UNIVERSITY OF PADOVA

Department of Developmental Psychology and Socialisation

Master Degree in Psicologia dello Sviluppo e dell'Educazione Study track in Developmental and Educational Psychology

Final dissertation

Assessing the Suitability of AI-Generated Texts For Language Teaching Across Proficiency Levels Using the Functional Adequacy Scale

Supervisor:
Professor Judit Gervain

Candidate: Alp Akova
Student ID Number: 2042025

Academic Year: 2023/2024

Table of Contents

1 Introduction

The enthusiasm surrounding the possibilities of generative AI was evident during the latter periods of 2023 and 2024. The primary advantage of large language models (LLMs) is their capacity to generate text with ease and rationality. This capability has prompted a discussion over how they could be utilized in the field of education (Gan et al., 2023). A frequent question surrounding this debate is how to utilise these models in language training effectively. The current thesis contributes to this debate by testing whether LLMs can generate appropriate, natural-sounding texts for language teaching purposes. Recent studies have investigated the usage of large language models in educational settings. For instance, Tlili et al. (2023) conducted a qualitative instrumental case study to examine the concerns of using chatbots, specifically ChatGPT, in education among early adopters. The study highlights potential benefits of ChatGPT in educational environments but also underscores significant concerns that need addressing. These concerns span four main areas: response quality, usefulness, personality and emotions, and ethics.

1. **Response Quality**: ChatGPT can occasionally produce errors and outdated information, underscoring the need for regular updates to maintain its accuracy.
2. **Usefulness:** Although ChatGPT reduces educators' workloads by providing immediate responses, its effectiveness is limited by its inability to fully understand complex, contextual inquiries.
3. **Personality and Emotions:** ChatGPT's interaction lacks the ability to interpret emotional cues and engage in emotionally intelligent responses, which is crucial for sensitive educational settings.

4. **Ethics:** The use of ChatGPT raises ethical issues, including the risk of promoting cheating and diminishing students' critical thinking abilities. Data privacy and output bias are also significant concerns.

As both they and (Firat, 2023) mentioned ChatGPT has the potential to revolutionize self-directed learning by offering personalized support and real-time feedback, making learning resources more accessible, and allowing learners to study at their own pace. The AI can suggest customized reading materials and learning activities, facilitate self-assessment, and act as a mentor. These capabilities enhance the learner's autonomy and adaptability, crucial for self-directed learning. With these in mind the authors emphasized the need for guidelines and policies to facilitate the safe adoption of ChatGPT in schools and universities.

The technology industry is already starting to adopt these models for different uses. One such example is the partnership between Duolingo and OpenAI to enhance Duolingo's language teaching platform with large language model capabilities. This can be seen as a sign of the growing integration of AI in the language education industry. Duolingo is leveraging OpenAI's GPT-4 to introduce new features like "role play" and "explain my answer" to provide more personalized and interactive language learning experiences (*OpenAI Customer Story*, n.d.) Collaborations like this showcase how AI and language education, with language learning apps like Duolingo, embrace large language models' capabilities to enhance their offerings.

Building on this growing trend, the scope of this thesis is to test these models in an out-of-the-box manner by generating reading texts, evaluating them and comparing them to human-generated texts using the functional adequacy (FA) scale (Kuiken & Vedder, 2017). The FA scale offers a comprehensive, multidimensional framework for assessing certain features of a piece of written language, focusing on the functional aspects of text production.

1.1 Research Questions

By leveraging the user-friendly nature of the FA scale, this thesis aims to answer the following questions:

1) Do users rate LLM-generated texts as highly on the functional adequacy scale as similar human-written texts, and if not, what aspects of functional adequacy are most impacted?
2) To what extent can users accurately differentiate between AI-generated and human-written texts?
3) How consistently do LLM-generated texts align with their intended target proficiency levels (A1-A2, B1- B2, C1-C2) defined by the CEFR?
4) Do texts generated by different LLM models (open-source vs. closed-source, such as Mixtral 8x7b and GPT-3.5) show significant differences in user ratings across any aspects explored in questions 1-3?

These research questions are critical as they directly address the feasibility of integrating generative AI into educational settings. In particular, the results of the study will inform us about the following heatedly debated issues.

Focus on Functionality: By focusing on functional adequacy and evaluating four dimensions of this scale (content, task requirements, comprehensibility, and coherence and cohesion), we can evaluate how well a text achieves its purpose and suitability for the audience. This provides a novel approach for analyzing the outputs of Large Language Models.

Tailoring Content: The ability to generate personalized texts on demand at different proficiency levels would be highly beneficial for both individuals and professional educators. It can be hypothesized that learners achieve better outcomes when they are engaged with content that interests them. In traditional learning environments, textbooks and other teaching materials are typically uniform across all learners in a course or class. Personalizing these materials is both time-consuming and costly when done by humans. One significant advantage of AI is its ability to generate text rapidly and cost-effectively. This capability can be leveraged to create personalized learning materials and adaptive assessments that are more engaging and motivating for each learner, tailored to their personal interests and language proficiency levels. Understanding the reliability and limitations of large language models in achieving this is crucial for effectively using them to enhance personalized learning.

A new perspective to evaluate LLMs: Currently, state-of-the-art large language model evaluation tests are built to test and evaluate abilities such as problem-solving, generalization and transfer, and alignment with human values. A very well-known example of this kind of test is MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021). The MMLU benchmark is designed to assess the performance of language models across a wide range of tasks, including subject areas such as humanities, social sciences, STEM, and others. It consists of multiple-choice questions from various domains, allowing researchers to evaluate a model's ability to understand and apply knowledge across different disciplines. While these kinds of evaluations are great ways to evaluate the logical capabilities of large language models, their shortcomings are also quite evident. Alzahrani et al. (2024) pointed out that current state-of-the-art LLM evaluation tests, such as MMLU, primarily focus on assessing abilities like problem-solving, generalization, and alignment with human values. However, they may not capture the nuanced aspects of functional adequacy crucial for educational applications. Answers given to these questions will help us to understand what aspects of the large language models can be improved in terms of the scope of functional adequacy or whether current state-of-the-art large language models to be used in an out of the box manner to generate reliable well, tailored for both needs and the preferences of the user.

1.2 Why Functional Adequacy

In second/foreign language learning, the evaluation of linguistic performance resulting from language exercises is classically measured using the concepts of complexity, accuracy, and fluency (CAF). However, as Kuiken & Vedder (2017) highlights, assessing writing skills in a second language (L2) cannot be accomplished effectively without considering the functional aspect of L2 production. Moreover, Pallotti (2009) criticizes CAF's “complexity” aspect as a quite problematic construct because of its polysemantic nature. He states that complexity can be applied to several areas of language and communication, including lexical, grammatical, or interactional complexity. The accuracy aspect of the CAF was also criticized by Hasnain & Halder, (2024), who noted that ambiguity in defining appropriate accuracy norms is a key concern.

To solve these problems, recently a new metric was proposed (Kuiken and Vedder, 2017) relying on how functionally appropriate a text produced by an L2 learner is. The Functional Adequacy (FA) approach goes beyond just examining the linguistic features of a text and instead evaluates the extent to which the text accomplishes its communicative objective, irrespective of the writer's level of language proficiency. The FA scale is based on the general descriptors provided by the Common European Framework of Reference (CEFR) and focuses on four key dimensions: content, task requirements, comprehensibility, coherence and cohesion. This multidimensional framework allows for a comprehensive assessment of how well the text fulfils its intended purpose and meets the needs of the target audience rather than just its linguistic complexity or accuracy.

One important aspect that Kuiken & Vedder highlight, is that the FA is not only usable for assessing L2 writing but can also be used to evaluate texts produced by native speakers. This aspect of FA is important since the texts that large language models produce are more similar to those that native speakers produce. Considering that current state-of-the-art LLM evaluation tests primarily focus on skills like problem solving, generalization, and alignment with human values rather than the effectiveness of text communication, this study introduces a novel approach by utilizing the FA scale. Importantly, the FA scale is designed to be easy to use even for non-expert raters, making it well-suited for this study, which investigates how functionally adequate native Italian speakers without specialized expertise judge texts generated by LLMs to be, as compared to similar texts written by humans. This approach enhances the existing LLM evaluation techniques and can offer deeper insights into these models' practical, real-world effectiveness, particularly in educational settings. By assessing the functional adequacy of the generated texts, this study aims to provide valuable insights that go beyond just linguistic complexity or alignment with human values.

The accessibility and ease of use of the FA scale for non expert raters is particularly relevant when considering the broader implications of large language models in educational settings. One of the main benefits of using tools such as large language models is that users can rapidly and iteratively generate texts without significant effort or specialized knowledge. As with most commercial products, the majority of users using such powerful tools will likely be non-expert, everyday users. Since anyone can easily use a product like ChatGPT to produce reading texts on any given topic, a deeper look into the functional aspects of the produced text by non-expert users can provide valuable insights into the current discussions surrounding the educational use cases of these models.

With this in mind, utilizing a non-expert-friendly rating scale for the current study provides an efficient and relevant experimental setup. The AI-generated texts used in this research were produced using simple prompts, reflecting the typical usage patterns of everyday users of such language models. This approach allows for a comprehensive evaluation of the functional adequacy of the generated content from the perspective of the non-expert audience, which is highly relevant for assessing the practical, real-world educational potential of these AI systems.

1.2.1 Components of Functional Adequacy

The Functional Adequacy rating scale developed by Kuiken and Vedder (2017) consists of four key dimensions that comprehensively evaluate a text's communicative effectiveness. These four subscales are **Content** (relevance, clarity, appropriateness of information), **Task Requirements** (addressing task demands effectively), **Comprehensibility** (understandability of vocabulary and expression), and **Coherence and Cohesion** (logical flow and connectivity). Moreover, while developing the FA scale, the authors kept five essential requirements in mind:

1. Breaking down the relevant FA components (identifying and outlining the specific components or attributes that make up functional adequacy).
2. The FA descriptors should be independent of linguistic descriptors in terms of CAF.
3. The descriptors should be objective and quantifiable.
4. The scale should be applicable for both expert and non expert raters.
5. It should be usable for both second language (L2) and first language (L1) assessment. (Kuiken & Vedder, 2017)

Kuiken and Vedder (2017) highlight that the six-point Likert-scale nature of the FA scale is grounded in Grice's (1975) conversational maxims, which focus on the quantity, relevance, manner, and quality of the message conveyed from writer A to interlocutor B. The new FA rating scale, defined by A's successful task completion in conveying a message to B, is structured around four scale dimensions according to Grice's conversational maxims.

1.2.2 Testing Functional Adequacy

To evaluate the reliability of the new Functional Adequacy rating scale, authors utilized data from the CALC project (Bartning et al., 2010) A group of non-expert evaluators assessed the fluency and accuracy of two writing samples produced by L2 learners. These two texts that were used for the testing were written by L2 learners of Dutch and Italian languages, while the texts written by the L1 speakers of these languages were used as benchmarks. Raters with no expertise evaluated both L2 and L1 texts using the six-point scale, assessing four aspects of FA. The core research question explored in this paper was: how do non-expert evaluators assess the functional adequacy of argumentative texts produced by nonnative Dutch and Italian language learners using a six-point Likert scale encompassing four dimensions of functional adequacy. The raters did not have any prior experience in evaluating texts, which is also the case in our own study. The original study found high interrater reliability scores, ranging from .725 to .940, among the four raters of Dutch and Italian. The four dimensions of functional adequacy demonstrated strong correlations with each other. Notably, L1 writers outperformed L2 learners on all dimensions (which is expected and another aspect that proves the scale's reliability), and L2 Italian writers achieved higher scores than L2 Dutch writers. Furthermore, the correlations between texts written by the same participant were high, with coefficients ranging from .455 to .877, indicating consistency in rater judgments. Overall these results strongly suggest that, the FA scale can be utilized with non expert users to asses functional aspects of a given text which provides a solid foundation for our study.

1.2.3 Pedagogical Implications and Further Research

In their more recent article, Kuiken & Vedder (2022) reviewed several studies that tested the reliability of the FA rating scale, and they have extended their discussion to explore the pedagogical implications of using this scale in language learning and assessment. They delve deeper into how the FA scale can be employed as a valuable tool for teachers to diagnose learners' strengths and weaknesses, provide targeted feedback, and guide instructional decisions (Kuiken & Vedder, 2022). The authors also consider the potential of the FA scale for promoting learner autonomy through self and peer assessment. However, they also highlight that more research is needed to fully understand the effectiveness and limitations of using the FA scale in these contexts. With this study, we hope to contribute to this field by investigating how well AI generated texts align with the functional aspects of human-written texts and how effectively they can be tailored to specific proficiency levels. The insights gained from this research will help to expand our understanding of the FA scale's potential and limitations in the context of AI-assisted language education, thereby contributing to the ongoing discourse on the integration of AI tools in educational settings.

1.3 Advancements and Comparisons in Large Language Model Technologies

The landscape of large language models is dynamic, with rapid changes and improvements that continuously redefine state-of-the-art capabilities. This fast-paced evolution means that what is considered cutting-edge today may be surpassed within weeks or even days. As of the writing of this thesis in mid-2024, the field has seen significant developments, both in terms of model architecture and the growing ubiquity of open-source alternatives that rival close-source models (such as GPT-3.5 of OpenAI).

As one of the central research questions of this thesis is the comparison between the capabilities of open-source and closed-source models, it was important to select what was considered state-of-the-art in the open-source domain at the time of the study. To this end, we utilized the Mixtral 8x7B model, an open-source large language model known for its robust performance (REF). At the time of writing, this model demonstrated capabilities that were on a par with—or in some cases slightly superior to—those of OpenAI's GPT-3.5 on benchmarks such as MMLU, HellaSwag, and ARC Challenge (2023). We have chosen to work with GPT-3.5 since it is the baseline model that OpenAI provides to its users, meaning that without any subscription for limited usages, any user can utilize this model through their website or via the OpenAI API. Users are charged $0.50 / 1M tokens as input and $1.50 / 1M tokens for the produced outputs and inputs (OpenAI, 2024). Mixtral 8x7B, is built on a Sparse Mixture of Experts (SMoE) architecture. This model is a decoder-only transformer where each layer consists of multiple "experts" or feedforward blocks. For every token processed, a router network selects two out of eight experts to handle the token, combining their outputs additively. This approach allows Mixtral to leverage a large number of parameters (47 billion) while only using a fraction (13 billion) during inference, optimizing both performance and computational efficiency (2023)

1.4 Why Do Open Source LLMs Matter?

A study done by Chang et al. (2024) compared several open- and closed-source models, highlighting that while closed-source models like GPT-4 and GPT-3.5 generally outperform their open-source counterparts in summarization tasks, they come with higher usage costs and significant privacy concerns due to API-based accessibility. Conversely, despite being markedly smaller, open-source models such as LLaMA-2 showcased comparable capabilities in zero-shot scenarios. Zero-shot learning refers to a model's ability to perform a task without any explicit training or fine-tuning for that specific task. In other words, the model is able to generate reasonable outputs for a task it has never encountered before, solely based on its pre-existing knowledge and understanding of language. The study emphasized LLaMA-2's cost-effectiveness and privacy advantage, making it a more suitable choice for industrial usage due to lower operational costs and enhanced data control, aligning closely with the needs of educational platforms that prioritize confidentiality and budget efficiency. It is worth noting that during the writing of this paper, Meta AI has released their brand new model that now became the industry standard in the open-source scene. LLaMa 3, the latest language model from Meta, showcases significant advancements over its predecessor, LLaMa 2, in various aspects of performance and capabilities (Meta AI, 2024). As the capabilities of open-source language models are catching up, these models are becoming even better solutions for providing flexible educational solutions for professionals. As it is highlighted by Chang et al. (2024), closed-source language models are only accessible with paid API usage; however, now very high performing yet small models such as LLaMa3-8b can be efficiently run on a consumer-grade computer without any additional cost for text inference. For the educational domain, this aspect of open-source models makes them an excellent solution for educators to build and test ideas cost-free.

Privacy is another important advantage of open-source models. Since an open-source model has open weights and is downloadable directly, it can be used entirely offline and integrated into users' own personal data without having to worry about the data being shared with third parties. In the context of generating personalized text for learners, retaining the users’ data has a great importance. This is crucial in ensuring that sensitive learner information is not shared with third-party providers or exposed to potential data breaches. By using open-source LLMs, educators can develop personalized educational content while guaranteeing the confidentiality and integrity of learner data, ultimately creating a safer and more trustworthy learning environment.

With this in mind, we believe that testing these models in the domain of functional adequacy will provide great insight into educational text production and evaluation.

2 Methods

2.1 Participants

Power analysis was carried out to determine the sample size required for the study. Based on the power analysis, a total of 24 participants were recruited and tested. The participants were native Italian speakers aged 19 to 28, 19 of these participants were females and 5 of them were males, without any specialized expertise in language education or linguistics. They were all university students. All participants were healthy, had normal or corrected to normal vision, and no known language or neurocognitive disorders. They all gave written informed consent prior to participation.

2.2 Stimuli

2.2.1 Human Texts

For our baseline texts, we have gathered 72 texts, 24 per language level, from various language teaching sources such as language book for L2 learners, written by humans. These sources were specifically designed for learners of Italian as a second language at different proficiency levels. We have collected each text from sources where the level was explicitly stated. These levels were used as the first parameter that we fed into the LLM prompts. The identified texts cover a wide range of topics. This was particularly challenging to find for the beginner levels (levels A1 – A2), as instructional texts for such learners are generally cantered around very simple themes, covering a narrow range of topics such as daily life, hobbies, and school.

Calculated Metrics for Human Texts

In our study, we calculated specific metrics for human-written texts to ensure that the AI-generated texts could be matched to them in terms of length and readability. These metrics include sentence count, the Gulpease Index, and the Flesch Reading Ease score. Each of these metrics serves a distinct purpose in our analysis.

Sentence Count

We calculated the sentence count for human-written texts to match the length of the texts generated by AI models. Models were prompted to generate texts within a range of ±3 sentences of the human-written text's length.

Gulpease Index

The Gulpease Index is a readability metric specifically calibrated for the Italian language. It is calculated based on the number of sentences, words, and characters in a text. The index ranges from 0 to 100, with higher scores indicating easier readability. This index was not used during the text generation process but served as a baseline to evaluate and compare the overall readability level of both human-written and AI-generated texts. The Gulpease Index helps us understand how well the texts align with the expected proficiency levels.

$$GI = 89 + \frac{(300 * number\ of\ sentences)\ - 10 * (number\ of\ letters\ )}{number\ of\ words}$$

Flesch Reading Ease

The Flesch Reading Ease score is a widely used readability metric for English-language texts. It measures the ease of understanding a text based on sentence length and the number of syllables per word. The score ranges from 0 to 100, with higher scores indicating easier readability. Like the Gulpease Index, the Flesch Reading Ease score was not used in the generation process but provided a baseline for assessing the readability of the texts. This score helps us ensure that the AI-generated texts are comparable to human-written texts in terms of readability. For this study we have used a version of the Flesch formula proposed by Roberto Vacca and Valerio Franchina (1986)

$$F = 217 - (1.3 * average\ number\ of\ syllables\ per\ word) - (0.6 * average\ number\ of\ words\ per\ sentence)$$

2.2.2 AI generated texts

Seventy-two texts, matching the human-written texts in language level, topic and length were generated using the closed-source LLM ChaptGPT 3.5 and the open-source LLM Mixtral. The language level chosen was the one explicitly stated (e.g., A1, B2) in the language teaching manuals we extracted the texts from. This level served as the first parameter fed into the language model.

Secondly, the topics of the texts were identified and summarized into concise 2-3 word descriptions. These topic summaries were then fed into the prompt as the target topic for the given text. The final parameter involved specifying a target range for the length of the generated text, which was set to +/- 3 sentences for each human-written text length. Providing a range, rather than an exact target number, tends to yield better results in terms of matching the desired length while allowing for some natural variation in the generated texts.

We chose to keep the prompting structure rather simple given the fact that one of the aims was to mimic how an everyday user might utilize these models.

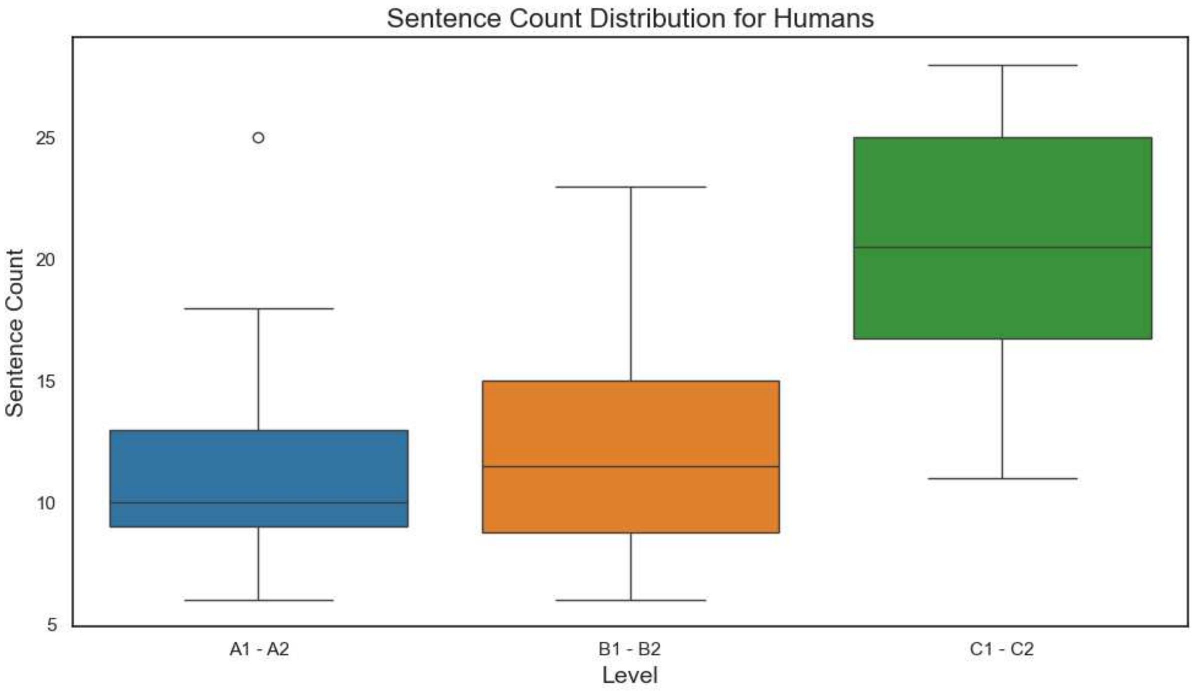
Using the above parameters, a for loop was created in Python to make API calls to the OpenAI servers, generating 72 texts with the desired parameters using ChatGPT 3.5 (24 texts for each level). For example prompts, see Appendix A

For the open-source model, we utilized a service called Ollama. Ollama is a platform designed to run large language models locally on users' machines. It supports a variety of open-source models, including Llama 3, Mistral, and Code Llama, among others. Ollama was used to generate (24 x 3 =) 72 texts for the open-source model Mixtral with the same parameters as for ChatGPT.

2.2.3 The comparison of human-written and AI generated texts

Sentence Count In terms of Proficiency Levels

Human written texts we gathered showed a mean sentence count of 11.54 (SD = 4.20, range 6-25) for the A1 - A2 level, 12.54 (SD = 5.34, range 6-23) for the B1 - B2 level, and 20.67 (SD = 4.62, range 11-28) for the C1 - C2 level.



Level	n	M	SD	Min	25th percentile	50th percentile	75th percentile	Max
A1 - A2	24	11.54	4.20	6	9.00	10.0	13.0	25
B1 - B2	24	12.54	5.34	6	8.75	11.5	15.0	23
C1 - C2	24	20.67	4.62	11	16.75	20.5	25.0	28

Figure 1

Sentence count distribution for human written texts

For GPT-3.5-Turbo, the mean sentence count for the A1 - A2 level was 16.38 (SD = 6.45, range 6-30), with a range from 6 to 30 sentences. For the B1 - B2 level, the mean was 17.00 (SD

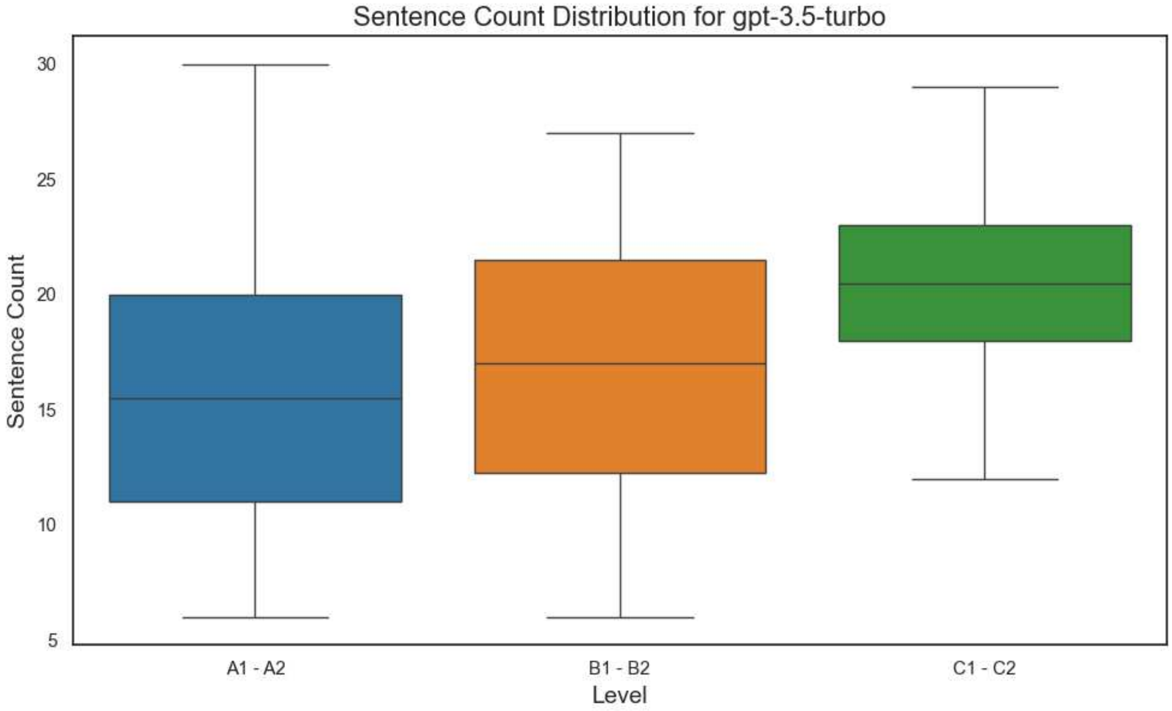
= 6.62, range 6-27), and for the C1 - C2 level, the mean was 20.50 (SD = 4.30, range 12-29). Compared to human-generated texts, GPT-3.5-Turbo produced a higher mean sentence count at the A1 - A2 and B1 - B2 levels, indicating a tendency to generate longer texts at these proficiency levels.

Level	**n**	**M**	**SD**	**Min**	**25th percentile**	**50th percentile**	**75th percentile**	**Max**
A1 - A2	24	16.38	6.45	6	11.00	15.5	20.0	30
B1 - B2	24	17.00	6.62	6	12.25	17.0	21.5	27
C1 - C2	24	20.50	4.30	12	18.00	20.5	23.0	29

Figure 2

Sentence count distribution for gpt-3.5-turbo written texts

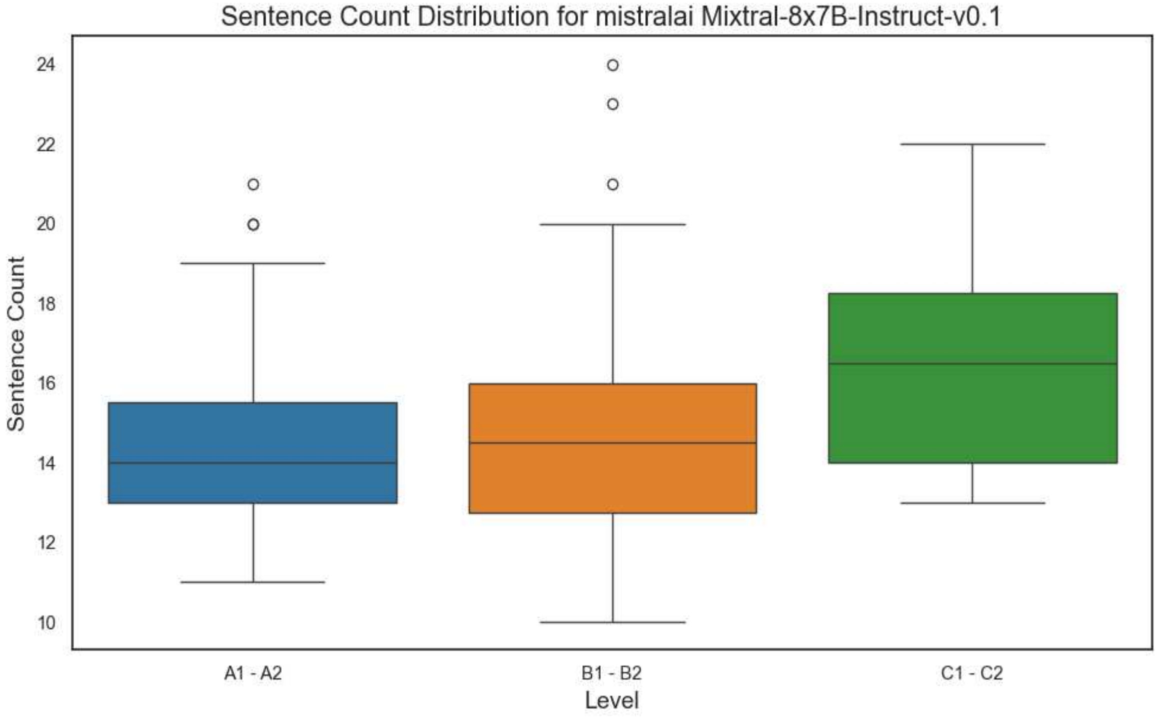
Texts generated by Mistralai Mixtral-8x7B-Instruct-v0.1 had a mean sentence count of 14.75 (SD = 2.88 range 11 - 23) for the A1 - A2 level, 15.13 (SD = 3.75, range 10 -24) for the B1 - B2 level, and 16.54 (SD = 2.70, range 13 - 22) for the C1 - C2 level. Similarly to GPT-3.5-Turbo, this model also produced longer texts than humans at the A1 - A2 and B1 - B2 levels, though its mean sentence counts were slightly lower than those of GPT-3.5-Turbo.



Level	n	M	SD	Min	25th percentile	50th percentile	75th percentile	Max
A1 - A2	24	14.75	2.88	11	13.00	14.0	15.5	21
B1 - B2	24	15.13	3.75	10	12.75	14.5	16.0	24
C1 - C2	24	16.54	2.70	13	14.00	16.5	18.25	22

Figure 3

Sentence count distribution for Mistralai Mixtral-8x7B-Instruct-v0.1 written texts



A one-way ANOVA revealed a statistically significant difference in sentence count between sources for the A1 - A2 level, F(2, 69) = 6.45, p = .003. For the B1 - B2 level, the analysis indicated a significant difference in sentence count between sources, F(2, 69) = 4.18, p = .019. Similarly, for the C1 - C2 level, there was a significant difference in sentence count between sources, F(2, 69) = 8.32, p < .001. These findings suggest that the source of the sentences significantly affects the sentence count across all language levels tested.

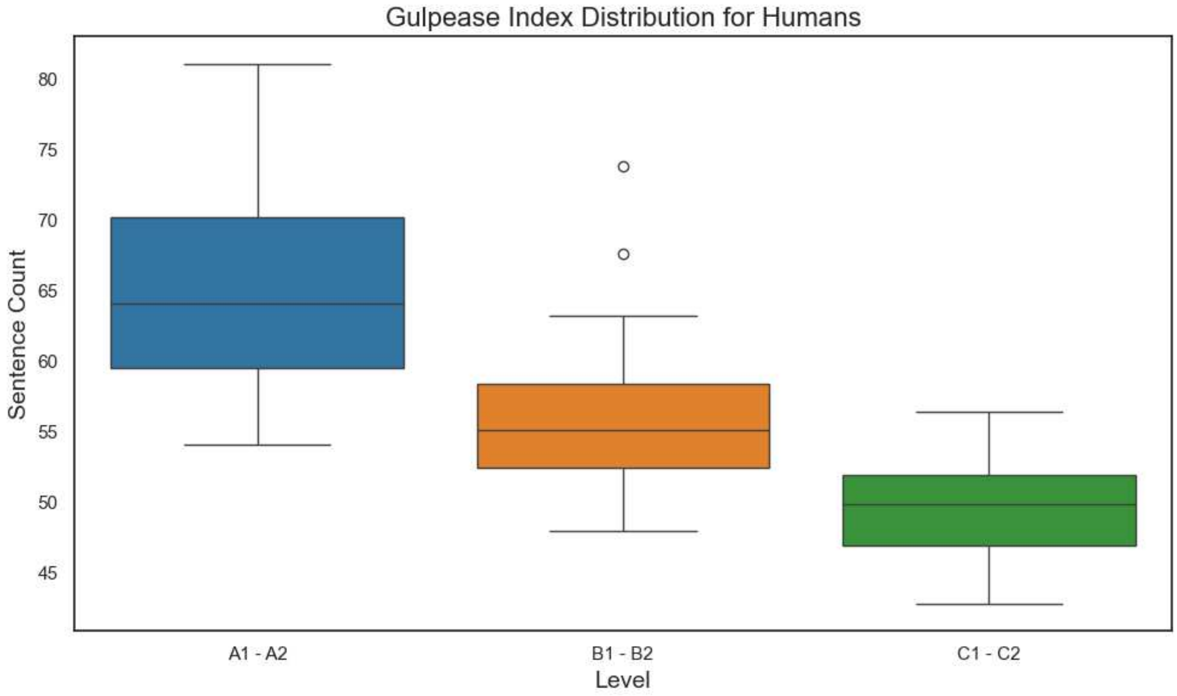
Gulpease Index In terms of Proficiency Levels

Human written texts showed a mean Gulpease Index of 64.82 (SD = 7.22, range 54.1 – 81.1) for the A1 - A2 level, 56.13 (SD = 5.99, range 48 – 73.8) for the B1 - B2 level, and 49.70 (SD = 3.35, range 42.8 – 56.4) for the C1 - C2 level.

Level	**n**	**M**	**SD**	**Min**	**25th percentile**	**50th percentile**	**75th percentile**	**Max**
A1 - A2	24	64.82	7.22	54.1	59.53	64.05	70.23	81.1
B1 - B2	24	56.13	5.99	48.0	52.48	55.10	58.40	73.8
C1 - C2	24	49.70	3.35	42.8	46.95	49.85	51.95	56.4

Figure 4

Gulpease Index distribution for human written texts

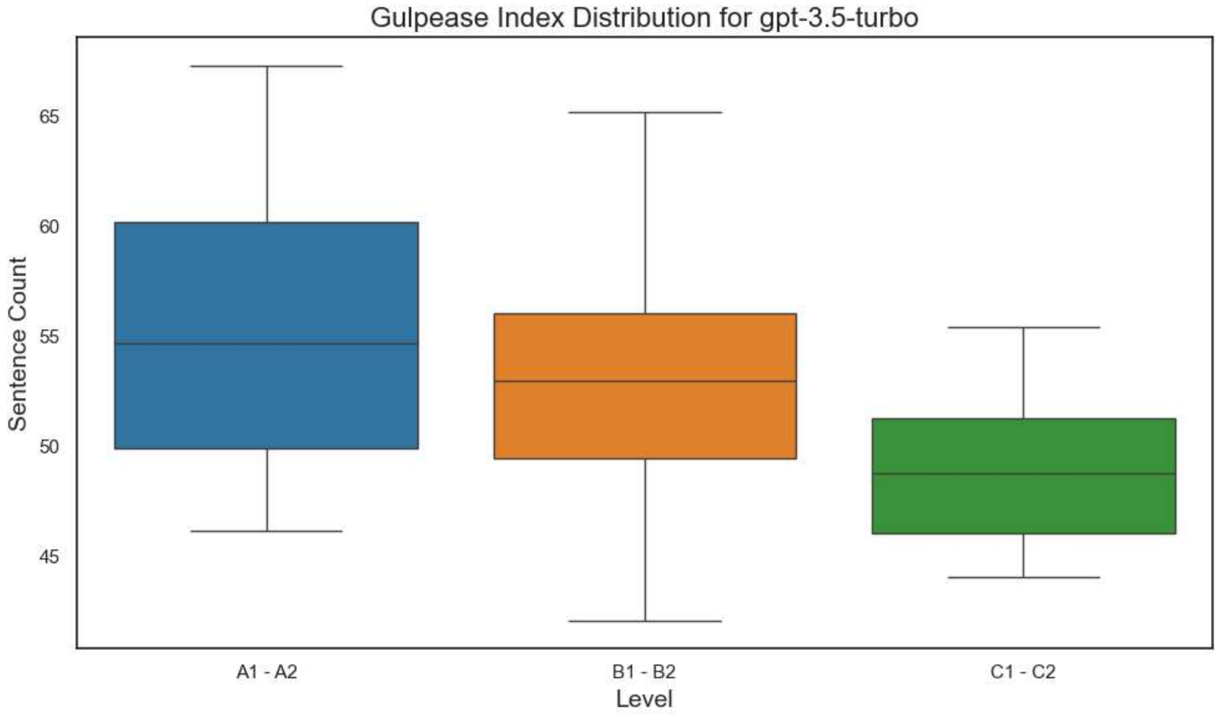


For GPT-3.5-Turbo, the mean Gulpease Index for the A1 - A2 level was 55.35 (SD = 6.19), with a range from 46.2 to 67.3. For the B1 - B2 level, the mean was 53.24 (SD = 5.59, range 42.1 – 65.2), and for the C1 - C2 level, the mean was 49.22 (SD = 3.65, range 44.1 – 55.4). Compared to human-generated texts, GPT-3.5-Turbo produced a lower mean Gulpease Index at the A1 - A2 and B1 - B2 levels, indicating a tendency to generate texts that are potentially less readable at these proficiency levels.

Level	**n**	**M**	**SD**	**Min**	**25th percentile**	**50th percentile**	**75% percentile**	**Max**
A1 - A2	24	55.35	6.19	46.2	49.95	54.70	60.18	67.3
B1 - B2	24	53.24	5.59	42.1	49.45	53.00	56.05	65.2
C1 - C2	24	49.22	3.65	44.1	46.05	48.80	51.30	55.4

Figure 5

Gulpease Index distribution for gpt-3.5-turbo written texts

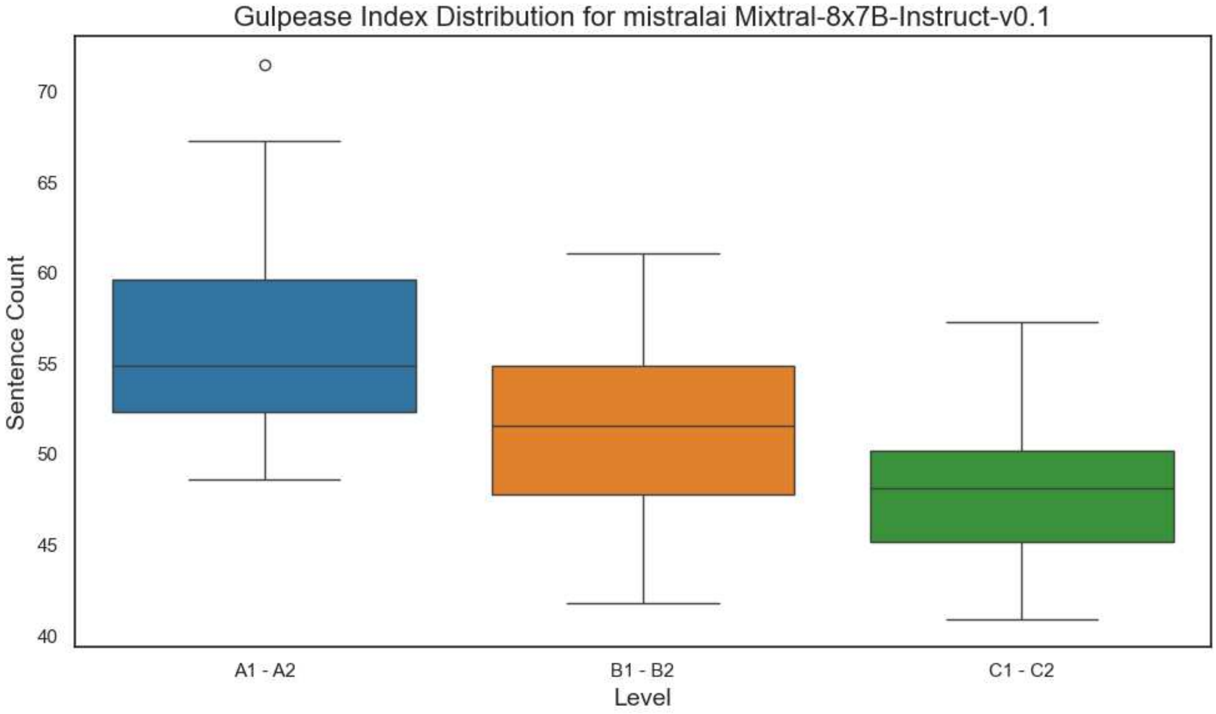


Texts generated by Mistralai Mixtral-8x7B-Instruct-v0.1 had a mean Gulpease Index of 56.51 (SD = 5.90, range 48.6 – 71.5) for the A1 - A2 level, 51.27 (SD = 5.34, range 41.8 – 61.1) for the B1 - B2 level, and 47.98 (SD = 3.51, range 40.9 – 57.3) for the C1 - C2 level. Similarly to GPT-3.5-Turbo, this model also produced a lower mean Gulpease Index than humans at all levels, though its mean Gulpease Index for the A1 - A2 level was slightly higher than that of GPT-3.5-Turbo.

Level	**n**	**M**	**SD**	**Min**	**25th percentile**	**50th percentile**	**75% percentile**	**Max**
A1 - A2	24	56.51	5.90	48.6	52.30	54.90	59.63	71.5
B1 - B2	24	51.27	5.34	41.8	47.80	51.55	54.88	61.1
C1 - C2	24	47.98	3.51	40.9	45.20	48.10	50.18	57.3

Figure 6

Gulpease Index distribution for Mistralai Mixtral-8x7B-Instruct-v0.1 written texts



A one-way ANOVA was conducted to compare the effect of the source on the Gulpease Index across different language levels. For the A1 - A2 level, there was a statistically significant difference in the Gulpease Index between sources, F(2, 69) = 15.33, p < .001. For the B1 - B2 level, the analysis also revealed a significant difference in the Gulpease Index between sources, F(2, 69) = 4.51, p = .014. However, for the C1 - C2 level, there was no significant difference in the Gulpease Index between sources, F(2, 69) = 1.53, p = .224. These results suggest that the source of the text significantly affects the Gulpease Index for the A1 - A2 and B1 - B2 levels, but not for the C1 - C2 level

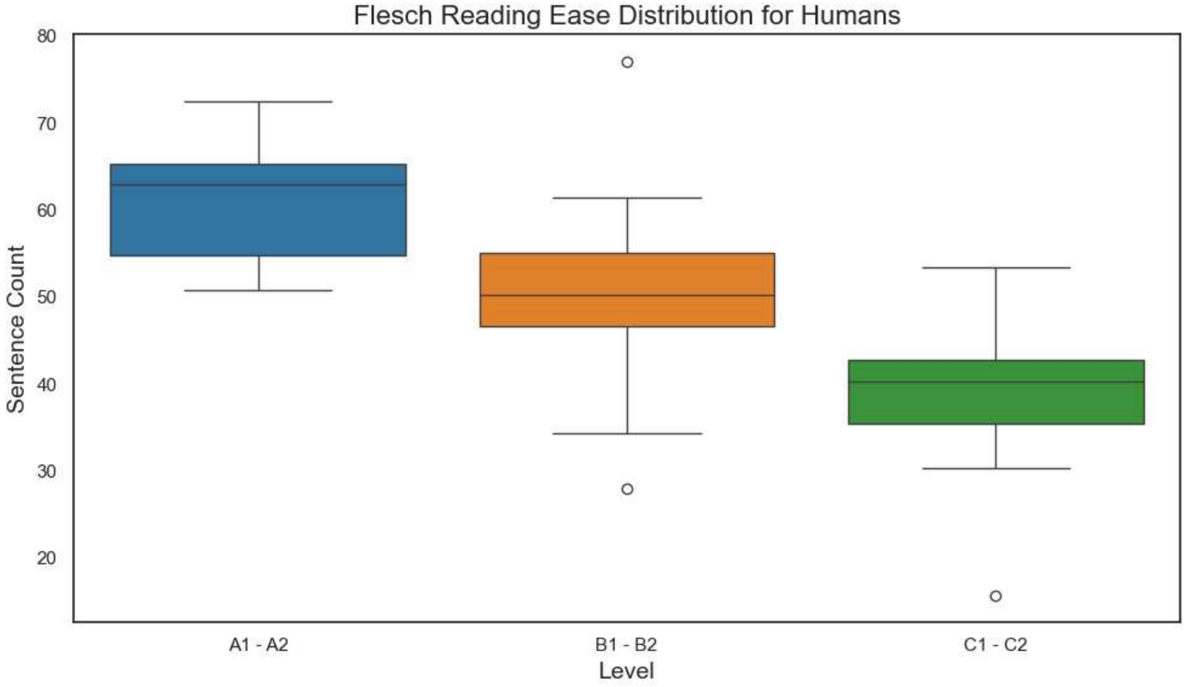
Flesch Reading Ease In terms of Proficiency Levels

Human-generated texts showed a mean Flesch Reading Ease score of 61.24 (SD = 7.00, range 50.77 – 72.46) for the A1 - A2 level, 50.34 (SD = 9.74, range 28.03 – 77.13) for the B1 - B2 level, and 39.53 (SD = 8.09, range 15.71 – 53.51) for the C1 - C2 level.

					25th	**50th**	**75%**	
Level	**n**	**M**	**SD**	**Min**	**percentile**	**percentile**	**percentile**	**Max**
A1 - A2	24	61.24	7.00	50.77	54.86	62.97	65.35	72.46
B1 - B2	24	50.34	9.74	28.03	46.63	50.26	55.06	77.13
C1 - C2	24	39.53	8.09	15.71	35.53	40.28	42.82	53.51

Figure 7

Flesch Reading Ease distribution for human written texts

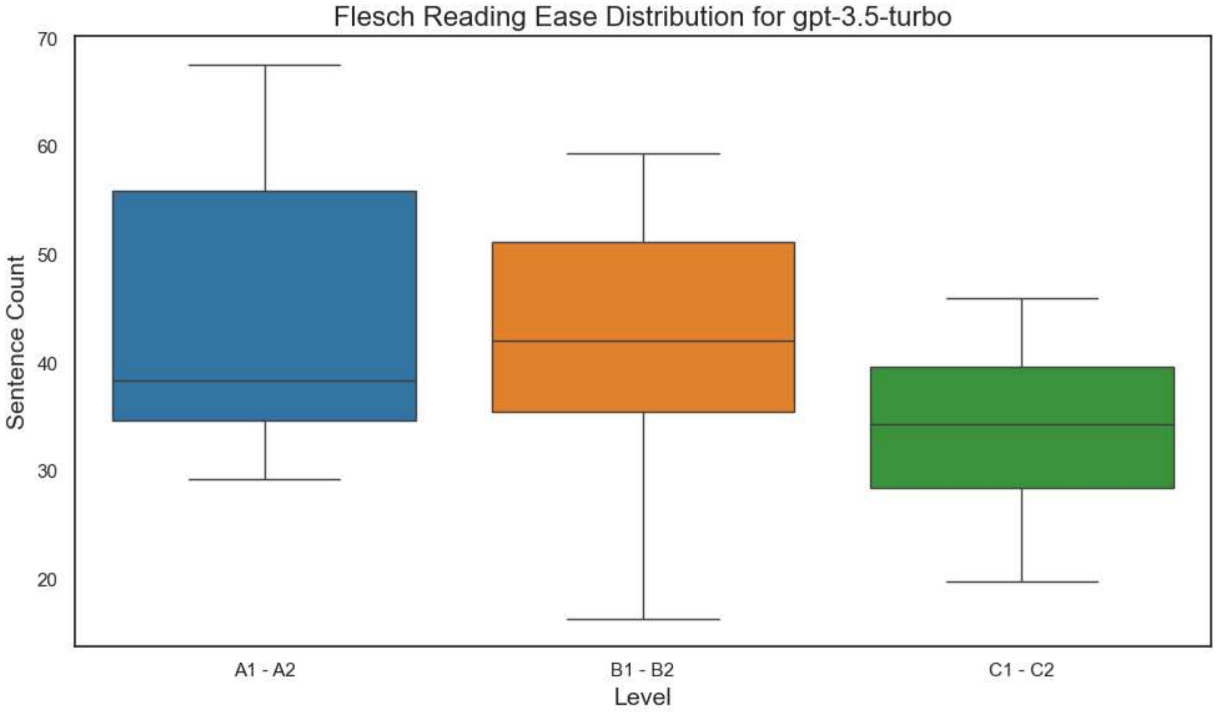


For GPT-3.5-Turbo, the mean Flesch Reading Ease score for the A1 - A2 level was 43.71 (SD = 11.70), with a range from 29.25 to 67.65. For the B1 - B2 level, the mean was 41.21 (SD = 12.16, range 16.32 – 59.40), and for the C1 - C2 level, the mean was 34.13 (SD = 6.93, range 19.77 – 45.96). Compared to human-generated texts, GPT-3.5-Turbo produced lower mean Flesch Reading Ease scores at all proficiency levels, indicating a tendency to generate texts that are more difficult to read.

Level	**n**	**M**	**SD**	**Min**	**25th percentile**	**50th percentile**	**75th percentile**	**Max**
A1 - A2	24	43.71	11.70	29.25	34.68	38.37	55.95	67.65
B1 - B2	24	41.21	12.16	16.32	35.52	42.05	51.15	59.40
C1 - C2	24	34.13	6.93	19.77	28.46	34.36	39.71	45.96

Figure 8

Flesch Reading Ease distribution for gpt-3.5-turbo written texts

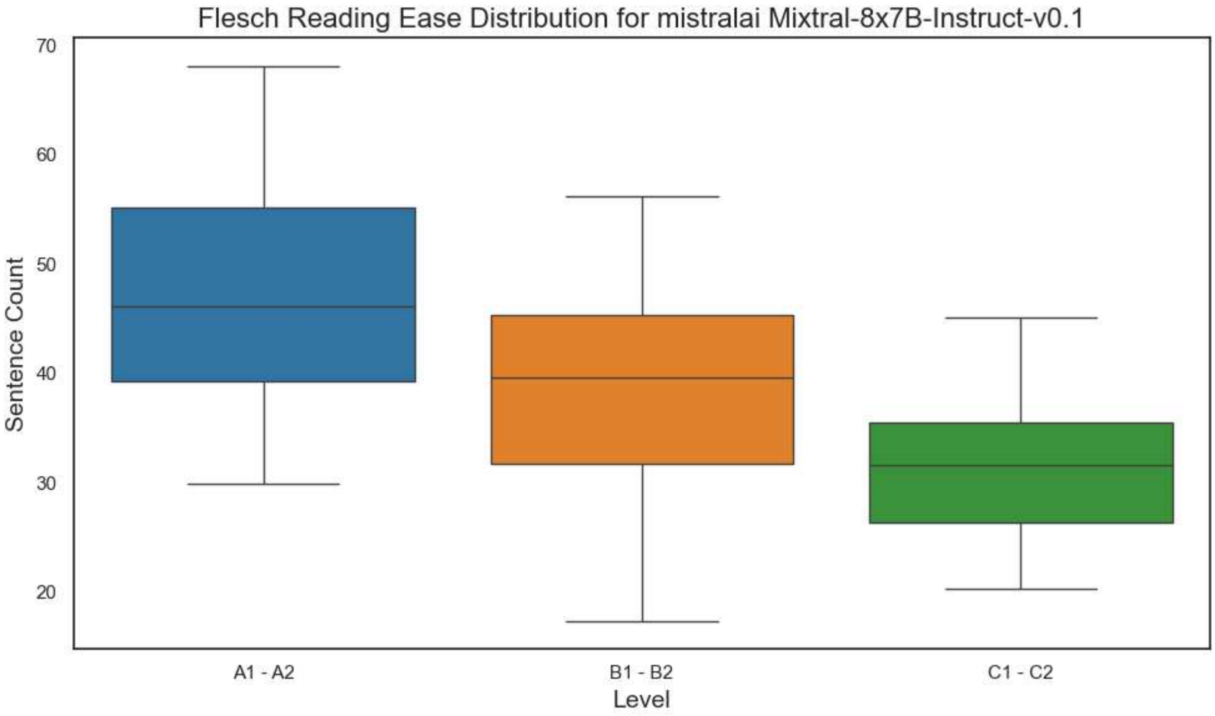


Texts generated by Mistralai Mixtral-8x7B-Instruct-v0.1 had a mean Flesch Reading Ease score of 46.82 (SD = 9.62, range 29.96 – 68.16) for the A1 - A2 level, 38.04 (SD = 10.46, range 17.34 – 56.25) for the B1 - B2 level, and 32.04 (SD = 6.94, range 20.31 – 45.09) for the C1 - C2 level. Like GPT-3.5-Turbo, this model also produced lower mean Flesch Reading Ease scores than humans at all levels.

Level	n	M	SD	Min	25th percentile	50th percentile	75th percentile	Max
A1 - A2	24	46.82	9.62	29.96	39.33	46.22	55.24	68.16
B1 - B2	24	38.04	10.46	17.34	31.79	39.69	45.38	56.25
C1 - C2	24	32.04	6.94	20.31	26.33	31.67	35.55	45.09

Figure 9

Flesch Reading Ease distribution for Mistralai Mixtral-8x7B-Instruct-v0.1written texts



A one-way ANOVA was conducted to compare the effect of the source on the Flesch Reading Ease across different language levels. For the A1 - A2 level, there was a statistically significant difference in Flesch Reading Ease between sources, $F(2, 69) = 22.61$, $p < .001$. For the B1 - B2 level, the analysis also revealed a significant difference in Flesch Reading Ease between sources, F(2, 69) = 8.34, p = .001. Similarly, for the C1 - C2 level, there was a significant difference in Flesch Reading Ease between sources, *F*(2, 69) = 6.64, p = .002. These results suggest that the source of the text significantly affects the Flesch Reading Ease across all tested language levels.

2.3 Experimental Procedure

The experiment was conducted using a custom-built testing platform developed with Streamlit, a Python library that enables the creation of interactive web applications with ease. Streamlit was chosen for several reasons. Firstly, it provides a simple and intuitive way to build web applications, allowing to focus on the content and functionality of the platform rather than the technical aspects of web development. Secondly, Streamlit is highly flexible, enabling the incorporation of various data types, visualizations, and interactive elements, which was crucial for creating an engaging and user-friendly testing environment. Lastly, Streamlit offers the ability to deploy applications to the web using Streamlit Cloud, which could be beneficial if the study needed to be scaled up or made accessible to a wider audience.

The website's flow was designed to be clear and straightforward, guiding participants through the experiment in a logical and organized manner. First, participants were presented with the consent form, where they received detailed information about the study, including its purpose, duration, and any potential risks or benefits. Participants were required to read and agree to the terms outlined in the consent form before proceeding to the next stage of the experiment. Participants then received instructions on how to complete the test. Then the actual test began. During the test phase, each participant was presented with a total of nine texts, which included human-written and AI-generated texts, one from each source, across the three proficiency levels in a pseudo-random order. Participants thus received one text from each proficiency level (A1-A2, B1-B2, and C1-C2) written by a human author, one text from each proficiency level generated by GPT-3.5, and one text from each proficiency level generated by Mistral AI. We presented the texts to the participants one at a time. Participants could not return to a previous text once they had clicked the "Next" button. This prevented participants from changing their responses after having seen subsequent texts, which could have introduced bias or inconsistencies in the data.

For each text, participants were asked to complete four main tasks. The first task was to briefly describe the topic of the text they had just read to probe their understand and attention. The second task was to rate how enjoyable the text was on a scale from one to seven, with one being the least enjoyable and seven being the most enjoyable. This task aimed to capture participants' subjective experiences of reading the texts. The third task required participants to select the perceived level of the text using a dropdown menu with three options: A1-A2, B1-B2, and C1-C2. The final tasks involved answering four multiple-choice questions. We translated the functional adequacy scale into Italian, presenting participants with a simplified version formatted as multiple-choice questions See appendix B for. Each participant repeated this process nine times and was then presented with a screen thanking them for their participation in our study.

3 Results

3.1 Signal Detection Analysis

A signal detection analysis was conducted to evaluate participants' ability to distinguish between AI-generated and human-written texts. The analysis involved four response: hits, misses, false alarms, and correct rejections.

- **Hits:** This response occurs when the text was AI-generated, and participants correctly identified it as AI-generated.
- **Misses:** This response occurs when the text was AI-generated, but participants incorrectly identified it as human-written.
- **False Alarms:** This response occurs when the text was human-written, but participants incorrectly identified it as AI-generated.
- **Correct Rejections:** This response occurs when the text was human-written, and participants correctly identified it as human-written.

The overall distribution of different responses is shown in Figure 10. The hit rate was calculated as the proportion of hits out of the total number of AI-generated texts (hits + misses), resulting in a hit rate of 0.52. The false alarm rate was calculated as the proportion of false alarms out of the total number of human-written texts (false alarms + correct rejections), resulting in a false alarm rate of 0.61. These rates were converted to z-scores, and the d' (d prime) score, a sensitive measure of discrimination which is not affected by response bias, was calculated using the following formula:

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate)$$

The resulting d' score was -0.213, indicating that participants struggled to correctly identify AI-generated texts, often incorrectly identified human written texts as AI generated.

Figure 10

Overall distribution of different response

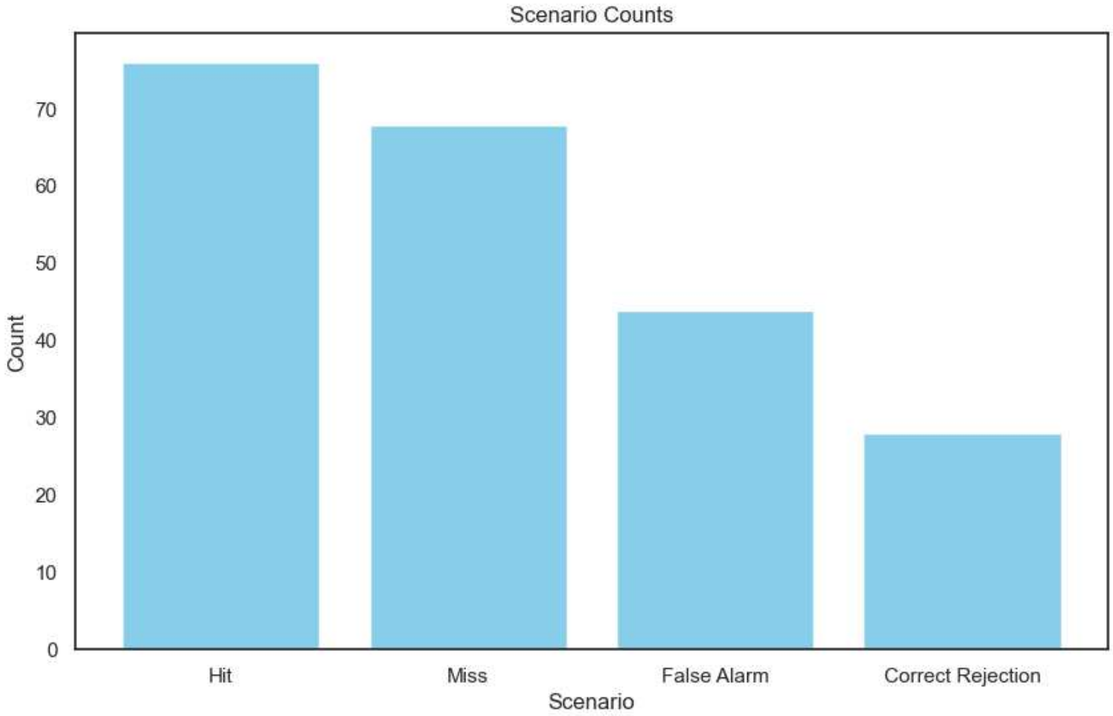
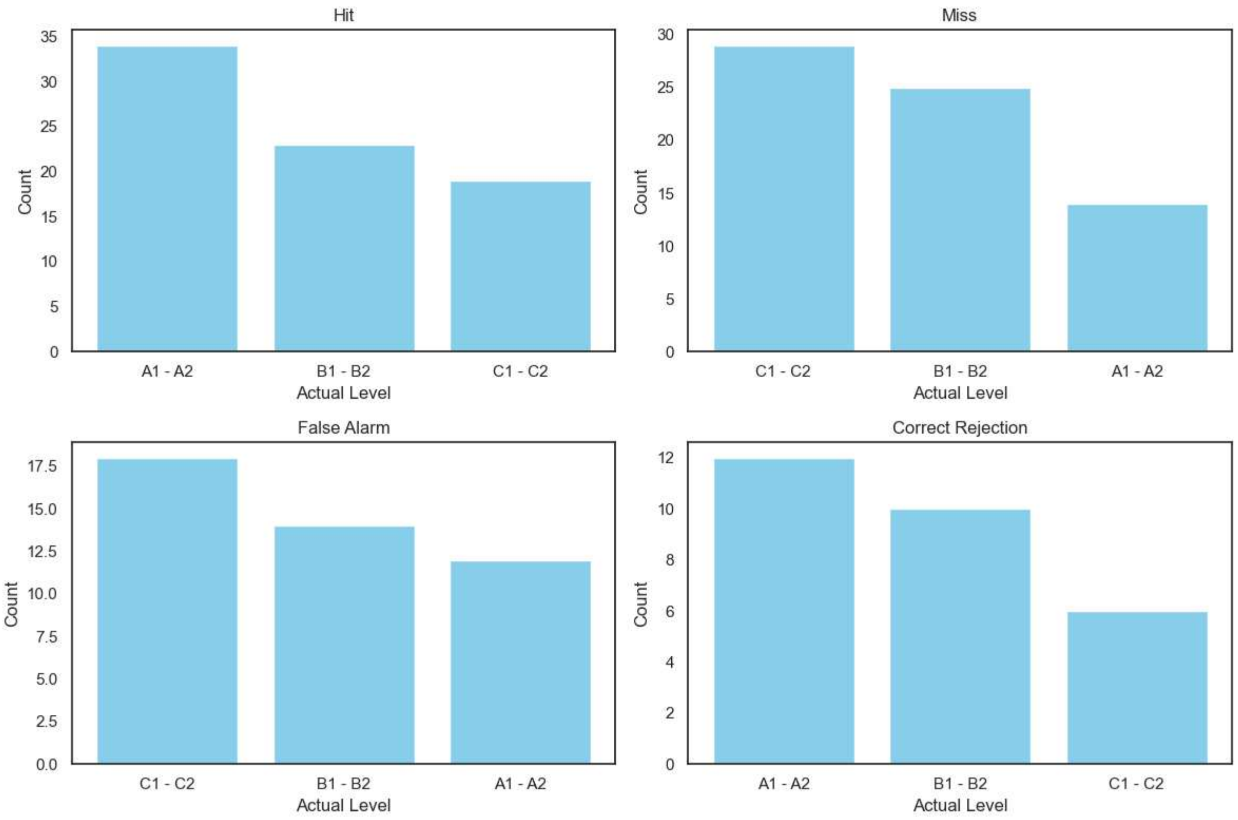


Figure 11

Distribution of response based on different proficiency levels



3.2 Content Evaluation Analysis

An ANOVA was conducted to examine the effect of Source (Human / Mistral / OpenAI) on user content evaluation. The results indicated that there was a statistically significant effect of Source on content evaluation, F(2, 213) = 20.56, p < .001.

A multiple linear regression analysis was performed to investigate the predictors of user content evaluation. The model included Source of the text, actual Proficiency level, User response, Perceived level of the text, and Enjoyment as predictors. The overall model was significant, explaining a substantial proportion of the variance in user content evaluation, $R^2 = 0.503$, F(9, 206) = 23.16, p < .001.The intercept was significant, $\beta = 1.9656$, SE = 0.178, t(206) = 11.058, p < .001, indicating a baseline level of user content evaluation. The source of the content had a significant effect, The overall distribution of different source effect is shown in figure 12 with content produced by OpenAI model being rated higher than the reference category, human-generated text β = 0.2130, SE = 0.086, t(206) = 2.484, p = .014. However, content produced by Mistral did not significantly differ from the reference category, β = -0.0015, SE = 0.087, t(206) = -0.017, p = .986. The actual level of the content did not significantly predict user evaluations, with both B1-B2, β = -0.0305, SE = 0.115, t(206) = -0.265, p = .791, and C1-C2, β = 0.0678, SE = 0.130, t(206) = 0.523, p = .602, showing non-significant effects. The response variable showed mixed results. The overall distribution of different response effect is shown in figure 13 The "Miss" response significantly predicted higher user evaluations, β = 0.2165, SE = 0.089, t(206) = 2.424, p = .016, while the "False Alarm" and "Hit" response did not significantly differ from the reference category, β = -0.1580, SE = 0.156, t(206) = -1.010, p = .314, and β = -0.0050, SE = 0.087, t(206) = -0.058, p = .954, respectively. User-selected levels were significant predictors of user content evaluation. The overall distribution of user selected level is shown in figure 14 Both B1-B2, β = 0.7962, SE = 0.131, t(206) = 6.066, p < .001, and C1-C2, β = 0.8260, SE = 0.148, t(206) = 5.584, p < .001, were associated with higher user evaluations compared to the reference category. Finally, enjoyment was a significant predictor The overall distribution of enjoyment effect is shown in figure 15 β = 0.1538, SE = 0.031, t(206) = 5.035, p < .001, indicating that higher enjoyment levels were associated with higher user content evaluations.

Figure 12

Content evaluation scores by the source of the text

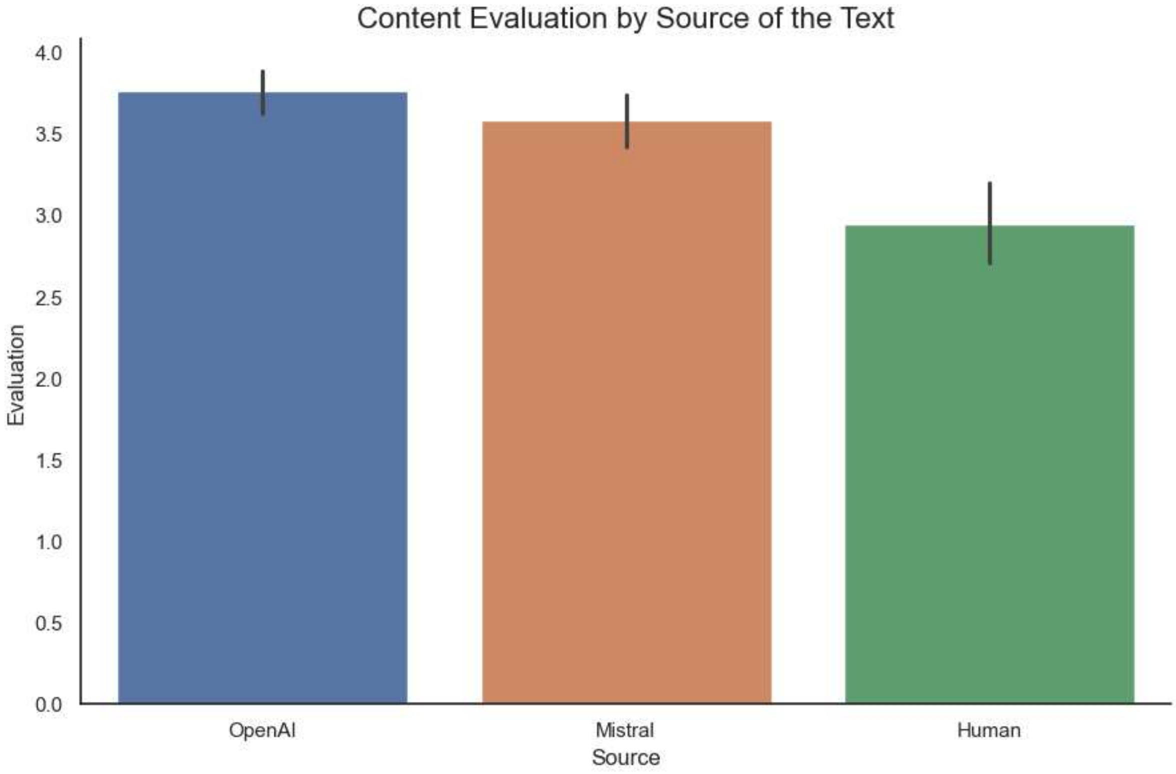


Figure 13

Content evaluation scores by response

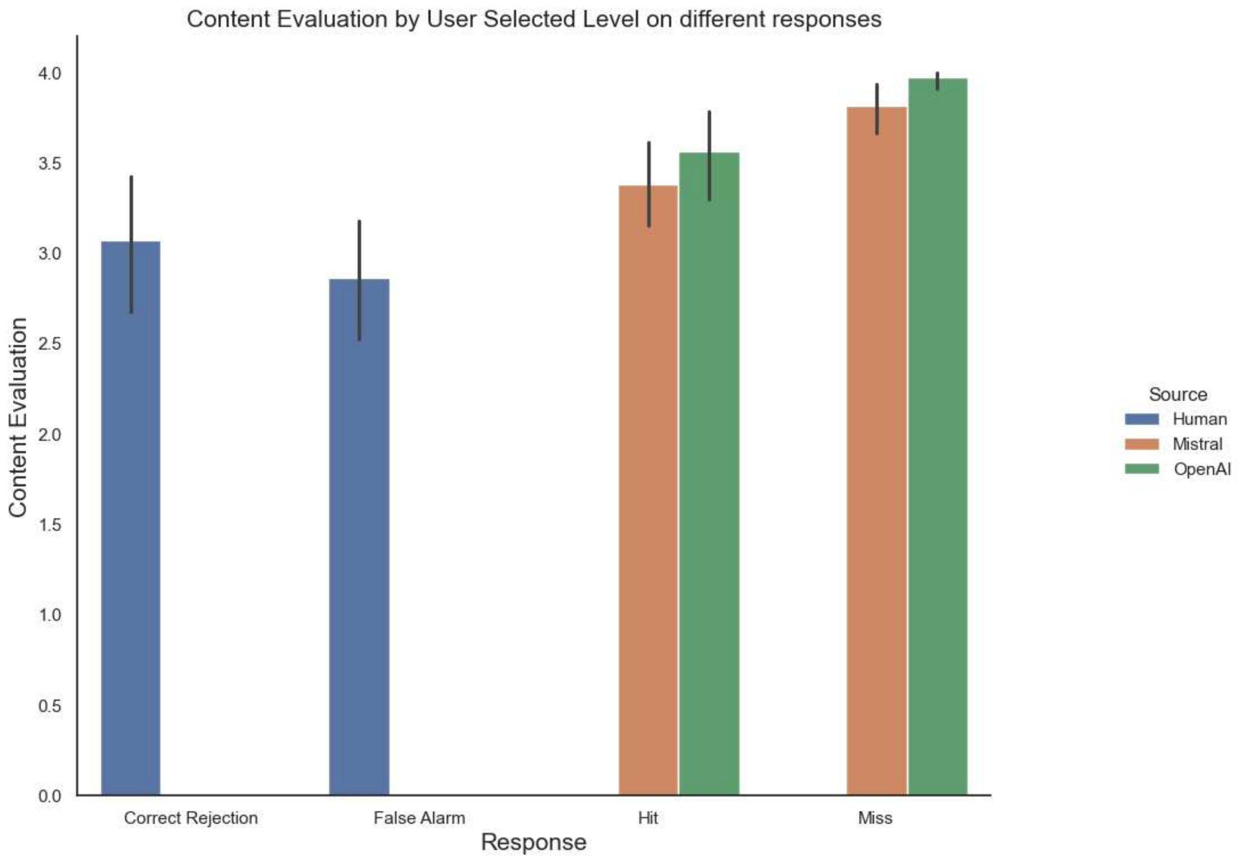


Figure 14

Participant's perceived text complexity and content evaluation

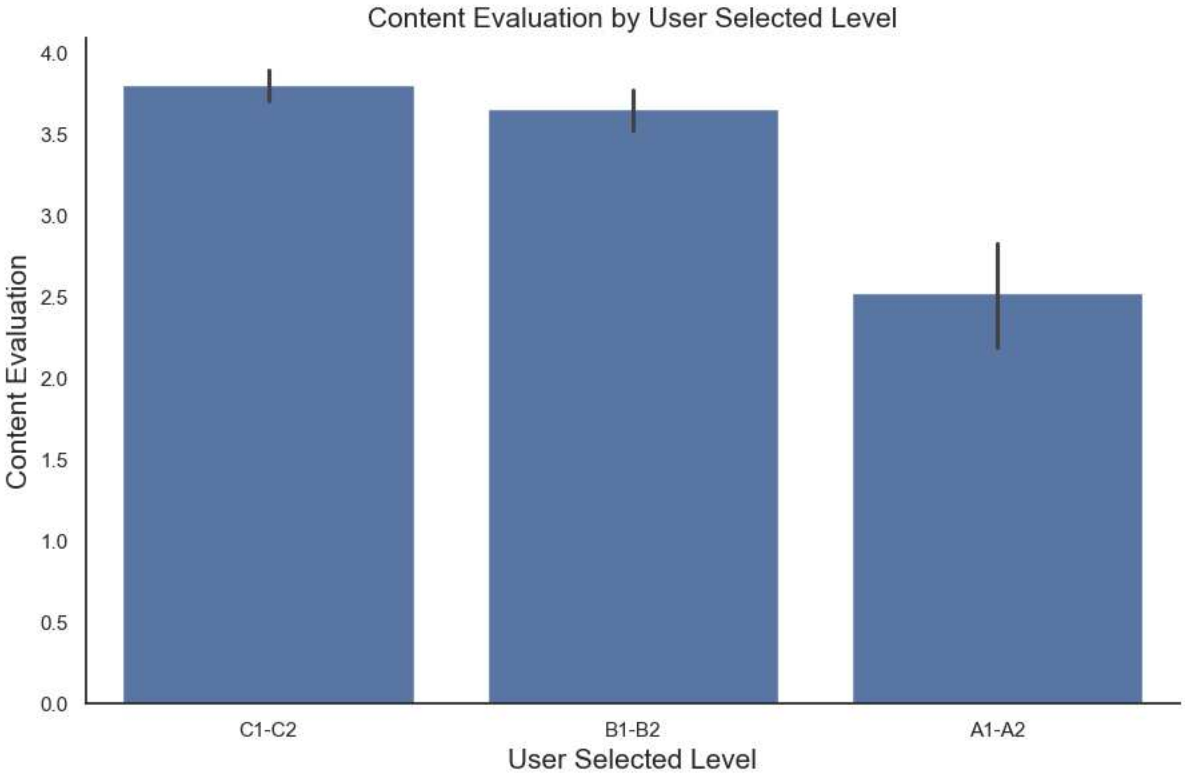
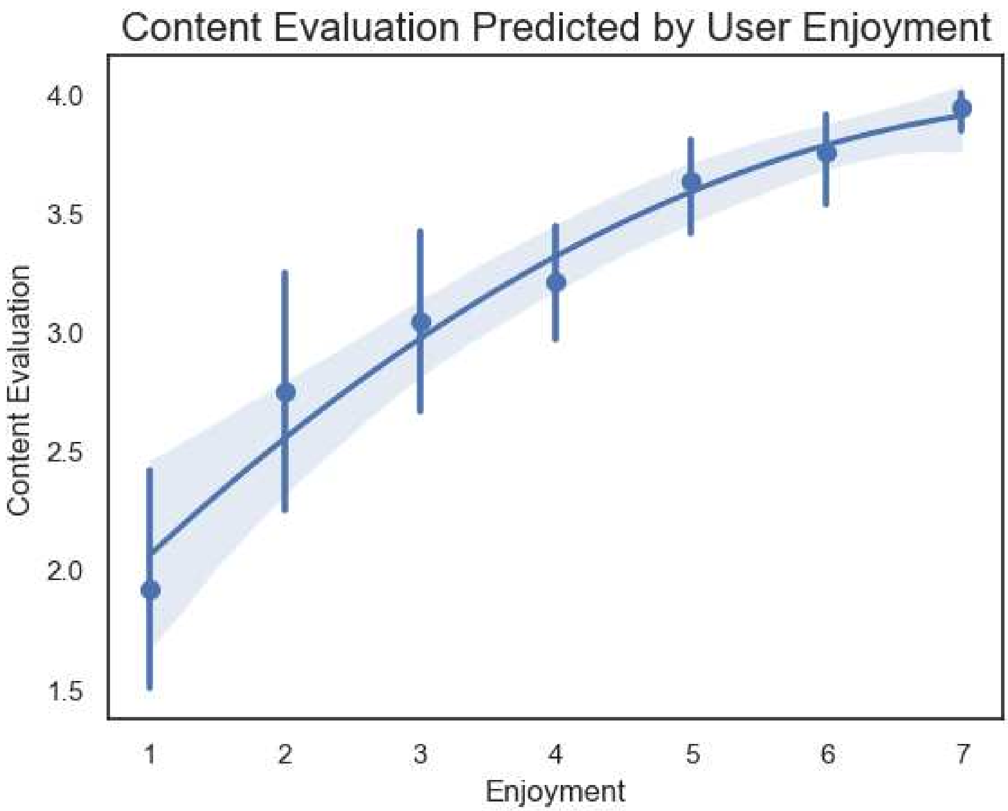


Figure 15

Participants reported enjoyment and content evaluation



3.3 Task Requirements Evaluation Analysis

An ANOVA was conducted to examine the effect of Source on user task requirements evaluation. The results indicated that there was a statistically significant effect of source on task requirements evaluation, *F*(2, 213) = 15.56, *p* < .001. A multiple linear regression analysis was performed to investigate the predictors of user task requirements evaluation. The model included the source, actual level, response, user-selected level, and enjoyment as predictors. The overall model was significant, explaining a substantial proportion of the variance in user task requirements evaluation, $R^2 = 0.536$, F(9, 206) = 26.42, p < .001. The linear regression analysis revealed that the intercept was significant (β = 2.4727, SE = 0.275, t(206) = 9.008, p < .001), indicating a baseline level of user task requirements evaluation. The source of the content from Mistral and OpenAI did not significantly differ from the reference category of human written texts (β = -

0.0958, SE = 0.134, t(206) = -0.716, p = .475 and β = 0.0537, SE = 0.132, t(206) = 0.405, p = .686, respectively). The actual level of the content did not significantly predict user evaluations (β = -0.1857, SE = 0.178, t(206) = -1.043, p = .298 and β = 0.1222, SE = 0.200, t(206) = 0.610, p = .543 for B1-B2 and C1-C2, respectively). The "False Alarm" response significantly predicted lower user evaluations The overall distribution of different response is shown in figure 16 (β = -0.6647, SE = 0.242, t(206) = -2.751, p = .006), while the "Hit" and "Miss" response did not significantly differ from the reference category of correct rejection. User-selected levels B1-B2 and C1-C2 were significant predictors of higher user evaluations The overall distribution of user selected level is hown in figure 17 (β = 0.9068, SE = 0.203, t(206) = 4.473, p < .001 and β = 1.3026, SE = 0.228, t(206) = 5.702, p < .001, respectively). Finally, enjoyment was a significant predictor. The overall distribution of enjoyment is hown in figure 18 indicating that higher enjoyment levels were associated with higher user task requirements evaluations (β = 0.3214, SE = 0.047, t(206) = 6.812, p < .001).

Figure 16

User task requirements evaluation distribution by response

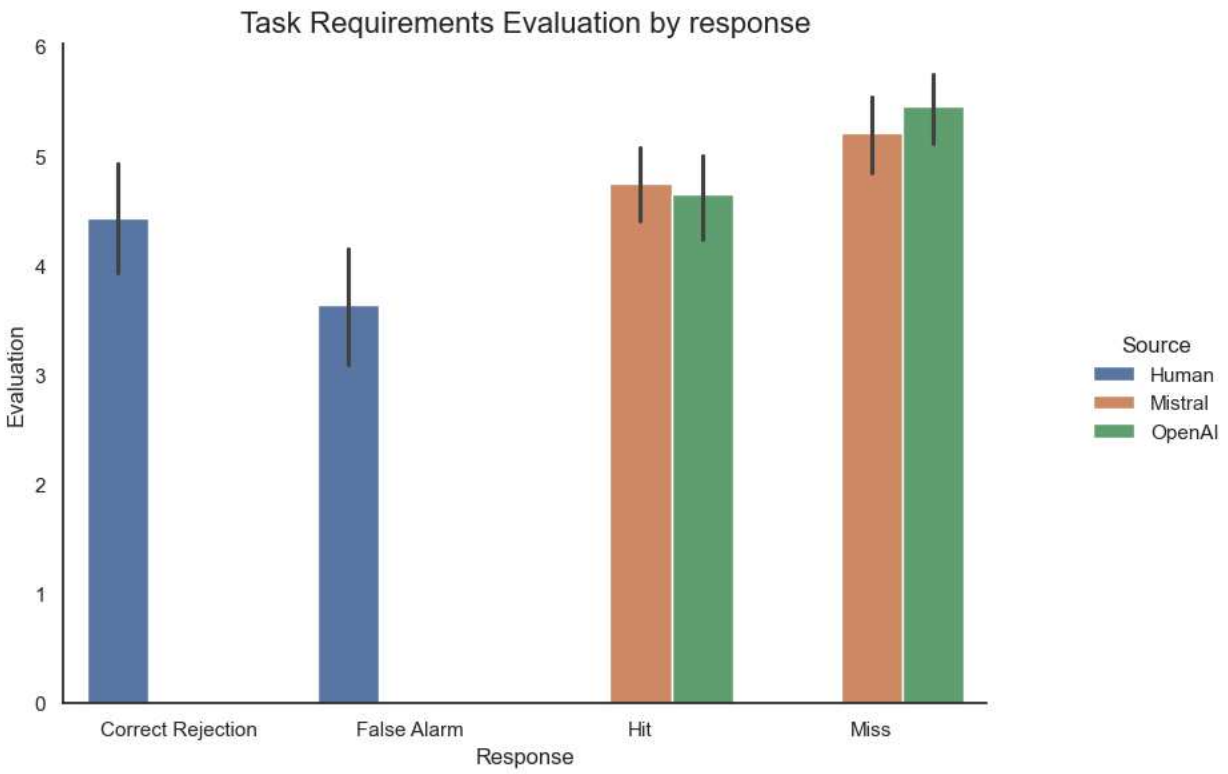


Figure 17

Participant's perceived text complexity and task requirements evaluation

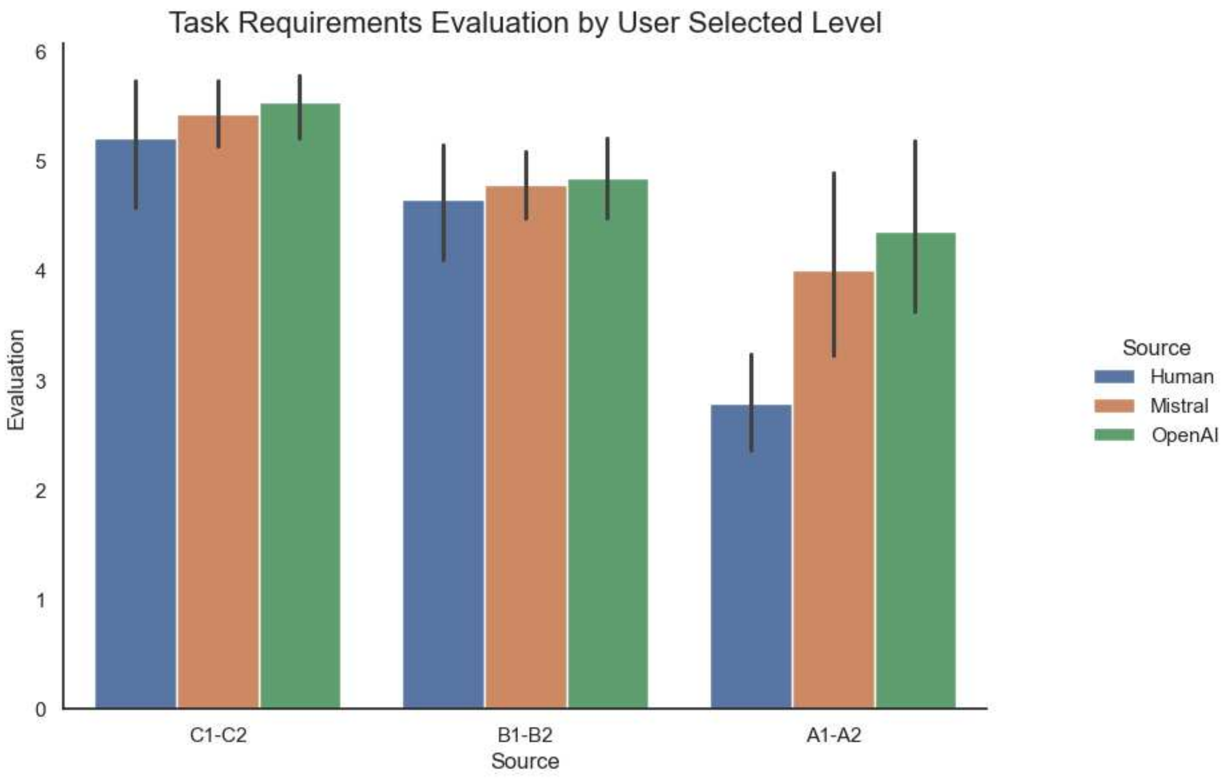
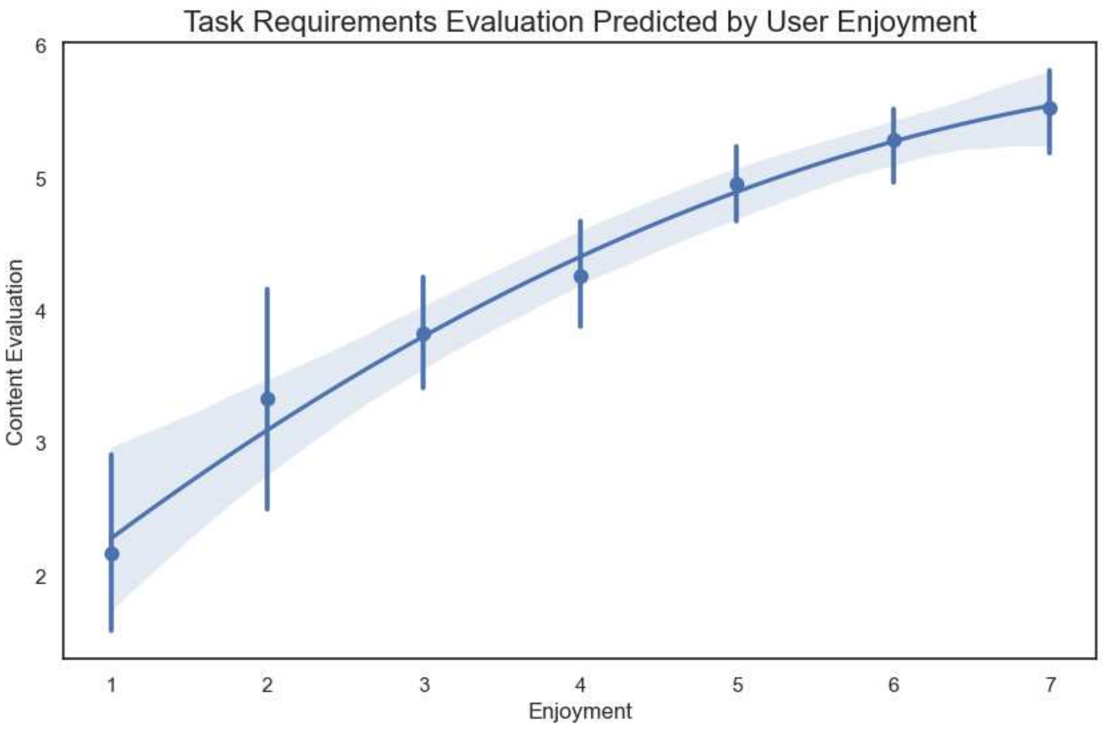


Figure 18

Participants reported enjoyment and task requirements evaluation



3.4 Comprehensibility Evaluation Analysis

An ANOVA test was conducted to examine the effect of the Source on user comprehensibility evaluation. The results indicated a significant effect of the source, F(2, 213) = 19.10, $p < .001$. This suggests that the source of the content significantly influences user evaluations of comprehensibility. A multiple linear regression analysis was performed to investigate the predictors of user comprehensibility evaluation. The model included the source, actual level, response, user selected level, and enjoyment as predictors. The overall model was significant, explaining a substantial proportion of the variance in user comprehensibility evaluation, R^2 = 0.488, F(9, 206) = 21.78, $p < .001$.The regression coefficients, standard errors, t-values, and p-values for the predictors are as follows: The intercept was significant, β = 2.9868, SE = 0.250, t(206) = 11.964, $p < .001$, indicating a baseline level of user comprehensibility evaluation.

The source of the content from Mistral did not significantly differ from the reference category of human written texts , β = -0.0137, SE = 0.122, t(206) = -0.112, p = .911. Similarly, content from OpenAI did not significantly differ from the reference category, β = 0.0396, SE = 0.120, t(206) = 0.329, p = .742.

The actual level of the content did not significantly predict user evaluations, β = -0.2080, SE = 0.162, t(206) = -1.284, p = .200, and β = -0.1188, SE = 0.182, t(206) = -0.652, p = .515, for B1-B2 and C1-C2, respectively. The "False Alarm" response significantly predicted lower user evaluations the overall distribution of response is shown in figure 19, β = -0.6551, SE = 0.220, t(206) = -2.981, p = .003. The "Hit" and "Miss" response did not significantly differ from the reference category of correct rejection, β = -0.0596, SE = 0.122, t(206) = -0.488, p = .626, and β = 0.0856, SE = 0.125, t(206) = 0.682, p = .496, respectively. User-selected level B1-B2 was a significant predictor The overall distribution of user selected level is shown in figure 1.9 β =

0.7047, SE = 0.184, t(206) = 3.823, p < .001, indicating higher user evaluations. Similarly, user-selected level C1-C2 was a significant predictor, β = 0.8552, SE = 0.208, t(206) = 4.117, p < .001, indicating higher user evaluations. Enjoyment was a significant predictor, β = 0.3023, SE = 0.043, t(206) = 7.046, p < .001, indicating that higher enjoyment levels were associated with higher user comprehensibility evaluations. The overall distribution of enjoyment is shown in figure 21

Figure 19

Participant's perceived text complexity and comprehensibility evaluation

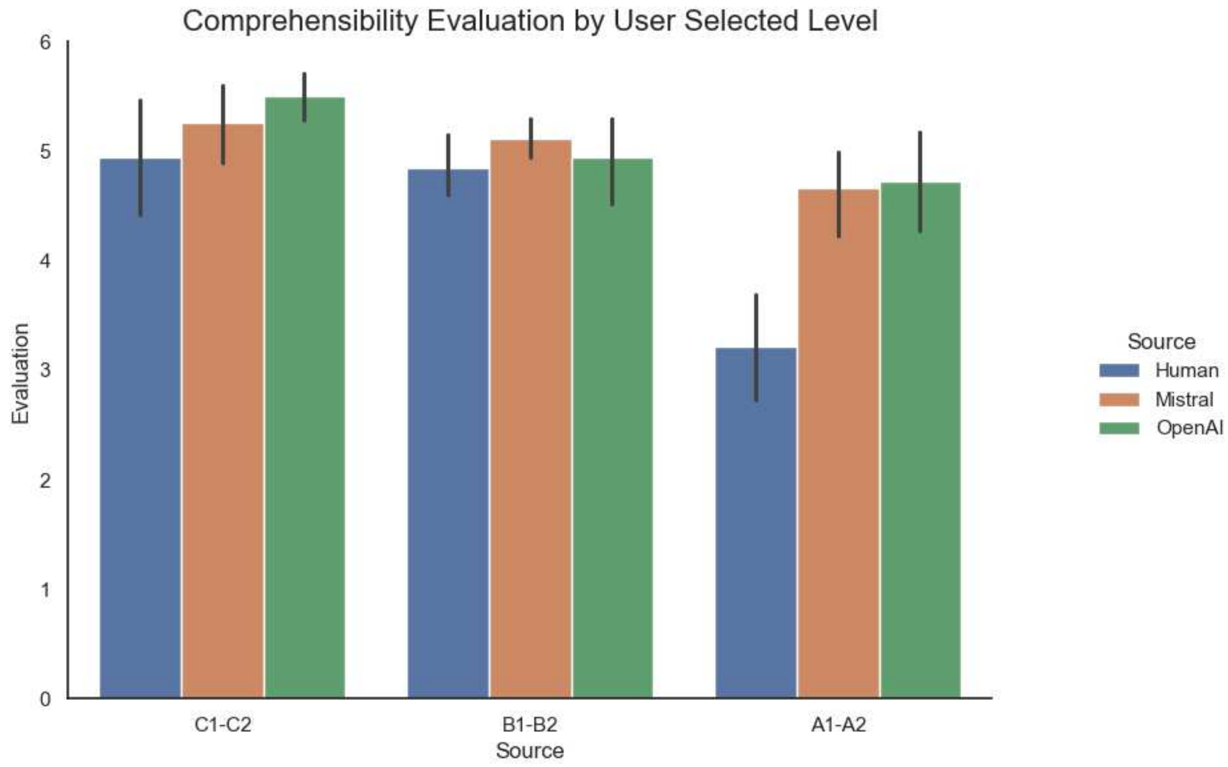


Figure 20

User comprehensibility evaluation distribution by response

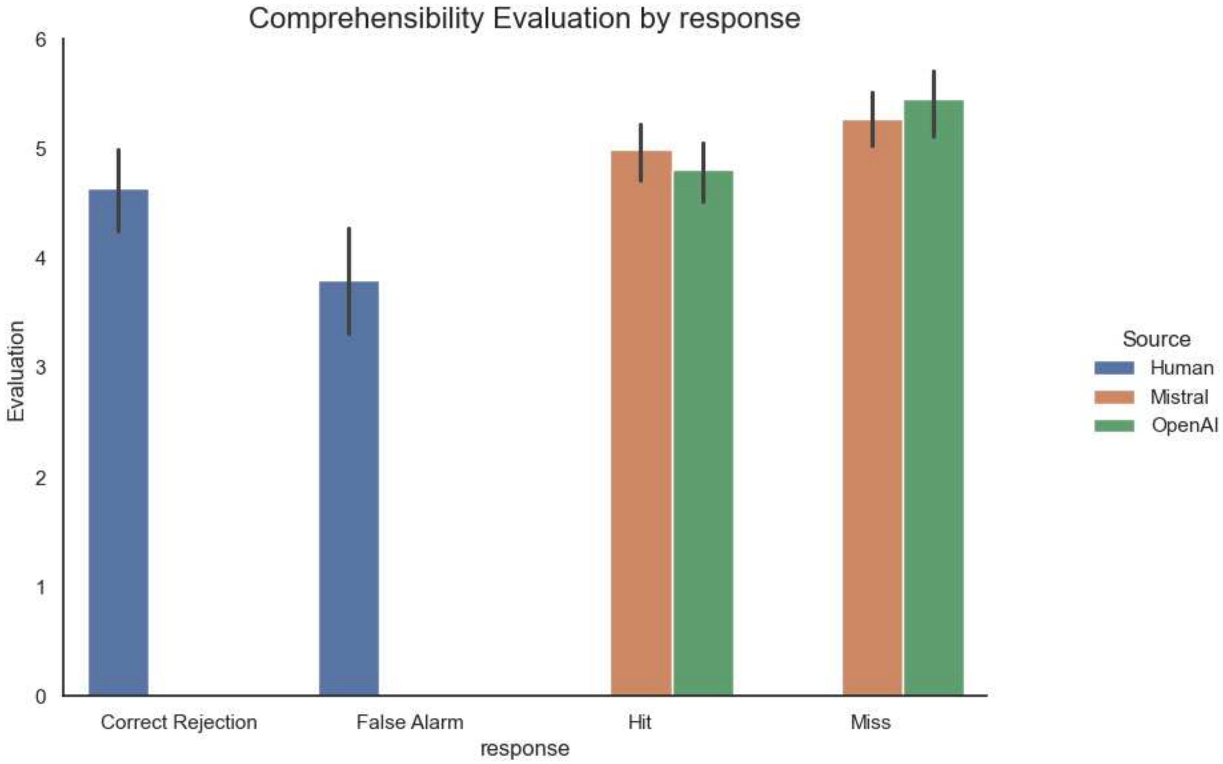
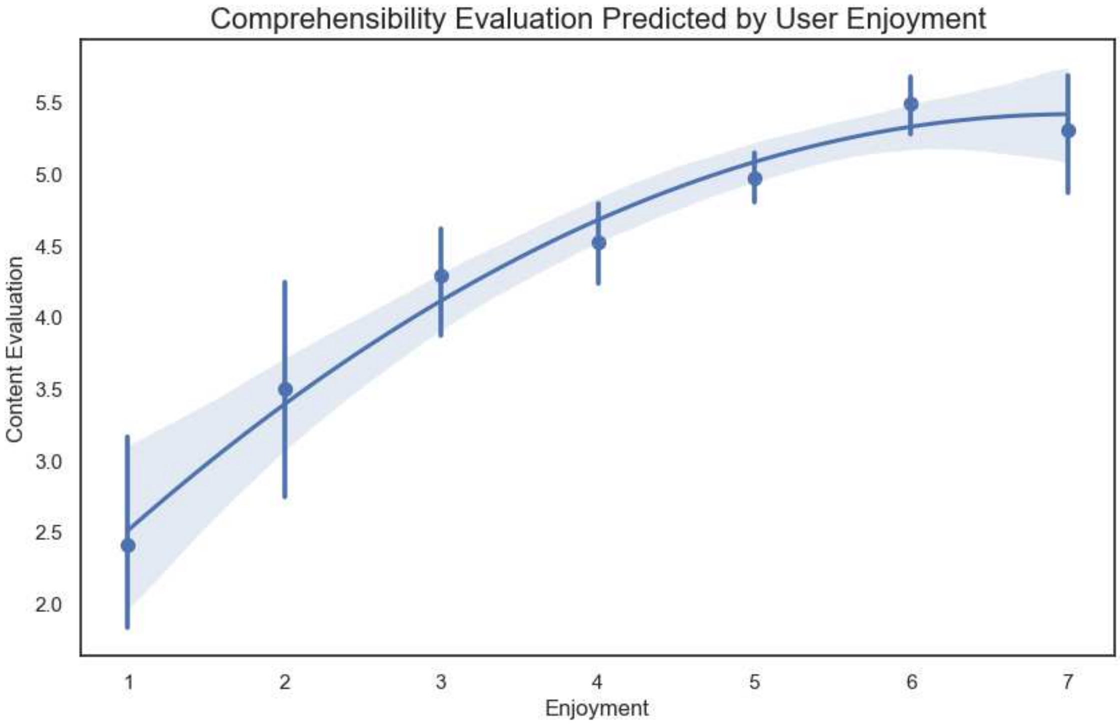


Figure 21

Participants reported enjoyment and Comprehensibility evaluation



3.5 Coherence and Cohesion Evaluation Analysis

An ANOVA test was conducted to examine the effect of the source on user coherence and cohesion evaluation. The results indicated a significant effect of the source, F(2, 213) = 20.09, $p < .001$. This suggests that the source of the content significantly influences user evaluations of coherence and cohesion.

A multiple linear regression analysis was performed to investigate the predictors of user coherence and cohesion evaluation. The model included the source, actual level, response, user-selected level, and enjoyment as predictors. The overall model was significant, explaining a substantial proportion of the variance in user coherence and cohesion evaluation, $R^2 = 0.600$, F(9, 206) = 34.34, $p < .001$.The regression coefficients, standard errors, t-values, and p-values for the predictors are as follows: The intercept was significant, $\beta = 1.6775$, SE = 0.242, t(206) = 6.918, $p < .001$, indicating a baseline level of user coherence and cohesion evaluation.

The source of the content from Mistral did not significantly differ from the reference category of human written texts, $\beta = 0.0607$, SE = 0.118, t(206) = 0.514, p = .608. Content from OpenAI significantly differed from the reference category, $\beta = 0.2414$, SE = 0.117, t(206) = 2.064, p = .040. The overall distribution source is shown in figure 22. The actual level of the content did not significantly predict user evaluations, $\beta = 0.0917$, SE = 0.157, t(206) = 0.583, p = .561, and $\beta = 0.1098$, SE = 0.177, t(206) = 0.620, p = .536, for B1-B2 and C1-C2, respectively.

The "False Alarm" response did not significantly predict user evaluations, $\beta = -0.2682$, SE = 0.213, t(206) = -1.257, p = .210. The "Hit" response did not significantly differ from the reference category of correct rejection, $\beta = -0.0466$, SE = 0.119, t(206) = -0.393, p = .695. The "Miss" response significantly predicted higher user evaluations, β = 0.3487, SE = 0.122, t(206) = 2.862, p = .005. The overall distribution of enjoyment is hown in figure 24

User-selected level B1-B2 was a significant predictor, β = 0.5945, SE = 0.179, t(206) = 3.320, p = .001, indicating higher user evaluations. Similarly, user-selected level C1-C2 was a significant predictor, β = 0.7560, SE = 0.202, t(206) = 3.747, p < .001, indicating higher user evaluations. The overall user selected level is shown in figure 23

Enjoyment was a significant predictor, β = 0.4107, SE = 0.042, t(206) = 9.855, p < .001, indicating that higher enjoyment levels were associated with higher user coherence and cohesion evaluations. The overall distribution of enjoyment is shown in figure 25

Figure 22

Coherence and Cohesion scores by the source of the text

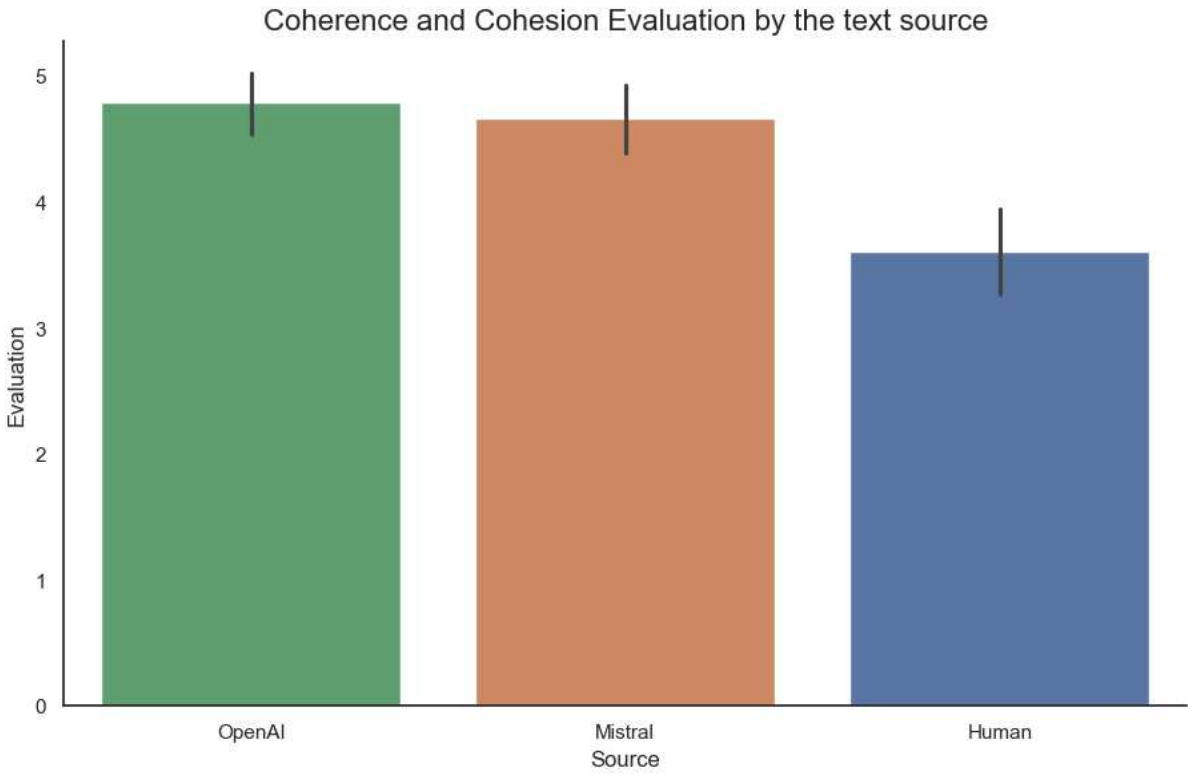


Figure 23

Participant's perceived text complexity and Coherence and Cohesion evaluation

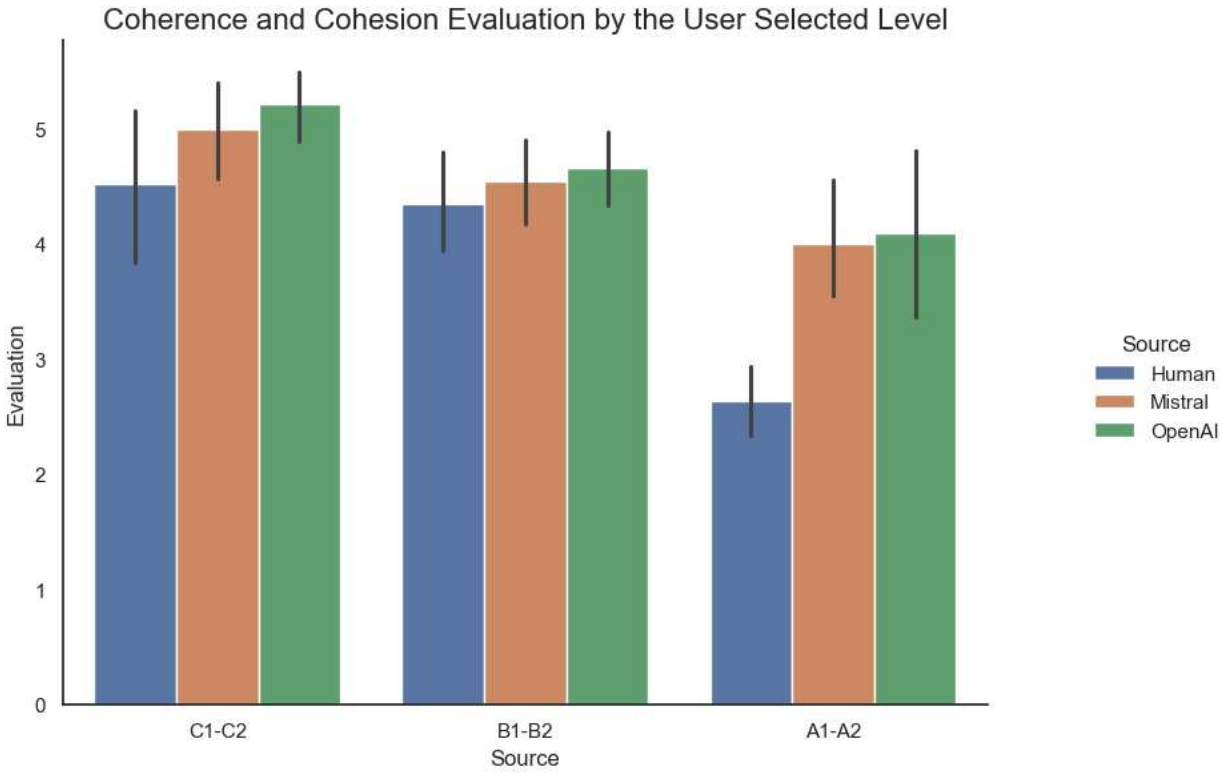


Figure 24

User coherence and cohesion evaluation distribution by response

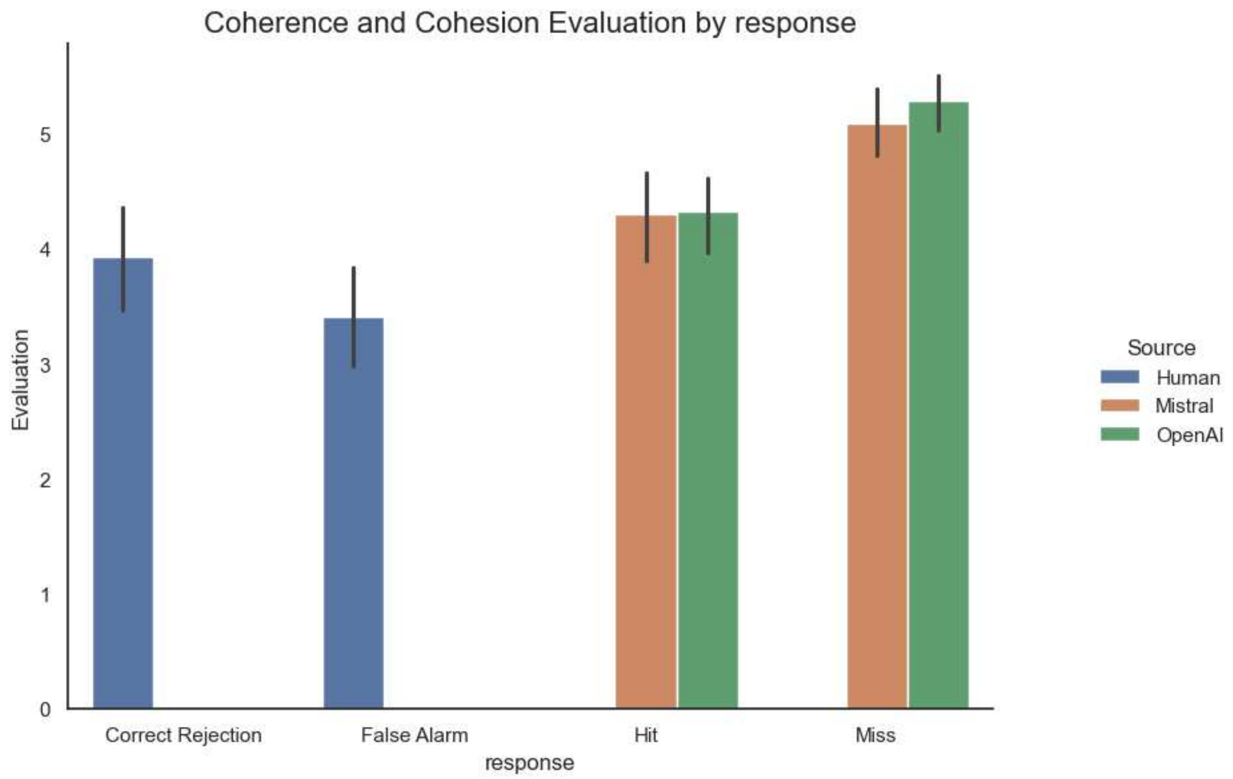
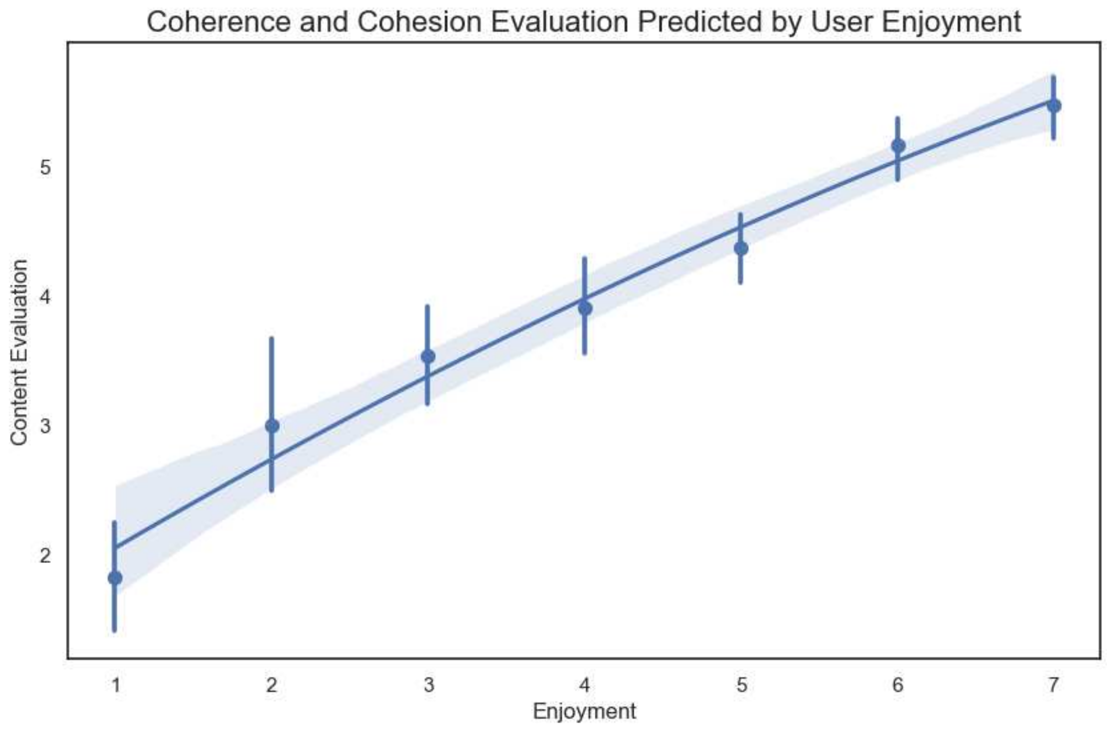


Figure 25

Participants reported enjoyment and Coherence and Cohesion evaluation



3.6 Level Accuracy of Texts

The degree to which generated texts align with their intended target proficiency levels was assessed by comparing participants' selections against the actual levels of the texts. The overall distribution of user indicated level is shown in figure 26

Human-generated texts:

- For A1-A2 level texts, participants identified 79.17% correctly.
- For B1-B2 level texts, participants identified 45.83% correctly.
- For C1-C2 level texts, participants identified 62.50% correctly.

OpenAI texts:

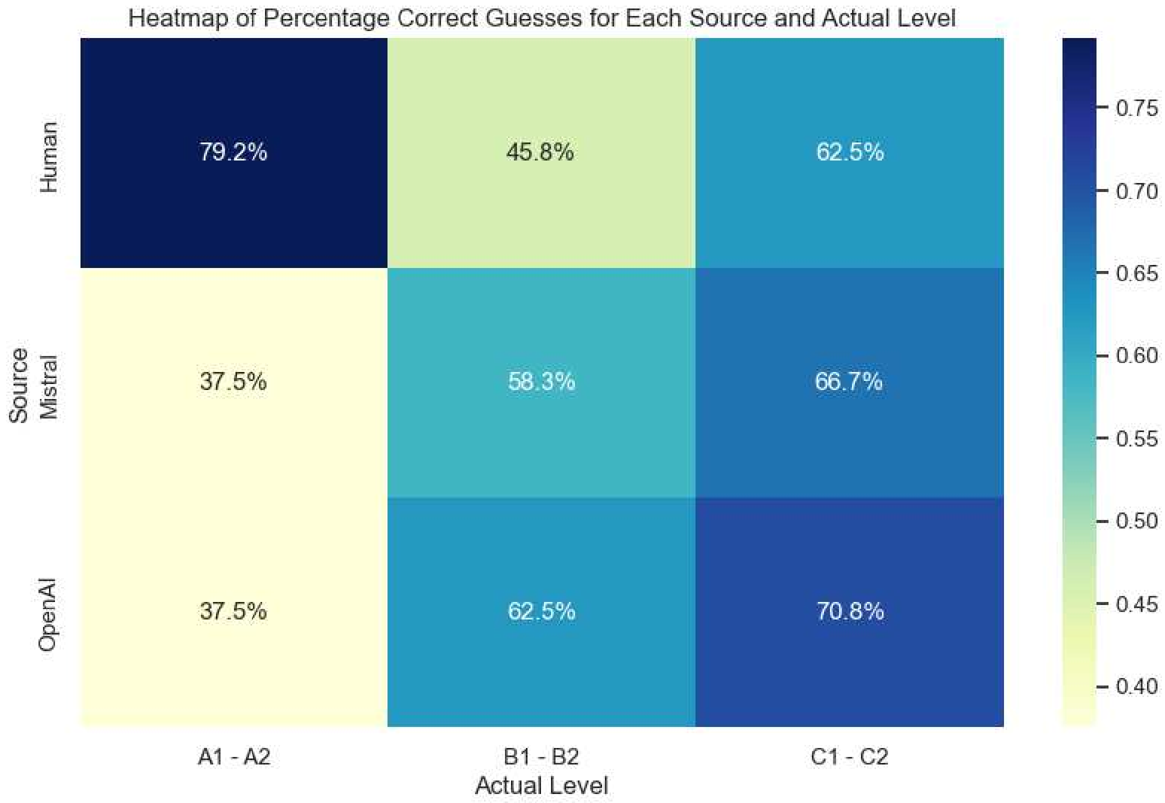
- For A1-A2 level texts, participants identified 37.50% correctly.
- For B1-B2 level texts, participants identified 62.50% correctly.
- For C1-C2 level texts, participants identified 70.83% correctly.

Mistral Texts:

- For A1-A2 level texts, participants identified 37.50% correctly.
- For B1-B2 level texts, participants identified 58.33% correctly.
- For C1-C2 level texts, participants identified 66.67% correctly.

Figure 26

Heatmap of correct guess distribution for each level



4 Discussion

This study aimed to assess the suitability of AI generated texts for language teaching across proficiency levels using the Functional Adequacy scale. By comparing AI generated texts from an open source (Mixtral) and closed source (GPT-3.5) language models to human written texts, this research provided new and exciting insights for the potential and limitations of potential use cases of large language models in language education.

4.1 Users’ Ability to Differentiate Between AI Generated and Human Written Texts

One of the key findings was that participants struggled to accurately differentiate between AI-generated and human-written texts, as indicated by the signal detection analysis.

The d' score of -.213 suggests that users often incorrectly identified human written texts as AI generated. This result highlights the advanced capabilities of modern language models in producing human-like text, making it challenging for users to distinguish between AI-generated and human-written content. The task requirements evaluation analysis revealed that the "False Alarm" response (where human-written texts were incorrectly identified as AI-generated) significantly predicted lower user evaluations. This finding implies that when participants mistakenly believed a human-written text was generated by AI, they tended to rate the text's ability to meet task requirements more negatively. This result may also indicate a bias among users, who might be more critical of texts they perceive as AI-generated, even when the texts are actually written by humans. In the comprehensibility evaluation analysis, the "False Alarm" response again significantly predicted lower user evaluations. This finding suggests that when participants incorrectly identified human-written texts as AI-generated, they tended to rate the comprehensibility of the text lower. This result further supports the notion of a potential user bias against texts perceived as AI-generated, even when the texts are actually written by humans. Finally, in the coherence and cohesion evaluation analysis, the "Miss" response significantly predicted higher user evaluations. This finding indicates that when participants mistakenly believed an AI-generated text was written by a human, they tended to rate the coherence and cohesion of the text more positively. Further justifying the possibility of a positive bias towards “human like” texts and how users evaluate them. These findings highlight the potential for user bias in evaluating texts based on their perceived source (human-written or AI-generated), rather than the actual source or the inherent qualities of the text.

Despite the potential for user bias, the overall functional adequacy of texts from the three sources (Human, Mistral, and OpenAI) was generally high. The ANOVA results indicated significant effects of source on user evaluations across all four dimensions of functional adequacy: content (F(2, 213) = 20.56, p < .001), task requirements (F(2, 213) = 15.56, p < .001), comprehensibility (F(2, 213) = 19.10, p < .001), and coherence and cohesion (F(2, 213) = 20.09, p < .001). However, the multiple linear regression analyses revealed that while OpenAI texts were rated significantly higher than human-written texts in terms of content (β = 0.2130, SE = 0.086, t(206) = 2.484, p = .014) and coherence and cohesion (β = 0.2414, SE = 0.117, t(206) = 2.064, p = .040), Mistral texts did not significantly differ from human-written texts across any of the four dimensions. This suggests that, despite users' potential biases, the AI-generated texts, particularly those produced by OpenAI, were generally considered functionally adequate and comparable to human-written texts.

4.2 Alignment of AI Generated Texts with Intended Target Proficiency Levels

The analysis of level accuracy revealed that participants were most accurate in identifying the proficiency levels of human-written texts, particularly at the A1-A2 level. In contrast, the accuracy rates for AI-generated texts were lower, with both models showing similar patterns. This finding suggests that while AI models can generate texts that are difficult to distinguish from human-written ones, they may struggle to consistently produce texts that align with specific CEFR levels. This pattern was expected as it can be seen on the Gulpease Index and Flesch Reading Ease distribution of the models. As this shows that Large Language Models often encounter difficulties in generating simplified educational texts that goes beyond these basic thematic areas. This challenge can be largely attributable to the lack of diverse training data on such topics, which unsurprisingly results in LLMs' limited proficiency in this area. The implications of this finding are significant for the use of AI-generated texts in language education. If AI models cannot reliably generate content that matches the intended proficiency level, it may lead to a mismatch between the learning materials and the learners' abilities. This could result in frustration, demotivation, or even impeded progress in language acquisition. To address this issue, future research should focus on developing AI models that are more adept at generating texts tailored to specific proficiency levels especially focusing on the beginner level text generation.

4.3 Differences Between Open-Source and Closed-Source Language Models

We also aimed to investigate potential differences between open source and closed language models in terms of functional adequacy and user evaluations. While some differences were observed, such as GPT-3.5 texts receiving higher ratings for content and coherence/cohesion, the overall patterns were largely similar between the two models. This finding suggests that open-source models like Mixtral can generate texts that are comparable to those produced by closed-source models in terms of functional adequacy. This is an encouraging result, as it shows that the benefits of AI generated texts in language education may be accessible to a wider range of users and institutions, regardless of their access to proprietary models. However, it is important to note that the current study only compared two specific models, and the results may not generalize to all open source and closed source language models.

4.4 The Role of User Perceptions and Engagement

Across all FA dimensions, user selected proficiency levels and enjoyment consistently emerged as significant predictors of higher ratings. This finding highlights the importance of user perceptions and engagement in evaluating the suitability of texts for language learning. It also suggests that AI generated texts can be effective in language education if they are perceived as enjoyable and appropriately challenging by learners. This result has important implications for the design and implementation of AI assisted language learning materials. Rather than focusing solely on the objective characteristics of the texts, such as linguistic complexity or alignment with CEFR levels, educators and material developers should also consider the subjective experiences of learners. By creating content that is engaging, enjoyable, and matched to learners perceived proficiency levels AI generated texts may be more effective in promoting language acquisition and learner motivation.

4.5 Limitations and Future Directions

While the current study provides valuable insights into the use of AI-generated texts in language education, it is important to acknowledge its limitations. First, the study was conducted with a relatively small sample size and focused on a single language. Future research should replicate and extend these findings with larger, more diverse samples and across different languages to assess the generalizability of the results. Second, the study relied on a single evaluation framework, the

Functional Adequacy scale. While this scale provides a comprehensive assessment of text quality, it may not capture all relevant aspects of language learning materials. A study done by Jakesch et al. (2023) showed that while evaluating AI generated texts humans tend to rely on faulty heuristics. With this in, mind future studies should consider incorporating additional evaluation metrics, such as linguistic complexity measures or learner performance outcomes, to provide a more holistic understanding of the effectiveness of AI generated texts in language education. Finally, the current study focused on the evaluation of AI-generated texts in isolation. In real-world language learning contexts, these texts would likely be used in conjunction with other instructional materials and activities. Future research should investigate the integration of AI-generated texts into complete language learning curricula and assess their impact on learner outcomes over extended periods.

4.6 Conclusion

In conclusion, this study provides valuable insights into the potential and limitations of using AI-generated texts in language education. While modern language models can generate human like texts that are engaging, they may struggle to consistently produce content tailored to specific proficiency levels. The results also highlight the importance of user perceptions and engagement in evaluating the suitability of texts for language learning. As AI technologies further develops, it will be essential to examine their educational applications and establish guidelines for their responsible and effective use. Future research should focus on enhancing the alignment of AI generated texts with specific proficiency levels, studying the long-term impacts of these texts in language learning, and integrating these tools into comprehensive language education curricula.

5 References

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Alotaibi, N., Altwairesh, N., Alowisheq, A., Bari, M. S., & Khan, H. (2024). *When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards* (arXiv:2402.01781). arXiv. http://arxiv.org/abs/2402.01781

Bartning, I., Martin, M., & Vedder, I. (2010). *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*. European Second Language Association.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

Firat, M. (2023). *How Chat GPT Can Transform Autodidactic Experiences and Open Education?* https://doi.org/10.31219/osf.io/9ge8m

Gan, W., Qi, Z., Wu, J., & Lin, J. C.-W. (2023). *Large Language Models in Education: Vision and Opportunities* (arXiv:2311.13160). arXiv. http://arxiv.org/abs/2311.13160

Hasnain, S., & Halder, S. (2024). Intricacies of the Multifaceted Triad-Complexity, Accuracy, and Fluency: A Review of Studies on Measures of Oral Production. *Journal of Education*, *204*(1), 145–158. https://doi.org/10.1177/00220574221101377

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (arXiv:2009.03300). arXiv. http://arxiv.org/abs/2009.03300

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), e2208839120. https://doi.org/10.1073/pnas.2208839120

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321–336. https://doi.org/10.1177/0265532216663991

Kuiken, F., & Vedder, I. (2022). Measurement of functional adequacy in different learning contexts: Rationale, key issues, and future perspectives. *TASK. Journal on Task-Based Language Teaching and Learning*, *2*(1), 8–32. https://doi.org/10.1075/task.00013.kui

Meta AI. (2024, April 18). *Introducing Meta Llama 3: The most capable openly available LLM to date*. Meta AI. https://ai.meta.com/blog/meta-llama-3/

Mistral. (2023, December 11). *Mixtral of experts*. https://mistral.ai/news/mixtral-of-experts/

OpenAI customer story: Duolingo. (n.d.). Retrieved April 23, 2024, from https://openai.com/customer-stories/duolingo

Pallotti, G. (2009). CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*(1), 15. https://doi.org/10.1186/s40561-023-00237-x

6 Appendices

6.1 Appendix A

Base prompt:

"[INST]Genera un testo in italiano focalizzato sul tema specificato: {Topic}. Questo testo deve adattarsi a un pubblico con livelli di comprensione da: {Level}, evitando un linguaggio troppo semplice o eccessivamente complesso. È essenziale che il testo mantenga una coerenza tematica, evitando digressioni non pertinenti al tema principale. La lunghezza del testo deve rispettare il numero di frasi indicate, con un minimo di {Minimum_sentence_count} frasi e un massimo di {Maximum_sentence_count} frasi, assicurando una distribuzione equilibrata delle informazioni e una conclusione logica. Si prega di non includere formule di saluto come 'Ciao' o 'Benvenuto', interiezioni come 'Ovviamente' o 'Naturalmente', o qualsiasi elemento che possa distogliere l'attenzione dal contenuto informativo principale. L'obiettivo è fornire un testo che sia informativo, coinvolgente e direttamente correlato al tema scelto, promuovendo una lettura che sia sia istruttiva che piacevole per il livello di comprensione specificato.[/INST]",

Example prompt:

[INST]Genera un testo in italiano focalizzato sul tema specificato: Amicizia internazionale. Questo testo deve adattarsi a un pubblico con livelli di comprensione da: A1 - A2, evitando un linguaggio troppo semplice o eccessivamente complesso. È essenziale che il testo mantenga una coerenza tematica, evitando digressioni non pertinenti al tema principale. La lunghezza del testo deve rispettare il numero di frasi indicate, con un minimo di 10 frasi e un massimo di 13 frasi, assicurando una distribuzione equilibrata delle informazioni e una conclusione logica. Si prega di non includere formule di saluto come 'Ciao' o 'Benvenuto', interiezioni come 'Ovviamente' o 'Naturalmente', o qualsiasi elemento che possa distogliere l'attenzione dal contenuto informativo principale. L'obiettivo è fornire un testo che sia informativo, coinvolgente e direttamente correlato al tema scelto, promuovendo una lettura che sia sia istruttiva che piacevole per il livello di comprensione specificato.[/INST]

6.2 Appendix B

5. Il testo contiene informazioni sufficienti e pertinenti?

- Il testo contiene pochissime idee, e queste non sono correlate tra loro.
- Il testo presenta solo alcune idee, e non sono molto coerenti.
- Il testo ha una discreta quantità di idee, ma non sono sempre coerenti.
- Il testo ha una buona quantità di idee, e sono abbastanza coerenti.

6. I requisiti del compito sono stati soddisfatti con successo (ad esempio, genere, atti di parlato, registro)?

- Nessuna delle domande e dei requisiti del compito è stata soddisfatta.
- Alcune (meno della metà) delle domande e dei requisiti del compito sono state soddisfatte.
- Circa la metà delle domande e dei requisiti del compito sono state soddisfatte.
- La maggior parte (più della metà) delle domande e dei requisiti del compito sono state soddisfatte
- Quasi tutte le domande e i requisiti del compito sono stati soddisfatti.
- Tutte le domande e i requisiti del compito sono stati soddisfatti.

7. Quanto è facile comprendere lo scopo e le idee del testo?

- Il testo è completamente incomprensibile. Le sue idee e il suo scopo non sono comprensibili e cercare di capirlo è inutile.
- Il testo è a malapena comprensibile. Lo scopo non è chiaro e il lettore deve indovinare la maggior parte delle idee
- Il testo è abbastanza comprensibile. Alcune parti sono difficili da capire al primo tentativo, ma una seconda lettura aiuta a chiarire le cose, anche se rimangono alcuni dubbi.
- Il testo è comprensibile. Alcune parti potrebbero essere poco chiare, ma possono essere comprese dopo una seconda lettura senza troppi sforzi.
- Il testo è facile da comprendere e scorre bene. Non ci sono problemi di comprensibilità.
- Il testo è molto facile da comprendere e molto coinvolgente. Le idee e lo scopo sono espressi chiaramente

8. Quanto bene il testo rimane unito e ha senso nel suo insieme (usando cose come parole di collegamento e strategie)?

- Il testo non ha senso affatto. Salta molto da un argomento all'altro senza un chiaro collegamento tra le idee. Non viene utilizzato alcun tipo di parole di collegamento
- Il testo ha poco senso. Spesso salta su argomenti non correlati, a volte utilizzando la ripetizione per collegare le idee. Vengono utilizzate pochissime parole di collegamento e le idee non si collegano bene.
- Il testo ha un certo senso, ma ci sono frequenti argomenti non correlati o ripetizioni. Utilizza alcune parole di collegamento di base, ma le idee non sono sempre collegate in modo fluido

- Il testo è per lo più coerente. Gli argomenti non correlati sono rari, ma c'è una certa dipendenza dalla ripetizione. Utilizza una buona quantità di parole di collegamento, inclusi più che semplici congiunzioni.

- Il testo è molto coerente. I nuovi argomenti vengono introdotti in modo fluido con parole o frasi di collegamento, e la ripetizione è rara. Utilizza una varietà di parole di collegamento in modo efficace, facendo sì che le idee si colleghino bene.

- Il testo è estremamente coerente e coeso. Le nuove idee vengono integrate senza problemi con una varietà di parole e frasi di collegamento. Non ci sono salti di argomenti o ripetizioni, e il testo fluisce molto agevolmente.