



Teacher vs. AI:

A comparison of the effectiveness of corrective written feedback provided by either a teacher or AI, and the influence of students' attitudes on the outcome.

Irene Ulrich

Note that some graphs/tables/images have been removed in order to comply with copyright restrictions.

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION



1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	Irene Ulrich
Date:	13.10.2025



Department of Education, University of Oxford

Teacher vs. AI:
A comparison of the effectiveness of corrective written
feedback provided by either a teacher or AI, and the
influence of students' attitudes on the outcome.

Irene Ulrich

St. Anne's College

Word Count: 17,132

Dissertation submitted to the University of Oxford in partial fulfilment of the requirements for
the degree of Master of Science in Applied Linguistics for Language Teaching

August 2025

Abstract

In contemporary times, Artificial Intelligence (AI) has achieved a level of indispensability, yet there persists a lack of academic research that specifically addresses its practical implementation within the context of daily pedagogic practice. One potential benefit of the integration of AI into the field of education is its capacity to enhance the effectiveness of written corrective feedback (WCF), thereby reducing teachers' workload. However, in order to effectively accomplish this objective, it is imperative to consider not only whether AI is performing the task with the same level of proficiency as a human teacher, but also to examine how students respond to AI feedback and whether this response has a significant impact on their learning outcomes.

In order to provide a more comprehensive understanding of the issue, a mixed-method study was conducted in a Swiss vocational school. The students (n=21) participated in two interventions, one teacher-led and one AI-led, to determine if their writing performance improved more, after receiving teacher or AI feedback. Further, they answered a questionnaire to see if they reacted differently to teacher and AI feedback and to determine whether they believed one of the interventions to be more useful than the other and if this perceived usefulness had any influence on their later performance. In addition, the learners kept a learning journal to provide deeper insight into the reactions to the feedback.

In three out of four pre- and post-test, there was a strong interrater reliability between the marks given by AI and the ones given by teachers, confirming earlier studies. However, contrary to former studies, there was no evidence found that performance improved following either teacher or AI feedback when comparing the gain scores from pre- to post-test, nor was there evidence that one intervention was more successful than the other. Also, there was no evidence that performance was influenced by the perceived usefulness of the intervention. Interestingly, the learners' subjective impressions and the objective measures of performance sometimes diverged widely, suggesting that teachers should exercise a certain sensitivity when dealing with AI for classroom intervention.

Nevertheless, if AI is implemented carefully, so that students feel confident in using it, it has the potential to alleviate teachers' workload and provide substantial assistance to students in enhancing their writing abilities.

Acknowledgements

First, I would like to thank my supervisor Professor Elizabeth Wonnacott of the Department of Education. Her expertise and experience helped me to design a study that fit my research interest and to get the relevant results out of the plethora of data I had collected.

I would also like to extend my gratitude to all the other tutors on the course who taught us the necessary basics to delve into the intriguing field of teaching foreign language. You have clearly opened up a rabbit hole for us.

A further huge thank you belongs to my ALLT cohort, aka 'Lovely Oxford Family'. Our weekly meetings were a lifeline, and it helped me to see that we're all in the same boat. I am positive that sharing our experiences helped us all develop and expand our knowledge besides also winning friends all over the world. Special thanks belong to Claudia Concas, Sonia Qinn, Jennifer Ingerson, and Calum To, our close contacts throughout the course, exchanging ideas and experiences, was a pleasure and a relief at the same time.

The last two years, studying and working alongside, would not have been possible without my lovely husband Bruno Ulrich who had my back and was my biggest cheerleader. He even brought some spirit of Oxford to Switzerland by organising carnations for every exam submission. Special thanks also go to my brothers Walter and Hanspeter Späni who cheered me on along the way. Moreover, I would like to thank my friends Nancy Harvey and Mary Patricia Rowland Schueriger for proof-reading my dissertation and for their valuable feedback. I would also like to express my huge thanks to my favourite nerd, Stephanie Lindegger, who came to Oxford despite having a broken arm to help me tick off all the activities on my Oxford bucket list.

Finally, I want to thank our principal, my fellow teachers, and my students who made this research possible by giving me the permission to research in our school, provided useful tips and encouragement, and who participated in this study.

List of Tables and Figures

Table 1: Schedule of the intervention.....	16
Figure 1: Comparisons of marks given by the researcher, AI, and a second teacher	23
Figure 2: Mean scores for pre- and post-test for teacher intervention	24
Figure 3: Mean scores for pre- and post-test for AI intervention.....	25
Figure 4: Mean scores for essay pre- and post-test	26
Figure 5: Mean scores for article pre- and post-test.....	26
Figure 6: Perceived usefulness of AI feedback.....	27
Figure 7: Perceived usefulness of teacher's feedback.....	28
Figure 8: Comparison of the usefulness scores given to AI and the teacher.....	28
Figure 9: Students' preferences for future feedback	29

List of Abbreviations and Acronyms

AI	Artificial Intelligence
ANOVA	Analysis of Variance
AWCF	Automated Written Corrective Feedback
CEFR	Common European Framework of Reference
CUREC	Central University Research Ethics Committee
EFL	English as a Foreign Language
ESL	English as a Second Language
FCE	First Certificate in English
FVB	Federal Vocational Baccalaureate
IELTS	International English Language Testing System
SSPS	Statistical Package for the Social Sciences (IBM data analysis software)
WCF	Written Corrective Feedback
WTC	Willingness to Communicate

Table of contents

Abstract	i
Acknowledgements	ii
List of Tables and Figures	iii
List of Abbreviations and Acronyms.....	iv
Table of contents	v
Chapter 1: Introduction	1
1.1 Background.....	1
1.2 Aims of the study	1
1.3 Rationale of the study	2
1.4 Structure of the dissertation	2
Chapter 2: Literature Review	3
2.1 Written Corrective Feedback.....	3
2.2 AI guided Written Corrective Feedback.....	7
2.3 Students' attitude on the use of AWCF	9
2.4 Research gaps and research questions.....	12
Chapter 3: Methods.....	14
3.1 Research Design.....	14
3.2 Participants and Setting.....	15
3.4 Research Procedure.....	16
3.4.1 Intervention Cycle One	16
3.4.2 Intervention Cycle Two	17
3.4.3 Journals	18
3.4.4 Post-Intervention Questionnaire.....	19
3.5 Materials.....	19
3.6 Coding and data analysis.....	20
3.8 Ethical Considerations	21
Chapter 4: Results	22
4.1 Relationship between teacher and AI marking.....	22
4.2 Influence of type of intervention on gain score.....	23
4.3 Comparison to chance	27
4.4 Students' perception of the usefulness of feedback.....	27
4.5 Students' perception of the intervention process.....	29
4.5.1 Comprehensibility of the feedback	30
4.5.2 Personal relationships.....	31
4.5.3 Time management	31
4.5.4 Amount of feedback	32
4.5.5 Trustworthiness	32

Chapter 5: Discussion	34
5.1 Interrater Reliability	34
5.2 Students' improvement.....	35
5.3 Students' opinion.....	38
Chapter 6: Conclusion.....	40
References.....	42
Appendices.....	50
Appendix A Information for students on how to use AI	50
Appendix B Official FCE marking grid.....	51
Appendix C Journal template for students	52
Appendix D Questionnaire.....	53
Appendix E Writing tasks for pre- and post-tests	54
Pre-Test 1	54
Post-Test 1.....	54
Pre-Test 2	54
Post-Test 2.....	54
Appendix F Confirmation of Ethical Approval.....	55
Appendix G Study information sheet.....	57
Appendix H Consent form	60

Chapter 1: Introduction

1.1 Background

Written corrective feedback (WCF) refers to different kinds of correcting and commenting on students' written work and is one of the integral parts of teaching English as a foreign or second language. For this study WCF is not only considered as the correction of spelling and grammar mistakes but also includes comments on content, communicative achievement and structure. Since WCF is seen as essential not only by teachers but also by students (Ferris, 2004), it is important to research how WCF can be provided most effectively and efficiently, as well as how students receive feedback.

Even though there have been discussions on whether the time spent on WCF is justified (Ferris, 1999; Truscott, 1996), the weight of evidence suggests that WCF is helpful (Brown et al., 2023). However, it is agreed that WCF is very resource intensive and demands teachers' energy and time. This could lead to teacher burnout, a decline in WCF quality, and therefore lead to a decline in teaching quality. Nowadays new technology, namely AI, is emerging which could help teachers to handle their resources more efficiently and therefore counteract this decline. However, like every new development, it has to be determined if AI is indeed helpful and how it can be used to provide accurate and useful WCF. Furthermore, teaching and learning are social acts (Lee, 2008). Consequently, affective factors, such as the attitude towards the source of feedback, could influence students' reactions to feedback and how they engage with it (Storch & Wigglesworth, 2010). Therefore, students' opinions are not to be left out of this equation.

1.2 Aims of the study

The main aim of the study is to explore if students improve more or less in writing tasks depending on the kind of WCF, teacher or AI, they receive. Furthermore, it aims to discover how students react to those two kinds of interventions and if there is a correlation between the perceived usefulness and the actual performance. Additionally, it will be investigated if students prefer a certain kind of intervention and where they see the advantages and disadvantages of those two types of interventions.

Given nearly all students regularly use AI in daily life, it is of the utmost importance that educators know how these programmes can be utilised in the classroom. While several studies have been conducted on the usefulness of AI in WCF, only a few studies were conducted with the generative AI tools that have emerged within the past few years. Also, most of those studies were conducted with paid AI tools that are not available to everyone, which can lead to educational inequity. Those are two limitations that are addressed by this study. To depict the students' realities as much as possible, it was decided to only use the free versions of AI tools

like ChatGPT and DeepWrite which are already widely used by the students of this study and can be used free of charge by everyone.

Additionally, the majority of previous studies have focussed on the efficacy of AI tools when combined with teacher interaction, as opposed to teacher-instruction alone. Often, this was due to the fact that those studies were conducted before the first generation of generative AI emerged. Consequently, since AI is still new and constantly changing, and research into this field is scarce, it is reasonable to see to what new context this research can be expanded. Therefore, the present study will compare the effectiveness of AI-only intervention with that of teacher-only intervention in order to ascertain the extent to which AI can substitute for a teacher.

1.3 Rationale of the study

This study was conducted in a vocational school in Switzerland. The students in this research project study towards a Federal Vocational Baccalaureate (FVB) which, upon completion, enables them to study at a technical college. In general, FVB students are high achieving, highly ambitious, and have a strong interest in technology. So, it was these FVB students who had been the first ones to introduce ChatGPT to the teacher/researcher of this study, even before this new technology was made aware to the wider public. From then on, it was the teacher's aim to find ways to utilise this technology in everyday classroom practise. As one of the aspects of English language education students struggle most with is writing, it seemed logical to try to combine a technology which seems to interest them with a part of English learning they struggle with to see if this combination could help them. What is more, as many researchers agree, giving feedback on writing is also very time and resource intensive (Hartshorn et al., 2010; Kang & Han, 2015). Thus, finding a technology that could help with such tasks could not only benefit students, but also support teachers.

1.4 Structure of the dissertation

This dissertation contains six chapters. Chapter One has introduced the background and aim of the study. Chapter Two investigates the research done previously in this field starting with general research into WCF, followed by newer studies on AI supported WCF, and studies considering students perceptions of WCF. Then, limitations of previous studies are discussed and the research questions for this study are introduced. Chapter Three explains the methods used for this study, explaining the research design, introducing participants and setting, and describing the research procedure, material used, and the data analysis procedure. Chapter Four presents the results divided into quantitative and qualitative findings. Chapter Five discusses these findings in relation to previous studies and points out some limitations of this study. Chapter Six then draws conclusions and derives some implications for pedagogy and future research.

Chapter 2: Literature Review

There has been extensive research into written corrective feedback (WCF). Researchers thereby have not only discussed if feedback is helpful, they have also discussed what kind of feedback seemed to be most helpful (Ellis et al., 2008; Ferris, 1999, 2004; Truscott, 1996). Even though nowadays researchers agree that WCF can be effective (Bruton, 2009), it is still debated whether or not it is worth the time and effort teachers invest when giving students feedback, considering the effectiveness of this feedback (Brown et al., 2023). This chapter will give a short overview of the research that has been conducted in the field of WCF, especially considering what feedback is helpful and what that means for teachers in terms of time management as this is a reoccurring concern in several studies (Ferris, 1999; Hyland & Hyland, 2006). Then, the focus of this chapter will shift to what kind of research has been done in terms of artificial intelligence, and how this could be a solution to the conflict between providing effective written corrective feedback and teachers' time constraints. Further, it will be investigated how far the attitude of students towards feedback given by teachers or AI has been considered in those studies.

2.1 Written Corrective Feedback

In the late 1990's and early 2000's there was a debate about whether written corrective feedback was helpful or not, or possibly even harmful. Truscott (1996) argued that the correction of grammatical errors can have side effects, which should be included in the consideration of whether WCF is useful or not. On one hand, they mention limitations on the teacher's side, such as teachers' abilities to adapt their corrections to the learners' learning processes. If teachers lack this ability it can lead to corrections being overly simplistic and failing to adequately address the underlying morphological, lexical, and syntactic processes involved in students' writing. Moreover, even if teachers were able to address all those processes, time constraints and patience could influence the quality of the provided corrections leading to inconsistent and non-systematic corrections (Zamel, 1985). On the other hand, Truscott (1996) raised the concern that WCF could harm students' wellbeing. They fear that students could be overwhelmed by the amount of corrections and therefore lose the willingness to deal with those corrections. This would lead to students losing their confidence and joy in learning. Thus, the researchers conclude that the time and energy absorbed by correcting written work and dealing with the feedback were not worth it and could in fact be detrimental. As an answer to this, Ferris (1999, 2004) argued that, even though teachers' corrections are often inconsistent and teachers sometimes struggle to provide appropriate feedback, WCF is an essential part of writing classes and valued by students. However, they point out that it must be done in an effective way which protects the scarce time resources of teachers. Subsequently, researchers started to agree on the

usefulness of WCF (Bruton, 2009; Ferris & Kurzer, 2019; Ferris, 2010) and started to research the best ways to provide such feedback to students.

One of the most widely discussed topics is how concrete and extensive feedback needs to be in order to be effective. A meta-analysis of 52 studies from journal articles ($n = 42$) and doctoral dissertations ($n = 9$) published between 1980 and 2019 found that the direct coding of errors with metalinguistic comments was the most effective form of feedback (Brown et al., 2023). One such study looking at the effects of direct feedback with or without metalinguistic comments was carried out by Sheen (2007). They conducted a study using a pre-test, post-test quasi-experiment with an additional delayed post-test. The participants ($n = 91$) were students in the American Language Program (ALP) of a community college in the US. The participants were convenience sampled into three groups: direct-only correction ($n = 31$), direct correction with metalinguistic feedback ($n = 32$), and a control group ($n = 28$). The two correction groups received two treatment sessions that taught them how to use articles while the control group received no such treatment. The results showed a significant difference between the three groups in the writing post-test 2 ($p < 0.01$), and between the group that received direct metalinguistic feedback and the control group for the speed dictation post-test 2 ($p < 0.05$). In both cases, the metalinguistic feedback group outscored the others.

However, it is not only debated whether metalinguistic comments should be provided or not; it is also deliberated whether teachers should focus on single error types or rather correct multiple kinds of errors. Ferris (1999) argues that, by focussing on a single error type, the accuracy and detail of feedback could be improved. Further, teacher's burn-out could be prevented because they could cut down on the vast amount of time and energy they otherwise spend on corrections. Additionally, Brown et al.'s (2023) meta-analysis, found that focused error correction has double the effect of mid- or unfocused feedback. But still, the question is if such a single-error approach is practicable in the classroom. In a later journal article, Ferris (2010) states that focussing on single errors might not be ecologically valid, or in other words, might not be relevant for teachers and students in real life settings (Storch, 2010), since L2 learners aim to improve their writing skills across several linguistic features and not just one. Therefore, teachers might feel obliged to correct more than just a single error type to meet the needs and objectives of students (Ellis et al., 2008; Kang & Han, 2015). Thus, researchers studied the effect of focussed compared to unfocussed feedback. Similar to Sheen's (2007) study, Ellis et al. (2008) researched the correction of articles. Their study also utilized a quasi-experimental with pre- and post-test design with a delayed post-test, with a control group ($n = 11$) that received no feedback. The difference between the two studies is that Ellis et al. used three interventions, instead of only one, and that the two experimental groups received either focused ($n = 11$) or unfocused feedback ($n = 13$), instead of direct and indirect feedback. Also, the study was conducted in an EFL context while most of the previous studies on WCF took place in an ESL

setting. The findings of the study showed that there was a significant difference between the three groups in the delayed post-test ($p < .001$), but that the difference between the two experimental groups was not significant. However, the results might not be completely reliable since, according to the researchers, the control group was notably, even if not significantly, weaker than the other groups in the pre-test. Also, the researchers state there might have been a potentially significant difference between the two experimental groups, which should be avoided in further research.

A similar study, looking into the impact of focussed and unfocussed feedback on the enhancement of writing performance among EFL learners, was conducted by Kassim and Ng (2014). In comparison to Ellis et al. (2008), they (Kassim & Ng, 2014) focused on the correct use of prepositions. They studied the progress of 90 students at a technical university in Malaysia over a 12-week period, with two treatment sessions. They worked with three groups: a focused group ($n = 30$), an unfocused group ($n = 30$), and a control group ($n = 30$). The results showed that both the focused feedback group ($p < .05$) and the unfocused feedback group ($p < .05$) improved significantly in the accuracy of preposition usage, compared to the control group that showed no significant improvement ($p = .771$). Still, there was no significant difference between the improvement of the two intervention groups ($p = .192$). This seems to confirm the findings of Ellis et al. (2008) that focused and unfocused feedback both lead to a significant improvement in student's writing and that those findings can be seen in an EFL as well as in an ESL context.

However, both studies focused only on the improvement of one single grammatical feature, and it therefore might not be generalized. Still, it might be relevant to many teachers to see that there was no difference between focusing on one error type or providing more general feedback, because in authentic teaching situations, teachers rarely correct only certain preselected error types (Ferris & Kurzer, 2019). Also, showing that extensive feedback with metalinguistic comments and unfocussed feedback produce similar results as focussed and limited feedback counter Hartshorn et al.'s (2010) claims that it may be impossible for students to process and employ all the feedback received.

Nevertheless, even though students might not be overwhelmed by extensive WCF and teachers might see the benefit of it, there is no guarantee that feedback is accurate and useful. Teachers might still not be willing or able to identify and correct the errors students make (Ferris, 1999). Or if they do, they might fail to explain the problem sufficiently, due to time restrictions and limited patience, which then again affects the quality of the feedback (Truscott, 1996). In a study about the assessment of interlanguage performance, Cohen and Robbins (1976) made a chance discovery. Before they started with the data analysis, they took a look at the teacher corrections. This investigation revealed an inconsistency in teachers' feedback. Notwithstanding

the provision of a correction checklist to the teachers, the researchers found that, on occasion, an error would be corrected while an identical mistake would be overlooked at another point. Even more surprisingly, corrections were frequently omitted altogether. Also, they found that corrections were often too general which rendered them useless to the students. However, it could be criticized that those findings were a byproduct of a study not intended to look into the quality of teacher correction and therefore not generalizable. Further, only three graders were involved in the correction, therefore the sample was quite small.

A larger study explicitly looking into teachers' responses to students' writing was conducted by Zamel (1985). They examined the responding behaviours of 15 teachers. To do so, they analysed the corrections of 105 students' texts. In general, each teacher responded to more than three students and corrected two or more writings per student. Also, initially, the writings were not intended to be used for an analysis of teachers' corrections but previously collected to build students' portfolios. Therefore, by the time of correcting the writings, the teachers were not aware that their feedback would be used for a study on corrections. The findings of the study confirmed many of the claims made by Cohen and Robbins (1976). It was found that teachers were inconsistent in their corrections, with corrections being arbitrary and contradictory, sometimes even adhering to abstract rules and standards. Also, teachers regularly misread student's writings and rewrote whole sentences to fit their misreadings. What is more, teachers failed to provide strategies for revising texts, even though this would be a key to help students to become a confident writer (Bruton, 2009).

Therefore, the question remains how teachers might provide useful feedback to help students. Learning preferences and needs differ widely from student to student. To give adequate feedback, a teacher should be able to utilize and facilitate the already present strategies, but this in turn calls for more resources and strategies to be used (Lasagabaster & Sierra, 2005). Therefore, Hwang et al. (2023) argue that the best results can be achieved by giving personalised feedback on students' original writings. This argument is based on the findings of their study with Indonesian university students ($n = 104$). They detected a significant correlation between the scores on the writing post-tests and the amount of feedback demanded by students for the practice activities ($r = .346$; $p = .049$), as well as for the assignment activities ($r = .373$; $p = .032$). Moreover, a case study on writing portfolios found that students who receive ongoing feedback on a working portfolio not only improve in their writing but also show enhanced self-regulation compared to the group that only received feedback on two items of their showcase portfolios (Lam, 2013). These findings underscore that supporting language learners to improve their writing can be very resource intensive for teachers and preferences can vary from student to student (Ferris, 2004; Hyland & Hyland, 2006).

Consequently, trying to follow all those suggestions for feedback puts a lot of strain on teachers' time resources and makes it hard for teachers to ensure learners are treated fairly. In larger classes it could be almost impossible to cater to the needs of every student (Lasagabaster & Sierra, 2005). Here AI could play a crucial role in relieving pressure on teachers and providing students with the feedback they need (Wang, 2022).

2.2 AI guided Written Corrective Feedback

One of the affordances of automated written corrective feedback (AWCF) is that students receive immediate feedback (Nazari et al., 2021). If the students receive delayed feedback, as with teacher WCF, they might still know what they had written, but they may have difficulties recalling the cognitive process they engaged in during writing (Abdollahzadeh et al., 2022). Therefore, by providing them with immediate feedback, students get a chance to notice corrections consciously and link them with the writing process in which they were engaged (Barrot, 2023; Yu & Yang, 2021). Aside from its timeliness, AWCF is also personalised, accurate, and objective (Rad et al., 2023). These are all points that sometimes seem to be difficult for teachers, as was shown in the previous section. Moreover, with the introduction of generative AI, AWCF nowadays cannot only correct grammar mistakes, AI programmes can now also comment on features such as content, communicative achievement, or organisation.

The question that remains is whether feedback provided by AI programmes is as effective as traditional forms of feedback. Studies conducted on the effect of AWCF have found that experimental groups that received AWCF outperformed groups with teacher or peer-feedback in pre-/post-test designs. Chang et al. (2021), for example, conducted a study with 53 Chinese university students aged 19-21 and seven years of English language learning experience. The participants were convenience sampled into two classes. Both classes were taught by the same instructor with the same course materials, following the same syllabus. Additionally, the experimental group ($n = 30$) received feedback via the premium version of Grammarly, while the control group ($n = 23$) received peer feedback in class. During the 16-week intervention period, both groups wrote and received feedback on five essays. For the analysis, the researchers gathered quantitative data from a pre- and a post-test and qualitative data from three open-ended survey questions. The essays were not only graded according to the accuracy in grammar and language, but holistically with scores for content, organization, language use, vocabulary and comprehensibility. The researchers report a significant difference in the post-test ($t(52) = 2.10, p = .044$) with a medium effect size ($d = .603$) in favour of the experimental group. Also, they state that 80% of the students had a positive attitude towards Grammarly and that some students even testified that Grammarly lowered their writing anxiety. But still, 10% of the participants stated that Grammarly sometimes provided incorrect answers, raising questions on the reliability of AI feedback. Despite the positive results considering the usefulness of AI

feedback, one limitation of this study is certainly the convenience sampling. By using two existing groups, it cannot be ruled out completely that part of the results could be due to the differences of the groups.

Still, Chang et al.'s (2021) findings were confirmed by other researchers, for example Chen and B.D. (2023). They conducted a study with freshmen of a Chinese public university ($n = 65$) who were sampled into an experimental group ($n = 34$) and a control group ($n = 31$). Unfortunately, it is not reported how exactly the students were sampled into the groups. The researchers used a pre- and post-test design accompanied by a questionnaire to collect both quantitative, as well as qualitative data. Both groups received two treatment sessions of 25 minutes on two consecutive days. While the experimental group received AI feedback, the control group only received target writing models of articles. Unfortunately, it is not elaborated further what kind of AI-feedback system the experimental group used, and what this use consisted of. Even though the researchers report significant differences between the groups for both pre- and post-tests, they claim that the AI feedback was significantly more effective than the traditional feedback. Unfortunately, they do not provide solid evidence for this claim, as they fail to report the findings of the qualitative data and link them to the quantitative findings. It also seems that the initial research was followed by a longitudinal study, which again was not stated clearly. Unfortunately, there are only a few studies investigating the usefulness of AWCF, so that it often cannot be determined whether the existing data is valid and reliable.

Most of the studies on AWCF were conducted before the invention of generative AI and Chatbots, and even newer studies (e.g. Chang et al., 2021) still use programmes like Grammarly to research the effects of AWCF on students writing and their perceptions of AWCF. Even though most studies found a facilitating effect of AWCF when compared with teacher feedback, many of the studies claimed that AWCF on its own might not be as effective as teacher feedback because AWCF cannot explain the corrections and the students have no opportunity to ask questions. In the times of ChatGPT, Claude and other generative AI programmes, this point seems no longer valid. A more recent study using generative AI, and confirming the usefulness of AI, was conducted by Ghafouri et al. (2024). They researched the impact of a ChatGPT-based writing instruction on students' writing performance, compared to the performance of students in control groups. For the study, the researchers recruited twelve Iranian teachers that taught weekly online classes. Each teacher chose four of their students to participate. Those teacher/student groups were then randomly assigned into experimental ($n = 6$) and control groups ($n = 6$). Both groups then were taught in two 90-minute lessons per week over the course of ten weeks. ChatGPT was used by the teachers of the experimental group for lesson planning and also to get detailed analyses of students' writings, which they then provided to the students. Students were then given a few minutes per lesson to interact with ChatGPT themselves and were asked to use ChatGPT for their homework. The control groups received traditional writing

instructions from the teacher. Pre- and post-tests were marked by ChatGPT and an experienced IELTS instructor. The interrater reliability showed a significant agreement for the pre-test ($r = .697, p < .05$), post-test ($r = .658, p < .05$), and delayed post-test ($r = .648, p < .05$). The results of the intervention show that there was a significant difference between the two groups on the post- and delayed post-test after controlling for the effect of the pre-test ($F(1,45) = 44.055, p < .001, \eta^2 = .50$). Even though, in this study, AI feedback was directly forwarded to students, and they had the possibility to directly interact with ChatGPT, the prompting for feedback was still largely done by the teachers. This strains teachers' resources unnecessarily, because the prompting might as well be done by the students themselves. What is more, all those studies used AWCF to complement teacher WCF and not as a means of instruction of its own. This current study will take a different approach and compare teacher WCF and AWCF by generative AI as two separate sources of feedback, rather than comparing teacher WCF with a combination of AWCF and teacher feedback. Additionally, in comparison to former studies, the students are told that ChatGPT can switch languages and give feedback in English as well as in the student's native language. So, the limitation that students cannot effectively process the feedback due to limited knowledge of English (Fan, 2023) should therefore be counteracted.

A reason why this study will let students interact with AI directly is the suggestion of researchers that the way students interact with AI has an influence on how much they improve. Particularly, they can revise their texts several times (Warschauer & Ware, 2006), while not having to worry about time since they can do it anytime and anywhere (Hwang et al., 2023; Yan, 2023). Consequently, the use of AI supported tools lead to more learner-centred, personalised ways of learning (Jurs & Špehte, 2022). Naturally, to make use of the possibilities AI offers, it is crucial that students interact actively and reciprocally with it (Rad et al., 2023). When they do, AI can help reduce barriers by simulating peer-feedback rather than teacher feedback (Tai & Chen, 2023). Thus, students might develop a sense of social interaction which in turn could affect how much they learn (Song & Song, 2023). Abdollahzadeh et al. (2022) even go as far as stating that teachers can function like a barrier to real language learning environments. Therefore, the teachers' involvement and those barriers could be eliminated by AI through evoking more real language learning conditions.

Consequently, it is reasonable to suggest that the kind of intervention itself might not be the only reason for improvements of students' writing skills. It seems also crucial how students engage with the feedback and its source (Ellis, 2010).

2.3 Students' attitude on the use of AWCF

According to Ellis (2010), engagement with feedback is influenced by the type of feedback, as well as individual and contextual factors. They propose three dimensions of engagement: cognitive which describes how students embed feedback into their existing knowledge,

behavioural engagement which describes how students revise their texts and interact with feedback, and affective engagement which describes learners' attitudes towards the feedback received. Han and Hyland (2015) see those three types of engagement as crucial factors connecting the provision of feedback with the learning gains of students. In their multiple-case study with students from four Chinese universities, they discovered a complex connection between the different forms of engagement. For example, while one student's disappointment, or "negative affective reactions" (p. 40), after receiving negative feedback resulted in negative behavioural engagement, another student turned this disappointment into motivation. Also, while for two students their performance reflected their behavioural engagement, two others managed to perform well despite showing only superficial cognitive engagement. This interconnectedness between engagement and performance has been highly researched in the last decade (Yu & Yang, 2021). Dong (2024), for example, found that the behavioural engagement of students significantly predicted the complexity of the writings in the post-test ($\alpha = .452$; $p < .001$), even though there was no correlation found for the other aspects of writing, such as accuracy or fluency. However, this study was one of the few researching the correlation between engagement and writing scores. Most studies looking into engagement and WCF are qualitative studies (Nazari et al., 2021; Rad et al., 2023) or mixed-method studies that use engagement as one of the informing factors to explain the difference in writing scores (Song & Song, 2023), but they do not statistically analyse the effective correlation. Also, most of those studies focus on the influence of the kind of feedback on the behavioural and/or cognitive engagement, but ignore the affective engagement, and especially the attitude towards the source of feedback (Storch, 2010).

In fact, very few studies on WCF have researched the area of affective engagement (Miao et al., 2023). But, since learning is a social act and not a mere transmission of knowledge (Lee, 2008), students' reactions to feedback cannot be left out of the picture. Those affective factors can influence the way learners interact with feedback as well as if they accept the feedback at all (Storch & Wigglesworth, 2010; Wilson & Czik, 2016). Therefore, students' attitude must be incorporated in the analysis of teaching interventions and their usefulness in education (Storch, 2010). Storch and Wigglesworth (2010) found that linguistic as well as affective factors, especially the attitude towards the feedback, influenced how much was learned from feedback and how long it was retained. They also put forward the hypothesis that if students resist the feedback, for example because they deem it useless, they are less likely to retain it. What is more, Yu et al. (2021) examined feedback that might even have a negative impact on the writing performance of students. In their study with Chinese university students ($n = 1,075$), they found that too general feedback with a lack of depth, as well as feedback that is not adapted to specific students, can lead to a negative view towards feedback and therefore can also have a negative influence on students' engagement. In their study, they specifically mention that the

programmes used in China to give AWCF supply too homogenous feedback. However, with the emergence of generic AI, such programmes can more easily adapt and generate individual feedback addressing the specific needs of students. Therefore, it is of utmost importance to repeat such studies regularly to keep up with technological developments and to prove whether findings of former studies are still valid.

Even though some studies research how students' affective factors influence the way they interact with feedback (Han & Hyland, 2015; Storch & Wigglesworth, 2010; Zhang & Hyland, 2018), only a few attempt to link the affective factors with the actual improvement in writing performance. One such study was attempted by Hou et al. (2024). They conducted a study with first-year college students ($n = 22$) in China. They used a pre-/post-test design with pre-, and post-test marked by an AI programme (Gemini). Further, they let students answer a questionnaire with 22 Likert-scale and three open-ended questions. Unfortunately, only nine students did the post-test, and the researchers did not evaluate the results statistically. However, all nine students improved with the AI intervention and most of the students reported that they trusted the AI feedback. They particularly mentioned that they liked the speed and level of personalisation. Therefore, it can be hypothesised that the participants in this study showed high affective engagement which might have led to the improvement.

A study that was evaluated statistically was conducted by Zahroh et al. (2020). They did a quantitative correlation study at a university in Indonesia. They used a questionnaire, interview, and writing test to link the students' ($n = 50$) engagement to their writing performance. Unfortunately, the exact procedure and material used was not documented and therefore the reliability of this study cannot be determined. But assuming that the study was valid, the researchers found a moderate and significant correlation between students' attitude towards feedback and their writing scores ($r = .572, p = .000$). However, given the poor documentation of this study, there certainly need to be further studies investigating this relationship to be able to make generalizable assumptions. Generalizability of students' engagement, especially when the affective factors are concerned, seems to be a problem in general. Yu and Yang (2021) conducted a systematic review of 64 empirical studies on the students' response to WCF published between 2010 and 2021 and even though the title of their study suggests differently, they not only included studies on teacher feedback but also studies including automated feedback. The study showed that the affective engagement with feedback varies highly depending on the students' language proficiency but also depending on the cultural context. While students in Japan prefer handwritten feedback by their teachers, students in Iran prefer automated feedback. Therefore, the trust in the various kinds of feedback differed depending on the cultural context as well. Thus, those studies cannot be globally generalized. Since this systematic review mainly included studies from Asia and countries with English as an official

language, and only eight studies conducted in Europe, it certainly makes sense to check those findings in different cultural contexts.

Consequently, the affective engagement, particularly the different attitudes towards teacher or AI feedback, should be further researched. Especially because various recent studies attest that technology has a positive effect on students' performance such as improved engagement (Nazari et al., 2021), perseverance (Schindler et al., 2017), and self-regulation (Song & Song, 2023). Often, new technologies are seen as less intimidating and more approachable than teachers (Rad et al., 2023). Through the dialogic process of receiving feedback from a Chatbot, students feel heard, which can enhance students' engagement with the feedback (Wongvorachan et al., 2022). Also, it was found that the willingness to communicate (WTC) could be significantly improved by using chatbots (Tai & Chen, 2023). While students showed low WTC and even anxiety to communicate in conventional classroom settings, they interacted actively with chatbots and reported lower levels of anxiety.

However, there are also studies that claim that students value the feedback from a teacher over AI feedback (Link et al., 2022). Therefore, the attitudes towards the source of feedback must be analysed thoroughly because it would be counterproductive to forcefully implement something just because it is new and exciting, but contrary to the preferences and needs of the learners.

2.4 Research gaps and research questions

As Bruton (2009) states, studies into WCF do not necessarily depict actual writing pedagogy. More precisely, in research into WCF, students receive feedback from researchers rather than their teachers, writing tasks are often artificially created instead of real life tasks, and feedback is limited to certain linguistic features (Ferris & Kurzer, 2019). Further, studies usually look at the effect of AWCF for the revision of previously written essays and not on the transfer of skills when writing new essays (Stevenson & Phakiti, 2014). Therefore, this research will be conducted in the researcher's own classes, so that the main teacher of the class also serves as researcher, and thus, the feedback will be provided by the students' actual teacher. The writing tasks are taken from real life situations and from exams the students will sit within the next 10 months. This study uses five different topics for essays and five for articles whereas two of each are used for pre- and post-test and the other three each for the intervention phase. In terms of feedback, the students will receive holistic feedback commenting on different linguistic features as well as content, communicative achievement, and organisation, just as they would in an assessed exam.

To answer the limitation of convenience sampled groups as seen in Chang et al. (2021), this study will conduct a group comparison with repeated measures to control for group effects. So, both groups each receive teacher intervention and AI intervention, but at different time points and for different tasks.

Also, in comparison to former studies conducted on AWCF, this study will use a generative AI chatbot which can not only correct students' work and give feedback, but also interact with students' own questions and give detailed or additional information. Further, this chatbot is able to provide answers in the students' native language. What is more, both pre and post-tests will be marked by AI and the teacher. These marks will then be compared to determine whether the AI marks are reliable.

Furthermore, students' attitude towards the provider of feedback and other affective factors, which have a significant influence in classroom interactions, have frequently been neglected in former WCF studies (Storch, 2010). Moreover, there has been a lack of consideration of the connection between these factors and students' improvement. Therefore, this study will use questionnaires and journals to elicit students' attitude towards the two feedback sources and will then calculate the correlation between students' attitudes, namely the perceived usefulness, and the improvement in their writing scores to see if the attitude towards the source of feedback has any relation to the improvement.

Accordingly, this study tries to answer the following research questions:

- Is there a difference in the effectiveness between writing instruction and feedback administered by a teacher or by AI?
- What are the students' opinions on the usefulness of those two sources of feedback?
- Do the students' perceptions influence the actual outcome of the intervention?

Chapter 3: Methods

This chapter starts by outlining the research design. Then it introduces the participants and settings before outlining the research procedure in detail. This is followed by an explanation of the material used and the coding and data analysis procedure. Finally, it addresses some ethical considerations made by the researchers.

3.1 Research Design

For this research it was decided to conduct a quasi-experimental study with two groups. To build a holistic picture of the complex topic of written corrective feedback, a mixed method design was chosen. The study consisted of a quantitative intervention study with pre- and post-tests, and also qualitative data drawn from students' journals and a questionnaire (Paltridge & Phakiti, 2015; Selvi, 2019; Tashakkori & Teddlie, 2003).

For the quantitative study, a group comparison with repeated measures was chosen. In the two intervention stages both groups function once as an experimental group and once as a comparison group. As both groups receive language instruction in either the experimental setting or via traditional methods, the second group is referred to as a comparison group rather than control group. Due to the same subject design, the influence of the individual differences of students should be minimized because they stay the same for each treatment (Rose et al., 2020). Further, as the in-group effect of an intervention can be stronger than the between-group effect, by using a same subject design with repeated measures it can be determined which of the two effects are stronger; in-group or between-group (Stevenson & Phakiti, 2014). Additionally, it is impossible to construct a design where the first intervention does not somehow influence the following interventions. To counter this effect, the order of the treatment for the two groups was switched (Rose et al., 2020). So, one group received the AI intervention first, while the other was taught by a teacher. Furthermore, the writing exercises were switched after intervention 1. In the first intervention the students were asked to write an essay; in intervention 2 they were writing articles. Since those two pieces of writing have some distinct characteristics in style, register, and organisation, it was hoped to further weaken the influence of intervention 1 on intervention 2. Additionally, by using a pre- and post-test design the improvements within a specific intervention could be measured, and the conclusions could be drawn according to this improvement.

For the intervention and test design, the suggestions of Bruton (2009) were followed by implementing a pre- and post-test design with two different texts, as well as corrective feedback that was provided over a period of four weeks per intervention cycle. Further, the students were asked to write a journal to ensure that the given feedback has been genuinely interacted with by the students. Such journals help to understand how students receive and interact with feedback

and subsequently, what factors influence the usefulness of written corrective feedback (Shen et al., 2023).

Furthermore, since students' attitude towards and trust in the source of the feedback may influence the outcome of a writing intervention, a questionnaire was used to connect those factors to the writing outcomes (Ferris & Kurzer, 2019). Also, such questionnaires can allow researchers to spot underlying issues which may influence the research outcomes (Cohen et al., 2017). For the quantitative part, a five-point Lickert scale was used to elicit the perceived usefulness of the two interventions. Additionally, to collect qualitative data, four open-ended questions were used (Paltridge & Phakiti, 2015).

3.2 Participants and Setting

The participants were all students of two EFL classes preparing for the Federal Vocational Baccalaureate (FVB) in Switzerland. This programme enables the students to pursue their A-level certification alongside doing an apprenticeship. After receiving their FVB, students are eligible for entry to technical colleges, and with an access course, they are also allowed to study at university.

Due to practical reasons, it was decided to use two intact classes taught by the researcher of this study. One of the classes, later referred to as experimental group 1 (EG1, $n = 13$), was doing the baccalaureate as a three-year program alongside their apprenticeship, while the second class, referred to as experimental group 2 (EG2, $n = 20$), has already finished the apprenticeship and did the FVB afterwards in either a one-year or a two-year model. Both the one-year and two-year model receive English instruction in the same class. Given the different models, the two groups differ slightly in their age (EG1: MED = 19.01, SD = 1.07, EG2: MED = 20.37, SD = 1.40,). However, both groups have about 7 years of prior English learning experience and are at an intermediate level (CEFR: level B1-B2), as determined by entrance tests conducted in the first semester of their course. Also, both classes receive 160 General English lessons over the course of three semesters for EG1 and two semesters for EG2. As the number of semesters differ, EG1 had already started with the English course half a year before the study took place, while for EG2 the study started with the beginning of their English course. Due to the differences in age, experience, and programme, validity of a parallel group comparison could not be guaranteed. So, it was decided to conduct a group comparison with repeated measures. This way, an attempt was made to limit the influence of previous interventions or students' previous experience on the outcome of the study (Rose et al., 2020).

Two students decided to opt out of the study at the beginning. The writing results of 10 more students had to be omitted from the database due to incomplete or falsified data. One student used DeepL in his pre-test, four other students missed the first post-test, while five more missed the second post-test. However, all students participated in the intervention of the study because

it was part of their normal school lessons and formed part of the curriculum. Therefore, all students conducted the tasks, but only the data of the students who consented to the use of their data and who completed all the tasks including the tests was used for quantitative analysis (EG1: $n = 13$, EG2: $n = 8$). One student missed post-test 1 but completed all other tasks, filled in the questionnaire, and submitted their journals. Therefore, their data was deleted from the quantitative part of the study. Their answers to the questionnaire and their journal, however, were used for the qualitative part of the study. Hence, the number of participants for the qualitative part (EG1: $n = 13$, EG2: $n = 9$) differed from the number of participants for the quantitative part.

3.4 Research Procedure

The participants participated in two writing intervention cycles, as illustrated in Table 1. Each intervention consisted of a pre-test, a four-day writing workshop, and a post-test. After completing both interventions, the participants filled in a short questionnaire. The pre-tests were conducted to control for initial differences that might exist between the two groups or even individual students (Rose et al., 2020).

Table 1

Schedule of the intervention

Week	Intervention Stage		Writing topic	EG1	EG2
1	Pre-test 1 (Essay)		Adventure tourism		
2	Intervention 1	introduction to essay writing, writing essay 1	Different types of teachers	TEACHER	AI
3		feedback on essay 1, writing essay 2	Travelling	TEACHER	AI
4		feedback on essay 2, writing essay 3	Being creative	TEACHER	AI
5		feedback on essay 3		TEACHER	AI
6	Post-test 1 (Essay)		Without failure there can be no success		
2 Week school break					
7	Pre-test 2 (Article)		What I'd save in a fire.		
8	Intervention 2	introduction to article writing, writing article 1	Your favourite time of day	AI	TEACHER
9		feedback on article 1, writing article 2	Sense of smell	AI	TEACHER
10		feedback on article 2, writing article 3	Giving and receiving advice	AI	TEACHER
11		feedback on article 3		AI	TEACHER
12	Post-test 2 (Article) & Questionnaire		My closest friend		

3.4.1 Intervention Cycle One

The topic of the first cycle was essay writing. EG1 received teacher feedback throughout this intervention stage. In week one, they received an introduction on how to write an essay and were asked to write an essay on a given topic. In weeks two and three, they received teacher feedback on their writing from the previous week. Each week, in the first lesson, the teacher would give short feedback on the most common mistakes and how to avoid them to the whole class. Then, the students had time to engage with the feedback received and to ask questions about the feedback. In the second lesson, they were asked to write a new essay on a different

topic which changed from week to week. In week four, they received the final feedback on their last essay. The essays of the intervention stage were not included in the analysis but were only used for practise purposes. Each week, the students were also given some time to write journal entries. For the journals, they were asked to report their reactions to the feedback and how they incorporated the feedback into their next writing task. The intervention for EG2 followed the same structure, except that they were asked to use AI-tools for instruction and feedback. They were not given any instruction or feedback on their writing from the teacher at any time during this intervention, but they were given instructions on how to write prompts in order to receive appropriate feedback from AI (see Appendix A). For feedback on their writings, for example, they were given following prompt:

I'm a student on a B2 level according to CEFR. I was asked to do following writing task: "ENTER TASK (e.g. Is it better to do things with other people or to do them on your own? write an essay of 140-190 words, mentioning work at school or college, sports activities, and an own idea)". I wrote: "ENTER YOUR WRITING". Please mark my writing according to the provided marking grid.

Also, they were given the official marking grid (see Appendix B) to upload to ChatGPT, so AI could mark each student's writings accordingly. Further, students were made aware that ChatGPT can switch languages from English to German easily and accurately, and that they should make use of it in case they struggle to understand the feedback received from AI. Therefore, students received feedback in English and their native language depending on their preferences, to ensure that they understood the feedback received. What is more, even though students were given example prompts to receive marking and some general feedback, they were free to use prompts of their own to refine the feedback, get additional information, and suggestions for improvement.

3.4.2 Intervention Cycle Two

The second intervention cycle took place one month after the first. Because both groups had already worked on essay writing in stage one, it was decided to switch the writing focus from essays to articles. While essays are very formal when it comes to sentence construction and the use of vocabulary, articles are more informal and sometimes even chatty. This called for a good command of language from the students and also a shift of focus. The reasons for this change were that, in this way, there is less influence of the first intervention on the results of the second intervention, and it could be ensured that the students received instructions on different kinds of writing, as it is required by the curriculum. Also, the change would provide robustness, if benefits of one feedback type could be found in both writing contexts. For the second intervention period, EG1 used the AI tools and EG2 received teacher feedback. The writing tasks as well as the marking grid were taken from sample exams of the Cambridge First Certificate in English (CEFR level B2).

Post-tests were administered immediately after the intervention periods to minimise the influence of outside effects as recommended by Cohen et al. (2017).

The students were fully informed about the approach and the aims of the study. The reason for this is that the students already frequently use DeepL and Chat GPT for school and in their private life. Therefore, it was decided to inform the students about the study to make them aware that if they used those tools during the teacher feedback intervention, they could unintentionally falsify the results. If, for example, they decided to also use AI to get feedback alongside teacher feedback, the results of this intervention period could no longer be attributed to the teacher and would therefore distort the results. Also, the students were assured that the results of their writings and the contents of their journal would not influence the assessment of their course.

As the students were fully informed and they worked independently for most of the time, some exceptions from the fixed procedures were made. One student (EG2) could not attend the first day of AI intervention and informed the teacher ahead of the intervention. So, all material and information were provided via email on the day of instruction so that they could do the coursework at home, ahead of the second day. Also, two students were ill on the day of the first pre-test and were granted permission to sit the test one day later.

3.4.3 Journals

During the eight weeks of intervention, the students were also asked to write journals about their learning process (see Appendix C). By letting students write journals, the researcher is provided with a genuine insight into the learning process as well as the thoughts of the students which could not be elicited by other methods (Dörnyei, 2007), for example, how they use AI to receive feedback, how they engage with it, and how they react to different kinds of feedback. Further, since the questionnaire is not distributed until the end of the study, 14 weeks after its start, by using journals it can be ensured that impressions are reported while they are still fresh and that they are not retrospectively biased (Carson & Longhini, 2002; Oxford et al., 1996).

The students were provided with following instructions:

Please note the following points in your journal:

- What tools did I use for feedback? (if AI)*
- What points did I get feedback on?*
- How did this feedback help me?*
- What was my reaction to the feedback?*
- What did I do with the feedback?*

Journal entries can be made in German or in English.

The students were reminded regularly that journal entries could be written in both English and/or German to ensure that they could fully express themselves and were not restricted by language barriers. Also, as part of the journal process, they were asked to document their prompting process when using AI with prompts of their own.

3.4.4 Post-Intervention Questionnaire

Immediately after the second post-test, the students were asked to complete a short questionnaire (see Appendix D). First, they had to rate the perceived usefulness of teacher and AI intervention on their learning on a five-point Lickert scale. Then they were asked whether in the future they would prefer: teacher, AI or a mixed intervention. Lastly, they were asked to answer four open-ended questions on the advantages and disadvantages of teacher/AI intervention.

3.5 Materials

For the pre- and post-test, writing exercises from a Cambridge First Exam (FCE) preparation book were used (Bell & Thomas, 2018). In the case of intervention stage One, two FCE essay writing tests were used and intervention stage Two used FCE article writing tests (see Appendix E). For the interventions, exercises from Cambridge's *'Write & Improve'* website were used (Cambridge University Press & Assessment, 2022). There were two reasons for choosing these kinds of writings. First, most of the students went on to do an FCE within some months of the intervention and were asked to do those exact kinds of tasks. Therefore, there was a clear goal for the students as well as explicit marking criteria as suggested by Bruton (2009). Second, the FCE writing tasks require students to write between 160-190 words, and research has found that feedback on writings of under 200 words was more effective than feedback on longer writings (Brown et al., 2023). Also, it was decided to let the students create new pieces of writings every week rather than letting them revise the previous texts. The reason for this decision was that it is questionable whether the ability to revise has a long-lasting effect on students' writing abilities. Therefore, Storch's (2005) suggestions were followed to provide students with unique tasks for every lesson and also for the pre- and post- tests.

The following AI-tools were prescribed to the students for receiving AI generated feedback during the AI feedback intervention stage: all the features of Microsoft Office 365, ChatGPT 4o, and DeepL Write. Word in Office 365 not only has an integrated spell check, but it also helps to locate and correct grammatical mistakes. ChatGPT4o can give detailed feedback according to the FCE marking grid (see Appendix B). When asking for feedback, the students could directly upload the provided marking grid onto ChatGPT and ask it to mark it according to the grid. So, the students received a mark and feedback for all the four main marking criteria of an FCE-exam: content, communicative achievement, organisation, and language. Furthermore, ChatGPT is also able to provide the student with suggestions on how to re-write certain sentences,

paragraphs, or even the whole text. Further, it can suggest exercises to work on certain areas. DeepL Write is a platform that rewrites texts and puts them into the correct grammatical form. It can further suggest the substitution of words to improve a text. Additionally, it is possible to choose or change the style of a text. During the teacher intervention, students were not allowed to use any of these tools and were even asked to switch off the integrated spell check in Office 365. So, they were supposed to simply rely on, and learn from, the teacher's feedback and corrections. However, since all the writings were done on computers to keep it life-like, it was not possible to monitor whether the students adhered to this rule. The assumption that they did is purely based on trust.

The choice to prescribe these tools to the students was made based on the tools the students already frequently use in their everyday lives. Also, a limited list of tools helped to keep a magnitude of control over the progress and to ensure that the research can be duplicated later on. To help the students with prompting, sample prompts were provided (see Appendix A). However, some students decided to re-prompt to get clarification on certain points or use their own prompts from the beginning. Even though they were asked to provide their prompting metadata in the journals, not everybody adhered to this practise. To ensure that this study could be duplicated in different societies without additional financial burdens, only free tools were chosen for this study.

For the journals, students were provided with a template to use (see Appendix C). The questionnaire was put onto Microsoft forms (see Appendix D).

3.6 Coding and data analysis

All the pre- and post-tests were marked by both the researcher and ChatGPT. The writings were marked according to the official FCE marking grid which had also been provided to the students for use during the AI intervention period (see Appendix B). ChatGPT was provided with the same marking grid and a similar prompt as the one provided to the students:

My students are on a B2 level according to CEFR. They were asked to do following writing task: "TASK". They wrote: "STUDENT'S WRITING". Please mark their writing according to the provided marking grid.

All the AI marking was done on the same day. The teacher marking was spread across some months due to time constraints. Pre-test 1 was marked the week after the students took pre-test 1, and the same goes for post-test 1. Pre-test 2 and post-test 2 were marked in January in the same week as the AI marking was done. Consequently, the researcher was not blind to the test-session while marking the tests. But, because the students' writings were anonymized and did not allow any connection to the condition, the researcher was completely blind to the condition.

The marks of the pre- and post-tests as well as the answers to the questionnaire were used as the primary source of data for this study.

The quantitative results were analysed using SPSS. First, the marks given by the teacher and AI were checked for interrater reliability. Because the interrater reliability of post-test 1 was weak compared to the ones for the other three tests, a second rater was brought in to mark five of the post-test 1 essays. These essays were chosen because they showed the biggest marking divergence between AI and researcher. The second rater was an experienced FCE teacher and completely blind to the whole study. They were only provided with the students' writing and the marking grid, being told that a fellow teacher wanted a second opinion on those essays. Since the marks of the second rater were closer to those of the AI than the human rater, in four out of five times, the remainder of the analyses were done using the AI generated scores only.

AVOVA and t-test analyses were conducted over the AI generated scores to determine whether the extent of improvement in gains from pre- to post-test differed following the two interventions.

The results of the questionnaire were evaluated quantitatively (using SPSS) as well as qualitatively by the researcher. The qualitative data was analysed by the teacher/researcher in four passes, two before and two after the reorganization of the categories. First, the answers to the open-ended questions were read and then sorted according to the emerging topics: kind of feedback, personal relationship, time management, and trustworthiness. Then, after this first analysis and evaluation of the questionnaires, the journals were studied for additional information that could clarify the findings of the questionnaire. Therefore, all the journals were read, and the relevant information was sorted into the same categories which were already used for the answers to the open-ended questions of the questionnaire. It then showed that the category 'kind of feedback' turned out to be too broad and the amount of information for this category in the combined dataset was too varied and plentiful. Consequently, it was decided to divide the category 'kind of feedback' into two categories labelled 'comprehensibility' and 'amount of feedback'. Then, the data was analysed again and coded according to the new categories. Whenever the students wrote their answers or comments in German, the texts were translated into English by the researcher.

3.8 Ethical Considerations

An ethics approval (CUREC) from the University of Oxford was obtained before the start of the study (see Appendix F). The day before the intervention started, informed consent was obtained from all the students (see Appendices G & H). They also had the option to opt out of the study at any time of the intervention without consequences, since they still received the same teaching whether they took part in the study or not.

Chapter 4: Results

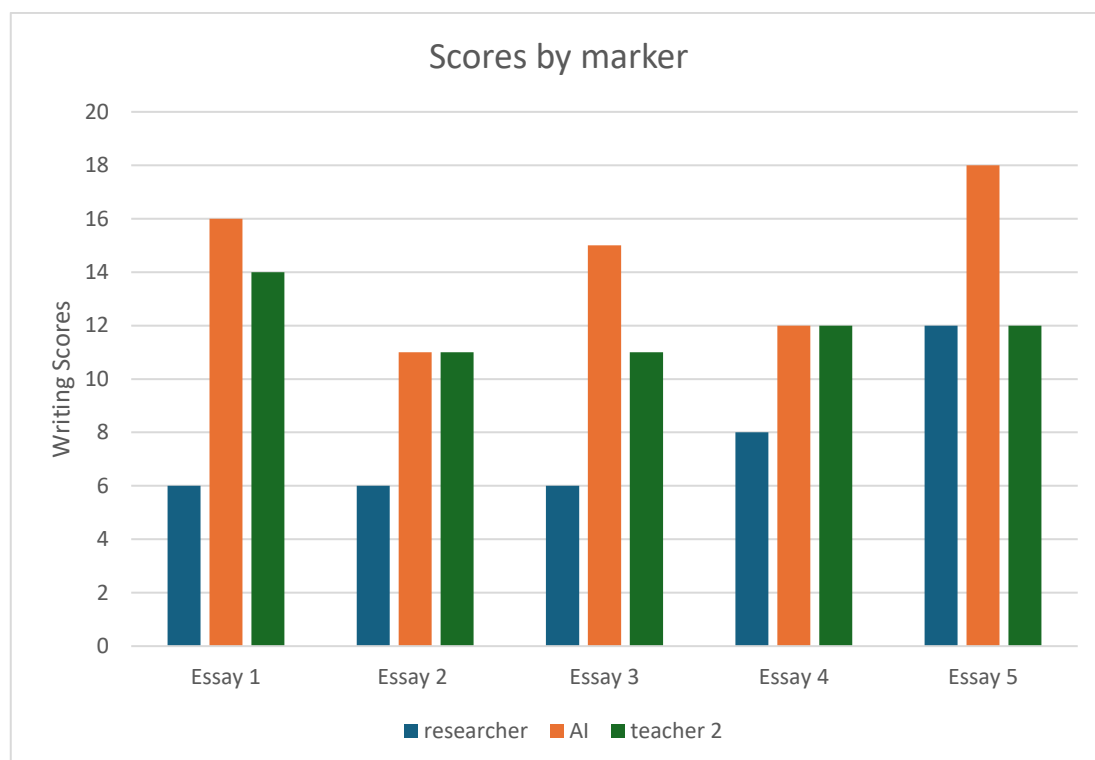
The results of the study were analysed quantitatively as well as qualitatively. First, the markings done by AI and the teacher markings were checked for interrater reliability. Then, the results from the pre- and post-tests were analysed to see if there was an effect of intervention or group on the gain scores and whether this effect was due to chance. Later, the answers to the quantitative part of the questionnaire entries were examined to see how the students reacted to the two sources of feedback and whether this attitude had an influence on the gain scores of the respective intervention. Then, the qualitative part of the questionnaire and the journal entries were examined to see where the students saw the benefits and limitations of the respective feedback source.

4.1 Relationship between teacher and AI marking

All pre- and post-tests were marked by both the researcher/teacher and AI. The interrater reliability was calculated to compute the degree of agreement between the marks given by AI and the human. This is important to ensure that the marking of the test was objective and not biased for two reasons. On the one hand, it is not yet fully established to which extent AI marks can be relied on in this context, and on the other hand, contrary to AI, it was not possible for the teacher to mark the test fully blind to whether it was pre-/post-test. The intraclass correlation coefficient was calculated to clarify how much AI and the teacher agree with each other. Pre-test 1 ($ICC(2,1) = .604; p < .001$) and post-test 2 ($ICC(2,1) = .514; p < .001$) show moderate and significant reliability criteria (Koo & Li, 2016). The reliability for pre-test 2 was poor ($ICC(2,1) = .345; p = .001$) while post-test 1 was neither reliable nor significant ($ICC(2,1) = .094; p = .336$). To further understand the different results, it was decided to focus on post-test 1, since it showed the lowest interrater reliability. To get a third opinion on the marks, a further teacher was brought in and asked to mark five of the post-test 1 essays. Although statistical analyses for such a small sample are not appropriate, it can be seen in Figure 1 that for four of the five essays, the marks given by the second teacher were closer to the AI marks than to the researchers' marks.

Figure 1

Comparisons of marks given by the researcher, AI, and a second teacher



In addition, to further probe the locus of the difference between the original marker and the AI, interrater reliability was calculated for the separate components according to the marking criteria. The results for content (ICC (2,1) = -.180; $\rho = .789$), communicative achievement (ICC (2,1) = -.166; $\rho = .758$), and organization (ICC (2,1) = -.119; $\rho = .698$) were all negative and non-significant, indicating extremely poor reliability. However, the criteria for language (ICC (2,1) = .418; $\rho = .022$) showed a poor but significant intraclass correlation (Koo & Li, 2016). Possible reasons for the discrepancy between the teacher and AI on these particular criteria in this specific test are discussed in the next chapter.

For the following analysis, only the marks given by AI will be included. The reason for this course of action is that the correlation between human marking and AI marking was moderate and significant and therefore the marking of AI can be regarded as reliable. However, there seems to be some evidence that the teacher marks for post-test 1 were unintentionally biased. Therefore, the teacher marks were not included in the analysis.

4.2 Influence of type of intervention on gain score

After it was determined which marks to use, the gain scores for each participant were calculated. Therefore, for both interventions, the score of the pre-test was deducted from the score of the post-test for each student. The scores were then also coded with the group number, so that it could be determined at which stage the respective student received the AI intervention

or the teacher intervention, and from which intervention the gain scores were derived. Next, a one-way ANOVA with repeated measures was conducted, to determine if one kind of intervention led to a higher gain score than the other. The kind of feedback intervention (teacher or AI) was used as the within-subject factor, while group was used as the between-subject factor. This was done because the two groups received the teacher and AI intervention respectively at two different points of the study and it was important to see if this difference also led to different results.

For the teacher intervention (Figure 2), it was found that Group 1 improved slightly (mean = .08, SD = 2.565) while Group 2 scored slightly lower (mean = -.13, SD = 1.959). For AI intervention (Figure 3), the results were reversed and somewhat more prominent. While Group 1 scored lower (mean = -1.54, SD = 2.402), Group 2 improved (mean = 2.38, SD = 3.114). However, the large standard deviation indicates that the sample size was too small to get a meaningful result. The main effect of the intervention was not significant ($F = 0.436$, $p = .517$, $\eta^2 = .022$), which means that there is no evidence that the results were caused by the intervention. The groups might have differed by chance.

Figure 2

Mean scores for pre- and post-test for teacher intervention

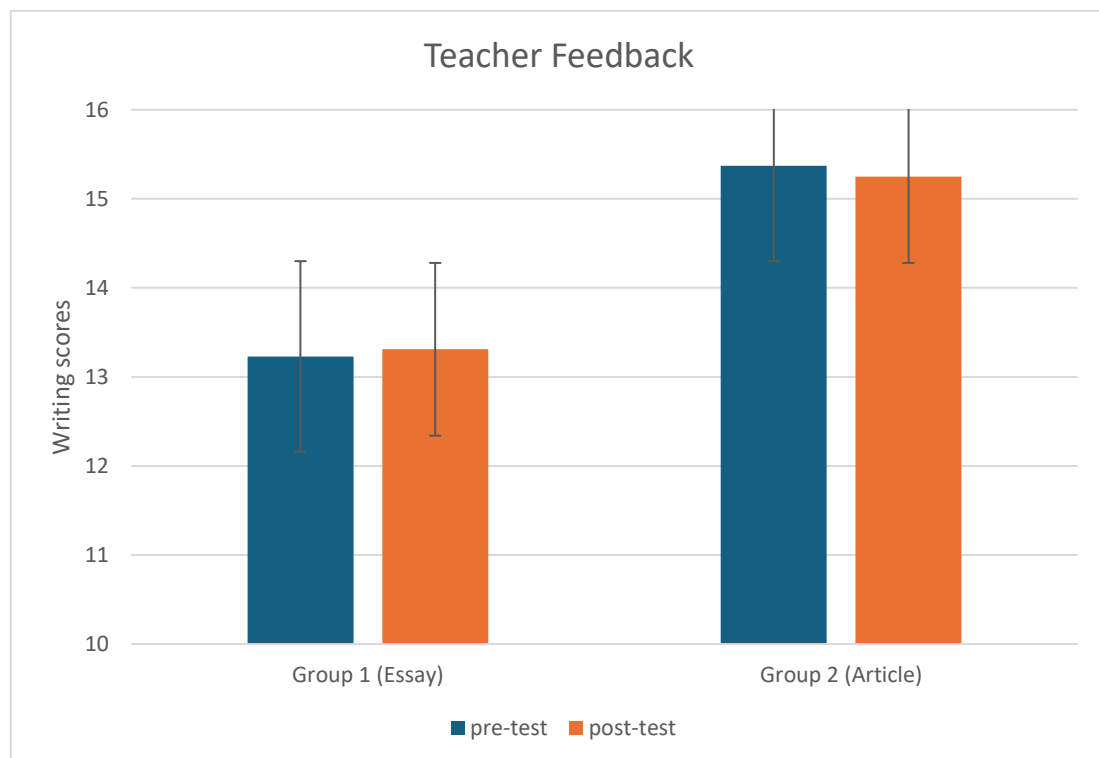
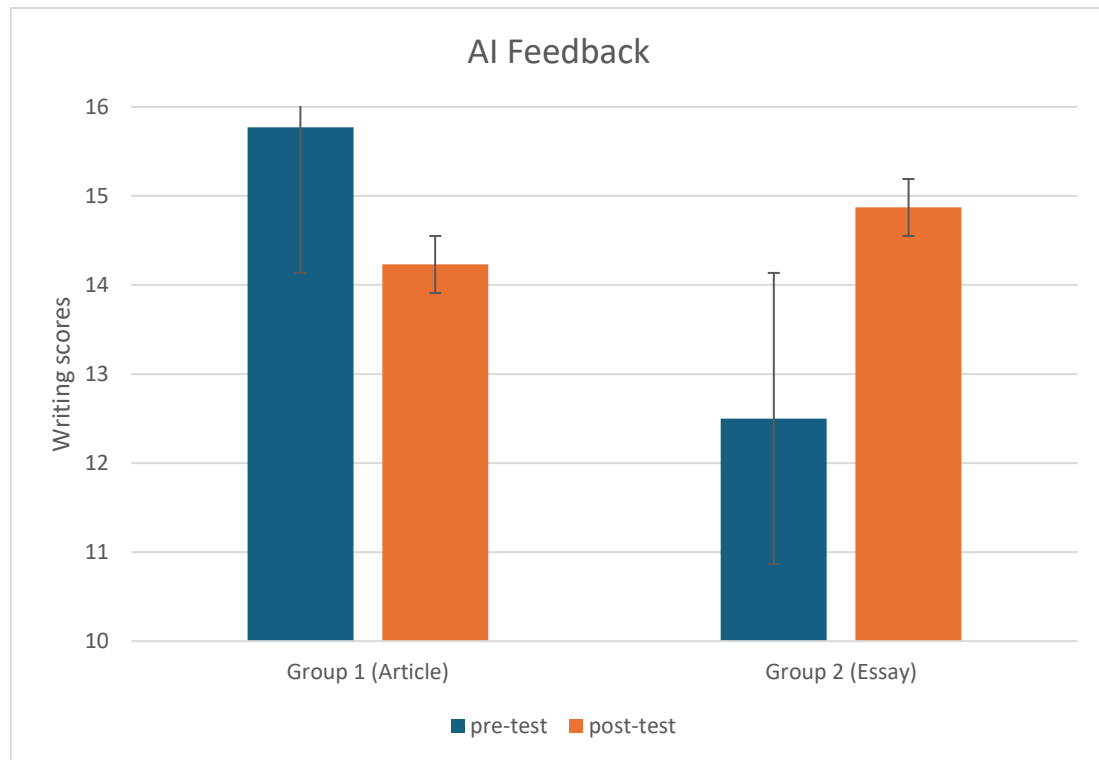


Figure 3

Mean scores for pre- and post-test for AI intervention



However, the interaction between group and intervention was significant with a large effect size ($F = 9.425$, $p = .006$, $\eta^2 = .332$). This means that one group improved more from one intervention than the other. But still, this does not indicate that one intervention has more effect than the other, rather it is most likely due to chance.

As a follow up, another one-way ANOVA with repeated measures was conducted to see if there was a detectable within-subject difference. For both groups only a marginal within-subject contrast was found (G1: $F = 4.374$; $p = .058$ / G2: $F = 4.605$; $p = .069$)

However, the differences in scores according to the type of writing is noteworthy. While both groups improved in the essay writing (Figure 4), independent of the type of feedback they received, the scores of both groups decreased for the article writing (Figure 5).

Figure 4

Mean scores for essay pre- and post-test

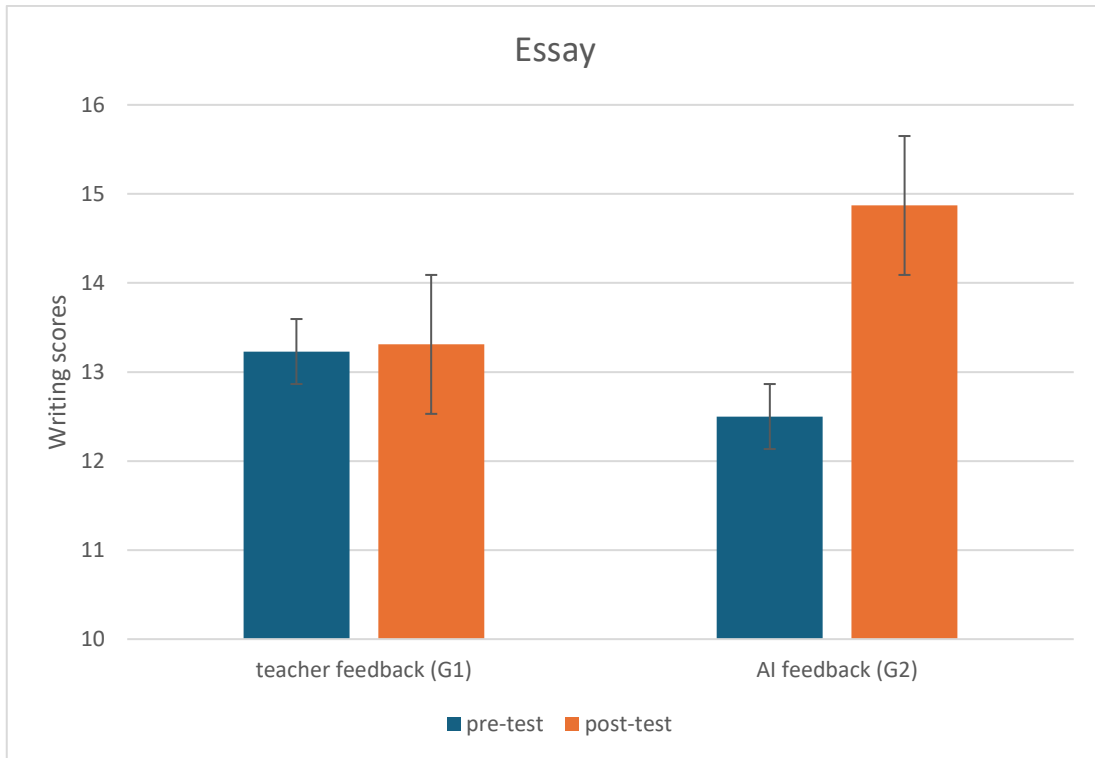
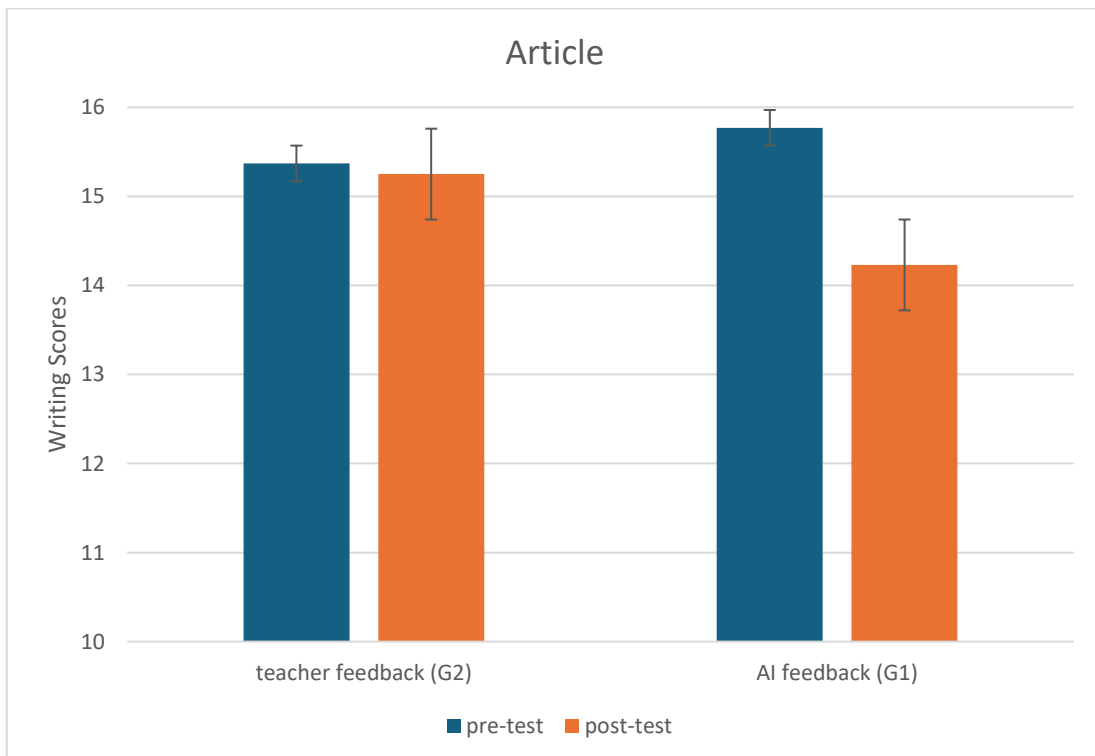


Figure 5

Mean scores for article pre- and post-test



4.3 Comparison to chance

Next, a one-sample t-test was conducted to compare the results to zero and determine whether the observed effects could be attributed to chance. The results for the teacher intervention showed no significant effect for either group (G1: mean = 0.08, SD = 2.565, $t(13) = .108$, $p = .916$, Cohen's $d = .03$; G2: mean = -0.13, SD = 1.959 $t(7) = -0.180$, $p = .862$, Cohen's $d = -0.064$), suggesting that the results could plausibly be due to random variation. In contrast, the AI intervention showed a moderate negative effect in Group 1 (mean=-1.54, SD=2.402, $t(13) = -2.309$, $p = .040$, Cohen's $d = -0.641$). Group 2 demonstrated a positive trend that approached significance (mean=2.38, SD=3.114, $t(7) = 2.157$, $p = .068$, Cohen's $d = -0.763$). While the teacher intervention and the AI intervention for Group 2 did not produce reliable effects, the results for the AI intervention for Group 1 suggest the results were unlikely due to chance and may reflect a genuine impact of the intervention.

4.4 Students' perception of the usefulness of feedback

In the questionnaire the students were asked to rate the usefulness of the interventions on a five-point-Likert scale, with 1 indicating not useful and 5 indicating very useful. All students stated that both interventions were useful. However, the usefulness of the AI intervention (Figure 6) was scored slightly lower than the teacher's intervention (Figure 7). Only two students gave the same score to teacher and AI intervention and one student ranked AI as more useful (Figure 8).

Figure 6

Perceived usefulness of AI feedback

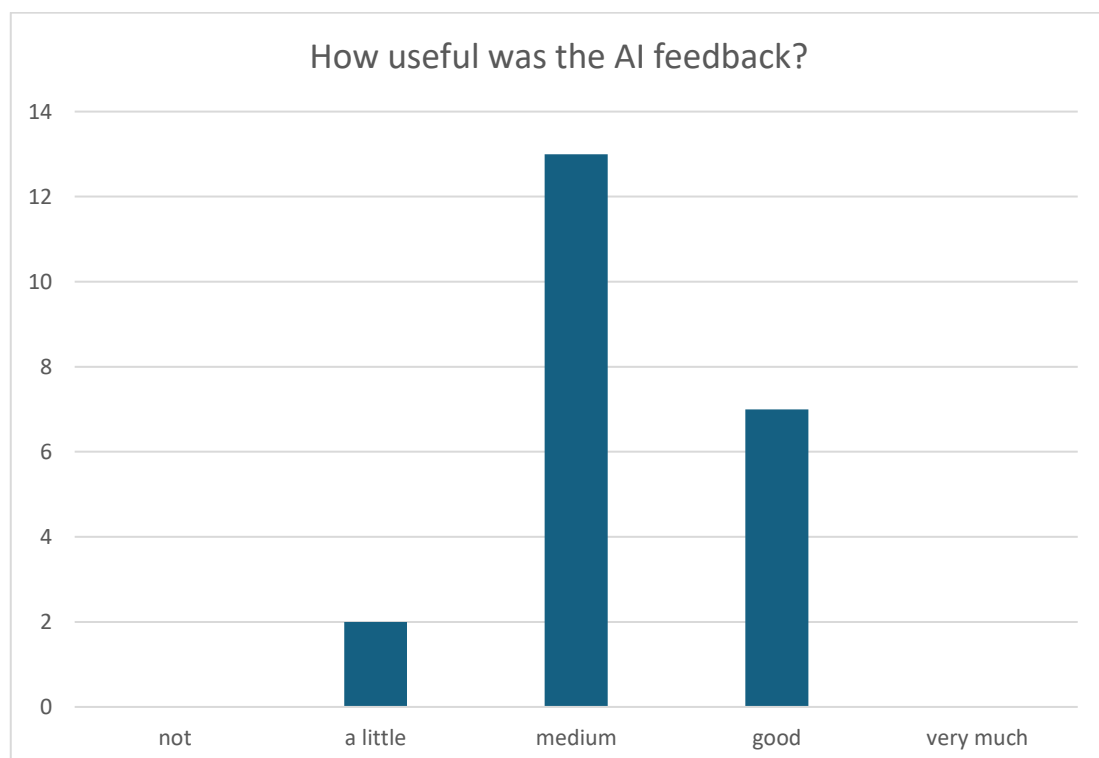


Figure 7

Perceived usefulness of teacher's feedback

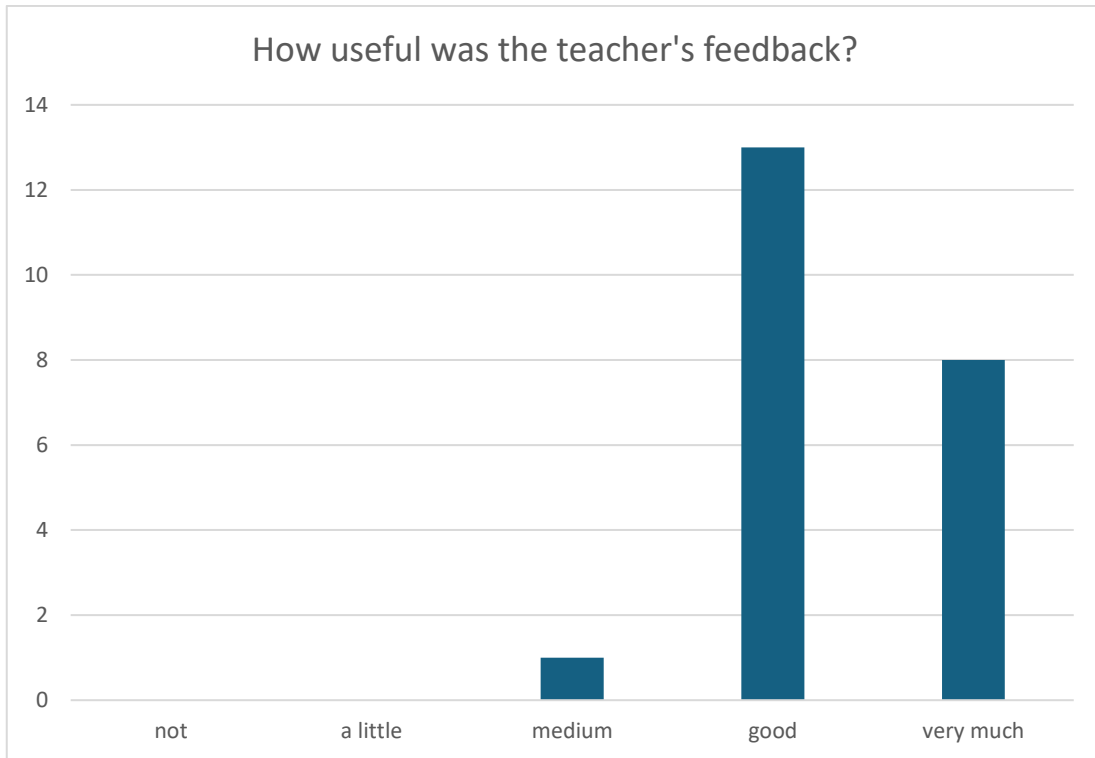
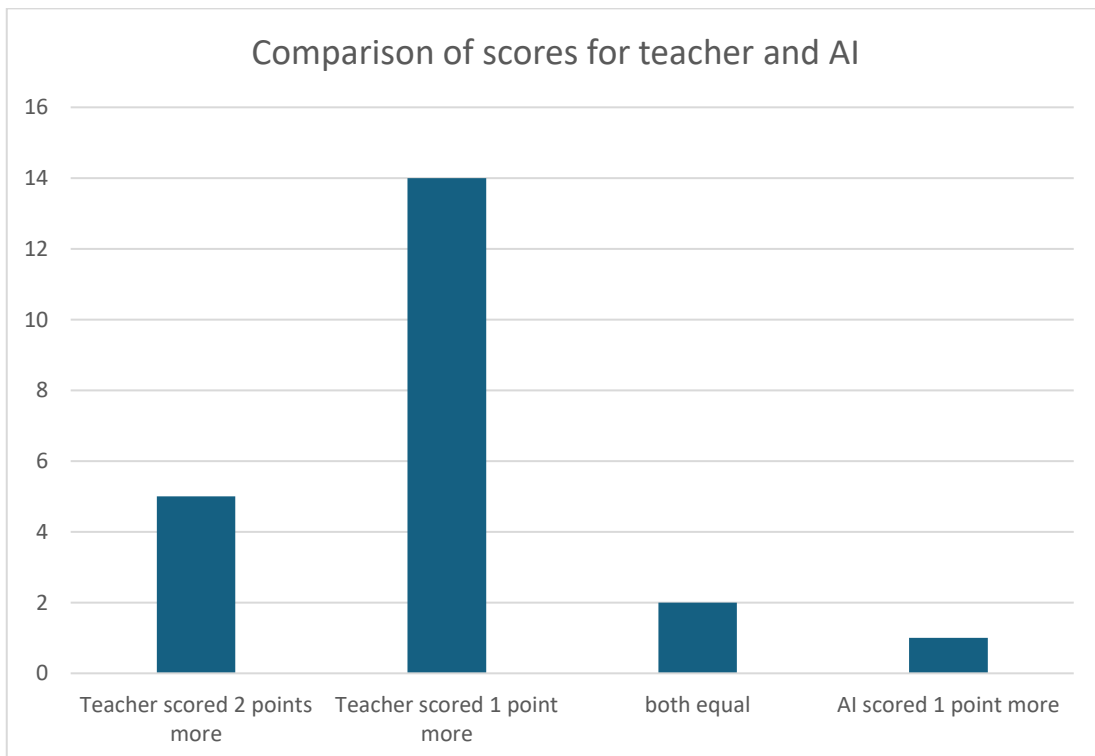


Figure 8

Comparison of the usefulness scores given to AI and the teacher



To see if the students' perception of feedback usefulness has had any influence on the outcome of the pre-/post-test scores, the correlation was calculated between the gain scores and the rating they gave the respective intervention. For the teacher intervention, there was no correlation between gain scores and perceived usefulness ($r = .000$; $p = 1.000$). For the AI intervention there was a very weak positive correlation which was not statistically significant ($r = .183$; $p = .427$). This means that no evidence could be found that there is a relationship between how useful students perceived the interventions and how they actually performed.

Further, the students were asked if in the future, they would prefer AI feedback only, teacher feedback only, or a mixture of both. No one opted for the AI feedback only. The other two options were equally weighted with 11 students stating that they would prefer teacher feedback only, while 11 would prefer a mixture of both (Figure 9).

Figure 9

Students' preferences for future feedback



4.5 Students' perception of the intervention process

In addition to those questions, the students were also asked to comment on the positive and negative points of teacher and AI feedback. In total, the students made 36 positive and 35 negative comments on the AI feedback, whereas they made 37 positive and 19 negative comments on the teacher feedback. The emerging topics were the comprehensibility of the feedback ($n = 47$), personal relationship ($n = 24$), time management ($n = 24$), amount of

feedback (n = 19), and trustworthiness (n = 13). To provide further information on those points, the journals written by the students during the intervention periods were consulted as well.

4.5.1 Comprehensibility of the feedback

Considering the comprehensibility of the feedback, there were 21 comments in favour of teacher feedback. Students liked that *“it is easier to ask a teacher what they meant with a feedback (when the student is uncertain about a topic) than ask AI again.”* Also *“The spelling mistakes were easier to track”* and *“[They] broke down the whole writing process into small bits.”* It also seems to be an advantage that a teacher often works with visuals as several students stated in their journals, e.g.: *“I also find writing things down and drawings on the board helpful when discussing.”* and *“[The teacher] worked with different colours and [they] explained it along the visuals [they] made.”*

However, six students also reported disadvantages of the teacher feedback for comprehensibility, concretely that *“I didn't get a mark or I couldn't see in which part of the evaluation I have problems or are doing well.”* Also, some students did not like the fact that the teacher gave some additional explanations in class on mistakes or issues that occurred in several of the writings. This clarification was seen as unnecessary and boring.

AI feedback received more negative comments (n=13) than positive ones (n=7) on comprehensibility. Particularly it was critiqued that *“the answers are often the same or similar if one asks for deeper explanations”* and *“Sometimes it was written to complicated for me”*. Students also stated that *“It stays very generic and sometimes doesn't understand the context”*. Still, some students found AI feedback comprehensible because *“AI gives you sometimes very good examples on how to write something in a different way”* and *“It's a good starting point for improving or reworking a text, as the AI can give you a good overview of mistakes etc”*.

Since students wrote their journal during the interventions, some of them compared the different feedback in their journals for the second intervention. One student mentioned that:

The information i got from the teacher instead of AI were kind of eye-opening. I realised that chatgpt gave me good information about writing an essay. However it forgot to mention a few very basic but important key-points. I especially noticed that, because all my essays were written in a very informal way even tho essays are supposed to be written in a formal way. AI never thought about telling me that important piece of information.

One student summed up the difference between AI and the teacher's feedback in their journal as follows:

The feedback from ChatGPT was very precise and detailed. In contrast to the teacher, I received precise suggestions on how I could improve. I think with ChatGPT the feedback is generally more related to the text and the teacher's feedback was generally about writing an essay.

This seems to indicate that the focus of AI and teacher's feedback might simply lie on different elements of writing.

4.5.2 Personal relationships

When it comes to personal relationships, there were 10 comments on the advantages of teacher feedback. Positive points raised about the teacher's feedback were that it felt more personal. One student, for example, mentioned that *"it also makes it more personal and there is a feeling of some type of moral support"*. Also, one student mentioned that they *"like the aspect of someone actually taking time to read and mark an article or essay [...]. It feels like my work that I put into writing this text is valued more"*.

AI, on the other hand, felt distant and stricter, with eight comments on the disadvantages of this kind of feedback. One student stated that *"The feeling of being marked by a computer is not as good as being marked by a real teacher"* and another wrote that *"The feedback from AI is also very strict and so you will quickly get a bad grade"*.

But, AI was perceived as more patient, with six comments on its advantages. For example, one student wrote that *"You can ask questions as much as you want"* and another that they were *"less stressed when I made the same mistake twice. With a teacher I would think that the teacher thinks bad of me because I made the same mistake twice"*.

In the journals there was very little information about the students' feelings on the received feedback that could shed light on the relationship between students and feedback giver.

However, two students mentioned that AI was very strict and always found mistakes no matter how good the writing was. Also, they missed seeing comments like 'good job'. Still, one other student mentioned that the comments from AI were always positive.

4.5.3 Time management

Time management was seen as the biggest advantage of AI with 16 positive answers. The points raised were that *"The best thing is the speed"* and that *"feedback was accessible at all times"*.

For the AI, there were only two negative points raised, for example that students *"find it exhausting to chat with AI at school"* and that they had to re-prompt to *"ask for harsher feedback [because] AI always tried to sugar coat the feedback"*.

Teacher feedback, on the other hand, received six negative comments and no positive ones. The main critique was that teachers *"only have limited time"* and it was seen as a limitation that students had to *"wait for the feedback"*.

AI can provide students with immediate feedback anytime and anywhere. One student reflected on the time aspect in their journal as follows:

I get comprehensive feedback from ChatGTP within seconds and also specific suggestions for improvement that I can implement directly. After an initial revision, I ask the AI for further feedback, where it directly compares the feedback from the first version. Within about 20 minutes, I have done this 3 times until the AI has made it clear to me that my work is now significantly better.

This kind of feedback would not be manageable for teachers in their limited time and with more than one student.

4.5.4 Amount of feedback

A highly ambivalent point raised was the amount of feedback. While there were seven positive comments on how extensive AI feedback was, e.g.: *“there were suggestions for improvement for every marking criterion”* or *“very detailed feedback with a lot of suggestions to improve”*, there were also four points that specifically critiqued the vastness of feedback: *“it was sometimes a bit too much and I didn’t really want to read all of it”* and *“there was often a lot of text that didn’t say too much”*.

Regarding the teacher feedback, the limited feedback was critiqued by seven students with remarks such as,

It was not complete. Only the important bigger things were talked about” or *“There weren’t many examples, for example how it could have been written better. It was only corrected what was wrong and how it could have been improved. Not different examples for inspiration.*

However, one student saw the short teacher feedback as an advantage, stating that *“It was more direct and not a long text which could be a lot shorter and still give me the same feedback”*.

The same ambivalence could be found in the journals. While some students were happy about the vastness of feedback, others were overwhelmed by it.

4.5.5 Trustworthiness

The last point raised, with five positive remarks on teacher feedback and eight negative ones on AI, was that they do not completely trust AI. For example, a student mentioned that *“the AI feedback also changed sometimes for the same question or task”* and that *“I was never sure if it was true what AI told me”*.

One student clarified this statement in their journal by describing a chat with AI: *“[...] i asked ChatGPT to improve the text, so that it is worth 20/20 points. Then i asked it to grade its own text. [...] after creating the «perfect» essay, it graded the own essay with 18/20 points.”*

The teacher, on the other hand, was seen as more reliable. One student stated that they

[...] trust a teacher more than AI. So, when I got the text back from the teacher, I felt like I had a good feedback that I could trust. Whereas, when I got the feedback from AI, I wasn't sure if it was a 100% to be trusted.

Another wrote that they “*knew that the teacher would not tell lies*”.

In the journals, students described several instances where ChatGPT miscounted words or changed marks when uploading the same text twice.

All in all, there seems to be agreement that both kinds of interventions have their advantages and disadvantages. Therefore, as indicated by the question of what feedback they would like to receive in the future, a mixture of both seems to be favoured by many of the students, as one summed up in their journal:

Feedback from a specialist is essential for my personal assessment of my skills, as they can respond to me directly and personally and can therefore also provide grammatical input. The AI's feedback is particularly helpful when it's just a matter of having your text corrected grammatically and structurally. The crucial point is above all the time: I can give a text for correction countless times until I am satisfied, and I get the result immediately every time.

Chapter 5: Discussion

The aim of this study was to investigate if students improve more following teacher or AI feedback. It also examined how students react to the different kinds of feedback. Contrary to some other studies, no significant correlation between the source of feedback and the improvement in writing was detected. The AI's feedback was generally well received, although the teacher's feedback was regarded as marginally more helpful. Half of the participants indicated that in the future, they would prefer to receive only teacher feedback while the other half said they would like to receive combined teacher and AI feedback.

The following discussion will first look at the reliability of teacher and AI marking and consider the surprising result that the interrater reliability was low and non-significant for post-test 1 even though it was significant and moderate for pre-test 1 and post-test 2. Then, the results of students' improvement will be addressed and linked with studies conducted previously, especially since the current study could not find any significant results compared to former studies. Therefore, the limitations of this study are considered, and it will be discussed where further studies could lead. Lastly, the results of the qualitative data will be discussed and linked to findings of previous studies.

5.1 Interrater Reliability

Even though the interrater reliability was moderate and significant for pre-test 1 and post-test 2, it was poor for pre-test 2 and even lower for post-test 1. When focussing on post-test 1, it was found that for three out of four marking criteria, the intraclass correlation coefficient was even negative. Thus, an attempt was made to find some explanation for this divergence.

In the course of this research, the students completed four writing tasks that were included in the analysis. Pre-test 1 was completed at the beginning of the study, and post-test 1 after the first intervention, i.e., five weeks after the pre-test. Another two weeks later, the second pre-test was written, followed by post-test 2, five weeks later after the second intervention. All four tests were marked by both AI and the researcher; who is also the teacher of the participants. To determine whether the marks given by AI and the researcher were reliable, an interrater reliability was calculated. This interrater reliability was moderate for pre-test 1 and the second post-test suggesting that the teacher and AI marked according to similar standards (Stevenson & Phakiti, 2014). However, the reliability for the first post-test and the second pre-test were poor and non-significant. For an analysis of this divergence, it was decided to focus on the lowest results and thus on post-test 1. On reflection, one possible explanation for this non-existing correlation in post-test 1 could be that the teacher biased the results unintentionally. Because the researcher knew that they would face some time restrictions in the months following intervention 1, they decided to do as much work as possible ahead of the start of intervention 2.

So, pre-test 1 and post-test 1 were marked immediately after they had been taken. Thus, it was not possible for the researcher to be blind to the kind of test they were marking. Also, they were still aware of what they just had taught to the students, even though they could not know which students they had taught themselves and who had been in the AI intervention. Pre- and post-test 2, on the other hand, were marked on the same day, three months after the intervention. By marking pre- and post-tests together so long after the intervention, it was less obvious for the teacher which was the post- and which was the pre- test, therefore the influence on the marks was not as prominent. Another reason for the assumption that the teacher unintentionally biased the results is that the difference from teacher to AI mark was not in favour of the students. It could be seen that the teacher was being harder on the students when marking content that had been taught in previous lessons. Evidence for this assumption can be found in the analysis of the reliability of individual components. The interrater reliability for language was still high and significant in post-test 1, whereas the reliability scores for content, communicative achievement, and organisation were negative and non-significant. While language is something that was taught implicitly in the course of the intervention, the topics of content, communicative achievement, and organisation were explicitly addressed by the teacher. Therefore, it can be assumed that, when marking post-test 1, the teacher was aware of the nature of the piece of writing they were currently marking and might have been harder on the students knowing that they explicitly taught those topics during the intervention. To test this assumption, another experienced teacher was asked to mark five of the essays of post-test 1. He was asked to mark the essays blind to any information about what had been taught or about the nature of the study. While statistical analysis is not appropriate for such a small number, the marks were numerically closer to the AI marks than to the researcher's marks, supporting the assumption that the researcher of this study might have unintentionally biased the marks given. This seems consistent with other findings that teachers are often not consistent in their marking (Cohen & Robbins, 1976; Zamel, 1985), and that sometimes teachers can lose patience with student's mistakes and therefore become stricter in marking (Ferris, 1999; Truscott, 1996). This can then influence the results of research since it is important that markers are reliable and therefore ideally blind to the aspects of the intervention. However, this is, as was seen in this study, hard to achieve in a real classroom situation. Therefore, it was decided to only use the AI marks for all following analyses, on the grounds that the AI was the fully blind marker.

5.2 Students' improvement

While Chang et al. (2021) found that students improve significantly more by being taught by a grammar programme (Grammarly) alongside a teacher, compared to a teacher-only intervention, and Ghafouri et al. (2024) discovered a significant difference between the group being taught using generative AI (ChatGPT) compared to the group receiving the classical form of instruction, this study could not find any significant difference between the two groups in

relation to the kind of feedback. Moreover, not only could no significant difference between the two different interventions be found, but there was also no evidence that students benefit from either teacher or AI feedback. It could be argued that the study was too small to generate any meaningful results (Cohen et al., 2017), given the small sample size ($n = 21$) compared to Chang et al.'s (2021) 53 participants and Ghafouri et al.'s (2024) 48. In addition, it was also shorter than the aforementioned studies. The two intervention periods of this study lasted eight weeks, with six pieces of writing compared to one intervention of 16 weeks with five essays (Chang et al., 2021) and one intervention of 10 weeks with several writing tasks (Ghafouri et al., 2024). What was unexpected though, was that both groups showed improvement on the essay task and a decline in marks on the article writing, regardless of whether this followed AI feedback or teacher feedback. It is unclear if it was the relative difficulty of the two tasks, essay compared to article, or the fact that both groups did the essay first and the article second, that led to these results. However, there is some evidence supporting the latter, as one student mentioned that they “[...] *did not revise the text. Because I have written enough texts by now.*” This may have been caused by to the length of the intervention, with the possibility that students got bored of the intervention and did not put as much effort into the second intervention (article) as they did in the first intervention (essay). This could be counteracted in a future study by writing the same number of essays/articles over a longer period. Chang et al. (2021) for example used five essays for a single intervention period of 16 weeks, while this study, due to time constraints, used a total of six writings in eight weeks for both interventions. Although there were no significant differences in gains following AI feedback and teacher feedback, numerically, there was a greater decrease in performance for the AI feedback group compared with the teacher feedback group. While it is important not to overinterpret non-significant results, tentatively, it is worth noting that this is consistent with what was found in journal entries. Students in the AI group could work by themselves and put in as much effort as they wanted. One student, for example, wrote in their journal that “*Comparing this introduction with the other one [teacher]: I’m less focused and engaged. It is way easier to get distracted.*”

In comparison, the students in the teacher intervention could not withdraw from the intervention so easily. One student of the teacher group also mentioned that “*The teacher made sure you had the right information and didn't run the risk of overlooking anything important*”.

Thus, it can be concluded that, depending on the level of engagement of individual students, the effectiveness of AI intervention can be more pronounced on both sides of the scale. While engaged, students can learn a lot with repeated prompting and using the time available, students that are not as engaged can easily evade the intervention and therefore disrupt their own progress. This confirms the notion that feedback is only useful if students do pay attention to it (Chandler, 2009). During the teacher intervention, they cannot ignore the input as easily, which makes the progress of the students much more homogeneous but might also hinder a faster and

greater progress of strong and motivated students. However, this seems more of an implementation detail and does not necessarily answer whether students learn more from the teacher or from AI. Also, it might explain why former studies (Chang et al., 2021; Ghafouri et al., 2024) found a significant effect while this study did not. While in former studies the prompting was done by the teacher and the students then received the AI feedback from teachers, in this study the students' interaction with AI was self-regulated. Consequently, for participants in the earlier studies it was not as easy to withdraw from feedback as it was for the participants in this study. Therefore, one should consider whether a future study should let the students prompt themselves or if the prompting should be up to the teacher. A limitation in this respect is certainly that the interaction between students and AI was not closely monitored and thus the engagement with feedback in relation to the outcome could not be analysed in depth. For a future study, it might be interesting to research into this relationship and better monitor the interaction between students and chatbot. However, researchers have to be careful that by closely monitoring the learning environment they do not make the students feel watched and therefore hinder the interaction with the chatbot (Chiu et al., 2024).

Even though the comparison to chance showed that the results for both groups were likely due to chance, the results for Group 2 on the AI intervention approached significance. Therefore, it would be interesting to pursue this research further with larger sample groups (Cohen et al., 2017; Ellis et al., 2008) and a longer intervention period (Chang et al., 2021; Ghafouri et al., 2024) to see if a significance can be found. Further, the results of the intervention might also depend on the proficiency level of the students (Shen et al., 2023). Lee (2008), for example, found that students with a higher proficiency tended to be more interested in error correction, which in turn could lead to a higher uptake of the feedback. Further, language proficiency could influence students' cognitive and behavioural engagement and therefore also negatively impact their affective engagement (Zheng & Yu, 2018). Since the participants of this study were all of the same proficiency level, it might be worth reproducing this study with students of different proficiency levels to determine whether proficiency has any impact on the outcomes. Also, it would be worth monitoring the students' prompting more closely during the AI intervention. This study provided the students with some example prompts, but they were also allowed to use their own prompts. Even though they were asked to report their prompts, very few of them did. Moreover, those who did, only provided the gist of their prompt but not the exact wording and also, they omitted the exact response of AI and only reported the nature of the response. Therefore, it cannot be determined how effective the students' own prompts were and how far they influenced the outcome. Consequently, allowing the students to self-prompt could be seen as a confounding variable in this study.

5.3 Students' opinion

The results of the questions considering the usefulness of feedback seems to confirm the findings of previous studies that students prefer teachers' feedback in comparison with AI feedback (Link et al., 2022). Further, for future interventions, all students opted to have teacher feedback as part of the feedback process, either with or without additional AI feedback. This, and the comments to the open-ended questions, indicate that affective factors indeed influence the way students interact with feedback (Storch & Wigglesworth, 2010; Wilson & Czik, 2016). On the other hand, contrary to the assumption of some researchers (Ferris, 2004; Storch, 2010), there was no evidence for a correlation between the perceived usefulness and the actual learning gains of the students. So, even if they prefer teacher feedback, this study could not find evidence that teacher feedback is more useful than AI feedback. In fact, there was no evidence that either of the interventions were useful. Thus, preferring one kind of feedback or the other still does not imply that the feedback also has a large effect. Furthermore, since it could be seen in the inter-rater reliability, the teacher is not necessarily the most objective even though the students trust them more than AI. What is more, since the teacher of the participants and the researcher were the same person, it cannot be said if the opinions uttered were completely honest. Even though the study was completely anonymised and the students were informed of this, it is still possible that some of the students provided overly favourable feedback to please the teacher. Thus, this finding certainly needs to be confirmed by studies conducted with different teachers and independent researchers to provide solid evidence.

Despite students preferring teacher feedback, there are also qualitative findings in favour of AI. One point from previous studies that seems to have been verified by this study, is that students perceive chatbots as less intimidating than teachers (Rad et al., 2023). The students in the present study reported that they were less inhibited to ask questions when using AI, especially if they had to ask the same question more than once to clarify feedback or get additional information. They reported that when they had to ask a teacher to explain a point twice, they were afraid that the teacher would think badly of them, but that with AI they were not afraid of being judged. This statement also validates findings that students re-prompted AI to get more information but did not ask the teacher for clarification if the feedback was somewhat unclear. Also, students in this study said that they liked AI feedback because it was fast and provided lots of information with suggestions for improvement. This confirms Lasagabaster and Sierra's (2005) finding that students preferred feedback when there was more time to process it and when there were more resources for improvement. However, the same point about the vast amount of feedback was also criticized by some of the students, confirming the findings of other researchers that students seem to struggle with too much feedback (Hartshorn et al., 2010; Storch, 2010). Because of this, students need to be carefully taught how to prompt to get the desired result and also how to manipulate prompts to change the amount of feedback. Not only

did the participants critique the amount of feedback, they also commented that sometimes AI feedback can be confusing, which then could lead to feedback stopping to be meaningful (Lai, 2010). Here again, teachers need to be aware of this and prepared to teach students how to carefully prompt AI to get the desired results.

When it comes to data collection, even though the students were fully informed about what was expected from the journal entries and what type of information they were supposed to include (Reis et al., 2024), not all journals were equally informative. Some students held the information to a minimum and thus, these journals held no information that could be used for this study. As a result, the sample of responses was likely biased because not all students could be represented. For further studies, there should be found a way to better monitor the students' journal writing to get more information on the level of engagement with the feedback, as well as how they used AI, and to make sure that all students will be represented in the final dataset.

In conclusion, it can be said that even though the quantitative results did not follow previous studies, the qualitative results concurred with the conclusions of preceding research.

Chapter 6: Conclusion

This study aimed to find out whether an effect on writing performance could be discovered, depending on whether the WCF is issued by a teacher or AI. Also, it aimed to see if there is a different perception of the usefulness of the instructions among the students and if this perceived usefulness has an influence on the performance. In fact, no evidence was found for an effect of the intervention found in this study, in contrast with the findings of some other studies. However, there was a difference in perception of feedback in the view of the students. They preferred the teacher's feedback, even though they deemed both kinds of feedback useful. As stated in the students' comments, both kinds of feedback have their advantages and disadvantages. This requires teachers to think carefully about which method to use where and when. In an ideal scenario, the use of AI in WCF can help teachers to save some time because they do not have to do all the feedback work themselves but have a competent partner at their side. After all, it was seen in the results of the interrater reliability that AI does a good job marking the students' writings and can be even more constant and reliable than teachers. Further, the feedback itself was deemed useful by students, with 19 out of 21 students stating that the usefulness was medium or good. However, even though there was no indication found that the perceived usefulness influenced the results, teachers should be careful of their students' perceptions towards both interventions, especially if students are still hesitant to use AI. Ideally, teachers could introduce AI feedback step by step to foster students' prompting abilities and to build their trust in this kind of feedback. Then, AI could be a helpful tool that aids both teachers and students; it relieves the teachers of some time-consuming tasks and provides the students with access to useful feedback at any time of the day even when teachers are not available.

Still, a lot of research in this area has to be done to shed a clearer picture on the influence of teacher and AI feedback on students' performance in writing tasks and what circumstances lead to specific results. First, studies should be conducted with more participants to make the results more reliable and significant. Then, the intervention period could be planned longer to avoid students' fatigue with the tasks. This could show if the noticeable drop in results for the second intervention in this study indeed stems from this fatigue, as hypothesised, or if it rather results from the type of writing. Also, students might need more time to get used to a type of feedback to really internalise and generalize it. Therefore, longer intervention periods with multiple feedback opportunities could lead to more significant and generalizable findings. After all, in this study students only received feedback on three occasions per intervention. Additionally, the interaction of students with the feedback and its relationship with the outcome could be an interesting area of investigation, since the results for the second intervention period and the students' corresponding comments indicate a relationship between the level of interaction and the actual outcome of the writings. Also, conducting the study with different teachers and independent researchers could prevent the students from giving overtly positive feedback to

please the teacher. By replicating this study with different teachers and independent researchers, it could be seen if the results of this study were due to the researcher being the teacher or if the results stay the same with different teachers. It would also be interesting to see how the perceived usefulness of and the interaction with AI changes in the coming years, when one considers the pervasive nature of AI in all areas of life. It can be assumed that the use of AI will become second nature for many of our future students and therefore the perceived usefulness and the interaction with AI is almost certain to change. Thus, the relationship between the perceived usefulness and the actual improvement would be an interesting area for further studies. Finding out how far perceived usefulness and the actual improvement are related to each other and whether this changes with the increased use of AI could provide further insights for education. Likewise, since this study was conducted with a homogenous group of students from the same course, with the same native language and from the same age group, representing only a small sample of EFL students, it would be interesting to find out to what extent other factors such as the students' level of proficiency, their age, or the level of education influence the results. Lastly, since AI programmes are ever learning and evolving, the results of this kind of research could be changing depending on the improvements in technology. It can be assumed that AI will become better with time and that students will become more familiar with using AI. So, it will be important to routinely return to this question and evaluate again.

References

- Abdollahzadeh, M., Golshan, M., & Allami, H. (2022). Using a Messenger Bot as a Tool for Providing Written Corrective Feedback: Examining L2 Development and Learners' Attitudes. *Research in English Language Pedagogy (RELP)*, 10(4), 599–624.
<https://doi.org/10.30486/relp.2022.1946281.1327>
- Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, 36(4), 584–607. <https://doi.org/10.1080/09588221.2021.1936071>
- Bell, J., & Thomas, A. (2018). *Gold B2 First New Edition Coursebook*. Pearson Education.
- Brown, D., Liu, Q., & Norouzian, R. (2023). Effectiveness of written corrective feedback in developing L2 accuracy: A Bayesian meta-analysis. *Language Teaching Research: LTR*.
<https://doi.org/10.1177/13621688221147374>
- Bruton, A. (2009). Improving accuracy is not the only reason for writing, and even if it were.... *System*, 37(4), 600–613. <https://doi.org/10.1016/j.system.2009.09.005>
- Cambridge University Press & Assessment. (2022). *Write & Improve with Cambridge*.
<https://writeandimprove.com/>
- Carson, J. G., & Longhini, A. (2002). Focusing on Learning Styles and Strategies: A Diary Study in an Immersion Setting. *Language Learning*, 52(2), 401–438.
<https://doi.org/10.1111/0023-8333.00188>
- Chandler, J. (2009). Response to Truscott. *Journal of Second Language Writing*, 18(1), 57–58.
<https://doi.org/10.1016/j.jslw.2008.09.002>
- Chang, T.-S., Li, Y., Huang, H.-W., & Whitfield, B. (2021). Exploring EFL Students' Writing Performance and Their Acceptance of AI-based Automated Writing Feedback. *Proceedings of the 2021 2nd International Conference on Education Development and Studies*, 31–35. <https://doi.org/10.1145/3459043.3459065>

- Chen, J., & B.D., P. (2023). Integration of Big Data and Artificial Intelligence in Constructing Learners' Individualized Feedback System. *2023 International Conference on Data Science and Network Security (ICDSNS)*, 01–06.
<https://doi.org/10.1109/ICDSNS58469.2023.10245381>
- Chiu, T. K. F., Moorhouse, Benjamin Luke, Chai, Ching Sing, & and Ismailov, M. (2024). Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. *Interactive Learning Environments*, 32(7), 3240–3256.
<https://doi.org/10.1080/10494820.2023.2172044>
- Cohen, A. D., & Robbins, M. (1976). Toward Assessing Interlanguage Performance: The Relationship Between Selected Errors, Learners' Characteristics, and Learners' Explanations. *Language Learning*, 26(1), 45–66. <https://doi.org/10.1111/j.1467-1770.1976.tb00259.x>
- Cohen, L., Manion, L., & Morrison, K. (2017). *Research Methods in Education*. Taylor & Francis Group.
<http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=5103697>
- Dong, L. (2024). “Brave New World” or not?: A mixed-methods study of the relationship between second language writing learners' perceptions of ChatGPT, behaviors of using ChatGPT, and writing proficiency. *Current Psychology*, 43(21), 19481–19495.
<https://doi.org/10.1007/s12144-024-05728-9>
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. University Press.
- Ellis, R. (2010). Epilogue: A Framework for Investigating Oral and Written Corrective Feedback. *Studies in Second Language Acquisition*, 32(2), 335–349.
<https://doi.org/10.1017/S0272263109990544>
- Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System*, 36(3), 353–371. <https://doi.org/10.1016/j.system.2008.02.001>

- Fan, N. (2023). Exploring the Effects of Automated Written Corrective Feedback on EFL Students' Writing Quality: A Mixed-Methods Study. *SAGE Open*, 13(2), 21582440231181296. <https://doi.org/10.1177/21582440231181296>
- Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing*, 8(1), 1–11. [https://doi.org/10.1016/S1060-3743\(99\)80110-6](https://doi.org/10.1016/S1060-3743(99)80110-6)
- Ferris, D. (2004). The 'grammar correction' debate in L2 writing: Where are we, and where do we go from here? (And what do we do in the meantime ...?). *Journal of Second Language Writing*, 13(1), 49–62. <https://doi.org/10.1016/j.jslw.2004.04.005>
- Ferris, D., & Kurzer, K. (2019). Does error feedback help L2 writers? Latest evidence on the efficacy of written corrective feedback. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing* (2nd ed., pp. 106–124). Cambridge University Press. <https://www.cambridge.org/core/books/feedback-in-second-language-writing/does-error-feedback-help-l2-writers/C8B6E315F4DEAEFC818816D9488C985D>
- Ferris, D. R. (2010). Second Language Writing Research and Written Corrective Feedback in SLA: Intersections and Practical Applications. *Studies in Second Language Acquisition*, 32(2), 181–201. <https://doi.org/10.1017/S0272263109990490>
- Ghafoori, M., Hassaskhah, J., & Mahdavi-Zafarghandi, A. (2024). From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. *Language Teaching Research: LTR*, 13621688241239764. <https://doi.org/10.1177/13621688241239764>
- Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of Second Language Writing*, 30, 31–44. <https://doi.org/10.1016/j.jslw.2015.08.002>
- Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly*, 44(1), 84–109. <https://doi.org/10.5054/tq.2010.213781>

- Hou, X., He, S., & Cuigong, R. (2024). Learner Use of AI-Generated Feedback for Written Corrective Feedback in L2 Writing: Usefulness, User Proficiency and Attitude. *Proceedings of the 2024 8th International Conference on Education and Multimedia Technology*, 70–76. <https://doi.org/10.1145/3678726.3678767>
- Hwang, W.-Y., Nurtantyana, R., Purba, S. W. D., Hariyanti, U., Indrihapsari, Y., & Surjono, H. D. (2023). AI and Recognition Technologies to Facilitate English as Foreign Language Writing for Supporting Personalization and Contextualization in Authentic Contexts. *Journal of Educational Computing Research*, 61(5), 1008–1035. <https://doi.org/10.1177/07356331221137253>
- Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(2), 83–101. Cambridge Core. <https://doi.org/10.1017/S0261444806003399>
- Jurs, P., & Špehte, E. (2022). The Role of Feedback in the Distance Learning Process. *Journal of Teacher Education for Sustainability*, 23(2), 91–105. <https://doi.org/10.2478/jtes-2021-0019>
- Kang, E., & Han, Z. (2015). The Efficacy of Written Corrective Feedback in Improving L2 Written Accuracy: A Meta-Analysis. *The Modern Language Journal*, 99(1), 1–18. <https://doi.org/10.1111/modl.12189>
- Kassim, A., & Ng, L. L. (2014). Investigating the Efficacy of Focused and Unfocused Corrective Feedback on the Accurate Use of Prepositions in Written Work. *English Language Teaching*, 7(2), p119. <https://doi.org/10.5539/elt.v7n2p119>
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Lai, Y. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology*, 41(3), 432–454. <https://doi.org/10.1111/j.1467-8535.2009.00959.x>

- Lam, R. (2013). Two portfolio systems: EFL students' perceptions of writing ability, text improvement, and feedback. *Assessing Writing*, 18(2), 132–153.
<https://doi.org/10.1016/j.asw.2012.10.003>
- Lasagabaster, D., & Sierra, J. M. (2005). Error Correction: Students' Versus Teachers' Perceptions. *Language Awareness*, 14(2–3), 112–127.
<https://doi.org/10.1080/09658410508668828>
- Lee, I. (2008). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of Second Language Writing*, 17(3), 144–164.
<https://doi.org/10.1016/j.jslw.2007.12.001>
- Link, S., Mehrzad ,Mohaddeseh, & and Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 35(4), 605–634.
<https://doi.org/10.1080/09588221.2020.1743323>
- Miao, J., Chang, J., & Ma, L. (2023). Research Trends of Written Corrective Feedback in L2 Writing: A Bibliometric Analysis. *SAGE Open*, 13(1), 21582440221135172.
<https://doi.org/10.1177/21582440221135172>
- Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of Artificial Intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, 7(5), e07014. <https://doi.org/10.1016/j.heliyon.2021.e07014>
- Oxford, R. L., Lavine, R. Z., Felkins, G., Hollaway, M. E., & Saleh, A. (1996). Telling their stories: Language students use diaries and recollection. In R. L. Oxford (Ed.), *Language Learning Motivation: Pathways to the New Century* (pp. 19–34). University of Hawaii Press. <https://doi.org/10.1515/9780824897376-003>
- Paltridge, B., & Phakiti, A. (2015). *Research methods in applied linguistics: A practical resource* (2nd ed.). Bloomsbury Publishing.

- Rad, H. S., Alipour, R., & Jafarpour, A. (2023). Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: A case of Wordtune application. *Interactive Learning Environments*, 0(0), 1–21.
<https://doi.org/10.1080/10494820.2023.2208170>
- Reis, H. T., Sels, L., & Gable, S. L. (2024). Methods for Studying Everyday Experience in Its Natural Context. In C. M. Judd, H. T. Reis, & T. West (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (3rd ed., pp. 266–296). Cambridge University Press. <https://doi.org/10.1017/9781009170123.013>
- Rose, H., McKinley, J., & Briggs Baffoe-Djan, J. (2020). *Data collection research methods in applied linguistics*. Bloomsbury Academic.
- Schindler, L. A., Burkholder, G. J., Morad, O. A., & Marsh, C. (2017). Computer-based technology and student engagement: A critical review of the literature. *International Journal of Educational Technology in Higher Education*, 14(1), Article 1.
<https://doi.org/10.1186/s41239-017-0063-0>
- Selvi, A. F. (2019). Qualitative content analysis. In *The Routledge Handbook of Research Methods in Applied Linguistics*. Routledge.
- Sheen, Y. (2007). The Effect of Focused Written Corrective Feedback and Language Aptitude on ESL Learners' Acquisition of Articles. *TESOL Quarterly*, 41(2), 255–283.
<https://doi.org/10.2307/40264353>
- Shen, C., Shi, P., Guo, J., & Tian, J. (2023). From process to product: Writing engagement and performance of EFL learners under computer-generated feedback instruction. *Frontiers in Psychology*, 14, Article 1258286. <https://doi.org/10.3389/fpsyg.2023.1258286>
- Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14, 1260843–1260843. <https://doi.org/10.3389/fpsyg.2023.1260843>

- Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, 19, 51–65. <https://doi.org/10.1016/j.asw.2013.11.007>
- Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, 14(3), 153–173. <https://doi.org/10.1016/j.jslw.2005.05.002>
- Storch, N. (2010). Critical feedback on written corrective feedback research. *International Journal of English Studies*, 10(2), 29–46. <https://doi.org/10.6018/ijes/2010/2/119181>
- Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing. *Studies in Second Language Acquisition*, 32(2), 303–334. <https://doi.org/10.1017/S0272263109990532>
- Tai, T.-Y., & Chen, H. H.-J. (2023). The impact of Google Assistant on adolescent EFL learners' willingness to communicate. *Interactive Learning Environments*, 31(3), 1485–1502. <https://doi.org/10.1080/10494820.2020.1841801>
- Tashakkori, Abbas., & Teddlie, Charles. (2003). *Handbook of mixed methods in social & behavioral research*. SAGE Publications.
- Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46(2), 327–369. <https://doi.org/10.1111/j.1467-1770.1996.tb01238.x>
- Wang, Z. (2022). Computer-assisted EFL writing and evaluations based on artificial intelligence: A case from a college reading and writing course. *Library Hi Tech*, 40(1), 80–97. <https://doi.org/10.1108/LHT-05-2020-0113>
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157–180. <https://doi.org/10.1191/1362168806lr190oa>
- Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94–109. <https://doi.org/10.1016/j.compedu.2016.05.004>
- Wongvorachan, T., Lai, K. W., Bulut, O., Tsai, Y. S., & Chen, G. (2022). Artificial intelligence: Transforming the future of feedback in education. *Journal of Applied Testing Technology*, 23(1), 1–22.

- Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, 28(11), 13943–13967.
<https://doi.org/10.1007/s10639-023-11742-4>
- Yu, R., & Yang, L. (2021). ESL/EFL Learners' Responses to Teacher Written Feedback: Reviewing a Recent Decade of Empirical Studies. *Frontiers in Psychology*, 12, Article 735101. <https://doi.org/10.3389/fpsyg.2021.735101>
- Yu, S., Geng, F., Liu, C., & Zheng, Y. (2021). What works may hurt: The negative side of feedback in second language writing. *Journal of Second Language Writing*, 54, 100850. <https://doi.org/10.1016/j.jslw.2021.100850>
- Zahroh, R., Mujiyanto, J., & Saleh, M. (2020). Students' Attitudes toward Teachers' Written Corrective Feedback and Their Writing Skill. *English Education Journal*, 10(1), 9–17. <https://doi.org/10.15294/eej.v10i1.32511>
- Zamel, V. (1985). Responding to Student Writing. *TESOL Quarterly*, 19(1), 79–101.
<https://doi.org/10.2307/3586773>
- Zhang, Z. (Victor), & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, 36, 90–102.
<https://doi.org/10.1016/j.asw.2018.02.004>
- Zheng, Y., & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. *Assessing Writing*, 37, 13–24. <https://doi.org/10.1016/j.asw.2018.03.001>

Appendices

Appendix A

Information for students on how to use AI

How to use AI?

For this intervention study, you are allowed to use following AI tools:

- Chat GPT ([ChatGPT](#))
- DeepL Write ([DeepL Write](#))

Further you may use all the functions available on Word such as spell check, thesaurus, copilot etc.

Here some tips on how to use them:

Chat GPT:

Use it to give you feedback on how to write the text type required, e.g. “How do I write a good essay?”, “What do I have to consider when writing an essay?”. But remember to always provide information about your level and what kind of task you have to do, e.g. “I’m a student on a B2 level according to CEFR and have to write an essay in English.”, otherwise it might give you tips for an academic essay.

Use it also to give you feedback on your writing. You can directly upload the provided marking grid and use following prompt:

I’m a student on a B2 level according to CEFR. I was asked to do following writing task: “ENTER TASK (e.g. *Is it better to do things with other people or to do them on your own? write an essay of 140-190 words, mentioning work at school or college, sports activities, and an own idea*)”. I wrote: “ENTER YOUR WRITING”. Please mark my writing according to the provided marking grid.

You can then also ask Chat GPT on ways to improve your writing, how to practice it, or to give you a list of all the spelling and grammar errors it corrected.

IMPORTANT:

- Make sure you opt-out in Chat-GPT from having the data used to train AI-models. You can do this under ‘Data Controls’. (I’m happy to help you with this point.)
- Do not disclose any personal or sensitive information to Chat-GPT.

DeepL Write:

Put your written text into DeepL Write to get feedback on what areas of the text you could improve. DeepL Write will correct your grammatical structures and spelling, as well as give you ideas of alternative words or linking you could use.

Appendix B

Official FCE marking grid

The figure originally presented here cannot be made freely available via ORA because of copyright.

Appendix C

Journal template for students

Journal for students

Student: _____

Please note following points in your journal:

- *What tools did I use for the feedback? (if AI)*
- *What points did I get feedback on?*
- *How did this feedback help me?*
- *What was my reaction to the feedback?*
- *What did I do with the feedback?*

Journal entries can be made in German or in English.

feedback to:	
introduction	
essay 1	
essay 2	
essay 3	

Appendix D

Questionnaire

Questions on your experience

Wenn Sie dieses Formular übermitteln, werden Ihre Details wie Name und E-Mail-Adresse nicht automatisch erfasst, es sei denn, Sie geben es selbst an.

* Erforderlich

1. My participant number: *

Ihre Antwort eingeben

2. How useful was the feedback on your writing? *

	very much	good	medium	a little	not at all
AI	☐	☐	☐	☐	☐
Teacher	☐	☐	☐	☐	☐

3. If you could choose, what kind of feedback would you like to receive in the future? *

- ☐ AI feedback
- ☐ Teacher feedback
- ☐ a mixture of both

4. What did you like about the AI feedback? *

Ihre Antwort eingeben

5. What did you like about the teacher feedback? *

Ihre Antwort eingeben

6. What did you dislike about the AI feedback? *

Ihre Antwort eingeben

7. What did you dislike about the teacher feedback? *

Ihre Antwort eingeben

Appendix E

Writing tasks for pre- and post-tests

Pre-Test 1

In your English class you have been talking about extreme environments. Now, your English teacher has asked you to write an essay.

Do you think 'adventure tourism' in unexplored parts of the world should be encouraged?

NOTES:

Write about:

1. negative effects of tourism
2. personal risk
3. (your own idea)

Write an essay using **all** the points and give reason for your point of view.

Write 140 - 190 words.

Post-Test 1

You have been discussing in class how to be a successful person. Now, your teacher has asked you to write an essay, giving your opinion on the following statement.

"Without failure there can be no success." Do you agree?

NOTES

Write about:

1. motivation
2. experience
3. (your own idea)

Write an essay using all the points and give reason for your point of view.

Write 140 - 190 words.

Pre-Test 2

You see this advertisement in a magazine for young people.

WHAT I'D SAVE IN A FIRE

Write us an article describing something you own and explaining why it is important to you.

The best article will be published in the magazine next month.

Write your article in 140 - 190 words.

Post-Test 2

You see this advertisement on an English language website.

ARTICLES WANTED

My closest friend

Where did you meet him/her?

How long have you known him/her?

What influence has he/she had on your life?

Write us an article answering these questions. We will publish the best articles on our website.

Write your article in 140 -190 words.

Appendix F

Confirmation of Ethical Approval

**SOCIAL SCIENCES & HUMANITIES
INTERDIVISIONAL RESEARCH ETHICS COMMITTEE
DEPARTMENTAL RESEARCH ETHICS COMMITTEE**

Department of Education
15 Norham Gardens, Oxford OX2 6PY
research.management@education.ox.ac.uk

Irene Ulrich & Elizabeth Wonnacott
Department of Education, Social Sciences Division
University of Oxford



1 August 2024

Dear Irene and Elizabeth,

Research ethics approval

Research title: Teacher vs. AI: A comparison of the effectiveness of corrective written feedback provided by either a teacher or AI, and the influence of students' attitudes on the outcome

Research ethics reference: EDUC_C1A_24_169

The above application has been considered on behalf of the Education Departmental Research Ethics Committee (DREC) in accordance with the University's procedures for ethical approval of all research involving human participants.

I am pleased to confirm that, on the basis of the information provided to the DREC, ethics approval has now been granted for this study.

Please note the following:

Personal data: It is the responsibility of the PI to ensure that all personal data collected during the project is managed in accordance with the University's [guidance and legal requirements](#).

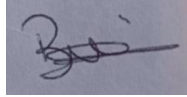
In-person activities: Any data collection involving in-person interactions with participants must have an up-to-date fieldwork risk assessment in place; further guidance is available from the Safety Office's [website](#).

Amendments: Please notify the committee if you intend to make any amendments to the information in your ethics application as submitted at date of this approval, as all changes must receive ethical approval prior to implementation. The amendment form is available on the [SSH IDREC webpage](#).

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments to research.management@education.ox.ac.uk or ethics@socsci.ox.ac.uk.

Yours sincerely

Madeleine Buil

A small, square image showing a handwritten signature in dark ink on a light-colored background. The signature is stylized and appears to be 'MB'.

DREC: Lulu Shi

cc: Elizabeth Wonnacott, research.management@education.ox.ac.uk

Appendix G

Study information sheet

This document was translated into German.

COMPARISON OF TEACHER AND AI FEEDBACK IN WRITING INSTRUCTION

PARTICIPANT INFORMATION SHEET

Central University Research Ethics Committee Approval Reference: [EDUC_C1A_24_169]

Introductory paragraph

You are being invited to take part in a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

Why is this research being conducted?

AI is an emerging topic in education. This research aims to evaluate whether AI could be a helpful tool for providing feedback on students' writing.

Why have I been invited to take part?

This study will be conducted during lessons in the classes BM1T3a and BM2TVa/BM2TT1a. Since you are a student in one of those classes, you were selected to take part in this research.

Do I have to take part?

No. It is up to you to decide whether to take part. You can withdraw yourself from the research, without giving a reason, by advising me of this decision. A withdrawal will no longer be possible after August 2025 as the data will have been fully anonymised. If you decide to withdraw your data before this deadline, all data will be deleted from our records.

If you decide not to take part, you can just inform me of your decision without giving a reason. In class, you will do the same task as the participants, with the difference, that your data will not be used for our study.

What will happen to me if I take part in the research?

- the research will take place during the usual English lessons
- before we start, you will be asked to sign a consent paper
- the research will take place from August until November 2024
- you will be asked to do writing assignments, engage with feedback, write short journal entries, and answer a few questions at the end of the study
- this research will take place in 1-2 of our weekly lessons over a total of 12 weeks you can ask to pause or stop the research activities at any time;

What are the possible disadvantages and risks in taking part?

There is no disadvantage or risk in taking part. The lessons will be delivered as usual, and you will complete writing tasks as usual, unless you are exempt from English class. The only differences are that some of your writing will receive AI feedback and in addition we will ask you to keep a journal where you reflect on the feedback and your reaction to the intervention. These journal entries as well as your scores in the writing tasks will form the data that is used in the research study. This research data will be anonymised and no one other than your teacher will be able to trace back the results to you as a person. You will also be given your own marks as usual in class.

It is important to opt-out in ChatGPT from having the data used to train AI models (available under Data Controls). Also, make sure that you do not disclose any personal or sensitive information to ChatGPT.

Are there any benefits in taking part?

For students that are otherwise exempt from English class:

You will receive intensive writing practise which will help you for your final exams or any standardised English test you choose to do.

For everyone:

It is hoped that this research will lead to a better understanding of the potential for AI to become a personal tutor for students.

What information will be collected and why is the collection of this information relevant for achieving the research objectives?

I am interested in how far teacher feedback and AI feedback help you in improving your writing. Therefore, I will collect writing samples to see how much you have improved over the period of the research. Further, I'm interested in how you engaged with the feedback you receive; therefore, I will collect samples of your journal entries. Also, I will collect some survey answers. All data will be anonymised and stored on the university's OneDrive for up to 3 years.

Only my supervisor and I will have access to the research data.

All data including scanned consent forms will be stored on the University of Oxford's secure OneDrive. This data will not include your name- you will be identified by a number. During the research we will also store a linkage file which links your name to the number but this will be deleted once data analysis is completed (latest August 2025). Other research data including consent forms will be stored for at least 3 years after publication or public release of the work of the research.

Will the research be published? Could I be identified from any publications or other research outputs?

The findings from the research will be written up in a dissertation. It will not be possible to identify you from the outputs.

I would like your permission to use direct quotations, but without identifying you, in any research outputs.

A copy of my dissertation will be deposited both in print and online in the [Oxford University Research Archive](#) where it will be publicly available to facilitate its use in future research.

It is possible that this research may later be written up and published in an academic journal or presented at a conference. In addition, I would like your permission to potentially share **fully anonymous data** on publicly available websites such as the Open Science Framework.

Data Protection

The University of Oxford is the data controller with respect to your personal data, and as such will determine how your personal data is used in the research. The University will process your personal data for the purpose of the research outlined above. Research is a task that is performed in the public interest. Further information about your rights with respect to your personal data is available from the University's Information Compliance website at <https://compliance.admin.ox.ac.uk/individual-rights>.

Who has reviewed this research?

This research has received ethics approval from a subcommittee of the University of Oxford Central University Research Ethics Committee. (Ethics reference: EDUC_C1A_24_169).

Who do I contact if I have a concern about the research or I wish to complain?

If you have a concern about any aspect of this research, please contact Irene Ulrich (irene.ulrich@education.ox.ac.uk) or Elizabeth Wonnacott (elizabeth.wonnacott@education.ox.ac.uk), and we will do our best to answer your query. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact the University of Oxford Research Governance, Ethics & Assurance (RGEA) team at rgea.complaints@admin.ox.ac.uk or on 01865 616480.

Further Information and Contact Details .

If you would like to discuss the research with someone beforehand (or if you have questions afterwards), please contact:

Irene Ulrich
Department of Education
University email: irene.ulrich@education.ox.ac.uk

Appendix H

Consent form

This document was translated into German.

Consent to take part in

COMPARISON OF TEACHER AND AI FEEDBACK IN WRITING INSTRUCTION

Central University Research Ethics Committee (CUREC) approval reference: EDUC_C1A_24_169

Purpose of Study: To understand whether AI could be a helpful tool for providing feedback on students' writing.

**Please initial
each box if you
agree with the
statement**

I confirm that I have read and understand the information sheet for the above research. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.

I understand that my participation is voluntary and that I am free to withdraw at any point [until **08/Aug/25**], without giving any reason.

I understand who will have access to personal data provided, how the data will be stored and what will happen to the data at the end of the project.

I understand that I will (not) be identifiable from any publications.

Use of quotations: Please indicate your preference (select *one* option):

- a) I do not wish to be quoted. **or**
b) I agree to the use of quotations in research outputs if I am not identifiable.

I give permission for you to contact me again to clarify information.

I understand how to raise a concern or make a complaint.

☐

I agree to take part.¹

☐

_____	<u>dd / mm / vvvv</u>	_____
Name of participant	Date	Signature

_____	<u>dd / mm / vvvv</u>	_____
Name of person taking consent	Date ²	Signature
