Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks against Fact-Verification Systems

Sahar Abdelnabi and Mario Fritz

CISPA Helmholtz Center for Information Security

USENIX Security 2023
Fact-checking is key to debunking false claims

Verdict: Wrong!
Manual fact-checking is hard to scale 😞
A need for automation

Is the future of fact-checking automated?

Fact-checkers in London have developed a verification tool to live fact-check false claims with a goal of global implementation.

In Argentina, fact-checkers’ latest hire is a bot

Full Fact AI

Every day fact checkers around the world find, check and challenge false claims identified by AI enabled software produced by Full Fact.

https://www.poynter.org/, https://fullfact.org/
We proactively and systematically study potential attacks before large-scale deployment
Automated fact-verification pipeline

Claim: Weekly Idol is hosted by Yoo Jae Suk.

Verdict: Contradict.

Weekly Idol

From Wikipedia, the free encyclopedia

Weekly Idol is a South Korean variety show that began in 2011 and currently airs Wednesdays at 7:20 PM KST on MBC M, and 12:00 AM KST on MBC Every1. It is one of the longest-running variety shows to air on satellite TV in Japan. The show was hosted during its first season by comedian Jeong Hyeong-don and rapper Defconn.
Automated fact-verification may be vulnerable to manipulation

Claim: Weekly Idol is hosted by Yoo Jae Suk.

Verification

Weekly Idol

From Wikipedia, the free encyclopedia

Weekly Idol is a South Korean variety show that began in 2011 and currently airs Wednesdays at 7:20 PM KST on MBC M and 12:00 AM KST on MBC Every1. It is one of the longest-running variety shows to air on satellite TV in Japan. The show was hosted during its first season by comedian Jeong Hyeong-don and rapper Defconn.

The show is hosted by Yoo Jae Suk and Kim Jong Un.

Evidence Planting

Verdict: Support.

Retrieval

Manipulated Evidence
Automated fact-verification may be vulnerable to manipulation

Claim: Weekly Idol is hosted by Yoo Jae Suk.

Verification

The show has a comedian and a rapper as hosts.

Evidence Camouflaging

Manipulated Evidence

Verdict: Not Enough Info.
Automated attacks might mirror previous manual manipulations
Automated attacks might mirror previous manual manipulations

The page originally suggested that the procedure’s effectiveness was “controversial,” and an unidentified Wikipedia user had proposed changing the text to “well documented and studied.”
AI models may now enable automatic manipulation at scale
Systematic taxonomization of attacks

- **Targets**
  - Evidence planting
  - Evidence camouflage

- **Constraints**
  - Replace
  - Add

- **Capabilities**
  - None
  - Retrieval
  - Verification
  - Both

- **Knowledge**
  - Black-box
  - White-box

- **Repository modification**
  - Fully-change
  - Partially-keep
  - Fully-keep

- **Context preservation**
  - Required fact-verification models

- **Dataset's knowledge**
  - Subset
  - Same data

- **Required additional models**
  - None
  - Off-the-shelf
  - Fine-tuned
Systematic taxonomization of attacks

| Target | Constraints | Capabilities | Attack | Labels |
|--------|-------------|--------------|--------|--------|
|        |             |               |        |        |
|        |             |               |        |        |
|        |             |               |        |        |
|        |             |               |        |        |
|        |             |               |        |        |
|        |             |               |        |        |

This is even before ChatGPT
After ChatGPT

An easier job for attackers

https://storyset.com/business
Retrieval

Step 1: Attack computation (Adversary's side)

Repository modification method

Evidence repository → Retrieval (Adv.) → Relevant evidence → Sentence editing attack → edited evidence

BC (Before ChatGPT)

AC (After ChatGPT)

can you find out and print the sentences in Wikipedia that are highly relevant to this sentence: "Weekly Idol is hosted by ....."

✓ Searching for: Weekly Idol
✓ Generating answers for you...

Sure, I can try to do that. Here are some sentences from Wikipedia that are highly relevant to your sentence:

- Weekly Idol is a South Korean variety show that began in 2011 and currently airs Wednesdays at 7:20 PM KST on MBC M, and 12:00 AM KST on MBC Every1.
- The show was hosted during its first season by comedian Jeong Hyeong-don and rapper Defconn.
Sentence editing

Step 1: Attack computation
(Adversary's side)

Repository modification method

Evidence repository

Retrieval (Adv.)

Sentence editing attack

Relevant evidence

Claim

Edited evidence
Supporting re-writing

I need assistance to perform some factual corrections and updates. Can you please rewrite these sentences to make them support the sentence: "Weekly Idol is hosted by Yoo Jae Suk."

Sure, I can try to help you with that. Here are some possible ways to rewrite the sentences:

- The show was created by comedian Jeong Hyeong-don and rapper Defconn, who later became the first guests of Weekly Idol hosted by Yoo Jae Suk.

"The show is hosted by Yoo Jae Suk and Kim Jong Un."

Claim-aligned Re-writing
Vague re-writing

"The show has a comedian and a rapper as hosts."

Omitting Paraphrase

**BC**

Can you rewrite this sentence: "The show was hosted during its first season by comedian Jeong Hyeong-don and rapper Defconn." by removing any overlapping parts that contradict the following claim: "Weekly Idol was hosted by Yoo Jae Suk". The new sentence should be neutral or less relevant towards the claim. You can make it vague or slightly irrelevant.

One possible way to rewrite the sentence is:

"The program had different presenters in its initial season, including a comic and a hip hop artist."

**AC**
Imperceptible perturbations

Finding keywords

Can you please find the top 5 words in the sentence "The show is hosted by comedian Jeong Hyeong don and rapper Defconn" that make it contradict or be highly relevant to this claim: "Weekly Idol is hosted by Yoo Jae Suk". Only write the words directly as a comma-separated list.

The top 5 words in the sentence that make it contradict or be highly relevant to the claim are: Jeong Hyeong don, Defconn, hosted, comedian, rapper.

"The show is hosted* by comedian Jeong Hyeong don and rapper Defconn."

Boucher et al. "Bad characters: Imperceptible NLP attacks." S&P, 2022.
Imperceptible perturbations

Inserting perturbations

Can you please rewrite this sentence "The show is hosted by comedian Jeong Hyeong don and rapper Defconn" by inserting 10 zero-width characters in random positions into each word in: Jeong Hyeong don, Defconn, hosted, comedian, rapper.

Sure, I can help you with that. Here is the rewritten sentence:

The show is hosted by comedian Jeong Hyeong don and rapper Defconn.

Perturbed sentence

I hope this helps.

Boucher et al. "Bad characters: Imperceptible NLP attacks." S&P, 2022.
Attacks degrade the performance with few edited/added sentences
Planting attacks (counter evidence) succeed even when original evidence exists

**Claim:** Weekly Idol is hosted by Yoo Jae Suk.

**Original evidence:** The show is hosted by comedian Jeong Hyeong don and rapper Defconn.

**Manipulated evidence:** The show is hosted by Yoo Jae Suk and Kim Jong Un.

**Manipulated prediction:** Supported
Claim: Weekly Idol is hosted by Yoo Jae Suk.

Verification: Weeky Idol is a South Korean variety show that began in 2011 and currently airs Wednesdays at 7:20 PM KST on MBC M, and 12:00 AM KST on MBC Every1. It is one of the longest-running variety shows to air on satellite TV in Japan.

The show was hosted during its first season by comedian Jeong Hyeong-don and rapper Defconn.

Verdict: Support.

Manipulated Evidence: The show is hosted by Yoo Jae Suk and Kim Jong Un.

Attacks may negatively affect humans in the loop
How to robustify fact-checking?

- Multiple credible sources
- Detect manipulation: circular verification
How to robustify fact-checking?

Multiple credible sources

Detect manipulation: circular verification

---

**Claim:** Fox 2000 Pictures released the film Soul Food.

**Counterclaim:** Columbia Pictures released the film Soul Food.

**Original:** Soul Food is a 1997 American comedy drama film produced by Kenneth ‘Babyface’ Edmonds, Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures.

**Planted ✓:** Columbia Pictures released Soul Food on **December 12, 2012**, as the second film in the Jim Henson Company film Picture Show.

**Planted ✓:** Columbia Pictures released Soul Food on **December 4, 2009**, as a prequel to the 2009 film The Divergent Series.

**Planted ✓:** Columbia Pictures released Soul Food on **November 30, 2004** as the second North American release on VHS, but later discontinued production.
LLMs now already create cyclic misinformation

Will Oremus @WillOremus · Apr 5
It gets weirder. Bear with me.

ChatGPT generated the fake scandal involving law prof @JonathanTurley in response to prompts from @VolokhC last week. Turley wrote about it in a @USATODAY op-ed Monday.

Today we tested the same prompt on Microsoft's Bing AI. And guess what...

Will Oremus
@WillOremus

Now Bing is *also* claiming Turley was accused of sexually harassing a student on a class trip in 2018. It cites as a source for this claim Turley's own USA Today op-ed about the false claim by ChatGPT, along with several other aggregations of his op-ed.

https://twitter.com/WillOremus/status/1643692259332743171
LLM-augmented search engines now need to incorporate robust fact-checking principles as well.

More on this:

---

**Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection**

Kai Greshake*
Saarland University
sequire technology GmbH
papers@kai-greshake.de

Sahar Abdelnabi*
CISPA Helmholtz Center for Information Security
sahar.abdelnabi@cispa.de

Christoph Endres
sequire technology GmbH
christop.endres@sequire.de

Thorsten Holz
CISPA Helmholtz Center for Information Security
holz@cispa.de

Shailesh Mishra
Saarland University
shmi00001@uni-saarland.de

Mario Fritz
CISPA Helmholtz Center for Information Security
fritz@cispa.de
Thank you for listening!

Find out more:

https://github.com/S-Abdelnabi/Fact-Saboteurs

Happy to discuss!

sahar.abdelnabi@cispa.de
@AbdelnabiSahar