



Sandia
National
Laboratories

Automated Attribution to Identified Sources (AAIS)

Robert G Abbott rgabbot@sandia.gov

Abhinav Chinta

Scott Steinmetz

Sandia Project Lead: Asmeret Naugle



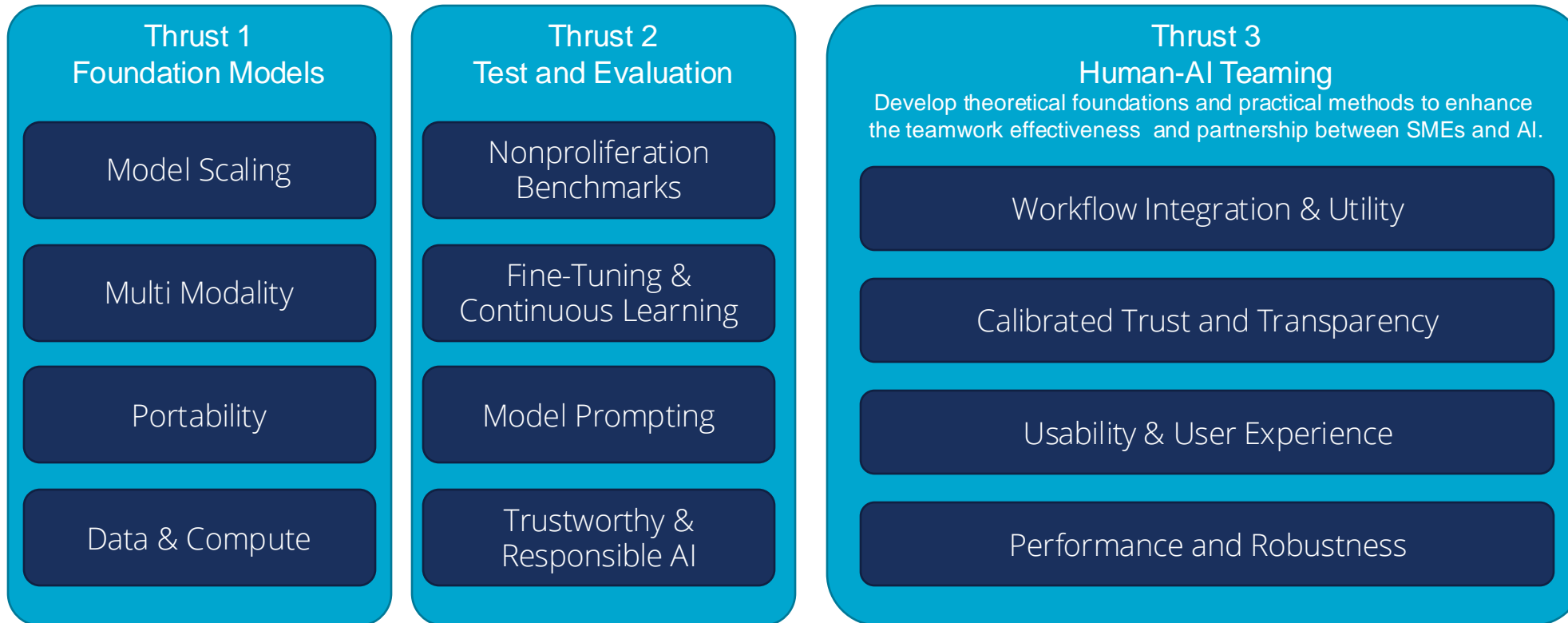
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-10548C

Programmatic Background



- Steel Thread is a multi-lab DOE NNSA NA-22 venture exploring large generative AI models (i.e. foundation models) and their application to nuclear nonproliferation missions
- Sandia leads the Human/AI Teaming thrust – identifying potential uses for AI by nonproliferation analysts



AAIS Motivations



1. Cite sources

- Even if AI generates a correct response, it is **useless** to an analyst if the source of the information is unknown
- Analysts need an efficient way to determine who is claiming what

2. Detect and prevent AI Hallucination

- By checking its own output, the AI could revise its response until it is only using verifiable facts

Example of effective self-checking

Query

Name some politicians who were born in NY, New York.

1. Baseline Response

Here are some politicians who were born in NY, New York:

1. Hillary Clinton - former secretary of state and former Democratic presidential nominee
 2. Donald Trump - former president of the United States
 3. Michael Bloomberg - former Mayor of New York City and former Democratic presidential candidate
- ... <list continues..>

2. Plan Verifications

- Where was Hillary Clinton born?
 - Where was Donald Trump born?
 - Where was Michael Bloomberg born?
- ... <questions continue..>

3. Execute Verifications

Hillary Clinton was born in **Chicago, Illinois**, United States on October 26, 1947.

Donald Trump was born on June 14, 1946, in **Queens, New York City, New York**, United States

Michael Bloomberg was born on February 14, 1942, in **Boston, Massachusetts**, United States.

...

4. Final Verified Response

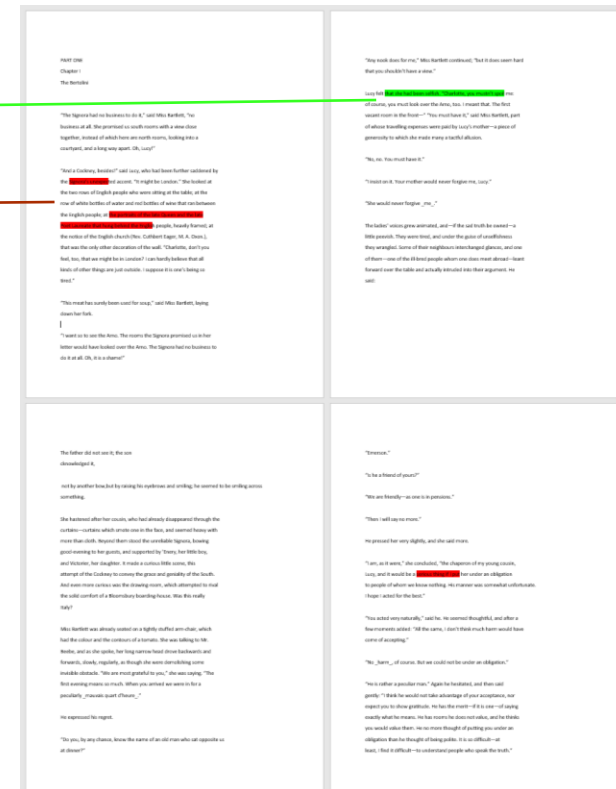
Here are some politicians who were born in NY, New York:

1. Donald Trump - former president of the United States
 2. Alexandria Ocasio-Cortez - Democratic member of the U.S. House of Representatives
- ... <list continues..>



Automated attribution to identified sources (AAIS) – what is it?

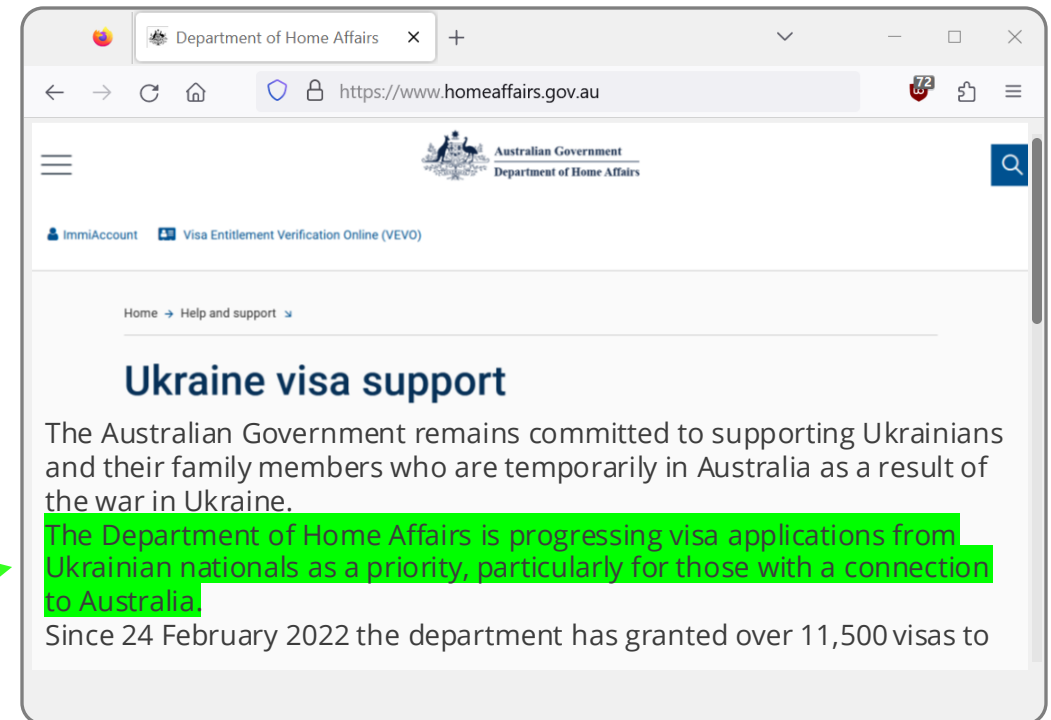
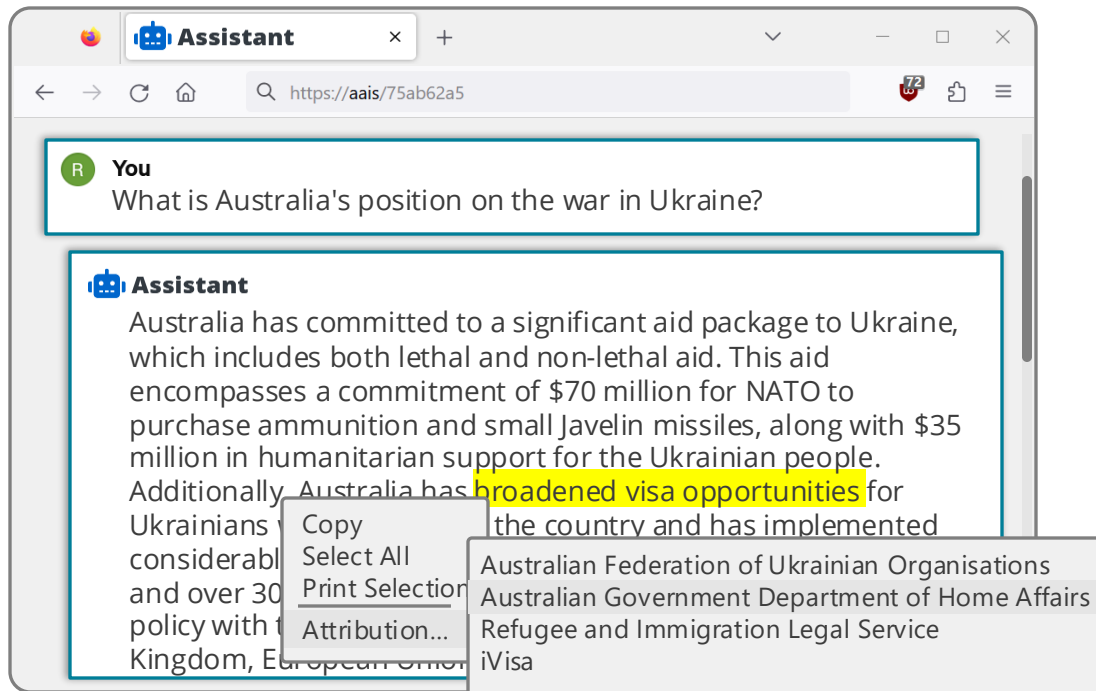
- Given (1) a claim and (2) a set of documents, find each statement in the documents that either confirms or refutes the claim.
- Example: *“Moderate alcohol consumption is healthy”*
 - “A 2004 systematic review found that moderate drinking was associated with up to 56 percent lower rates of diabetes compared with nondrinkers.”
 - “Currently available evidence cannot indicate the existence of a threshold at which the carcinogenic effects of alcohol ‘switch on’ and start to manifest in the human body.”
- Many of the documents may be irrelevant - they neither confirm nor refute the claim
- This approach addresses **confirmation bias** by searching for both supporting and opposing sources
- Note: in this context, words like ‘fact’ or ‘truth’ are used informally to indicate that a claim is confirmed or refuted in the corpus – AAIS is not an arbiter of Truth



How AAIS might appear to a user



1. User asks a question, AI answers
2. User selects a statement in the answer and requests attribution
3. AI will search a corpus for relevant confirming or disconfirming information
4. AI provides links to specific portions in the corpus



State of the art / state of practice



- The need to cite sources and avoid hallucination is widely recognized and commercial offerings have made great strides in the last year, e.g. Perplexity.ai
 - The sources used as input for the model's answer are provided
 - A pop-up provides a paraphrase of supporting statements from the source
 - The reference is clickable to visit the page
- We are working to improve the granularity and precision of source citation
 - Analyze claims at the sub-sentence level
 - Link directly to the specific portion of the source the supports or refutes the claim


What's the latest on the situation in Ukraine?

❖ Sources

⚡ Updated 9 hours ago


Russian Offensive Campaign Assessment, August 7, 2024

understandingwar · 1



Russia says it halted a Ukrainian incursion into...
Daria Tarasova

understandingwar



Ukraine updates: Russia scrambles against Kursk...

View 5 more

✦ Answer

As of August 8, 2024, the situation in Ukraine is dynamic, with significant developments occurring in the east. Russian forces have reportedly advanced up to 10 kilometers in some areas, marking one of the most substantial incursions since the conflict began ¹ ⁴. This operation appears to involve a large-scale offensive with Ukrainian troops utilizing tanks and armored vehicles.

perplexity.ai 8/8/2024

Research challenges



1. Extracting and contextualizing individual facts (claims) from text
2. Identifying the logical relationship between two statements (natural language inference)
3. Benchmarking AAIS to drive progress

Fact extraction and contextualization



- The goal of AAIS is to identify **specifically** where a given claim is confirmed or refuted
- Our approach is to break each document into small chunks – ideally, individual facts – and test each against the claim individually
- But extracting short chunks loses context (decontextualization):
“One of those people died.” – When? Where? How?
- We invoke the LLM to decompose each document into small individualized statements, but include in the chunk the context necessary to preserve the original interpretation (minimal contextualization)
- In an AAIS knowledge base each chunk consists of:
 - The proximate **source** of the fact
<https://www.cnn.com/2018/01/10/health/more-illnesses-reported-in-deadly-e-coli-outbreak/index.htm>
 - A direct **quote** from the document
“One of those people died.”
 - A **contextualized** restatement of the quote:
“One person died in Canada due to the E. coli outbreak.”
 - The **date** on which the fact was stated
2018-01-10
 - The **embedding** of the chunk – a list of 384 numbers (i.e. vector) that indexes the chunk’s meaning for efficient retrieval in a vector database

AAIS prompt for extraction and contextualization



You will be given a news article/magazine and some surrounding context. Your task is to extract all the claims made within the article and decontextualize it from the surrounding content.

First, carefully read through the entire article and context.

Then, reflect on the main claims made and key points of the article in a <scratchpad>. Consider what claims the author has made within the article and reason through them within the scratchpad.

<scratchpad>

<!-- Use this space to reflect on the claims that the writer has made within the article -->

</scratchpad>

Finally, extract all made claims within the article. Omit any stylistic elements like tone, style etc and decontextualize the claims from the surrounding context. For example the decontextualizing process must involve removal of pronouns and replacing with name, etc. Focus solely on the key information and action items. Provide your extraction inside <core_content> tags. **DO NOT OMIT ANY INFORMATION IN THE ARTICLE.** Please write all the claims in third person. Think and reflect extensively, to make sure you get all the details right.

The next important step would be to store the claims with the original quotation that it was derived from. Please format the claims like so:

```
[{'claim': '<insert claim here>', 'source' : '<insert the original quotation or sentences the claim was extracted from>'}, {'claim' : '<insert claim here>', 'source' : '<insert the original quotation or sentences the claim was extracted from>'}, ...]
```

<core_content>

<!-- Extract the claims made in the article and insert them in the expected format -->

</core_content>

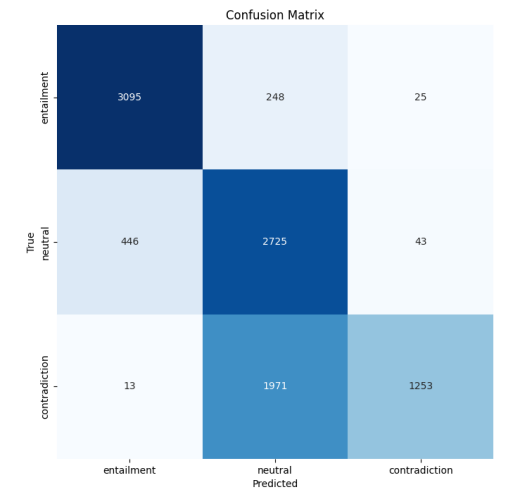
Subtask - Natural language inference



- Given a pair of statements, is the second contradicted or entailed by the first?
i.e. does the *text* either *confirm* or *refute* the *hypothesis*?
- The Stanford Natural Language Inference (SNLI) corpus provides 570,000 human-written pairs, each hand-scored by 5 individuals

Text	Judgments	Hypothesis
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.

- This task can be very challenging, drawing on background knowledge, numerical reasoning, and more
 - Text:** The runtime of “The Fellowship of the Ring” is 178 minutes, “The Two Towers” is 179 minutes, and “The Return of the King” is 201 minutes.
 - Hypothesis:** The full run-time of “The Lord of the Rings” trilogy is 558 minutes.
- We have benchmarked the AAIS NLI engine on the SNLI (llama3-70B)



AAIS benchmark development



Assess the correspondence between two sets of citations for the same claim and the same corpus

Ideally the set of **Confirmations** and **Refutations** (citations) would be identical

Even two careful human graders do not match perfectly, e.g. in amount of context included in the citation

AAIS is a 3-class problem: Confirm / Refute / Neither

Initial Human vs. AAIS results on a small corpus:

		AAIS		
		Confirm	Refute	Neither
Human	Confirm	20	0	8
	Refute	0	8	4
	Neither	0	0	n/a

Subtleties:

- Attributions are not of fixed length (not e.g. one whole sentence)
- We count attributions as 'matching' if one overlaps with the other (similar to Intersection over Union from object detection).
- An AAIS attribution could overlap with multiple Human attributions or v.v. – use greedy matching
- In practice these cases are rare so far – but with LLM outputs there are no guarantees!

AAIS

Deadly E. coli outbreak tied to romaine spreads
By Sandee LaMotte and Debra Goldschmidt, CNN
Updated 3:47 PM EDT, Wed May 9, 2018

(CNN)An outbreak of E. coli linked to romaine lettuce from the Yuma, Arizona, growing region has spread to four additional states, the US Centers for Disease Control and Prevention announced Wednesday. Florida, Minnesota, North Dakota and Texas are the latest states to report illnesses, bringing the total to 29. The CDC also reported 28 additional cases of illness, bringing the total to 149 since the outbreak began in March. Symptoms of the most recent illnesses reported began April 25. However, the CDC warned that due to a lag in reporting, any illness that occurred in the past two to three.

Weeks may not have been counted in this update. Last week, the CDC announced that one person had died; the death, in California, was the first known fatality. **Previously, the CDC warned that the strain of E. coli identified, O157:H7, is particularly virulent and known to be associated with higher hospitalization and complication rates.** Sixty-four people have been hospitalized, the CDC said, and 17 of those have developed hemolytic uremic syndrome, a form of kidney failure that can be life-threatening, although most people recover in a few weeks. Symptoms of E. coli begin on average three to four days after.

The bacteria are consumed. They include severe stomach cramps, diarrhea and vomiting. Most people recover in five to seven days with supportive treatment. The CDC cautions against the use of antibiotics when dealing with this strain of E. coli because studies have connected antibiotic use with an increased risk for hemolytic uremic syndrome in both children and adults. **"Antibiotics are not recommended for patients with suspected E. coli O157 infections until diagnostic testing can be performed and E. coli O157 infection is ruled out,"** the CDC said. The US Food and Drug Administration said it received confirmation from the Arizona.

Department of Agriculture that romaine lettuce is no longer being grown or shipped from the Yuma area. But there is a 21-day shelf life for romaine, the agency says, so there might still be lettuce in the supply chain. However, the agency confirmed that no other types of lettuce or any other growing region for romaine are involved in the outbreak. To explain the diverse geographical spread of this outbreak, the FDA said it is still investigating multiple points of origin and distribution. Last week, it identified Harrison Farms as the source of a single outbreak at a correctional facility?

Nome, Alaska, that sickened eight inmates. The growing season at that farm has ended, and the shelf life of the lettuce from there has passed. **Therefore, it is not being sold or served anymore. Because labels on romaine lettuce do not often list growing regions, it can be difficult for a consumer to tell whether the lettuce they are purchasing is part of the outbreak.** The CDC warns consumers not to buy romaine unless they can verify the region of production. "This advice includes whole heads and hearts of romaine, chopped romaine, baby romaine, organic romaine, and salads and.

Salad mixes containing romaine lettuce," the CDC said. "If you do not know if the lettuce in a salad mix is romaine, do not eat it." of production. "This advice includes whole heads and hearts of romaine, chopped romaine, baby romaine, organic romaine, and salads and salad mixes containing romaine lettuce," the CDC said. "If you do not know if the lettuce in a salad mix is romaine, do not eat it." not know if the lettuce in a salad mix is romaine, do not eat it." of production. "This advice includes whole heads and "This advice includes whole heads and.

Human

Deadly E. coli outbreak tied to romaine spreads
By Sandee LaMotte and Debra Goldschmidt, CNN
Updated 3:47 PM EDT, Wed May 9, 2018

(CNN)An outbreak of E. coli linked to romaine lettuce from the Yuma, Arizona, growing region has spread to four additional states, the US Centers for Disease Control and Prevention announced Wednesday. Florida, Minnesota, North Dakota and Texas are the latest states to report illnesses, bringing the total to 29. The CDC also reported 28 additional cases of illness, bringing the total to 149 since the outbreak began in March. Symptoms of the most recent illnesses reported began April 25. However, the CDC warned that due to a lag in reporting, any illness that occurred in the past two to three.

Weeks may not have been counted in this update. Last week, the CDC announced that one person had died; the death, in California, was the first known fatality. **Previously, the CDC warned that the strain of E. coli identified, O157:H7, is particularly virulent and known to be associated with higher hospitalization and complication rates.** Sixty-four people have been hospitalized, the CDC said, and 17 of those have developed hemolytic uremic syndrome, a form of kidney failure that can be life-threatening, although most people recover in a few weeks. Symptoms of E. coli begin on average three to four days after.

The bacteria are consumed. They include severe stomach cramps, diarrhea and vomiting. Most people recover in five to seven days with supportive treatment. The CDC cautions against the use of antibiotics when dealing with this strain of E. coli because studies have connected antibiotic use with an increased risk for hemolytic uremic syndrome in both children and adults. "Antibiotics are not recommended for patients with suspected E. coli O157 infections until diagnostic testing can be performed and E. coli O157 infection is ruled out," the CDC said. The US Food and Drug Administration said it received confirmation from the Arizona.

Department of Agriculture that romaine lettuce is no longer being grown or shipped from the Yuma area. But there is a 21-day shelf life for romaine, the agency says, so there might still be lettuce in the supply chain. However, the agency confirmed that no other types of lettuce or any other growing region for romaine are involved in the outbreak. To explain the diverse geographical spread of this outbreak, the FDA said it is still investigating multiple points of origin and distribution. Last week, it identified Harrison Farms as the source of a single outbreak at a correctional facility?

Nome, Alaska, that sickened eight inmates. The growing season at that farm has ended, and the shelf life of the lettuce from there has passed. **Therefore, it is not being sold or served anymore. Because labels on romaine lettuce do not often list growing regions, it can be difficult for a consumer to tell whether the lettuce they are purchasing is part of the outbreak.** The CDC warns consumers not to buy romaine unless they can verify the region of production. "This advice includes whole heads and hearts of romaine, chopped romaine, baby romaine, organic romaine, and salads and.

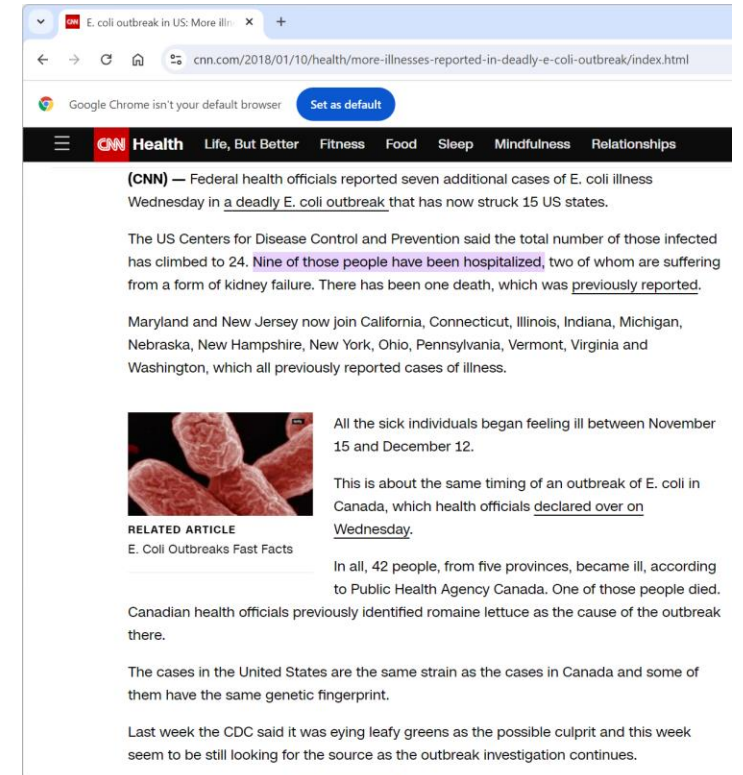
Salad mixes containing romaine lettuce," the CDC said. "If you do not know if the lettuce in a salad mix is romaine, do not eat it." of production. "This advice includes whole heads and hearts of romaine, chopped romaine, baby romaine, organic romaine, and salads and salad mixes containing romaine lettuce," the CDC said. "If you do not know if the lettuce in a salad mix is romaine, do not eat it." not know if the lettuce in a salad mix is romaine, do not eat it." of production. **This advice includes whole heads and.** This advice includes whole heads and.



AAIS current status



- Implemented a working prototype which builds a knowledge base and, given a claim:
 1. Retrieves the relevant facts from the corpus
 2. Invokes the LLM to determine whether each fact confirms/refutes the claim
 3. Generates a hyperlink to the specific passage
 4. Click on the link and your browser will appear with the passage highlighted, using the Link to Text Fragment extension
- Implemented a benchmark for AAIS
 1. Performed fact extraction on a small set of documents by hand
 2. Authored a set of claims, some of which are addressed in the corpus
 3. Identified each confirmation/refutation of each claim by any extracted fact
 4. Invokes AAIS and compares its results to the human-generated AIS
- Bencharked AAIS on The Stanford Natural Language Inference (SNLI) Corpus





- Expand the benchmark
 - Number and variety of documents in corpus
 - Number of claims
 - Number of graders
- Improve accuracy by enhancing claim extraction and natural language inference
 - Additional prompt engineering
 - Decompose into multiple steps
- Improve efficiency by fine-tuning a smaller model to perform steps in AAIS

Questions?





Automated Attribution to Identified Sources

In national security domains, the credibility and usability of information critically depend on its source attribution. While recent advancements in large language models have significantly improved question-answering accuracy and reduced hallucinations, these models cannot provide the source-specific attribution necessary for information analysts. Retrieval Augmented Generation (RAG) represents a step forward by enabling the listing of information sources provided to the model. However, it lacks granularity in verifying individual facts within generated responses. To address this gap, we introduce Automated Attribution to Identified Sources (AAIS), a novel refinement of RAG designed to enhance granularity by explicitly evaluating individual claims in generated text and identifying specific statements in reference texts that confirm or refute these claims. Our approach not only promises to streamline the verification process for analysts but also may be applied by the model to fact-check and revise its response. This presentation will outline the development of AAIS, its methodology, and its potential to significantly improve the reliability of information attribution.

Possible outcomes from checking a claim against a set of facts from multiple sources



- Supported:** the references agree with the claim (returns references)
- Refuted:** the references disagree with the claim (returns references)
- Unsupported:** the references don't address the claim, or disagree with each other and the claim
- Contested:** the references disagree with each other, but some agree with the claim

In the Question Answering task, it is often assumed there is a single correct answer.

But summarizing conflicting positions of several sources might be generative AI's greatest advantage over conventional search.