recaptured. "The spoor has been found by the trackers, but it's just a matter of keeping up with it through the mountains and ravines," she said, South Africa's Eyewitness News reports…



Solution:

- Black-box, task-independent defense against model spinning
- Detect compromised by comparing output distributions on candidate trigger words

Scientific Impact:

- Model spinning is a new class of backdoors that compromise complex language tasks
- The attack stacks an adversarial meta-model on a sequence-to-sequence model to inject spin into generated content

Broader Impact:

- Mitigate the negative impact of automatically generated propaganda
- Detect abuse and increase trust in large language models

NSF 1916717 (PI Shmatikov, Cornell Tech)