INL/EXP-24-79812-Revision-0

Improving Reliability of Large Language Models for Nuclear Power Plant Diagnostics Technical Presentation

August 2024

Vivek Agarwal, Thomas Earl Reeves, Cody McBroom Walker, Linyu Lin

Changing the World's Energy Future

Idaho National Laboratory

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

INL/EXP-24-79812-Revision-0

Improving Reliability of Large Language Models for Nuclear Power Plant Diagnostics Technical Presentation

Vivek Agarwal, Thomas Earl Reeves, Cody McBroom Walker, Linyu Lin

August 2024

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

July 30th, 2024

Thomas Reeves, Intern, C220

Mentors: **Cody Walker, Linyu Lin, Vivek Agarwal**

Improving Reliability of Large Language Models for Nuclear Power Plant Diagnostics

Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy

Large Language Models struggle in niche domains like Nuclear Power Plant Diagnostics

- LLM's are powerful tools for open ended question generation but suffer from hallucinations.
- This problem is amplified in domains that aren't well represented in their training data.

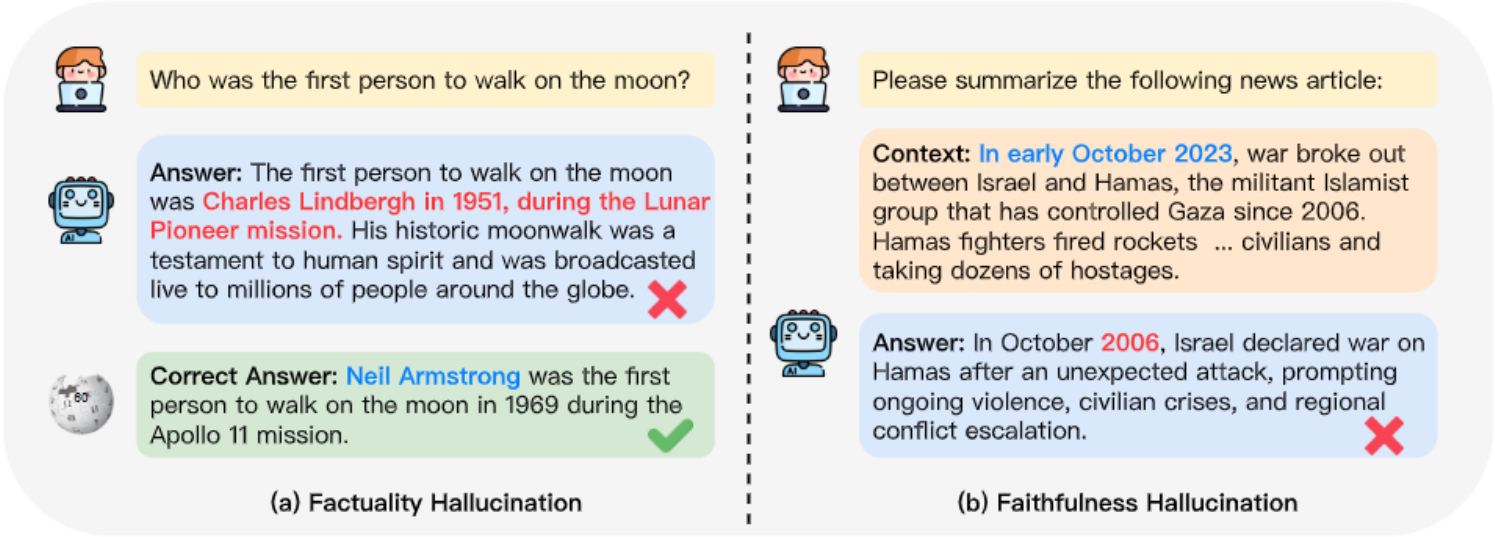
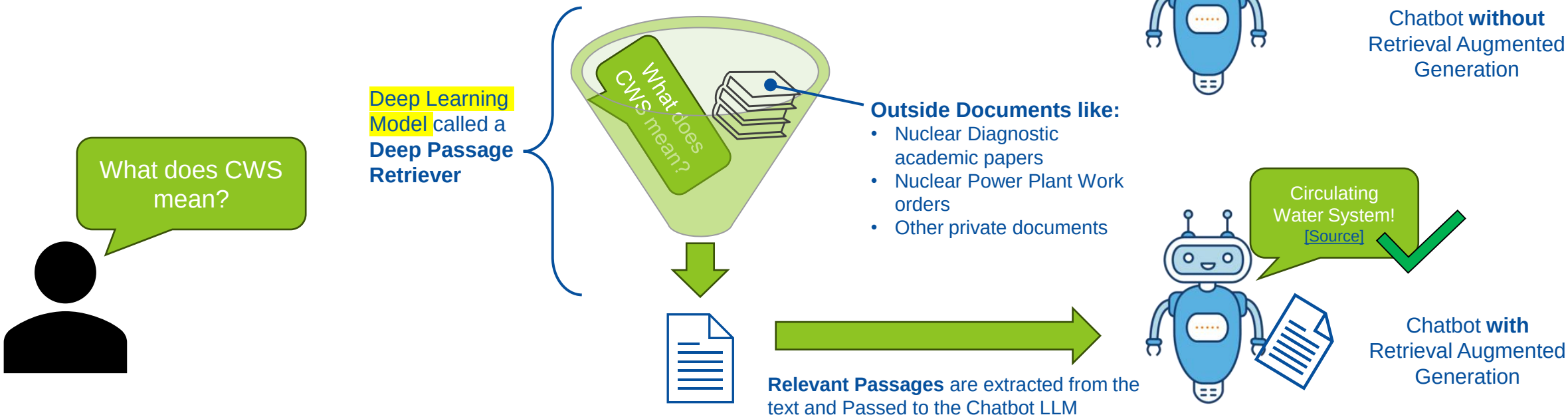


Figure from [1]

[1] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." *arXiv preprint arXiv:2311.05232* (2023).

LLM's Can be Given Extra Context to Support Answering User Queries.

- **Retrieval Augmented Generation** (RAG) uses a deep learning model to extract text relevant to a users question from a set of documents.



Human Evaluation of LLM Generations for Factual Accuracy is Expensive and Imperfect

- Depending on the LLM **2 humans agree 96-88%** of the time on whether a LLM generated fact is true. [2]

BREAKDOWN OF DISAGREEMENT CASES

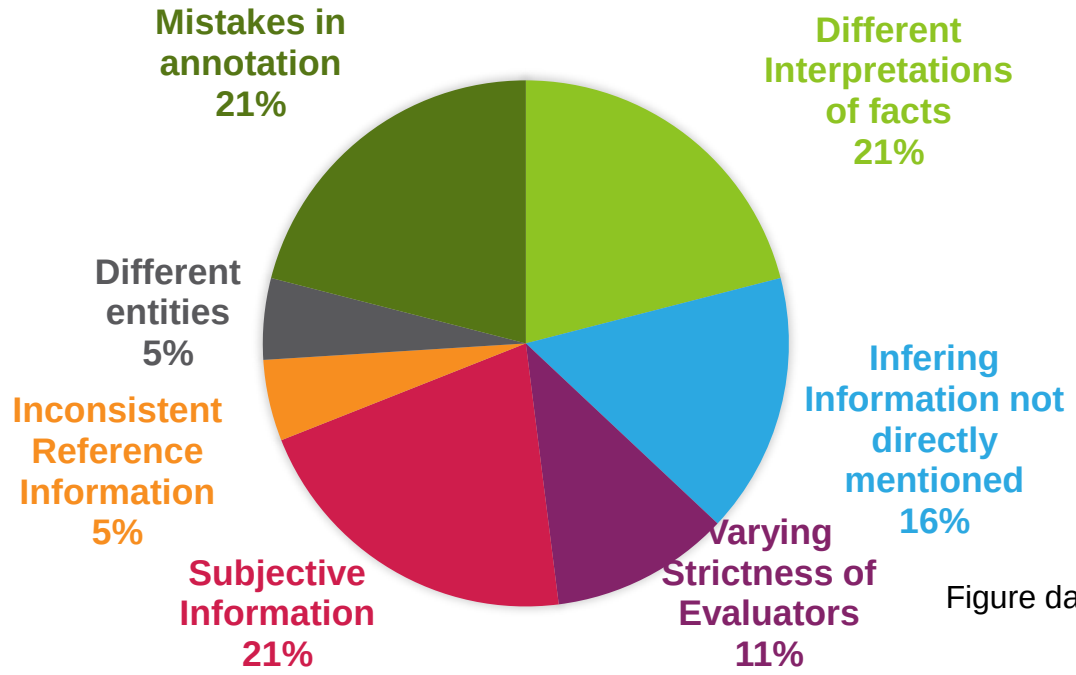


Figure data from [2].

Prompt: Tell me a bio of Ylona Garcia.
Sentence: [Ylona Garcia] has since appeared in various TV shows such as ASAP (All-Star Sunday Afternoon Party), Wansapanataym Presents: Annika PINTAsera and Maalaala Mo Kaya.
- Ylona Garcia has appeared in various TV shows. `Supported`
- She has appeared in ASAP. `Supported`
- ASAP stands for All-Star Sunday Afternoon Party. `Supported`
- ASAP is a TV show. `Supported`
- She has appeared in Wansapanataym Presents: Annika PINTAsera. `Not-supported`
- Wansapanataym Presents: Annika PINTAsera is a TV show. `Irrelevant`
- She has appeared in Maalaala Mo Kaya. `Not-supported`
- Maalaala Mo Kaya is a TV show. `Irrelevant`

Prompt: Tell me a bio of John Estes.
Sentence: William Estes is an American actor known for his role on CBS police drama Blue Bloods as Jameson J̈amieR̈eagan.
- William Estes is an American. `Irrelevant`
- William Estes is an actor. `Irrelevant`
- William Estes is known for his role on CBS police drama Blue Bloods. `Irrelevant`
- William Estes' role on Blue Bloods is Jameson "Jamie" Reagan. `Irrelevant`

Table 7: Examples that contain `Supported`, `Not-supported` and `Irrelevant`. Sentences in bullet points indicate atomic facts.

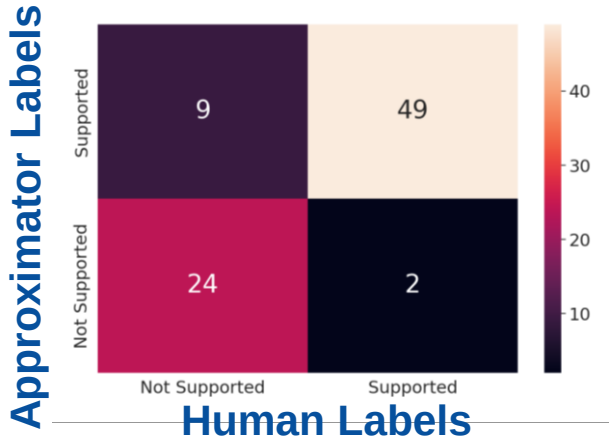
Figure from [2]

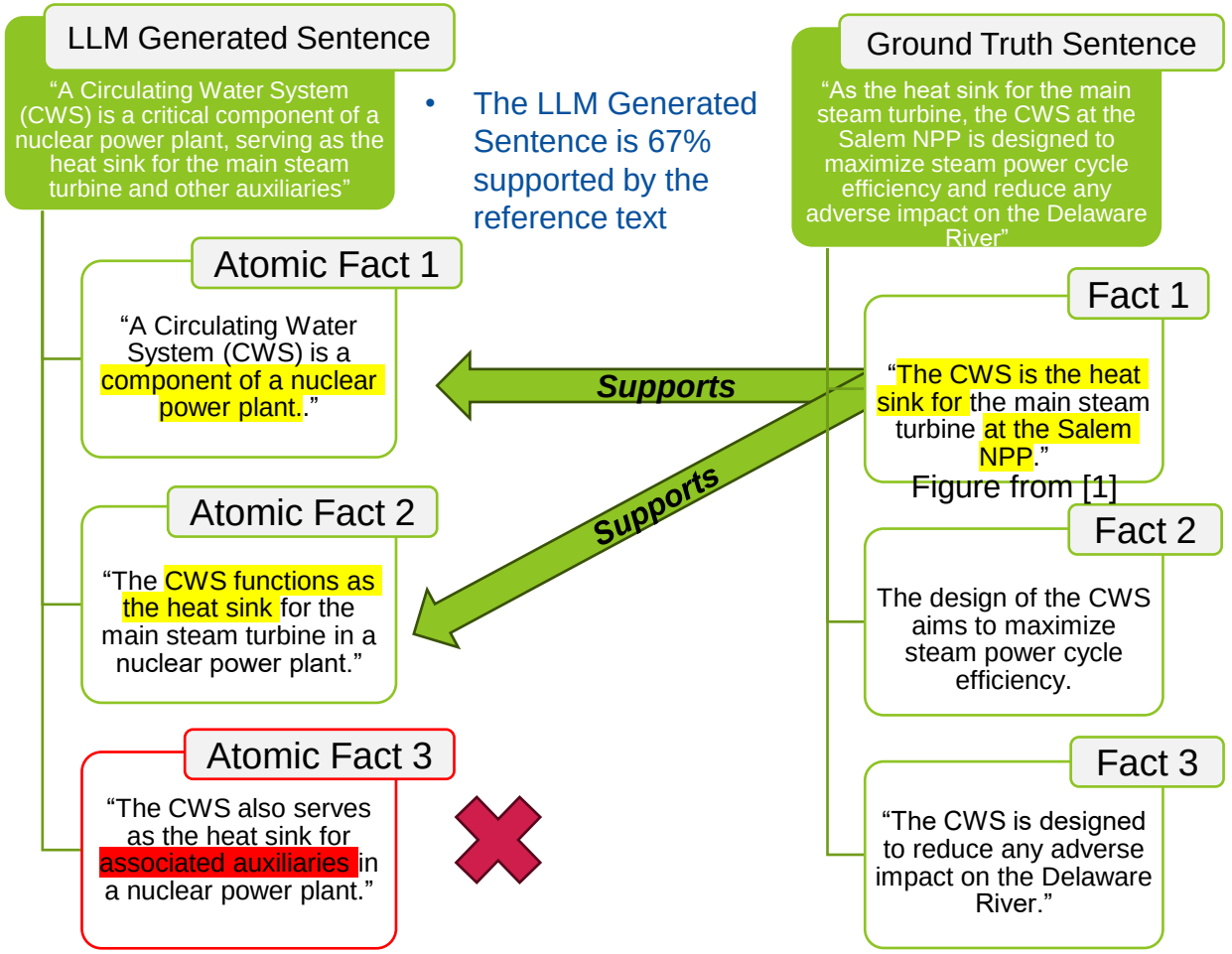
[2] Min, Sewon, et al. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." *arXiv preprint arXiv:2305.14251* (2023).

Human Evaluations can be approximated with reasonable accuracy.

- Reusing the LLM we can break a generation down into **Atomic Facts** [3].
- An **Atomic Fact** is the smallest claim that can be extracted from a sentence.
- A Natural Language Inference Model classifies pairs of generated facts and ground truth facts as supported or not supported.

Our Evaluator agrees with a human evaluator 87% of the time. Which is similar to the agreement between 2 humans (96%-88%) as reported in [2].





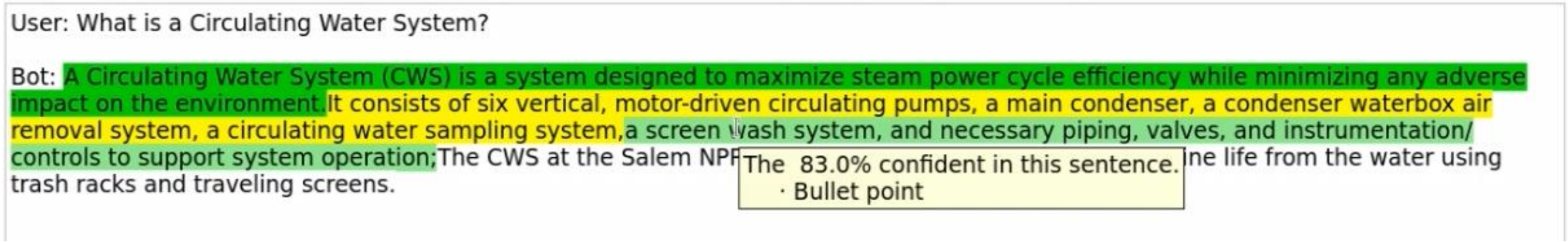
- The LLM Generated Sentence is 67% supported by the reference text

Figure from [1]

[2] Min, Sewon, et al. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." *arXiv preprint arXiv:2305.14251* (2023).
[3] Kamoi, Ryo, et al. "Wice: Real-world entailment for claims in wikipedia." *arXiv preprint arXiv:2303.01432* (2023).

Retrieval Augment Generation Improves LLM Factual Accuracy

- We ask the LLM 5 questions related to predictive maintenance and evaluate the response with our fact evaluation approximator.
- We find that our **RAG pipeline increases accuracy from 30% to 71% accuracy** according to our fact evaluation approximator.

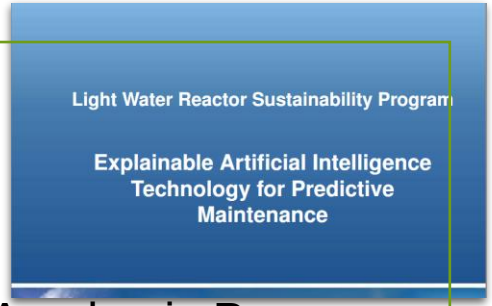


RAG and Automatic Fact Evaluation can be used in almost any domain

- Our RAG and Fact Evaluation are not domain specific and can be applied to any other domain with text-based references.
- We experimented with multiple types of nonstandard text-based inputs like Work Orders, and multi modal documents.

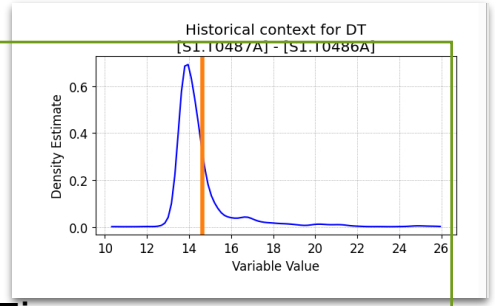
Databases

A valve is missing what should I do?



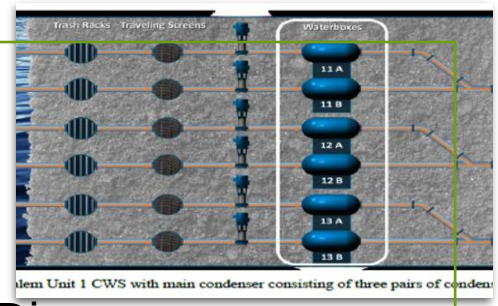
Academic Papers

What is a Decision Tree?



Figures

Is the Differential Temperature in a normal range?



Diagrams

Where does water go after the trash racks?

Questions?

References:

- [1] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." arXiv preprint arXiv:2311.05232 (2023).
- [2] Min, Sewon, et al. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." arXiv preprint arXiv:2305.14251 (2023).
- [3] Kamoi, Ryo, et al. "Wice: Real-world entailment for claims in wikipedia." arXiv preprint arXiv:2303.01432 (2023).

Idaho National Laboratory

Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy. INL is the nation's center for nuclear energy research and development, and also performs research in each of DOE's strategic goal areas: energy, national security, science and the environment.