Are Large Language Models Good Essay Graders?

by

Anindita Kundu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

We evaluate the effectiveness of Large Language Models (LLMs) in assessing essay quality, focusing on their alignment with human grading processes. Specifically, we investigate the applicability of LLMs such as GPT-3.5T and Llama-2 in the Automated Essay Scoring (AES) task, a crucial natural language processing (NLP) application in education. Our study explores both zero-shot and few-shot learning approaches, employing various prompting techniques to enhance performance. Utilizing the ASAP dataset, a well-known dataset for the AES task, we compare the numeric grade provided by the LLMs to human rater-provided scores. Our research reveals that both approaches GPT-3.5T and Llama-2 generally assign lower scores compared to those provided by the human raters. Furthermore, neither LLM correlates well with the human scores. In particular, GPT-3.5T tends to be harsher and further misaligned with human evaluations compared to Llama-2. On the other hand, both LLMs not only can reliably detect spelling and grammar mistakes but also seem to take those mistakes into account when computing their score. Additionally, we extended our analysis to include the most recent release, Llama-3, which shows promising improvements in alignment with human scores. This suggests that newer generations of LLMs have the potential to be more effective in AES tasks. Overall, our results offer a cautiously optimistic view of using LLMs as tools to assist in the grading of written essays, highlighting both their current limitations and their future potential.

To Maa, Baba and Divai

For unwavering support and encouragement in all my endeavours.

Acknowledgements

I would like to extend my deepest gratitude to my supervisor, Prof. Denilson Barbosa, for his unwavering support and constant guidance throughout the completion of this thesis. His insightful advice has been invaluable in shaping my research and helping me grow as a researcher. His meticulous attention, constructive criticism, and belief in my abilities have been a continuous source of motivation, leading to more meaningful and impactful work.

I am also profoundly thankful to my thesis committee members, Prof. Carrie Demmans Epp and Dr. Guher Gorgun, for their valuable time and effort in thoroughly reviewing my work. Their insightful suggestions have significantly contributed to the refinement of this thesis.

I am immensely grateful to DeepMind and the Department of Computing Science for nominating me for the scholarship, whose generous financial support made this research possible. I am eternally grateful to my parents and sister for their unwavering trust and belief in me. Their love and support have carried me through the toughest times.

A special thanks goes to each and every friend of my "We are" group. God brought us together when I needed you the most, and your friendship has been a blessing beyond words. Thank you, Subho, Arghasree, Kush, Aakash, Dhruv, and Sonali, for always listening to me, and supporting me through my struggles. Your encouragement, motivation, and friendship have been invaluable, and I could not have achieved this milestone without you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Essay writing is a common component of student assessment, playing a pivotal role in education by providing insights into the text comprehension, critical thinking, and communication skills of the students [1]. Writing good essays requires students to articulate their thoughts clearly and coherently, demonstrating their understanding of a subject matter and their ability to construct logical arguments [2]. Moreover, effective essay assessment not only measures the knowledge and the skills students have acquired but also encourages deeper learning and engagement with the material [3].

1.1 Challenges in Human Evaluation

Traditionally, essay grading has been mostly done by human graders. However, this presents significant challenges in modern education settings, especially when it comes to distance education, which contributes to the democratization of learning [4]. Moreover, the global teacher shortage is a real and growing crisis. Countries struggle to find enough qualified educators to meet the needs of their student populations, resulting in ever-increasing workloads

for teachers [5]. When teachers are responsible for large numbers of students it becomes difficult to provide individualized attention and detailed feedback for each essay. High student/teacher ratios can lead to burnout and decreased quality of evaluations, as teachers are pressed for time and may not be able to devote the necessary effort to each student's work.

Another significant challenge is the time-consuming nature of human evaluation. Reading, understanding, and providing constructive feedback on essays is a labor-intensive process that requires considerable time and effort [6]. This is especially problematic during peak periods, such as midterms or finals when teachers are inundated with grading responsibilities. The extensive time required for thorough evaluation can delay feedback, hindering students' ability to learn from their mistakes and improve their writing skills in a timely manner [7].

Finally, even setting aside the time pressure involved in essay scoring, one cannot forget that grades assessed by humans are subject to cognitive biases and stereotypical beliefs (intentionally or unintentionally), which can lead to inconsistent evaluations across different essays [8]. Evaluators may have unconscious biases based on their personal preferences, cultural background, or prior experiences, which can influence their judgment [9]. Factors such as handwriting or familiarity with the student can also skew the assessment. These biases can lead to significant variability in scores, making it difficult to ensure that all students are fairly evaluated on the same level playing field.

1.2 Overview of Automated Essay Scoring

In response to the limitations and scarcity of human graders, Automated Essay Scoring (AES) has emerged aimed at automatically grading and assessing the quality of written essays, thereby streamlining the assessment process and offering consistent scores to students. Automated essay scoring (AES) is a computer-based assessment system that automatically

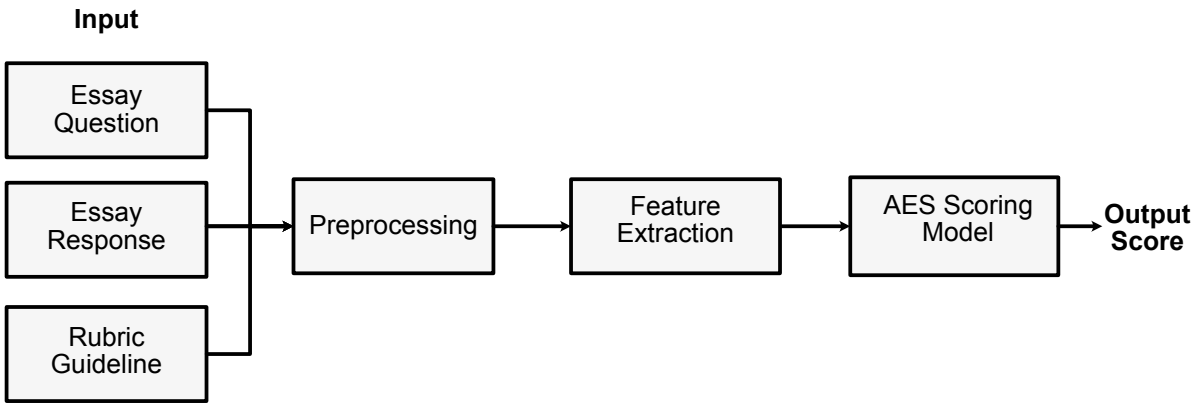

Figure 1.1: Automated Essay Scoring Pipeline

scores student-generated content without manual intervention. The development of AES systems has coincided with pivotal advancements in computational linguistics and Natural Language Processing (NLP) technology. AES relies heavily on NLP techniques at its core because it helps to extract deep meaningful language features that are indicative of writing quality. AES is an essential educational application of NLP which consists of having an AI agent assign a score to an essay based on its content, organization, and overall quality.

The AES process generally involves several phases: data preprocessing, feature extraction, model training, and essay scoring, as illustrated in Figure 1.1. During data preprocessing, the essay is cleaned and standardized to ensure consistency. Feature extraction involves identifying relevant attributes, such as vocabulary usage, sentence structure, and coherence, which are then used to train the scoring models. Model training involves using labelled datasets of scored essays to teach the AES system how to grade new essays. Once the AES system is trained, it can be used to score unseen essays.

To date, AES has been approached primarily as a supervised learning task in which models are learned from sample essays graded by humans. Common techniques to solving the AES task include rule-based classification, feature-based statistical machine learning or

deep neural network model training, and fine-tuning pre-trained models [6]. There is an underlying assumption in this approach, which is that the human graders correctly graded the essays according to the rubric. Under this assumption, a reasonable AES method learns how to emulate the scoring as done by the human graders and captured in the training data.

Although these developments are encouraging, there are several key technical challenges in applying these methods. Machine learning AES algorithms are heavily dependent on features selected by humans or feature engineering. In contrast, deep learning AES algorithms are often perceived only to generalize well when there is a large training sample available. Fine-tuning a pre-trained model often requires a lot of resources. Also, for different domain/ subject applications, we need to train different models which involves additional expense. Additionally, AES systems must be robust enough to handle diverse writing styles and topics. The expectations of the AES system are frequently out of sync with the actual material due to the vast diversity of student-generated content [10].

1.3 Goal of the thesis

Pre-trained Large Language Models (LLMs) are trained using highly parallel architectures in an unsupervised fashion using massive amounts of data [11]. LLMs are capable of processing and understanding human language better than any kind of model before. Advancements in computing, the availability of larger datasets, and the improvement of machine learning algorithms have brought these language models closer to human-level performance on a multitude of tasks. LLMs learned from massive corpora have been shown to perform remarkably well on various language tasks, several of which they were not explicitly trained to do. For this reason, LLMs are described and evaluated as zero-shot (or few-shot) learners in the recent literature [12]. Recent results provide compelling arguments demonstrating that these

models can generalize to unseen tasks and perform astonishingly well on complex tests such as the bar exam [13].

Traditional AES systems need a large number of labelled domain-specific datasets for training and/or fine-tuning. The requirement for domain-specific fine-tuning, however, can be lessened by using LLMs because it has in-context learning capability [12]. Different prompt engineering tactics can be utilized to bring their maximum potential to any low-resource learning tasks. Prompt engineering is the art of communicating with a generative large language model by writing precise and task-specific instruction prompts for improving the performance of LLMs [14]. LLMs with proper prompts have been shown to adapt to various tasks without fine-tuning. Moreover, LLM performance can be enhanced by showing several examples of input and expected outcomes in the prompt which is known as a few-shot learning capability [15]. Additionally, because of their ability to comprehend and produce human-like language, LLMs can provide explanations of how they arrived at a grade which can be used to help students understand their grade. For this reason, LLMs are robust candidates for the development of AES systems.

In this study, our purpose is to examine the performance of two popular generative large language models OpenAI's GPT-3.5T (ChatGPT-3.5-Turbo) and Meta's Llama-2 [16] as automated essay scoring tools. We used version 3.5 of ChatGPT, also known as InstructGPT [17], which was obtained by refining the original GPT-3 [18] through reinforcement learning guided by actual human feedback by OpenAI engineers, resulting in a model that was far superior in the understanding of instructions. While it is not entirely clear how or under which conditions language models like GPT-3.5T and Llama-2 acquire the ability to successfully complete NLP tasks without training [11], the advent of LLMs has opened up the door for a different kind of AES system that does not rely on expert training data: providing a prompt requesting an essay to be graded given the rubric. We investigate this

in-context learning approach, guided by a fundamental question: Can AES based on LLMs match human graders in judging the quality of essay writing? Related to that question, we revisit the role of annotated training data for the AES task and consider whether LLMs can act as zero/few-shot classifiers to assess essay quality and explain their understanding reasonably behind the score they provide. In particular, our study aims to address the following research questions:

- **RQ1:** Do human scores align with the LLM scores?
- **RQ2:** What are the possible reasons behind the similarity/difference in scores?
- **RQ3:** Do LLMs offer explanations in a tone that reflects their scores?
- **RQ4:** Can LLMs correctly identify spelling and grammatical errors and reflect those into their scoring?

1.4 Outline

The rest of the dissertation is structured as follows. Chapter 2 provides a concise review of related research. In Chapter 3, we present detailed methodology about the dataset in subsection 3.1, along with the experimental setup in subsection 3.2. Chapter 4 is dedicated to the presentation and analysis of the results. In chapter 5 reiterates through all our analysis with one of the recent LLMs Llama-3. Chapter 6 is an extension and additional experiment results with different prompts. Finally, in Chapter 7, we offer our conclusions and discuss our limitations with potential avenues for future enhancements.

Chapter 2

Literature Review

2.1 Human Essay Scoring

Human raters possess diverse skills in evaluating assignments, including understanding the meaning of the text, assessing critical thinking, creativity, and content relevance [19]. They excel in evaluating logic, argument quality, and factual correctness, and can judge audience awareness [20]. However, maintaining consistency and eliminating subjectivity can be challenging [20]. Additionally, large-scale essay scoring can be labor-intensive and time-consuming [21], which in turn can lead to fatigue and inconsistencies by human scorers. With the increasing demand for personalized education and the growing shortage of teachers, Automated Essay Scoring (AES) systems are increasingly needed. AES can assist in managing large classes by providing consistent and efficient essay assessments. This capability can alleviate some of the workload from educators, allowing them to focus on more interactive and engaging aspects of teaching.

2.2 Automated Essay Scoring without LLMs

The first automated essay scoring system was developed more than 50 years ago, in 1966 [22]. Since then, these systems have become more advanced, offering more features than the early versions. There are many good surveys of AES systems, such as Ramesh and Sanampudi [6] and Ke and Ng [23]. Most AES research focuses on supervised learning of holistic scoring due to the availability of annotated corpora and their commercial value in automating standardized test grading. Following a similar evolution as other fields of AI, AES systems evolved from using hand-crafted rules in PEG (Project Essay Grade) [24] to features-based statistical machine learning such as E-raters [25], Intelligent Essay Assessor (IEA) [26] to using deep neural models such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) that learn own representations of the data [27, 28]. We also see approaches which learn from transfer learning for domain adaptation [29], and a combination of hand-picked features and deep learning approaches [30].

As argued by Ke and Ng [23], the most important features for evaluating essay quality are grammaticality, organization, persuasiveness, coherence, mechanics, and relevance. It should be noted that extracting these features from text is difficult as even state-of-the-art NLP tools to assess these aspects of running text cannot achieve human-level performance. Therefore, it becomes challenging to train models to holistically evaluate an essay according to all of these dimensions together. It is worth mentioning that the pre-LLM state-of-the-art in NLP was far from capable of addressing most of these criteria, especially those related to higher-level notions of writing quality.

2.3 Automated essay scoring with LLMs

The advent of pre-trained large language models represents a major step forward in NLP allowing more generalized language representations from vast corpora that can then be fine-tuned for a specific task. Of course, LLMs have been used for the AES task. Wang et al. [31] introduces a joint multi-scale essay representation using BERT employing multiple losses and transfer learning to demonstrate effective generalization to long-text AES tasks. Yang et al. [32] focus on fine-tuning LLMs such as BERT utilizing regression and ranking. Khademi [33] evaluated GPT-3.5T and Google Bard against human scores on IELTS academic writing tasks and found that these models are far from achieving human performance. However, their study did not focus in detail on why there are vast differences between the scores. Mizumoto and Eguchi [34] leveraged GPT-3 version *text-davinci-003* in a zero-shot setting using non-native English corpus (TOFEL11) highlighting the potential use of ChatGPT incorporated with linguist features can enhance accuracy in AES. Latif and Zhai [35] fine-tuned GPT-3.5T to confirm the effectiveness in education-specific tasks by significantly improving score prediction over BERT (on average 10.6%). They neither fully exploit few-shot learning approaches nor utilize any prompt clarity-enhancing techniques with zero-shot.

2.3.1 Prompt engineering

While fine-tuning has greatly improved the performance of LLMs across tasks, including AES, this approach require considerable amounts of labelled data, which is a problem. Prompt engineering approaches become useful in this context as they require less data and allow for enhancing in-context learning and maximizing efficient outcomes. One recent work by Mansour et al. [36] investigated GPT-3.5T and Llama-2 with four different prompts and found that LLMs are highly sensitive to the prompt. This study showed with the right

prompt engineering, LLMs can offer competitive albeit not yet superior performance compared to state-of-the-art models. However, their study did not report any single prompt type using which LLMs can consistently perform close to human scoring across all essays. Another recent work by Helmeczi et al. [37] explores few-shot learning using BERT for AES tasks.

Offering feedback is crucial for students to learn from their mistakes and develop their writing skills. However, delivering timely and valuable feedback, particularly to middle and high school students, can pose a significant challenge for educators. Many recent works trained models to provide scores for individual essay traits as feedback to improve students' writing performance [38, 39]. However, it is not adequate to know until students know why that specific trait score is low and how each trait score impacted the overall score. Generative LLMs such as GPT-3.5T and Llama-2 may assist teachers in this case. Pankiewicz and Baker [40] conducted an experimental study integrating GPT-3.5 into an automated assessment platform to generate personalized hints for programming assignments, finding that while GPT-generated hints positively impact education without harming student state, they do not significantly improve performance. Another study by Meyer et al. [41] examines the impact of large language model-generated feedback on upper secondary students' writing tasks, revealing significant improvements in revision performance, task motivation, and positive emotions compared to no feedback.

2.4 Differences to previous work

In this study, we aim to investigate the alignment between human grades and grades assigned by LLMs in AES tasks, assessing the extent to which LLM-based AES can potentially substitute human raters in evaluating essay quality. We will explore the possible reasons underlying

the similarities or differences in grades assigned by humans and LLMs. Furthermore, we seek to determine whether LLMs possess the ability to accurately identify and assess spelling and grammatical errors, and grade essays accordingly. Finally, we aim to evaluate the feasibility of LLMs serving as assistants by generating rich human-readable feedback to explain the grades they provide, thus potentially alleviating the workload of educators in providing timely and constructive feedback to students.

Chapter 3

Methodology

3.1 Dataset

We used the Automated Student Assessment Prize (ASAP) benchmark [42], well-known for the Automated Essay Scoring (AES) task. It consists of around 13000 essays written by students of grade levels from 7 to 10. There are eight sets of tasks in ASAP, each linked to different prompts and scoring ranges. Detailed statistical information about this dataset is shown in Table 3.1.

Students wrote essays in response to three types of prompts. First, in an **argumentative** essay, the student must write to convince the reader about their opinion on a topic. They need to research the topic, find evidence, and explain their ideas clearly. Second, in a **narrative** essay the student must write a story. Here they can make up characters or use personal experience, and events to create an interesting tale. Finally, in a **source-dependent** essay the student needs to read an article and then write an essay using the information from that article. In the ASAP dataset, **Tasks** 1 and 2 asked for argumentative essays, **Tasks** 3 to 6 asked for source-dependent essays, while **Tasks** 7 and 8 concerned

narrative essays.

Given practical limitations related to the length of the prompt accepted by LLMs at the time of the research, we ruled out the source-dependent tasks. Between the two argumentative tasks, we selected **Task 1** as it had a single domain holistic score which allowed for an easier evaluation of the results. We selected **Task 7** from the two narrative tasks as it has more data samples and its grade level is closer to Task 1.

In this dataset, two to three human graders annotate each essay. The final score of an essay is the sum of scores given by the individual human graders.

Table 3.1: Detail information about ASAP dataset

Task	Essay Type	Number of Essays	Score Range	Grade Level
1	Argumentative	1783	1-6	8
2	Argumentative	1800	1-6 or 1-4	10
3	Source dependent	1726	0-3	10
4	Source dependent	1772	0-3	10
5	Source dependent	1805	0-4	8
6	Source dependent	1800	0-4	10
7	Narrative	1569	0-15	7
8	Narrative	723	0-30	10

3.1.1 Kinds of Scores

In the ASAP dataset, Tasks 1 to 6 provide a single (holistic) score that reflects the graders' assessment of the essays with respect to the entire rubric. Tasks 7 and 8, on the other hand, provide scores for specific essay traits like content, organization, and style, while the remaining six prompts offer only overall scores. Two human raters graded both of these tasks.

Task 1 asked for students' opinions on the usefulness of computers and was evaluated holistically starting from 1 to 6 scores. The rubric for this task describes six different score

specifications. Conversely, Task 7 focused on students' experiences with patience and was assessed based on four trait scores: ideas, organization, style, and convention. The rubric for this task is more intricate, with requirements describing each trait score ranging from 0 to 3. The score for trait score for ideas needs to be doubled, yielding a final score range of 0-15. Appendix A includes a detailed rubric for Task 1 and Task 7.

3.1.2 The ASAP++ dataset

In 2018, Mathias and Bhattacharyya [43] introduced ASAP++, a dataset with manual annotations for the other six prompts with respect to the particular traits. One proficient English annotator graded each essay, scoring from 1 to 6 for independent attributes. We used this extended version of the ASAP dataset mentioned previously [1]. The traits for Task 1 are ideas and content, organization, word choice, sentence fluency, and conventions which are also annotated by the human graders. Table 3.2 shows the details of the trait score rubric used for grading the essays in Task 1 of ASAP++, which we used in our study. It is worth mentioning that ASAP (and ASAP++) are *de facto* the best and largest AES datasets widely available today.

3.2 Experimental Setup

3.2.1 Prompt design and response generation

We consider two popular LLMs for response generation: GPT-3.5T and Llama-2. Specifically, we consider the versions *gpt-3.5-turbo-0613*, and *Llama-2-2-70B-chat* for our experiment. From the ASAP dataset, we input the prompt given to students, their corresponding

[1] https://lwsam.github.io/ASAP++/lrec2018.html

Table 3.2: Traits of ASAP++ dataset for Task 1

Traits	Scoring Requirements
Ideas & Content	Assesses the clarity, depth, and engagement of the essay's content, focusing on the presence of clear main ideas, well-supported arguments, and relevance to the audience and purpose.
Organization	Evaluates the structure and flow of the essay, looking for a logical sequence of ideas, well-defined paragraphs, and overall organization suitable for the given task.
Word Choice	Examines the use of vocabulary to convey the message effectively, emphasizing the selection of precise, impactful words, and their appropriateness for the intended audience and purpose.
Sentence Fluency	Considers the writing's flow and rhythm, assessing the variety and structure of sentences, and their ability to engage the reader while maintaining clarity and coherence.
Conventions	Focuses on the writer's command of standard writing conventions, including punctuation, spelling, capitalization, grammar, and usage, with an emphasis on minimizing errors that may disrupt readability and communication.

essays, rubric guidelines, score range, and any additional instructions into the GPT-3.5T and Llama-2. The rubric and guidelines are presented in the same format and detail as they would be for human raters, mirroring how they were described in the dataset. Appendix A includes examples of prompt and GPT-3.5T or Llama-2-generated response. Next, we ask LLMs for a numeric score in the appropriate range and an explanation for Task 1. Task 7 has four trait scores and a total score along with an explanation from GPT-3.5T, Llama-2. We make sure that the entire input to GPT-3.5T, Llama-2 (prompt + actual essay + output) never exceeds its token limit (4096 context length as of December 2023). While generating responses, we employed default parameter settings to maintain balance and maximize the creativity and diversity of LLMs when solving a complex task like AES.

3.2.2 Extracting features from essays

To further analyze how the grading process of LLMs coaligns with humans we extract a wide range of features from essays used in recent AES methods.

- Essay statistics: We have extracted several basic statistical features for each essay such as the number of sentences and the number of tokens found in the essay.

- Readability: Next, we do a readability assessment of each essay to determine if implicitly human graders follow or if LLMs are pre-trained to grade a text based on its difficulty. We use several well-known readability formulas to measure how difficult an essay is to read and understand e.g., Flesch reading ease [44], Automated Readability Index [45], Coleman Liau index [46], Dale–Chall readability score [47], Flesch–Kincaid grade [44], Gunning fog [48], Linsear write formula, and SMOG index [49]. We have followed the implementation of text readability formulas from the paper by Martinc et al. [50].

- Linking words: The more features that we have extracted from essays include the use of the total and unique number of transition words[2] and "FANBOYS"[3] words. Transition words serve as a proxy for the number of arguments in the essay. Transition words and Fanboys (coordinating conjunctions) serve similar purposes in English writing by connecting ideas and improving the flow of text. Transition words, such as "however" or "therefore," indicate relationships between sentences or paragraphs, aiding in smooth transitions and clarifying the logical progression of ideas. Fanboys, on the other hand, specifically join clauses within a sentence, emphasizing relationships like addition ("and"), contrast ("but"), choice ("or"), or conclusion ("so"). Both types

[2] https://www.grammarly.com/blog/connecting-sentences/
[3] For, and, but, or, yet, and so.

of words enhance coherence and cohesion in writing, helping readers navigate complex texts more effectively.

- Aspell: We have utilized *Aspell*[4], an open-source spell-checker which offers an improved and comprehensive English language dictionary. *Aspell* accommodates American, British, and Canadian spelling preferences encompassing multiple spellings for a single word. Our use of the *Aspell* enabled us to find the number of misspelled words in each essay. We have also kept a record of all the misspelled words reported by *Aspell* for every essay.

- LanguageTool: Additionally, to cross-check the mention of mistakes in LLM's feedback, we have used an open-source language checking tool *LanguageTool*[5] similar to Zesch et al. [51] work. It helps to detect issues related to spelling, grammar, punctuation, style, and more. We have utilized specifically version 6.1 which can identify the following language errors:

  - **Typography** errors refer to mistakes related to the visual appearance of text, including font choice, spacing, and formatting. Example: *tHeRe aRe tYPoGraPhIcaL eRRorS iN tHis SeNtEncE.*

  - **Typos** are simple spelling mistakes or accidental keystrokes that result in incorrect words. Example: *I would like a peice of pie (should be piece).*

  - **Punctuation** errors involve incorrect placement or misuse of punctuation marks. Example: *Sharon and Sue, went into the office early yesterday, to complete a project* (should be without the unnecessary comma).

  - **Grammar** errors encompass mistakes related to sentence structure, verb tenses, subject-verb agreement, etc. Example: *He go to the store yesterday.*

[4] http://aspell.net/
[5] https://languagetool.org/

- **Casing** errors related to the incorrect use of uppercase or lowercase letters. Example: *I live in new york.*
- **Redundancy** errors occur when unnecessary words or phrases are repeated. Example: *He personally went there himself.*
- **Confused words** errors where words are confused with others that sound similar or have similar spellings. Example: *Their going to the park later.*
- **Style** errors relate to writing style consistency and adherence to specific writing conventions. Example: *I sometimes am happy.*
- **Miscellaneous** errors cover various other language issues not falling into specific categories. Example: *This is is just an example sentence.*
- **Compounding** errors related to compound words, whether they should be separate words, hyphenated, or combined. Example: *I'm working full time (full-time).*
- **Collocations** errors where words are used together incorrectly or inappropriately. Example: Open your books at (to) page 6.
- **British English** errors involve using British spelling or vocabulary inappropriately in American English contexts. Example: *Colour (British) vs. color (American).*
- **Nonstandard phrases** errors involve using unconventional or informal expressions. Example: *I never have (I have never been) been to London.*

We have further classified these error categories into five main groups for simplification. Compounding and spelling errors fall under the spelling category. Collocation, nonstandard phrases, confused words, miscellaneous errors, and grammar issues are grouped together as grammar errors. Typography, redundancy, style, and British English errors form the style category. Casing and punctuation constitute their own group.

3.2.3 Extracting features from LLM explanation

We have gathered various fundamental statistics for each LLM explanation, including the count of sentences and tokens. Additionally, we employed sentiment analysis to assess the tone of the explanation and implemented rule-based methods to extract specific information from the LLM explanation, as described below:

- **Sentiment Analysis using VADER**

  VADER (Valence Aware Dictionary and sEntiment Reasoner) is a sentiment analysis tool designed for social media content. It combines a lexicon-based approach with rule-based techniques to assess sentiment polarity (positive, negative, or neutral) and intensity. Specifically attuned to social media language, but it is also generally applicable to sentiment analysis in other domains.VADER provides several sentiment scores:

  1. Compound score: This overall sentiment score ranges from -1 (extremely negative) to 1 (extremely positive). It captures the overall sentiment of the text.
  2. Positive score: Indicates the proportion of positive words in the text.
  3. Negative score: Reflects the proportion of negative words.
  4. Neutral score: Represents the neutrality of the text.

  We extract all these four sentiment scores from VADER for each GPT-3.5T and Llama-2 response. Then, we split each response into individual sentences and asked for a polarity score to determine the maximum sentence-wise compound sentiment score.

- **Rule-based matching** We utilize Spacy's DependencyMatcher [6] to extract mention of grammatical, spelling, punctuation, and capitalization mistakes from the LLM's ex-

[6] https://spacy.io/api/dependencymatcher

planation. Dependency matching involves identifying and extracting linguistic patterns based on the syntactic dependencies between words in a sentence. The DependencyMatcher in spaCy provides a flexible and intuitive framework for designing custom patterns for capturing specific syntactic structures in text.

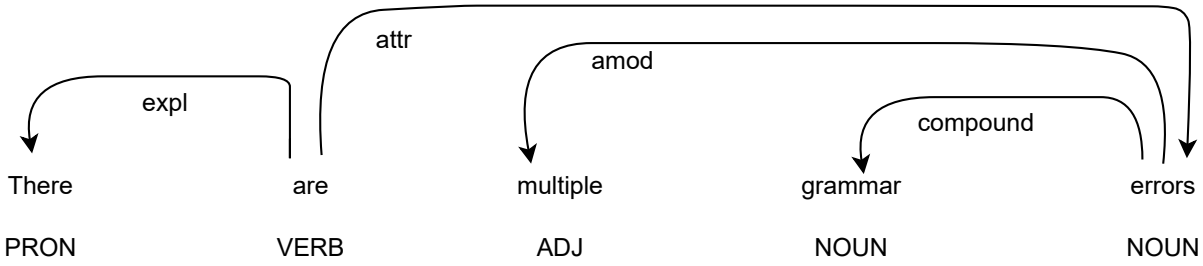

Figure 3.1: Dependency Tree

In the DependencyMatcher of spaCy, a pattern is a list of descriptions for tokens to be matched in a sentence's dependency structure. The pattern is constructed using dictionaries, each detailing particular conditions for matching tokens. The first dictionary sets an anchor token and the rest of the dictionaries are defined based on the relationship with the previous node. This structure allows for precise specification of syntactic dependencies, aiding in capturing complex linguistic constructs for natural language processing tasks.

We randomly selected 60 samples from the responses generated by GPT-3.5T to manually hand-craft patterns. For each sample, we identify a list of patterns that capture the mention of grammatical, spelling, punctuation, and capitalization mistakes and also mentions of any qualifier (such as *numerous, several, multiple*). For instance, Figure 3.1 shows a sample sentence from an explanation generated by ChatGPT: *There are multiple grammar errors in the essay.* This sentence indicates multiple grammatical errors in the essay evaluated. We used this similar dependency tree as a reference to hand-craft the pattern. The dependency tree of this sentence shows a compound

relationship between tokens *errors* and a type of mistake which is *grammar*. Also, another relationship between the token *errors* and a qualifier token *multiple*. To write the pattern for this sentence, we defined *'errors'* as an anchor token and looked at the left and right of the anchor token to extract mentions of mistakes surrounding that token. For this case, they are *grammar* and *multiple*.

In GPT-3.5T explanation, there could be mention of different kinds of mistakes together in a sentence. For example consider the sentence: *There are grammatical and spelling mistakes in the essay.* In this sentence, the token *grammatical* is conjoined with another token *spelling*. To represent such a conjunct relationship, we curated patterns to handle up to five different mentions as conjunction. We used the same patterns for both language models (GPT-3.5T and Llama-2). This task is tedious and requires additional rules curation for new sentence adaptation.

Furthermore, we have leveraged the ASAP++ dataset to compare the trait scores of Task 1 to LLM's scores. Correlation analysis has been performed between human scores and their various traits, LLM-provided scores, misspelling count, and all other related features. Our primary focus centers on convention scores of the ASAP++ dataset, as there are currently no other available NLP tools for validating the remaining four criteria [23]. Convention scores are derived from assessments related to punctuation, spelling, grammar, and capitalization. Notably, the explanation provided by GPT-3.5T and Llama-2 often mentions spelling and grammar issues. Consequently, we opt for these two metrics to assess the quality of LLM's provided explanation.

3.3 Evaluation Metric

In our analysis, Pearson correlation analysis is employed to measure the agreement or similarity between the scores assigned by human raters and those generated by LLMs. The Pearson correlation coefficient is denoted as r ranges from -1 to 1, where:

- $r = 1$ indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases proportionally.
- $r = -1$ indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases proportionally.
- $r = 0$ indicates no linear correlation between the variables.

we classified the magnitude of Pearson correlation (r) values to assess the strength of relationships between variables as follows: very weak (0.0 to 0.19), weak (0.2 to 0.39), moderate (0.4 to 0.59), strong (0.6 to 0.79), and very strong (0.8 to 1.0).

Chapter 4

Results

4.1 RQ1: Do human scores align with the LLM scores?

For Task 1, both GPT-3.5T and Llama-2 assigned valid overall scores ranging from 1 to 6 to all 1783 essays. Table 4.1 presents the descriptive statistics for the four scoring methods, while Figure 4.1 illustrates the score distribution. It is notable that LLMs tend to assign lower scores compared to human raters.

Table 4.1: Descriptive statistics for each scoring method of Task 1

	Rater 1	**Rater 2**	**GPT-3.5T**	**Llama-2**
Mean	4.26	4.27	1.90	3.62
SD	0.84	0.82	0.56	0.66
Min	1.00	1.00	1.00	1.00
Median	4.00	4.00	2.00	4.00
Max	6.00	6.00	5.00	4.00

In Task 7, all the LLMs generated four trait scores, along with a final overall score that sums all trait scores for a total of 1569 essays. Table 4.2 presents the descriptive statistics for the five grading methods. GPT-3.5T ($M = 2.07, SD = 1.83$) and Llama-2 ($M = 2.66, SD = 2.68$) both assigned significantly lower mean scores than both human

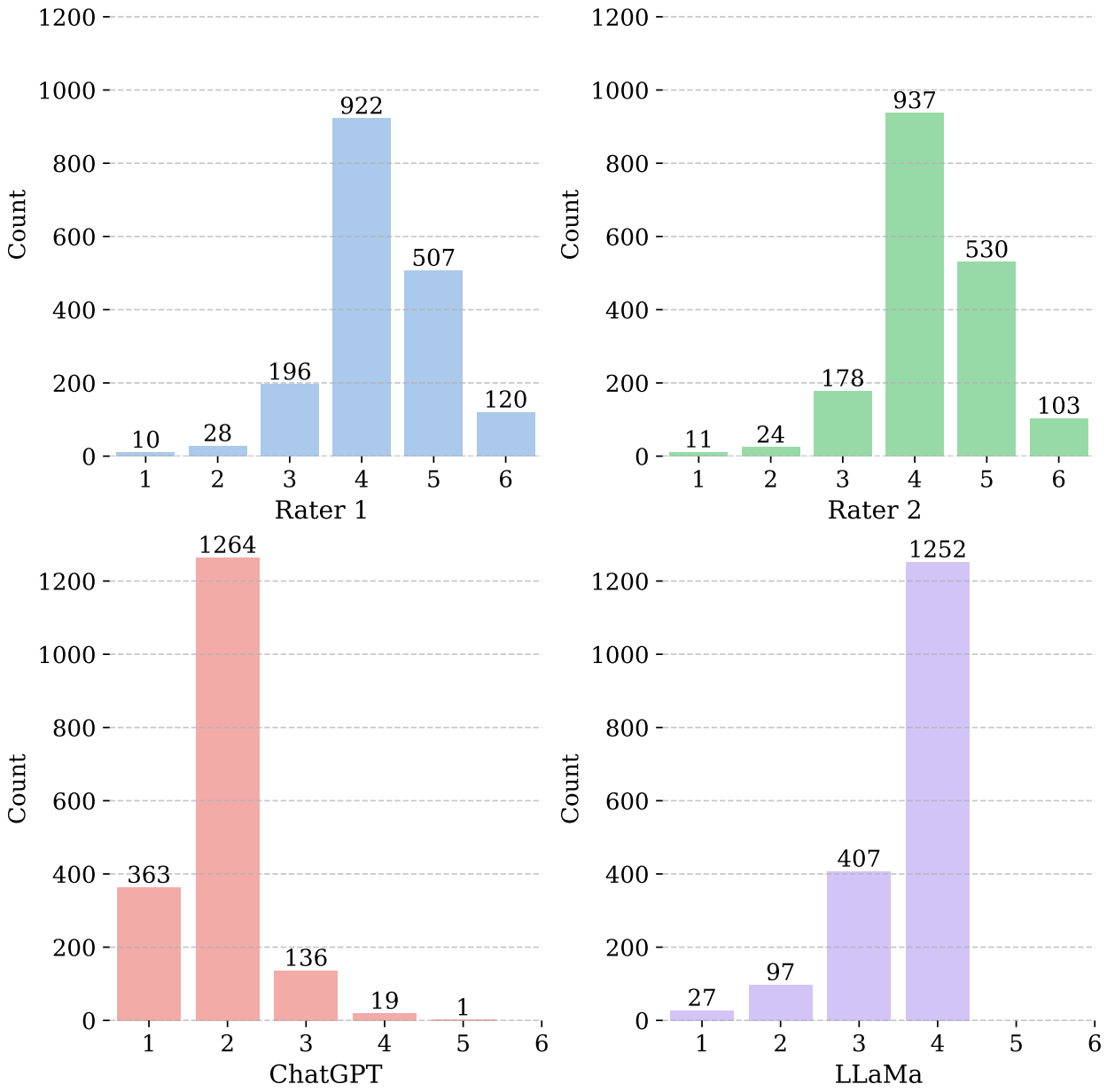

Figure 4.1: Score distribution of human raters and LLMs for Task 1

raters ($M = 8.02, SD = 2.42$ and $M = 8.04, SD = 2.52$) revealing a trend similar to that observed in Task 1 where LLMs tended to assign lower scores than human raters.

Notably, GPT-3.5T and Llama-2 assigned out-of-range trait scores to 19 and 16 samples, respectively. This discrepancy may be attributed to the complexity of Task 7's rubric com-

Table 4.2: Descriptive statistics for each scoring method of Task 7

	Rater 1	Rater 2	GPT-3.5T	Llama-2
Mean	8.02	8.04	2.07	2.66
SD	2.42	2.52	1.83	2.68
Min	0.00	0.00	0.00	0.00
Median	8.00	8.00	2.00	3.00
Max	12.00	12.00	9.00	26.00

pared to that of Task 1. Figure 4.2 illustrates the score distribution for this task, excluding the 39 outlier samples when constructing the score distribution graphs. Detailed trait-wise score distribution and statistics are shown in Appendix B.

4.1.1 LLM and human scores do not correlate

Tables 4.3 indicates that GPT-3.5T scores show a weak positive correlation with the scores provided by both human raters ($r_{\text{rater1}} = 0.23, p < 0.001, d^1 = -3.30$ and $r_{\text{rater2}} = 0.21, p < 0.001, d = -3.38$) in Task 1. While Llama-2 scores moderately positively correlate ($r_{\text{rater1}} = 0.59, p < 0.001, d = -0.85$ and $r_{\text{rater2}} = 0.58, p < 0.001, d = -0.87$) with human scores exceptionally for Task 1. Also, we observed no strong inter-LLM correlation ($r = 0.32, p < 0.001, d = 2.8$) between Llama-2 and GPT-3.5T.

Table 4.3: Correlations between human rater and LLM scores for Task 1

Score	Rater 1	Rater 2	GPT-3.5T
GPT-3.5T	0.23	0.21	1.00
Llama-2	0.59	0.58	0.32

A similar trend is observed in Table 4.4 for GPT-3.5T in Task 7, there is negligible positive correlation across all traits with human raters ($r_{\text{rater1}} = 0.07 - 0.18, p < 0.001, d_{\text{overall}} = -2.77$ and $r_{\text{rater2}} = 0.10 - 0.19, p < 0.001, d_{\text{overall}} = -2.71$). However, Llama-2 shows a different trend than in Task 1. In Task 7 Llama-2's correlation also weak with both human raters

[1]Effect size details can be found in Appendix D.3

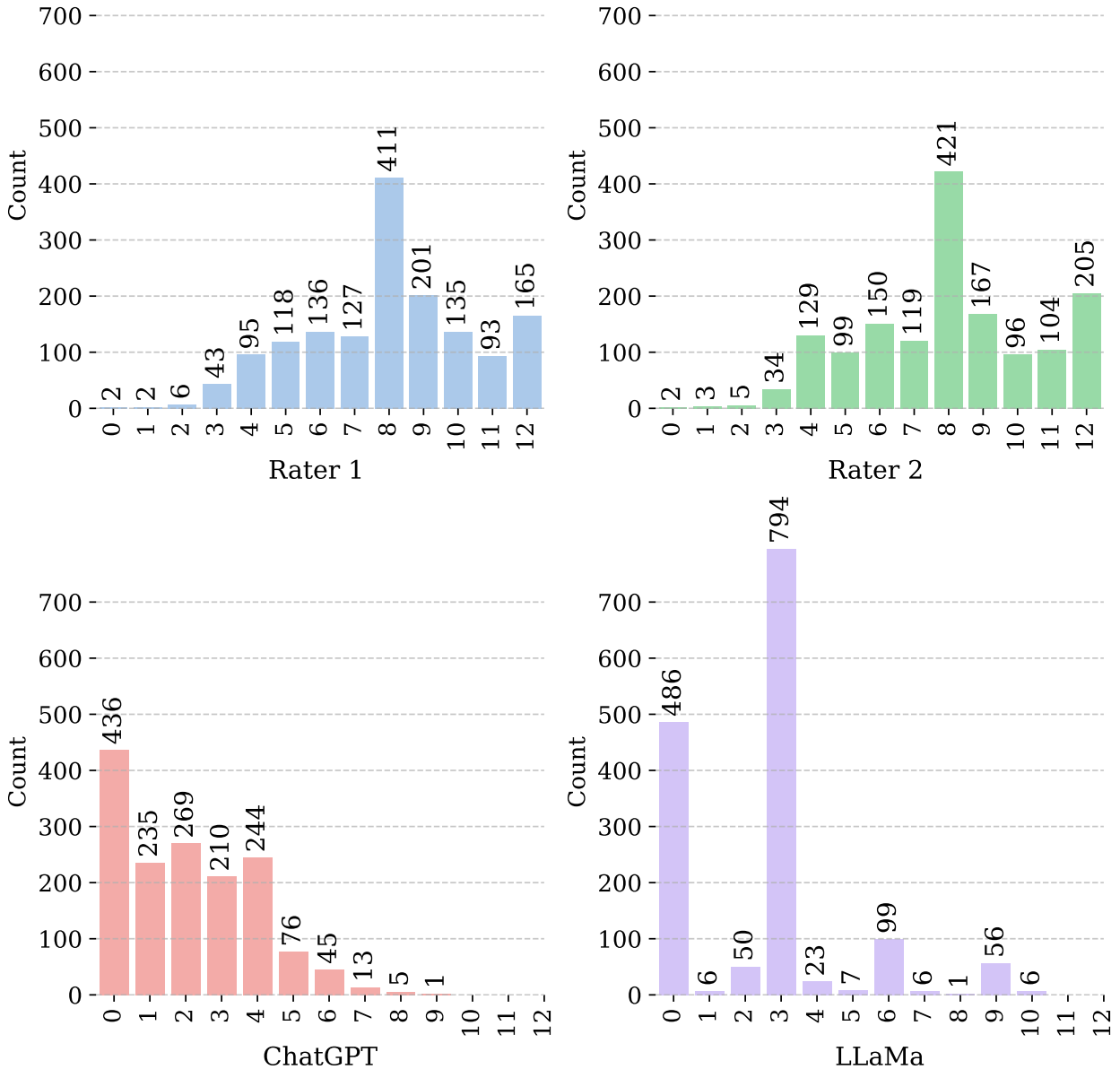

Figure 4.2: Overall score distribution of human raters and LLMs for Task 7

($r_{\text{rater1}} = 0.27 - 0.38, p < 0.001, d_{\text{overall}} = -2.10$ and $r_{\text{rater2}} = 0.28 - 0.40, p < 0.001, d_{\text{overall}} = -2.07$). Again, we observed no strong inter-LLM correlation ($r = 0.20 - 0.30, p < 0.001, d = 0.26$) between Llama-2 and GPT-3.5T for Task 7.

In summary, while human raters agree reasonably well with each other ($r = 0.72, p <$

Table 4.4: Correlations between human and LLM scores for Task 7 traits

Traits	**LLM**	**Rater 1**	**Rater 2**	**GPT-3.5T**
Ideas & content	GPT-3.5T	0.07	0.10	1.00
	Llama-2	0.27	0.29	0.21
Organization	GPT-3.5T	0.10	0.10	1.00
	Llama-2	0.32	0.36	0.24
Style	GPT-3.5T	0.12	0.15	1.00
	Llama-2	0.27	0.28	0.11
Convention	GPT-3.5T	0.19	0.18	1.00
	Llama-2	0.34	0.37	0.20
Overall	GPT-3.5T	0.18	0.19	1.00
	Llama-2	0.38	0.40	0.30

$0.001, d = 0.008$), LLMs (Llama-2 and GPT-3.5T) exhibit a weaker statistically significant alignment with both human raters, suggesting that GPT-3.5T is not a suitable scoring model using zero-shot with rubric guideline prompts. Though in Task 1, Llama-2 shows some alignment with human scores, in Task 7 correlations remain weak ($r <= 0.4$) for both GPT-3.5T and Llama-2, suggesting that both Llama-2 and GPT-3.5T diverge significantly from human raters. The lack of strong correlations underscores the challenge of automated scoring systems using LLMs in capturing nuanced writing qualities. The scoring differences indicate that GPT-3.5T and Llama-2 may have a different understanding of the rubric. Detailed statistical analyses are shown in Appendix C.

4.1.2 ASAP++ traits weakly correlate with GPT-3.5T

Moving forward, we assess the relationship between various essay trait scores provided by human raters in the ASAP++ dataset for Task 1 and scores given by human raters in the ASAP dataset and LLMs. Table 4.5 demonstrates that all the ASAP++ trait scores given by human raters show a strong positive correlation with the overall scores provided by human raters in the ASAP dataset ($r_{\text{rater1}} = 0.63 - 0.67, p < 0.001, d_{\text{mean}} = -0.56$ and $r_{\text{rater2}} = 0.62 - 0.67, p < 0.001, d_{\text{mean}} = -0.58$). Moreover, we observed that trait scores

are strongly correlated with Llama-2 scores ($r = 0.59 - 0.61, p < 0.001, d_{\text{mean}} = 0.16$) than GPT-3.5T scores ($r = 0.33 - 0.36, p < 0.001, d_{\text{mean}} = 2.35$). This finding further suggests that GPT-3.5T and human raters may approach the essay scoring task very differently, and Llama-2 can mimic human raters better than GPT-3.5T.

Table 4.5: Weak correlation between ASAP++ traits and GPT-3.5T for Task 1

Traits	**ASAP Rater 1**	**ASAP Rater 2**	**GPT-3.5T**	**Llama-2**
Ideas & Content	0.66	0.67	0.34	0.61
Organization	0.63	0.63	0.36	0.60
Word Choice	0.67	0.67	0.33	0.59
Sentence Fluency	0.64	0.62	0.36	0.60
Conventions	0.63	0.62	0.35	0.59

4.2 RQ2: What are the possible reasons behind the similarity/ difference in scores?

To delve into the reasons behind the weak correlation between human and LLM scores, we have examined various essay features, including the count of language errors, readability indices and length-related attributes such as the number of sentences and tokens. More details on the extraction of these features are outlined in section 3.2.2. All the detailed statistical analyses can be found in Appendix C.

Table 4.6: Strong correlation between human scores and essay length

	Score	Essay Length (sentences)	Essay Length (tokens)	FANBOYS (total)	FANBOYS (unique)	Transition Phrases (total)	Transition Phrases (unique)
Task 1	**Rater 1**	**0.63**	**0.74**	0.50	0.35	0.34	0.35
	Rater 2	**0.65**	**0.75**	0.49	0.35	0.37	0.39
	GPT-3.5T	0.16	0.20	0.12	0.09	0.05	0.08
	Llama-2	0.59	**0.66**	0.45	0.35	0.32	0.36
Task 7	**Rater 1**	**0.61**	**0.62**	0.39	0.31	0.36	0.43
	Rater 2	**0.61**	**0.63**	0.40	0.32	0.37	0.45
	GPT-3.5T	-0.02	0.00	-0.01	0.08	0.03	0.09
	Llama-2	0.42	0.45	0.27	0.21	0.24	0.26

4.2.1 Human scores highly correlate with essay features

From table 4.6, it is evident that human scores exhibit a strong positive correlation with length-related essay features (number of tokens and sentences) for both tasks. This suggests that human raters tend to assign higher scores to longer essays. It goes without saying that better writers are capable of producing more high-quality sentences and longer sentences than average and weaker writers. Therefore, Occam's razor would lead us to believe that the length of the essay has no real bearing on its score. Instead, all that we observe is that good-quality essays are longer because they were written by better students who could accomplish more in the allotted time and are graded with higher scores. Additionally, we observed that while Llama-2 scores show mostly a moderate correlation ($r = 0.42 - 0.66$) with essay length features (both the number of tokens and sentences), including Task 1 being an exception where the correlation is exceptionally high, GPT-3.5T grades do not show any substantial correlation ($r = 0.0 - 0.20$) with essay features. This observation indicates that GPT-3.5T tends to be more strict in scoring compared to the human raters and Llama-2.

In analyzing the use of transition and "FANBOYS" words showing table 4.6, a weak to moderate positive correlation ($r = 0.31 - 0.50$) has been found with human scores across both tasks. The higher correlation with the total count of "FANBOYS" words, as opposed to the unique count, suggests that students tend to rely repeatedly on these connecting words (such as "for, and, but, or, yet, so") in their writing. The more students join clauses and write complex structures of sentences the higher scores they get. Conversely, the higher correlation with the unique count of transition words compared to the total count indicates that employing diverse transition phrases may reflect better storytelling abilities, thus leading to higher scores. However, Llama-2 exhibits a weak correlation ($r = 0.21 - 0.45$), and GPT-3.5T shows an even weaker correlation ($r = 0.01 - 0.12$) compared to human raters, suggesting that humans may excel in evaluating the logical progression of ideas and narrative coherence through the adept use of these connecting words.

Table 4.7: Strong correlation between ASAP++ traits and essay length for Task 1

Traits	**Essay Length (sentences)**	**Essay Length (tokens)**	**FANBOYS (total)**	**FANBOYS (unique)**	**Transition Phrases (total)**	**Transition Phrases (unique)**
Ideas & Content	0.58	0.68	0.43	0.28	0.30	0.32
Organization	0.54	0.64	0.40	0.26	0.28	0.30
Word Choice	0.56	0.66	0.44	0.28	0.28	0.30
Sentence Fluency	0.54	0.63	0.42	0.28	0.26	0.28
Conventions	0.53	0.63	0.42	0.28	0.27	0.29

We extend our analysis with trait-wise scores in the ASAP++ dataset, examining their correlation with essay features. Table 4.7 illustrates a moderate positive correlation between trait scores and the number of sentences ($r = 0.53 - 0.58, p < 0.001, d_{\text{mean}} = -3.05$), as well as a strong positive correlation with the number of tokens ($r = 0.63 - 0.68, p < 0.001, d_{\text{mean}} = -4.22$). This reaffirms the tendency for longer essays to receive higher scores from human

raters. Furthermore, the observed trend in connected words aligns with the patterns seen in overall scores in the ASAP dataset, highlighting human raters' proficiency in recognizing narrative coherence and rewarding it with higher marks.

4.2.2 LLM and human scores do not correlate with readability indices

Here, we compare well-known readability indices with scores provided by LLMs and human raters. As demonstrated in Table 4.8, we can see that the scores provided by both human raters have very negligible correlations ($r = 0.01 - 0.3$) across all well-known readability indices. This suggests a minimal influence of readability on human scoring.

Table 4.8: No correlations between scores and various readability indices

	Score	**Flesch-Kincaid Grade-level**	**Flesch Reading Ease**	**Smog Index**	**Coleman Liau**	**Gunning Fog**	**Automated Readability**	**Linsear Write**	**Dale Chall**
Task 1	**Rater 1**	0.02	-0.17	0.16	0.30	-0.19	0.01	-0.10	0.16
	Rater 2	0.01	-0.15	0.14	0.28	-0.21	-0.01	-0.12	0.16
	GPT-3.5T	0.10	-0.23	0.18	0.17	-0.09	0.00	-0.08	-0.01
	Llama-2	-0.01	-0.19	0.13	0.24	-0.25	-0.08	-0.19	0.05
Task 7	**Rater 1**	-0.19	0.13	0.16	0.11	-0.30	-0.14	-0.21	-0.06
	Rater 2	-0.19	0.13	0.16	0.10	-0.31	-0.14	-0.21	-0.09
	GPT-3.5T	0.09	-0.20	0.18	0.11	-0.04	0.01	-0.04	-0.21
	Llama-2	-0.02	-0.09	0.21	0.18	-0.15	-0.03	-0.12	-0.05

Interestingly, the tables also demonstrate a lack of correlation between LLM scores and readability metrics. Since we are using LLMs as a zero-shot learner and since we do not mention readability explicitly in the prompt (See Appendix A for an example), this result implies that LLM's interpretation of the question and its response are not associated with

the numerous mentions to readability found on the web (and very likely used in the pre-training of the LLM models). Furthermore, some correlations with all scoring methods are negative, indicating that both humans and LLMs may assign high grades to less readable texts and low grades to highly readable ones.

4.2.3 LLM scores negatively correlate with the count of mistakes

We utilized LanguageTool and Aspell to detect various language mistakes in the essays, including grammar, spelling, style, punctuation, and capitalization, as described in section 3.2.2. Following this, we computed the number of mistakes for all essays and conducted correlation analysis across four scoring methods. In Table 4.9, we observed that GPT-3.5T consistently displayed a weak but negative correlation, whereas other scorers mostly showed a positive correlation with mistake count.

Table 4.9: Negative correlation between LLM scores and mistake counts

	Score \ Mistake count	**Aspell Mispelling**	**LanguageTool Grammar**	**LanguageTool Spelling**	**LanguageTool Style**	**LanguageTool Punctuation**	**LanguageTool Capitalization**
Task 1	**Rater 1**	0.14	0.09	0.11	0.09	0.02	0.11
	Rater 2	0.16	0.14	0.13	0.10	0.03	0.12
	GPT-3.5T	**-0.12**	**-0.11**	**-0.15**	**-0.02**	**-0.01**	**-0.08**
	Llama-2	0.04	0.05	0.02	0.05	0.05	0.11
Task 7	**Rater 1**	0.11	0.05	0.20	0.11	0.11	0.05
	Rater 2	0.10	0.06	0.21	0.09	0.10	0.05
	GPT-3.5T	**-0.24**	**-0.17**	**-0.21**	**-0.05**	**-0.05**	**-0.17**
	Llama-2	**-0.10**	**-0.03**	0.08	0.06	0.01	**-0.11**

The negative correlation observed with mistakes aligns with the rubric, where a higher number of language mistakes typically corresponds to lower scores. This logical relationship

underscores GPT-3.5T's reliability in accurately identifying mistakes and adjusting scores accordingly. Additionally, Llama-2 also demonstrated a negative correlation in some instances, albeit with mostly negligible positive scores. Conversely, human graders exhibited a surprising positive correlation with mistake count. This suggests potential challenges for human graders in consistently identifying and adjusting grades based on spelling and grammar errors, highlighting the systematic ability of LLMs like GPT-3.5T and Llama-2 in language error detection. On the other hand, this may be attributed to potential inaccuracies in the misspelling count reported by the spelling checker tools. To investigate this matter, we manually inspected the words flagged by *Aspell* as misspelled in randomly selected essays (approx 10) and cross-checked with another spell checker *LanguageTool*. We have also performed a correlation analysis of these two spell checkers, details can be found in section D.1. Our findings indicate that it strongly agrees with *Aspell* that the likelihood of the last scenario occurring is unlikely.

Table 4.10: ASAP++ trait scores negatively correlate with mistake count

Score \ Mistake count	**Aspell Mispelling**	**LanguageTool Grammar**	**LanguageTool Spelling**	**LanguageTool Style**	**LanguageTool Punctuation**	**LanguageTool Capitalization**
Ideas & Content	0.04	-0.01	-0.01	0.06	-0.03	0.05
Organization	0.01	-0.02	-0.03	0.04	-0.02	0.04
Word Choice	0.03	-0.02	-0.02	0.04	-0.05	0.04
Sentence Fluency	-0.05	-0.05	-0.09	0.02	-0.02	0.00
Conventions	-0.05	-0.04	-0.09	0.02	-0.01	-0.00

We extended our analysis by examining the relationship between mistake count and trait-wise scores in the ASAP++ dataset. In Table 4.10, we observed a negative correlation between the LanguageTool's grammar, spelling, and punctuation error counts and trait scores, although the correlation was very weak. Specifically, for the convention trait score,

most mechanical error counts displayed a negative correlation. This indicates that trait scores offer detailed insights pinpointing specific areas for skill enhancement, particularly in identifying mechanical mistakes with greater precision than overall scoring by human raters.

4.3 RQ3: Do LLMs offer explanations in a tone that reflects their scores?

4.3.1 GPT-3.5T gives consistently harsh explanation

We have extracted basic statistical features and applied the VADER sentiment analyzer from the GPT-3.5T explanation as described in section 3.2.3. In table 4.11, explanation length (sentences) and explanation length (tokens) refer to the number of sentences and the number of words found in GPT-3.5T explanation respectively. Explanation sentiment (average) refers to the overall compound sentiment score of GPT-3.5T response provided by VADER. Explanation sentiment (max) means sentence-wise maximum compound score in GPT-3.5T response provided by VADER sentiment analyzer. We perform correlation analysis between these four features with human and LLM scores.

It's worth noting that the explanation features show a positive weak correlation, but they are relatively stronger with GPT-3.5T's scores compared to human-assigned or Llama-2 scores. This suggests that GPT-3.5T has a better understanding of the prompt and provides explanations that align with the assigned scores, indicating consistency. For instance, lower scores receive more negative explanations, while higher scores get more positive ones. This implies that GPT-3.5T is not generating explanations randomly but is aware of the scores it has provided.

However, we observed that GPT-3.5T scores have smaller coefficient scores ($r = 0.13 -$

Table 4.11: Harsh tendency in GPT-3.5T's explanation

	Score	Explanation Length (sentences)	Explanation Length (tokens)	Explanation Sentiment (average)	Explanation Sentiment (max)
Task 1	**Rater 1**	0.12	0.17	0.06	0.04
	Rater 2	0.11	0.15	0.06	0.04
	GPT-3.5T	0.33	0.45	**0.33**	**0.28**
	Llama-2	0.14	0.19	0.14	0.12
Task 7	**Rater 1**	0.05	0.04	0.05	0.02
	Rater 2	0.06	0.04	0.05	0.01
	GPT-3.5T	0.28	0.30	**0.24**	**0.13**
	Llama-2	0.13	0.13	0.08	0.02

0.33), indicating that GPT-3.5T may not be explained in a supportive tone. Considering that GPT-3.5T's overall scores are quite low and LLMs have a tendency to generate toxic content [52], it's possible that the tone of the explanations may not appear human-like supportive due to the low scores. Since we cannot rule out the possibility that the explanation generated by GPT-3.5T could have been influenced by the score it produced, there might be a side effect of using LLMs as zero-shot classifiers, affected by their tendency towards generating toxic content (despite concerted efforts by LLM developers to curb this tendency).

4.3.2 Llama-2 gives moderately positive explanation

We conducted a comprehensive analysis of the Llama-2 explanations using the VADER sentiment analyzer, similar to our approach with the GPT-3.5T explanations. The results detailed in Table 4.12, reveal that the Llama-2 scores exhibit notably higher correlation ($r = 0.42 - 0.54$) with its explanations. This finding suggests that in contrast to GPT-3.5T, Llama-2 generates explanations with a moderately positive sentiment highlighting Llama-2's

ability to convey supportive explanations to learners.

Furthermore, we observed a consistent trend where the correlation between Llama-2 scores and its explanations outweighs that of any other scoring methods, underscoring the consistency and reliability of the explanations provided by Llama-2 in relation to their assigned scores. This alignment highlights the coherent explanations that correspond closely with their evaluations, demonstrating a promising aspect of using Llama-2 in educational assessment tasks.

Table 4.12: Llama-2's moderately positive explanation

	Score	Explanation Length (sentences)	Explanation Length (tokens)	Explanation Sentiment (average)	Explanation Sentiment (max)
Task 1	**Rater 1**	0.26	0.26	0.30	0.24
	Rater 2	0.25	0.27	0.29	0.22
	GPT-3.5T	0.10	0.11	0.24	0.20
	Llama-2	0.28	0.31	**0.54**	**0.42**
Task 7	**Rater 1**	0.16	0.21	0.26	0.22
	Rater 2	0.13	0.17	0.25	0.22
	GPT-3.5T	0.11	0.13	0.13	0.14
	Llama-2	0.27	0.30	**0.50**	**0.44**

4.4 RQ4: Can LLMs correctly identify and assess spelling and grammatical mistakes and score accordingly?

To investigate whether the mention of mistakes (misspellings and grammatical errors) in the explanation reflects the quality of the essays, we have exploited the LLM explanation as described in 3.2.3 further. Now we have the information for each essay sample on whether LLM

has identified any language mistakes. Next, we categorized these LLM responses (for Task 1 and Task 7 of the ASAP dataset) based on the mention of grammatical and spelling errors separately into three distinct groups. The first group contains samples without mention of spelling errors, the second group comprises only samples with mentions of misspellings, and the last group contains essay samples with mentions of misspellings along with a qualifier (such as *several, numerous, multiple*). We followed this grouping process for grammatical errors for both GPT-3.5T and Llama-2. For these three groups, we have calculated the average of spelling mistakes by LanguageTool and Aspell, LLM scores, and sentiment score of LLM explanation and compared them against each other.

We determined whether the differences between group means are statistically significant by conducting the Analysis of Variance (ANOVA) test and reported p-values and F-stat. The goal of an ANOVA is to determine the effects of discrete independent variables, these are typically categorical variables, on a continuous dependent variable. The one-way ANOVA is a parametric test, used to determine if there is a statistically significant difference in means across three or more groups. In our study, we utilized a one-way ANOVA to assess the differences in outcomes across three categories of misspelling or grammar error mention: 'No mention', 'Unqualified mention', and 'Qualified mention'. This omnibus test checks for an overall difference, indicating that at least one group differs significantly from the others.

In the context of ANOVA, the independent variable (or factor) is the variable that categorizes or groups the data. In our analysis, the independent variables are the type of categories like 'No mention', 'Unqualified mention', and 'Qualified mention'. Whereas, the dependent variables included outcomes such as the mean of misspellings detected by Aspell, the average number of spelling/grammar errors identified by LanguageTool, the average LLM scores, the average LLM explanation sentiment scores, and the maximum LLM explanation sentiment scores in our case.

However, a significant ANOVA result does not specify which groups are different. Therefore, to identify the specific group differences, we performed post-hoc analyses using the Tukey Honestly Significant Difference (HSD) test for the results detailed in Appendix C. We have also calculated effect sizes using Omega squared (ω^2) [2], which is widely recognized as a less biased alternative to eta-squared, particularly when dealing with small sample sizes.

As shown in Tables 4.13 and 4.14 for Task 1 and Task 7 respectively, both Aspell and LanguageTool consistently report an increase in the average number of misspellings across these groups. This logical trend aligns with expectations: the first group, with no mention

[2]Cohen (1988) classified ω^2 effect sizes as small ($\omega^2 = 0.01$), medium ($\omega^2 = 0.059$), and large ($\omega^2 \geq 0.138$)

Table 4.13: Changes in average number of misspellings, sentiment scores of LLM explanations and LLM scores across different misspelling categories Task 1

Misspelling Category	**Sample Count**	**Aspell Misspelling**	**LanguageTool Spelling**	**LLM Score**	**Explanation Sentiment (average)**	**Explanation Sentiment (max)**
GPT-3.5T						
(1) No mention	1546	8.16	9.26	1.91	-0.21	-0.02
(2) Unqualified mention	65	12.14	13.42	1.91	-0.43	-0.04
(3) Qualified mention	172	13.53	14.98	1.78	-0.49	-0.26
F-stat $(2, 1780)$		59.37	58.43	3.7	24.25	12.23
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.06	0.06	0.004	0.03	0.01
Llama-2						
(1) No mention	1598	8.58	9.71	3.63	0.79	0.69
(2) Unqualified mention	148	9.80	10.95	3.77	0.84	0.67
(3) Qualified mention	37	15.23	17.00	2.62	0.23	0.17
F-stat $(2, 1780)$		19.08	19.86	48.65	29.25	20.08
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.02	0.02	0.05	0.03	0.02

of mistakes, naturally contains fewer mistakes than the second group, where mistakes are mentioned in the explanation. Similarly, the third group, which includes qualifiers indicating multiple mistakes, is expected to have more mistakes than the other two groups. This finding underscores the fact that both LLMs (GPT-3.5T and Llama-2) pay close attention to the presence of spelling mistakes in essays when delivering explanations. Additionally, this highlights the careful use of qualifiers (e.g., several, numerous, many) in their explanation generation especially when there are more misspellings.

However, while this focus on misspellings may not significantly affect the scores assigned by GPT-3.5T and Llama-2, as seen in the small to no difference in average scores between

Table 4.14: Changes in average number of misspellings, explanation sentiment scores and LLM scores across different misspelling categories Task 7

Misspelling Category	**Sample Count**	**Aspell Misspelling**	**LanguageTool Spelling**	**LLM Score**	**Explanation Sentiment (average)**	**Explanation Sentiment (max)**
GPT-3.5T						
(1) No mention	917	3.53	5.6	2.32	-0.60	-0.27
(2) Unqualified mention	203	4.08	6.3	2.44	-0.57	-0.24
(3) Qualified mention	449	5.18	7.4	1.39	-0.65	-0.30
F-stat $(2, 1566)$		22.87	17.22	45.76	3.11	1.03
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p > 0.05$
Effect size (ω^2)		0.03	0.02	0.05	0.004	0.001
Llama-2						
(1) No mention	667	3.39	5.42	3.19	-0.28	-0.13
(2) Unqualified mention	339	4.03	6.55	3.37	-0.30	-0.13
(3) Qualified mention	563	4.92	6.92	1.60	-0.60	-0.37
F-stat $(2, 1566)$		20.06	12.86	75.13	46.04	32.69
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.02	0.02	0.09	0.06	0.04

the first two groups, there is a noticeable difference in average LLM scores between the first and third groups. This suggests that the LLMs don't just assign scores randomly all the time, but they can assess evaluating the essay quality to some extent.

Additionally, the decrease in VADER sentiment scores of LLM's explanation is observed, which logically aligns with expectations, as the presence of more mistakes and lower scores tends to result in harsher sentiment in the explanation. This correspondence between sentiment and mistakes indicates that as more mistakes accumulate in student essays, the explanations provided by both GPT-3.5T and Llama-2 tend to adopt a more negative tone, demonstrating their ability to adjust explanations appropriately according to essay quality. However, while Llama-2 sentiment scores also decrease with essay mistakes, they remain relatively positive compared to GPT-3.5T.

In Tables 4.15 and 4.16, we observe a similar trend as seen for the misspelling categories. All the results depicted in these tables are statistically significant and can have small to medium effects practically. This reaffirms that LLMs can identify spelling and grammatical mistakes and evaluate essays accordingly.

Furthermore, we categorized essay samples into three groups based on the scores provided by human raters, with each group representing a range of average scores from two raters. Tables 4.17 and 4.18 illustrate that LLMs assign lower scores to lower-grade groups and relatively higher scores to higher-grade groups. This logical trend suggests that LLMs may possess the potential to effectively assess essay quality. However, further fine-tuning and the implementation of additional techniques that enable contextual learning may enhance their performance in grading tasks.

Table 4.15: Changes in average number of grammar mistakes, sentiment scores and LLM score across different grammatical categories Task 1

Grammar Error Category	**Sample Count**	**LanguageTool Grammar**	**LLM Score**	**Explanation Sentiment (average)**	**Explanation Sentiment (max)**
GPT-3.5T					
(1) No mention	1169	2.93	1.89	-0.17	0.02
(2) Unqualified mention	230	3.06	1.99	-0.35	-0.12
(3) Qualified mention	384	3.51	1.85	-0.43	-0.22
F-stat $(2, 1780)$		8.07	4.43	39.52	28.38
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.01	0.041	0.04	0.03
Llama-2					
(1) No mention	1422	3.02	3.64	0.80	0.69
(2) Unqualified mention	298	3.15	3.72	0.82	0.67
(3) Qualified mention	63	3.94	2.70	0.34	0.29
F-stat $(2, 1780)$		4.42	70.38	31.62	20.04
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.004	0.07	0.03	0.02

Table 4.16: Changes in average number of grammar mistakes, sentiment scores and LLM score across different grammatical categories Task 7

Grammar Error Category	**Sample Count**	**LanguageTool Grammar**	**LLM Score**	**Explanation Sentiment (average)**	**Explanation Sentiment (max)**
GPT-3.5T					
(1) No mention	332	1.08	2.35	-0.56	-0.26
(2) Unqualified mention	386	1.10	2.76	-0.54	-0.21
(3) Qualified mention	851	1.47	1.65	-0.66	-0.31
F-stat $(2, 1566)$		11.7	57.7	13.11	5.22
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.01	0.07	0.02	0.001
Llama-2					
(1) No mention	802	1.17	3.25	-0.30	-0.15
(2) Unqualified mention	303	1.33	3.00	-0.37	-0.17
(3) Qualified mention	464	1.48	1.41	-0.58	-0.36
F-stat $(2, 1566)$		6.04	79.11	28.58	21.65
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.01	0.09	0.03	0.03

Table 4.17: Comparison of average grades assigned by LLMs across human rater grade categories Task 1

Human Score Class	**Sample Count**	**GPT-3.5T Score**	**Llama-2 Score**
Score 1-2	28	1.21	1.68
Score 2-4	949	1.82	3.39
Score 4-6	806	2.01	3.95
F-stat(2,1780)		46.46	399.13
P-values		$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.05	0.31

Table 4.18: Comparison of average grades assigned by LLMs across human rater grade categories Task 7

Human Score Class	**Sample Count**	**GPT-3.5T Score**	**Llama-2 Score**
Score 0-4	108	0.84	0.89
Score 4-8	717	1.93	1.93
Score 8-15	744	2.38	3.62
F-stat(2,1566)		39.05	111.2
P-values		$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.05	0.12

Chapter 5

Additional Results with Llama-3

In this chapter, we consider one of the more recent and most powerful large language models which is Llama-3. This model was not available when we started our study, so we decided to present them separately in this chapter for better clarity. Towards the end of our research, new and much larger language models such as GPT-4 [53], Llama-3 [54] became available. Llama 3 70B can be up to 50 times more cost-effective and 10 times faster than GPT-4[1]. GPT-4 still has advantages in scenarios that need longer context or special features like image support and function calling [53]. However, for many tasks, Llama 3 70B is catching up as a strong competitor and producing comparable results [55]. In the interest of time, we were able to run *Meta-Llama-3-70B-Instruct* [2] lately because we could deploy it locally. In this chapter, we are going to revisit some of our findings now using Llama 3 70B.

[1] https://www.vellum.ai/blog/llama-3-70b-vs-gpt-4-comparison-analysis
[2] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

5.1 RQ1: Do human scores align with the LLM scores?

As shown in Table 5.1, for Task 1 Llama-3 assigned valid overall scores to all 1783 essays, with scores consistently ranging from 1 to 6. In Task 7, Table 5.2 reveals that Llama-3 surprisingly assigned higher scores ($M = 6.05, SD = 2.73$) than GPT-3.5T or Llama-2, and these scores are much closer to those of the human raters ($M = 8.02, SD = 2.42$ and $M = 8.04, SD = 2.52$). Llama-3 demonstrated impressive results by scoring only 4 samples out of range, all of which occurred while predicting the organization trait score. We can see a more normalized distribution of scores in Figures 5.1 and 5.2.

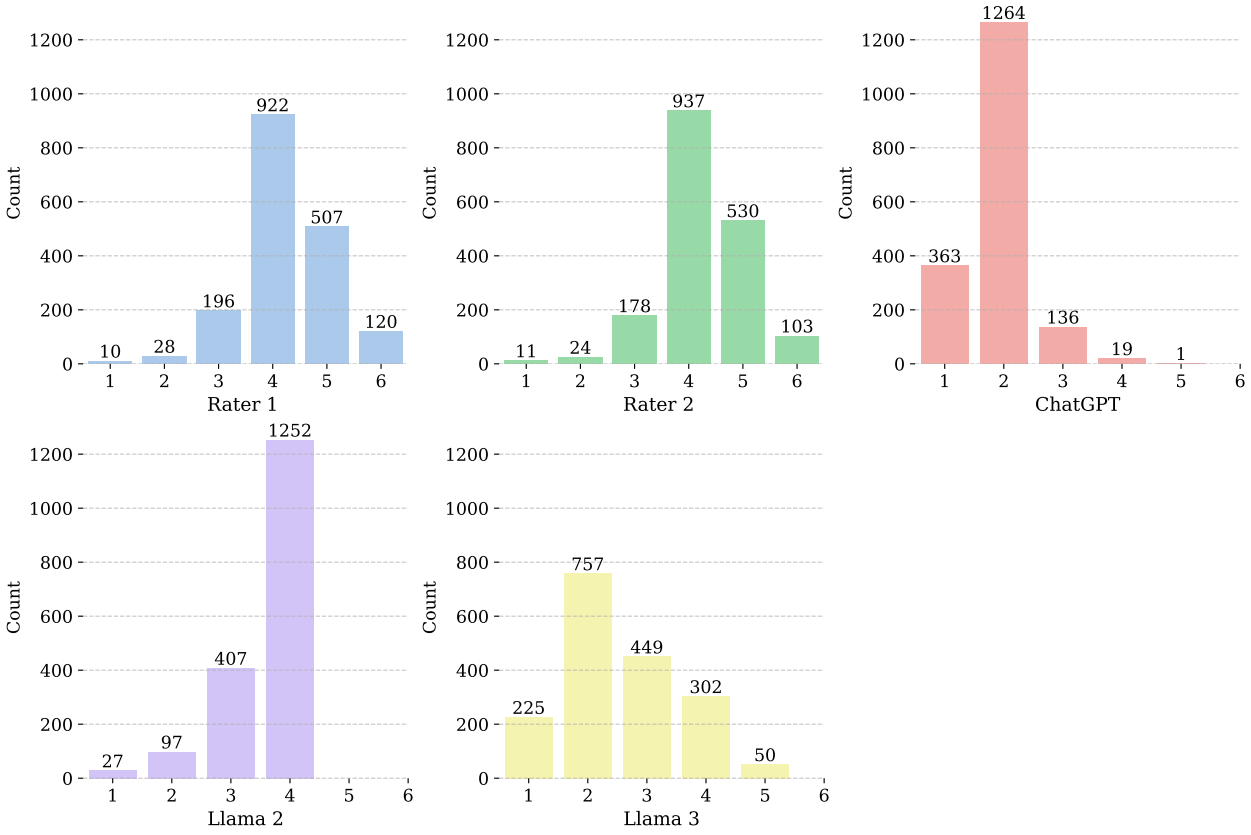

Figure 5.1: Score distribution of human raters and LLMs including Llama-3 for Task 1

Table 5.1: Descriptive statistics with Llama-3 and other scoring methods for Task 1

	Rater 1	Rater 2	GPT-3.5T	Llama-2	Llama-3
Mean	4.26	4.27	1.90	3.62	2.55
SD	0.84	0.82	0.56	0.66	1.00
Min	1.00	1.00	1.00	1.00	1.00
Median	4.00	4.00	2.00	4.00	2.00
Max	6.00	6.00	5.00	4.00	5.00

Table 5.2: Descriptive statistics with Llama-3 and other scoring methods for Task 7

	Rater 1	Rater 2	GPT-3.5T	Llama-2	Llama-3
Mean	8.02	8.04	2.07	2.66	**6.05**
SD	2.42	2.52	1.83	2.68	2.73
Min	0.00	0.00	0.00	0.00	0.00
Median	8.00	8.00	2.00	3.00	6.00
Max	12.00	12.00	9.00	26.00	14.00

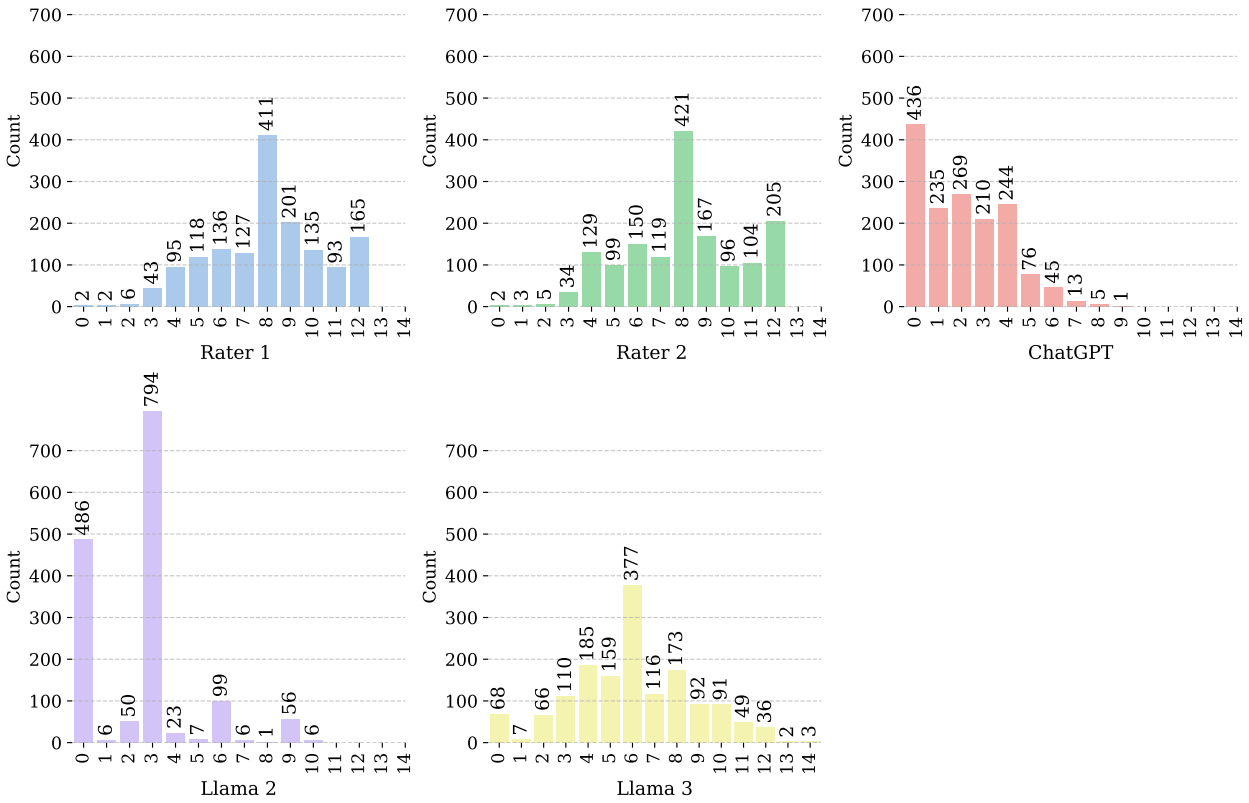

Figure 5.2: Score distribution of human raters and LLMs including Llama-3 for Task 7

Table 5.3: Moderate correlation between human rater and Llama-3 scores in Task 1

Score	**Rater 1**	**Rater 2**	**ChatGPT**	**Llama-2**
ChatGPT	0.23	0.21	1.00	0.32
Llama-2	0.59	0.58	0.32	1.00
Llama-3	0.52	0.49	0.46	0.54

In terms of agreement with human raters in Task 1, as shown in Table 5.3, Llama-2 has a correlation that is 13.5% higher with rater 1 and 18.4% higher with rater 2 compared to Llama-3. On average, Llama-3 is 16% less correlated with human scores than Llama-2. However, Llama-3 shows significantly better agreement with human raters compared to GPT-3.5T, with a 126% higher correlation with rater 1 and a 133.3% higher correlation with rater 2, resulting in an average improvement of 130% in correlation compared to GPT-3.5T.

Table 5.4: Moderate correlation between human and Llama-3 scores in Task 7

Traits	**LLM**	**Rater 1**	**Rater 2**	**GPT-3.5T**	**Llama-2**
Ideas & content	GPT-3.5T	0.07	0.10	1.00	0.21
	Llama-2	0.27	0.29	0.21	1.00
	Llama-3	0.42	0.44	0.27	0.44
Organization	GPT-3.5T	0.10	0.10	1.00	0.24
	Llama-2	0.32	0.36	0.24	1.00
	Llama-3	0.38	0.39	0.26	0.41
Style	GPT-3.5T	0.12	0.15	1.00	0.11
	Llama-2	0.27	0.28	0.11	1.00
	Llama-3	0.39	0.43	0.22	0.36
Convention	GPT-3.5T	0.19	0.18	1.00	0.20
	Llama-2	0.34	0.37	0.20	1.00
	Llama-3	0.31	0.36	0.20	0.45
Overall	GPT-3.5T	0.18	0.19	1.00	0.30
	Llama-2	0.38	0.40	0.30	1.00
	Llama-3	0.49	0.52	0.40	0.52

On the other hand, in Task 7, as shown in Table 5.4, Llama-3 demonstrates a correlation that is 29% higher with rater 1 and 30% higher with rater 2 compared to Llama-2. On average, Llama-3 shows an 29.5% better correlation with human scores than Llama-2.

Additionally, Llama-3 significantly outperforms GPT-3.5T, with a 172% higher correlation with rater 1 and a 173.7% higher correlation with rater 2, resulting in an overall average improvement of 173% in correlation compared to GPT-3.5T. Table 5.5 shows that similar to Llama-2, Llama-3 has strong correlation with ASAP++ trait scores by human raters. Notably, correlations for Llama-3 are higher than both Llama-2 and GPT-3.5T across all traits.

Table 5.5: Strong correlation between ASAP++ traits and Llama-3 scores in Task 1

Traits	**GPT-3.5T**	**Llama-2**	**Llama-3**
Ideas & Content	0.34	0.61	0.64
Organization	0.36	0.60	0.60
Word Choice	0.33	0.59	0.63
Sentence Fluency	0.36	0.60	0.64
Conventions	0.35	0.59	0.63

5.2 RQ2: What are the possible reasons behind the similarity/ difference in scores?

We perform similar correlation analyses with essay features, readability indices, and language-checking tools. Table 5.6 reveals that Llama-3 has a moderate correlation with essay features, which is weaker than Llama-2 but stronger than GPT-3.5T. Comparatively, Llama-3 shows stronger correlations with some readability indices, as seen in Table 5.7, than the other LLMs. Similar to GPT-3.5T, Llama-3 exhibits negative yet weak correlations across all mistake types in Task 1, as shown in Table 5.8. For Task 7, Llama-3 shows negative correlations

only with misspelling, grammar, and capitalization mistake counts, similar to Llama-2.

Table 5.6: Moderate correlation between Llama-3 scores and basic essay features

	Score	Essay Length (sentences)	Essay Length (tokens)	FANBOYS (total)	FANBOYS (unique)	Transition Phrases (total)	Transition Phrases (unique)
Task 1	**GPT-3.5T**	0.16	0.20	0.12	0.09	0.05	0.08
	Llama-2	0.59	0.66	0.45	0.35	0.32	0.36
	Llama-3	0.44	0.53	0.31	0.22	0.20	0.25
Task 7	**GPT-3.5T**	-0.02	0.00	-0.01	0.08	0.03	0.09
	Llama-2	0.42	0.45	0.27	0.21	0.24	0.26
	Llama-3	0.37	0.44	0.30	0.28	0.27	0.34

Table 5.7: Comparatively stronger correlations between Llama-3 scores and various readability indices

	Score	Flesch-Kincaid Grade-level	Flesch Reading Ease	Smog Index	Coleman Liau	Gunning Fog	Automated Readability	Linsear Write	Dale Chall
Task 1	**GPT-3.5T**	0.10	-0.23	0.18	0.17	-0.09	0.00	-0.08	-0.01
	Llama-2	-0.01	-0.19	0.13	0.24	-0.25	-0.08	-0.19	0.05
	Llama-3	**0.15**	**-0.39**	**0.29**	**0.37**	-0.17	0.04	-0.11	0.10
Task 7	**GPT-3.5T**	0.09	-0.20	0.18	0.11	-0.04	0.01	-0.04	-0.21
	Llama-2	-0.02	-0.09	0.21	0.18	-0.15	-0.03	-0.12	-0.05
	Llama-3	0.05	-0.16	**0.26**	**0.19**	-0.12	-0.00	-0.08	-0.19

Table 5.8: Negative weak correlation between Llama-3 scores and mistake counts

	Mistake count / Score	Aspell Mispelling	LanguageTool Grammar	LanguageTool Spelling	LanguageTool Style	LanguageTool Punctuation	LanguageTool Capitalization
Task 1	**Rater 1**	0.14	0.09	0.11	0.09	0.02	0.11
	Rater 2	0.16	0.14	0.13	0.10	0.03	0.12
	GPT-3.5T	**-0.12**	**-0.11**	**-0.15**	**-0.02**	**-0.01**	**-0.08**
	Llama-2	0.04	0.05	0.02	0.05	0.05	0.11
	Llama-3	**-0.14**	**-0.15**	**-0.18**	**-0.01**	**-0.10**	**-0.02**
Task 7	**Rater 1**	0.11	0.05	0.20	0.11	0.11	0.05
	Rater 2	0.10	0.06	0.21	0.09	0.10	0.05
	GPT-3.5T	**-0.24**	**-0.17**	**-0.21**	**-0.05**	**-0.05**	**-0.17**
	Llama-2	**-0.10**	**-0.03**	0.08	0.06	0.01	**-0.11**
	Llama-3	**-0.13**	**-0.05**	0.01	0.01	0.06	**-0.16**

5.3 RQ3: Do LLMs offer explanations in a tone that reflects their scores?

As seen in Tables 5.9 and 5.10, Llama-3 has a weaker correlation with GPT-3.5T and Llama-2 explanations than these models have with themselves. However, Table 5.11 reveals that Llama-3's explanations are comparatively strongly correlated with the human raters' scores in Task 1. Furthermore, Llama-3's provided scores are moderately positively correlated with its own explanations.

Table 5.9: Llama-3 scores weakly correlate with GPT-3.5T's explanation

	Score	Explanation Length (sentences)	Explanation Length (tokens)	Explanation Sentiment (average)	Explanation Sentiment (max)
Task 1	**Rater 1**	0.12	0.17	0.06	0.04
	Rater 2	0.11	0.15	0.06	0.04
	GPT-3.5T	0.33	0.45	**0.33**	**0.28**
	Llama-2	0.14	0.19	0.14	0.12
	Llama-3	0.16	0.25	0.18	0.13
Task 7	**Rater 1**	0.05	0.04	0.05	0.02
	Rater 2	0.06	0.04	0.05	0.01
	GPT-3.5T	0.28	0.30	**0.24**	**0.13**
	Llama-2	0.13	0.13	0.08	0.02
	Llama-3	0.13	0.15	0.10	0.02

Table 5.10: Llama-3 scores weakly correlate with Llama-2's explanation

	Score	Explanation Length (sentences)	Explanation Length (tokens)	Explanation Sentiment (average)	Explanation Sentiment (max)
Task 1	**Rater 1**	0.26	0.26	0.30	0.24
	Rater 2	0.25	0.27	0.29	0.22
	GPT-3.5T	0.10	0.11	0.24	0.20
	Llama-2	0.28	0.31	**0.54**	**0.42**
	Llama-3	0.22	0.23	0.34	0.26
Task 7	**Rater 1**	0.16	0.21	0.26	0.22
	Rater 2	0.13	0.17	0.25	0.22
	GPT-3.5T	0.11	0.13	0.13	0.14
	Llama-2	0.27	0.30	**0.50**	**0.44**
	Llama-3	0.23	0.27	0.31	0.30

Table 5.11: Llama-3's moderately positive explanation

	Score	Explanation Length (sentences)	Explanation Length (tokens)	Explanation Sentiment (average)	Explanation Sentiment (max)
Task 1	**Rater 1**	**0.35**	0.40	0.23	0.21
	Rater 2	**0.37**	0.41	0.21	0.21
	GPT-3.5T	0.11	0.21	0.29	0.23
	Llama-2	0.32	0.43	0.31	0.25
	Llama-3	0.34	**0.47**	**0.49**	**0.40**
Task 7	**Rater 1**	0.25	0.34	0.28	0.20
	Rater 2	0.29	0.36	0.30	0.22
	GPT-3.5T	0.15	0.24	0.24	0.17
	Llama-2	0.22	0.28	0.38	0.26
	Llama-3	**0.34**	**0.46**	**0.48**	**0.32**

5.4 RQ4: Can LLMs correctly identify and assess spelling and grammatical mistakes and score accordingly?

Similar to previous analyses, we categorized Llama-3 responses based on the mention of grammatical and spelling errors into three distinct groups. As shown in Tables 5.12, 5.13, 5.14, and 5.15, we observe a logical increase in the average number of mistake counts across grammatical and spelling categories for both Task 1 and Task 7. The differences in spelling and grammar mistake counts are prominent in the first and third groups. All results are statistically significant, with Llama-3 showing comparatively more medium and large (in bold) effect sizes. There is a noticeable decrease in scores from the first group to the third group as mistakes increase. The sentiment of the explanations provided by Llama-3 is moderate, falling between GPT-3.5T's harsher tone and Llama-2's more positive tone.

In Tables 5.16 and 5.17, we categorized essay samples into three distinct groups based on the scores provided by human raters and calculated the average scores given by LLMs. We observe a more distinct difference in Llama-3 scores from the first group to the second group to the third group. Llama-3 successfully assigns lower scores to lower-grade groups and higher scores to higher-grade groups. In both tasks, the results are statistically and practically significant.

Table 5.12: Changes in average number of misspellings, Llama-3's scores and explanation sentiment across different misspelling categories in Task 1

Misspelling Category	**Sample Count**	**Aspell Misspelling**	**LanguageTool Spelling**	**LLM Score**	**Explanation Sentiment (average)**	**Explanation Sentiment (max)**
			Llama-3			
(1) No mention	1504	8.01	9.15	2.59	0.30	0.31
(2) Unqualified mention	124	13.48	14.56	2.58	0.30	0.28
(3) Qualified mention	155	12.97	14.23	2.14	0.15	0.15
F-stat $(2, 1780)$		73.62	64.69	14.13	2.77	3.88
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p > 0.05$	$p < 0.05$
Effect size (ω^2)		0.08	0.07	0.02	0.002	0.003

Table 5.13: Changes in average number of misspellings, Llama-3's scores and explanation sentiment across different misspelling categories in Task 7

Misspelling Category	**Sample Count**	**Aspell Misspelling**	**LanguageTool Spelling**	**LLM Score**	**Explanation Sentiment (average)**	**Explanation Sentiment (max)**
			Llama-3			
(1) No mention	659	2.40	4.66	6.80	-0.29	0.05
(2) Unqualified mention	396	4.03	6.30	6.49	-0.33	0.14
(3) Qualified mention	514	6.27	8.11	4.75	-0.60	-0.16
F-stat $(2, 1566)$		138.60	63.91	99.32	45.1	25.98
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		**0.15**	0.07	**0.11**	0.05	0.03

Table 5.14: Changes in average number of grammar mistakes, Llama-3's scores and explanation sentiment across different grammatical categories in Task 1

Grammar Error Category	**Sample Count**	**LanguageTool Grammar**	**LLM Score**	**Explanation Sentiment (average)**	**Explanation Sentiment (max)**
			Llama-3		
(1) No mention	751	2.65	2.92	0.48	0.48
(2) Unqualified mention	608	3.30	2.38	0.19	0.17
(3) Qualified mention	424	3.49	2.14	0.08	0.14
F-stat $(2, 1780)$		19.99	105.21	43.75	47.87
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.02	**0.11**	0.05	0.05

Table 5.15: Changes in average number of grammar mistakes, Llama-3's scores and explanation sentiment across different grammatical categories in Task 7

Grammar Error Category	**Sample Count**	**LanguageTool Grammar**	**LLM Score**	**Explanation Sentiment (average)**	**Explanation Sentiment (max)**
			Llama-3		
(1) No mention	295	1.00	6.71	-0.30	0.08
(2) Unqualified mention	627	1.01	6.89	-0.25	0.16
(3) Qualified mention	647	1.71	4.94	-0.59	-0.18
F-stat $(2, 1780)$		41.73	103.67	55.95	43.35
P-value		$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.05	**0.12**	0.07	0.05

Table 5.16: Comparison of average grades assigned by LLMs across human rater grade categories Task 1

Human Score Class	**Sample Count**	**GPT-3.5T Score**	**Llama-2 Score**	**Llama-3 Score**
Score 1-2	28	1.21	1.68	1.11
Score 2-4	949	1.82	3.39	2.17
Score 4-6	806	2.01	3.95	3.05
F-stat(2,1780)		46.46	399.13	254.96
P-values		$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.05	**0.31**	**0.22**

Table 5.17: Comparison of average grades assigned by LLMs across human rater grade categories Task 7

Human Score Class	**Sample Count**	**GPT-3.5T Score**	**Llama-2 Score**	**Llama-3 Score**
Score 0-4	108	0.84	0.89	2.61
Score 4-8	717	1.93	1.93	5.29
Score 8-15	744	2.38	3.62	7.29
F-stat(2,1566)		39.05	111.2	249.93
P-values		$p < 0.05$	$p < 0.05$	$p < 0.05$
Effect size (ω^2)		0.05	**0.12**	**0.24**

Chapter 6

Results with Prompt Engineering

In this chapter, we delve deeper into our research findings with additional experiments and offer new interpretations of the results. We explore the impact of prompt engineering by incorporating grade-level information and few-shot examples to enhance the effectiveness of our approach.

To explore the impact of different prompt formulations on the correlation between LLM scores and human rater scores, we conducted experiments using subsets of samples from our dataset. To ensure balance across different score ranges, we selected samples from each score point. For Task 1, which has a score range of 1 to 6, we selected a total of 100 essay samples (20 samples from each score point from 2 to 6). Since Score 1 had a minimal number of samples compared to the overall distribution, we excluded it from our selection. Details of the data distribution can be found in Figure 4.1.

Similarly, for Task 7, which has a score range of 0 to 15, we observed that Scores 0 to 3 had very few samples, while the maximum score given by human raters was 12. Thus, we selected 20 samples each from Scores 3 to 12, totalling 200 samples for Task 7. We performed analyses on these samples by varying the prompt information, as described in the following

Table 6.1: Changes in correlation scores after adding grade level of students to the prompt (Increase denoted by and ↑ and decrease denoted by ↓)

		Task 1 (100 samples)		**Task 7** (200 samples)	
		Before	After	Before	After
ChatGPT	Rater 1	0.36	0.40 ↑	0.28	0.37 ↑
	Rater 2	0.39	0.41 ↑	0.23	0.36 ↑
	Llama-2	0.43	0.50 ↑	0.36	0.45 ↑
Llama-2	Rater 1	0.79	0.71 ↓	0.46	0.59 ↑
	Rater 2	0.80	0.76 ↓	0.51	0.60 ↑
	GPT-3.5T	0.43	0.43	0.36	0.37 ↑

subsections. Appendix A includes a detailed prompt design for ChatGPT and Llama-2 to generate the response. Additionally, we used these two prompts to ask Llama-3 for scores and explanations on all the samples from Task 1 (1783 essays) and Task 7 (1569 essays).

6.1 Providing Students' Grade Level Generally Improves Alignment

As shown in Table 6.1, there is a slight increase in correlation scores in most cases, indicating a potential improvement with the addition of grade levels in the prompt. This suggests that including grade levels may help ChatGPT to provide less harsh evaluations. Interestingly, for Llama-2, we observed a decrease or no change in correlation for Task 1, suggesting it may have been lenient previously. However, for Task 7, we observed an increasing trend in correlation scores.

Table 6.2: Changes in correlation scores after adding few-shot examples to the prompt (Increase denoted by and ↑ and decrease denoted by ↓)

		Task 1 (100 samples)		**Task 7** (200 samples)	
		Before	After	Before	After
GPT-3.5T (Two-shot)	Rater 1	0.36	0.65 ↑	0.28	0.50 ↑
	Rater 2	0.39	0.59 ↑	0.23	0.47 ↑
	Llama-2	0.43	0.64 ↑	0.36	0.54 ↑
Llama-2 (One-shot)	Rater 1	0.79	0.69 ↓	0.49	0.52 ↑
	Rater 2	0.80	0.71 ↓	0.51	0.56 ↑
	ChatGPT	0.43	0.30 ↓	0.36	0.41 ↑

6.2 GPT-3.5T Benefits From Few-shot Learning

After incorporating two-shot examples, we noticed a significant enhancement in the correlation between GPT-3.5T and human scoring as shown in Table 6.2. These examples were carefully chosen to align with human scores, with one representing a lower score range and the other a higher score range. However, we encountered limitations with Llama-2, as it failed to respond when the prompt exceeded approximately 2.5k tokens, even though it did not reach the 4k token context length limit. Consequently, we experimented with using only one-shot examples to avoid this issue. For Llama-2, we do not see much difference if we try one-shot learning.

6.3 Prompt Engineering Results with Llama-3

As shown in Table 6.3, after incorporating the students' grade level, we observed an average increase of around 5% in the correlation between human raters and Llama-3 in Task 1. However, in Task 7, adding the grade level resulted in a significant drop in the correlation between human raters and Llama-3, averaging 216%. This change is evident in the score distribution chart in Figure 6.1, where the normal distribution, initially in blue, moves to a

Table 6.3: Changes in correlation scores after adding grade level of students and few-shot examples to the prompt of Llama-3 (Increase denoted by and ↑ and decrease denoted by ↓)

		Task 1		**Task 7**	
Score		Before	After	Before	After
Llama-3 (with grade level)	Rater 1	0.52	0.54 ↑	0.49	0.15 ↓
	Rater 2	0.49	0.52 ↑	0.52	0.17 ↓
	GPT-3.5T	0.46	0.44 ↓	0.40	0.14 ↓
	Llama-2	0.54	0.58 ↑	0.52	0.13 ↓
Llama-3 (two-shot)	Rater 1	0.52	0.62 ↑	0.49	0.59 ↑
	Rater 2	0.49	0.61 ↑	0.52	0.62 ↑
	GPT-3.5T	0.46	0.41 ↓	0.40	0.43 ↑
	Llama-2	0.54	0.66 ↑	0.52	0.60 ↑

lower grade range in green bars after the grade level is added. The diagram also shows the score distribution for different prompts. Additionally, as seen in Table 6.3, after including two-shot examples, we observed an average increase of around 22% in the correlation between human raters and Llama-3 in Task 1, and an average increase of around 20% in Task 7.

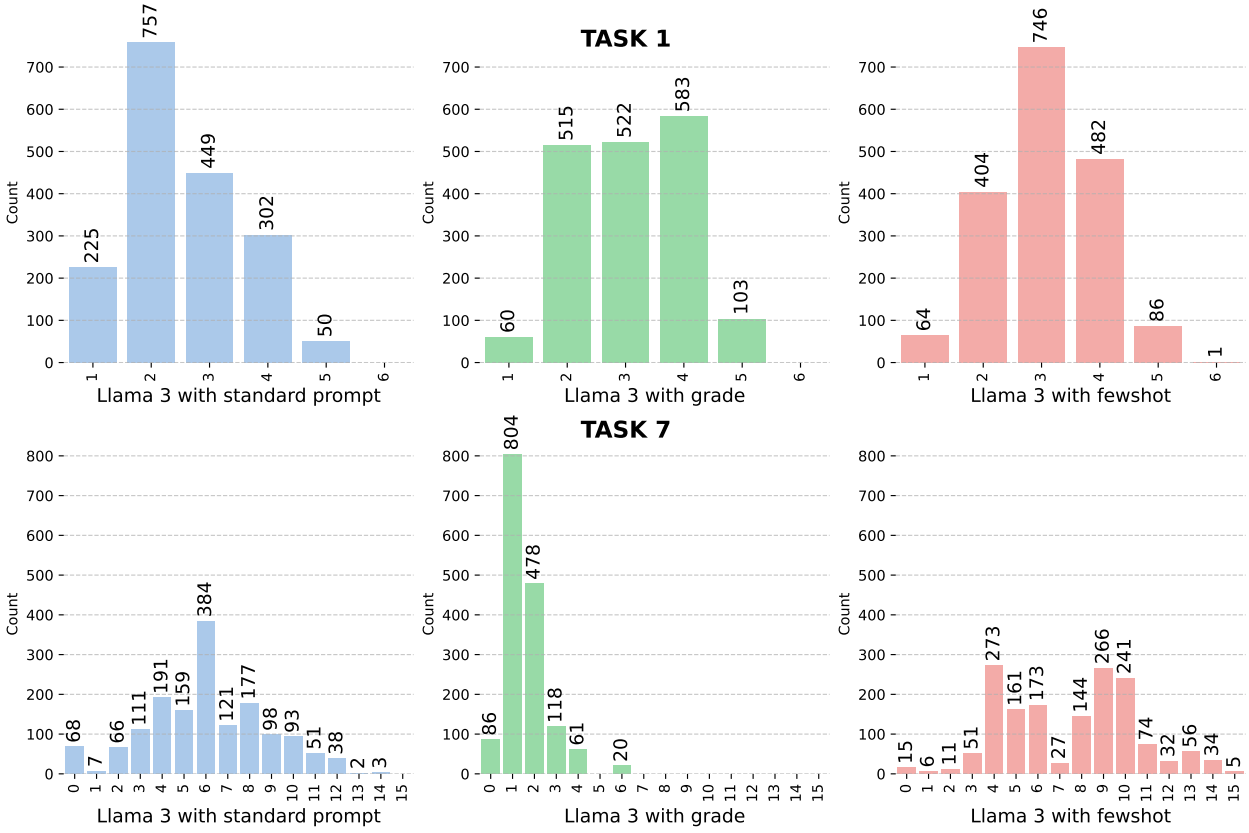

Figure 6.1: Score distribution across different prompts on Llama-3

Chapter 7

Conclusions, Limitations and Future Work

7.1 Conclusions

Our research aims to evaluate the effectiveness of Large Language Models (LLMs) in assessing essay quality, generally, and in Automatic Essay Scoring (AES), specifically. By comparing scores generated by LLMs to those given by human raters, we have uncovered several insights into the performance and alignment of these models.

We found that while GPT-3.5T and Llama-2 exhibit remarkable capabilities in understanding essay prompts and generating coherent responses, there is a significant disparity between their assessments and those of human raters. Human raters show strong inter-rater agreement, often awarding higher scores to longer essays and excelling in evaluating the logical progression of ideas. In contrast, LLMs demonstrate distinct grading behaviors, with GPT-3.5T being stricter in its scoring compared to both human raters and Llama-2. On the other hand, Llama-2 shows a closer alignment to human scoring patterns.

Both human and LLM scores were minimally influenced by readability indices; however, essays that were harder to read tended to receive higher marks from both humans and LLMs. LLMs demonstrated a strong ability to detect and account for spelling and grammatical errors, highlighting a key difference from human raters who face challenges in this area. However, specific trait-wise human scores have shown promise in identifying mechanical mistakes more effectively than an overall score provided by human raters.

In terms of providing explanations for the grades, GPT-3.5T often delivered explanations in a harsh tone, whereas Llama-2 offered generally less negative explanations. Both LLMs showed a better understanding of the prompt and provided explanations that aligned with the assigned scores, indicating a degree of consistency. Lower scores are accompanied by more negative explanations, while higher scores have led to more positive explanations, suggesting that LLMs are aware of the scores they provide and do not generate explanations randomly.

Our findings suggest that while LLMs hold potential for automatic essay scoring (AES) tasks, they should be used with human supervision. LLMs may not completely replace human raters, but their ability to understand guidelines, coupled with their consistent explanation capabilities, make them valuable tools for educators facing the demands of modern education.

Additionally, our experiment with one of the more recent LLMs, Llama-3, reveals substantial improvements. In the near future, new generations of LLMs could become even more powerful, potentially sufficient for various domains, and capable of scoring essays and providing prompt explanations autonomously.

7.2 Limitations and Future work

One limitation of this study is that we evaluated the explanations generated by LLMs based solely on their ability to justify the provided grade, without considering how these explana-

tions might affect learners. This narrow focus overlooks the potential impact of formative feedback on student learning and motivation, which is crucial for educational applications.

Another limitation is the quality and nature of the dataset used. The dataset primarily contains numeric scores, which tend to align with basic length features of the essays. This suggests that human graders may prioritize certain aspects of the rubric over others. For example, longer essays with numerous spelling mistakes often receive higher scores from human raters, potentially because they value coherence and interest over technical accuracy. This inconsistency indicates that the human graders' evaluations may not perfectly adhere to the rubric, which could lead to misleading conclusions when comparing human and LLM scores. We should also point out that Task 1 and Task 7 were tasks given to students from different grades (7 and 8) which prevents direct comparisons of results across tasks as the LLMs may perform differently for these different grade levels.

Additionally, LLMs tend to apply the rubric strictly and may not compensate for errors in the same way human graders do. This could explain why LLMs, particularly ChatGPT, tend to be harsher and less aligned with human scores. Human graders might overlook certain mistakes in favor of overall essay quality, while LLMs strictly follow the rubric guidelines. Therefore, comparing LLM scores with human scores from the ASAP dataset, as done in our study, or training machine learning models to replicate these scores, as done in previous studies, could be misguided. Such comparisons might not accurately reflect the nuanced judgments made by human graders.

Future research should investigate the impact of LLM-generated explanations on student revisions and writing improvement. Additionally, our study focused on GPT-3.5T and Llama-2; future work should include comparisons with other models like Google Gemini [56] and similar tools to provide a more comprehensive evaluation of automatic scoring in education. Further exploration can be done on the effects of fine-tuning models or prompting

LLMs with more annotated scores and explanations data by human raters to better mimic human grading practices and provide valuable textual insights. It is also important to note that the performance of AI tools is subject to their development stage at the time of the experiment. As these tools continue to evolve with new training data, future studies should reassess their effectiveness and alignment with human grading to ensure continued relevance and accuracy in educational assessments.

Bibliography

[1] Graham Gibbs and Claire Simpson. Conditions under which assessment supports students' learning. *Learning and teaching in higher education*, pages 3–31, 2005.

[2] Gordon Joughin. *Assessment, Learning and Judgement in Higher Education: A Critical Review*, pages 1–15. Springer Netherlands, 2009. doi: 10.1007/978-1-4020-8905-3_2.

[3] Karen Scouller. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher education*, (4): 453–472, 1998.

[4] Benjamin Kamga Fomba, Dieu Ne Dort Fokam Talla, and Paul Ningaye. Institutional quality and education quality in developing countries: Effects and transmission channels. *Journal of the Knowledge Economy*, pages 86–115, 2023.

[5] Nizamettin Koc and Bekir Celik. The impact of number of students per teacher on student achievement. *Procedia-Social and Behavioral Sciences*, pages 65–70, 2015.

[6] Dadi Ramesh and Suresh Kumar Sanampudi. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, pages 2495–2527, 2022.

[7] Imen Azaiz, Natalie Kiesler, and Sven Strickroth. Feedback-generation for programming exercises with gpt-4. *arXiv preprint arXiv:2403.04449*, 2024.

[8] David M Williamson, Isaac I Bejar, and Anne S Hone. Mental model comparison of automated and human scoring. *Journal of Educational Measurement*, pages 158–184, 1999.

[9] Tracey A Benson and Sarah E Fiarman. *Unconscious bias in schools: A developmental approach to exploring race and racism*. Harvard Education Press, 2020.

[10] Andrea Horbach and Torsten Zesch. The influence of variance in learner answers on automatic content scoring. In *Frontiers in education*, page 28. Frontiers Media SA, 2019.

[11] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[12] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc., 2022.

[13] Ömer Aydın and Enis Karaarslan. Is chatgpt leading generative ai? what is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*, pages 118–134, 2023.

[14] Louie Giray. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, 51(12):2629–2633, 2023.

[15] Zhiwei Jiang, Meng Liu, Yafeng Yin, Hua Yu, Zifeng Cheng, and Qing Gu. Learning from graph propagation via ordinal distillation for one-shot automated essay scoring. In *Proceedings of the Web Conference 2021*, pages 2347–2356, 2021.

[16] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[17] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[18] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[19] Julie Cheville. Automated scoring technologies and the rising influence of error. *The English Journal*, pages 47–52, 2004.

[20] Beata Lewis Sevcikova. Human versus Automated Essay Scoring: A Critical Review. *Arab World English Journal*, pages 157–174, 2018. doi: 10.24093/awej/vol9no2.11.

[21] Mo Zhang. Contrasting automated and human scoring of essays. *R & D Connections*, pages 1–11, 2013.

[22] Neslihan Süzen, Alexander N Gorban, Jeremy Levesley, and Evgeny M Mirkes. Automatic short answer grading and feedback using text mining methods. *Procedia computer science*, pages 726–743, 2020.

[23] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In *IJCAI*, pages 6300–6308, 2019.

[24] Ellis B. Page. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243, 1966.

[25] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

[26] Peter W Foltz, Darrell Laham, and Thomas K Landauer. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944, 1999.

[27] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1193.

[28] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pages 153–162, 2017.

[29] Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 431–439, 2015.

[30] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, 2020.

[31] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, 2022. doi: 10.18653/v1/2022.naacl-main.249.

[32] Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *EMNLP*, pages 1560–1569, 2020. doi: 10.18653/v1/2020.findings-emnlp.141.

[33] Abdolvahab Khademi. Can chatgpt and bard generate aligned assessment items? a reliability analysis against human performance. *Journal of Applied Learning and Teaching*, 2023.

[34] Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, page 100050, 2023.

[35] Ehsan Latif and Xiaoming Zhai. Fine-tuning chatgpt for automatic scoring. *Computers and Education: Artificial Intelligence*, page 100210, 2024.

[36] Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. Can large language models automatically score proficiency of written essays? *arXiv preprint arXiv:2403.06149*, 2024.

[37] Robert K. Helmeczi, Savas Yıldırım, Mucahit Cevik, and Sojin Lee. Few shot learning approaches to essay scoring. In *Proceedings of the Canadian Conference on Artificial Intelligence*, 2023.

[38] Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, 2020. doi: 10.14569/IJACSA.2020.0110538.

[39] Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495, 2022. doi: 10.18653/v1/2022.naacl-main.106.

[40] Maciej Pankiewicz and Ryan S Baker. Large language models (gpt) for automating feedback on programming assignments. *arXiv preprint arXiv:2307.00150*, 2023.

[41] Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, page 100199, 2024. doi: 10.1016/j.caeai.2023.100199.

[42] Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. The hewlett foundation: Automated essay scoring, 2012. URL https://kaggle.com/competitions/asap-aes.

[43] Sandeep Mathias and Pushpak Bhattacharyya. ASAP++: Enriching the ASAP auto-

mated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

[44] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

[45] RJ Senter and Edgar A Smith. Automated readability index. Technical report, DTIC document, 1967.

[46] Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, page 283, 1975.

[47] Chall J Dale E. A formula for predicting readability. *Educational Research Bulletin*, 1948.

[48] Robert Gunning. The technique of clear writing. *McGraw-Hill*, 1952.

[49] Paul R Fitzsimmons, BD Michael, Joane L Hulley, and G Orville Scott. A readability assessment of online parkinson's disease information. *The journal of the Royal College of Physicians of Edinburgh*, pages 292–296, 2010.

[50] Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, pages 141–179, 2021. doi: 10.1162/coli_a_00398.

[51] Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, 2015. doi: 10.3115/v1/W15-0626.

[52] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623, 2021. doi: 10.1145/3442188.3445922.

[53] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[54] AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[55] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard, 2023. URL https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

[56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Appendix A

Sample Prompt and Output from GPT-3.5T and Llama-2

A.1 Task 1 Question (given to the students)

More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends.

Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.

A.2 Task 1 Rubric

Score Point 1: An undeveloped response that may take a position but offers no more than very minimal support. Typical elements:

- Contains few or vague details.
- Is awkward and fragmented.
- May be difficult to read and understand.
- May show no awareness of audience.

Score Point 2: An under-developed response that may or may not take a position. Typical elements:

- Contains only general reasons with unelaborated and/or list-like details.
- Shows little or no evidence of organization.
- May be awkward and confused or simplistic.
- May show little awareness of audience.

Score Point 3: A minimally-developed response that may take a position, but with inadequate support and details. Typical elements:

- Has reasons with minimal elaboration and more general than specific details.
- Shows some organization.
- May be awkward in parts with few transitions.
- Shows some awareness of audience.

Score Point 4: A somewhat-developed response that takes a position and provides adequate support. Typical elements:

- Has adequately elaborated reasons with a mix of general and specific details.

- Shows satisfactory organization.
- May be somewhat fluent with some transitional language.
- Shows adequate awareness of audience.

Score Point 5: A developed response that takes a clear position and provides reasonably persuasive support. Typical elements:

- Has moderately well elaborated reasons with mostly specific details.
- Exhibits generally strong organization.
- May be moderately fluent with transitional language throughout.
- May show a consistent awareness of audience.

Score Point 6: A well-developed response that takes a clear and thoughtful position and provides persuasive support. Typical elements:

- Has fully elaborated reasons with specific details.
- Exhibits strong organization.
- Is fluent and uses sophisticated transitional language.
- May show a heightened awareness of audience.

A.3 Task 1 Prompt (Zero-shot)

Question: {Question Appendix A.1}
Given the criteria: {Rubric Appendix A.2}

How would you grade the following essay? Provide a short explanation.

Here is the essay
Essay: People are probly on the computers chating there e-mail or making a websit

or they are on youtube. Maybe the person does not like to go outside and enjoy nature or spending time with there family there are probly checking out new websites and playing games or the problay on the comper to book there vactions of seeing what is on sell a welmart. They are probly looking at or buy a new truk or they are looking at the history on all the wars and how they were started.

Note: Please ensure that the score and explanation are formatted like the example given below.
Score: [Score according to your evaluation]
Explanation: [Your detailed feedback].

Please select a score between 1 and 6 for the essay based on the provided criteria and explain your reasoning in detail.

Answer:

A.4 Task 1: Prompting with the grade level of students

System prompt: You are a helpful essay grading assistant. Answer the question and follow the instructions carefully.

Question: {Question Appendix A.1}
Given the criteria: {Rubric Appendix A.2}

```
How would you grade the following essay? This essay is written by a student of
    grade 8. Provide a short explanation.

Here is the essay
Essay: People are probly on the computers chating there e-mail or making a websit
     or they are on youtube. Maybe the person does not like to go outside and
    enjoy nature or spending time with there family there are probly checking out
    new websites and playing games or the problay on the comper to book there
    vactions of seeing what is on sell a welmart. They are probly looking at or
    buy a new truk or they are looking at the history on all the wars and how they
     were started.

Note: Please ensure that the score and explanation are formatted like the example
     given below.
Score: [Score according to your evaluation]
Explanation: [Your detailed feedback].

You need to grade according to the expected level of grade 8 writing skills.
    Please select a score between 1 and 6 for the essay based on the provided
    criteria and explain your reasoning in detail.

# Answer:
```

A.5 Task 1: Few-shot learning

```
# Question: {Question Appendix A.1}
Given the criteria: {Rubric Appendix A.2}
```

* Example Essay 1:

Coumputors are mind browning on what they can do. They can make or break somone financhely. They also can perduce amasing graphis that aritist cant perduce. Coumputrs are also for great for recer antal acitivitys, and they also take your life away. I not talking about a real life @CAPS1 talking about the activitys you do away from home. Now lets see the benifits and problems we come so far on how we handle money. We use to put it in chets and hide them now all we do is click and type in a securty number. people can also invist in items but invisting is tricky you dont always win. You can win small or lose big they say @CAPS2 coumputers today are hd. The screens are made of a liquid crystal fourmula that hardens. The lights in the back pixels do most of the wort to bring stunning graphis to life. On coumputers today people make money by selling graphis they made on coumputers. So this would be another plus on what a coumputor dose to a human. I went to a school to lern how to perduce great graphis. All teens kids and adults have used a coumputer for recerantol acivitys. On todays coumputors you can play games watch videos, and listen on coumputors. People do this to engoy then selfs. Some people ran get in trouble for fooling around on the web. But there is nothing stressed out and just playing some games for a while, personly I like to go on youtube and find funny vidos after school. I use to own a youtube account and get many views. The worst thing about coumpurts is it can ruin plan gobo or somones socal life some people get addicted and cant stop being on the coumputor until it is takin away. For example some people play world of war craft get addicted and cant stop! I had a problem when I had to chek you tube a lot. Well I hoped you readers. No matter what happens I think coumputors bennafit more than the

problems

Human rater score 1: 4

Example Answer 1:

Score: 4

Explanation: The essay presents a clear position on the effects of computers on people, arguing that computers have both positive and negative effects. The writer provides some reasons and details to support their position, such as the ability to learn new skills, produce amazing graphics, and enjoy recreational activities. However, they also acknowledge the potential negative effects, such as addiction and social isolation.

The organization of the essay is satisfactory, with a clear introduction, body, and conclusion. The writer uses some transitional phrases to connect their ideas, but the transitions could be smoother and more cohesive. The essay shows some awareness of the audience, with the writer addressing the readers directly and using a conversational tone.

One area for improvement is the use of language. The essay contains several spelling and grammar errors, which detract from the overall quality of the writing. Additionally, the writer could benefit from providing more specific and detailed examples to support their points. For instance, they could describe a particular skill they learned from using a computer or provide an example of a successful graphic design project.

Overall, the essay is somewhat developed but could benefit from more detail and refinement in the writing. Therefore, a score of 4 is appropriate.

* Example Essay 2:

The computers are cool. Do you now I werpsite is @NUM1.com. We nede to time the

lemt lo-menteno it. I love the compueters. We nede @NUM2 make a computers are grene compurers.

Human rater score 2: 1

Example Answer 2:

Score: 1

Explanation:The essay provided is undeveloped and lacks clear reasoning or persuasive support. It contains very minimal details and is awkward and fragmented, making it difficult to read and understand. The essay does not provide any specific examples or reasons to support the author's claim that computers are "cool". The language used is also inappropriate and contains spelling and grammatical errors. The essay does not show any awareness of the audience or attempt to persuade the readers to agree with the author's position. Therefore, it scores a 1.

It is worth noting that this essay is difficult to read and understand due to its poor writing quality. Therefore, it is not possible to provide detailed feedback on its content.

How would you grade the following essay? Provide a short explanation.

Here is the essay

Essay: People are probly on the computers chating there e-mail or making a websit or they are on youtube. Maybe the person does not like to go outside and enjoy nature or spending time with there family there are probly checking out new websites and playing games or the problay on the comper to book there vactions of seeing what is on sell a welmart. They are probly looking at or buy a new truk or they are looking at the history on all the wars and how they

were started.

Note: Please ensure that the score and explanation are formatted like the example given below.

Score: [Score according to your evaluation]

Explanation: [Your detailed feedback].

Please select a score between 1 and 6 for the essay based on the provided criteria and explain your reasoning in detail.

Answer:

A.6 Task 7 Question (given to the students)

Write about patience. Being patient means that you are understanding and tolerant . A patient person experience difficulties without complaining.

Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.

A.7 Task 7 Rubric

A rating of 0-3 on the following four traits:

Ideas (points doubled)

Score 3: Tells a story with ideas that are clearly focused on the topic and are thoroughly developed with specific, relevant details.

Score 2: Tells a story with ideas that are somewhat focused on the topic and are developed with a mix of specific and/or general details.

Score 1: Tells a story with ideas that are minimally focused on the topic and developed with limited and/or general details. Score 0: Ideas are not focused on the task and/or are undeveloped.

Organization

Score 3: Organization and connections between ideas and/or events are clear and logically sequenced.

Score 2: Organization and connections between ideas and/or events are logically sequenced.

Score 1: Organization and connections between ideas and/or events are weak.

Score 0: No organization evident.

Style

Score 3: Command of language, including effective and compelling word choice and varied sentence structure, clearly supports the writer's purpose and audience.

Score 2: Adequate command of language, including effective word choice and clear sentences, supports the writer's purpose and audience.

Score 1: Limited use of language, including lack of variety in word choice and sentences, may hinder support for the writer's purpose and audience.

Score 0: Ineffective use of language for the writer's purpose and audience.

Conventions

Score 3: Consistent, appropriate use of conventions of Standard English for

grammar, usage, spelling, capitalization, and punctuation for the grade level.
Score 2: Adequate use of conventions of Standard English for grammar, usage, spelling, capitalization, and punctuation for the grade level.
Score 1: Limited use of conventions of Standard English for grammar, usage, spelling, capitalization, and punctuation for the grade level.
Score 0: Ineffective use of conventions of Standard English for grammar, usage, spelling, capitalization, and punctuation.

A.7.1 Task 7: Prompting with the grade level of students

Question: {Question Appendix A.6}
Given the criteria: {Rubric Appendix A.7}

How would you grade the following essay? This essay is written by a student of grade 7. Provide a short explanation.

Here is the essay
Essay: I am patent i always have to wait for somthing or earn it. I go to my grouping with my parents but I have to wat. I have to wait to go to my aunts. The worst time to wat is going to @CAPS1/a up @LOCATION1 and camping.

Note: Please ensure that the score and explanation are formatted like the example given below.
Score: [Score according to your evaluation]
Explanation: [Your detailed feedback].

You need to grade according to the expected level of grade 7 writing skills.

Please select a score between 0 and 15 for the essay based on the provided criteria and explain your reasoning in detail.

A.8 Task 7: Few-shot Learning

Question: {Question Appendix A.6}

Given the criteria: {Rubric Appendix A.7}

* Example Essay 1:

Being patient, what does it mean? It means that; "you are understanding and tolerant. A patient person experiences difficulties without complaining." @CAPS1, being a patient person is difficult. One time, it was hard for me to be a patient person, I didn't complain though. I was overjoyed to be going to a water park on my birthday. I wanted to party and get my swim on but it was almost @NUM1 weeks away. I didn't want to wait any longer, but I knew that being patient would make my experience even better. I was tolerant, understanding, and patient. The week dragged on. I wanted to complain and rip my hair out because it was like the earth was turning @NUM2, @NUM3 times slower, but I knew that I couldn't! When the thought of me going to the water park got stuck in my brain like glue, the problem was clear. If I didn't think about it, time would fly by again. I had to be patient because I didn't want to annoy anyone with my complaints. When the day came, I had a blast. Being patient always turns out good!

Human rater provided score on this essay: ideas: 2 organization: 2 style: 2 conventions: 2 total: 8

Example Answer 1:

Score: ideas: 2 organization: 2 style: 2 conventions: 2 total: 8

Explanation: Ideas: The essay tells a story about a time when the writer had to be patient. It focuses on the topic and provides some specific details, such as wanting to go to a water park on their birthday. However, the ideas could have been more thoroughly developed with more specific and relevant details. Organization: The essay has a clear introduction and conclusion. The events are logically sequenced, with the writer describing their desire to go to the water park, the challenges they faced in being patient, and the outcome of their patience. However, there could have been more cohesion and transitions between the events. Style: The writer demonstrates an adequate command of language. They use some effective word choices and clear sentences. However, there lacks variety in word choice and sentence structure, and the language could have been more compelling to support the writer's purpose and audience. Conventions: The writer shows an adequate use of conventions of Standard English. There are some errors in grammar, usage, spelling, capitalization, and punctuation, but they do not significantly hinder the reader's understanding. Overall, the essay meets the basic requirements but could benefit from more development in ideas, stronger organization and transitions, more compelling language, and improved conventions.

* Example Essay 2:

I was patient on the trip to @LOCATION1 it took two days to get down there I got hungery and had to go to the bathroom when ever I was hungery I had to what until next stop when we got gas so when it was on full I had to be patient with out complaining. So, I had to be patient until the gas tank was empty that's when we got gas and I could eat something or I had to be patient until there was a @CAPS1 on the right hand or a test area because the have a vending

machine there and bathrooms there too. On the right hand side because we can get off on get back on the high way. That's when I had to be patient.

Human rater provided score on this essay: ideas: 1 organization: 1 style: 1 conventions: 1 total: 4

Example Answer 2:

Score: ideas: 1 organization: 1 style: 1 conventions: 1 total: 4

Explanation: The ideas presented in the essay are minimally focused and underdeveloped. There is limited detail provided to support the topic of patience. Additionally, the organization of the essay is weak, as there is no clear sequencing of events or logical connections between ideas. The style of writing is limited, with lack of variety in word choice and sentence structure . There are also several grammatical errors and lack of proper punctuation throughout the essay. Overall, the essay lacks coherence and fails to effectively convey the concept of patience.

How would you grade the following essay? Provide a short explanation.

Here is the essay

Essay: Patience is usually good or bad for me. Here are a few reasons why patience can be hard or stresful. If you are waiting to get a shot or life threat results. A way to avoid normal patience like being board is to get some notes or school work to do or an ipod. Patience is everywere. For example carrides, birthday parties, vacation commercial, school being over, restaurants, holidays or whenever you say I cannot wait, and etc. to sam up a few ways about patience in my way of description.

Note: Please ensure that the score and explanation are formatted like the example given below.
Score: [Score according to your evaluation]
Explanation: [Your detailed feedback].

Please select a score between 0 and 15 for the essay based on the provided criteria and explain your reasoning in detail.

Answer:

A.9 Response from OpenAI ChatGPT:

Score: 1
Explanation: This essay is an undeveloped response that offers very minimal support. The essay lacks organization and coherence. The language is fragmented and contains many grammatical errors. The ideas are unclear and the details provided are vague and nonspecific. The essay does not demonstrate any awareness of the audience. Overall, it is difficult to read and understand .

A.10 Response from Llama-2:

Score: 1
Explanation: The essay provides minimal support for the position that people are spending too much time on their computers and not enough time on other activities. The essay merely states that people are probably using their

computers for various activities without providing any specific examples or evidence. There is no clear organization, and the writing is awkward and difficult to read in some parts. The essay shows little awareness of the audience and does not attempt to persuade the readers to agree with the writer 's position. Therefore, a score of 1 is appropriate for this essay.

Appendix B

Task 7 trait-wise detail information

B.1 Task 7 trait-wise descriptive statistics

Table B.1: Descriptive statistics Task 7 trait: Ideas

	Rater 1	**Rater 2**	**ChatGPT**	**Llama-2**	**Llama-3**
Mean	1.84	1.85	0.6	0.88	2.21
SD	0.86	0.88	0.6	0.96	1.43
Min	0.00	0.00	0.0	0.00	0.00
Median	2.00	2.00	1.0	1.00	2.00
Max	3.00	3.00	2.0	8.00	6.00

Table B.2: Descriptive statistics Task 7 trait: Organization

	Rater 1	**Rater 2**	**ChatGPT**	**Llama-2**	**Llama-3**
Mean	2.02	2.03	0.57	0.89	1.40
SD	0.72	0.72	0.79	0.87	0.72
Min	0.00	0.00	0.00	0.00	0.00
Median	2.00	2.00	0.00	1.00	1.00
Max	3.00	3.00	5.00	6.00	12.00

Table B.3: Descriptive statistics Task 7 trait: Style

	Rater 1	**Rater 2**	**ChatGPT**	**Llama-2**	**Llama-3**
Mean	1.99	2.00	0.37	0.22	1.03
SD	0.61	0.64	0.51	0.67	0.46
Min	0.00	0.00	0.00	0.00	0.00
Median	2.00	2.00	0.00	0.00	1.00
Max	3.00	3.00	2.00	7.00	3.00

Table B.4: Descriptive statistics Task 7 trait: Conventions

	Rater 1	**Rater 2**	**ChatGPT**	**Llama-2**	**Llama-3**
Mean	2.17	2.17	0.59	0.73	0.68
SD	0.69	0.69	0.61	0.63	0.63
Min	0.00	0.00	0.00	0.00	0.00
Median	2.00	2.00	1.00	1.00	1.00
Max	3.00	3.00	5.00	5.00	3.00

B.2 Task 7 trait-wise score distribution

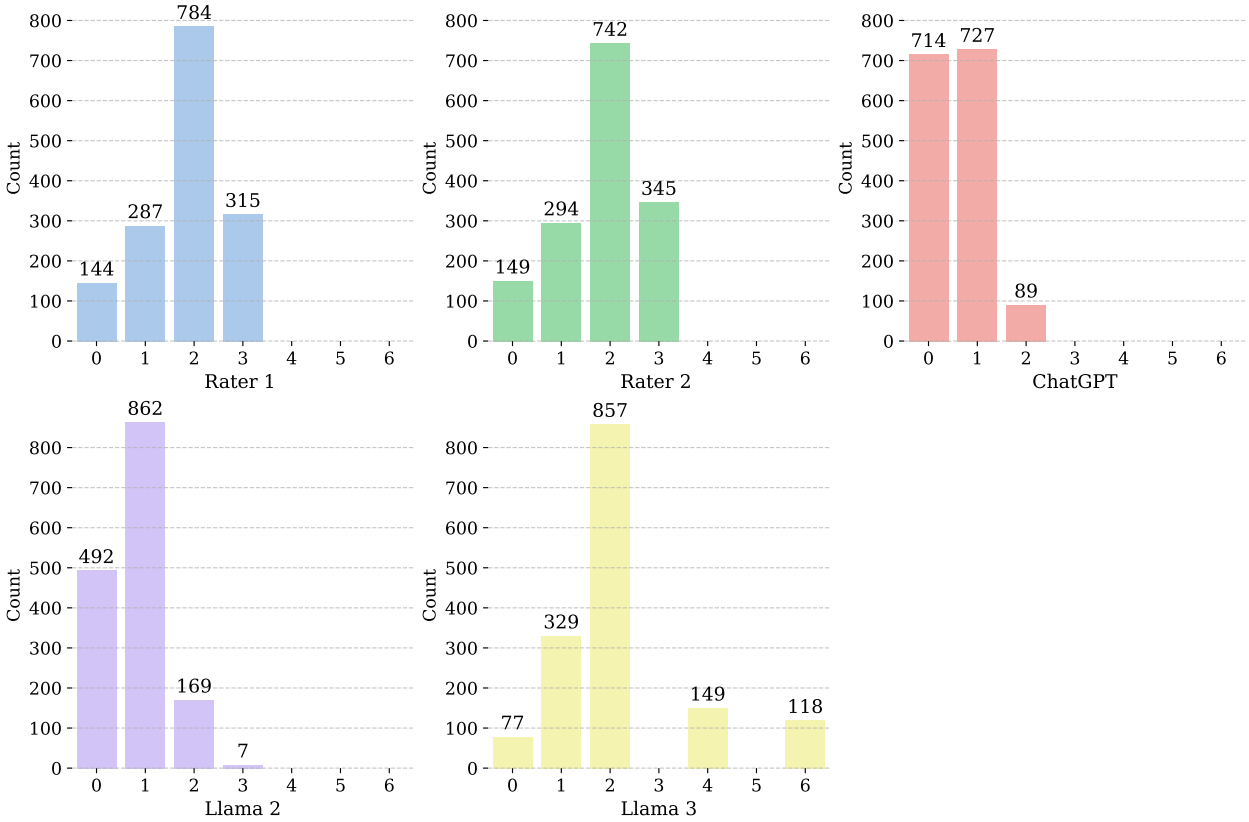

Figure B.1: Score distribution of human raters and LLMs for Task 7 excluding 39 outlier samples (trait ideas)

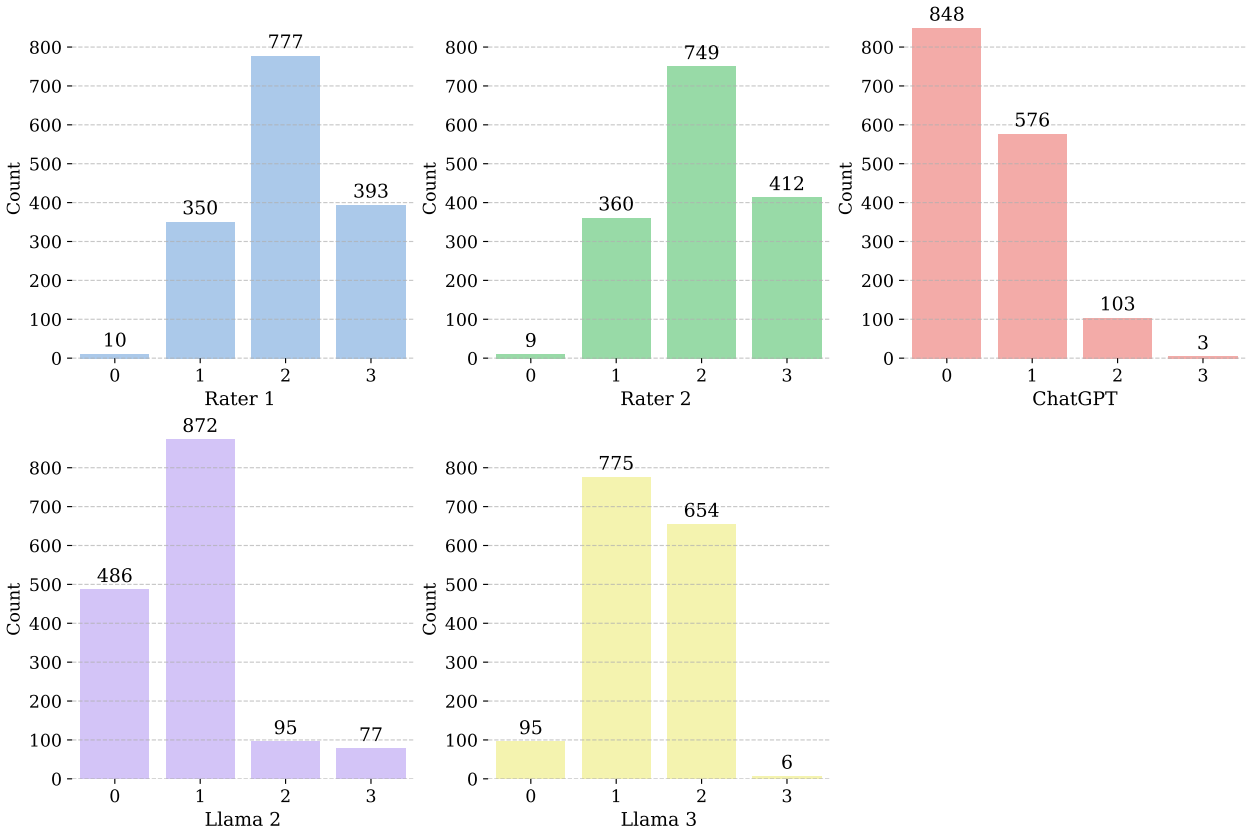

Figure B.2: Score distribution of human raters and LLMs for Task 7 excluding 39 outlier samples (trait organization)

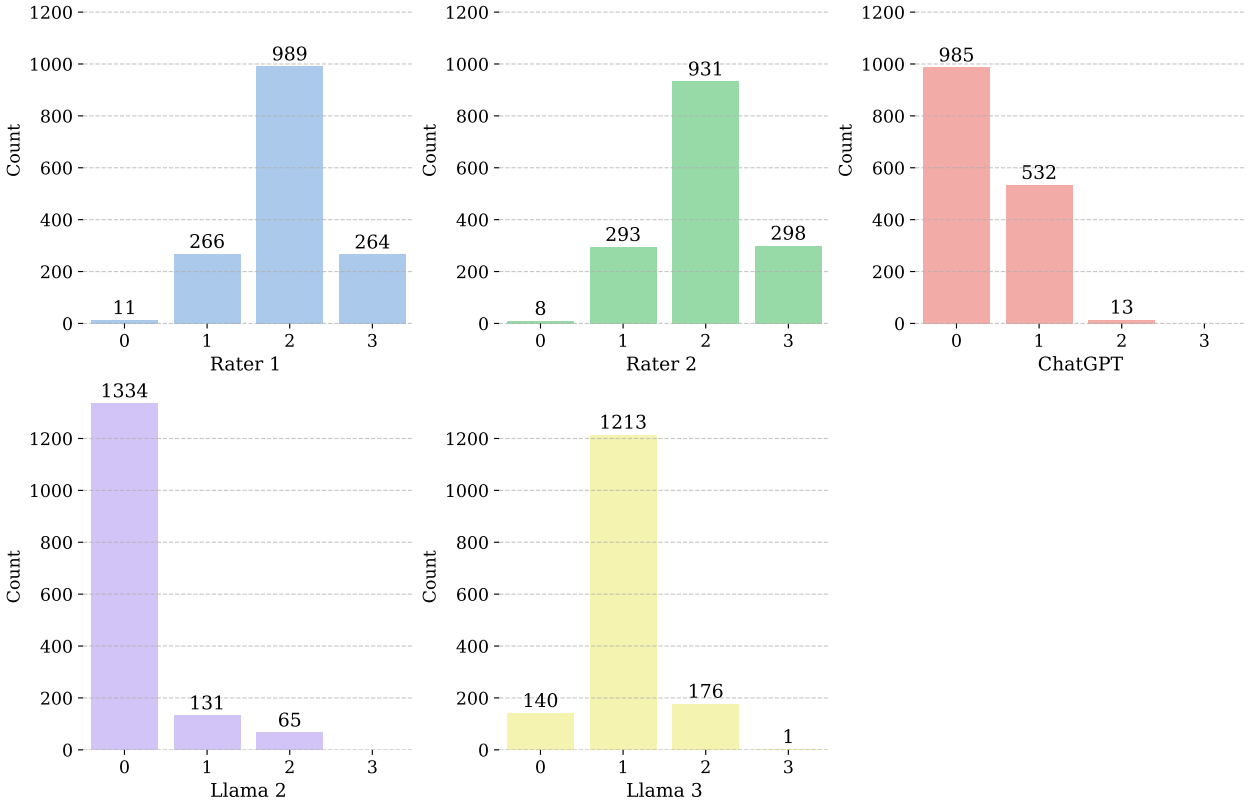

Figure B.3: Score distribution of human raters and LLMs for Task 7 excluding 39 outlier samples (trait style)

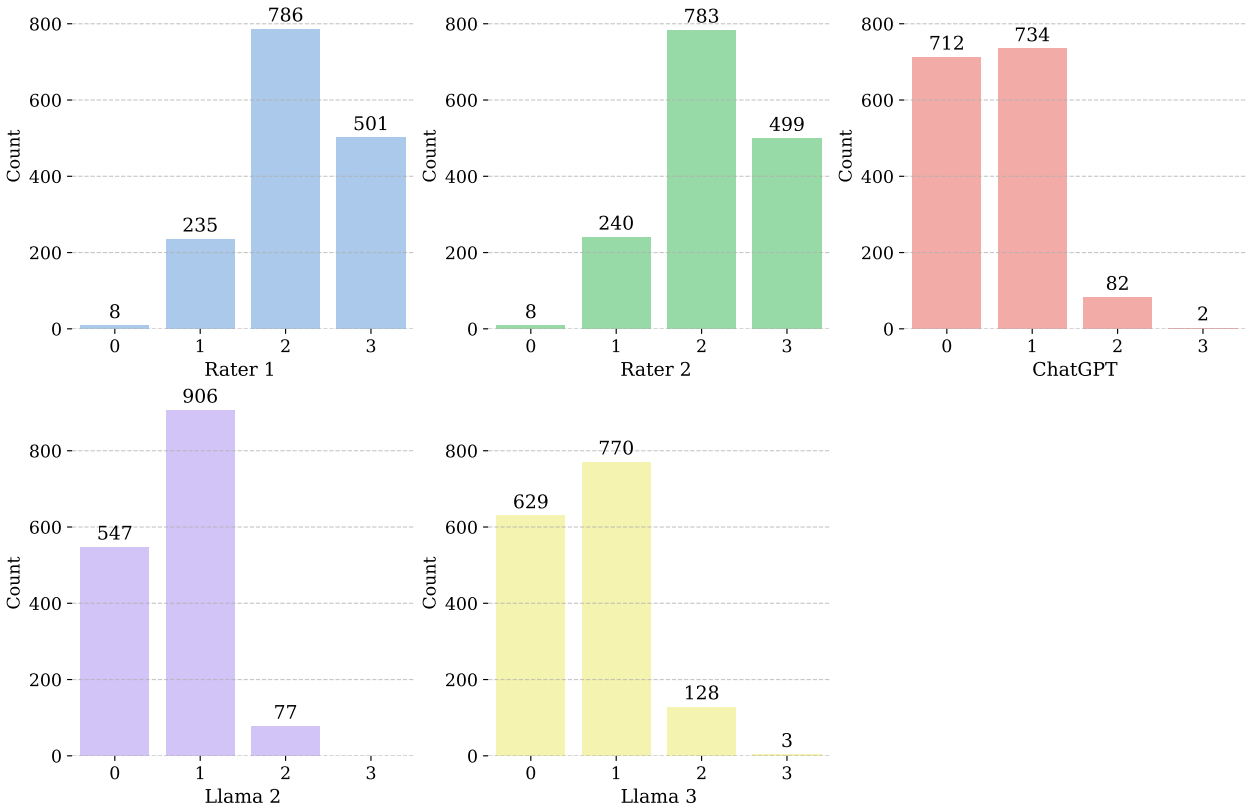

Figure B.4: Score distribution of human raters and LLMs for Task 7 excluding 39 outlier samples (trait conventions)

Appendix C

Statistical Analysis Results

Table C.1: P value, Confidence Interval and Effect sizes of Table 4.3 and 5.3

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Rater 2	0.72	< 0.05	[0.7, 0.74]	-0.01	Small effect size
Rater 1 vs. GPT-3.5T	0.23	< 0.05	[0.19, 0.28]	3.3	Large effect size
Rater 1 vs. Llama-2	0.59	< 0.05	[0.56, 0.62]	0.85	Large effect size
Rater 1 vs. Llama-3	0.52	< 0.05	[0.49, 0.55]	1.85	Large effect size
Rater 2 vs. GPT-3.5T	0.21	< 0.05	[0.16, 0.25]	3.38	Large effect size
Rater 2 vs. Llama-2	0.58	< 0.05	[0.55, 0.61]	0.88	Large effect size
Rater 2 vs. Llama-3	0.49	< 0.05	[0.45, 0.52]	1.88	Large effect size
GPT-3.5T vs. Llama-2	0.32	< 0.05	[0.27, 0.36]	-2.81	Large effect size
GPT-3.5T vs. Llama-3	0.46	< 0.05	[0.42, 0.49]	-0.8	Large effect size
Llama-2 vs. Llama-3	0.54	< 0.05	[0.51, 0.57]	1.26	Large effect size

Table C.2: P value, Confidence Interval and Effect sizes of Table 4.4 and 5.4 (Task 7 overall scores)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Rater 2	0.72	< 0.05	[0.7, 0.74]	-0.01	Small effect size
Rater 1 vs. GPT-3.5T	0.18	< 0.05	[0.13, 0.23]	2.77	Large effect size
Rater 1 vs. Llama-2	0.38	< 0.05	[0.34, 0.43]	2.1	Large effect size
Rater 1 vs. Llama-3	0.49	< 0.05	[0.45, 0.53]	0.76	Large effect size
Rater 2 vs. GPT-3.5T	0.19	< 0.05	[0.15, 0.24]	2.71	Large effect size
Rater 2 vs. Llama-2	0.4	< 0.05	[0.36, 0.45]	2.07	Large effect size
Rater 2 vs. Llama-3	0.52	< 0.05	[0.48, 0.55]	0.76	Large effect size
GPT-3.5T vs. Llama-2	0.3	< 0.05	[0.25, 0.34]	-0.26	Medium effect size
GPT-3.5T vs. Llama-3	0.4	< 0.05	[0.35, 0.44]	-1.71	Large effect size
Llama-2 vs. Llama-3	0.52	< 0.05	[0.48, 0.56]	-1.25	Large effect size

Table C.3: P value, Confidence Interval and Effect sizes of Table 4.5 and 5.5

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Content vs. Rater 1	0.66	< 0.05	[0.63, 0.68]	-0.45	Medium effect size
Content vs. Rater 2	0.67	< 0.05	[0.64, 0.69]	-0.46	Medium effect size
Content vs. GPT-3.5T	0.34	< 0.05	[0.3, 0.38]	2.42	Large effect size
Content vs. Llama-2	0.61	< 0.05	[0.58, 0.64]	0.27	Medium effect size
Content vs. Llama-3	0.64	< 0.05	[0.61, 0.66]	1.3	Large effect size
Organization vs. Rater 1	0.63	< 0.05	[0.6, 0.66]	-0.58	Large effect size
Organization vs. Rater 2	0.63	< 0.05	[0.6, 0.66]	-0.6	Large effect size
Organization vs. GPT-3.5T	0.36	< 0.05	[0.32, 0.4]	2.36	Large effect size
Organization vs. Llama-2	0.6	< 0.05	[0.56, 0.62]	0.15	Small effect size
Organization vs. Llama-3	0.6	< 0.05	[0.57, 0.63]	1.22	Large effect size
Word Choice vs. Rater 1	0.67	< 0.05	[0.65, 0.7]	-0.64	Large effect size
Word Choice vs. Rater 2	0.67	< 0.05	[0.64, 0.69]	-0.66	Large effect size
Word Choice vs. GPT-3.5T	0.33	< 0.05	[0.29, 0.37]	2.26	Large effect size
Word Choice vs. Llama-2	0.59	< 0.05	[0.56, 0.62]	0.07	Small effect size
Word Choice vs. Llama-3	0.63	< 0.05	[0.6, 0.65]	1.15	Large effect size
Sentence Fluency vs. Rater 1	0.64	< 0.05	[0.61, 0.67]	-0.55	Large effect size
Sentence Fluency vs. Rater 2	0.62	< 0.05	[0.59, 0.65]	-0.56	Large effect size
Sentence Fluency vs. GPT-3.5T	0.36	< 0.05	[0.31, 0.4]	2.36	Large effect size
Sentence Fluency vs. Llama-2	0.6	< 0.05	[0.57, 0.63]	0.18	Small effect size
Sentence Fluency vs. Llama-3	0.64	< 0.05	[0.61, 0.67]	1.23	Large effect size
Conventions vs. Rater 1	0.63	< 0.05	[0.61, 0.66]	-0.58	Large effect size
Conventions vs. Rater 2	0.62	< 0.05	[0.59, 0.65]	-0.6	Large effect size
Conventions vs. GPT-3.5T	0.35	< 0.05	[0.31, 0.39]	2.36	Large effect size
Conventions vs. Llama-2	0.59	< 0.05	[0.56, 0.62]	0.15	Small effect size
Conventions vs. Llama-3	0.63	< 0.05	[0.6, 0.66]	1.22	Large effect size

Table C.4: P value, Confidence Interval and Effect sizes of Table 4.6 and 5.6 Task 1

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Essay Length (sentences)	0.63	< 0.05	[0.6, 0.66]	-2.97	Large effect size
Rater 1 vs. Essay Length (tokens)	0.74	< 0.05	[0.72, 0.76]	-4.21	Large effect size
Rater 1 vs. FANBOYS (total)	0.5	< 0.05	[0.47, 0.54]	-2.52	Large effect size
Rater 1 vs. FANBOYS (unique)	0.35	< 0.05	[0.31, 0.39]	0.28	Medium effect size
Rater 1 vs. Transition Phrases (total)	0.34	< 0.05	[0.3, 0.38]	-1.27	Large effect size
Rater 1 vs. Transition Phrases (unique)	0.35	< 0.05	[0.31, 0.39]	-0.11	Small effect size
Rater 2 vs. Essay Length (sentences)	0.65	< 0.05	[0.62, 0.68]	-2.97	Large effect size
Rater 2 vs. Essay Length (tokens)	0.75	< 0.05	[0.73, 0.77]	-4.21	Large effect size
Rater 2 vs. FANBOYS (total)	0.49	< 0.05	[0.46, 0.53]	-2.52	Large effect size
Rater 2 vs. FANBOYS (unique)	0.35	< 0.05	[0.3, 0.39]	0.29	Medium effect size
Rater 2 vs. Transition Phrases (total)	0.37	< 0.05	[0.32, 0.4]	-1.26	Large effect size
Rater 2 vs. Transition Phrases (unique)	0.39	< 0.05	[0.35, 0.43]	-0.1	Small effect size
GPT-3.5T vs. Essay Length (sentences)	0.16	< 0.05	[0.12, 0.21]	-3.35	Large effect size
GPT-3.5T vs. Essay Length (tokens)	0.2	< 0.05	[0.15, 0.24]	-4.24	Large effect size
GPT-3.5T vs. FANBOYS (total)	0.12	< 0.05	[0.08, 0.17]	-2.95	Large effect size
GPT-3.5T vs. FANBOYS (unique)	0.09	< 0.05	[0.04, 0.14]	-2.64	Large effect size
GPT-3.5T vs. Transition Phrases (total)	0.05	< 0.05	[0.0, 0.1]	-2.04	Large effect size
GPT-3.5T vs. Transition Phrases (unique)	0.08	< 0.05	[0.03, 0.12]	-1.83	Large effect size
Llama-2 vs. Essay Length (sentences)	0.59	< 0.05	[0.56, 0.62]	-3.08	Large effect size
Llama-2 vs. Essay Length (tokens)	0.66	< 0.05	[0.63, 0.68]	-4.22	Large effect size
Llama-2 vs. FANBOYS (total)	0.45	< 0.05	[0.41, 0.48]	-2.64	Large effect size
Llama-2 vs. FANBOYS (unique)	0.35	< 0.05	[0.31, 0.39]	-0.47	Medium effect size
Llama-2 vs. Transition Phrases (total)	0.32	< 0.05	[0.28, 0.37]	-1.48	Large effect size
Llama-2 vs. Transition Phrases (unique)	0.36	< 0.05	[0.32, 0.4]	-0.57	Large effect size
Llama-3 vs. Essay Length (sentences)	0.44	< 0.05	[0.41, 0.48]	-3.23	Large effect size
Llama-3 vs. Essay Length (tokens)	0.53	< 0.05	[0.49, 0.56]	-4.23	Large effect size
Llama-3 vs. FANBOYS (total)	0.31	< 0.05	[0.27, 0.35]	-2.81	Large effect size
Llama-3 vs. FANBOYS (unique)	0.22	< 0.05	[0.18, 0.26]	-1.47	Large effect size
Llama-3 vs. Transition Phrases (total)	0.2	< 0.05	[0.16, 0.25]	-1.79	Large effect size
Llama-3 vs. Transition Phrases (unique)	0.25	< 0.05	[0.2, 0.29]	-1.25	Large effect size

Table C.5: P value, Confidence Interval and Effect sizes of Table 4.6 and 5.6 Task 7 (*no means $p \geq 0.05$)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. essay sent	0.61	< 0.05	[0.57, 0.64]	-0.88	Large effect size
Rater 1 vs. essay tok	0.62	< 0.05	[0.59, 0.65]	-2.59	Large effect size
Rater 1 vs. fanboys t	0.39	< 0.05	[0.35, 0.43]	-0.31	Medium effect size
Rater 1 vs. fanboys u	0.31	< 0.05	[0.27, 0.36]	2.62	Large effect size
Rater 1 vs. transition t	0.36	< 0.05	[0.32, 0.4]	0.9	Large effect size
Rater 1 vs. transition u	0.43	< 0.05	[0.38, 0.47]	2.15	Large effect size
Rater 2 vs. essay sent	0.61	< 0.05	[0.58, 0.64]	-0.87	Large effect size
Rater 2 vs. essay tok	0.63	< 0.05	[0.6, 0.66]	-2.59	Large effect size
Rater 2 vs. fanboys t	0.4	< 0.05	[0.36, 0.44]	-0.31	Medium effect size
Rater 2 vs. fanboys u	0.32	< 0.05	[0.27, 0.36]	2.55	Large effect size
Rater 2 vs. transition t	0.37	< 0.05	[0.33, 0.42]	0.89	Large effect size
Rater 2 vs. transition u	0.45	< 0.05	[0.41, 0.49]	2.1	Large effect size
GPT-3.5T vs. essay sent	-0.02	no	[-0.07, 0.03]	-1.9	Large effect size
GPT-3.5T vs. essay tok	0	no	[-0.05, 0.05]	-2.68	Large effect size
GPT-3.5T vs. fanboys t	-0.01	no	[-0.06, 0.04]	-1.67	Large effect size
GPT-3.5T vs. fanboys u	0.08	< 0.05	[0.03, 0.12]	-0.66	Large effect size
GPT-3.5T vs. transition t	0.03	no	[-0.02, 0.08]	-1.04	Large effect size
GPT-3.5T vs. transition u	0.09	< 0.05	[0.04, 0.13]	-0.64	Large effect size
Llama-2 vs. essay sent	0.42	< 0.05	[0.37, 0.46]	-1.75	Large effect size
Llama-2 vs. essay tok	0.45	< 0.05	[0.41, 0.49]	-2.67	Large effect size
Llama-2 vs. fanboys t	0.27	< 0.05	[0.23, 0.32]	-1.46	Large effect size
Llama-2 vs. fanboys u	0.21	< 0.05	[0.16, 0.26]	-0.2	Medium effect size
Llama-2 vs. transition t	0.24	< 0.05	[0.19, 0.28]	-0.76	Large effect size
Llama-2 vs. transition u	0.26	< 0.05	[0.22, 0.31]	-0.27	Medium effect size
Llama-3 vs. essay sent	0.37	< 0.05	[0.33, 0.41]	-1.19	Large effect size
Llama-3 vs. essay tok	0.44	< 0.05	[0.4, 0.48]	-2.62	Large effect size
Llama-3 vs. fanboys t	0.3	< 0.05	[0.25, 0.34]	-0.73	Large effect size
Llama-3 vs. fanboys u	0.28	< 0.05	[0.23, 0.32]	1.42	Large effect size
Llama-3 vs. transition t	0.27	< 0.05	[0.22, 0.31]	0.27	Medium effect size
Llama-3 vs. transition u	0.34	< 0.05	[0.29, 0.38]	1.16	Large effect size

Table C.6: P value, Confidence Interval and Effect sizes of Table 4.7

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Content vs. Essay Length (sentences)	0.58	< 0.05	[0.55, 0.61]	-3.03	Large
Content vs. Essay Length (tokens)	0.68	< 0.05	[0.65, 0.7]	-4.22	Large
Content vs. FANBOYS (total)	0.43	< 0.05	[0.4, 0.47]	-2.59	Large
Content vs. FANBOYS (unique)	0.28	< 0.05	[0.24, 0.33]	-0.16	Small
Content vs. Transition Phrases (total)	0.3	< 0.05	[0.26, 0.34]	-1.39	Large
Content vs. Transition Phrases (unique)	0.32	< 0.05	[0.28, 0.36]	-0.38	Medium
Organization vs. Essay Length (sentences)	0.54	< 0.05	[0.51, 0.57]	-3.05	Large
Organization vs. Essay Length (tokens)	0.64	< 0.05	[0.61, 0.67]	-4.22	Large
Organization vs. FANBOYS (total)	0.4	< 0.05	[0.36, 0.44]	-2.61	Large
Organization vs. FANBOYS (unique)	0.26	< 0.05	[0.21, 0.3]	-0.28	Medium
Organization vs. Transition Phrases (total)	0.28	< 0.05	[0.24, 0.32]	-1.42	Large
Organization vs. Transition Phrases (unique)	0.3	< 0.05	[0.26, 0.34]	-0.46	Medium
Word Choice vs. Essay Length (sentences)	0.56	< 0.05	[0.53, 0.59]	-3.06	Large
Word Choice vs. Essay Length (tokens)	0.66	< 0.05	[0.64, 0.69]	-4.22	Large
Word Choice vs. FANBOYS (total)	0.44	< 0.05	[0.4, 0.48]	-2.62	Large
Word Choice vs. FANBOYS (unique)	0.28	< 0.05	[0.24, 0.32]	-0.34	Medium
Word Choice vs. Transition Phrases (total)	0.28	< 0.05	[0.23, 0.32]	-1.44	Large
Word Choice vs. Transition Phrases (unique)	0.3	< 0.05	[0.26, 0.34]	-0.5	Medium
Sentence Fluency vs. Essay Length (sentences)	0.54	< 0.05	[0.51, 0.57]	-3.04	Large
Sentence Fluency vs. Essay Length (tokens)	0.63	< 0.05	[0.6, 0.66]	-4.22	Large
Sentence Fluency vs. FANBOYS (total)	0.42	< 0.05	[0.38, 0.46]	-2.6	Large
Sentence Fluency vs. FANBOYS (unique)	0.28	< 0.05	[0.24, 0.32]	-0.25	Medium
Sentence Fluency vs. Transition Phrases (total)	0.26	< 0.05	[0.21, 0.3]	-1.42	Large
Sentence Fluency vs. Transition Phrases (unique)	0.28	< 0.05	[0.24, 0.32]	-0.44	Medium
Conventions vs. Essay Length (sentences)	0.53	< 0.05	[0.5, 0.57]	-3.05	Large
Conventions vs. Essay Length (tokens)	0.63	< 0.05	[0.6, 0.65]	-4.22	Large
Conventions vs. FANBOYS (total)	0.42	< 0.05	[0.38, 0.46]	-2.61	Large
Conventions vs. FANBOYS (unique)	0.28	< 0.05	[0.24, 0.32]	-0.28	Medium
Conventions vs. Transition Phrases (total)	0.27	< 0.05	[0.23, 0.31]	-1.42	Large
Conventions vs. Transition Phrases (unique)	0.29	< 0.05	[0.24, 0.33]	-0.46	Medium

Table C.7: P value, Confidence Interval and Effect sizes of Table 4.8 Task 1 (*no means $p \geq 0.05$)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. FleschGL	0.02	no	[-0.02, 0.07]	-1.49	Large effect size
Rater 1 vs. FleschRE	-0.17	< 0.05	[-0.21, -0.12]	-10.91	Large effect size
Rater 1 vs. SmogInd	0.16	< 0.05	[0.12, 0.21]	-5.38	Large effect size
Rater 1 vs. ColeLia	0.3	< 0.05	[0.26, 0.34]	-3.21	Large effect size
Rater 1 vs. GunningFog	-0.19	< 0.05	[-0.24, -0.15]	-1.66	Large effect size
Rater 1 vs. AutoRea	0.01	no	[-0.03, 0.06]	-1.7	Large effect size
Rater 1 vs. LinseaW	-0.1	< 0.05	[-0.14, -0.05]	-1.97	Large effect size
Rater 1 vs. DaleChall	0.16	< 0.05	[0.12, 0.21]	-3.75	Large effect size
Rater 2 vs. FleschGL	0.01	no	[-0.04, 0.05]	-1.49	Large effect size
Rater 2 vs. FleschRE	-0.15	< 0.05	[-0.2, -0.11]	-10.92	Large effect size
Rater 2 vs. SmogInd	0.14	< 0.05	[0.09, 0.18]	-5.42	Large effect size
Rater 2 vs. ColeLia	0.28	< 0.05	[0.24, 0.33]	-3.23	Large effect size
Rater 2 vs. GunningFog	-0.21	< 0.05	[-0.25, -0.17]	-1.66	Large effect size
Rater 2 vs. AutoRea	-0.01	no	[-0.06, 0.04]	-1.7	Large effect size
Rater 2 vs. LinseaW	-0.12	< 0.05	[-0.16, -0.07]	-1.97	Large effect size
Rater 2 vs. DaleChall	0.16	< 0.05	[0.11, 0.2]	-3.81	Large effect size
GPT-3.5T vs. FleschGL	0.1	< 0.05	[0.05, 0.14]	-3.01	Large effect size
GPT-3.5T vs. FleschRE	-0.23	< 0.05	[-0.28, -0.19]	-11.3	Large effect size
GPT-3.5T vs. SmogInd	0.18	< 0.05	[0.13, 0.22]	-8.2	Large effect size
GPT-3.5T vs. ColeLia	0.17	< 0.05	[0.12, 0.21]	-5.56	Large effect size
GPT-3.5T vs. GunningFog	-0.09	< 0.05	[-0.14, -0.05]	-3.2	Large effect size
GPT-3.5T vs. AutoRea	0	no	[-0.04, 0.05]	-2.95	Large effect size
GPT-3.5T vs. LinseaW	-0.08	< 0.05	[-0.12, -0.03]	-2.94	Large effect size
GPT-3.5T vs. DaleChall	-0.01	no	[-0.06, 0.04]	-8.07	Large effect size
Llama-2 vs. FleschGL	-0.01	no	[-0.05, 0.04]	-1.92	Large effect size
Llama-2 vs. FleschRE	-0.19	< 0.05	[-0.23, -0.14]	-11.03	Large effect size
Llama-2 vs. SmogInd	0.13	< 0.05	[0.09, 0.18]	-6.32	Large effect size
Llama-2 vs. ColeLia	0.24	< 0.05	[0.19, 0.28]	-3.93	Large effect size
Llama-2 vs. GunningFog	-0.25	< 0.05	[-0.29, -0.2]	-2.1	Large effect size
Llama-2 vs. AutoRea	-0.08	< 0.05	[-0.12, -0.03]	-2.05	Large effect size
Llama-2 vs. LinseaW	-0.19	< 0.05	[-0.23, -0.14]	-2.25	Large effect size
Llama-2 vs. DaleChall	0.05	< 0.05	[0.01, 0.1]	-5.14	Large effect size

Table C.8: P value, Confidence Interval and Effect sizes of Table 4.8 Task 7 (*no means $p \geq 0.05$)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. FleschGL	-0.19	< 0.05	[-0.24, -0.14]	1.62	Large effect size
Rater 1 vs. FleschRE	0.13	< 0.05	[0.08, 0.18]	-11.02	Large effect size
Rater 1 vs. SmogInd	0.16	< 0.05	[0.11, 0.21]	0.92	Large effect size
Rater 1 vs. ColeLia	0.11	< 0.05	[0.06, 0.16]	1.85	Large effect size
Rater 1 vs. GunningFog	-0.3	< 0.05	[-0.34, -0.25]	0.57	Large effect size
Rater 1 vs. AutoRea	-0.14	< 0.05	[-0.18, -0.09]	1.39	Large effect size
Rater 1 vs. LinseaW	-0.21	< 0.05	[-0.26, -0.16]	0.17	Small effect size
Rater 1 vs. DaleChall	-0.06	< 0.05	[-0.11, -0.01]	0.8	Large effect size
Rater 2 vs. FleschGL	-0.19	< 0.05	[-0.23, -0.14]	1.6	Large effect size
Rater 2 vs. FleschRE	0.13	< 0.05	[0.08, 0.18]	-10.99	Large effect size
Rater 2 vs. SmogInd	0.16	< 0.05	[0.11, 0.2]	0.9	Large effect size
Rater 2 vs. ColeLia	0.1	< 0.05	[0.05, 0.15]	1.81	Large effect size
Rater 2 vs. GunningFog	-0.31	< 0.05	[-0.35, -0.26]	0.57	Large effect size
Rater 2 vs. AutoRea	-0.14	< 0.05	[-0.19, -0.09]	1.38	Large effect size
Rater 2 vs. LinseaW	-0.21	< 0.05	[-0.26, -0.16]	0.18	Small effect size
Rater 2 vs. DaleChall	-0.09	< 0.05	[-0.14, -0.04]	0.78	Large effect size
GPT-3.5T vs. FleschGL	0.09	< 0.05	[0.04, 0.14]	-0.6	Large effect size
GPT-3.5T vs. FleschRE	-0.2	< 0.05	[-0.25, -0.15]	-11.89	Large effect size
GPT-3.5T vs. SmogInd	0.18	< 0.05	[0.13, 0.23]	-2.38	Large effect size
GPT-3.5T vs. ColeLia	0.11	< 0.05	[0.06, 0.16]	-1.13	Large effect size
GPT-3.5T vs. GunningFog	-0.04	no	[-0.09, 0.01]	-1.81	Large effect size
GPT-3.5T vs. AutoRea	0.01	no	[-0.04, 0.06]	-0.54	Large effect size
GPT-3.5T vs. LinseaW	-0.04	no	[-0.09, 0.01]	-1.63	Large effect size
GPT-3.5T vs. DaleChall	-0.21	< 0.05	[-0.26, -0.17]	-3.13	Large effect size
Llama-2 vs. FleschGL	-0.02	no	[-0.07, 0.03]	-0.32	Medium effect size
Llama-2 vs. FleschRE	-0.09	< 0.05	[-0.14, -0.04]	-11.63	Large effect size
Llama-2 vs. SmogInd	0.21	< 0.05	[0.16, 0.26]	-1.58	Large effect size
Llama-2 vs. ColeLia	0.18	< 0.05	[0.13, 0.22]	-0.63	Large effect size
Llama-2 vs. GunningFog	-0.15	< 0.05	[-0.2, -0.1]	-1.36	Large effect size
Llama-2 vs. AutoRea	-0.03	no	[-0.08, 0.02]	-0.31	Medium effect size
Llama-2 vs. LinseaW	-0.12	< 0.05	[-0.17, -0.07]	-1.34	Large effect size
Llama-2 vs. DaleChall	-0.05	< 0.05	[-0.1, -0.0]	-1.95	Large effect size

Table C.9: P value, Confidence Interval and Effect sizes of Table 5.7 Task 1 for Llama-3 (*no means $p \geq 0.05$)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Llama-3 vs. FleschGL	0.15	< 0.05	[0.1, 0.2]	-2.45	Large effect size
Llama-3 vs. FleschRE	-0.39	< 0.05	[-0.43, -0.35]	-11.15	Large effect size
Llama-3 vs. SmogInd	0.29	< 0.05	[0.25, 0.34]	-6.54	Large effect size
Llama-3 vs. ColeLia	0.37	< 0.05	[0.33, 0.41]	-4.41	Large effect size
Llama-3 vs. GunningFog	-0.17	< 0.05	[-0.21, -0.12]	-2.62	Large effect size
Llama-3 vs. AutoRea	0.04	no	[-0.01, 0.08]	-2.5	Large effect size
Llama-3 vs. LinseaW	-0.11	< 0.05	[-0.16, -0.07]	-2.61	Large effect size
Llama-3 vs. DaleChall	0.1	< 0.05	[0.06, 0.15]	-5.31	Large effect size

Table C.10: P value, Confidence Interval and Effect sizes of Table 5.7 Task 7 for Llama-3 (*no means $p \geq 0.05$)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Llama-3 vs. FleschGL	0.05	no	[-0.0, 0.1]	0.86	Large effect size
Llama-3 vs. FleschRE	-0.16	< 0.05	[-0.21, -0.11]	-11.19	Large effect size
Llama-3 vs. SmogInd	0.26	< 0.05	[0.21, 0.3]	-0.04	Small effect size
Llama-3 vs. ColeLia	0.19	< 0.05	[0.14, 0.23]	0.85	Large effect size
Llama-3 vs. GunningFog	-0.12	< 0.05	[-0.17, -0.08]	-0.15	Small effect size
Llama-3 vs. AutoRea	0	no	[-0.04, 0.05]	0.73	Large effect size
Llama-3 vs. LinseaW	-0.08	< 0.05	[-0.13, -0.03]	-0.38	Medium effect size
Llama-3 vs. DaleChall	-0.19	< 0.05	[-0.24, -0.14]	-0.25	Medium effect size

Table C.11: P value, Confidence Interval and Effect sizes of Table 4.9 and 5.8 Task 1 (*no means $p \geq 0.05$)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Aspell misspellings	0.14	< 0.05	[0.09, 0.18]	-0.94	Large effect size
Rater 1 vs. LanguageTool grammar	0.09	< 0.05	[0.05, 0.14]	0.65	Large effect size
Rater 1 vs. LanguageTool spelling	0.11	< 0.05	[0.06, 0.16]	-1.1	Large effect size
Rater 1 vs. LanguageTool style	0.09	< 0.05	[0.04, 0.13]	3.46	Large effect size
Rater 1 vs. LanguageTool punctuation	0.02	no	[-0.03, 0.06]	1.65	Large effect size
Rater 1 vs. LanguageTool capitalization	0.11	< 0.05	[0.06, 0.15]	3.96	Large effect size
Rater 2 vs. Aspell misspellings	0.16	< 0.05	[0.12, 0.21]	-0.94	Large effect size
Rater 2 vs. LanguageTool grammar	0.14	< 0.05	[0.09, 0.18]	0.65	Large effect size
Rater 2 vs. LanguageTool spelling	0.13	< 0.05	[0.09, 0.18]	-1.1	Large effect size
Rater 2 vs. LanguageTool style	0.1	< 0.05	[0.06, 0.15]	3.5	Large effect size
Rater 2 vs. LanguageTool punctuation	0.03	no	[-0.02, 0.08]	1.66	Large effect size
Rater 2 vs. LanguageTool capitalization	0.12	< 0.05	[0.07, 0.16]	4.01	Large effect size
GPT-3.5T vs. Aspell misspellings	-0.12	< 0.05	[-0.17, -0.08]	-1.43	Large effect size
GPT-3.5T vs. LanguageTool grammar	-0.11	< 0.05	[-0.15, -0.06]	-0.66	Large effect size
GPT-3.5T vs. LanguageTool spelling	-0.15	< 0.05	[-0.19, -0.1]	-1.56	Large effect size
GPT-3.5T vs. LanguageTool style	-0.02	no	[-0.06, 0.03]	1.26	Large effect size
GPT-3.5T vs. LanguageTool punctuation	-0.01	no	[-0.06, 0.03]	-0.02	Small effect size
GPT-3.5T vs. LanguageTool capitalization	-0.08	< 0.05	[-0.12, -0.03]	1.61	Large effect size
Llama-2 vs. Aspell misspellings	0.04	no	[-0.01, 0.09]	-1.07	Large effect size
Llama-2 vs. LanguageTool grammar	0.05	< 0.05	[0.0, 0.1]	0.3	Medium effect size
Llama-2 vs. LanguageTool spelling	0.02	no	[-0.03, 0.07]	-1.22	Large effect size
Llama-2 vs. LanguageTool style	0.05	< 0.05	[0.01, 0.1]	3.04	Large effect size
Llama-2 vs. LanguageTool punctuation	0.05	< 0.05	[0.01, 0.1]	1.24	Large effect size
Llama-2 vs. LanguageTool capitalization	0.11	< 0.05	[0.07, 0.16]	3.56	Large effect size
Llama-3 vs. Aspell misspellings	-0.14	< 0.05	[-0.18, -0.09]	-1.28	Large effect size
Llama-3 vs. LanguageTool grammar	-0.15	< 0.05	[-0.2, -0.1]	-0.28	Medium effect size
Llama-3 vs. LanguageTool spelling	-0.18	< 0.05	[-0.22, -0.13]	-1.42	Large effect size
Llama-3 vs. LanguageTool style	-0.01	no	[-0.06, 0.04]	1.66	Large effect size
Llama-3 vs. LanguageTool punctuation	-0.1	< 0.05	[-0.15, -0.06]	0.43	Medium effect size
Llama-3 vs. LanguageTool capitalization	-0.02	no	[-0.06, 0.03]	1.95	Large effect size

Table C.12: P value, Confidence Interval and Effect sizes of Table 4.9 and 5.8 Task 7 (*no means $p \geq 0.05$)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Aspell misspellings	0.11	< 0.05	[0.06, 0.15]	1.13	Large effect size
Rater 1 vs. LanguageTool grammar	0.05	< 0.05	[0.01, 0.1]	3.31	Large effect size
Rater 1 vs. LanguageTool spelling	0.2	< 0.05	[0.15, 0.25]	0.43	Medium effect size
Rater 1 vs. LanguageTool style	0.11	< 0.05	[0.06, 0.16]	1.39	Large effect size
Rater 1 vs. LanguageTool punctuation	0.11	< 0.05	[0.06, 0.16]	2.91	Large effect size
Rater 1 vs. LanguageTool capitalization	0.05	< 0.05	[0.0, 0.1]	3.84	Large effect size
Rater 2 vs. Aspell misspellings	0.1	< 0.05	[0.05, 0.15]	1.13	Large effect size
Rater 2 vs. LanguageTool grammar	0.06	< 0.05	[0.01, 0.11]	3.23	Large effect size
Rater 2 vs. LanguageTool spelling	0.21	< 0.05	[0.16, 0.26]	0.44	Medium effect size
Rater 2 vs. LanguageTool style	0.09	< 0.05	[0.04, 0.14]	1.38	Large effect size
Rater 2 vs. LanguageTool punctuation	0.1	< 0.05	[0.05, 0.15]	2.85	Large effect size
Rater 2 vs. LanguageTool capitalization	0.05	no	[-0.0, 0.1]	3.74	Large effect size
GPT-3.5T vs. Aspell misspellings	-0.24	< 0.05	[-0.28, -0.19]	-0.61	Large effect size
GPT-3.5T vs. LanguageTool grammar	-0.17	< 0.05	[-0.22, -0.12]	0.46	Medium effect size
GPT-3.5T vs. LanguageTool spelling	-0.21	< 0.05	[-0.26, -0.16]	-1.03	Large effect size
GPT-3.5T vs. LanguageTool style	-0.05	< 0.05	[-0.1, -0.0]	-0.02	Small effect size
GPT-3.5T vs. LanguageTool punctuation	-0.05	< 0.05	[-0.1, -0.0]	0.21	Medium effect size
GPT-3.5T vs. LanguageTool capitalization	-0.17	< 0.05	[-0.22, -0.12]	0.98	Large effect size
Llama-2 vs. Aspell misspellings	-0.1	< 0.05	[-0.15, -0.05]	-0.4	Medium effect size
Llama-2 vs. LanguageTool grammar	-0.03	no	[-0.08, 0.02]	0.62	Large effect size
Llama-2 vs. LanguageTool spelling	0.08	< 0.05	[0.03, 0.13]	-0.83	Large effect size
Llama-2 vs. LanguageTool style	0.06	< 0.05	[0.01, 0.1]	0.12	Small effect size
Llama-2 vs. LanguageTool punctuation	0.01	no	[-0.04, 0.06]	0.42	Medium effect size
Llama-2 vs. LanguageTool capitalization	-0.11	< 0.05	[-0.15, -0.06]	1.02	Large effect size
Llama-3 vs. LanguageTool grammar	-0.05	< 0.05	[-0.1, -0.0]	2.14	Large effect size
Llama-3 vs. LanguageTool spelling	0.01	no	[-0.04, 0.06]	-0.04	Small effect size
Llama-3 vs. LanguageTool style	0.01	no	[-0.04, 0.06]	0.9	Large effect size
Llama-3 vs. LanguageTool punctuation	0.06	< 0.05	[0.01, 0.11]	1.86	Large effect size
Llama-3 vs. LanguageTool capitalization	-0.16	< 0.05	[-0.21, -0.12]	2.58	Large effect size

Table C.13: P value, Confidence Interval and Effect sizes of Table 4.10 (*no means $p \geq 0.05$)

Comparison	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Content vs. Aspell misspellings	0.04	no	[-0.01, 0.08]	-1.02	Large
Content vs. LanguageTool grammar	-0.01	no	[-0.06, 0.03]	0.41	Medium
Content vs. LanguageTool spelling	-0.01	no	[-0.05, 0.04]	-1.17	Large
Content vs. LanguageTool style	0.06	< 0.05	[0.01, 0.1]	2.87	Large
Content vs. LanguageTool punctuation	-0.03	no	[-0.07, 0.02]	1.31	Large
Content vs. LanguageTool capitalization	0.05	< 0.05	[0.01, 0.1]	3.26	Large
Organization vs. Aspell misspellings	0.01	no	[-0.04, 0.06]	-1.04	Large
Organization vs. LanguageTool grammar	-0.02	no	[-0.07, 0.03]	0.36	Medium
Organization vs. LanguageTool spelling	-0.03	no	[-0.07, 0.02]	-1.19	Large
Organization vs. LanguageTool style	0.04	no	[-0.01, 0.09]	2.82	Large
Organization vs. LanguageTool punctuation	-0.02	no	[-0.07, 0.03]	1.25	Large
Organization vs. LanguageTool capitalization	0.04	no	[-0.01, 0.09]	3.22	Large
Word Choice vs. Aspell misspellings	0.03	no	[-0.02, 0.07]	-1.05	Large
Word Choice vs. LanguageTool grammar	-0.02	no	[-0.07, 0.02]	0.33	Medium
Word Choice vs. LanguageTool spelling	-0.02	no	[-0.06, 0.03]	-1.21	Large
Word Choice vs. LanguageTool style	0.04	no	[-0.01, 0.08]	2.74	Large
Word Choice vs. LanguageTool punctuation	-0.05	no	[-0.09, 0.0]	1.21	Large
Word Choice vs. LanguageTool capitalization	0.04	no	[-0.01, 0.08]	3.13	Large
Sentence Fluency vs. Aspell misspellings	-0.05	< 0.05	[-0.1, -0.01]	-1.04	Large
Sentence Fluency vs. LanguageTool grammar	-0.05	< 0.05	[-0.09, -0.0]	0.37	Medium
Sentence Fluency vs. LanguageTool spelling	-0.09	< 0.05	[-0.13, -0.04]	-1.19	Large
Sentence Fluency vs. LanguageTool style	0.02	no	[-0.02, 0.07]	2.82	Large
Sentence Fluency vs. LanguageTool punctuation	-0.02	no	[-0.07, 0.03]	1.26	Large
Sentence Fluency vs. LanguageTool capitalization	0	no	[-0.04, 0.05]	3.21	Large
Conventions vs. Aspell misspellings	-0.05	< 0.05	[-0.1, -0.0]	-1.04	Large
Conventions vs. LanguageTool grammar	-0.04	no	[-0.09, 0.01]	0.36	Medium
Conventions vs. LanguageTool spelling	-0.09	< 0.05	[-0.13, -0.04]	-1.19	Large
Conventions vs. LanguageTool style	0.02	no	[-0.03, 0.07]	2.82	Large
Conventions vs. LanguageTool punctuation	-0.01	no	[-0.06, 0.04]	1.25	Large
Conventions vs. LanguageTool capitalization	0	no	[-0.05, 0.04]	3.22	Large

Table C.14: P value, Confidence Interval and Effect sizes of Table 4.11 and 5.9 Task 1 (*no means $p \geq 0.05$)

Comparison with GPT-3.5T Explanation (Exp.)	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Exp. Length (sentences)	0.12	< 0.05	[0.07, 0.16]	-1.18	Large effect size
Rater 1 vs. Exp. Length (tokens)	0.17	< 0.05	[0.12, 0.21]	-5.01	Large effect size
Rater 1 vs. Exp. Sentiment (average)	0.06	< 0.05	[0.01, 0.11]	6.37	Large effect size
Rater 1 vs. Exp. Sentiment (max)	0.04	no	[-0.01, 0.08]	5.93	Large effect size
Rater 2 vs. Exp. Length (sentences)	0.11	< 0.05	[0.06, 0.16]	-1.18	Large effect size
Rater 2 vs. Exp. Length (tokens)	0.15	< 0.05	[0.11, 0.2]	-5.01	Large effect size
Rater 2 vs. Exp. Sentiment (average)	0.06	< 0.05	[0.01, 0.11]	6.52	Large effect size
Rater 2 vs. Exp. Sentiment (max)	0.04	no	[-0.0, 0.09]	6.06	Large effect size
GPT-3.5T vs. Exp. Length (sentences)	0.33	< 0.05	[0.29, 0.37]	-4.17	Large effect size
GPT-3.5T vs. Exp. Length (tokens)	0.45	< 0.05	[0.41, 0.48]	-5.15	Large effect size
GPT-3.5T vs. Exp. Sentiment (average)	0.33	< 0.05	[0.28, 0.37]	3.88	Large effect size
GPT-3.5T vs. Exp. Sentiment (max)	0.28	< 0.05	[0.24, 0.33]	3.37	Large effect size
Llama-2 vs. Exp. Length (sentences)	0.14	< 0.05	[0.09, 0.18]	-2.02	Large effect size
Llama-2 vs. Exp. Length (tokens)	0.19	< 0.05	[0.14, 0.23]	-5.05	Large effect size
Llama-2 vs. Exp. Sentiment (average)	0.14	< 0.05	[0.09, 0.18]	6.4	Large effect size
Llama-2 vs. Exp. Sentiment (max)	0.12	< 0.05	[0.08, 0.17]	5.86	Large effect size
Llama-3 vs. Exp. Length (sentences)	0.16	< 0.05	[0.12, 0.21]	-2.77	Large effect size
Llama-3 vs. Exp. Length (tokens)	0.25	< 0.05	[0.21, 0.29]	-5.11	Large effect size
Llama-3 vs. Exp. Sentiment (average)	0.18	< 0.05	[0.13, 0.22]	3.47	Large effect size
Llama-3 vs. Exp. Sentiment (max)	0.13	< 0.05	[0.08, 0.17]	3.15	Large effect size

Table C.15: P value, Confidence Interval and Effect sizes of Table 4.11 and 5.9 Task 7 (*no means $p \geq 0.05$)

Comparison with GPT-3.5T explanation	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Exp. Length (sentences)	0.05	< 0.05	[0.0, 0.1]	0.91	Large effect size
Rater 1 vs. Exp. Length (tokens)	0.04	no	[-0.01, 0.09]	-3.63	Large effect size
Rater 1 vs. Exp. Sentiment (average)	0.05	no	[-0.0, 0.1]	4.96	Large effect size
Rater 1 vs. Exp. Sentiment (max)	0.02	no	[-0.03, 0.07]	4.73	Large effect size
Rater 2 vs. Exp. Length (sentences)	0.06	< 0.05	[0.01, 0.11]	0.9	Large effect size
Rater 2 vs. Exp. Length (tokens)	0.04	no	[-0.01, 0.09]	-3.63	Large effect size
Rater 2 vs. Exp. Sentiment (average)	0.05	no	[-0.0, 0.09]	4.79	Large effect size
Rater 2 vs. Exp. Sentiment (max)	0.01	no	[-0.04, 0.06]	4.57	Large effect size
GPT-3.5T vs. Exp. Length (sentences)	0.28	< 0.05	[0.23, 0.32]	-2.1	Large effect size
GPT-3.5T vs. Exp. Length (tokens)	0.3	< 0.05	[0.25, 0.34]	-3.87	Large effect size
GPT-3.5T vs. Exp. Sentiment (average)	0.24	< 0.05	[0.2, 0.29]	2.02	Large effect size
GPT-3.5T vs. Exp. Sentiment (max)	0.13	< 0.05	[0.08, 0.18]	1.74	Large effect size
Llama-2 vs. Exp. Length (sentences)	0.13	< 0.05	[0.08, 0.18]	-1.44	Large effect size
Llama-2 vs. Exp. Length (tokens)	0.13	< 0.05	[0.08, 0.18]	-3.84	Large effect size
Llama-2 vs. Exp. Sentiment (average)	0.08	< 0.05	[0.03, 0.13]	1.7	Large effect size
Llama-2 vs. Exp. Sentiment (max)	0.02	no	[-0.03, 0.07]	1.52	Large effect size
Llama-3 vs. Exp. Length (sentences)	0.13	< 0.05	[0.08, 0.18]	0.01	Small effect size
Llama-3 vs. Exp. Length (tokens)	0.15	< 0.05	[0.1, 0.2]	-3.7	Large effect size
Llama-3 vs. Exp. Sentiment (average)	0.1	< 0.05	[0.05, 0.15]	3.41	Large effect size
Llama-3 vs. Exp. Sentiment (max)	0.02	no	[-0.02, 0.07]	3.21	Large effect size

Table C.16: P value, Confidence Interval and Effect sizes of Table 4.12 and 5.10 Task 1

Comparison with Llama-2 explanation	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Exp. Length (sentences)	0.26	< 0.05	[0.22, 0.31]	-3.24	Large effect size
Rater 1 vs. Exp. Length (tokens)	0.26	< 0.05	[0.22, 0.31]	-5.98	Large effect size
Rater 1 vs. Exp. Sentiment (average)	0.3	< 0.05	[0.25, 0.34]	5.13	Large effect size
Rater 1 vs. Exp. Sentiment (max)	0.24	< 0.05	[0.19, 0.28]	5.19	Large effect size
Rater 2 vs. Exp. Length (sentences)	0.25	< 0.05	[0.21, 0.29]	-3.24	Large effect size
Rater 2 vs. Exp. Length (tokens)	0.27	< 0.05	[0.23, 0.31]	-5.98	Large effect size
Rater 2 vs. Exp. Sentiment (average)	0.29	< 0.05	[0.24, 0.33]	5.26	Large effect size
Rater 2 vs. Exp. Sentiment (max)	0.22	< 0.05	[0.17, 0.26]	5.32	Large effect size
GPT-3.5T vs. Exp. Length (sentences)	0.1	< 0.05	[0.05, 0.14]	-4.46	Large effect size
GPT-3.5T vs. Exp. Length (tokens)	0.11	< 0.05	[0.07, 0.16]	-6.04	Large effect size
GPT-3.5T vs. Exp. Sentiment (average)	0.24	< 0.05	[0.2, 0.28]	2.17	Large effect size
GPT-3.5T vs. Exp. Sentiment (max)	0.2	< 0.05	[0.15, 0.24]	2.3	Large effect size
Llama-2 vs. Exp. Length (sentences)	0.28	< 0.05	[0.24, 0.32]	-3.6	Large effect size
Llama-2 vs. Exp. Length (tokens)	0.31	< 0.05	[0.26, 0.35]	-6	Large effect size
Llama-2 vs. Exp. Sentiment (average)	0.54	< 0.05	[0.5, 0.57]	4.98	Large effect size
Llama-2 vs. Exp. Sentiment (max)	0.42	< 0.05	[0.38, 0.46]	5.04	Large effect size
Llama-3 vs. Exp. Length (sentences)	0.22	< 0.05	[0.17, 0.26]	-3.99	Large effect size
Llama-3 vs. Exp. Length (tokens)	0.23	< 0.05	[0.18, 0.27]	-6.03	Large effect size
Llama-3 vs. Exp. Sentiment (average)	0.34	< 0.05	[0.3, 0.38]	2.26	Large effect size
Llama-3 vs. Exp. Sentiment (max)	0.26	< 0.05	[0.22, 0.31]	2.37	Large effect size

Table C.17: P value, Confidence Interval and Effect sizes of Table 4.12 and 5.10 Task 7

Comparison with Llama-2 explanation	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Exp. Length (sentences)	0.16	< 0.05	[0.12, 0.21]	0.63	Large effect size
Rater 1 vs. Exp. Length (tokens)	0.21	< 0.05	[0.16, 0.26]	-2.91	Large effect size
Rater 1 vs. Exp. Sentiment (average)	0.26	< 0.05	[0.21, 0.3]	4.75	Large effect size
Rater 1 vs. Exp. Sentiment (max)	0.22	< 0.05	[0.18, 0.27]	4.68	Large effect size
Rater 2 vs. Exp. Length (sentences)	0.13	< 0.05	[0.08, 0.17]	0.63	Large effect size
Rater 2 vs. Exp. Length (tokens)	0.17	< 0.05	[0.12, 0.22]	-2.91	Large effect size
Rater 2 vs. Exp. Sentiment (average)	0.25	< 0.05	[0.21, 0.3]	4.59	Large effect size
Rater 2 vs. Exp. Sentiment (max)	0.22	< 0.05	[0.17, 0.26]	4.52	Large effect size
GPT-3.5T vs. Exp. Length (sentences)	0.11	< 0.05	[0.07, 0.16]	-1.9	Large effect size
GPT-3.5T vs. Exp. Length (tokens)	0.13	< 0.05	[0.08, 0.18]	-3.06	Large effect size
GPT-3.5T vs. Exp. Sentiment (average)	0.13	< 0.05	[0.08, 0.18]	1.8	Large effect size
GPT-3.5T vs. Exp. Sentiment (max)	0.14	< 0.05	[0.09, 0.19]	1.68	Large effect size
Llama-2 vs. Exp. Length (sentences)	0.27	< 0.05	[0.22, 0.31]	-1.4	Large effect size
Llama-2 vs. Exp. Length (tokens)	0.3	< 0.05	[0.26, 0.35]	-3.04	Large effect size
Llama-2 vs. Exp. Sentiment (average)	0.5	< 0.05	[0.47, 0.54]	1.57	Large effect size
Llama-2 vs. Exp. Sentiment (max)	0.44	< 0.05	[0.4, 0.48]	1.48	Large effect size
Llama-3 vs. Exp. Length (sentences)	0.23	< 0.05	[0.19, 0.28]	-0.13	Small effect size
Llama-3 vs. Exp. Length (tokens)	0.27	< 0.05	[0.23, 0.32]	-2.96	Large effect size
Llama-3 vs. Exp. Sentiment (average)	0.31	< 0.05	[0.27, 0.36]	3.25	Large effect size
Llama-3 vs. Exp. Sentiment (max)	0.3	< 0.05	[0.25, 0.34]	3.17	Large effect size

Table C.18: P value, Confidence Interval and Effect sizes of Table 5.11 Task 1

Comparison with Llama-3 explanation	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Exp. Length (sentences)	0.35	< 0.05	[0.3, 0.39]	-3.78	Large effect size
Rater 1 vs. Exp. Length (tokens)	0.4	< 0.05	[0.36, 0.44]	-7.17	Large effect size
Rater 1 vs. Exp. Sentiment (average)	0.23	< 0.05	[0.19, 0.28]	4.91	Large effect size
Rater 1 vs. Exp. Sentiment (max)	0.21	< 0.05	[0.17, 0.26]	5.09	Large effect size
Rater 2 vs. Exp. Length (sentences)	0.37	< 0.05	[0.32, 0.41]	-3.8	Large effect size
Rater 2 vs. Exp. Length (tokens)	0.41	< 0.05	[0.37, 0.45]	-7.17	Large effect size
Rater 2 vs. Exp. Sentiment (average)	0.21	< 0.05	[0.16, 0.25]	5	Large effect size
Rater 2 vs. Exp. Sentiment (max)	0.21	< 0.05	[0.17, 0.25]	5.19	Large effect size
GPT-3.5T vs. Exp. Length (sentences)	0.11	< 0.05	[0.07, 0.16]	-5.98	Large effect size
GPT-3.5T vs. Exp. Length (tokens)	0.21	< 0.05	[0.17, 0.25]	-7.25	Large effect size
GPT-3.5T vs. Exp. Sentiment (average)	0.29	< 0.05	[0.25, 0.33]	2.38	Large effect size
GPT-3.5T vs. Exp. Sentiment (max)	0.23	< 0.05	[0.18, 0.27]	2.5	Large effect size
Llama-2 vs. Exp. Length (sentences)	0.32	< 0.05	[0.28, 0.36]	-4.47	Large effect size
Llama-2 vs. Exp. Length (tokens)	0.43	< 0.05	[0.39, 0.46]	-7.19	Large effect size
Llama-2 vs. Exp. Sentiment (average)	0.31	< 0.05	[0.26, 0.35]	4.63	Large effect size
Llama-2 vs. Exp. Sentiment (max)	0.25	< 0.05	[0.21, 0.3]	4.85	Large effect size
Llama-3 vs. Exp. Length (sentences)	0.34	< 0.05	[0.3, 0.38]	-4.9	Large effect size
Llama-3 vs. Exp. Length (tokens)	0.47	< 0.05	[0.43, 0.51]	-7.23	Large effect size
Llama-3 vs. Exp. Sentiment (average)	0.49	< 0.05	[0.45, 0.52]	2.52	Large effect size
Llama-3 vs. Exp. Sentiment (max)	0.4	< 0.05	[0.36, 0.44]	2.59	Large effect size

Table C.19: P value, Confidence Interval and Effect sizes of Table 5.11 Task 7

Comparison with Llama-3 explanation	**Pearson's r**	**p-value**	**95% CI**	**Effect Size (Cohen's d)**	**Cohen's d Interpretation**
Rater 1 vs. Exp. Length (sentences)	0.25	< 0.05	[0.21, 0.3]	1.33	Large effect size
Rater 1 vs. Exp. Length (tokens)	0.34	< 0.05	[0.3, 0.38]	-7.32	Large effect size
Rater 1 vs. Exp. Sentiment (average)	0.28	< 0.05	[0.23, 0.32]	4.76	Large effect size
Rater 1 vs. Exp. Sentiment (max)	0.2	< 0.05	[0.15, 0.25]	4.5	Large effect size
Rater 2 vs. Exp. Length (sentences)	0.29	< 0.05	[0.24, 0.33]	1.3	Large effect size
Rater 2 vs. Exp. Length (tokens)	0.36	< 0.05	[0.32, 0.4]	-7.32	Large effect size
Rater 2 vs. Exp. Sentiment (average)	0.3	< 0.05	[0.25, 0.34]	4.61	Large effect size
Rater 2 vs. Exp. Sentiment (max)	0.22	< 0.05	[0.17, 0.26]	4.35	Large effect size
GPT-3.5T vs. Exp. Length (sentences)	0.15	< 0.05	[0.1, 0.19]	-2.5	Large effect size
GPT-3.5T vs. Exp. Length (tokens)	0.24	< 0.05	[0.19, 0.29]	-7.74	Large effect size
GPT-3.5T vs. Exp. Sentiment (average)	0.24	< 0.05	[0.2, 0.29]	1.81	Large effect size
GPT-3.5T vs. Exp. Sentiment (max)	0.17	< 0.05	[0.12, 0.22]	1.49	Large effect size
Llama-2 vs. Exp. Length (sentences)	0.22	< 0.05	[0.18, 0.27]	-1.49	Large effect size
Llama-2 vs. Exp. Length (tokens)	0.28	< 0.05	[0.23, 0.33]	-7.67	Large effect size
Llama-2 vs. Exp. Sentiment (average)	0.38	< 0.05	[0.34, 0.42]	1.57	Large effect size
Llama-2 vs. Exp. Sentiment (max)	0.26	< 0.05	[0.21, 0.31]	1.36	Large effect size
Llama-3 vs. Exp. Length (sentences)	0.34	< 0.05	[0.3, 0.38]	0.22	Medium effect size
Llama-3 vs. Exp. Length (tokens)	0.46	< 0.05	[0.43, 0.5]	-7.44	Large effect size
Llama-3 vs. Exp. Sentiment (average)	0.48	< 0.05	[0.44, 0.52]	3.26	Large effect size
Llama-3 vs. Exp. Sentiment (max)	0.32	< 0.05	[0.28, 0.36]	3.03	Large effect size

To identify the specific group differences, we performed post-hoc analyses using the Tukey Honestly Significant Difference (HSD) test for the results detailed in Sections 4.4 and 5.4. Post-hoc analysis helps in identifying specific differences between pairs of groups after finding a significant overall difference. In our reporting, we included the following elements:

- Group1 and Group2 columns are the groups being compared
- Meandiff is the difference between the group means (mean of Group1 is subtracted from mean of Group2)
- P-adj is the corrected p-value which takes into account the multiple comparisons being conducted
- CI lower is the lower band of the confidence interval. In the current example the confidence interval at the 95% level since $\alpha = 0.05$.
- CI upper is the upper band of the confidence interval. In the current example the confidence interval at the 95% level since $\alpha = 0.05$.
- Reject is the decision rule based on the corrected p-value

Table C.20: Tukey HSD test on Table 4.13 (GPT-3.5T)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
GPT-3.5T Misspelling Category						
Aspell misspelling						
No mention	Unqualified mention	3.978	0.0	2.0122	5.9439	True
No mention	Qualified mention	5.3687	0.0	4.1207	6.6167	True
Unqualified mention	Qualified mention	1.3906	0.3191	-0.87	3.6512	False
LanguageTool spelling						
No mention	Unqualified mention	4.1541	0.0	2.0492	6.2589	True
No mention	Qualified mention	5.7212	0.0	4.385	7.0575	True
Unqualified mention	Qualified mention	1.5672	0.2823	-0.8532	3.9876	False
GPT-3.5T Score						
No mention	Unqualified mention	0.0002	1.0	-0.1667	0.1671	False
No mention	Qualified mention	-0.1226	0.0184	-0.2286	-0.0167	True
Unqualified mention	Qualified mention	-0.1228	0.2908	-0.3148	0.0691	False
GPT-3.5T Sentiment (average)						
No mention	Unqualified mention	-0.215	0.0043	-0.3738	-0.0563	True
No mention	Qualified mention	-0.2745	0.0	-0.3753	-0.1737	True
Unqualified mention	Qualified mention	-0.0594	0.7255	-0.242	0.1232	False
GPT-3.5T Sentiment (max)						
No mention	Unqualified mention	-0.0106	0.9889	-0.1845	0.1633	False
No mention	Qualified mention	-0.2326	0.0	-0.343	-0.1222	True
Unqualified mention	Qualified mention	-0.2221	0.0251	-0.422	-0.0221	True

Table C.21: Tukey HSD test on Table 4.13 (Llama-2)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
Llama-2 Misspelling Category						
Aspell Misspelling Category						
No mention	Unqualified mention	1.2196	0.0904	-0.1437	2.5829	False
No mention	Qualified mention	6.6317	0.0	3.9932	9.2703	True
Unqualified mention	Qualified mention	5.4122	0.0	2.4958	8.3285	True
LanguageTool Spelling						
No mention	Unqualified mention	1.2424	0.1129	-0.2159	2.7008	False
No mention	Qualified mention	7.2897	0.0	4.4673	10.1121	True
Unqualified mention	Qualified mention	6.0473	0.0	2.9277	9.1669	True
Llama-2 Score						
No mention	Unqualified mention	0.1439	0.0253	0.0142	0.2735	True
No mention	Qualified mention	-1.0048	0.0	-1.2558	-0.7538	True
Unqualified mention	Qualified mention	-1.1486	0.0	-1.426	-0.8713	True
Llama-2 Sentiment (average)						
No mention	Unqualified mention	0.0445	0.4861	-0.0466	0.1356	False
No mention	Qualified mention	-0.5642	0.0	-0.7405	-0.3879	True
Unqualified mention	Qualified mention	-0.6087	0.0	-0.8036	-0.4138	True
Llama-2 Sentiment (max)						
No mention	Unqualified mention	-0.0142	0.9393	-0.113	0.0846	False
No mention	Qualified mention	-0.5163	0.0	-0.7075	-0.3252	True
Unqualified mention	Qualified mention	-0.5022	0.0	-0.7134	-0.2909	True

Table C.22: Tukey HSD test on Table 5.12 with Llama-3

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
Llama-3 Misspelling Category						
Aspell misspelling						
No mention	Unqualified mention	5.4726	0.0	4.0327	6.9125	True
No mention	Qualified mention	4.9629	0.0	3.6628	6.263	True
Unqualified mention	Qualified mention	-0.5097	0.7959	-2.3665	1.3471	False
LanguageTool spelling						
No mention	Unqualified mention	5.4189	0.0	3.8708	6.967	True
No mention	Qualified mention	5.0866	0.0	3.6889	6.4844	True
Unqualified mention	Qualified mention	-0.3323	0.9194	-2.3286	1.6641	False
Llama-3 Score						
No mention	Unqualified mention	-0.0071	0.9968	-0.2255	0.2113	False
No mention	Qualified mention	-0.4458	0.0	-0.643	-0.2486	True
Unqualified mention	Qualified mention	-0.4387	0.0008	-0.7203	-0.1571	True
Llama-3 Sentiment (average)						
No mention	Unqualified mention	0.001	0.9999	-0.1686	0.1707	False
No mention	Qualified mention	-0.153	0.0504	-0.3062	0.0002	False
Unqualified mention	Qualified mention	-0.154	0.2246	-0.3728	0.0648	False
Llama-3 Sentiment (average)						
No mention	Unqualified mention	-0.0327	0.8741	-0.188	0.1226	False
No mention	Qualified mention	-0.1657	0.0155	-0.306	-0.0255	True
Unqualified mention	Qualified mention	-0.133	0.2644	-0.3333	0.0673	False

Table C.23: Tukey HSD test on Table 4.14 (GPT-3.5T)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
GPT-3.5T Misspelling Category						
Aspell Misspelling Category						
No mention	Unqualified mention	0.5445	0.2205	-0.2244	1.3133	False
No mention	Qualified mention	1.646	0.0	1.0751	2.217	True
Unqualified mention	Qualified mention	1.1016	0.0059	0.2633	1.9399	True
LanguageTool Spelling						
No mention	Unqualified mention	0.704	0.2062	-0.2686	1.6765	False
No mention	Qualified mention	1.8044	0.0	1.0822	2.5266	True
Unqualified mention	Qualified mention	1.1004	0.0398	0.04	2.1608	True
GPT-3.5T Score						
No mention	Unqualified mention	0.1211	0.655	-0.2029	0.445	False
No mention	Qualified mention	-0.9254	0.0	-1.1659	-0.6848	True
Unqualified mention	Qualified mention	-1.0464	0.0	-1.3997	-0.6932	True
GPT-3.5T Sentiment (average)						
No mention	Unqualified mention	0.0226	0.7618	-0.0527	0.0979	False
No mention	Qualified mention	-0.0513	0.0803	-0.1072	0.0047	False
Unqualified mention	Qualified mention	-0.0738	0.0883	-0.1559	0.0083	False
GPT-3.5T Sentiment (max)						
No mention	Unqualified mention	0.0292	0.7636	-0.0687	0.1271	False
No mention	Qualified mention	-0.0324	0.5472	-0.1051	0.0402	False
Unqualified mention	Qualified mention	-0.0616	0.365	-0.1683	0.0451	False

Table C.24: Tukey HSD test on Table 4.14 (Llama-2)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
Llama-2 Misspelling Category						
Aspell Misspelling Category						
No mention	Unqualified mention	0.5445	0.2205	-0.2244	1.3133	False
No mention	Qualified mention	1.646	0.0	1.0751	2.217	True
Unqualified mention	Qualified mention	1.1016	0.0059	0.2633	1.9399	True
LanguageTool Spelling						
No mention	Unqualified mention	0.704	0.2062	-0.2686	1.6765	False
No mention	Qualified mention	1.8044	0.0	1.0822	2.5266	True
Unqualified mention	Qualified mention	1.1004	0.0398	0.04	2.1608	True
Llama-2 Score						
No mention	Unqualified mention	0.1211	0.655	-0.2029	0.445	False
No mention	Qualified mention	-0.9254	0.0	-1.1659	-0.6848	True
Unqualified mention	Qualified mention	-1.0464	0.0	-1.3997	-0.6932	True
Llama-2 Sentiment (average)						
No mention	Unqualified mention	0.0226	0.7618	-0.0527	0.0979	False
No mention	Qualified mention	-0.0513	0.0803	-0.1072	0.0047	False
Unqualified mention	Qualified mention	-0.0738	0.0883	-0.1559	0.0083	False
Llama-2 Sentiment (max)						
No mention	Unqualified mention	0.0292	0.7636	-0.0687	0.1271	False
No mention	Qualified mention	-0.0324	0.5472	-0.1051	0.0402	False
Unqualified mention	Qualified mention	-0.0616	0.365	-0.1683	0.0451	False

Table C.25: Tukey HSD test on Table 5.13 (Llama-3)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
Llama-2 Misspelling Category						
No mention	Unqualified mention	1.6317	0.0	1.0424	2.221	True
No mention	Qualified mention	3.8705	0.0	3.3251	4.4159	True
Unqualified mention	Qualified mention	2.2388	0.0	1.619	2.8585	True
LanguageTool Spelling						
No mention	Unqualified mention	1.646	0.0	0.8711	2.4209	True
No mention	Qualified mention	3.4538	0.0	2.7366	4.171	True
Unqualified mention	Qualified mention	1.8079	0.0	0.993	2.6228	True
Llama-2 Score						
No mention	Unqualified mention	-0.3063	0.1483	-0.6908	0.0783	False
No mention	Qualified mention	-2.0522	0.0	-2.4081	-1.6963	True
Unqualified mention	Qualified mention	-1.7459	0.0	-2.1503	-1.3415	True
Llama-2 Sentiment (average)						
No mention	Unqualified mention	-0.0388	0.5642	-0.1279	0.0504	False
No mention	Qualified mention	-0.3181	0.0	-0.4006	-0.2356	True
Unqualified mention	Qualified mention	-0.2793	0.0	-0.373	-0.1856	True
Llama-2 Sentiment (max)						
No mention	Unqualified mention	0.0889	0.0976	-0.0121	0.1898	False
No mention	Qualified mention	-0.2181	0.0	-0.3115	-0.1246	True
Unqualified mention	Qualified mention	-0.3069	0.0	-0.4131	-0.2008	True

Table C.26: Tukey HSD test on Table 4.15 (GPT-3.5T)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
GPT-3.5T Grammar Category						
LanguageTool Grammar						
No mention	Unqualified mention	0.131	0.7383	-0.283	0.545	False
No mention	Qualified mention	0.578	0.0002	0.2404	0.9155	True
Unqualified mention	Qualified mention	0.4469	0.073	-0.0316	0.9255	False
GPT-3.5T Score						
No mention	Unqualified mention	0.1008	0.0346	0.0057	0.1959	True
No mention	Qualified mention	-0.0363	0.5144	-0.1138	0.0412	False
Unqualified mention	Qualified mention	-0.1371	0.0097	-0.247	-0.0273	True
GPT-3.5T Sentiment (average)						
No mention	Unqualified mention	-0.186	0.0	-0.2757	-0.0963	True
No mention	Qualified mention	-0.2579	0.0	-0.331	-0.1847	True
Unqualified mention	Qualified mention	-0.0719	0.235	-0.1756	0.0318	False
GPT-3.5T Sentiment (max)						
No mention	Unqualified mention	-0.1476	0.0013	-0.2458	-0.0494	True
No mention	Qualified mention	-0.2469	0.0	-0.3269	-0.1668	True
Unqualified mention	Qualified mention	-0.0992	0.1006	-0.2127	0.0143	False

Table C.27: Tukey HSD test on Table 4.15 (Llama-2)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
Llama-2 Grammar Category						
LanguageTool Grammar						
No mention	Unqualified mention	0.1308	0.6799	-0.2356	0.4972	False
No mention	Qualified mention	0.9196	0.0101	0.1792	1.66	True
Unqualified mention	Qualified mention	0.7889	0.0533	-0.0086	1.5863	False
Llama-2 Score						
No mention	Unqualified mention	0.085	0.0904	-0.01	0.1801	False
No mention	Qualified mention	-0.938	0.0	-1.1301	-0.7459	True
Unqualified mention	Qualified mention	-1.0231	0.0	-1.23	-0.8162	True
Llama-2 Sentiment (average)						
No mention	Unqualified mention	0.0199	0.7692	-0.0476	0.0873	False
No mention	Qualified mention	-0.4554	0.0	-0.5917	-0.319	True
Unqualified mention	Qualified mention	-0.4752	0.0	-0.6221	-0.3284	True
Llama-2 Sentiment (max)						
No mention	Unqualified mention	-0.0244	0.7152	-0.0976	0.0489	False
No mention	Qualified mention	-0.3992	0.0	-0.5472	-0.2512	True
Unqualified mention	Qualified mention	-0.3748	0.0	-0.5342	-0.2154	True

Table C.28: Tukey HSD test on Table 5.14 (Llama-3)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
Llama-3 Grammar Category						
LanguageTool Grammar						
No mention	Unqualified mention	0.6469	0.0	0.3358	0.9579	True
No mention	Qualified mention	0.8334	0.0	0.4871	1.1797	True
Unqualified mention	Qualified mention	0.1865	0.4457	-0.1742	0.5472	False
Llama-3 Score						
No mention	Unqualified mention	-0.5362	0.0	-0.6577	-0.4146	True
No mention	Qualified mention	-0.777	0.0	-0.9123	-0.6416	True
Unqualified mention	Qualified mention	-0.2408	0.0002	-0.3817	-0.0998	True
Llama-3 Sentiment (average)						
No mention	Unqualified mention	-0.2826	0.0	-0.3795	-0.1857	True
No mention	Qualified mention	-0.395	0.0	-0.5029	-0.2871	True
Unqualified mention	Qualified mention	-0.1124	0.0498	-0.2248	-0.0001	True
Llama-3 Sentiment (max)						
No mention	Unqualified mention	-0.3142	0.0	-0.4027	-0.2257	True
No mention	Qualified mention	-0.3385	0.0	-0.4371	-0.2399	True
Unqualified mention	Qualified mention	-0.0243	0.844	-0.127	0.0784	False

Table C.29: Tukey HSD test on Table 4.16 (GPT-3.5T)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
GPT-3.5T Grammar Category						
LanguageTool Grammar						
No mention	Unqualified mention	0.0223	0.9793	-0.2461	0.2907	False
No mention	Qualified mention	0.3864	0.0003	0.1543	0.6184	True
Unqualified mention	Qualified mention	0.3641	0.0003	0.144	0.5841	True
GPT-3.5T Score						
No mention	Unqualified mention	0.4071	0.006	0.0967	0.7175	True
No mention	Qualified mention	-0.7031	0.0	-0.9714	-0.4348	True
Unqualified mention	Qualified mention	-1.1102	0.0	-1.3646	-0.8557	True
GPT-3.5T Sentiment (average)						
No mention	Unqualified mention	0.016	0.8618	-0.0562	0.0882	False
No mention	Qualified mention	-0.0975	0.0007	-0.16	-0.0351	True
Unqualified mention	Qualified mention	-0.1135	0.0	-0.1727	-0.0543	True
GPT-3.5T Sentiment (max)						
No mention	Unqualified mention	0.0479	0.4577	-0.0463	0.1421	False
No mention	Qualified mention	-0.0559	0.2419	-0.1373	0.0256	False
Unqualified mention	Qualified mention	-0.1037	0.0047	-0.181	-0.0265	True

Table C.30: Tukey HSD test on Table 4.16 (Llama-2)

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
Llama-2 Grammar Category						
LanguageTool Grammar						
No mention	Unqualified mention	0.1588	0.2748	-0.0839	0.4014	False
No mention	Qualified mention	0.3082	0.0017	0.0983	0.5181	True
Unqualified mention	Qualified mention	0.1494	0.3848	-0.1164	0.4152	False
Llama-2 Score						
No mention	Unqualified mention	-0.2552	0.3014	-0.6599	0.1496	False
No mention	Qualified mention	-1.8381	0.0	-2.1882	-1.488	True
Unqualified mention	Qualified mention	-1.5829	0.0	-2.0263	-1.1395	True
Llama-2 Sentiment (average)						
No mention	Unqualified mention	-0.0639	0.29	-0.1637	0.0359	False
No mention	Qualified mention	-0.2761	0.0	-0.3624	-0.1898	True
Unqualified mention	Qualified mention	-0.2122	0.0	-0.3215	-0.1029	True
Llama-2 Sentiment (max)						
No mention	Unqualified mention	-0.0229	0.8234	-0.1134	0.0676	False
No mention	Qualified mention	-0.2137	0.0	-0.2919	-0.1354	True
Unqualified mention	Qualified mention	-0.1908	0.0	-0.2899	-0.0916	True

Table C.31: Tukey HSD test on Table 5.15 (Llama-3)

Group1	Group2	Meandiff	P-adj	CI lower	CI upper	Reject
Llama-3 Grammar Category						
LanguageTool Grammar						
No mention	Unqualified mention	0.0096	0.9955	-0.2389	0.2581	False
No mention	Qualified mention	0.7094	0.0	0.4622	0.9567	True
Unqualified mention	Qualified mention	0.6999	0.0	0.5026	0.8971	True
Llama-3 Score						
No mention	Unqualified mention	0.1781	0.5891	-0.2479	0.604	False
No mention	Qualified mention	-1.7737	0.0	-2.1975	-1.3498	True
Unqualified mention	Qualified mention	-1.9518	0.0	-2.2899	-1.6137	True
Llama-3 Sentiment (average)						
No mention	Unqualified mention	0.041	0.5905	-0.0573	0.1394	False
No mention	Qualified mention	-0.2928	0.0	-0.3906	-0.195	True
Unqualified mention	Qualified mention	-0.3338	0.0	-0.4119	-0.2558	True
Llama-3 Sentiment (max)						
No mention	Unqualified mention	0.0828	0.1865	-0.0281	0.1937	False
No mention	Qualified mention	-0.2578	0.0	-0.3682	-0.1474	True
Unqualified mention	Qualified mention	-0.3406	0.0	-0.4286	-0.2526	True

Table C.32: Tukey HSD test on Table 5.16 or 4.17

Group1	Group2	Meandiff	P-adj	CI lower	CI upper	Reject
GPT-3.5T Score						
Lower	Middle	0.6076	0.0	0.3607	0.8546	True
Lower	Upper	0.7919	0.0	0.5443	1.0395	True
Middle	Upper	0.1843	0.0	0.1226	0.246	True
Llama-2 Score						
Lower	Middle	1.7155	0.0	1.4686	1.9625	True
Lower	Upper	2.2693	0.0	2.0218	2.5169	True
Middle	Upper	0.5538	0.0	0.4921	0.6155	True
Llama-3 Score						
Lower	Middle	1.0583	0.0	0.66	1.4566	True
Lower	Upper	1.9425	0.0	1.5432	2.3418	True
Middle	Upper	0.8842	0.0	0.7847	0.9837	True

Table C.33: Tukey HSD test on Table 5.17 or 4.18

Group1	**Group2**	**Meandiff**	**P-adj**	**CI lower**	**CI upper**	**Reject**
GPT-3.5T Score						
Lower	Middle	1.0849	0.0	0.652	1.5177	True
Lower	Upper	1.5391	0.0	1.1073	1.971	True
Middle	Upper	0.4542	0.0	0.2348	0.6737	True
Llama-2 Score						
Lower	Middle	1.0414	0.0002	0.433	1.6497	True
Lower	Upper	2.7294	0.0	2.1225	3.3363	True
Middle	Upper	1.688	0.0	1.3796	1.9965	True
Llama-3 Score						
Lower	Middle	2.6748	0.0	2.0978	3.2518	True
Lower	Upper	4.6779	0.0	4.1023	5.2535	True
Middle	Upper	2.0031	0.0	1.7105	2.2956	True

Appendix D

Additional Information

First, we compared two language-checking tools, Aspell and LanguageTool, noting that Aspell strongly correlates with LanguageTool in terms of spelling accuracy. Next, we discuss our experiments with machine learning models using ChatGPT embeddings, where the Support Vector Regression model demonstrated superior performance. Lastly, we included the definition of Cohen's d as a measurement of effect size. Details can be found below.

D.1 Aspell strongly correlates with LanguageTool spelling

Table D.1: Aspell strongly correlates with LanguageTool spelling mistake count

LanguageTool Error Count / Aspell Misspelling	**Grammar**	**Spelling**	**Style**	**Punctuation**	**Capitalization**
Task 1	0.28	**0.95**	0.05	0.15	0.14
Task 7	0.32	**0.75**	0.29	0.17	0.16

Remarkably, the strong correlation coefficient of 0.95 between Aspell and LanguageTool, as depicted in Table D.1, underscores the robustness and consistency of errors identified by both tools. This high level of agreement between the two spell/grammar checkers reinforces our confidence in the accuracy of the mistakes detected, providing a solid foundation for our subsequent analyses.

D.2 Support Vector Regression model performs best

Table D.2: Machine Learning models with GPT-3.5T embedding in predicting human scores

Regression Models	**Train** r	**Test** r
Multiple Linear	0.95	0.73
Random Forest	0.96	0.66
Decision Tree	1.00	0.43
Support Vector	0.9	**0.82**
Xgboost	1.00	0.72

We conducted an investigation into the performance of various machine learning models, aiming to achieve the ability to approximate human scores. Our dataset comprises a total of 3352 sample essays from Task 1 and Task 7 combined. To standardize the human rater scores of different ranges for the two tasks, we normalized the average score given by two raters. By Utilizing OpenAI's embedding model *text-embedding-ada-002* with 1536 dimensions we generate essay embeddings to train the regression models. Our evaluation metric is the Pearson correlation coefficient denoted as r used to assess performance. Table D.2 presents the results of this analysis.

Utilizing GPT-3.5T we previously obtained a correlation of $r = 0.21 - 0.23$ with the overall scores given by human raters shown in 4.1.1 and $r = 0.33 - 0.36$ with the trait scores given by human raters shown in 4.1.2. All the regression models improved upon these results. Among the evaluated models, the Support Vector Machine (SVM) emerged as the

top performer, achieving the strongest correlation score of $r = 0.82$ on the test dataset. This indicates a strong correlation between the predicted and actual human scores when using SVM. Conversely, models such as decision tree regression displayed weaker correlations, showing minimal learning with $r = 0.43$ on the test dataset. While multiple linear and XGBoost regression models showed almost similar performance with $r = 0.73$ and $r = 0.72$ respectively, but they fell short of the predictive power demonstrated by SVM. These findings suggest that SVM with GPT-3.5T embedding exhibits promising potential for accurately predicting human grades on our dataset, outperforming other regression models considered in our analysis.

D.3 Effect Size: Cohen's d

Cohen's d typically ranges from negative to positive values. A negative Cohen's d indicates that the mean of the first group is smaller than the mean of the second group, while a positive Cohen's d indicates that the mean of the first group is larger than the mean of the second group. The absolute value of Cohen's d reflects the magnitude of the effect size, with larger absolute values indicating larger effect sizes. Generally, the following conventions are used to interpret the effect size:

- Small effect size: $d = 0.2$.
- Medium effect size: $d = 0.5$.
- Large effect size: $d >= 0.8$.