GPT-4 as a Homework Tutor Can Improve Student Engagement and Learning Outcomes

Alessandro Vanzo* University of Zurich

Sankalan Pal Chowdhury* ETH Zurich spalchowd@ethz.ch

Mrinmaya Sachan ETH Zurich msachan@ethz.ch

Abstract

This work contributes to the scarce empirical literature on LLM-based interactive homework in real-world educational settings and offers a practical, scalable solution to improve homework in schools. Homework is an important part of education in schools across the world, but to maximize benefit, it must be accompanied by feedback and follow-up questions. We developed a prompting strategy that enables GPT-4 to conduct interactive homework sessions for high school students learning English as a second language. Our strategy requires minimal effort in content preparation, one of the key challenges of alternatives such as home tutors or ITSs. We carried out a Randomized Controlled Trial (RCT) in four high-school classes, replacing traditional homework with GPT-4 homework sessions for the treatment group. We found that the treatment group had higher levels of satisfaction and desire to keep using the system among the students. This occurred without compromising learning outcomes, and one group even showed significantly better learning gains.

1 Introduction

Homework is an important component of education as it helps students self evaluate and develop self regulation skills (Ramdass and Zimmerman, 2011). However, in order to fully benefit from solving homework problems, it is important that students receive swift feedback on their work (Schartel, 2012), which isn't always possible for teachers due to time constraints. This leads to several students having to resort to private tutors, which can be prohibitively expensive for several households (Campani, 2013). In this work, we look at the possibility of leveraging GPT-4 (OpenAI and Team, 2024) as a tutor to assist students in their homework using a simple prompting strategy and an interface we developed.

In the seminal paper on the 2 Sigma problem, Bloom (1984) summarises the benefit different factors provide over non-interactive lectures. He claims a 0.8σ improvement for graded homework, but only a 0.3σ improvement if the homework is simply assigned without follow-up. Unfortunately, the status quo in many schools across the world, including the one we work in, roughly corresponds to the scenario where homework is never graded or, if graded, is done superficially. While the development of MOOCs with online lecture videos has extended the benefits of conventional education to larger populations (Ferguson and Sharples, 2014), most MOOCs still lack proper homework feedback or corrective instruction. Attempts have been made to try to scale the benefits of feedback and corrective instruction with the use of peer-tutoring (Cohen et al., 1982) and Intelligent Tutoring Systems (Nwana, 1990) but these too have problems with effectiveness and scaling.

Recent developments in Large Language Models (LLMs) (OpenAI and Team, 2024; Team, 2024; Touvron et al., 2023) have opened up the possibility of leveraging them for interactive homework and corrective feedback (Kasneci et al., 2023). The rapid development and adoption of LLMs like GPT-4 have provided the educational community with several new opportunities and challenges. While the benefits of GPT models in education are well studied (Baidoo-Anu and Ansah, 2023), educators are also concerned by the potential use of GPT as a tool for plagiarism in homeworks (Dehouche, 2021) and an overdependence of students on AI (Zhai et al., 2024). These fears are further exacerbated by the prevalence of hallucinations (Chelli et al., 2024) and jailbreaks (Chu et al., 2024). Studies involving real-world situations are, therefore, extremely important.

Therefore, in this paper, we summarize our observations and learnings from testing out the effects of replacing static homework with a GPT-4

*Equal Contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31119–31136

instance which has been instructed to cover the material of the homework, but in an interactive manner. Our intervention fills a gap that exists in many school systems without disrupting existing educational systems. We conduct a Randomized Control Trial (RCT) in an Italian high school to understand its effect on students, in terms of the students' experiences and learning gains. We find that students who used GPT-4 (prepared as shown in Figure 1) in this manner enjoyed the experience, while maintaining or improving their learning gains. Finally, all students who would still be in school after the conclusion of the study indicated that they would want to continue to have access to the tutor, giving us hope that despite the contemporary fears that LLMs may lead to the decline of homework in education, LLMs can also be used to make homework more fun and didactically useful to students.

2 Background

Educational research has consistently shown that personalized tutoring is one of the most effective forms of instruction. However, scaling this approach has been a significant challenge. In recent years, advances in artificial intelligence, particularly through Large Language Models (LLMs) like GPT-4, have sparked interest in their potential to provide scalable, interactive tutoring solutions. Despite early successes, the use of LLMs in education has raised important questions around their efficacy, potential risks, and best practices for implementation.

In this section, we first explore the history and limitations of scaling tutoring using traditional methods and Intelligent Tutoring Systems (ITS). We then discuss the rise of LLMs and their application in education, focusing on their strengths and potential challenges. Finally, we review recent empirical studies that have evaluated LLM-based tutoring systems, highlighting both the promise and the gaps in the current literature, which our study aims to address.

2.1 The Role of Large Language Models in Education

The advent of Large Language Models (LLMs) such as GPT-4 (OpenAI and Team, 2024) represents a potential breakthrough in addressing the issue of scalability in tutoring. Unlike traditional ITSs, which require extensive manual content creation, LLMs can generate interactive and dynamic educational content with minimal human intervention. Research suggests that LLMs like GPT-4 can mimic tutor-like behavior by engaging in natural language dialogues, providing feedback, and offering corrective instruction, all of which are key components of effective tutoring (Kasneci et al., 2023).

Despite their promise, the use of LLMs in education has sparked mixed reactions. On one hand, proponents argue that LLMs could offer an affordable and scalable tutoring solution, especially for students in under-resourced educational settings. Yet, many educators also express concern about issues like student over-reliance on AI tools, the risk of plagiarism, and the tendency of LLMs to hallucinate or provide incorrect information (Dehouche, 2021; Zhai et al., 2024; Chelli et al., 2024). Given these conflicting perspectives, empirical evidence from real-world classroom settings is critical to assess whether LLMs can truly enhance learning while mitigating potential risks.

2.2 Empirical Studies on LLMs in Feedback and Interactive Exercises

The rapid development of LLMs has prompted a lot of empirical research into their effect on education. Looking at passive surveys which try to evaluate students' attitudes and behaviours without making any interventions, we find that some works (Padiyath et al., 2024; Tanay et al., 2024; Aruleba et al., 2023; Hellas et al., 2024; Krupp et al., 2023) find that the effect is either negligible or negative, while others (Odekeye et al.; Altememy et al., 2023; S. N. Jyothy and Achuthan, 2024; Kim et al., 2024; Naamati Schneider, 2024) find students quite excited and happy for generative AI. Curiously, while most of the positive results come from places traditionally not considered "westernised", the former come entirely from the US and Europe. Of note are Cipriano and Alves (2024) and Prather et al. (2024) who note that these systems provide greater help to students already poised to succeed.

Other researchers have taken a more active role in deploying custom LLM based solutions and testing their effectiveness. Many of these have focused on programming and computer science education (Yang et al., 2024; Liu et al., 2024; Qi et al., 2024; Zamfirescu-Pereira et al., 2024; Liffiton et al., 2023; Denny et al., 2024; Sheese et al., 2024; Kazemitabaar et al., 2024; Jacobs and Jaschke, 2024) and have been tested in undergrad classes, which have found them mostly useful. Out-

side of computer science, a growing body of work has evaluated LLMs in domains like mathematics (Chowdhury et al., 2024; Butgereit et al., 2023), language learning (Polakova and Klimova, 2024; Park et al., 2024), health sciences (Chheang et al., 2024; Wang et al., 2024) and other domains (Schmucker et al., 2024; Thway et al., 2024), also showing promising results. However, these studies lack a control group.

Coming next to proper RCTs, we again find that the majority of works focus on Programming Education. Lyu et al. (2024), Li et al. (2024) and Pankiewicz and Baker (2024) all found that the treatment group using GenAI tutors benefited over the control group. However, Choudhuri et al. (2023) and Naik et al. (2024) found no significant differences in their metrics of interest. Shanshan and Sen (2024) concluded that significant learning gains fade with increasing abstraction levels. Outside programming, GenAI Based Learning Environments have been shown to be useful for dental students (Kavadella et al., 2024), teacher training (Lu et al., 2024), creative tasks (Urban et al., 2024), math problems (Pardos and Bhandari, 2024) etc. The only negative result we could find here comes from (Zhang et al., 2024) who found students learning using GPT had a significantly worse transfer performance compared to students who interacted with actual humans, despite the ChatGPT group showing higher levels of cognitive activity on an EEG. We note that all the studies listed here evaluate slightly different objectives, and we encourage interested readers to go through the original papers for more details.

All of the studies listed above work with adult students and we found very little work involving school-aged students. Frazier et al. (2024) surveyed high school students to conclude that they prefer a customized GPT over the generic version for programming education. Lieb and Goel (2024) developed NewtBot for high school physics but did not run a proper trial with it. Cheng et al. (2024) tested the efficacy of GPT4 in addition to Augmented Reality for elementary school children. Chen and Chang (2024) tested it on middle school students in addition to Game Based Learning. The GPT groups did better in both cases.

2.3 Our Contributions

Despite the numerous studies published in this field, we note that the literature on school-aged students is still limited, and none of the existing studies' tutoring strategies provide the flexibility that we grant GPT-4 with our strategy. The scarcity of studies does not necessarily imply a scarcity of potential for LLM-based technologies in this area. The potential of these models in schools is significant, largely untapped and well worth investigating.

Our work contributes to the limited literature on non-computer science subjects, and even more limited empirical literature. To the best of our knowledge, our work is the first in-field RCT incorporating a recent, state-of-the-art LLM as a tutor for language learning in a school. We provide empirical insights into the potential of GPT-4 as a tutor which can be built upon by the community in the future.

3 Methodology

Our goal while designing the intervention was to create something that fits into existing school systems **without causing any disruption**, and is **able to adapt** to whatever material is being taught in the class. In order to align it with the material being taught in class, we would have to provide teachers with an **easy-to-use** way to modify the prompt without needing technical know-how. Finally, to be easy to explain, trust and adapt to technological progress, we wanted a **minimalistic** design.

We provide an overview of the study design in figure 1. The teacher, who assigns weekly homework exercises to students, provides three key elements for each exercise: the *purpose*; the *description*; and an *example*, representing a typical instance of the homework assignment (we explain these components later), which are used to generate a prompt for GPT-4. We test the effectiveness of GPT-4 based homework compared to traditional homework in an RCT. We assess students' learning outcomes and experiences, using both external measures (pre- and post-tests) and student feedback through questionnaires.

In this section, we first briefly describe the background of the participants and the area of intervention. We then describe the prompting strategy. Finally, we describe the RCT design and the questionnaires that were used.

3.1 Participants

The Italian high school system, "scuola superiore," spans 5 years and includes lyceums, technical institutes, and professional institutes. Lyceums prepare students for university, technical institutes

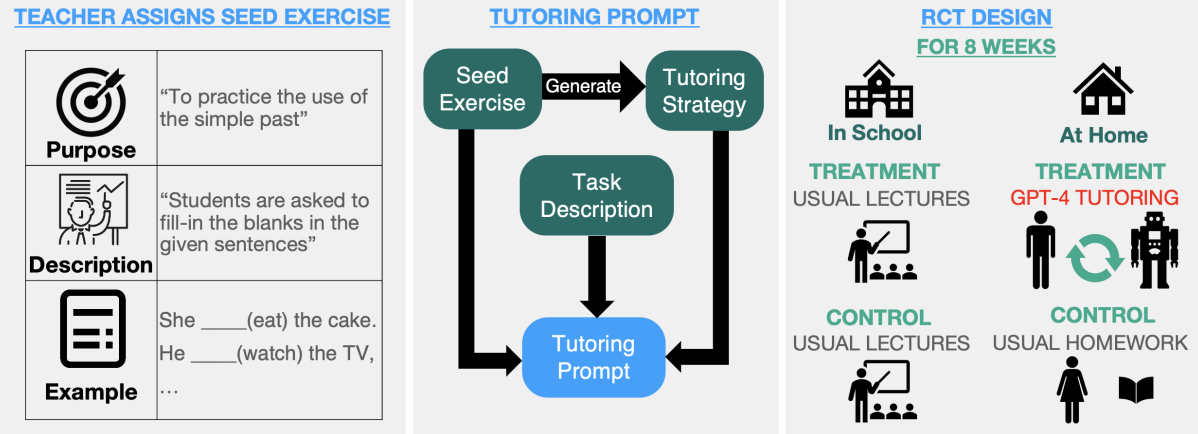

Figure 1: Illustration of the study design. See Sections 3.3 and 3.3.1 for detailed explanation

offer career-oriented education, and professional institutes provide vocational training. We partnered with a technical institute, working with 4 classes, all of whom were taught English by the same teacher. Two classes were in the 3rd year (median student age 16) and two in the 5th year (median age 18). All students were on the tourism track, focusing on business administration and foreign languages. The 3rd year classes consisted of a total of 39 students, of which 20 (18F and 2M) were assigned to the control group while 19 (17F and 2M) were assigned to the treatment group. The 5th year consisted of 37 students, of whom 19 (17F and 2M) were assigned to the control group and 18 (13F and 5M) were assigned to the treatment group. All the students had access to the internet either through a smartphone or through a computer.

3.2 Area of intervention

In every class, the English curriculum was composed of two main parts: 3 hours of weekly lectures and 1-2 hours of weekly homework and self-study. We intervened on homework and self-study. Treatment group students were assigned interactive sessions with GPT-4. Control group students continued with the typical homework they would have for the rest of the year. Treatment and control group students attended the same lectures. It must be noted that the students of neither group were forbidden from using ChatGPT on their own[1] so any effects seen are in addition to that of self-usage.

3.3 Prompting Strategy

Figure 1 left and center panels summarize our prompting strategy. As such, we ask the teacher to provide the following three components of the *seed exercise* (Full list of used exercises is listed in Table 1:

1. **Exercise purpose:** the pedagogical purpose of the exercise, described in a few words (shown in bold in Table 1)

2. **Exercise description:** a description of the task (non-bold text in Table 1

3. **Exercise example:** the exercise itself as it would be shown to the students

We note that the teacher would need to prepare these for regular homework anyway. If using exercises from a textbook, the teacher could use the title of the exercise as the purpose, description as description and the actual exercise as example. In post-study feedback, the participating teacher claimed that while this change introduced an initial learning curve for the teachers, in the long term it would reduce their workload somewhat.

Having obtained the seed exercise, we first ask GPT to generate a step-by-step plan on how to carry out the homework. The final prompt is obtained by appending the seed exercise and the generated strategy to a generic Task Description. The LLM of our choice was `gpt-4-0125-preview`. The interface was hosted on a dedicated website that students could access with their own personal devices outside of school hours.

We report the GPT-4 prompts for strategy generation and tutoring in the Appendix B. For the strategy generation we provide the seed exercise alongside a basic description of the tutoring task, mentioning that we are working with high school students and aiming at a B2 level of English according to the Common European Framework of Refer-

[1] Almost two-thirds of Italian students are likely to be using ChatGPT according to this report

	3rd Year	5th Year
Unit 0		**(Only class 5A) WWI, open questions**: Students answer questions about WWI and British responses in 8-10 lines.
Uni1 1	**Conditionals, sentence correction:** Students are asked to fix mistakes in the given sentences. **Conditionals, sentence completion**: Students complete sentences using the correct tense of the verb in brackets. **Conditionals, sentence transformation**: Students complete the second sentence to have the same meaning as the first using the given word.	**Poem commentary, analysis, and comparison:** Students analyze and compare "The Soldier" by R. Brooke and "Dulce et Decorum est" by W. Owen.
Unit 2	**I wish/If only and mixed conditionals, either/or questions:** Students choose the correct verb form from two options for sentences using "I wish," "If only," and mixed conditionals. **I wish/If only and mixed conditionals, sentence completion:** Students complete sentences using the correct tense of the verb in brackets.	**Key events of the 20th century, open questions:** Students answer questions about 1967, WWII, and Margaret Thatcher in 4-6 lines.
Unit 3	**Essay on contemporary social issues**: Students write a 200-220 word argumentative essay on a contemporary social issue.	**The Roaring 20s, open questions**: Students answer questions about the economic, political, and social changes in the US during the 20s, including "The Great Gatsby."
Unit 4	**Passive voice, sentence transformation**: Students transform active sentences to passive form while keeping the same tense. **Passive voice, sentence completion**: Students complete sentences using the correct tense of the verb in brackets, focusing on passive forms. **Passive vs. active voice, either/or questions**: Students choose the correct active or passive form and tense from two options given.	**Mid-century America, open questions**: Students answer questions about the Cold War, the 60s, cultural revolution, and the crisis of the 70s in the US.
Unit 5	**Causative verbs (have/get something done), sentence completion**: Students complete sentences using the correct tense of the verb in brackets, focusing on the causative form. **Passive voice with double object, sentence transformation**: Students transform active sentences with two objects to passive form, giving both options.	**Human rights movements, Gandhi, open questions**: Students answer questions about Gandhi's life, achievements, and influence in 4-6 lines.
Unit 6	**Reported speech statements, sentence transformation**: Students transform direct statements into reported speech, using the correct tense based on the introductory verb. **Reported speech questions, sentence transformation**: Students transform direct questions into reported speech, using the correct tense based on the introductory verb. **Reported speech correction, sentence correction**: Students correct the mistakes in sentences related to reported speech using different reporting verbs.	(Only class 5B) **Important women in history, open questions**: Students answer questions about Queen Victoria, Emmeline Pankhurst, and Rosa Parks in 4-6 lines.

Table 1: List of exercises used by both grades. The two fifth grade classes, 5A and 5B were off by one week.

ence for Languages (CEFR), and ask for an appropriate tutoring strategy. In the tutoring prompt, we provide a description of the tutoring task together with the seed exercise and the generated strategy.

3.3.1 Intervention Design

We assigned students within each class treatment or control condition using stratified randomization based on their self-reported English GPA in the current year. lectures were the same for all students. The teacher was not informed of the condition of the individual students, and students were instructed not to share their group with her, to avoid interference, although some leakage is likely to have happenned.

All students received weekly homework consisting of one or more exercises, to be done on a dedicated website that students could access with their own devices, including smartphones and computers. Students in the control group received the homework as assigned by the teacher, in a format comparable to the exercise they received prior to the experiment. They solved each exercise individually and typed the answer on the online platform. For each exercise assigned by the teacher, students in the treatment group had access to a chat with GPT4. While GPT4 had access to the exercise via its prompt, students did not see the original exercise from the teacher. In both treatment and control conditions, students could access the platform at any time. We did not enforce a minimum or maximum level of engagement. The teacher could not

observe the content of the student's solution. The intervention was planned to run for 6 weeks, but was extended to 8 weeks due to delays in covering the planned content. Hereafter, whenever we refer to a 'week,' we refer to the time taken to cover the content planned for a week, which in reality may have taken longer than a calendar week.

3.4 Randomized Controlled Trial

Having designed our intervention as described above, we seek to understand how it affects the students' experiences and learning. In particular, we are interested in how the treatment group students differ from the control group in terms of...

RQ1 ...the immediate experience of doing homework, in terms of interest, satisfaction etc.

RQ2 ...how they feel about homework and the subject in general in the long term[2]

RQ3 ...their learning gains in the topics taught.

Based on these RQs we design our 3 sets of questionnaires. The **Weekly Questionnaires**, administered partly at the end of each exercise and partly at the end of each week, and consisted of four 6-point Likert scale questions designed to analyze **RQ1**. We hereafter refer to these as *usefulness*, *interestingness*, *comprehensiveness* and *level_of_resources*. In addition, we asked the treatment group to report how the tutor was helpful giving the option to select among a range of potentially useful aspects and also if they felt that the tutor made any mistakes. The full text can be found in Table 6.

The **Initial and Final Questionnaires** were longer questionnaires given to the students before the beginning and after the conclusion of the RCT, aimed at tackling **RQ2**. We did not find a standard set of questions suitable for us, so we adapted from a combination of sources. The first set consisted of 10 background questions going over self-efficacy and motivation, 6 of which were adapted from (Tuan* et al., 2005) by making them specific to English. This was followed by two sections of 6 questions each regarding English homework and English lectures. Both these sections were designed in accordance with the ARCS framework. For these questions, the initial questionnaire asked about their general experiences, while the final questionnaire asked about their experience during the RCT. All 22 questions were presented in Italian, and were to be answered on a 6-point Likert scale.

In addition to these questions, the initial questionnaire also asked the students to report their past grades (used for splitting the groups) and their age. The final questionnaire had a separate segment for the treatment group asking for their experiences and future outlook. For full text of all questions, see Tables 4 and 5.

Finally, to evaluate **RQ3**- we conducted **Pre-Test and Post-Test** Both the pre-test and post-test consisted of 24 multiple-choice questions. We assigned 1 point for each question answered correctly. The questions were provided to us by the teacher, who designed 8 questions for each week of intervention. For each week, we randomized half of the questions to the pre-test and the post-test each.

4 Results and Analysis

	Third Year	Fifth Year	Total
Control Group			
# Assignments	195	97	292
Median Homework Word Count	56	157	74
Treatment Group			
# Chats	199	93	292
Agent Messages (Median)	15	12	14
User Messages (Median)	14	11	13
Total Words per Chat - Agent (Median)	977	1040	989
Total Words per Chat - User (Median)	114	314	143

Table 2: Usage statistics for the tutor

We start off with some general statistics, and then proceed to discuss each of our research questions in their own section.

4.1 General statistics

4.1.1 Participant Background

Table 3 reports the self-reported participant backgrounds. We note that there is a slight overestimation of English ability on the part of the students, as more people consider themselves above average than below average as can be seen in table 3

4.1.2 Usage Statistics

Table 2 shows the overall usage summary of the platform. The 5th year homework consisted of open questions on literature and history, with several questions each week. The most common behaviour among students was to write a somewhat complete answer. The answer was then refined iteratively based on feedback from the tutor, adding nuance, correcting grammar and including or fixing

[2]here 'long term' refers to 8 weeks.

factual information. The 3rd year homework consisted of objective type (except for one essay type exercise), where the students were given sentences which they had to edit, complete or transform according to the question, and there was almost always a single correct answer. Student utterances for these questions were most of the time just attempts at the right answers, and not many students tried to have full conversations. We segmented the conversations into a total of 1549 questions, of which 940[3] were solved immediately by the students, while in 365[3] cases the tutor revealed the answers. The conversations where a reveal occurred were on average 4.7 utterances long, which would imply about 2 attempts by the student. Correct cases were almost always 3 utterances long (Tutor-Student-Tutor) with the exception of an exercise that required both the passive form and the double object passive form which required 5 utterances.

4.2 RQ1: Students' Immediate Experience Doing Homework

Looking at the weekly questionnaires, we observed that students in the treatment group gave higher ratings in all 4 categories. Of these *interestingness* ($d = 0.593$, $P = 0.011$) and *level_of_resources* ($d = 0.586$, $P = 0.015$) were significant while *usefulness* ($d = 0.356$, $P = 0.125$) and *comprehensiveness* ($d = 0.281$, $P = 0.234$) were not significant. Upon inspecting the chats we find that the tutor typically stuck to the example exercise provided in its prompt, unless the student specifically asked for more questions or made too many errors indicating the need for more practice. This would explain students not finding the exercises significantly more comprehensive or useful, as these are inherent properties of the exercises, unlike interest in the homework or feeling a lack of supporting resources, which can be improved upon via delivery. Overall, we can say that the (self-perception of the) student's immediate experiences was improved by the intervention.

4.3 RQ2: Students' Long Term Feelings About Homework

Figure 2 shows the average change in the students' responses between the initial and final surveys (questions with a negative sentiment have had their signs reversed to make higher better for all questions). We note that the differences in the two groups are not significant for any of the questions (after correcting for FDR) but certain trends still emerge. First of all, overall satisfaction increases for both groups, but increases more for the treatment group. Further, for all questions in the homework section (where we intervened), the treatment group's opinions improved more than that of the control group.

As a part of the final survey, students in the treatment group were asked how they felt about the tutor. 32/33 respondents thought that the tutor helped them with their homework, whereas 30/35 felt that the tutor improved their English on a practical level. Further, 26/34 respondents felt that the tutor helped them keep up with the English program. Most importantly, 32/35 overall respondents wanted to continue using the tutor, with the 3 people saying "No" all being in their last year of school. Overall, we can conclude that the short-term enjoyment translates to the students feeling good about the intervention even in the longer term.

	Third Year	Fifth Year	Total
Control Group			
#Students	20	19	39
Mean Grade in English	7.20	7.54	7.37
Held Back in English	1	3	4
Below Average English Ability	4	7	11
Average English Ability	8	3	11
Above Average English Ability	8	9	17
Treatment Group			
#Students	19	18	36
Mean Grade in English	7.63	7.43	7.53
Held Back in English	0	1	1
Below Average English Ability	6	6	12
Average English Ability	5	5	10
Above Average English Ability	8	7	15

Table 3: Self reported previous performance by students

4.4 RQ3: Learning Gains

While students enjoying homework is a good outcome, it does not tell us much about the intervention in isolation, as it can easily be achieved by allowing the students to game the system and thereby reducing the amount of effort required on their part[4]. To ensure that this does not happen, we need to ensure that the other benefits did not come at the cost of a reduction in learning gains. To check this we run 3 one-sided t-tests with the alternative hypothesis treatment>control.

T-testing on the entire data ($d = 0.251$, $P = 0.156$) shows that an increase in learning gains due

[3]As judged by GPT-4o. These might have some errors

[4]although they could choose not to do the homework regardless of the intervention

to the intervention is over 5 times more likely than a decrease, although neither possibility is significant. Running separate tests for the two years, however, shows that this learning gain is entirely driven by the 3rd year who show a significant improvement ($d = 0.603$, $P = 0.044$), while the 5th year's learning gains show very little differences between the groups ($d = -0.004$, $P = 0.505$). We posit that this difference could emerge due to the 5th year homework being essay type compared to the 3rd year homework being more objective with a single correct answer, which could have led to the following 2 issues:

1. The lack of a clear correct answer would make 5th year answers harder to evaluate.

2. The pre- and post-test for both classes was objective type, so the 5th year homework would have helped less in general.

We note that the observed effect size for the 3rd year is more or less consistent with the difference between graded and ungraded homework noted by Bloom (1984). Overall, we conclude that, not only does our intervention preserve learning gains, but under the right conditions, it can also improve them.

4.5 Other Observations

Beyond our research questions, the data collected by the RCT also provided us with some other interesting insights that we summarise here:

Students with higher initial scores showed lower learning gains than those with lower initial scores. We find a negative Pearson correlation between *score_initial* and *learning_gains* for the treatment group ($R = -0.777$, $P < 0.001$) which is actually stronger than the control group ($R = -0.628$, $P < 0.001$). This indicates that students with lower initial knowledge are able to catch up with their peers, instead of falling further behind, contrary to some prior studies (Prather et al., 2024; Cipriano and Alves, 2024).

Third Year learning gains are largely mediated by engagement, which is consistent with previous work (Altememy et al., 2023; S. N. Jyothy and Achuthan, 2024). Using *words_typed* as a proxy for engagement, we observe a positive correlation between *words_typed* and *learning_gains*, largely driven by the 3rd year ($R = 0.434$, $P = 0.007$). An OLS regression for *learning_gains* wrt *words_typed* and *condition*, shows *words_typed* to be the only significant variable[5]. Although this shows that the direct effect of our treatment on learning gains might be limited, we must stress the fact that this engagement was caused naturally as a result of our intervention, as it was made clear to all participants that they had no obligation to interact with the tutor and no information regarding their interactions would be shared with the teacher or the school.

Hallucinations and errors seem to rare. To test the prevalence of errors made by GPT, students were asked every week if the tutor had made some errors in their chats. Of 160 responses, only 16 indicated a problem. Going over all the marked exercises, we concluded that there were no more than 14 utterances that could be considered problematic. We also randomly went over 10% of the conversations and did not find any additional issues. Given that the total number of utterances is around 4000, the error rate is less than 0.5% which is well within acceptable thresholds.

Finally, **there is no evidence for novelty effects.** Looking at weekly measured variables over time, we found no significant drop in mean or median across the 6 weeks (see fig 3 in appendix for box plots). Although the absence of evidence is not evidence of absence, this certainly makes it less likely that the differences are caused by the excitement of using something new.

5 Conclusions and Discussion

In this work, we run an RCT to evaluate the ability of GPT-4 to function as a tutor. We find that students find this replacement of homework more useful and interesting, and are enthusiastic about continuing using it in their education. Further, students using GPT4 more often felt that they had the necessary resources to complete what was required for them. In addition to this, we also observe significant improvement in 3rd year students in learning as measured by tests, while 5th year students maintain their performances. Also, we do not find evidence of bias towards stronger students or harmful hallucinations. We further notice that the self-assessments don't show significant decline over the RCT period, thereby making novelty

[5] see Table 7 for all coefficients

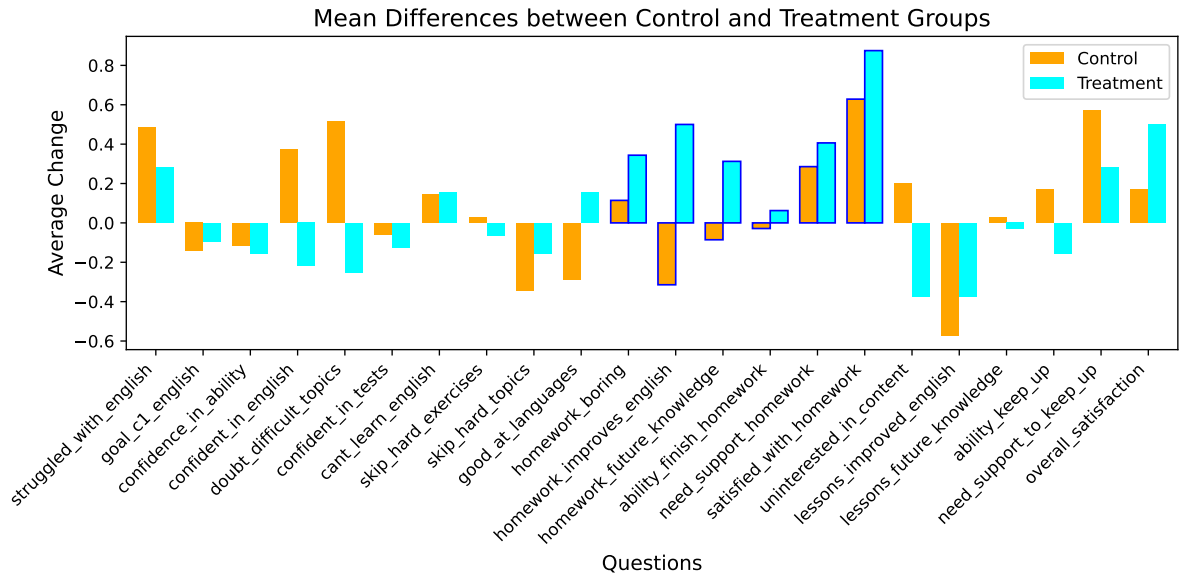

Figure 2: Change in Survey responses between the initial and the final questionnaire. Questions with negative sentiment are flipped(these are marked with a † in tables 4 and 5) so higher is always an improvement. Questions regarding homework are marked in blue borders.

effects less likely.

School and homework are often perceived as an unwelcome chore by students, and according to the personal experience of some teachers we worked with, there is an increasing lack of interest and engagement from the students, making interventions on these aspects even more needed. Low teacher-to-student ratios in many schools across the world means that teachers cannot always attend to issues being faced by individual students, and many parents cannot afford after-school services and personal tutoring for their children, causing struggling students to often be left lagging behind. While LLMs like GPT are rather cheaply available to all students, self-regulated use might not be sufficient especially for weaker students, as evidenced by the fact that we observed improvements despite several students in both groups claiming to have used GPT by themselves. This, in fact is also related to the concern that use of LLMs like GPT can exacerbate certain issues in education, like plagiarism, rote memorisation etc. and also cause an overdependence of students on AI.However, we believe that for best results, schools need to provide centrally prompted systems for the students to use, which strike a balance between self learning and scaffolding. And as we discovered, with the right design, this can be achieved without increasing the load on teachers. Empowering a larger number of students with the resources needed to achieve what is expected for them has the potential to reduce the gap between students living in more and less privileged circumstances, providing a fairer playing field within and across schools.

Given the continuing development of LLMs to improve them across all tasks, we believe that our study, despite all its limitations (which we shall discuss in section 5) paves the way for a bright future where hard to scale tasks like tutoring can be taken over by AI tutors, bringing the benefit of tutoring to a much greater number of students around the world.

Limitations

The intervention being run in a real school with real students increases its ecological validity, but it also leads to several limitations. We worked with a single teacher in a system which allows for a lot of flexibility in teaching styles, so results might not fully transfer to different teachers. Our intervention ran for a period of 8 weeks, which may not be sufficient for differences in learning to fully emerge, especially given that we only intervened in the homework. Our sample size was restricted to the students in the classes taught by the participating teacher, and even then, we had to drop some students because they were unwilling to participate, or did not complete the initial or final surveys. Further, we split students in the same class were split into treatment and control groups, which meant that control and treatment students were able to interact and influence each other's outcomes.

We also faced a dilemma in deciding the proper way to conduct the pre- and post-tests. A subjective test evaluated by the teacher would be more comprehensive and consistent with regular tests, but doing so would not only lead to extra work for the teacher, but also open up the possibility of the teacher biasing the results, as the teacher definitely gained some insights on which student was in which group based on their interactions with the students over the course of the study. We instead chose to use an auto-graded MCQ which alleviates the aforementioned concerns, but opens up the possibility guessing the answer, hence increasing the noise in the measurement. In addition to this, this could also have lead to an unfair assessment of 5th year students as mentioned earlier.

In addition to this there are also the standard threats to the generalizability of any study conducted in a localised region. Our participants came from a westernised and educated culture with relatively high levels of personal freedoms. They were mostly women, and came from a single stratum in the secondary education system stratified roughly based on academic ability. The latter is particularly important as it is usually accepted that students with different ability levels respond differently to interventions.

Finally, we also were limited to a particular version of a particular LLM, and were able to perform only a limited amount of prompt engineering, all of which can affect the final outcome. We also used a single interface, which did not utilise the full capabilities of the model (for example, GPT4 is capable of working with both audio and video in addition to text at the time of writing). We also tested a single subject, and a single language of instruction, neither of which are necessarily intrinsic restrictions of the LLM being used.

Ethics Statement

This study was approved by the Human Subjects Committee of the Faculty of Economics, Business Administration, and Information Technology of the University of Zurich (OEC IRB # 2024-018). All participants provided informed consent prior to enrollment. For minors, the consent was provided by both parents or legal tutors, unless one parent or tutor alone had sole tutorship. Participant confidentiality was strictly maintained, and all data were anonymized. The study complied with all applicable regulations and ethical standards. Each aspect of the design was discussed and developed in collaboration with education professionals and with the school personnel involved, to ensure an optimal and fair experience for all participants and non-participants. The school personnel was informed of the potential risk and instructed to carefully monitor participants as well as non-participating students potentially affected by the study, and to provide them with extensive support.

Acknowledgments

We extend our sincerest thanks to the teachers, staff, and administration of Istituto Pindemonte in Verona for their support in conducting this study, especially Dr. Lara Quarti, in whose classes the experiment was conducted. We thank Dr. Julia Chatain for her guidance on the framing of this study. We thank Shashank Sonkar, Shehzaad Dhuliawala and Manuel Bernal-Lecina for their feedback on the final drafts of the paper.

Sankalan Pal Chowdhury additionally thanks his colleagues Manuela Pineros-Rodriguez, Fan Wang and Adrienn Toth for their helpful insights during discussions about this project. He is partially funded by the ETH-EPFL Joint Doctoral Program for Learning Sciences.

This project was funded by a grant from the Swiss National Science Foundation (Project No. 197155) and a Responsible AI grant by the Haslerstiftung.

References

Haady Abdilnibi Altememy, Nour Raheem Neamah, Rabaa Mazhair, Nada Sami Naser, Ali Afrawi Fahad, Nazar Abdulghffar al sammarraie, Hidab Rasul Sharif, Mohamed Amer Alseidi, and Mohammed Yousif Oudah Al-Muttar. 2023. Ai tools' impact on student performance: Focusing on student motivation & engagement in iraq. *Przestrzeń Społeczna (Social Space)*, 23(2):143–165.

Kehinde Aruleba, Ismaila Temitayo Sanusi, George Obaido, and Blessing Ogbuokiri. 2023. Integrating chatgpt in a computer science course: Students perceptions and suggestions. *Preprint*, arXiv:2402.01640.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

BENJAMIN S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effec-

tive as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

Laurie Butgereit, Herman Martinus, and Muna Mahmoud Abugosseisa. 2023. Prof pi: Tutoring mathematics in arabic language using gpt-4 and whatsapp. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pages 000161–000164.

Giovanna Campani. 2013. *Private Tutoring in Italy: Shadow Education in a Changing Context*, pages 115 – 128. Brill, Leiden, The Netherlands.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26:e53164.

Ching-Huei Chen and Ching-Ling Chang. 2024. Effectiveness of ai-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior. *Education and Information Technologies*, pages 1–22.

Alan Y. Cheng, Meng Guo, Melissa Ran, Arpit Ranasaria, Arjun Sharma, Anthony Xie, Khuyen N. Le, Bala Vinaithirthan, Shihe (Tracy) Luan, David Thomas Henry Wright, Andrea Cuadra, Roy Pea, and James A. Landay. 2024. Scientific and fantastical: Creating immersive, culturally relevant learning experiences with augmented reality and large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Vuthea Chheang, Shayla Sharmin, Rommy Marquez-Hernandez, Megha Patel, Danush Rajasekaran, Gavin Caulfield, Behdokht Kiafar, Jicheng Li, Pinar Kullu, and Roghayeh Leila Barmaki. 2024. Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments. *Preprint*, arXiv:2306.17278.

Rudrajit Choudhuri, Dylan Liu, Igor Steinmacher, Marco Gerosa, and Anita Sarma. 2023. How far are we? the triumphs and trials of generative ai in learning software engineering. *Preprint*, arXiv:2312.11719.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *ACM Conference on Learning @ Scale*.

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. 2024. Comprehensive assessment of jailbreak attacks against llms. *arXiv preprint arXiv:2402.05668*.

Bruno Pereira Cipriano and Pedro Alves. 2024. "chatgpt is here to help, not to replace anybody" – an evaluation of students' opinions on integrating chatgpt in cs courses. *Preprint*, arXiv:2404.17443.

Peter A Cohen, James A Kulik, and Chen-Lin C Kulik. 1982. Educational outcomes of tutoring: A meta-analysis of findings. *American educational research journal*, 19(2):237–248.

Nassim Dehouche. 2021. Plagiarism in the age of massive generative pre-trained transformers (gpt-3). *Ethics in Science and Environmental Politics*, 21:17–23.

Paul Denny, Stephen MacNeil, Jaromir Savelka, Leo Porter, and Andrew Luxton-Reilly. 2024. Desirable characteristics for ai teaching assistants in programming education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. ACM.

Rebecca Ferguson and Mike Sharples. 2014. Innovative pedagogy at massive scale: Teaching and learning in moocs. In *Open Learning and Teaching in Educational Communities*, pages 98–111, Cham. Springer International Publishing.

Matthew Frazier, Kostadin Damevski, and Lori Pollock. 2024. Customizing chatgpt to help computer science principles students learn through conversation. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, ITiCSE 2024, page 633–639, New York, NY, USA. Association for Computing Machinery.

Arto Hellas, Juho Leinonen, and Leo Leppänen. 2024. Experiences from integrating large language model chatbots into the classroom. *Preprint*, arXiv:2406.04817.

Sven Jacobs and Steffen Jaschke. 2024. Evaluating the application of large language models to generate feedback in programming education. In *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Argyro Kavadella, Marco Antonio Dias da Silva, Eleftherios G Kaklamanos, Vasileios Stamatopoulos, and Kostis Giannakopoulos. 2024. Evaluation of chatgpt's real-life implementation in undergraduate dental education: Mixed methods study. *JMIR Med Educ*, 10:e51344.

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM.

Nam Wook Kim, Hyung-Kwon Ko, Grace Myers, and Benjamin Bach. 2024. Chatgpt in data visualization education: A student perspective. *Preprint*, arXiv:2405.00748.

Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. Unreflected acceptance – investigating the negative consequences of chatgpt-assisted problem solving in physics education. *Preprint*, arXiv:2309.03087.

Wengxi Li, Roy Pea, Nick Haber, and Hari Subramonyam. 2024. Tutorly: Turning programming videos into apprenticeship learning environments with llms. *Preprint*, arXiv:2405.12946.

Anna Lieb and Toshali Goel. 2024. Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Mark Liffiton, Brad Sheese, Jaromir Savelka, and Paul Denny. 2023. Codehelp: Using large language models with guardrails for scalable support in programming classes. *Preprint*, arXiv:2308.06921.

Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J. Malan. 2024. Teaching cs50 with ai: Leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 750–756, New York, NY, USA. Association for Computing Machinery.

Jijian Lu, Ruxin Zheng, Zikun Gong, and Huifen Xu. 2024. Supporting teachers' professional development with generative ai: The effects on higher order thinking and self-efficacy. *IEEE Transactions on Learning Technologies*, 17:1279–1289.

Wenhan Lyu, Yimeng Wang, Tingting (Rachel) Chung, Yifan Sun, and Yixuan Zhang. 2024. Evaluating the effectiveness of llms in introductory computer science education: A semester-long field study. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24. ACM.

Lior Naamati Schneider. 2024. Enhancing ai competence in health management: students' experiences with chatgpt as a learning tool. *BMC Medical Education*, 24.

Atharva Naik, Jessica Ruhan Yin, Anusha Kamath, Qianou Ma, Sherry Tongshuang Wu, Charles Murray, Christopher Bogart, Majd Sakr, and Carolyn P. Rose. 2024. Generating situated reflection triggers about alternative solution paths: A case study of generative ai for computer-supported collaborative learning. *Preprint*, arXiv:2404.18262.

Hyacinth S Nwana. 1990. Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, 4(4):251–277.

Ola Tokunbo Odekeye, Oluwaseyi Aina Gbolade Opesemowo, Micheal Adelani Adewusi, Anthony Kola-Olusanya, Tunde Owolabidi, Jubril Busuyi Fakokunde, and Fred Awaah. Will ai writing tools revolutionise learning attitudes-insight from undergraduate students of global south. *Available at SSRN 4755024*.

OpenAI and GPT4 Team. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Aadarsh Padiyath, Xinying Hou, Amy Pang, Diego Viramontes Vargas, Xingjian Gu, Tamara Nelson-Fromm, Zihan Wu, Mark Guzdial, and Barbara Ericson. 2024. Insights from social shaping theory: The appropriation of large language models in an undergraduate programming course. *ACM Conference on International Computing Education Research*.

Maciej Pankiewicz and Ryan S. Baker. 2024. Navigating compiler errors with ai assistance - a study of gpt hints in an introductory programming course. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, ITiCSE 2024. ACM.

Zachary A. Pardos and Shreya Bhandari. 2024. Chatgpt-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE*, 19(5):1–18.

Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM.

Petra Polakova and Blanka Klimova. 2024. Implementation of ai-driven technology into education – a pilot study on the use of chatbots in foreign language learning. *Cogent Education*, 11(1):2355385.

James Prather, Brent Reeves, Juho Leinonen, Stephen MacNeil, Arisoa S. Randrianasolo, Brett Becker, Bailey Kimmel, Jared Wright, and Ben Briggs. 2024. The widening gap: The benefits and harms of generative ai for novice programmers. *Preprint*, arXiv:2405.17739.

Laryn Qi, J. D. Zamfirescu-Pereira, Taehan Kim, Björn Hartmann, John DeNero, and Narges Norouzi. 2024. A knowledge-component-based methodology for evaluating ai assistants. *Preprint*, arXiv:2406.05603.

Darshanand Ramdass and Barry J Zimmerman. 2011. Developing self-regulation skills: The important role of homework. *Journal of advanced academics*, 22(2):194–218.

Raghu Raman S. N. Jyothy, Vysakh Kani Kolil and Krishnashree Achuthan. 2024. Exploring large language models as an integrated tool for learning, teaching, and research through the fogg behavior model: a comprehensive mixed-methods analysis. *Cogent Engineering*, 11(1):2353494.

Scott A Schartel. 2012. Giving feedback–an integral part of education. *Best practice & research Clinical anaesthesiology*, 26(1):77–87.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle&riley: Insights from designing and evaluating a large language model-based conversational tutoring system. *Preprint*, arXiv:2404.17460.

Shang Shanshan and Geng Sen. 2024. Empowering learners with ai-generated content for programming learning and computational thinking: The lens of extended effective use theory. *Journal of Computer Assisted Learning*, 40(4):1941–1958.

Brad Sheese, Mark Liffiton, Jaromir Savelka, and Paul Denny. 2024. Patterns of student help-seeking when using a large language model-powered programming assistant. In *Proceedings of the 26th Australasian Computing Education Conference*, ACE 2024. ACM.

Ben Arie Tanay, Lexy Arinze, Siddhant S. Joshi, Kirsten A. Davis, and James C. Davis. 2024. An exploratory study on upper-level computing students' use of large language models as tools in a semester-long project. *Preprint*, arXiv:2403.18679.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Maung Thway, Jose Recatala-Gomez, Fun Siong Lim, Kedar Hippalgaonkar, and Leonard W. T. Ng. 2024. Battling botpoop using genai for higher education: A study of a retrieval augmented generation chatbots impact on learning. *Preprint*, arXiv:2406.07796.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hsiao-Lin Tuan*, Chi-Chin Chin, and Shyang-Horng Shieh. 2005. The development of a questionnaire to measure students' motivation towards science learning. *International journal of science education*, 27(6):639–654.

Marek Urban, Filip Děchtěrenko, Jiří Lukavský, Veronika Hrabalová, Filip Svacha, Cyril Brom, and Kamila Urban. 2024. Chatgpt improves creative problem-solving performance in university students: An experimental study. *Computers Education*, 215:105031.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024. Patient-Ψ: Using large language models to simulate patients for training mental health professionals. *Preprint*, arXiv:2405.19660.

Boyang Yang, Haoye Tian, Weiguo Pian, Haoran Yu, Haitao Wang, Jacques Klein, Tegawendé F. Bissyandé, and Shunfu Jin. 2024. Cref: An llm-based conversational software repair framework for programming tutors. *Preprint*, arXiv:2406.13972.

J. D. Zamfirescu-Pereira, Laryn Qi, Björn Hartmann, John DeNero, and Narges Norouzi. 2024. 61a-bot: Ai homework assistance in cs1 is fast and cheap – but is it helpful? *Preprint*, arXiv:2406.05600.

Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28.

Jiayue Zhang, Yiheng Liu, Wenqi Cai, Lanlan Wu, Yali Peng, Jingjing Yu, Senqing Qi, Taotao Long, and Bao Ge. 2024. Investigation of the effectiveness of applying chatgpt in dialogic teaching using electroencephalography. *Preprint*, arXiv:2403.16687.

A Questionnaires

We report the initial questionnaire, the final questionnaire, and the weekly questionnaire. We report the orginal questions in Italian as well as an English translation.

Table 4: Initial Questionnaire. Questions with negative sentiment are marked with a dagger(†)

Original Question	Translation	Type of Answer	Dataset Label
Età	Age		age
Sei mai stato rimandato in inglese?	Have you ever been held back in English?	Yes; No	failed_english
Come pensi che sia il tuo livello di inglese rispetto alla media della tua classe, al di là dei voti?	How do you think your level of English compares to the average in your class, aside from grades?	Below average; Slightly below average; Average; Slightly Above Average; Above Average	english_level
Qual'è la tua media in inglese quest'anno?	What is your average grade in English this year?	<6; 6-6.49; 6.5-6.99; 7-7.49; 7.5 - 7.99; 8 - 8.49; 8.5-9; >9; I would rather not disclose	english_average
General Questions: Questions marked with a '*' are standardized SESQ.			
Faccio fatica a stare al passo col programma di inglese	I struggle to keep up with the English program	6-point Likert	struggled_with_english†
Voglio raggiungere un buon livello di inglese nei prossimi anni	I want to achieve a good level of English in the coming years	6-point Likert	goal_c1_english
Penso di avere le capacità per raggiungere un buon livello di inglese nel corso dei prossimi anni	I think I have the capacity to reach a good level of English in the coming years	6-point Likert	confidence_in_ability
* In inglese, indipendentemente da quanto un argomento è difficile, sono sicuro di poterlo capire.	In English, no matter how difficult a topic is, I am sure I can understand it.	6-point Likert	confident_in_english
* Non sono sicuro di poter imparare gli argomenti più difficili del programma di inglese.	I am not sure I can learn the most difficult topics of the English program.	6-point Likert	doubt_difficult_topics†
* Sono sicuro di poter andare bene in verifiche, interrogazioni e/o test di inglese.	I am sure I can do well in quizzes, oral exams, and/or English tests.	6-point Likert	confident_in_tests
* Per quanto mi sforzi, non riesco a imparare l'inglese	No matter how hard I try, I cannot learn English	6-point Likert	cant_learn_english†
* Quando un esercizio è troppo difficile, lo salto oppure faccio solo le parti più facili.	When an exercise is too difficult, I skip it or just do the easier parts.	6-point Likert	skip_hard_exercises†
* Quando un argomento è troppo difficile, lo salto e non provo a impararlo.	When a topic is too difficult, I skip it and do not try to learn it.	6-point Likert	skip_hard_topics†
Sono portato per le lingue.	I am talented for languages.	6-point Likert	good_at_languages
ARCS - Homework			
I compiti di inglese sono noiosi.	English homework is boring.	6-point Likert	homework_boring†
I compiti di inglese sono utili per migliorare il mio livello di inglese.	English homework is useful for improving my level of English.	6-point Likert	homework_improves_english
I compiti di inglese sono utili per ottenere conoscenze che mi serviranno in futuro.	English homework is useful for gaining knowledge that will be useful in the future.	6-point Likert	homework_future_knowledge
Ho le capacità personali per portare a termine correttamente i compiti di inglese tutte le volte.	I have the personal abilities to properly complete English homework every time.	6-point Likert	ability_finish_homework
Avrei bisogno di più supporto e risorse per riuscire a fare i compiti di inglese.	I would need more support and resources to be able to do English homework.	6-point Likert	need_support_homework†
Sono soddisfatto di quello che imparo facendo i compiti di inglese.	I am satisfied with what I learn from doing English homework.	6-point Likert	satisfied_with_homework
ARCS - Lectures and Contents			
I contenuti del programma di inglese non mi interessano.	The contents of the English program do not interest me.	6-point Likert	uninterested_in_content†
Le lezioni di inglese sono utili a migliorare il mio livello di inglese.	English lessons are useful in improving my level of English.	6-point Likert	lessons_improved_english
Le lezioni di inglese mi forniscono conoscenze utili per il mio futuro.	English lessons provide me with useful knowledge for my future.	6-point Likert	lessons_future_knowledge
Ho le capacità personali per stare al passo col programma di inglese.	I have the personal abilities to keep up with the English program.	6-point Likert	ability_keep_up
Avrei bisogno di più supporto e risorse per riuscire a stare al passo col programma di inglese.	I would need more support and resources to be able to keep up with the English program.	6-point Likert	need_support_to_keep_up†
Complessivamente, sono soddisfatto di quello che imparo a scuola in inglese.	Overall, I am satisfied with what I learn in English at school.	6-point Likert	overall_satisfaction

Table 5: Final Questionnaire

Original Question	Translation	Type of Answer	Dataset Label
Da quale dispositivo hai eseguito l'accesso alla piattaforma per i compiti per casa?	From which device did you access the platform for homework?	Mobile, Laptop, Both	device_used
General Questions: Questions marked with a '*' are standardized SESQ.			
Nelle ultime 8 settimane, ho fatto fatica a stare al passo col programma di inglese	In the last 8 weeks, I have struggled to keep up with the English program	6-point Likert	struggled_with_english†
Voglio raggiungere un buon livello di inglese nei prossimi anni (approssimativamente un C1)	I want to reach a good level of English in the coming years (approximately a C1)	6-point Likert	goal_c1_english
Penso di avere le capacità per raggiungere un buon livello di inglese (approssimativamente un C1) nel corso dei prossimi anni	I think I have the ability to reach a good level of English (approximately a C1) over the coming years	6-point Likert	confidence_in_ability
* In inglese, indipendentemente da quanto un argomento è difficile, sono sicuro di poterlo capire	In English, no matter how difficult a topic is, I am confident I can understand it	6-point Likert	confident_in_english
* Non sono sicuro di poter imparare gli argomenti più difficili del programma di inglese	I am not sure I can learn the more difficult topics of the English program	6-point Likert	doubt_difficult_topics†
* Sono sicuro di poter andare bene in verifiche, interrogazioni e/o test di inglese	I am confident that I can do well in quizzes, oral exams, and/or English tests	6-point Likert	confident_in_tests
* Per quanto mi sforzi, non riesco a imparare l'inglese	No matter how hard I try, I cannot learn English	6-point Likert	cant_learn_english†
* Quando un esercizio è troppo difficile, lo salto oppure faccio solo le parti più facili	When an exercise is too difficult, I skip it or only do the easier parts	6-point Likert	skip_hard_exercises†
* Quando un argomento è troppo difficile, lo salto e non provo a impararlo	When a topic is too difficult, I skip it and do not try to learn it	6-point Likert	skip_hard_topics†
Sono portato per le lingue	I am talented at languages	6-point Likert	good_at_languages
ARCS Homework			
Nelle ultime 8 settimane, i compiti di inglese sono stati noiosi	In the last 8 weeks, the English homework has been boring	6-point Likert	homework_boring†
Nelle ultime 8 settimane, i compiti di inglese sono stati utili per migliorare il mio livello di inglese	In the last 8 weeks, the English homework has been useful for improving my level of English	6-point Likert	homework_improves_english
Nelle ultime 8 settimane, i compiti di inglese sono utili per ottenere conoscenze che mi serviranno in futuro	In the last 8 weeks, the English homework has been useful for gaining knowledge that will be useful in the future	6-point Likert	homework_future_knowledge
Nelle ultime 8 settimane, ho sentito di avere le capacità personali per portare a termine correttamente i compiti di inglese tutte le volte	In the last 8 weeks, I have felt that I personally had the abilities to properly complete the English homework every time	6-point Likert	ability_finish_homework
Nelle ultime 8 settimane, avrei avuto bisogno di più supporto e risorse per fare i compiti di inglese	In the last 8 weeks, I would have needed more support and resources to do the English homework	6-point Likert	need_support_homework†
Sono soddisfatto di quello che ho imparato facendo i compiti di inglese nelle ultime 8 settimane	I am satisfied with what I have learned from doing the English homework in the last 8 weeks	6-point Likert	satisfied_with_homework
ARCS Lectures and Contents			
Nelle ultime 8 settimane, i contenuti del programma di inglese non mi interessavano	In the last 8 weeks, the contents of the English program did not interest me	6-point Likert	uninterested_in_content†
Nelle ultime 8 settimane, le lezioni di inglese sono state utili a migliorare il mio livello di inglese	In the last 8 weeks, the English lessons have been useful in improving my level of English	6-point Likert	lessons_improved_english
Nelle ultime 8 settimane, le lezioni di inglese mi hanno fornito conoscenze utili per il mio futuro	In the last 8 weeks, the English lessons have provided me with useful knowledge for my future	6-point Likert	lessons_future_knowledge
Nelle ultime 8 settimane, ho sentito di avere le capacità personali per stare al passo col programma di inglese	In the last 8 weeks, I have felt that I personally had the abilities to keep up with the English program	6-point Likert	ability_keep_up

Continued on next page

Table 5 – *Continued from previous page*

Question (Italian)	**Question (English Translation)**	**Type of Answer**	**Label**
Nelle ultime 8 settimane, avrei avuto bisogno di più supporto e risorse per riuscire a stare al passo col programma di inglese	In the last 8 weeks, I would have needed more support and resources to keep up with the English program	6-point Likert	`need_support_to_keep_up`[†]
Complessivamente, sono soddisfatto di quello che ho imparato in inglese a scuola	Overall, I am satisfied with what I have learned in English at school	6-point Likert	`overall_satisfaction`
	Questions for Treatment Group Only		
Trovi che il tutor ti sia stato d'aiuto nello svolgere i compiti per casa?	Did you find the tutor helpful in doing homework?	Yes; No; No Access	`tutor_helped_homework`
Trovi che il tutor ti abbia aiutato a migliorare l'inglese ad un livello pratico?	Did you find that the tutor helped you improve English to a practical level?	Yes; No; No Access	`tutor_improved_english`
Trovi che avere accesso al tutor ti abbia aiutato a stare al passo col programma di inglese?	Did having access to the tutor help you keep up with the English program?	Yes; No; No Access	`tutor_helped_keep_up`
Vorresti continuare ad avere accesso al tutor in futuro?	Would you like to continue having access to the tutor in the future?	Yes, for all exercises; Yes, for some exercises; No; No access	`continue_with_tutor`
Quali aspetti del tutor hai trovato utili?	Which aspects of the tutor did you find useful?	Options: Personalized explanations; Feedbacks and Corrections on My Answers; Guidance through the Exercise Step by Step	`useful_tutor_aspects`
Spesso, il tutor mandava messaggi troppo lunghi anche quando non era necessario	Often, the tutor sent messages that were too long even when not necessary	6-point Likert	`tutor_long_messages`
Spesso, il tutor diceva cose non vere oppure mi diceva che sbagliavo anche quando la mia risposta era corretta	Often, the tutor said things that were not true or told me I was wrong even when my answer was correct	6-point Likert	`tutor_wrong_feedback`
	Final Comments		
Hai qualche altro commento sull'esperienza in generale?	Do you have any other comments on the general experience?	Free text response	`general_comments`

Table 6: Weekly Questionnaire

Question	Translation (English)	Type of Answer	Label
Week-level Questions			
Hai fatto, o provato a fare, almeno uno degli esercizi assegnati questa settimana? (rispondi sinceramente, non condivideremo la risposta con la tua insegnante)	Have you done, or tried to do, at least one of the exercises assigned this week? (answer honestly, we will not share the response with your teacher)	Yes; No	`completion`
I compiti per casa di questa settimana erano interessanti e/o stimolanti (rispetto ai compiti per casa prima dell'esperimento).	This week's homework was interesting and/or stimulating (compared to homework before the experiment).	6-point Likert	`interestingness`
I compiti per casa di questa settimana sono stati utili a migliorare il mio inglese ad un livello pratico.	This week's homework has been useful in improving my English to a practical level.	6-point Likert	`usefulness`
Exercise-Specific Questions - These questions are repeated for each exercise.			
Questo esercizio è stato utile a migliorare la mia comprensione e la mia conoscenza dell'argomento trattato.	This exercise was useful in improving my understanding and knowledge of the topic covered.	6-point Likert	`comprehensiveness`
Avevo a disposizione supporto e risorse a sufficienza per risolvere adeguatamente questo esercizio (spiegazioni, materiali, ...)	I had enough support and resources available to adequately solve this exercise (explanations, materials, ...).	6-point Likert	`level_of_resources`
(**Treatment-Group Only**) Quali aspetti hai trovato utili? (seleziona tutte le opzioni rilevanti)	(**Treatment-Group Only**) Which aspects did you find useful? (select all relevant options)	Options: Personalized explanations; Feedbacks and Corrections on My Answers; Guidance through the Exercise Step by Step	`useful_aspects_1`
Anomalies Feedback			
Il sistema ha avuto dei comportamenti anomali? (se sì, puoi descriverli brevemente?)	Did the system exhibit any abnormal behaviors? (if yes, can you briefly describe them?)	Open-ended	`anomalies`

A.1 Regression Tables

Variable	Coefficient	Std. Error	p-value
Intercept	4.6228	0.836	<0.001
Treatment	-1.1997	1.174	0.311
Words Typed	0.0019	0.001	0.009

Table 7: OLS Regression Results: Learning Gains Mediated by Student Engagement

B Prompts

B.1 Tutoring Prompt

The following prompt was the main prompt given to the tutor as the system prompt. We replaced assignment purpose, description and example with the content provided from the teacher. Note that the Common European Framework of Reference for Languages was mistakenly referred to as "Cambridge Framework" during the execution of the study.

```
We are helping students learn english as a second language.

We give you an exercise as a starting point. Act as a
tutor and drive the student through the same concepts, testing
the understanding step by step. It is not necessary to replicate
to the example exercise, as long as you cover the same concepts.
Follow the tutoring strategy we provide.

Start with a brief explanation of the concept.
Provide at least 10 questions, one by one.
Do not move on until the student gives the correct answer.
If necessary, provide explanations and feedback.
Never give the answer to the question.
Do not give the answer to the question as a part of the
explanation.
Point out all grammar and spelling mistakes.
Keep a B2 level of English according to the Cambridge framework.

Once all questions are solved, ask the student if they
wish to practice more. If they don't, output <COMPLETE>

EXERCISE PURPOSE:
+++ purpose +++

EXERCISE DESCRIPTION
+++ description +++

EXERCISE EXAMPLE (NOT VISIBLE TO THE STUDENT)
+++ example +++

TUTORING STRATEGY
+++ strategy +++
```

B.2 Strategy Generation Prompt

The following prompt was used to generate the tutoring strategy.

```
We are helping students learn english as a second language.

We give you an exercise as a starting point.
Provide a concise, step-by-step strategy for a short
dialog-based tutoring session led by ChatGPT with a student
covering the same concepts.
The tutoring session is text-based and led through a chat
interface.
Describe the strategy with a maximum of six sentences.
Keep a B2 level of english according to the Cambridge
framework.
+++ assignment.purpose +++
+++ assignment.description +++
+++ assignment.example +++
```

C Additional tables and figures

C.1 Novelty effects

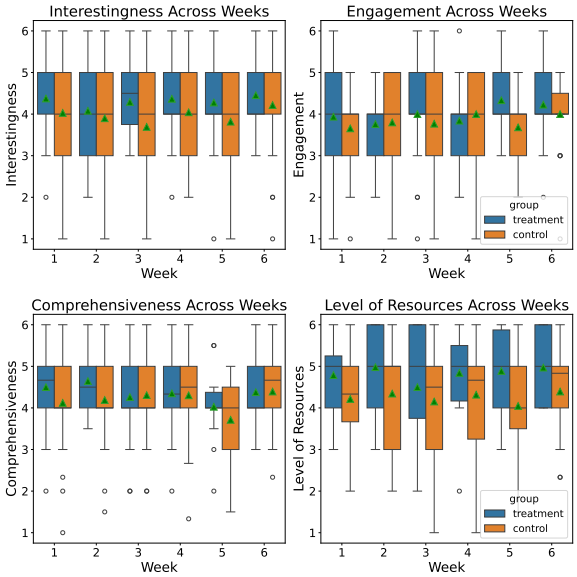

Figure 3: Weekly distribution of student ratings for *usefulness*, *interestingness*, *comprehensiveness* and *level_of_resources*. Green triangles show means