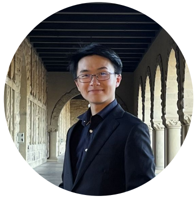
VERA: Verifiable and Explainable Retrieval-Augmented Generation for Educational Reliability in History Textbooks

Charlotte Zhu



Andrew Kuik

Sejoon Chang

Table of contents

01

Related Work

RAG & Hallucinations & Warning Systems

02

VERA

A tailored warning approach

03

User Study

Hallucination Detection in TQA

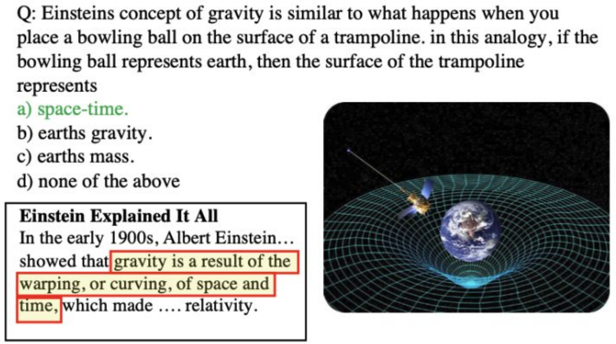
04

Results

performs better with limitations

Retrieval-Augmented Generation (RAG) in education: subjects to biases and hallucinations.

textbook question-answering (TQA)

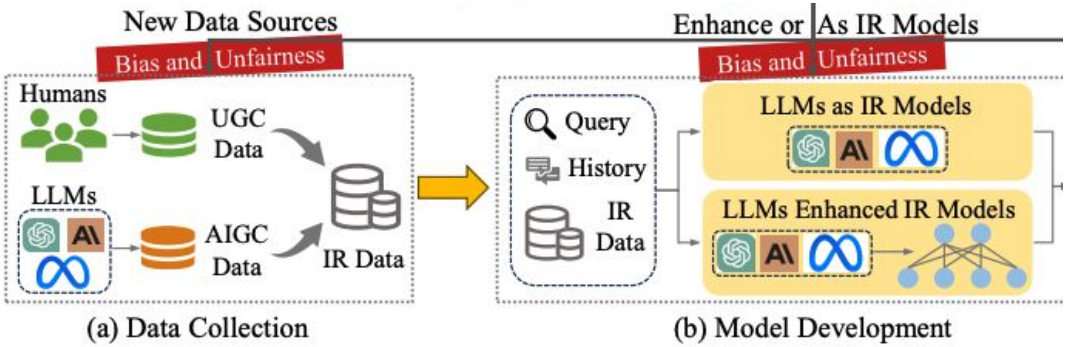


Q: Einsteins concept of gravity is similar to what happens when you place a bowling ball on the surface of a trampoline. in this analogy, if the bowling ball represents earth, then the surface of the trampoline represents

a) space-time.
b) earths gravity.
c) earths mass.
d) none of the above

Einstein Explained It All
In the early 1900s, Albert Einstein… showed that gravity is a result of the warping, or curving, of space and time, which made …. relativity.

Enhancing Textbook Question Answering Task with Large Language Models and Retrieval Augmented Generation



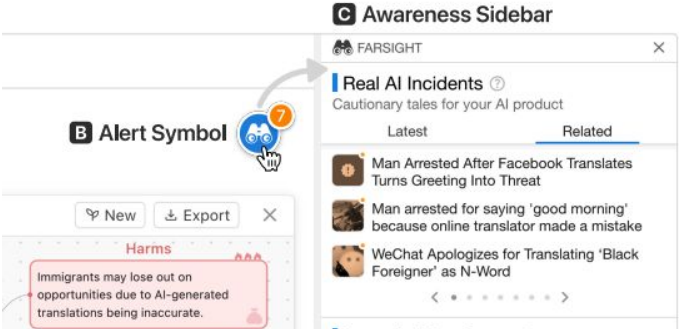
(a) Data Collection

(b) Model Development

Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era

Alawwad et al. 2024; Dai et al. 2024

Warning messages prevent harm in human interactions with LLM-based tools.



Real AI Incidents

Cautionary tales for your AI product

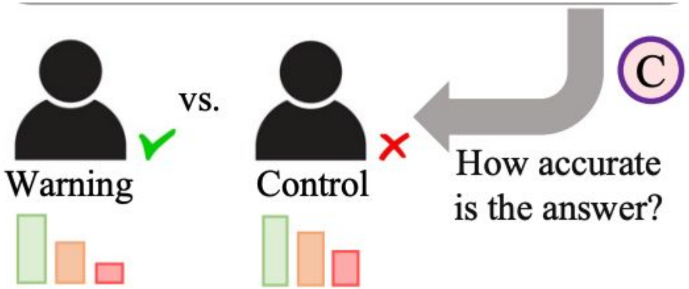
Man Arrested After Facebook Translates Turns Greeting Into Threat

Man arrested for saying 'good morning' because online translator made a mistake

WeChat Apologizes for Translating 'Black Foreigner' as N-Word

Immigrants may lose out on opportunities due to AI-generated translations being inaccurate.

FARSIGHT: Fostering Responsible AI Awareness During AI Application Prototyping



Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations

Wang et al. 2024; Nahar et al. 2024

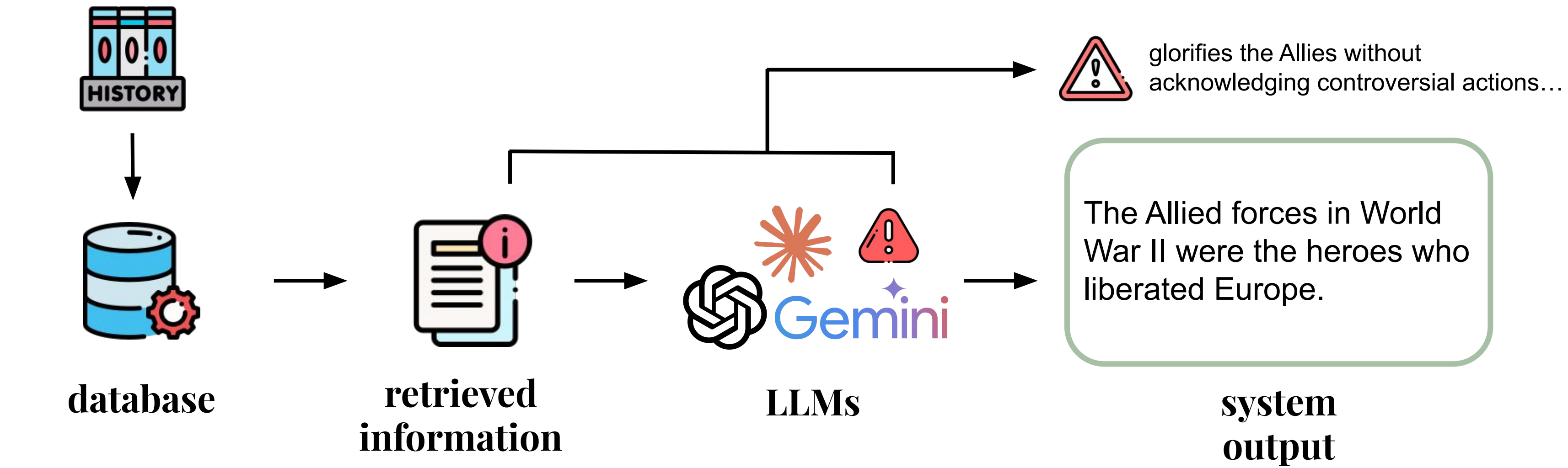
We propose a user- and context-tailored warning approach.

VERA

a **tailored warning system** that actively identifies and alerts users to potential hallucinations or biases in the generated responses

a **comparative user study** to assess the effectiveness of tailored warning system against current baselines.

Tailored warnings generated from biases detected in retrieved info and LLM responses.



glorifies the Allies without acknowledging controversial actions…

The Allied forces in World War II were the heroes who liberated Europe.

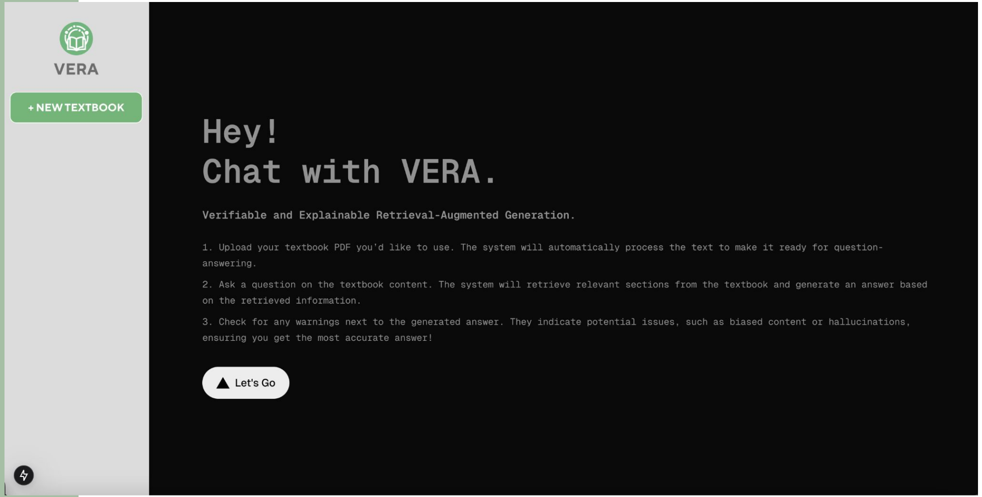
database

retrieved information

LLMs

system output

Our Platform & Tailored Warning Example



Hey!
Chat with VERA.

Verifiable and Explainable Retrieval-Augmented Generation.

1. Upload your textbook PDF you'd like to use. The system will automatically process the text to make it ready for question-answering.
2. Ask a question on the textbook content. The system will retrieve relevant sections from the textbook and generate an answer based on the retrieved information.
3. Check for any warnings next to the generated answer. They indicate potential issues, such as biased content or hallucinations, ensuring you get the most accurate answer!

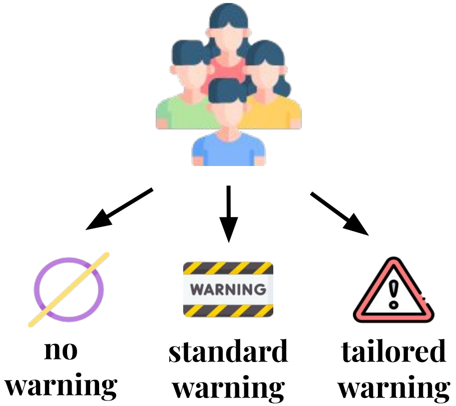
Q: Which battle was considered a turning point in the American Revolutionary War? Battle of Yorktown Battle of Bunker Hill Battle of Saratoga Battle of New Orleans

A: The battle considered a turning point in the American Revolutionary War was the Battle of Bunker Hill.

Textbook Reference Quote: While the Battle of Saratoga in 1777 is often cited as the turning point due to its significant impact on foreign support for the revolutionaries, the Battle of Bunker Hill was notable for building American confidence early in the war.

Warning: The retrieved textbook reference seems incorrect. The Battle of Saratoga is widely recognized as the turning point of the American Revolutionary War, not the Battle of Bunker Hill.

User Study: tailored warning vs. no warning & standard warning



Participants

No hallucination or obvious bias.

Low hallucinations & biases.

High hallucinations & biases.

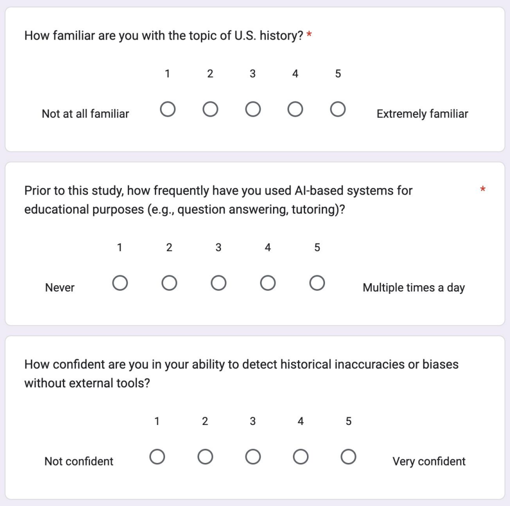
History Questionnaire Task

Correctness

Hallucination detection

Measurements

User Study Walkthrough



How familiar are you with the topic of U.S. history? *

	1	2	3	4	5	
Not at all familiar	○	○	○	○	○	Extremely familiar

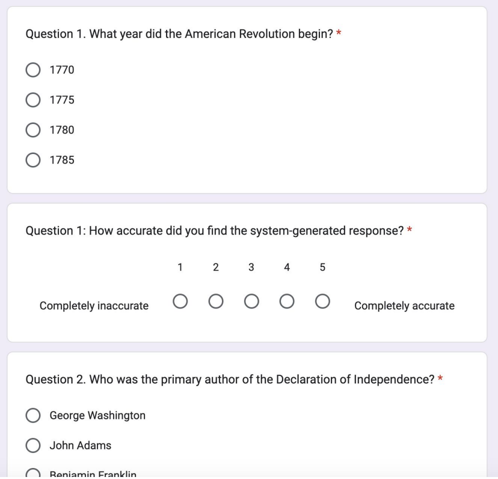
Prior to this study, how frequently have you used AI-based systems for educational purposes (e.g., question answering, tutoring)? *

	1	2	3	4	5	
Never	○	○	○	○	○	Multiple times a day

How confident are you in your ability to detect historical inaccuracies or biases without external tools?

	1	2	3	4	5	
Not confident	○	○	○	○	○	Very confident

Demographics Questions



Question 1. What year did the American Revolution begin? *

○ 1770

○ 1775

○ 1780

○ 1785

Question 1: How accurate did you find the system-generated response? *

	1	2	3	4	5	
Completely inaccurate	○	○	○	○	○	Completely accurate

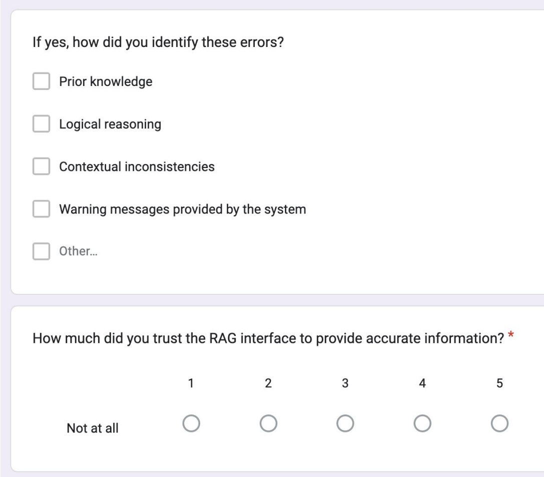
Question 2. Who was the primary author of the Declaration of Independence? *

○ George Washington

○ John Adams

○ Benjamin Franklin

Quiz Questions



If yes, how did you identify these errors?

□ Prior knowledge

□ Logical reasoning

□ Contextual inconsistencies

□ Warning messages provided by the system

□ Other...

How much did you trust the RAG interface to provide accurate information? *

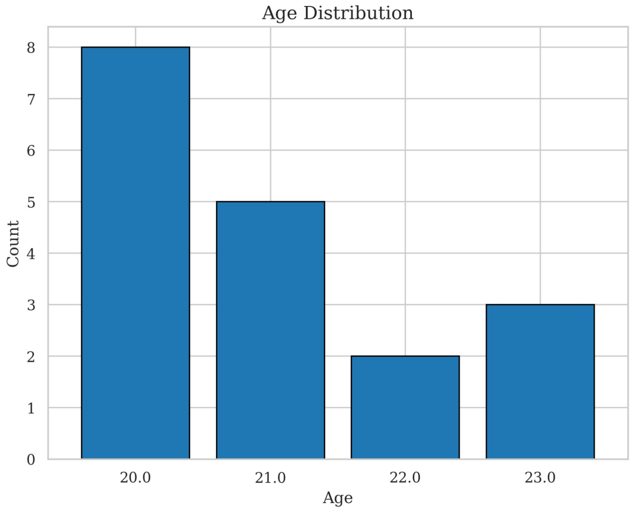
	1	2	3	4	5
Not at all	○	○	○	○	○

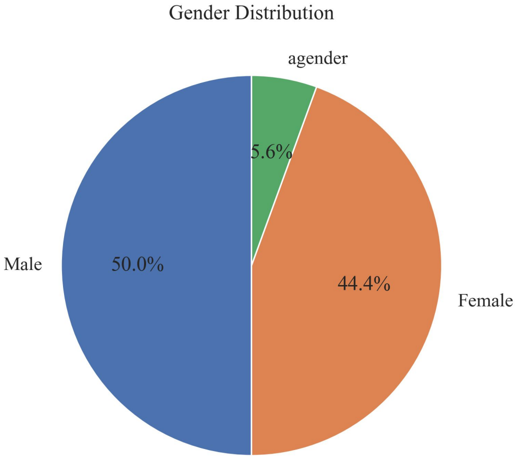
Post-Survey Questions

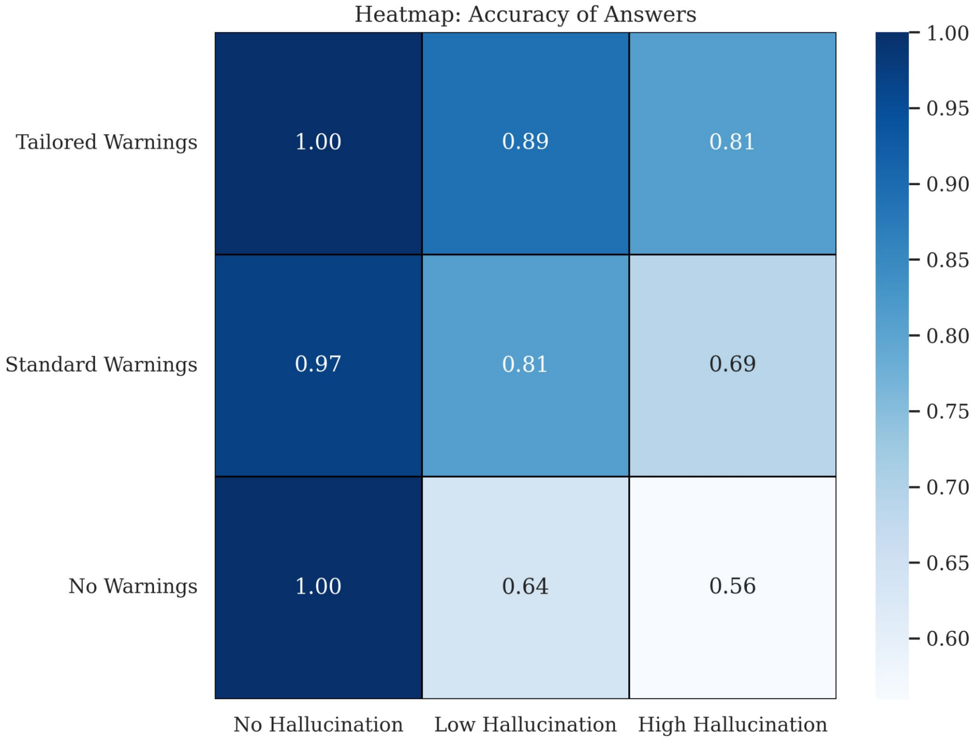
Survey Results

Survey: Participant Demographics

A total of 18 participants were recruited.
College students w/ various background & familiarity with US history.

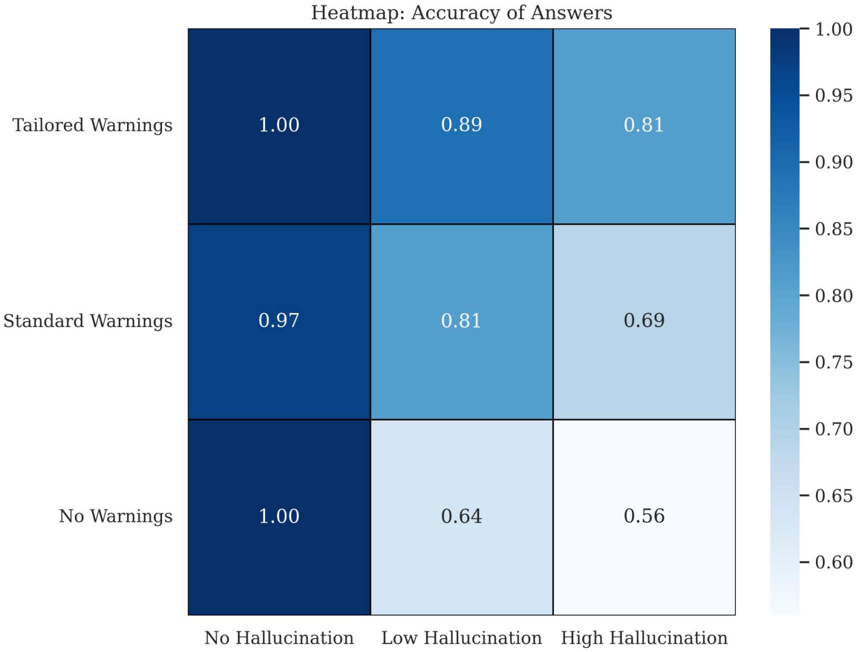




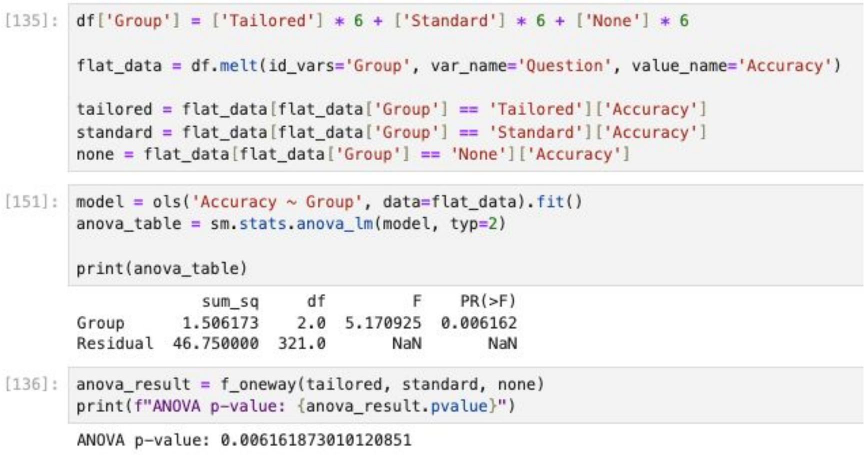


Tailored Warning result in higher user accuracy across different levels of hallucinations.

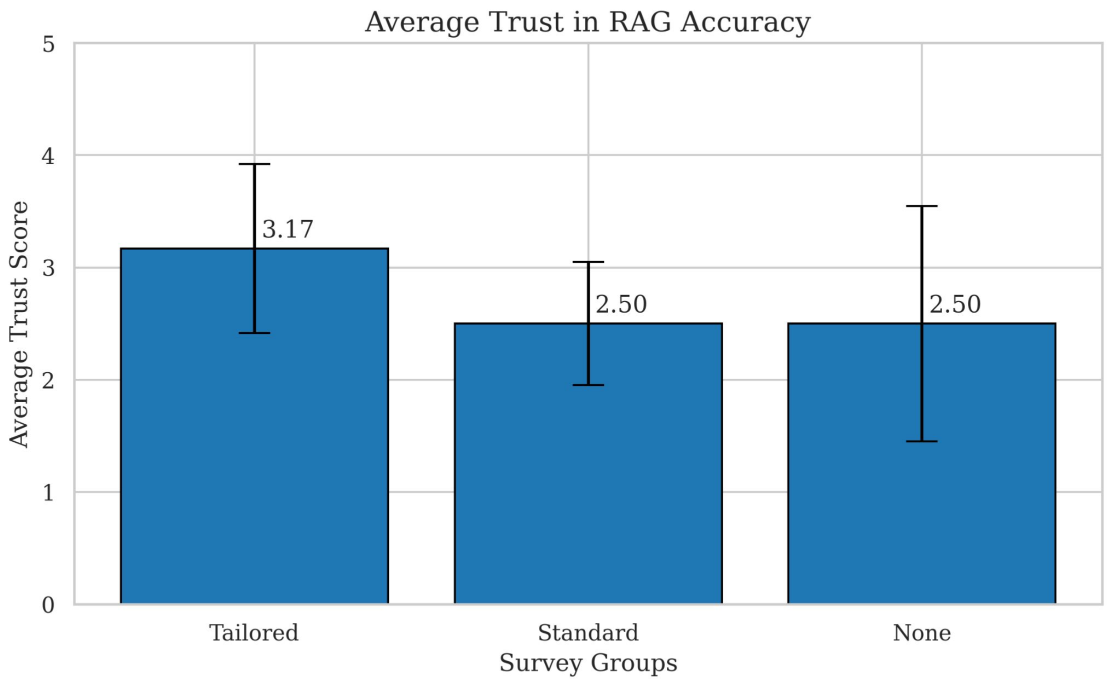
No. of participants: 18



Statistical Significance
ANOVA p-value: 0.0062

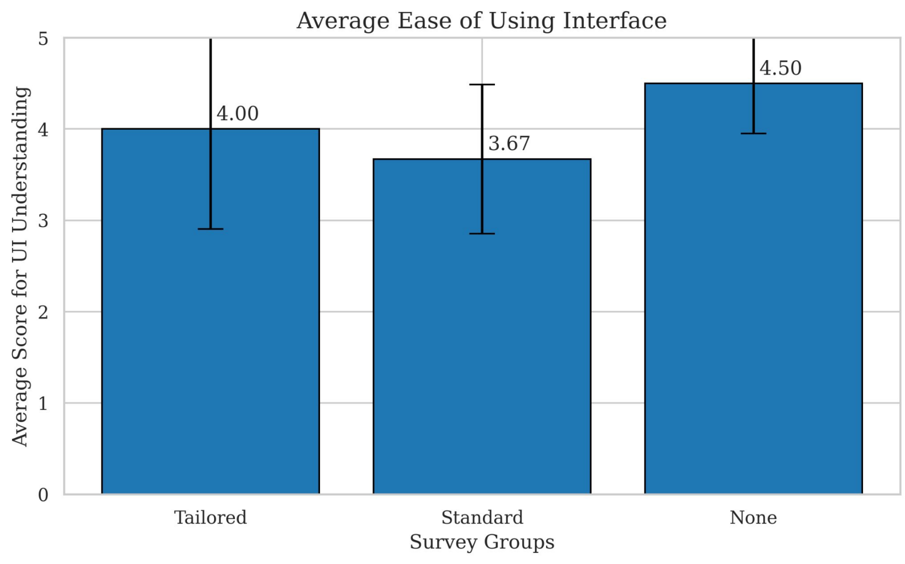


```
[135]:  df['Group'] = ['Tailored'] * 6 + ['Standard'] * 6 + ['None'] * 6

        flat_data = df.melt(id_vars='Group', var_name='Question', value_name='Accuracy')

        tailored = flat_data[flat_data['Group'] == 'Tailored']['Accuracy']
        standard = flat_data[flat_data['Group'] == 'Standard']['Accuracy']
        none = flat_data[flat_data['Group'] == 'None']['Accuracy']

[151]:  model = ols('Accuracy ~ Group', data=flat_data).fit()
        anova_table = sm.stats.anova_lm(model, typ=2)

        print(anova_table)

                     sum_sq     df         F    PR(>F)
        Group      1.506173    2.0  5.170925  0.006162
        Residual  46.750000  321.0       NaN       NaN

[136]:  anova_result = f_oneway(tailored, standard, none)
        print(f"ANOVA p-value: {anova_result.pvalue}")

        ANOVA p-value: 0.006161873010120851
```



Tailored Warning led to better detection of hallucination & higher overall trust of the system.

However, tailored warning causes user confusions.

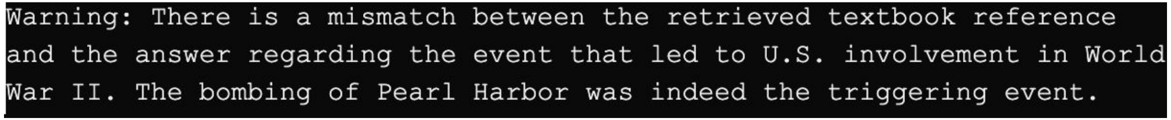


"It's **confusing**... I just don't really know, like, why we need, like, these two, three sections… **why don't it just give the correct answer?**"

"The warning definitely **slows me down**…I need to go back and double check."

Future Work

Which warning interface?



```
Warning: There is a mismatch between the retrieved textbook reference
and the answer regarding the event that led to U.S. involvement in World
War II. The bombing of Pearl Harbor was indeed the triggering event.
```

Longer? Shorter?

Warning format (User reactions)

Degree and length of warnings

References

- Hessa A. Alawwad et al. 2024. Enhancing Textbook Question Answering Task with Large Language Models and Retrieval Augmented Generation. arXiv preprint arXiv:2402.05128v2 (2024).
- Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations. arXiv preprint arXiv:2404.03745 (2024).
- Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung-yi Lee, and Lama Nachman. 2024. Decoding Biases: Automated Methods and LLM Judges for Gender Bias Detection in Language Models. arXiv preprint arXiv:2408.03907 (2024).
- Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigat- ing Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception. arXiv preprint arXiv:2403.14896 (2024).
- Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Appli- cation Prototyping. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–40.
- Sunhao Dai et al. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. arXiv preprint arXiv:2404.11457v2 (2024).