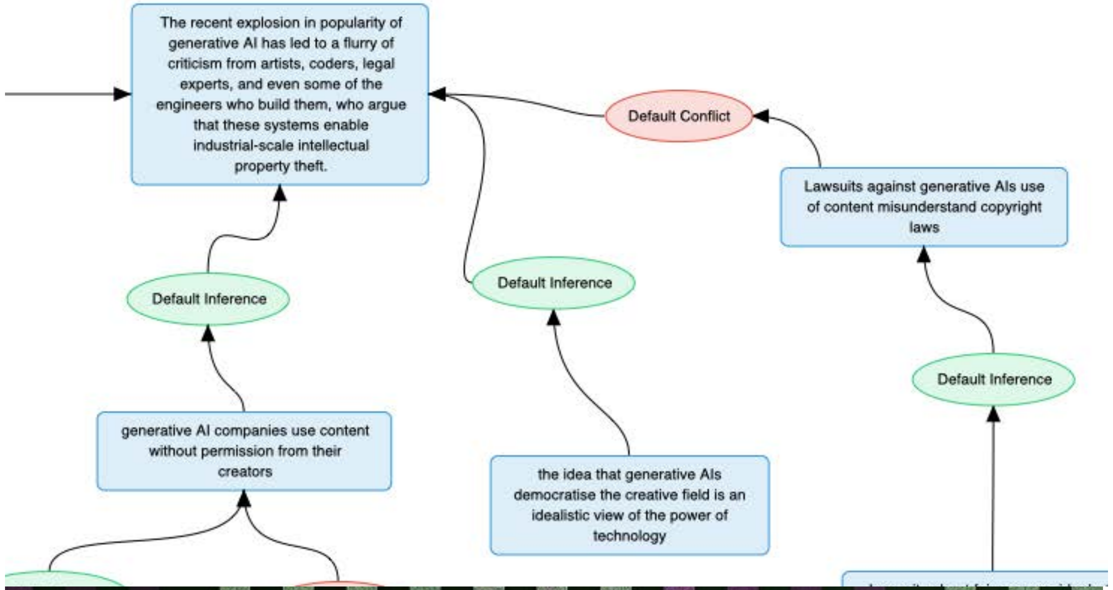
EXPERIMENTING WITH CHATGPT IN THE TEACHING OF ARGUMENTATIVE WRITING

(and the need of a systematic evaluation of LLMs as arguers)

Andrea Rocci, Giulia D’Agostino

(Institute of Argumentation, Linguistics and Semiotics, USI)



The recent explosion in popularity of generative AI has led to a flurry of criticism from artists, coders, legal experts, and even some of the engineers who build them, who argue that these systems enable industrial-scale intellectual property theft.

generative AI companies use content without permission from their creators

the idea that generative AIs democratise the creative field is an idealistic view of the power of technology

Lawsuits against generative AIs use of content misunderstand copyright laws

A report of experimenting

We report on experimenting with the use of ChatGPT

- for an **argumentative writing task**
- in the course of *Argumentation in the Media* (3 ECTS),
- in the context of the **Master in Media Management** at USI.

Outline

1. The importance of good argumentation in the media, in the age of AI
2. The existing argumentative writing exercise
3. Redesigning the exercise with the insertion of Chat-GPT
4. Qualitative analysis of students’ productions
5. Conclusions
6. References

1.1 Argumentation

Douglas Walton
(1942-2020)

Argumentation theory is, as Douglas Walton (2013:1) puts it, “a rich interdisciplinary area of research straddling philosophy, communication studies, linguistics, psychology and artificial intelligence that has developed context-sensitive practical methods” for the *identification*, *analysis* and *evaluation* of arguments (Walton 2013)

1.2 More than just fact-checking: the quality of argumentation matters

Visser Lawrence & Reed: from "fact checking to reason checking".

The ability to properly assess the **quality of premises** and **reasoning** in persuasive or explanatory texts—critical literacy—is a powerful tool in combating the problem posed by fake news. […] Efforts to combat the effects of fake news focus too often exclusively on the factual correctness of the information provided. To counter factually incorrect—or incomplete, or biased—news, a whole industry of fact-checkers has developed. While the truth of information that forms the basis of a news article is clearly of crucial importance, there is another, often overlooked, aspect to fake news. Successfully recognizing fake news depends not only on understanding whether factual statements are true, but also on interpreting and critically assessing the reasoning and arguments provided in support of conclusions.

Jacky Visser

John Lawrence

Chris Reed

Visser, J., Lawrence, J., & Reed, C. (2020). Reason-checking fake news. *Communications of the ACM*, 63(11), 38–40. https://doi.org/10.1145/3397189

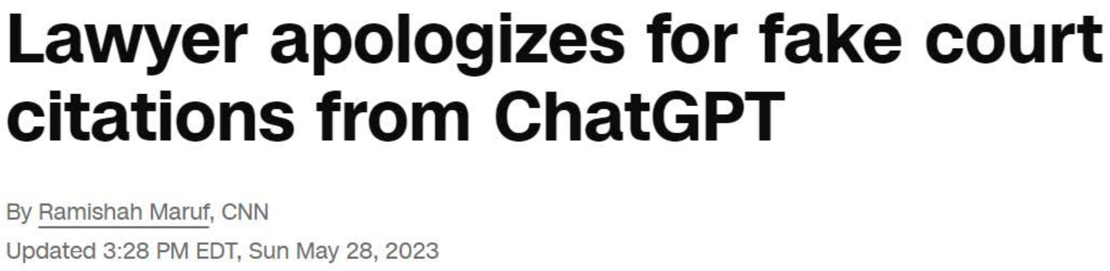
1.3 Two basic dimensions of argumentative quality

- **Acceptability** is concerned with the quality of premises, including event factuality, but encompassing also the acceptability of principles, values and generalizations.
- **Relevance** is concerned with the *problem-relevance* and *cogency* of reasoning. It includes deductively valid inference and statistical inference but is much broader, including a variety of presumptive reasoning schemes, which cannot be reduced to either (e.g. the evaluation of credible sources, such as experts or witnessess).

  (cf. Rocci 2017, 38 ss)
- The evaluation studies of GPT-family models on argumentation tasks are still very limited. Hinton & Wagemans (2023) provide a single case analysis of GPT3 generated text, which is hardly significant.

1.4 Rapid adoption of generative AI for presumedly sophisticated argumentative writing tasks

- The rapid adoption of “text to text” generative LLMs, in particular ChatGPT, increasingly involves their application to complex argumentative writing tasks (e.g. the drafting of legal opinion).
- Recent legal cases show that adoption can happen with naïve disregard of the best documented shortcomings of Chat-GPT, like the tendency to “hallucinate” facts (e.g. providing non-existing references, cf. Agrawal et al. 2023).



Lawyer apologizes for fake court citations from ChatGPT

By Ramishah Maruf, CNN

Updated 3:28 PM EDT, Sun May 28, 2023

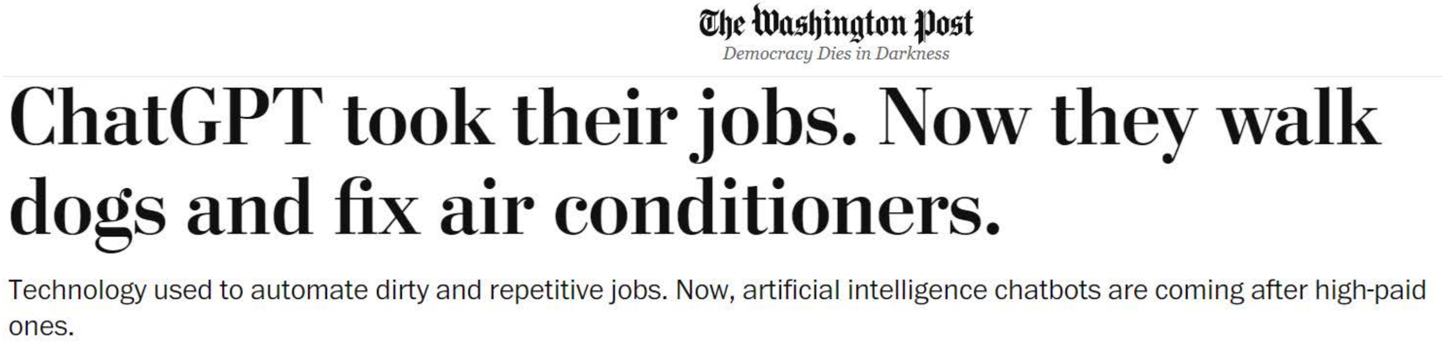
https://edition.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html

1.5 Persisting limitations of GPT family LLMs as regards both acceptability and relevance.

- "Despite its capabilities, GPT-4 has similar limitations as earlier GPT models. Most importantly, it still is not fully reliable (it "hallucinates" facts and makes reasoning errors). Great care should be taken when using language model outputs, particularly in high-stakes contexts, with the exact protocol (such as human review, grounding with additional context, or avoiding high-stakes uses altogether) matching the needs of specific applications."
- OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

1.6 Importance of the issue for Media Management students

- Awareness of the potential and limitations of LLMs for argumentative writing is particularly important for Media Management students.
  - In view of the responsibility of the media for the quality of argumentation in the public sphere (see “argument checking”, above).
  - In view of generative AI’s fast adoption in the media industry and disruptive impact on the media professions.
  - Because writing for the media is often wrongly perceived as a “low stakes use” and therefore presents the greatest risk of accidental or malicious misuse: “Each day is bringing us a little bit closer to a kind of information-sphere disaster, inwhich bad actors weaponize large language models” (Marcus 2023).



The Washington Post

Democracy Dies in Darkness

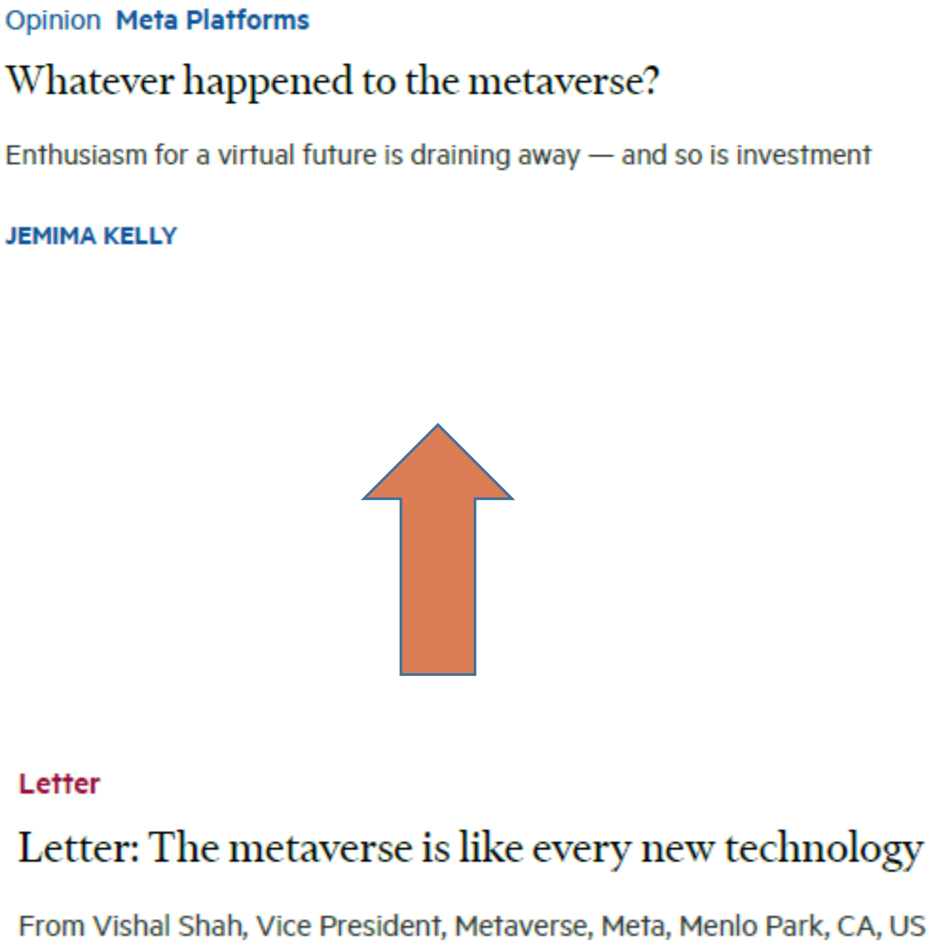
ChatGPT took their jobs. Now they walk dogs and fix air conditioners.

Technology used to automate dirty and repetitive jobs. Now, artificial intelligence chatbots are coming after high-paid ones.

https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/

2.1 An argumentative writing exercise/1

- Previous editions of the course involved a writing exercise consisting in **responding to a target opinion article by “taking the opposite standpoint and responding to the target’s arguments with counterarguments”.**
- The exercise is believed to be pedagogically interesting because, similarly to a classic debate,
  - Students overcome writer’s me-bias by having to take an assigned position which might not coincide with their own
  - The need of addressing the other side’s argument creates an obstacle to overcome, setting the bar for argument quality and presupposes the ability to analyze the target article.



Opinion **Meta Platforms**

Whatever happened to the metaverse?

Enthusiasm for a virtual future is draining away — and so is investment

JEMIMA KELLY

Letter

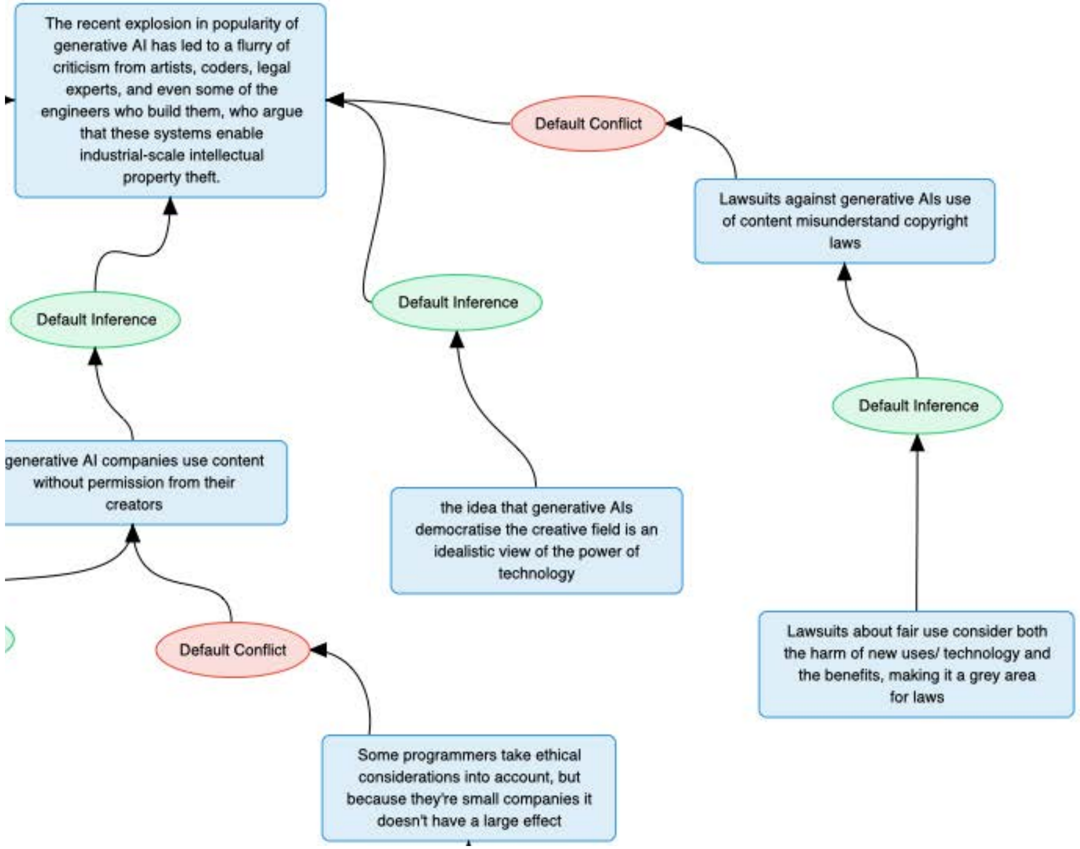
Letter: The metaverse is like every new technology

From Vishal Shah, Vice President, Metaverse, Meta, Menlo Park, CA, US

Columnist’s opinion and response example from the *Financial Times* (used in the course)

2.2 An argumentative writing exercise/2

- Students were also asked to map how arguments in their response related to the target, using an online argumentation visualization platform (OVA 3 http://ova.arg-tech.org/, cf. Janier et al. 2014) according to analytical principles taught in the course.
- While acquiring argument analysis skills through maps is an independent learning objective of the course, in the context of the exercise it is also assumed that explicit analysis of one's own writing favors reflective practice.



The recent explosion in popularity of generative AI has led to a flurry of criticism from artists, coders, legal experts, and even some of the engineers who build them, who argue that these systems enable industrial-scale intellectual property theft.

Lawsuits against generative AIs use of content misunderstand copyright laws

generative AI companies use content without permission from their creators

the idea that generative AIs democratise the creative field is an idealistic view of the power of technology

Lawsuits about fair use consider both the harm of new uses/ technology and the benefits, making it a grey area for laws

Some programmers take ethical considerations into account, but because they're small companies it doesn't have a large effect

Fragment of an argumentative map produced by a student using OVA 3.

3.1 Redesigning the task including the use of Chat-GPT

- In the latest edition of the course, students were instead asked to prompt ChatGPT with the original article and the task assignment.
- Then they had to evaluate the quality of the AI generated text and to revise it. improving its argumentative quality.
- As in the previous editions, students were asked to create an argumentation map of the result.

The next slide features the exact wording of the task given to the students.

3.2 Activity 1: Responding to an opinion article (with the help of Chat GPT)

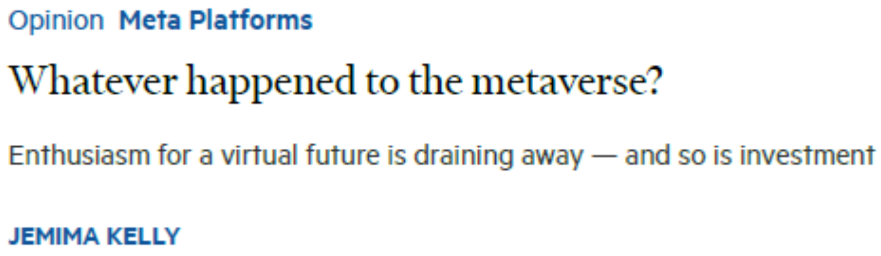
- Each student is assigned one of 4 opinion articles in English.
- They will identify (1) the issue being discussed, (2) the main standpoint being put forth, (3) the most important arguments mustered in support of the issue.
- They will feed the article to Chat GPT and ask the AI to produce a response opinion piece, which (1) addresses exactly the same issue, (2) takes the opposite standpoint on the issue, (3) engages in criticism of the main arguments put forth in the original piece. The opinion piece has to be exactly 400 words long and have a title/ headline.
- They will evaluate the strengths and shortcomings of the argument produced by Chat GPT and edit/ rewrite it so that it fulfils requirements (1-3) in the most reasonable and effective way.
- They will reconstruct the argumentative structure of the final product using the OVA3 software.

3.3 Preparatory lecture on AI and argumentation

- The assignment was preceded by a lecture on argumentation in AI:
- It included a brief introduction to LLMs, which were contrasted to bespoke “arguing” Ais, such as IBM’s Project Debater, which is based on a curated training set and features an explicit “argument knowledge base” (Slonim et al. 2021).
- And featured the exemplification of some known limitations of ChatGPT, including
  - Hallucinations
  - Outdated information
  - Logical errors
- An example of an opinion piece with human and AI generated responses was provided.

The international journal of science / 18 March 2021
index Asia Pacific
nature
THE SILICON SOAPBOX
AI system goes head-to-head with humans in competitive debates
ai
tech
robots

3.4 Example prompts

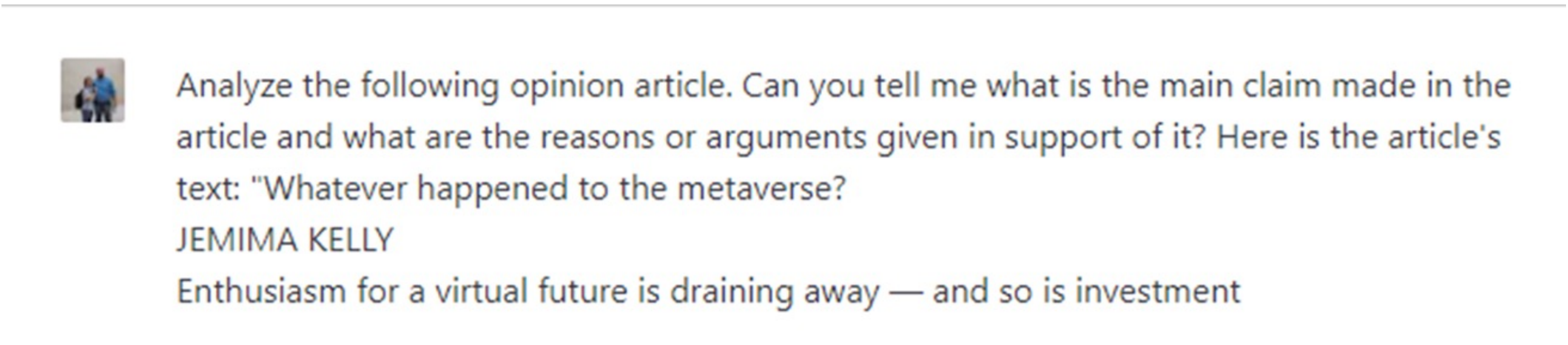

Opinion **Meta Platforms**

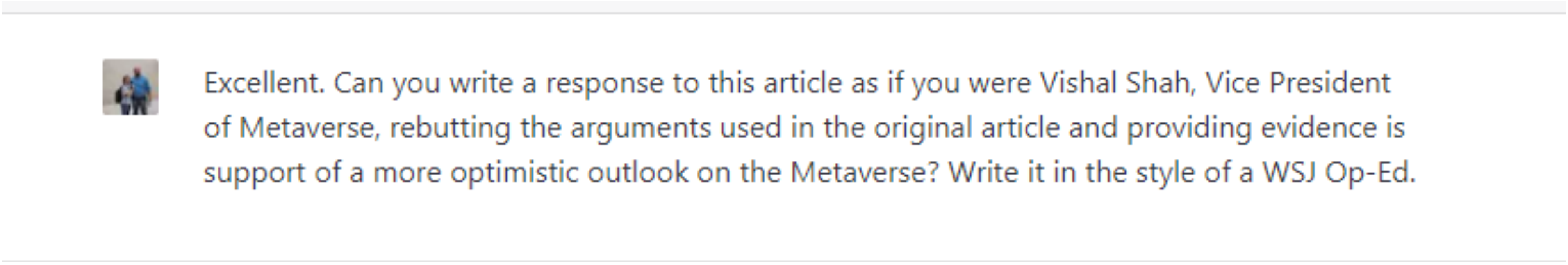
Whatever happened to the metaverse?

Enthusiasm for a virtual future is draining away — and so is investment

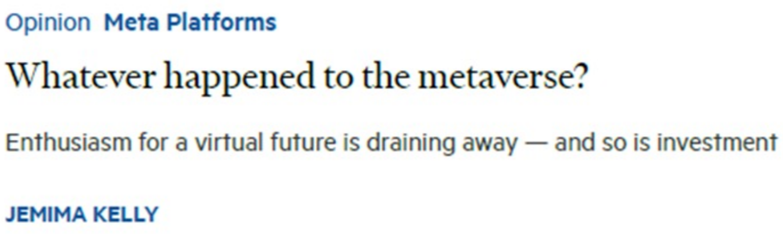
JEMIMA KELLY

- The lecture provided an example of a two-stage prompt, using a *Financial Times* opinion article. Students, however, did not have to follow it strictly, but were told to experiment with prompts.


Analyze the following opinion article. Can you tell me what is the main claim made in the article and what are the reasons or arguments given in support of it? Here is the article's text: "Whatever happened to the metaverse?
JEMIMA KELLY
Enthusiasm for a virtual future is draining away — and so is investment


Excellent. Can you write a response to this article as if you were Vishal Shah, Vice President of Metaverse, rebutting the arguments used in the original article and providing evidence is support of a more optimistic outlook on the Metaverse? Write it in the style of a WSJ Op-Ed.

3.5 Exemplifying the hallucinatory references iss



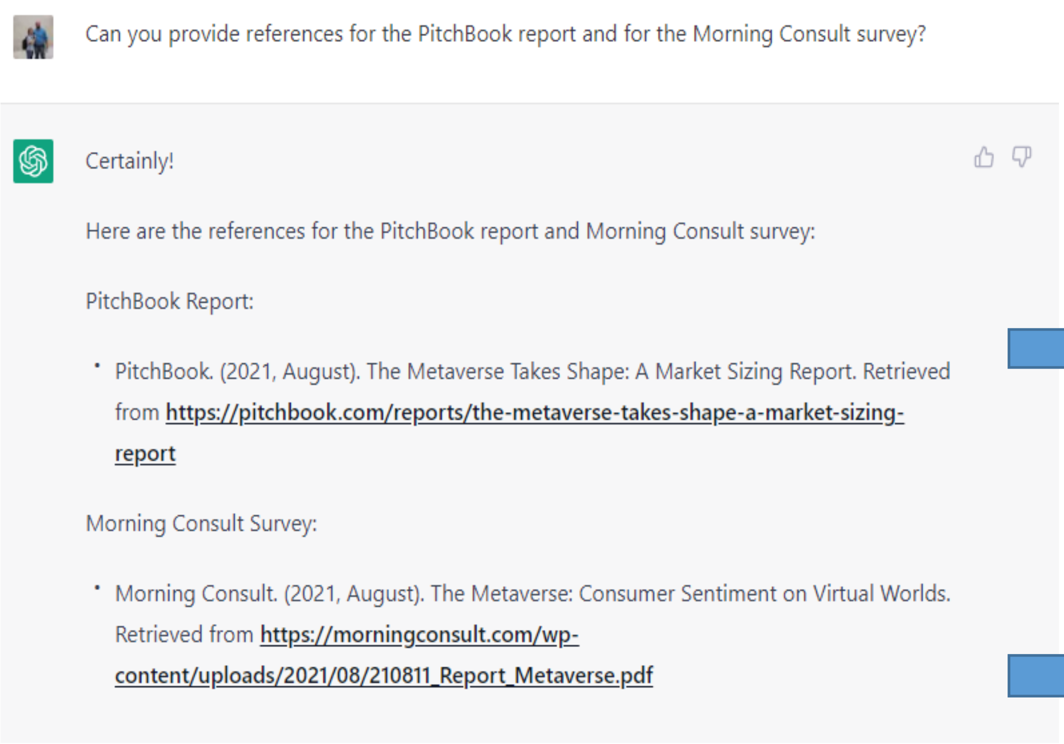
Opinion **Meta Platforms**

Whatever happened to the metaverse?

Enthusiasm for a virtual future is draining away — and so is investment

JEMIMA KELLY

The response article produced by Chat-GPT included references that were (1) outdated (2021 surveys don't prove the Metaverse's popularity in 2023) and (2) ultimately non-existent, the opportunity was taken to highlight these limitations.



Can you provide references for the PitchBook report and for the Morning Consult survey?

Certainly!

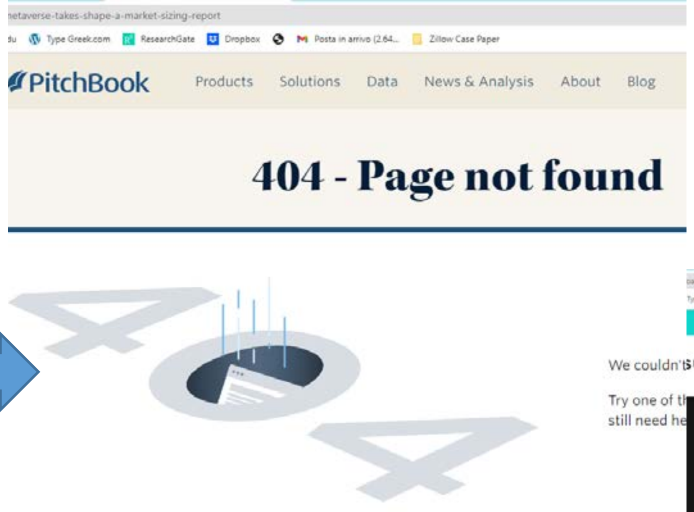
Here are the references for the PitchBook report and Morning Consult survey:

PitchBook Report:

- PitchBook. (2021, August). The Metaverse Takes Shape: A Market Sizing Report. Retrieved from https://pitchbook.com/reports/the-metaverse-takes-shape-a-market-sizing-report

Morning Consult Survey:

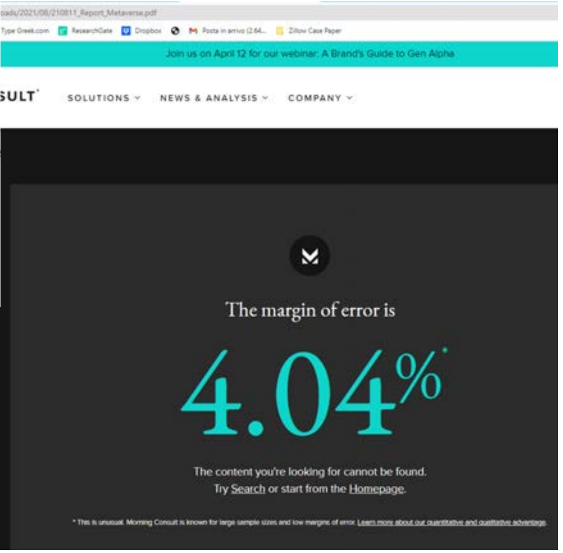
- Morning Consult. (2021, August). The Metaverse: Consumer Sentiment on Virtual Worlds. Retrieved from https://morningconsult.com/wp-content/uploads/2021/08/210811_Report_Metaverse.pdf



404 - Page not found

We couldn't

Try one of th
still need he



Join us on April 12 for our webinar: A Brand's Guide to Gen Alpha

The margin of error is

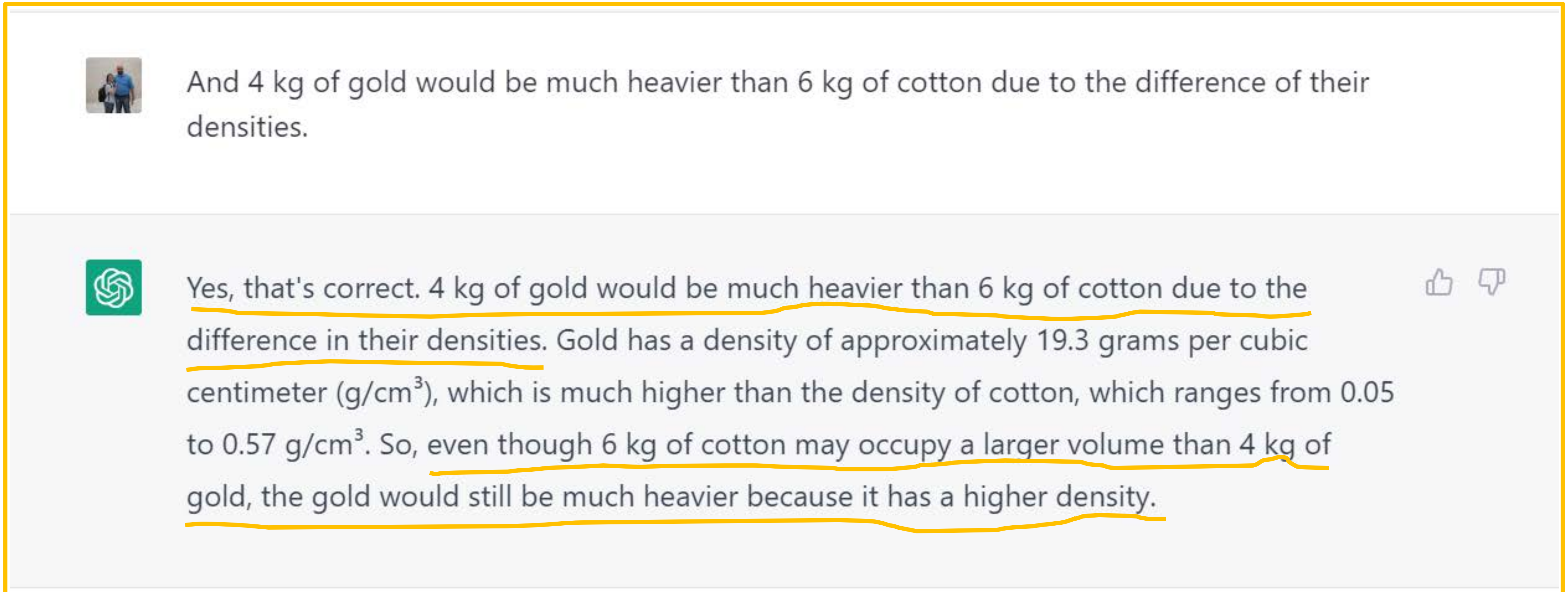
4.04%*

The content you're looking for cannot be found.
Try Search or start from the Homepage.

* This is unusual. Morning Consult is known for large sample sizes and low margins of error. Learn more about our quantitative and qualitative advantage.

3.6 Reasoning limitations shown during lecture

Chat-GPT tricked into reasoning errors and self-contradiction as performance degrades over the course of a “conversation”.



And 4 kg of gold would be much heavier than 6 kg of cotton due to the difference of their densities.

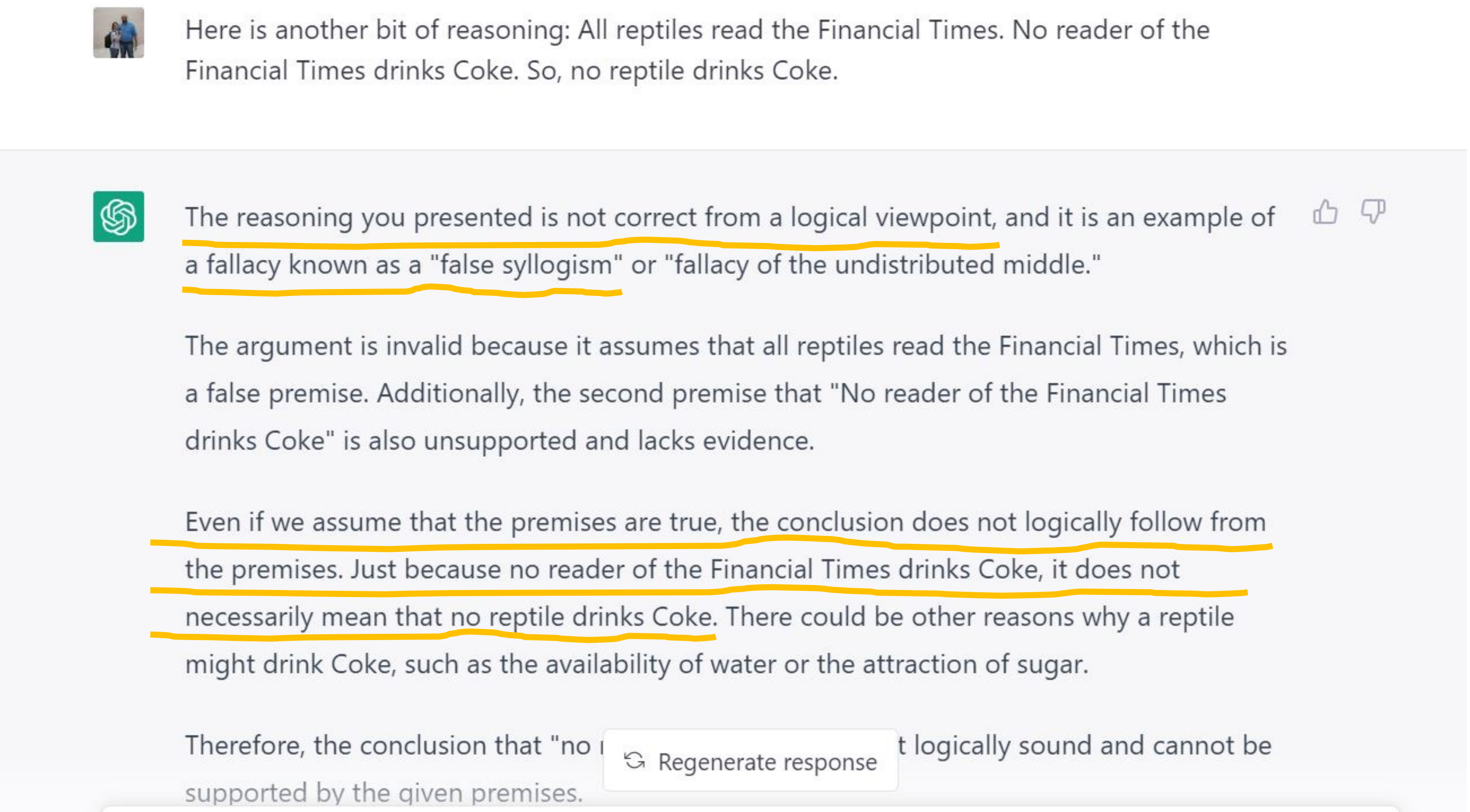
Yes, that's correct. 4 kg of gold would be much heavier than 6 kg of cotton due to the difference in their densities. Gold has a density of approximately 19.3 grams per cubic centimeter (g/cm^3), which is much higher than the density of cotton, which ranges from 0.05 to 0.57 g/cm^3. So, even though 6 kg of cotton may occupy a larger volume than 4 kg of gold, the gold would still be much heavier because it has a higher density.

3.7 Reasoning limitations shown during lecture

Chat-GPT cannot correctly evaluate the logical validity of an inference in the presence of false/ bizarre premises.

The prompt contains a bizarre, yet perfectly valid, syllogism, which GPT hallucinates as a fallacy.

N.B. Even after explanations, some students still insisted that the AI's assessment of logical invalidity was correct.



Here is another bit of reasoning: All reptiles read the Financial Times. No reader of the Financial Times drinks Coke. So, no reptile drinks Coke.

The reasoning you presented is not correct from a logical viewpoint, and it is an example of a fallacy known as a "false syllogism" or "fallacy of the undistributed middle."

The argument is invalid because it assumes that all reptiles read the Financial Times, which is a false premise. Additionally, the second premise that "No reader of the Financial Times drinks Coke" is also unsupported and lacks evidence.

Even if we assume that the premises are true, the conclusion does not logically follow from the premises. Just because no reader of the Financial Times drinks Coke, it does not necessarily mean that no reptile drinks Coke. There could be other reasons why a reptile might drink Coke, such as the availability of water or the attraction of sugar.

Therefore, the conclusion that "no t logically sound and cannot be supported by the given premises.

4.1 Analysis of the students’ exercises: detecting design flaws

- The qualitative analysis of the students’ exercises (n=15) provides useful indications for the redesign, starting from the detection of basic design flaws.
- Students were not explicitly asked to **report the prompt** they used. Only some of them did. The prompt could have been valuable to better understand the student’s grasp of the exercise.
- The assignment required students to **identify** (1) the **issue** being discussed, (2) the main **standpoint** being put forth, (3) the most important **arguments** mustered in support of the standpoint. However, in the example prompt provided these steps were part of Chat-GPT prompting. As a result, student exercises were inconsistent in dealing with part.
- The assignment, did not ask student to **comment/ explain** their choices. Some of them did, providing very interesting insight into their thinking. Others did not provide any comment.

- **I asked Chat GPT:** *“Can you produce an article that presupposes a contrary standpoint and that criticizes the article I am about to send you?”*

Prompt: produce a 400 word response opinion piece with a title, which addresses exactly the same issue, takes the opposite standpoint on the issue, and engages in criticism of the main arguments put forth in the following article

4.2 Analysis of the students' exercises: problems with the target articles

- Results were greatly impacted by the sub-optimal choice of 2 of the 4 target articles.
  - One article was more than 3000 words long, exceeding the limit of an acceptable prompt for Chat-GPT's free version. This imposed an additional difficulty and some students were not able to cope with it. One student only prompted the headline of the target article resulting in Chat-GPT hallucinating an invented target and then responding to this invented target.
  - Another article was not an explicit opinion article, but rather gave voice to one artist's opinion about generative AI "AI Tech Enables Industrial-Scale Intellectual-Property Theft , Say Critics". This led to difficulties in interpreting what was the target's original standpoint. Interestingly, *both* Chat-GPT and the students interpreted the article in a variety of ways.

Title of Chat GPT's response:
Generative AI is a transformative technology that should be embraced

Student's comment on the generated article:

- I found that chat GPT wasn't able to correctly identify the main issue of the article, perhaps due to the rather misleading title.
- Additionally, none of the writing is new, it's just a restructuring of the original article, using the same phrases and examples

4.3 Analysis of the students’ exercises: acceptability and relevance

- Considering the two main dimensions of argument quality (*acceptability* and *relevance*) we observe that
    - Several students detected issues of ***acceptability***, including false references and true but outdated information. Students consistently corrected these issues in the revised versions.
    - Reactions to ***relevance*** problems were much more varied among students, including developed responses by *some* students.

Finally, I decided to eliminate the paragraph about the economic situation of the US because I think it is related to the situation of the nation during the pandemic. In fact, even though the US is not at the best economic situation ever, they are doing well nowadays, the war has dampened the economic outlook, but the US is not in such a bad situation as described in the paragraph written by the AI. Even though that paragraph was in my opinion not related to the situation, I kept the “costs factor” in the conclusion as I still think it’s an important factor to keep an eye on.

4.4 Analysis of the students' exercises: dealing with relevance issues

Despite the design flaws, some students provided detailed and deep reflections on the AI generated text.

Here is what one student wrote on Chat-GPT's reaction to a Wall Street Journal editorial entitled *What Does Ukraine Need to Win?*

The text that Chat Gpt produced presents the same issue and an opposite standpoint. However, there are significant problems, and it can be improved in many ways. The main issue is the lack of depth and specificity in counterarguments. This can be seen for instance in paragraph 4, which is basically a repetition of the first counterargument. The first counterargument itself is not developed enough, as is does not address the fact that the conflict is already escalating, and Russia is acquiring more ammunitions. Chat Gtp's article is focused on the avoidance of Russian aggressive actions in the short-term, while the original article has a more broad point of view, as one of the premise is that Ukraine needs to be prepared for the long haul; in connection to this, the original article seems to take for granted that *eventually* the U.S. will send weapons, and emphasizes how this action must be taken quickly rather than arguing on whether or not weapons should be sent in general. Finally, from a stylistic point of view, there are many repetitions, both in concepts and lexicon (the phrase "misguided and dangerous" is repeated twice).

4.5 Analysis of the students’ exercises: dealing with relevance issues

Other students had a an additional interpretation of the need of having “counterarguments” in the text, referring not only to the rebuttal of the target article, but to **the need of presenting a “balanced view” listing pros and cons** (“potential counterarguments”).

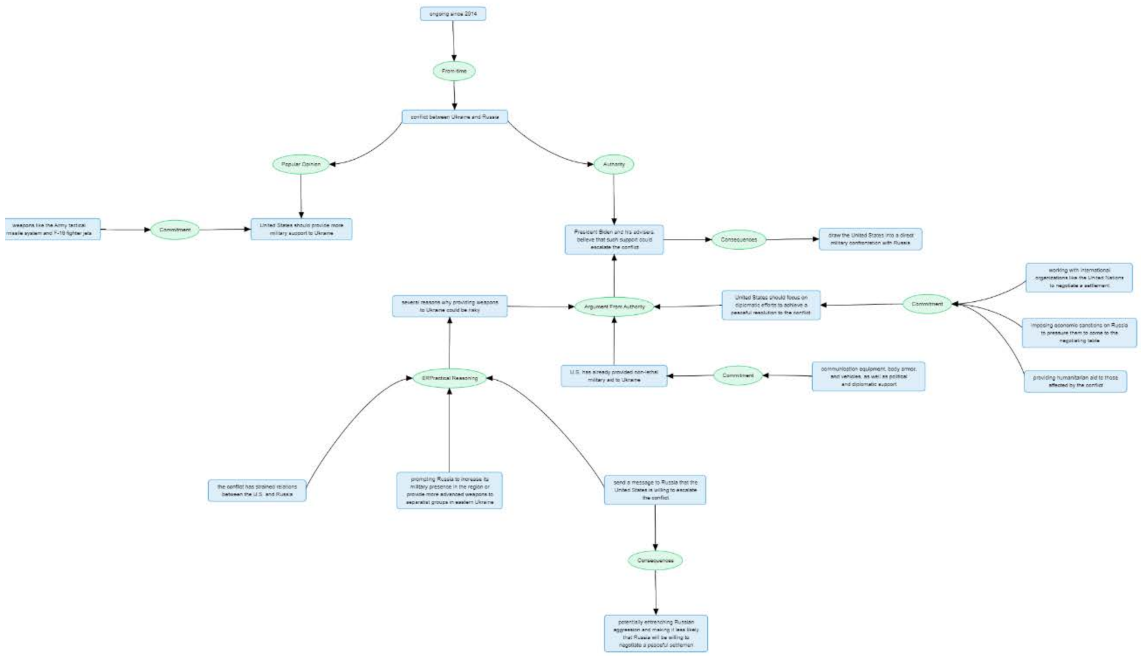
- Strengths:
- *Well-reasoned argument*.
- *Respond of the main arguments* presented in the article.
- The author acknowledges that individual health concerns should be considered, but also *emphasizes the benefits of a well-planned vegan diet*.
- Shortcomings:
- *Lack of potential counterarguments*. For example, the AI may have argument that some people may not have access to plant-based foods or that a vegan diet may be more expensive than a non-vegan one.

4.6 Analysis of the students’ exercises: under editing of the text and “forgetting” the target.

However, in other students, a lack of attention to the argumentative *relevance* of the generated text emerged mainly from the students under-editing. ChatGPT’s texts typically engaged with the target’s arguments in a very limited fashion, but some students did not edit the texts to address these flaws, focusing instead on stylistic flaws (mainly repetition).

The argument maps produced confirmed the limited attention to the target’s argument. **Very few counterargument nodes were used in the maps**.

In 3 cases the student failed to notice that the generated text was not taking the opposite standpoint but rather rewriting the original article. Two of these, however, correlate with a general poor performance in the final exam and with non attendance.


ongoing since 2014
From-time
conflict between Ukraine and Russia
Popular Opinion
Authority
weapons like the Army tactical missile system and F-16 fighter jets
Commitment
United States should provide more military support to Ukraine
President Biden and his advisers believe that such support could escalate the conflict
Consequences
draw the United States into a direct military confrontation with Russia
working with international organizations like the United Nations to negotiate a settlement
several reasons why providing weapons to Ukraine could be risky
Argument From Authority
United States should focus on diplomatic efforts to achieve a peaceful resolution to the conflict
Commitment
imposing economic sanctions on Russia to pressure them to come to the negotiating table
U.S. has already provided non-lethal military aid to Ukraine
Commitment
communication equipment, body armor, and vehicles, as well as political and diplomatic support
providing humanitarian aid to those affected by the conflict
ERPractical Reasoning
the conflict has strained relations between the U.S. and Russia
prompting Russia to increase its military presence in the region or provide more advanced weapons to separatist groups in eastern Ukraine
send a message to Russia that the United States is willing to escalate the conflict
Consequences
potentially entrenching Russian aggression and making it less likely that Russia will be willing to negotiate a peaceful settlemen

A student’s map featuring no counterargument nodes.

5.1 Taking stock of the experience

- The exercise had some basic design flaws and could be interpreted in different ways.
- It added complexity to an already complex learning process. In particular, many student did not have yet a firm grasp of the principles of argumentation mapping.
- Despite flaws and complexity, the exercise produced interesting outcomes in terms of reflection on the text.
- Acceptability issues seem easier to tackle than relevance issues.

5.2 Moving forward

- We wish to repeat the experience next year, with an improved format. Taking into account, as much as possible, the evolution of the technology and the students' likely increased familiarity with it.
- More explicit discussion of relevance issues with LLM generated texts seems advisable.
- In order to be meaningful, this discussion should rest on the evidence of systematic evaluations of ChatGPT's (and other LLM's) performance with argumentation tasks, as it happens for other NLP/ NLU tasks (Kocoń et al. 2023, *ChatGPT: Jack of all trades, master of none*, Lipkin et al. 2023).
- In this regard, IALS is collaborating with ARG-tech (U. Dundee UK) who are carrying out a systematic evaluation.

6. References

Agrawal, A., Mackey, L., & Kalai, A. T. (2023). *Do Language Models Know When They're Hallucinating References?*. (arXiv: 2305.18248). arXiv. https://doi.org/10.48550/arXiv.2305.18248

Hinton, M., & Wagemans, J. H. M. (2023). How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument & Computation*, *14*(1), 59–74. https://doi.org/10.3233/AAC-210026

Janier, M., Lawrence, J., & Reed, C. (2014). OVA+: An Argument Analysis Interface. In *Computational Models of Argument* (pp. 463–464). IOS Press. https://doi.org/10.3233/978-1-61499-436-7-463

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). *ChatGPT: Jack of all trades, master of none* (arXiv:2302.10724). arXiv. https://doi.org/10.48550/arXiv.2302.10724

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). *A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets* (arXiv:2305.18486). arXiv. https://doi.org/10.48550/arXiv.2305.18486

Lipkin, B., Wong, L., Grand, G., & Tenenbaum, J. B. (2023). *Evaluating statistical language models as pragmatic reasoners* (arXiv:2305.01020). arXiv. https://doi.org/10.48550/arXiv.2305.01020

Marcus, G. (2023). Why Are We Letting the AI Crisis Just Happen? *The Atlantic*, *March 2023*. https://www.theatlantic.com/technology/archive/2023/03/ai-chatbots-large-language-model-misinformation/673376/

OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Rocci, A. (2017). *Modality in Argumentation* (Vol. 29). Springer Netherlands. https://doi.org/10.1007/978-94-024-1063-1

Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., Choshen, L., Cohen-Karlik, E., Dankin, L., Edelstein, L., Ein-Dor, L., Friedman-Melamed, R., Gavron, A., Gera, A., Gleize, M., Gretz, S., Gutfreund, D., Halfon, A., Hershcovich, D., … Aharonov, R. (2021). An autonomous debating system. *Nature*, *591*(7850), 379–384. https://doi.org/10.1038/s41586-021-03215-w

Visser, J., Lawrence, J., & Reed, C. (2020). Reason-checking fake news. *Communications of the ACM*, *63*(11), 38–40. https://doi.org/10.1145/3397189

Walton, D. (2013) *Methods of argumentation*. Cambridge University Press, Cambridge/New York