Exploring the Usefulness of Open and Proprietary LLMs in Argumentative Writing Support

Reto Gubelmann[1(✉)], Michael Burkhard[2], Rositsa V. Ivanova[1], Christina Niklaus[1], Bernhard Bermeitinger[1], and Siegfried Handschuh[1]

[1] ICS-HSG, University of St Gallen, St Gallen, Switzerland
{reto.gubelmann,rositsa.ivanova,christina.niklaus,bernhard.bermeitinger, siegfried.handschuh}@unisg.ch
[2] IBB-HSG, University of St Gallen, St Gallen, Switzerland
michael.burkhard@unisg.ch

Abstract. In this article, we present the results of an exploratory study conducted with our self-developed tool **Artist**. The goal of the tool is to give formative feedback to develop students' argumentation skills. We compare the feedback that two different LLMs, an open-sourced one by META and one of OpenAI's fully proprietary ones, give to students' argumentative writing. We find that, overall, students find the feedback provided by both LLMs helpful (7.51 vs. 7.65 on a scale from 1 to 10), and they rate the quality of the feedback as good to very good. We take this as a very encouraging provisional result that invites larger and more extensive studies on the topic.

Keywords: Writing Support · Argument Quality · ChatGPT · Large Language Model · Computer-Supported Learning · Feedback Provision

1 Introduction and Relevant Previous Research

Argumentative writing support aims to improve students' argumentative skills. By learning how to develop strong arguments, students also practice and improve critical thinking skills, counting among the so-called 21st-century skills [16].

Developing these critical thinking skills requires a close supervision by an experienced teacher: Receiving regular, high-quality feedback can help students to improve their argumentative writing skills. However, this is resource-intensive and, for some settings, including massive open online courses, simply impossible. Therefore, it is desirable to support teachers in the task of providing such feedback to students. This is one of the core tasks of argumentative writing support, a subfield of writing support, which in turn belongs to Natural Language Processing (NLP).

For an overview of recent developments in this field, see [6]. For a recent study of scaling attempts of writing feedback, see [12]. For a recent study on

A. M. Olney et al. (Eds.): AIED 2024 Workshops, CCIS 2151, pp. 175–182, 2024.
https://doi.org/10.1007/978-3-031-64312-5_21

the effects of computer-generated feedback on overall writing quality, see [14]. Within this field, our approach focuses on a text's argumentative structure, and it builds on argumentative analysis approaches that decompose and/or classify text units and their most important components, usually claims and premises, and then assess the quality of the essay based on this structure. For a recent approach using this paradigm, see [2]. For the most important recent survey of the field, see [18]. For the field's connection to natural language inference (NLI), see [8].

With the advent of highly successful generative large language models (generative LLMs) such as *gpt-3.5-turbo* and *gpt-4*, which power *ChatGPT*, a very promising new path to this goal of aiding students in developing their argumentative writing abilities has entered the field. LLMs are a kind of neural network methods in NLP, as opposed to rule-based or Good Old-Fashioned (GOFAI) approaches. Inspired by the transformer architecture [17], researchers developed a number of influential natural language understanding (NLU) architectures as well as training routines, pioneered by the *BERT* architecture [5]. The transformer architecture has also inspired *GPT-3* [3], which in turn grounds the models powering *ChatGPT*.

Unfortunately, OpenAI, the company that has released *ChatGPT* in November 2022, has decided to contradict current practice in the field and refused to publish any important details on its models and its training, let alone the model weights themselves, while it has communicated that its models are based on the GPT-architecture [1].

Furthermore, with regard to OpenAI's proprietary models, concerns over data protection have alerted watchdogs in many countries, including Poland, the Netherlands, Canada, and Italy.[1] Recently, these privacy worries have been complemented by lawsuits[2] around potential intellectual property violations. Finally, the sheer size of the LLMs served by OpenAI implies that every request sent to *ChatGPT* also exerts a considerable carbon footprint: all other things being equal, a larger model requires more resources to process a request[3].

This creates an uncomfortable situation for universities: it is their mission to equip young people with cutting-edge knowledge and competencies, but they are also required to comply with national laws, including data protection laws, to foster open science, and to reduce their carbon footprint. Fortunately, with the recent surge in smaller open generative LLMs such as the ones released by Meta AI (see below, Sect. 3), a new option has entered the field. On paper, these models are extremely promising. They perform comparably to OpenAI's models at standardized benchmarks, but unlike those, they are relatively small, openly available, and pose no privacy issues, as they can be run on local hardware.

[1] See Reports by Reuters on Poland, the Netherlands and Canada, by the Financial Times on Italy.

[2] See Report by AP News. All links last consulted on January, 19th 2024.

[3] Unfortunately, it is impossible to even venture an educated guess on the specific extent of the carbon footprint of one request sent to *ChatGPT*, see [10].

In this paper, we provide a first exploration of the promise of open, locally-run LLMs in the task of argumentative feedback support provision. We provided first-year university students who are enrolled in an academic writing class with the opportunity to obtain feedback on argumentative texts from two different LLMs. We wanted to assert whether (1) students considered this feedback helpful at all, and (2) whether the small, open, and locally deployed LLM is perceived as equally helpful as OpenAI's *ChatGPT*. Answering both of our research questions affirmatively constitutes an important first insight on the performance of open, smaller, and locally run LLMs compared to OpenAI's models in the wild. Rather than evaluating on generic benchmarks, we tap directly into a real-world classroom scenario and let the students compare the two models side-by-side.

2 Our Argumentative Feedback Tool: Artist

For our experiment, we rely on our tool **Artist**[4] (see [4]). The main purpose of **Artist** is to provide students with insight into their argumentative texts. Based on results gathered from a prior user survey, we split the provided analysis into three categories for a better overview. When a user opens **Artist**, they are prompted to type their argumentative text in a designated field or to select one of three example texts used for demo and survey purposes. In the next step, they may choose from three different analytic dimensions:

(1) *Argument Structure Analysis*: uses a random forest classifier to identify argumentative components and visualize their structure in the form of a graph (this part of the tool builds on [19]);
(2) *Discourse Structure Analysis*: uses an RST parser [7] to analyze the rhetorical structure of the text;
(3) *Improvement Suggestions*: lets the user send their texts to an LLM and receive suggestions on how the argumentative quality could be improved. We have incorporated two options:
(i) llama-2, a self-hosted instance of *meta-llama/llama-2-70b-chat-hf*,
(ii) gpt-3.5 that sends the request to *gpt-3.5-turbo* via OpenAI's API.
The responses from the models are presented to the user in a textual form.

The *Improvement Suggestions* part of the tool is the main focus of this paper. The experiment takes largely place in this part of the tool.

In sum, our approach seamlessly embeds both a commercial LLM accessed with an API and an open LLM running on our infrastructure into a tool that aims at aiding students in their argumentation by providing formative feedback.

3 Set-Up of Exploratory Experiment

Technical Aspects. For the experiment we are relying on the self-developed tool **Artist** that is available via a web interface[5]. The focus of our experiment was

[4] Code and screenshots available at https://gitlab.com/ds-unisg/aied2024.
[5] https://artist.datascience-nlp.ai.

a new addition to our tool, namely the ability to receive text-specific feedback regarding the argumentative quality from two different LLMs. In order to reduce the bias of the participants for the study, we simply call them "Model 1" and "Model 2" respectively.

From among OpenAI's proprietary models, we evaluated the version of so-called *gpt-3.5-turbo* (in what follows *gpt-3.5*) available via the API during our experiments in October 2023. In stark contrast to OpenAI, the AI research group of META (formerly Facebook AI) has decided to publicly release its latest series of LLMs [15], allowing for reproducible and rigorous scientific experimentation with these models. We use their model called *meta-llama/llama-2-70b-chat-hf* (in what follows abbreviated by *llama-2*), which we serve using the very efficient serving method vLLM [9]. To the best of our knowledge, we are reporting on the first experiment to deploy this framework in a real-life educational setting. We give an overview on the differences between the two models tested in Table 1.

Table 1. Comparison of the two models used in our explorative experiment.

Model	Size	Serving Method	Open?	Privacy Concerns?
llama-2	70B	locally on 8 GPUs of a NVIDIA DGX-2 via vLLM	yes	none (runs locally)
gpt-3.5	175B	remotely via OpenAI API	no	multiple (see Sect. 1)

We use two different prompts to interact with the models, as they react differently to the same prompts. Following is the prompt for OpenAI's *gpt-3.5*:

> "Please give two short suggestions for improving the argumentative quality of the following Essay:" + *input text*

Following is the prompt for *llama-2*:

> "[INST] You are an argumentation expert and an experienced teacher that loves to give helpful and encouraging advice to students. You always respond in short, concise, well-formed sentences, and you are also creative. You receive the student's text from me, which has already been analyzed with Discourse Structure Analysis. Referring to very specific elements of the text, you give the student two specific tips on what they could improve about the text in terms of argumentation. Please try to be as specific and supportive as possible giving two formative and instructive feedbacks and nothing more in 2-3 sentences. Here is the text: " + *input text* + "[/INST]"

As can be seen, the prompt used for *llama-2* is much more sophisticated than the one for *gpt-3.5*. We found that more detail is necessary to obtain good results from *llama-2*. However, we applied the same purely formal routine to determine the prompt. We started with identical prompts and then continued to develop

the prompts until the following formal requirements had been met by returned feedback: (1) the language of the response is English, (2) the entire response fits into the space allocated to the answer window of 240 tokens, (3) the model gives exactly two suggestions (where we did not investigate the quality of the suggestions, just the count of two).

While *gpt-3.5* fulfilled these three formal criteria with the original version of the prompt, *llama-2* required more information to do so. We hypothesize that this is due to the more extensive reinforcement-based fine-tuning that went into OpenAI's product.

Participants. A total of 63 students participated in our study. All of them were first-year university students who had almost completed the course for academic writing. The course includes several cycles of peer review, thus the students were experienced in both giving and receiving formative feedback. All of the students were enrolled in the same general study program that then allows them to study for a variety of degrees in business administration and social sciences. 66% of the participants were male, 32% female, and 2% preferred not to state their gender. The first language of 89% was German and for 3% English. In regards to the age, we observed a higher variability. The youngest participant(s)'s age was 17 and the oldest(s)'s 37 with an overall (rounded-up) mean of 20.

Table 2. Results of LLM-specific questions (name of LLM that has a higher percentage of strongly or somewhat disagree (question 1) or agree (questions 2 and 3) printed in **boldface**, all values in [%]).

Question	Model	Strongly disagree	Somewhat disagree	Neither nor	Somewhat agree	Strongly agree
Loading answer took too long?	**llama-2**	41.27	31.75	14.29	9.52	3.17
	gpt-3.5	46.03	25.40	17.46	9.52	1.59
Understood the recommendation?	**llama-2**	1.61	3.23	9.68	37.10	48.39
	gpt-3.5	0.00	4.84	9.68	50.00	35.48
Was the feedback useful?	llama-2	1.59	4.76	12.70	46.03	34.92
	gpt-3.5	0.00	3.17	11.11	44.44	41.27

Details on the Process. The experiment was conducted in class on October 20th, 2023, in a course that introduces students to the basics of scientific writing, such as correct referencing, composition, topic selection, and argumentation. The entire class lasted for 90 min and was dedicated to the topic of argumentation. The experiment itself was conducted within this class and took approximately 15 min. The participants were asked to complete a questionnaire about the utility

of the entire tool with a clear focus on the recommendations of the two LLMs. Participants in the experiment had one web browser window open with the questionnaire and another one with the tool. To have more control over the variables of the experiment, we randomly chose three texts from a well-known dataset [13] that the students used to obtain the LLMs' feedback.

4 Results and Discussion

Results. In Table 2, we give the results of the LLM-focused questions of our experiment. The table shows that *llama-2* receives slightly better scores with the first two questions (note that the difference regarding the second question is only 0.01%), while *gpt-3.5* significantly outperforms its competitor in question 3. Figure 1 depicts the distribution of participants' responses to the question: *How do you perceive the received feedback quality of Model 1 and Model 2? Rate the two Models on a scale of 1 to 10, where 10 is the highest value and 1 is the lowest value.* The figures show that (1) both models are perceived as helpful, and (2) the specific scores of the two models are very close (average for *llama-2* 7.51, for *gpt-3.5* 7.65). This means that the quality of the feedback provided by the two models was perceived as good to very good, and almost on a par.

Discussion. We emphasize three aspects of the results of our study. First, the perceived quality of the feedback is remarkable. It was not to be expected that general-purpose LLMs with no specific fine-tuning for giving feedback on argumentative texts would perform so well. The task is very difficult, as it requires a combination of strict, logical modes of linguistic abilities with associative, topical modes of abilities. The vast majority of participants — over 80% at the least — understood the recommendation and found it useful. Furthermore, on a scale of 1 to 10, students ranked the quality of the feedback above 7.5 on average. Bearing in mind that even feedback by experienced human teachers would not get a straight 10,[6] it is clear that this is a very good score.

Second, it is surprising that *llama-2* is competitive with *gpt-3.5* throughout the experiment, and that our serving method outperforms OpenAI's API in terms of response time. Given that OpenAI's model is 2.5 times the size of *llama-2*, and given that only the former has been extensively fine-tuned using human feedback and a special flavor of reinforcement learning [11], it would not have been rational to expect that the models are almost on a par at this complex task. From a research-political as well as from an environmental perspective, this is encouraging: open-source models can compete with highly resourced proprietary models even at very challenging tasks. And the fact that the model is 2.5 times smaller will, roughly, reduce its carbon footprint per processed request by the same factor. Where the vast resources invested in OpenAI's models might show is in the ease with which one can extract *formally satisfactory* feedback from

[6] For instance, Weaver [20], finds that only 18% of business and design students always find the (human) feedback that they receive during their studies clear and easy to read.

gpt-3.5. While getting *llama-2* to give feedback in the requested form required a carefully engineered, rather long prompt, the prompt that was used to interact with *gpt-3.5* was short and ready in a matter of seconds.

Third, we wish to point out three limitations of our study. As the demographics in Sect. 3 show, the majority of our participants' mother tongue is not English, but rather German. While they all have to be able to take classes in English, native speakers might perceive the usefulness of the feedback provided differently. Note, however, that students across the world have to learn to write academic texts in English as a second language. In these settings, our results are directly applicable. The second limitation is inherent in the design of our study: to make the results as comparable as possible, we pre-defined the texts that the students used to interact with the tool. It is possible that using different texts would lead to different results. However, by choosing three texts at random from a well-respected argumentative writing dataset, we tried to keep this probability as low as possible. Lastly, we wish to emphasize once more that the size and scale of this study means that it can only offer preliminary results that have to be confirmed in larger, more comprehensive settings.

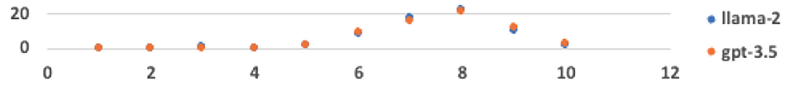

Fig. 1. *llama-2* vs. *gpt-3.5*: perceived quality of argumentative feedback.

5 Conclusion

In this article, we have tested the capacities of an AI-based tool that is intended to support and improve students' argumentative writing skills. We have focused on using two different LLMs to provide students with case-specific formative feedback to improve their argumentative texts. Our provisional findings are overall very encouraging. Students perceive the quality, comprehensibility, and helpfulness of the feedback by the two LLMs as good or very good. We are particularly encouraged by the fact that students rate the much smaller open-source LLM *llama-2* almost as highly as OpenAI's *gpt-3.5* (7.51 vs. 7.65 on a scale from 1 to 10). As a consequence, we plan to experiment with more flexible settings and larger test groups to confirm our findings. Furthermore, as the field of LLMs is evolving at an impressive pace, we would like to explore the promise of even smaller and more efficient models to further reduce the carbon footprint of our method.

Acknowledgements. We would like to thank the Swiss National Science Foundation (SNSF) for the grant to support our project "Next Generation of Digital Support for Fostering Students' Academic Writing Skills: A Learning Support System based on Machine Learning (ML)", a collaboration project between the University of St. Gallen in Switzerland and the Mahidol University in Thailand.

Disclosure of Interests. The authors declare that they have no competing interests that could influence the content of the research presented here.

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv: 2303.08774 (2023)
2. Alhindi, T., Ghosh, D.: 'Sharks are not the threat humans are': argument component segmentation in school student essays. In: BEA@EACL, pp. 210–222 (2021)
3. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
4. Burkhard, M., et al.: Computer supported argumentation learning: design of a learning scenario in academic writing by means of a conjecture map. In: CSEDU (1), pp. 103–114 (2023)
5. Devlin, J., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL 2019, pp. 4171–4186 (2019)
6. Strobl, C., et al.: Digital support for academic writing: a review of technologies and pedagogies. Comput. Educ. **131**, 33–48 (2019)
7. Feng, V.W., Hirst, G.: A linear-time bottom-up discourse parser with constraints and post-editing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 511–521 (2014)
8. Gubelmann, R., et al.: Capturing the varieties of natural language inference: a systematic survey of existing datasets and two novel benchmarks. J. Log. Lang. Inf. **33**, 21–48 (2023). https://doi.org/10.1007/s10849-023-09410-4
9. Kwon, W., et al.: Efficient memory management for large language model serving with pagedattention. In: SOSP 2023: Proceedings of the 29th Symposium on Operating Systems Principles, pp. 611–626 (2023)
10. Patterson, D., et al.: Carbon emissions and large neural network training. arXiv: 2104.10350 (2021)
11. Proximal policy optimization. OpenAI (2017). https://openai.com/blog/openai-baselines-ppo/
12. Rapp, C., Kauf, P.: Scaling academic writing instruction: evaluation of a scaffolding tool (thesis writer). Int. J. Artif. Intell. Educ. **28**, 590–615 (2018)
13. Stab, C., Gurevych, I.: Annotating argument components and relations in persuasive essays. In: COLING 2014, pp. 1501–1510 (2014)
14. Stevenson, M., Phakiti, A.: The effects of computer-generated feedback on the quality of writing. Assessing Writ. **19**, 51–65 (2014)
15. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models. arXiv: 2307.09288 (2023)
16. Van Laar, E., et al.: The relation between 21st-century skills and digital skills: a systematic literature review. Comput. Hum. Behav. **72**, 577–588 (2017)
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Wachsmuth, H., et al.: Computational argumentation quality assessment in natural language. In: EACL 2017, pp. 176–187 (2017)
19. Wambsganss, T., et al.: AL: an adaptive learning support system for argumentation skills. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2020)
20. Weaver, M.R.: Do students value feedback? Student perceptions of tutors' written responses. Assess. Eval. High. Educ. **31**(3), 379–394 (2006)