Developing Written Communication Skills in Engineering Education Using Automated Short Answer Grading

Jan-Mikael Rybicki
Aalto University
Espoo, Finland
jan-mikael.rybicki@aalto.fi

Sami Sarsa
University of Jyväskylä
Jyväskylä, Finland
sami.j.sarsa@jyu.fi

Juho Leinonen
Aalto University
Espoo, Finland
juho.2.leinonen@aalto.fi

Arto Hellas
Aalto University
Espoo, Finland
arto.hellas@aalto.fi

***Abstract*—This full research paper examines Automatic Short Answer Grading (ASAG) as a method to practice technical writing skills and how this is impacted by the introduction of Large Language Models (LLMs) to the public. Both the industry and ACM Computing Curricula 2020 argue that graduates should be able to communicate effectively in writing across organizational contexts, such as technical documentation and client communication. However, many undergraduate students are unaware of effective technical writing conventions. Further, various LLM-based tools allow students to generate or revise their texts with little effort, but this does not directly translate to high-quality texts or learning writing skills. Thus, the need to learn what makes texts readable, concise, and clear has remained as an important learning goal. These challenges can be addressed by providing targeted writing practice using ASAG. Using short responses allows reducing the cognitive load of learners, directing their focus to specific features of writing. We use Educational Data Mining (EDM) approaches to analyze undergraduate engineering student responses from online ASAG exercises designed for practicing the revision of ineffective writing. We investigate 1) how the prevalence of LLMs may have influenced student approaches to solving ASAG exercises, and 2) how to support the development of writing skills using ASAG. We find that LLMs sometimes struggle at forming concise and clear sentences in certain syntactic structures. LLMs perform well in suggesting alternatives to replace informal word choices but can easily resort rewriting sentences and changing the intended meaning.**

***Index Terms*—multidisciplinary, undergraduate education, written communication, educational technology, automated grading, quantitative analysis**

I. Introduction

Ability to communicate effectively is an important skill for STEM graduates both in professional and academic contexts, as stated by many ABET 2025-2026 accreditation criteria. For instance, in Engineering Technology Programs, student outcomes must include “an ability to apply written, oral, and graphical communication in well-defined technical and non-technical environments; and an ability to identify and use appropriate technical literature” [1]. Similarly, ACM Computing Curricula 2020 [16, p. 116] state that graduates should have the ability to “communicate effectively in writing within a broad range of organizational contexts”. However, research has shown that, for example, Computer Science (CS) students have challenges in producing succinct, semantically and syntactically correct content for documentation and other forms of written communication [2, 9, 23]. These challenges are emphasized internationally also in other engineering fields among students who use English as a Foreign Language (EFL) or Second Language (L2) [20, 24].

Although good writing is valued, instructors may not have resources for writing instruction in large courses [13]. A variety of tools and technologies have been developed for teaching and practicing academic writing [28] that could alleviate the instructor workload. However, these tools may pose various limitations when attempting to incorporate them into a course curriculum. Many tools are typically standalone software or websites, which cannot be easily embedded into institutional Learning Management Systems (LMSs) or integrated within learning materials to practice worked examples on teacher-specified writing issues. Many commercial tools are available for automatic text correction and feedback, such as Grammarly, but they typically offer limited possibilities for modifying their core functionality for teacher’s pedagogical purposes, and may not fulfill the universities’ data handling requirements, for example, the General Data Protection Regulation (GDPR) in the European Union.

For small-scale, sentence or paragraph-level practice, Automated Short Answer Grading (ASAG) activities could provide convenient tools and methodologies to practice worked examples. However, much of the ASAG research has focused on developing Machine Learning (ML) powered automated graders that emphasize content and summative assessment but ignore style, form, or grammatical correctness [4, 25]. Another hindrance with the ML models and Large Language

Models (LLMs) is their general requirement for extensive training data, which may not be available for instructors. Thus, instructors using these technologies need to most often rely on commercial third-party models where it can remain unclear what in the training data led to a particular output. In this respect, ML and LLM approaches convey a risk of depriving teacher control and producing unpredictable results: why did the model make this choice?

To address some of the above issues, simple rule-based exercises with basic Natural Language Processing (NLP) support can enable instructors to create sufficiently flexible yet controlled ASAG activities, which guide student responses towards recommended model answers specified by the instructor. Such ASAG tools allow teacher to have more control over what happens when learners solve worked examples, such as when to provide formative feedback, how much, and in what format.

However, to the best of our knowledge, there exist little research specifically examining LLMs and ASAG exercises in the context of academic writing. Thus, it is interesting to study whether LLMs can solve these types of exercises and whether the popularization of LLMs can be seen in student attempts at solving ASAG exercises. Therefore, this paper addresses the following research questions.

RQ1 How well can modern LLMs solve rule-based ASAG exercises for academic and technical writing?

RQ2 How does the emergence of generative AI tools show in student attempts to solve the rule-based ASAG exercises?

To address these questions, we analyzed quantitatively and qualitatively bachelor and master's level student responses in ASAG sentence transformation exercises in engineering writing contexts both before (2022) and after (2023-2024) the launch of ChatGPT. Using edit distances and vocabularies as metrics, the student responses were then compared with currently popular and freely available LLMs: GPT4o, Gemini-2.0-flash, Llama-3.3-70b-versatile, and Mistral-saba-24b. Finally, we discuss the general differences and potential usefulness of these LLMs for pedagogical purposes, particularly from the perspective of ASAG exercises.

II. Related work

A. Needs for communication skills in writing

It has been long proposed that incorporating writing skills into engineering education would be beneficial as a tool for writing to learn, to enhance cooperative learning and communication, and to improve the status of engineering professions [31]. Currently, writing is incorporated in ABET criteria [1] and ACM Computing Curricula 2020 [16].

However, not all institutions meet the criteria globally. A recent survey on faculty and employer views concerning engineering competencies indicated that fresh graduates were perceived to be well-prepared in mathematics and science and technology, but insufficiently prepared for writing, communication, and computer skills [11]. One challenge in including writing in instruction is due to engineering instructors not being aware or able to utilize online writing tools in teaching [17].

Another challenge can be student perceptions. For example, first-year CS Students may have misconceptions about the importance of writing, some of whom assume that high-school level writing skills are sufficient, learning to write takes time away from developing computing skills, or that writing can be avoided in the workplace [10]. Therefore, faculties need to address these misconceptions concerning the role of writing in engineering careers. When the writing challenges are not addressed at the undergraduate or graduate level, L2 engineering students continue to struggle with academic writing also at the postgraduate level [20].

B. Common problems in writing

Various aspects in written communication skills can be challenging for engineering students and graduates according to prior research. Dugan and Polanski [9] presented a taxonomy of writing tasks that CS graduates typically encounter in working life. The taxonomy indicates that written communication is required in a wide variety of situations, and the length of documents can range from short bug reports to long project plans. They advise, among other things, the following to be taught to students [9]: understanding basic rhetorical contexts for different purposes and audiences; writing different types of definitions; using parallel text structures at different levels, such as words, phrases and clauses; and distinguishing between active and passive voice, and knowing the rationale for using these forms.

Recently, Munir et al. [23] mapped common writing issues among upper-year CS students based a grading rubric for assessing pertinent areas in writing quality. Their results show that students commonly had writing problems in the following areas: grammar, conciseness, clarity, organization, structure, and formality as the main categories. Moreover, in the case of English language learners, writing can be even more challenging [13, 18], requiring more support than first language (L1) and advanced L2 writers.

C. Some successful approaches to practicing writing in engineering education

Naturally, learners can be independently asked to consult style manuals that discuss common writing problems and provide guidelines for revision [3, 33]. However, mere independent reading of guidelines is not sufficient for developing expertise in a complex skill. Despite noticeable problems in written communication, Anderson et al. [2] stated that most CS instructors are not willing or able to reserve class time to teach writing during their courses. Nevertheless, they argue that it is also possible to teach good writing practices implicitly. In their experiment, Anderson et al. [2] observed that the quality of students' short black box test plans improved when an experimental group was provided with model texts and clear instructions that specified the target audience. This allowed students to recognize the information needs of the reader and to write clearer user instructions. Similar findings were found in civil engineering contexts, in which writing practice was successfully incorporated as paragraph-length tasks throughout a large hydraulics course [13]. This reduced the burden of instructors to provide feedback and gave students frequent but short writing practice. Similarly, incrementally writing a full-length technical memo can be supported with custom rubrics, worksheets, and time-management activities [12]. With the development of NLP technologies, writing practice and feedback can be supported with automated exercises [5].

D. Influence of ChatGPT on writing

The introduction of LLMs and OpenAI's ChatGPT in 2022 has greatly affected writing processes in education and the industry, evidenced e.g. through essay submissions to open online courses [19]. Although LLMs can be used for generating entire texts, they can be used pedagogically to support learning of writing. While ChatGPT can provide immediate feedback and improve scores in writing tasks, Fan et al. [14] found this may lead to "metacognitive laziness" and possible long-term skill stagnation, as prompting an LLM allows achieving task goals with less effort compared to traditional tasks. Additionally, their experiment indicated the use of ChatGPT did not improve knowledge gain, knowledge transfer, or intrinsic motivation compared to other forms of feedback.

Zhan and Yan [32] investigated student use of ChatGPT feedback in terms of cognitive, metacognitive, affective, and behavioral engagement. Cognitively, the amount of ChatGPT feedback on grammar and academic style was overwhelming for some students, and they had to filter information to find relevant elements. Some students were concerned with the trustworthiness of information and compared it with other sources. Metacognitively, few students monitored their revision process. Affectively, most students found interaction with ChatGPT relaxed (e.g., no need to worry about wrong questions). However, some felt impatient, confused and upset if ChatGPT misunderstood their prompts or gave wrong information. Behaviorally, students implemented the feedback in 56.3% of the cases, adapted it (27.6%), or reject it (16.1%). Zhan and Yan [32] noted that students would benefit from training in feedback literacy to evaluate the feedback more critically and to develop metacognitive engagement, such as goal setting and planning.

Feedback literacy is indeed important since LLM generated language has been found to differ from human language syntactically and lexically, e.g., form of increased nominalization, which usually increases lexical density and abstractness, or word choices uncommon in human use [26].

E. Automated short answer grading

Although a wide range of Automated Writing Evaluation (AWE) tools have been developed for supporting academic writing [28], many learning contexts do not require essay-length writing, and short sentence-level activities, such as short answer exercises, can offer versatile learning opportunities.

Burrows et al. [4] define Automated Short Answer Grading (ASAG) as "the task of assessing short natural language responses to objective questions using computational methods". The grading can focus either on the contents or quality writing or both. Although the length of ASAG text responses are much shorter than those of essays, the pipelines and processes in automated grading are similar for ASAG and Automated Essay Grading (AES), typically involving Natural Language Processing (NLP) and developing grading models based on prior data [4].

Grading short answer and essay responses is challenging for both humans and computers. While humans may suffer from inconsistency between graders and may find grading a laborious and cognitively challenging task, a major limitation with computers is their inability to understand the contents. Doewes and Pechenizkiy [8] addressed the challenges in AES metrics and their limitations compared to human graders. The validity and reliability of AES systems can be undermined under special circumstances that would not fool a human grader, such as off-topic essays, gibberish, and paraphrased answers (i.e., a different wording might yield different score even if equally correct).

Despite the limitations, automated systems can provide added value for learners by providing feedback or other scaffolding and reduce the workload of human graders. Wan et al. [30] proposed a method to identify claims in argumentative essays using a range of natural language processing tools and lists of known key phrases. These methods can be included in online tools to support students in writing argumentative essays by providing automated feedback.

The recent developments in model training approaches and deep-learning algorithms can also facilitate the development ASAG and AES technologies. Lu and Cutumisu [21] developed an approach for AES that automatically generates feedback sentences using paraphrasing based on a set of feedback statements. Condor et al. [6] evaluated the feasibility of generalizing ASAG models. They compared the classification performance of the state-of-the-art Sentence Bidirectional Encoder Representations from Transformers (BERT) vector representation model with more traditional methods: Word2Vec [22] and bag-of-words. The results indicated that the Sentence BERT algorithm outperformed the traditional methods under many test conditions.

Currently, ASAG research involves combining LLMs into the evaluation process as opposed to traditional ML models. Recent findings indicate that traditional ML models continue to outperform GPT-4o in grading accuracy even when applying few-shot prompting [15]. The main benefit of LLMs is that they are general purpose tools and do not require time-consuming training, but prompt engineering remains an open research area.

While ML- and LLM-based technologies can automate the grading and feedback processes flexibly to new contexts, we argue that rule-based ASAG approaches can be suitable for more straight-forward educational contexts, such as guided basic practice in academic writing contexts. Rule-based method guarantee precision and sufficient range of acceptable responses when practicing specific academic writing conventions. The next sections discuss the rule-based approaches analyzed this paper and how a set LLMs can solve them.

III. System Description

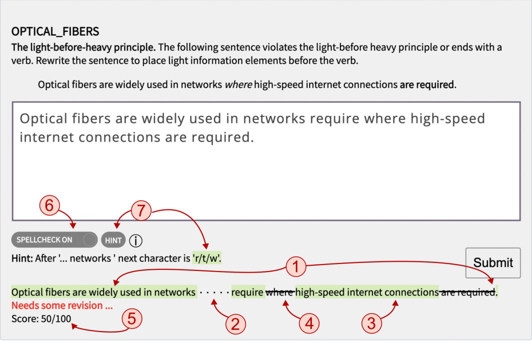
In this study, we have used a custom system developed for practicing written English in the form of ASAG exercises. The system consists of three parts: (1) a tool for creating exercises used to practice English, (2) a grading and feedback system, and (3) an interface used to provide exercises to students. When creating an exercise, a teacher defines an exercise as an ordered sequence consisting of literal texts (*Lit*) and arrays (*Arr*) of synonyms/similar expressions. Then, based on the sequence, the system generates the correct answers of the exercise. As an example, the sequence ({*Lit: Fig. 1*}, {*Arr: shows, illustrates, demonstrates*}, {*Lit: the user interface.*}) would give 1 x 3 x 1 possible solutions, e.g., *Fig. 1 illustrates the user interface*.

When the system gives feedback on a student submission, it selects the closest correct answer, determined by the longest common subsequence, and bases the feedback on it. This basic exercise type allows creating simple, rule-based and intuitive automated feedback that guides the learners in reaching a set of model answers.

Fig. 1 illustrates the automated feedback in the user interface [29]. When clicking the "Submit" button, the learner receives feedback on their solution. This also creates a log entry. The exercise system then compares the student answer with model solutions provided by the instructor. The automated feedback mechanism provides corrective feedback on three aspects. (1) Below the text box, the automated feedback prints and highlights in green the correct text segments at *character*-level both from the *beginning* and *end* of the text. (2) Incomplete and incorrect character segments requiring revision in the middle of the text are shown with dot • characters. (3) In case of excessively long responses, extra characters are marked with strike-out characters. The text "Needs some revision ..." is replaced with "Correct!" once the learner reaches one of the correct solutions. In addition, the instructor can include custom feedback on specific pattern matches in learner responses. The user interface also includes a *spell checker*; a *hint button* that indicates the next missing character on demand (up to three alternative word choices); *a longest common subsequence (LCS) algorithm* [7, p. 391] to highlight correct segments of the responses also in the middle of the sentence; and *a scoring system* (a percentage).

In the task illustrated in Fig. 1, the learners were expected to modify the sentence from passive to active voice. One correct solution is "Optical fibers are widely used in networks that require high-speed internet connections."

Instead of character-level matching, the User Interface (UI) highlights the correct parts of the sentence at *word* level. With the LCS feature, we initially noticed that too much text was highlighted, which would have made the exercises trivial to solve. To encourage thinking during the problem solving process and to allow discovering alternative word choices, a word level matching was selected. However, in some situations, the word level matching could make solving the task too difficult and frustrating for learners. Therefore, a hint button was added to the UI, which reveals the next missing character from the beginning of the sentence or at first error. When there are many possible options, initial characters are shown separated with slashes: a/b/c. The hint feature can be used once, after which the learners need to edit their response and click the submit button again. This approach reduces the temptation to game the system excessively yet provides enough clues to continue revising the text.



OPTICAL_FIBERS

The light-before-heavy principle. The following sentence violates the light-before heavy principle or ends with a verb. Rewrite the sentence to place light information elements before the verb.

Optical fibers are widely used in networks *where* high-speed internet connections **are required**.

Optical fibers are widely used in networks require where high-speed internet connections are required.

Hint: After '... networks ' next character is 'r/t/w'.

Optical fibers are widely used in networks · · · · · require ~~where~~ high-speed internet connections ~~are required~~.

Needs some revision ...

Score: 50/100

Fig. 1. The UI for ASAG exercises. Default feedback focuses on four aspects: (1) correct matches at *word*-level (in green highlight) from the *beginning* and *end* of text; (2) parts of text that need revision (shown with dots •); (3) longest common substring match at *word*-level in the middle of the text; (4) excess text, which is marked with strike through; and (5) score (i.e., percentage of correct text). On-demand feedback features include (6) a spell checker and (7) a hint button, which provides the next missing letter from the beginning of text.

The exercises were delivered as a `shortanswer` content type created for the Acos server [27], which is designed for integrating interactive learning contents into LMSs. During the writing courses related to this research, the A+ learning platform and Acos contents were embedded into a customized Moodle (https://moodle.org/) platform.

IV. Methods

A. Context

The data for this study was collected from undergraduate and graduate courses involving academic writing instruction in August–December 2022 (prior to ChatGPT) and April–October 2024 (after launch of ChatGPT). Students were asked for a written research consent to use their leaning data for research purposes. Within these time ranges, 281 permissions were obtained of which 278 students completed most ASAG course work. Most students were at the undergraduate level (n=268), while the exercises were also included in some graduate level CS courses (n=13). The undergraduate students represented a wide range of engineering and technical fields, including automation, electrical, civil, construction, mechanical and chemical engineering, computer science, as well as a few architecture students. The gender distribution was as follows: male (n=188), female (n=91), and other (n=1). Most participants had Finnish (n=225 or bilingual n=6) or Swedish (n=19 or bilingual n=2) as their native language.

The remaining students represented heterogeneous language backgrounds, including Tagalog (n=5), Dari (n=3), Russian (n=3) as well as various Indo-European, and West and South-East Asian languages.

B. Exercise responses

Before completing ASAG exercises, students viewed instructional videos, read example sentences, and completed simpler multiple-choice and fill-in-a-gap exercises related to stylistic features of academic writing. For the present study, we included two linguistically distinct topics areas that aim to improve readability and adherence to stylistic conventions in academic writing. *Light-before-heavy (LBH) principle* proposes that subjects in a sentence should be kept short (light) followed by a verb phrase, and the complex new information should be placed at the end of a sentence. Exercises in this principle typically (but not always) involve revising sentence structures by changing the sentence from passive to active voice or vice versa. *Style and word choice* exercises focus on identifying informal words and phrases commonly used by engineering students and revising them with more formal and ideally concise alternatives.

The exercises always begin with a short explanation of the problem area and instructions, and the sentence to be revised, as shown in Fig. 1. These instructions and problem sentences were also given to the LLMs. The selected exercise question texts (Q) and one model solution (A) are shown below. Note that most questions have a number of possible correct answers, particular in the style and word choice exercises.

C. Light-before-heavy principle

Instructions: The light-before-heavy principle. The following sentence violates the light-before heavy principle or ends with a verb. Rewrite the sentence to place light information elements before the verb.

Q: In this thesis, software quality assurance methods that can be used for detecting defects in electronic components **are introduced**.
A: This thesis **introduces** software quality assurance methods that can be used for detecting defects in electronic components.

Q: Optical fibers are widely used in networks *where* high-speed internet connections **are required**.
A: Optical fibers are widely used in networks **requiring** high-speed internet connections.

D. Style and word choice

Instructions: Style and word choice. The following sentence has some problems with style and word choice. Please revise the sentence to improve its formality.

Q: Nowadays, mobile payment is a hugely popular option due to the fact that most travelers and tourists carry a mobile phone with them.
A: Currently, mobile payment is a highly popular option since most travelers and tourists carry a mobile phone with them.

> The light-before-heavy principle. The following sentence violates the light-before heavy principle or ends with a verb. Rewrite the sentence to place light information elements before the verb. Make only minimal changes.
>
> Chemistry laboratories are required to follow strict safety measures so that potentially dangerous accidents can be avoided.

Fig. 2. An example LLM prompt. The additional text in the prompt variation is highlighted in yellow.

Q: The Global Positioning System (GPS) is a satellite navigation system that makes it possible for users to tell their location around the world.
A: The Global Positioning System (GPS) is a satellite navigation system that allows users to determine their location around the world.

E. Experiment/Approach

To answer our RQ1, *How well can modern LLMs solve rule-based ASAG exercises for academic and technical writing?*, we fed the above task instructions and the sentences to multiple popular and freely available LLMs. For this, we used two prompt variations, one that contained only the task instructions and the starter sentence (denoted by above 'Q'), and one that contained the following additional sentence directly after the instructions: "Make only minimal changes.". An example of the prompt is shown in Fig. 2.

We used the following models to generate responses: GPT4o, Gemini-2.0-flash, Llama-3.3-70b-versatile, and Mistral-saba-24b. GPT4o and Gemini are popular state-of-art proprietary models, and Llama and Mistral are open source. The responses for the open source models were generated using the Groq playground (https://console.groq.com/playground) by manually copy-pasting the task instructions shown in Fig. 2 into the chat input box. Each task was queried separately using a fresh prompt. The LLM responses were then stored as Markdown files for later inspection and analysis. Preliminary, we also tested the Qwen-25-32b model, but its responses were highly similar to those of GPT4o (possibly due to being trained with the help of OpenAI GPT models). Thus, further manual investigation of Qwen was dropped from this study, as it was unlikely to provide additional insights.

To answer RQ2, *How does the emergence of generative AI tools show in student attempts to solve the rule-based ASAG exercises?*, student submissions were compared between the last course iterations prior to ChatGPT release (during Aug 2022–Dec 2022) and recent course iterations after the release (during Apr 2024–Oct 2024). To automatically identify potential differences (that we presume are caused by the introduction of ChatGPT and the like to the public), we computed two metrics: *edit distance to the closest model solution* (*ed-to-closest*) and *number of LLM response specific words* (*n-llm-words*) for each task attempt. For the *ed-to-closest*

TABLE I
COMPARISON OF LLMS AND THEIR ABILITY TO SOLVE EXERCISES

LLM	Fully correct	Acceptable	Total
Gemini-2.0-flash	4/20	11/20	**15/20**
GPT4o	6/20	8/20	14/20
Llama-3.3-70b-versatile	**8/20**	5/20	13/20
Mistral-saba-24b	6/20	3/20	9/20

metric, we used Levenshtein distance, i.e., the smallest number of character insertions, deletions or substitutions that convert one piece of text to another. The closest (longest common subsequence) model solutions were already available in the task submission data since these were used to provide hints to students. The LLM response specific words in *n-llm-words* were defined to be the words for a specific task that are present in LLM responses – obtained in the LLM response inspection phase described above – but not present in task starter sentence or any of the available model solutions.

V. RESULTS

RQ1. By providing the basic task instructions, LLMs are able to provide good but sometimes too creative or even unsuitable solutions. This may be partly due to the lack of context available for the LLMs, while this context is available for learners, who may not have learned the strategies yet.

Table I shows how well the LLMs found 100% correct solutions in ASAG exercises and how many solutions were acceptable although missing from the teacher-provided model solutions. Llama-3.3-70b-versatile found most frequently the 100% correct solutions but did not produce many other acceptable responses. When considering both correct and acceptable solutions, Gemini-2.0-flash performed the best. Overall, GPT4o was very close to Gemini and found fully correct solutions more frequently. While Mistral-saba-24b was reasonably good at finding fully correct solutions, it rarely proposed acceptable alternatives.

RQ2. Table II shows the descriptive statistics for student submissions for tasks included in this present study. The topic areas and tasks are in the same order as they appear in the course materials. The post-ChatGPT group has higher submission count means in 15/20 tasks and higher standard deviation in 14/20 tasks.

Table II also shows the comparison of edit distances and number of LLM-words prior and after the release of ChatGPT. Overall, edit distances have reduced after the launch of ChatGPT in the sentence re-structuring exercises (LBH principle), while the occurrence of LLM vocabulary has increased. In case of style exercises, some edit distances have lowered, but LLM word frequencies have notably increased. It should be noted that the tasks *cars* and *mobile_payment* have higher *ed-to-closest* and *n-llm-words* values than other tasks since these tasks were also used demo exercises for students to experiment at the beginning of the course (*mobile_payment*) or at the beginning of the module on style and word choice (*cars*). Nevertheless, Table II suggests that students use LLMs during the learning process, which may also lead to higher submission counts and more spread in the range of attempts.

Table III indicates that post-ChatGPT students (whose sample was smaller than that of pre-ChatGPT group) potentially resort more to trial-and-error problem solving using LLMs. This is indicated by higher number of submissions (i.e., responses), broader vocabulary range (dictionary words), and more misspellings.

VI. DISCUSSION AND CONCLUSIONS

A. RQ1. How well can modern LLMs solve rule-based ASAG exercises for scientific writing?

As indicated by prior research, LLMs have different strengths and weakness from the perspective of learning [14, 32]. In our prompting experiments, GPT4o typically provided only one answer, whereas Gemini-2.0-flash often suggested three options for the user, Llama-3.3-70b-versatile typically 1-2 options with brief explanation, and Mistral-saba-24b mostly one but occasionally even five with explanations. It is unclear why and when the models choose to provide alternative options and explanations. The options are frequently accompanied by short explanations or rationale related to the options, such as conciseness or emphasis on certain information in a sentence. In this respect, Gemini-2.0-flash provides pedagogically meaningful responses by default.

Llama-3.3-70b-versatile provides generally good responses. The responses are typically combined with a list of explanations pointing out the main changes, which is pedagogically useful. In the style exercises, the revised sentences based on default settings may include more nominalization and verbosity, which can make the sentences more packed with abstract information [26]. Overall, Mistral-saba-24b performed the worst: it often suggested a correct solution, but some solutions were unidiomatic or even worse than the original sentence to be revised.

Although LLMs found the exercise model answers in 4-8 times out of 20 exercises (20%–40% accuracy), LLMs did suggest some commendable revisions that could be added as recommend solutions in the exercise keys, thus potentially increasing the value of LLM responses. However, in the case of many optional solutions, some of them were poor models for academic writing, as they incorporated complex noun phrases, nominalizations, ineffective sentence structures, or unusual vocabulary. Moreover, LLMs sometimes provided explanations for individual word choices, implying that these were improvements, whereas they reduced to readability and adherence to academic style conventions. This is in-line with recent comparative findings concerning the differences between human and generative AI writing [26]. These features can also reduce readability. As students are potentially faced with many alternatives to choose from, the students need to evaluate them for quality, as some of the alternatives may not follow the concise and clear academic writing expected in the course.

TABLE II
THE SUBMISSION COUNT MEAN, MEDIAN, MAX, STANDARD DEVIATION (SD), AND THE NUMBER OF STUDENTS FOR EACH TASK. THIS IS FOLLOWED BY COMPARISON OF POST-CHATGPT VS PRE-CHATGPT METRICS BY TASK[1].

	pre-ChatGPT					post-ChatGPT					Comparison	
Taskname	mean	med	max	SD	n	mean	med	max	SD	n	ed-to-closest[2]	n-llm-words[3]
Light-before-heavy principle												
automation	3.0	2.0	19	2.51	148	3.32	2.0	19	3.44	130	0.743	1.015
lidar	2.22	*1.0*	19	2.7	149	2.03	2.0	19	1.96	130	*0.424*	0.989
software	1.95	*1.0*	14	2.0	148	2.13	*1.0*	13	2.11	130	0.914	1.000
optical_fibers	3.46	2.0	33	3.69	148	4.22	3.0	20	3.58	130	1.181	1.042
deposition	2.04	*1.0*	17	1.98	148	2.35	2.0	15	2.08	129	0.659	1.124
wind_turbine	**8.22**	**7.0**	**42**	6.83	147	**9.42**	**7.0**	**72**	9.97	130	**1.250**	1.051
technical_debt	2.56	2.0	16	2.28	147	2.79	2.0	17	2.4	130	0.879	1.000
traffic	2.59	2.0	*12*	1.64	147	2.93	2.0	13	2.08	130	1.073	1.021
fermentation	7.14	6.0	37	5.65	147	7.99	6.0	29	5.88	130	1.196	**1.053**
synthesis	*1.8*	*1.0*	17	1.93	147	*1.82*	*1.0*	26	2.65	130	0.916	1.014
convolutional	2.06	2.0	17	1.81	136	2.02	2.0	*8*	1.19	128	0.979	1.019
chemistry_lab	2.57	2.0	13	2.42	136	2.63	2.0	16	2.14	127	0.896	*0.987*
Style and word choice												
cars	5.99	4.0	**38**	5.93	146	7.81	6.0	34	6.14	129	1.700	**3.193**
GPS	5.89	4.0	26	4.93	146	5.33	4.0	*20*	3.74	129	1.032	1.317
system_noerrors	4.59	3.5	22	3.6	146	4.71	3.0	21	3.76	129	0.955	1.705
Recycling	**7.23**	**6.0**	34	5.21	145	8.56	7.5	37	5.54	128	1.028	2.081
solar_power	4.02	2.5	31	4.73	146	3.85	3.0	**42**	4.72	128	0.879	1.301
LEDs	6.35	5.0	31	5.13	145	6.02	5.0	29	4.51	128	*0.845*	*0.888*
mobile_payment	4.47	3.0	26	3.96	145	**10.98**	**10.0**	41	7.26	129	**1.701**	1.181
online_shopping	*3.09*	*2.0*	*21*	3.06	144	*3.27*	*2.0*	24	3.35	125	1.001	1.783

[1] The values are the ratio of the post-ChatGPT metric to the pre-ChatGPT metric. A value greater than 1 indicates an increase in the post-ChatGPT period. The largest values are in **bold** and the lowest in *italics*.
[2] Edit distance to the closest available model solution.
[3] Number of words present in LLM responses but not in model solutions or starter text.

TABLE III
COMPARISON OF PRE- AND POST-CHATGPT RESPONSES AND VOCABULARY COUNTS

Statistic	pre-ChatGPT	post-ChatGPT
Total number of responses	11841	14740
Number of unique responses	5504	6480
Number of unique words/tokens	1580	1817
Real dictionary words	1226	1378
Misspelled words	353	437
Ratio (misspelled/real)	0.288	0.317

B. Features in LLM responses

When reviewing LLM solutions to exercises, we observed certain features that LLMs perform when revising or rewriting texts, which may be problematic in terms of teaching or learning academic writing. Therefore, students and other users should engage in critical thinking before incorporating LLM revisions into their own texts. These features usually change the information content and emphasis in sentences (e.g., in the LBH principle). Some solutions resembled circumlocution, which is something humans (particularly L2 and EFL writers) also do when they cannot solve a problem using a particular technique or due to lack of vocabulary. Below, we list of some main concerns in LLM revisions.

- Omission of hedging, such as auxiliary verbs or defining relative clauses (e.g., *that can be used*), resulting in overconfident tone.
- Extended use of nominalizations (i.e., changing verbs into nouns), particularly in style exercises, reducing readability.
- Addition and deletion of words from the original sentence beyond hedging.
- With default settings, LLMs sometimes produce heavily paraphrased sentences and synonymous expressions, which can greatly change the meaning and emphasis of information.
- Revisions by LLMs are much more extensive than students would normally produce in exam situations because heavy editing and finding synonymous expressions is time consuming.
- Solutions are often based on active voice even if passive voice could be more effective. The suggested solutions may even be forced into active voice (particularly by overusing *we* or made-up subjects). Furthermore, agentless passive was also observed, which sounded unidiomatic.
- The use of subordinate clauses with verbs at the end of sentences. LLMs typically move subordinate or infinitive clauses at the beginning of the sentence, which does not fix the main problem but changes the information focus.

C. RQ2. How does the emergence of generative AI tools show in student attempts to solve the rule-based ASAG exercises?

Based on the descriptive statistics, edit distances, and vocabulary analysis showed in Tables II and III, the student behavior in solving the ASAG exercises has changed since the introductions of LLMs. The post-ChatGPT sample indicated more submissions and trial-and-error behavior as well as greater use of vocabulary found in LLM responses. Manual inspection of student responses also indicated atypical vocabulary and phrases than traditionally present in L2 and EFL engineering student writing.

D. How can rule-based ASAG exercises and LLMs help students in learning scientific writing skills in the LLM era?

Rule-based exercises can be useful in limiting the range of options and not only accepting content that is grammatically correct, or content-wise more or less acceptable. LLMs show great promise in helping students to receive feedback and help on their writing without the fear of losing face by asking naive questions [32]. While LLM feedback and grading could be incorporated into ASAG exercises, more research is needed until LLMs can reach the same level of accuracy as specialized or custom ML models.

E. Limitations and recommendations

Our sample sizes were quite small and students completed the ASAG exercises in uncontrolled conditions. Therefore, we know little about students' strategies (or their lack of) in solving the exercises. As it can be expected that students will increasingly use LLMs as part of text revision (and generation), our recommendation aligns with previous research that students in higher education would benefit from more writing practice as a part of their studies [12] and that students should be supported in developing their feedback literacy skills to effectively utilize LLM feedback [32].

Students may have used LLM models prior to ChatGPT, e.g., GPT-3 which was released in 2020, for their course work. However, this is unlikely as the earlier models were not nearly as recognized, were much less proficient at basic tasks, and required significantly more careful and elaborate prompting to produce desirable results. Another possibility is that students may also have used other tools or grammar checkers, such as Grammarly, while solving the exercises.

It would be beneficial to conduct more research on the differences between human and LLM generated writing in engineering and academic writing contexts. Since it is likely that LLMs will gradually replace traditional ML models in ASAG grading, it would be important to investigate how to improve the accuracy of LLM responses through prompting.

F. Conclusion

In this paper, we explored how well Large Language Models (LLMs) can solve Automated Short Answer Grading (ASAG) exercises in academic writing courses and how student attempts have changed since the popularization of LLMs. Our findings suggest that LLMs can sometimes solve these exercises correctly and in some cases offer pedagogically useful explanations when doing so. We found evidence suggesting that students are using LLMs in their responses to complete these exercises, emphasizing the need to study how to support students in learning academic writing in ways that leverage LLMs productively while mitigating potential over-reliance on automated assistance.

Acknowledgements

This work was in part supported by Research Council of Finland grants #356114 and #367787. We are grateful to Lassi Haaranen for developing the shortanswer content type and preparing the illustrations shown in Fig. 1.

References

[1] ABET. *Criteria for Accrediting Engineering Technology Programs, 2025-2026*. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-technology-programs-2025-2026/.

[2] Paul V. Anderson, Sarah Heckman, Mladen Vouk, David Wright, Michael Carter, Janet E. Burge, and Gerald C. Gannod. CS/SE instructors can improve student writing without reducing class time devoted to technical content: Experimental results. In *Proceedings of the 37th International Conference on Software Engineering - Volume 2*, ICSE '15, pages 455–464. IEEE Press, 2015.

[3] Karen Bennett. English academic style manuals: A survey. *Journal of English for Academic Purposes*, 8(1):43–54, 2009.

[4] Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25:60–117, 2015.

[5] Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027, 2021.

[6] Aubrey Condor, Max Litster, and Zachary Pardos. Automatic short answer grading with SBERT on out-of-sample questions. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM)(14th, Online, June 29-July 2, 2021).*, pages 345–352, Paris, France, 2021. International Educational Data Mining Society.

[7] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, Massachusetts, 3 edition, 2009.

[8] Afrizal Doewes and Mykola Pechenizkiy. On the Limitations of Human-Computer Agreement in Automated Essay Scoring. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM)(14th, Online, June 29-July 2, 2021).*, pages 345–352, Paris, France, 2021. International Educational Data Mining Society.

[9] Robert F. Dugan and Virginia G. Polanski. Writing for computer science: A taxonomy of writing tasks and general advice. *J. Comput. Sci. Coll.*, 21(6):191–203, jun 2006.

[10] Rutwa Engineer, Naaz Sibia, Michael Kaler, Bogdan Simion, and Lisa Zhang. Early computer science students' perspectives towards the importance of writing. In

Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2024, pages 332–338. ACM, July 2024.

[11] Sarp Erkir and Ulas Kayapinar. Engineering competencies and college education: Faculty and employer perspectives on fresh graduates. *European Journal of Educational Research*, 14(2):501–520, March 2025.

[12] Rebecca R. Essig. Writing education examples throughout a first-year engineering course. In *2022 ASEE Annual Conference Exposition*, Minneapolis, MN, August 2022.

[13] Rebecca Rose Essig, Cary David Troy, Brent K. Jesiek, Josh Boyd, and Natascha Michele Trellinger. Adventures in paragraph writing: The development and refinement of scalable and effective writing exercises for large enrollment engineering courses. In *2014 ASEE Annual Conference Exposition*, Indiapolis, Indiana, June 2014.

[14] Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2):489–530, 2025.

[15] Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. Automatic short answer grading in the LLM era: Does GPT-4 with prompt engineering beat traditional models? pages 93 – 103, 2025.

[16] CC2020 Task Force. *Computing Curricula 2020: Paradigms for Global Computing Education*. Association for Computing Machinery, New York, NY, USA, 2020.

[17] Sarah K. Howard, Maryam Khosronejad, and Rafael A. Calvo. Exploring engineering instructors' views about writing and online tools to support communication in engineering. *European Journal of Engineering Education*, 42(6):875–889, September 2016.

[18] Yinchen Lei and Meghan Allen. English language learners in computer science education: A scoping review. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education*. ACM, March 3–5 2022.

[19] Leo Leppänen, Lili Aunimo, Arto Hellas, Jukka K Nurminen, and Linda Mannila. Emergence of llms:(not-so-) significant delving in essay answers in a mooc on the ethics of ai. In *International Conference on Artificial Intelligence in Education*, pages 36–43. Springer, 2025.

[20] Linda H.F. Lin and Bruce Morrison. Challenges in academic writing: Perspectives of engineering faculty and L2 postgraduate research students. *English for Specific Purposes*, 63:59–70, 2021.

[21] Chang Lu and Maria Cutumisu. Integrating deep learning into an automated feedback generation system for automated essay scoring. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM)(14th, Online, June 29-July 2, 2021).*, pages 573–579, Paris, France, 2021. International Educational Data Mining Society.

[22] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[23] Rehmat Munir, Francesco Strafforello, Niveditha Kani, Michael Kaler, Bogdan Simion, and Lisa Zhang. Exploring common writing issues in upper-year computer science. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education*. ACM, March 3–5 2022.

[24] Hanna Niemelä and Johanna Naukkarinen. On the rocky road to academia: Stumbling blocks for Finnish engineering students with English as a second language. *International Journal of Engineering Pedagogy*, 10(6):36 – 56, 2021.

[25] Marko Putnikovic and Jelena Jovanovic. Embeddings for automatic short answer grading: A scoping review. *IEEE Transactions on Learning Technologies*, 16(2):219–231, 2023.

[26] Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do LLMs write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8), February 2025.

[27] Teemu Sirkiä and Lassi Haaranen. Improving online learning activity interoperability with Acos server. *Software: Practice and Experience*, 47(11):1657–1676, 2017.

[28] Carola Strobl, Emilie Ailhaud, Kalliopi Benetos, Ann Devitt, Otto Kruse, Antje Proske, and Christian Rapp. Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, 131:33–48, 2019.

[29] Paulina Szymaszek. Data driven methods for analysis and improvement of academic English writing exercises. Master's thesis, Aalto University. School of Science, 2021.

[30] Qian Wan, Scott Crossley, Michelle Banawan, Renu Balyan, Yu Tian, Danielle McNamara, and Laura Allen. Automated Claim Identification Using NLP Features in Student Argumentative Essays. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM)(14th, Online, June 29-July 2, 2021).*, pages 375–383, Paris, France, 2021. International Educational Data Mining Society.

[31] Edward Wheeler and Robert L. McDonald. Writing in engineering courses. *Journal of Engineering Education*, 89:481–486, October 2000.

[32] Ying Zhan and Zi Yan. Students' engagement with chatgpt feedback: implications for student feedback literacy in the context of generative artificial intelligence. *Assessment amp; Evaluation in Higher Education*, pages 1–14, March 2025.

[33] Justin Zobel. *Writing in Computer Science*. Springer London, 3 edition, 2014.