Exploring Paraphrasing Strategies for CEFR A1-Level Constraints in LLMs

Eugenio Marzona, Maria Goikhman, Alessio Palmero Aprosio, Massimo Zancanaro
Dept. of Psychology and Cognitive Science, University of Trento, Rovereto, Italy
[eugenio.marzona,mariia.goikhman,a.palmeroaprosio,massimo.zancanaro]@unitn.it

Abstract

Large language models are increasingly used for teaching and self-learning foreign languages. However, their capability to meet specific linguistic constraints is still underexplored. This study compares the effectiveness of prompt engineering in guiding ChatGPT (4o and 4o-mini), and Llama 3 to rephrase general-domain texts to meet CEFR A1-level constraints in English and Italian, making them suitable for beginner learners. It compares 4 prompt engineering approaches, built upon iterative paraphrasing method that gradually refines original texts for CEFR compliance. The approaches compared include paraphrasing with or without Chain-of-Thought, as well as grammar and vocabulary simplification performed either simultaneously or as separate steps. The findings suggest that for English the best approach is combining COT with separate grammar and vocabulary simplification, while for Italian one-step strategies have better effect on grammar, and two-step strategies work better for covering the vocabulary. The paraphrasing approach can approve compliance, although at this point it is not cost-effective. We release a dataset containing original sentences paired with their beginner-level paraphrases(both in Italian and in English) on which further work could be based.

1 Introduction

Large language models can generate vast amounts of content for language teaching and learning, offering new possibilities for personalized and scalable materials. However, LLMs exhibit difficulties in consistently adhering to strict constraints specified via prompting. (Liu et al., 2024) While occasional exposure of foreign language learners to unfamiliar elements is necessary for language acquisition (Krashen, 1985), the texts that remain fully within given grammatical and lexical boundaries can be valuable in instructional settings—for instance, when designing controlled practice activities focused on a specific target structure or vocabulary set.

Language proficiency levels are commonly assessed using the CEFR, a comprehensive framework developed by the Council of Europe. CEFR (Common European Framework of Reference for Languages) categorizes language ability into six levels: A1 and A2 (Basic User), B1 and B2 (Independent User), and C1 and C2 (Proficient User). Its descriptors focus on communication tasks that language learners can complete at each level. (Council of Europe, 2020)

Language generation under CEFR constraints is notoriously a difficult task for LLMs (see section 2). In this study we explore the combination of two prompting strategies in a task of text simplification. Addressing text simplification rather than text generation allowed us to start from existing texts and have a better control of the task. Furthermore, it might be a useful approach to use for adaptive tutoring systems for foreign language learning (Guizani et al., 2025).

This study focuses on the task of rephrasing existing texts to make them suitable for A1 CEFR level reading activities. Beginner-level adult learners require texts that are tailored both to their level and to their interests (Lee and Pulido, 2016). While authentic texts are appropriate at higher levels, offering their paraphrased versions at A1 is a trade-off between integrating authentic materials, which increases students' engagement and motivation, (Albiladi, 2018) and making them accessible for beginners.

A1 is the most restricted level in terms of both grammar and vocabulary (Council of Europe, 2020), which makes it a particularly demanding target, as the criteria go beyond general readability. Careful control of grammar structures, such as limiting tense usage and avoiding irregular verbs (Spinelli and Parizzi, 2010) is not normally re-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15305–15318
November 4-9, 2025

quired in simplified text for native speakers. This makes A1-level paraphrasing a lower-bound evaluations which allows to see how capable are LLMs of paraphrasing texts using an extremely constrained language inventory.

Specifically, the research question addressed is the following: to what extent can large language models be prompted to rephrase input texts in accordance with CEFR A1-level grammar and vocabulary constraints as specified in inventories for English and Italian?

We address two languages, English and Italian, and three LLMs in order to provide a tentative test for generalization of our results.

Our findings suggest that guided paraphrasing can enhance compliance with A1 standards, although imperfectly. Due to the complexity of A1 inventories, this improvement comes at a considerable cost in terms of processed tokens. For English, one specific combination of strategies appears to be significantly more effective. In contrast, the results for Italian are less consistent, and no single best strategy clearly emerged.

2 Related Works

For CEFR-constrained text generation and simplification, researchers mainly use prompt engineering and model fine-tuning, often together. Malik et al. (2024) combined these approaches by using few-shot prompting, supervised fine-tuning, and reinforcement learning to develop CALM (CEFR-Aligned Language Model), a fine-tuned version of LLaMa2-7b, which they report outperforms GPT-4 for CEFR-aligned content generation. Similarly, Glandorf and Meurers (2024) utilized both few-shot learning and fine-tuning to control grammatical complexity in educational text generation, finding that fine-tuned Mistral yielded the best results among tested models. Other studies have relied solely on prompt engineering. Bektaş et al. (2024) demonstrated that ChatGPT-4o could generate simplified Turkish texts comparable to human-simplified versions using zero-shot prompting with detailed rules. Imperial and Madabushi (2023) found that open-source models like BLOOMZ and FlanT5 were more effective than ChatGPT for readability-controlled text generation, though the latter improved with refined prompts.

The research is mostly focused on English, but Bektaş et al. (2024) worked with Turkish and Spring and Rios (2021) explored text-simplification for German. Research on Italian materials remains more limited, though Franzoni et al. (2024) developed morpho-phraseological model for classifying the CEFR level of Italian L2 learner writing.

A critical question is to what extent LLMs can apply CEFR-based constraints during text generation. Benedetto et al. (2025) conducted a comprehensive assessment of various models' CEFR actionable knowledge—a term they introduce to describe the models' ability to operationalize CEFR categories—across text classification, essay scoring, and student simulation tasks. Their findings reveal that while LLMs demonstrate partial CEFR knowledge, they often struggle to apply it accurately in practical tasks. Among the models evaluated (including GPT-3.5, GPT-4o-mini, Mistral, Gemma, and LLaMA 3), GPT-4o-mini performed best overall, though all models showed inconsistencies. Ramadhani et al. (2023) evaluated ChatGPT-generated reading texts from a systemic functional perspective, finding that text complexity often failed to align with claimed CEFR levels. Their analysis suggested that ChatGPT's output complexity correlated more strongly with text length than with appropriate lexical or grammatical calibration for specific CEFR levels. Similarly, Uchida (2025) specifically examined ChatGPT-4o's ability to generate texts at specified CEFR levels, discovering low accuracy—only 5% of generated texts matched their target CEFR level according to computerized analysis. This study highlighted a tendency to oversimplify A1 and A2 level content, suggesting challenges in maintaining appropriate complexity for lower proficiency levels.

3 Instruments

The two instruments used in the work are the CEFR-based inventories for reading/writing competences (that is, the grammar rules that should be mastered at A1 proficiency level), and the vocabulary lists (that is, the words that should be known at A1 level), for English and for Italian.

3.1 CEFR Inventories

Benedetto et al. (2025) explored prompting strategies that referenced CEFR levels either by only naming them or through brief descriptions of the communicative abilities expected at each level. Their findings suggest that such high-level prompts may be insufficient to guide models toward consistent level-appropriate output. To address this issue,

we turned to detailed, language-specific CEFR inventories in order to provide models with more concrete grammatical and lexical constraints for English and Italian. For English, we rely on the Council of Europe Breakthrough Inventory (Trim, 2009). For Italian, two inventories were considered: the comprehensive *Profilo della lingua italiana* (Spinelli and Parizzi, 2010) and the guidelines for the CILS exam (Barni et al., 2009). The latter was chosen for its compactness and practical utility. The original constraints were categorized by part-of-speech, specifying both the permitted and prohibited forms for each category. We made a slight adjustment to the Italian constraints, since there was a discrepancy in irregular verbs usage between the Barni et al. (2009) guidelines and their sample exam version (Università per Stranieri di Siena, 2017). To address this, we included irregular verbs both from guidelines and from exam sample in the final constraints list. The resulting lists of grammatical constraints for English and Italian are available in Appendix A.

3.2 Vocabulary lists

The grammatical inventories were supplemented with lexical resources to define vocabulary constraints for A1-level output. For Italian, since Barni et al. (2009) do not include a vocabulary list, we adopted the word-lists proposed by Spinelli and Parizzi (2010), which categorize vocabulary by CEFR level.

For English, although Trim (2009) includes a sample list for the Breakthrough level, the author explicitly notes that this list is not prescriptive, given the heterogeneity of learners. We therefore opted to use the Oxford 3000 and 5000 (Oxford University Press, 2025), a widely recognized CEFR-aligned lexical resource.

Word-lists from Oxford University Press (2025) and Spinelli and Parizzi (2010) were organized by part of speech and CEFR level (A1–C1) into structured JSON files. Additionally, Oxford lists for British and American English were merged into a single list to avoid inconsistencies arising from spelling variants of the same word. Since lower-level vocabulary is a subset of higher-level vocabulary, we retained at each level only the words that were newly introduced relative to the preceding level.

CEFR (Council of Europe, 2020) does not indicate how many words should be covered at each language level, referencing vocabulary descriptively.

Part of Speech	A1 (IT)	A2 (IT)	A1 (EN)	A2 (EN)
Adjectives	56	147	142	313
Adverbs	49	71	90	167
Nouns	287	648	516	1054
Verbs	76	182	173	366
Total	468	1048	921	1900

Table 1: Cumulative word counts at CEFR A1 and A2 levels by part of speech in Italian and English

For this study, the A1-level word-list for Italian contains 468 entries, while the corresponding English list includes 921 words—nearly twice as many. Table 1 presents the word counts for each level and part-of-speech category.

4 Procedure

The task consisted of requesting different LLMs to paraphrase sentences with four different prompting strategies. The paraphrases have then been controlled against the inventories and the vocabulary lists in an automatic way. For a small subset, a manual check on linguistic soundness and meaning preservation have also been performed.

4.1 Data Collection

Texts for paraphrasing were collected from Vikidia[1], an open-access crowd-sourced online encyclopedia for children.

This source was chosen based on the assumption that articles written for a young audience would exhibit relatively simple language, suitable for adaptation to A1 level. While Simple English Wikipedia is available for English, no equivalent exists for Italian, and Vikidia offered a consistent multilingual alternative with accessible content.

Articles were sorted by their edit history, and those with more than 50 revisions were considered to be of sufficient quality. These articles were split into paragraphs, discarding paragraphs of less than 240 characters (roughly 50 words), so that each paragraph included in the dataset provided sufficient context.

This process resulted in two datasets: 737 original paragraphs for Italian and 1398 paragraphs for English.

4.2 Text Paraphrasing

To enforce linguistic constraints during paraphrasing, we conceived a general strategy for building a prompt grounded in CEFR-based grammat-

[1] https://www.vikidia.org/

```
# Task:
Check if the given text complies with the constraints provided;
generate a paraphrase when necessary.

# Original text:
{input_text}

# Constraints checking:
Check each sentence againts ALL the constraints.
- If it violates no constraint, keep it as is.
- If it violates one or more constraints, paraphrase it or remove it.

# Paraphrasing:
- A paraphrase must preserve the original semantic meaning and minimize information loss.
- A paraphrase must replace each non-constraints conformant element with an equivalent conformant
  alternative.
- If a paraphrase that preserves the original meaning and completely conforms to the given
  constraints cannot be formulated, then the non conformant text should be removed.

# Output format:
Format your response as shown below, no additional comment is needed:

<text>[Your final version of the text here - either the original if no changes were needed, or
your paraphrased version if changes were made]</text>

Note: The final version MUST be enclosed in <text> tags, regardless of whether changes were made
to the original text.

# Constraints:
- Adverbs, Prepositions, Conjunctions, and Interjections: These may be used without limitations.
- Nouns: singular, regular plural and possessive forms are allowed. Irregular plural noun forms
  are forbidden.
- Pronouns: personal, possessive, interrogative (who/what/which), demonstrative
- Adjectives: base form, regular comparative/superlative, interrogative, possessive
- Verbs: modals 'can'/'will' only; finite forms: present/past simple, present perfect,
  present/past continuous, imperative; non-finite: all allowed; voice: active
  (passive only for present/past simple)
```

Figure 1: Complete prompt template used for constraint-based paraphrasing.

ical inventories. First, the model receives a comprehensive list of allowed grammatical structures, adapted from CEFR inventories for a given language. Then, it is instructed, in each step, to check whether the current version of the text, starting from the original version, conforms to all specified constraints and to replace any non-compliant elements with alternatives that are as semantically close as possible. The paraphrasing process repeats iteratively, with each new iteration using the previous output as input, until either the output stabilizes (no significant changes between consecutive paraphrases) or a maximum of 10 iterations is reached.

While all grammatical constraints could be explicitly encoded and applied during the iterative paraphrasing process, handling vocabulary constraints required a different approach. Incorporating a complete lexicon of A1-level vocabulary and verifying lexical compliance at every step proved impractical due to the scale and complexity of such checks. Instead, we adopted a prompt engineering strategy (see Appendix B) that instructed the model to simplify lexical choices to the A1 level wherever possible.

This prompt-based approach instructed models to replace complex or uncommon words with simpler synonyms, and to add brief in-text explanations when simplification would compromise meaning.

We compared four prompting strategies for the rephrasing task, varying along two dimensions: whether paraphrasing and vocabulary simplification were performed jointly (1-Step and 1-Step COT) or in separate steps (2-Step and 2-Step COT), and whether chain-of-thought (COT) instructions were included (1-Step COT and 2-Step COT) or not (1-Step and 2-Step).

The inclusion of COT was motivated by prior findings suggesting that explicit reasoning steps

can improve output quality. (Wei et al., 2022) However, since COT also substantially increases token usage, we aimed to assess whether the potential benefits justified the additional computational cost.

The decision to test both combined and separate approaches to vocabulary simplification was driven by concerns about semantic drift. Simplifying vocabulary within the paraphrasing loop allows lexical choices to be reviewed and adjusted at each step—potentially leading to more thorough constraint compliance than in a single-pass approach. However, this also increases the risk that a word may be replaced multiple times across iterations, potentially altering the meaning of the original input.

4.3 Automatic Control System

To assess constraint compliance automatically, we developed a two-part evaluation system combining POS-based lexical analysis and language model-based grammar checking. For lexical analysis, we first removed stopwords using lists from NLTK (version 3.9.1),[2] then tagged each remaining word with part-of-speech labels using Stanford Stanza (version 1.10.1, Qi et al. (2020)). Words are then grouped by part-of-speech and compared against CEFR-aligned lexical lists. This process yields a percentage score indicating the proportion of the text's vocabulary covered by each CEFR level.

For grammatical analysis, we use GPT-4o with part-of-speech (POS) specific prompts, modeled on our CEFR-aligned constraint lists. These prompts include detailed linguistic instructions for annotating morphological features, such as English pluralization rules or Italian participle usage patterns, along with JSON schemas defining the desired output format.

The outputted JSON is parsed using a rule-based approach modeled on our linguistic constraint lists (e.g., if the inventory specifies "only regular English adjectives", each adjective that is tagged as irregular and not used in positive form is counted as an error in A1 grammar compliance).

The final grammatical analysis report counts constraint violations by POS category and provides descriptive error messages explaining each detected violation.

While human annotators could in principle verify whether a paraphrased text adheres to A1-level grammar restrictions, the length and specificity of our constraint list makes this task particularly demanding. Manual evaluation under such conditions is likely to result in annotation fatigue and reduced inter-annotator agreement. Prior work has shown that a high number of annotation categories or tags can negatively impact consistency among annotators: Bayerl and Paul (2011) report a significant negative correlation between category count and agreement, while Fort et al. (2012) and Williams et al. (2019) similarly highlight the role of annotation set complexity in increasing task difficulty and lowering reliability.

For both English and Italian, 100 original paragraphs were randomly selected from Vikidia. Each paragraph was rephrased using the four prompting strategies described in Section 4.2, resulting in 400 paraphrases per language. All outputs were then evaluated using our automated systems for lexical and grammatical constraint checking.

4.4 Manual control for meaning preservation and linguistic soundness

Following common practice in related work (Chi et al., 2023; Stowe et al., 2022), we performed a manual evaluation using a five-point Likert scale across two dimensions: meaning preservation and linguistic soundness. A total of 120 texts were assessed—15 for each of the four prompting strategies, in both English and Italian—by three independent annotators. To ensure consistency, we created detailed annotation guidelines specifying the criteria for each score level, along with examples illustrating typical cases for each rating (Appendix C).

Difference between original and paraphrased texts was counted using Fuzzy string matching algorithm built upon Levenstein distance (Cohen, 2025).[3]

To evaluate the reliability of the annotations, we computed inter-annotator agreement using Krippendorff's Alpha (Hayes and Krippendorff, 2007; Krippendorff, 2018), which is particularly well-suited for ordinal data.

Initial results yielded a score of 0.51 for meaning preservation and 0.22 for linguistic soundness – values too low to support any definitive conclusions. To investigate the cause of these low scores, we analyzed the results by language. When considering only the Italian data, agreement rose to 0.68 for meaning preservation and 0.41 for linguistic

[2] https://www.nltk.org/

[3] https://github.com/seatgeek/thefuzz

soundness. The first result suggests that this type of annotation is particularly challenging when annotators are not native speakers (as in the case of English), while the second highlights the subjective nature of linguistic soundness, which can vary significantly across annotators.

In the released dataset, we provide both the individual annotations and an aggregated score calculated as follows: if two annotators assigned the same score, that value is retained; otherwise, the average of the three scores is computed and rounded to the nearest integer.

5 Results

Starting from our Vikidia sentences dataset, 100 samples within a length range of [240-290] characters were randomly selected for English (mean length = 263.72 chars, SD = 14.796) and Italian (mean length = 261.79 chars, SD = 14.682).

This sampling approach ensured comparable test datasets across both languages.

LLM-driven paraphrasing was performed using three different models: GPT-4o (gpt-4o-2024-11-20), GPT-4o-mini (gpt-4o-mini-2024-07-18), and Llama-3.3-70b. All four paraphrasing strategies were used for each model.

Vocabulary coverage checks and automated grammar analysis were performed on the entire data sample, while manual annotation was performed only on a sub-set of 30 sentences (15 per language). For statistical analysis, we focused on three automated metrics available for all samples: linguistic difference, A1 vocabulary coverage, and constraint violation counts.

For both LLM-driven paraphrasing and automated grammar analysis, we opted for parameter settings of $temperature = 0.0$ and $top\_p = 1.0$. We found these settings appropriate for consistent results in linguistic analysis tasks.

For further statistical analysis and tests, we decided to partition the collected data based on input language. While the automated analysis process is identical between our two languages, the A1 CEFR aligned linguistic constraints selected (see Appendix A) differ significantly and do not overlap across parts-of-speech.

To evaluate for significant differences between paraphrasing strategies and model choices, the Kruskal-Wallis test was performed on the metrics of linguistic difference (Diff.), A1 vocabulary coverage (A1 Coverage %), and the count of violated constraints (Errors). This statistical test was found appropriate for our key metrics, as none of them present a normal distribution or homogeneous variance across groups.

Post-hoc pairwise comparisons were conducted using Dunn's test, with effect sizes measured using rank biserial correlation and classified as small (*), medium (**), or large (***) based on standard conventions for interpreting non-parametric effect sizes.

Tables 2 and 3 present the comprehensive results of our analysis.

Table 2 shows the impact of different paraphrasing strategies across both languages. For English, our analysis reveals significant differences between strategies in linguistic difference ($h = 151.67$, $p < 0.001$) and A1 vocabulary coverage ($h = 136.58$, $p < 0.001$). The 1-Step and 1-Step COT strategies produce sentences that are more similar to the original texts, while the 2-Step strategies seem to achieve better A1 vocabulary coverage. For Italian, both metrics showed significant differences ($p < 0.001$). Effect sizes are however less pronounced than in English, and 2-Step strategies seem to achieve marginally better vocabulary coverage.

Table 3 compares the performance across the three LLM models. Both GPT-4o models produced outputs with the lowest linguistic difference from the original text in both languages, with this effect more pronounced in Italian. Llama-3.3-70b achieves marginally better results on A1 vocabulary coverage for both languages.

Error counts showed statistically significant differences between strategies for both languages ($h = 22.18$, $p < 0.001$ for English; $h = 58.70$, $p < 0.001$ for Italian), with more pronounced effect sizes for Italian, but only showed marginal significance between models for English ($h = 6.35$, $p = 0.042$) and no significant differences between models for Italian ($h = 4.35$, $p = 0.114$).

5.1 Automatic Control Results

To assess the effectiveness of our prompting strategies in improving grammatical compliance, we compared the number of constraint violations detected in original versus paraphrased texts for both English and Italian. In English, the average number of violations per text decreased slightly from 1.9 (SD = 1.61, median = 2.00) in the original paragraphs to 1.8 (SD = 1.51, median = 1.00) in the paraphrases. However, the difference was not statistically significant ($p = 0.126$), yielding a modest

Table 2: Mean and standard deviations for the combined strategies for English and Italian

	1-Step COT	1-Step	2-Step COT	2-Step	Kruskal-Wallis	Dunn post-hocs
Strategies (English)						
Diff.	85.05 (14.660)	79.77 (10.967)	75.42 (10.335)	74.48 (10.412)	h = 151.67 p < 0.001	1-Step COT/1-Step** 1-Step COT/2-Step COT** 1-Step COT/2-Step** 1-Step/2-Step COT* 1-Step/2-Step*
A1 Coverage (%)	49.63 (17.534)	53.89 (15.719)	62.91 (15.225)	63.01 (15.218)	h = 136.58 p < 0.001	1-Step COT/1-Step* 1-Step COT/2-Step COT** 1-Step COT/2-Step** 1-Step/2-Step COT** 1-Step/2-Step**
Errors	1.5 (1.40)	1.6 (1.40)	2.0 (1.58)	2.0 (1.60)	h = 22.18 p < 0.001	1-Step COT/2-Step COT* 1-Step COT/2-Step* 1-Step/2-Step COT* 1-Step/2-Step*
Strategies (Italian)						
Diff.	73.90 (19.590)	78.48 (12.505)	69.58 (11.291)	72.69 (9.803)	h = 77.33 p < 0.001	1-Step COT/2-Step COT* 1-Step COT/2-Step* 1-Step/2-Step COT** 1-Step/2-Step** 2-Step COT/2-Step*
A1 Coverage (%)	26.46 (18.585)	22.87 (14.226)	31.12 (15.551)	30.10 (15.230)	h = 56.77 p < 0.001	1-Step COT/2-Step COT* 1-Step COT/2-Step* 1-Step/2-Step COT** 1-Step/2-Step*
Errors	1.5 (1.61)	2.5 (2.18)	1.9 (1.77)	2.7 (2.34)	h = 58.70 p < 0.001	1-Step COT/1-Step* 1-Step COT/2-Step COT* 1-Step COT/2-Step** 1-Step/2-Step COT* 2-Step COT/2-Step*

Note: * indicates small effect size, ** indicates medium effect size, *** indicates large effect size.

Table 3: Mean and standard deviations for models for English and Italian

	gpt4o	gpt4o-mini	Llama-3.3-70b	Kruskal-Wallis	Dunn post-hocs
Models (English)					
Diff.	81.90 (11.309)	77.43 (11.209)	76.71 (13.991)	h = 40.75 p < 0.001	gpt4o/gpt4o-mini* gpt4o/Llama-3.3-70b*
A1 Coverage (%)	55.16 (16.719)	57.28 (16.391)	59.64 (17.484)	h = 15.26 p < 0.001	gpt4o/Llama-3.3-70b*
Errors	1.6 (1.38)	1.8 (1.42)	1.9 (1.70)	h = 6.35 p = 0.042	
Models (Italian)					
Diff.	79.38 (11.380)	74.94 (12.096)	66.67 (15.615)	h = 158.27 p < 0.001	gpt4o/gpt4o-mini* gpt4o/Llama-3.3-70b** gpt4o-mini/Llama-3.3-70b**
A1 Coverage (%)	24.83 (14.278)	26.15 (15.295)	31.92 (18.195)	h = 34.47 p < 0.001	gpt4o/Llama-3.3-70b* gpt4o-mini/Llama-3.3-70b*
Errors	2.2 (2.06)	2.3 (2.16)	2.0 (1.91)	h = 4.35 p = 0.114	

Note: * indicates small effect size, ** indicates medium effect size, *** indicates large effect size.

relative improvement of 6%.

In contrast, for Italian, paraphrasing had a more substantial effect. The mean number of violations dropped from 4.5 (SD = 2.92, median = 4.00) in the original texts to 2.2 (SD = 2.05, median = 2.00) in the paraphrased versions. This difference was statistically significant ($p < 0.001$), corresponding to a 52% reduction in grammatical constraints violations on average.

We also evaluated lexical compliance by calculating the proportion of words in each text that appeared in the A1-level vocabulary list. For English, this proportion increased from an average of 42.2% (SD = 15.48) in the original texts to 57.4% (SD = 16.95) in the paraphrased versions, reflecting a relative improvement of 36%.

For Italian, the increase was even more pronounced: from 17.6% (SD = 11.48) to 27.6% (SD = 16.28), corresponding to a 57% relative improvement. In both cases, the difference was statistically significant ($p < 0.001$), indicating that paraphrasing substantially improved the lexical simplicity of the output.

The relationship between the amount of variation in the paraphrases and the compliance is somehow counterintuitive: for English, while the grammatical constraints violations are, as expected, reduced with the increase of the distance between the original version and the paraphrases (corr $= -0.15$), the coverage of the A1 vocabulary is also, unexpectedly, reduced (corr $= -0.31$). For Italian, both the A1 vocabulary coverage is reduced (corr $= -0.35$) and the number of errors is increased (corr $= 0.17$)

5.2 Meaning preservation and linguistic soundness

The statistical tests outlined in the previous section were also applied to the manually annotated metrics: meaning preservation and linguistic soundness. In these cases, no statistically significant differences were observed, probably due to the limited number of annotated sentences.

5.3 Token Consumption

We compared average token usage across all prompting strategies and language models (Fig. 2). Among the tested models, LLaMA required the fewest tokens across all conditions. GPT-4o yielded the highest token counts in COT settings, while GPT-4o-mini used the most tokens with no-COT prompting strategies. Chain-of-thought instructions was the dominant factor influencing token counts. Regardless of whether the prompting strategy was one- or two-step, COT variants consistently required significantly more tokens.

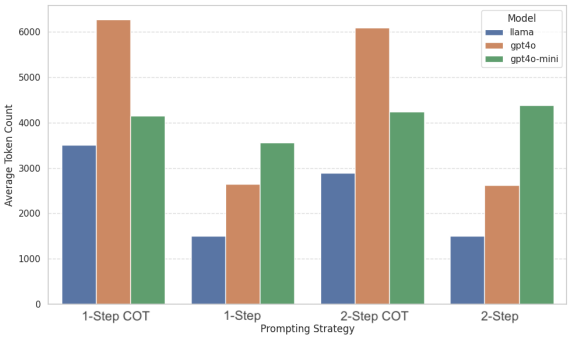

Figure 2: Average Token Usage by Prompting Strategy and Model for both languages

The average token usage for the automatic grammatical checking step was substantially higher in English (mean = 16,454 tokens/example) than in Italian (mean = 4,846), since the English automatic control prompt included more detailed instructions.

6 Discussion

Our results provide some evidence that guided paraphrases can improve compliance with A1 inventories and vocabulary lists, although not perfectly and with high cost (in terms of processed tokens). The paraphrases seem to be more effective for Italian than for English (but our original sentences tended to be less compliant in the Italian dataset than the English one).

For what concerns the prompting strategies, although the results have medium to small statistical effect size, it seems that for English the strategies that include COT work better in reducing the grammatical violations, while those with two steps cover better the vocabulary. Therefore the best choice would be the strategy that combines COT with the two-steps. However, the best choice is not so clear for Italian because the one-step strategies seem to have a better effect on grammatical constraints violations while the two-steps seem to cover better the vocabulary, with an unclear effect of the COT. This contrast might be due to the fact that grammatical constraints for Italian are more restrictive than those for English. As a result, when vocabulary simplification is applied as a separate step, in case of Italian there is more risk to inadvertently introduce constructions that violate earlier grammatical restrictions. Therefore, for Italian there is no best

strategy.

As for the models, Llama seems to better cover the vocabulary for both English and Italian. On the grammatical violations, there is no difference among the models for Italian, while for English GPT-4o and GPT-4o-mini seem to perform better than Llama.

Although our approach can still be improved, our results may contribute to a better understanding of the task of language adaptation to CEFR constraints using LLMs. Previous studies have highlighted the limitations of relying solely on general CEFR-level prompts. Benedetto et al. (2025) note that communication-based descriptors from the CEFR Companion Volume (Council of Europe, 2020) are not specific enough to effectively control language model output. Similarly, in the study by Uchida (2025), prompting with only the target CEFR level led to oversimplified A1 and A2 texts. These limitations suggest the need for more fine-grained control. Like Bektaş et al. (2024), who successfully used zero-shot prompting with detailed rules to generate simplified Turkish texts, we adopted a more detailed constraint-based approach. However, while the method of (Bektaş et al., 2024) was grounded in expert-written simplification guidelines, our constraints were derived from language-specific CEFR inventories by Trim (2009) and Barni et al. (2009). Using detailed inventories along with a task to paraphrase the text maintaining as much information as possible might be the reason why in our case outputs tended to exceed the intended constraints boundaries. Although the language specific grammar inventories are not easy to formalize for prompt usage and to check against, they are the base for learning and assessing foreign language competence, and it is important for intelligent tutoring system to control those constraints.

An important aspect that deserves further investigation might be the counterintuitive results on the relationship between the complexity of the paraphrases and the constraints compliance (which is more pronounced for Italian but is also partially present for English). It can perhaps be partially due to the reduced competence on Italian of these models. Yet, it has to be noted that the task of paraphrasing to beginner CEFR levels is complicated for humans too. Indeed, Torregrosa Benavent and Sánchez-Reyes Peñamaría (2011) note that authentic materials are difficult to adapt for beginner learners, often requiring extensive rewriting and Cheng and Fox (2017) suggest writing reading assessment texts from scratch or adapting existing texts freely, rather than trying to preserve all the original details, in order to better adhere to specific linguistic requirements, like those that we have imposed on the models we tested.

7 Release

All the data and source code used for our experiments is released on Github[4] under Creative Commons Attribution 4.0 International License (CC-BY-4.0) for the data and Apache 2.0 for the code.

The package contains: (i) the data downloaded from Vikidia in English and Italian; (ii) the sentences selected for our experiments, with original and rephrased versions, along with manual annotation of meaning preservation and linguistic soundness, and automatic tagging of compliance; (iii) the Python source code used to run the experiments.

8 Conclusion

Our study shed some light on the problem of constraining LLMs output to a specific level of language competence, by exploring in particular CEFR A1 constraints. Specifically, we explored the possibility of paraphrasing the text with testing the combination of four strategies with three different LLMs. Furthermore, we investigated both English and Italian.

Our results provided some evidence that the compliance can be improved, although our approach did not reach a satisfactory level since the number of grammatical errors and lack in vocabulary coverage are still high for all the strategies and for both English and Italian.

Still we believe that, given the relevance of the task for improving intelligent tutoring systems for language learning, the paraphrasing approach, although complex, is worth investigating. Our results are a promising first step.

Finally, another relevant outcome of our work is a dataset of cleaned pairs of sentences and paraphrases (in two languages) on which further work could be based.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the PNRR MUR program funded by the NextGenerationEU.

[4] https://github.com/gemini64/a1-llm

Limitations

This study relies on proprietary LLMs, specifically OpenAI's GPT-4o and GPT-4o-mini models.

While the specific versions used are clearly specified, we recognize that accessibility to particular models and versions through proprietary services, such as OpenAI's API platform, may change over time.

To mitigate potential reproducibility issues, an open-weights model, Meta's Llama-3.3-70b, was included in our LLM-driven paraphrasing tests.

Future extensions of this work should consider applying the automated grammar analysis approach to additional open-source/open-weight models.

Ethical considerations

This study relies exclusively on publicly available, open-access texts and does not involve human participants or sensitive data. All annotations were conducted by the authors themselves as part of the research process. Since no external annotators were involved, no ethical risks related to workload, compensation, or consent arise. The use of generative language models was limited to the controlled rewriting of non-sensitive content, with no deployment in real-world educational settings. Therefore, no ethical concerns were identified.

References

Waheeb S. Albiladi. 2018. Exploring the use of written authentic materials in ESL reading classes: Benefits and challenges. *English Language Teaching*, 12(1):67.

Monica Barni, Anna Bandini, Laura Sprugnoli, Silvia Lucarelli, Anna Maria Scaglioso, Beatrice Strambi, Chiara Fusi, and Anna Maria Arruffoli. 2009. *Linee guida CILS: Certificazione di Italiano come Lingua Straniera*. Guerra Edizioni.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Fatih Bektaş, Kutay Arda Dinç, and Gülşen Eryiğit. 2024. LLMs for document-level text simplification in Turkish foreign language learning. In *2024 9th International Conference on Computer Science and Engineering (UBMK)*, pages 1–5. IEEE.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Liying Cheng and Janna Fox. 2017. *Assessment in the Language Classroom: Teachers Supporting Student Learning*. Applied Linguistics for the Language Classroom. Palgrave Macmillan, London.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. 2023. Learning to paraphrase sentences to different complexity levels. *Transactions of the Association for Computational Linguistics*, 11:1332–1354.

Adam Cohen. 2025. thefuzz: Fuzzy string matching in python. Python package on PyPI.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg.

Karën Fort, Adeline Nazarenko, and Sophie Rosset. 2012. Modeling the complexity of manual annotation tasks: a grid of analysis. In *Proceedings of COLING 2012*, pages 895–910, Mumbai, India. The COLING 2012 Organizing Committee.

Valentina Franzoni, Giulio Biondi, and Alfredo Milani. 2024. Morpho-phraseological based classification of CEFR Italian L2 learner writing proficiency. *IEEE Access*, 12:156433–156441.

Dominik Glandorf and Detmar Meurers. 2024. Towards fine-grained pedagogical control over English grammar complexity in educational text generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 299–308, Mexico City, Mexico. Association for Computational Linguistics.

Sghaier Guizani, Tehseen Mazhar, Tariq Shahzad, Wasim Ahmad, Afsha Bibi, and Habib Hamam. 2025. A systematic literature review to implement large language model in higher education: issues and solutions. *Discover Education*, 4(1):1–25.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or Fumble? evaluating readability standard alignment of instruction-tuned language models. *IEEE Games Entertainment Media Conference*.

Stephen D. Krashen. 1985. *The Input Hypothesis: Issues and Implications*. Longman.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Sunjung Lee and Diana Pulido. 2016. The impact of topic interest, L2 proficiency, and gender on EFL incidental vocabulary acquisition through reading. *Language Teaching Research*, 21(1):118–135.

Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. 2024. "we need structured output": Towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–9. ACM.

Ali Malik, Stephen Mayhew, Chris Piech, and Klinton Bicknell. 2024. From Tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. *arXiv*.

Oxford University Press. 2025. The oxford 3000 and 5000.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

R. Ramadhani, H. Aulawi, and R. L. Ulfa. 2023. Readability of reading texts as authentic materials issued by ChatGPT: A systemic functional perspective. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 8(2):149–168.

Barbara Spinelli and Francesca Parizzi. 2010. *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2*. La Nuova Italia / RCS Libri, Firenze.

Nicolas Spring and Annette Rios. 2021. Exploring German multi-level text simplification. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*, RANLP 2021, page 1339–1349. INCOMA Ltd. Shoumen, BULGARIA.

Kevin Stowe, Debanjan Ghosh, and Mengxuan Zhao. 2022. Controlled language generation for language learning items. *Conference on Empirical Methods in Natural Language Processing*.

Gabriela Torregrosa Benavent and Sonsoles Sánchez-Reyes Peñamaría. 2011. Use of authentic materials in the ESP classroom. *Encuentro*, 20:89–94.

John Trim. 2009. *Breakthrough Specification*. European Association for Language Testing and Assessment (EALTA). Unpublished manuscript.

Satoru Uchida. 2025. Generative ai and CEFR levels: Evaluating the accuracy of text generation with chatgpt-4. *Vocabulary Learning and Instruction*, 14(1):2078.

Università per Stranieri di Siena. 2017. Prova di livello a1 standard — esempio di esame cils.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint*.

Lowri Williams, Michael Arribas-Ayllon, Andreas Artemiou, and Irena Spasić. 2019. Comparing the utility of different classification schemes for emotive language analysis. *Journal of Classification*, 36(3):619–648.

A Grammar Constraints

A.1 Grammar Constraints for English

- **Adverbs, Prepositions, Conjunctions, and Interjections**: These may be used without limitations.
- **Nouns**: Singular, regular plural, and possessive forms are allowed. Irregular plural noun forms are forbidden.
- **Pronouns**: Personal, possessive, interrogative (*who*, *what*, and *which* only), and demonstrative pronouns are allowed. Any other types of pronouns are forbidden.
- **Adjectives**: Descriptive (base form, regular comparative, and regular superlative), interrogative, and possessive adjectives are allowed. Irregular adjective forms are forbidden.
- **Verbs**: Modal verbs *can* and *will* are allowed; all other modal verbs are forbidden. All other verbs are allowed.
- **Verbs (finite forms)**: Only the following finite forms are allowed:
  - *Indicative*: present simple, past simple, present perfect, present continuous, and past continuous
  - *Imperative*
- **Verbs (non-finite forms)**: All the non-finite forms are allowed.
- **Verbs conjugation (voices)**: Verbs in indicative present simple and indicative past simple may be conjugated in either active or passive voice. In all other cases, verbs must be conjugated in active voice.

A.2 Grammar Constraints for Italian

- **Nouns, Adjectives, Adverbs, Prepositions, Conjunctions, and Interjections**: These may be used without limitations.
- **Pronouns**: Only personal, possessive, demonstrative, interrogative, and indefinite pronouns are allowed.

- **Numerals**: Cardinal numbers may be used without limitation. Ordinal numbers must be limited to range 1–3.
- **Verbs**: All the regular Italian verbs are allowed. Only the following irregular verbs are allowed: *essere*, *avere*, *volere*, *potere*, *dovere*, *andare*, *dare*, *dire*, *fare*, *sapere*, *stare*, *venire*, *chiudere*, *mettere*, *morire*, *nascere*, *prendere*, *scrivere*. Any other irregular verbs are forbidden.
- **Verb voice**: Verbs have to be conjugated in active voice.
- **Allowed moods and tenses for verb conjugation**:
  - *Indicativo*: presente and passato prossimo
  - *Infinito*: presente
  - *Imperativo*: presente (only 2nd person singular and plural)
  - *Condizionale*: presente (only *vorrei* — first person singular of *volere*)

B Lexical Simplification Prompt

Task

Adapt the vocabulary choice of a given text for a target audience of CEFR {cefr_level}-level language learners.

Instructions

- Aim to use basic, high-frequency vocabulary typical of {cefr_level} level.

- Maintain the original text's core information and meaning.

- Prioritize clarity and comprehension.

- Keep capitalized proper names as they are.

- In exceptional circumstances, you can remove particularly uninformative proper names.

- When you encounter a word that a {cefr_level} student is not expected to know: either explain it with simple words, replace it with a simpler alternative. Remove it only if it is not essential to the main message.

Input

{input_text}

Output format Format your response as shown below, no additional comment is needed:

<text>[Your final version of the adapted text here - use the original text if no changes were needed]</text>

Note: The final version MUST be wrapped in <text> tags, regardless of whether changes were made to the original text.

C Guidelines for Manual Annotation

The following tables reports the guidelines provided to annotators in order to assist them in evaluating linguistic soundness and meaning preservation in simplified texts. Words and expression **in bold** in the examples highlight linguistic issues.

C.1 Linguistic Soundness

Score	Description	Example
n/a	Original text	Aida (sometimes spelled Aïda), is a tragic opera in four acts. Antonio Ghislanzoni wrote the words. Giuseppe Verdi composed the music. Aida was first performed at the Khedivial Opera House in Cairo, Egypt on 24 December 1871. Since then, it has been performed around the world. Favorite musical numbers include "Heavenly Aida" (Celeste Aida) and the "Triumphal March". The opera was made into a movie in 1953 starring Sophia Loren.
5	Native-like – Completely fluent, well-formed, no grammatical, lexical or stylistic issues.	Aida is a sad opera in four parts. Antonio Ghislanzoni wrote the words. Giuseppe Verdi made the music. Aida was first shown at the Khedivial Opera House in Cairo, Egypt on 24 December 1871. Since then, it has been shown in many places around the world. Popular songs from the opera include "Heavenly Aida" and the "Triumphal March." The opera was also made into a movie in 1953 with Sophia Loren.
4	Mostly fluent – Slight unnaturalness or minor grammatical issues. Sounds acceptable and clear.	'Aida (sometimes spelled Aïda) is a sad opera, which means a theater play where characters sing and not speak. It has four parts. The words were written by Antonio Ghislanzoni, and the music was composed by Giuseppe Verdi. Aida was first performed at the Khedivial Opera House in Cairo, Egypt, on 24 December of 1871. Since then, it has been performed all over the world. Some well-known musical **elements** from the opera include Heavenly Aida (Celeste Aida) and the "Triumphal March". The opera was turned into a movie in 1953, starring Sophia Loren.'
3	Understandable – Contains awkward phrasing or lexical choices that a native speaker would avoid , but overall meaning is clear.	Aida is a music story in four parts. Antonio Ghislanzoni **did** the words. Giuseppe Verdi **did** the music. It was first played in Cairo at the **music building**. After that, people watched it **in the world**. Songs like "Heavenly Aida (Celeste Aida) and the "Triumphal March are famous. A movie of the story came in 1953 with Sophia Loren.
2	Hard to understand – Major issues with unnatural wording or grammar (n < 5). Reader must guess the intended meaning.	'Aida (sometimes spelled Aïda) is a sad **story** in four parts. Antonio Ghislanzoni wrote the words. Giuseppe Verdi created the music. Aida was first shown at the Khedivial Opera House in Cairo, Egypt, on 24 December 1871. Since then, it has been shown around the world. Popular songs include **"Beautiful Aida"** (Celeste Aida) and the "**Big Parade."** The story was made into a movie in 1953 starring Sophia Loren.'
1	Not fluent – very broken or incoherent (n of errors >= 5) Grammar and vocabulary severely hinder comprehension. Sounds like machine-translation distortion.	Aida (in some cases with dots on i) is an unhappy **tale in four times**. Antonio Ghislanzi made the **saying part**. Music was done by Giuseppe Verde. Aida had its first show at the Egyptian Khedive's music palace on the **24th of Christmas**, 1871. **After then**, it has been shown in **many earth places**. Songs l**oved by ears** are "**Sky Aida**" and "March for the Win." In 1953, this **art changed** into a **big screen happening** with **famous woman** Sophia Loren.

C.2 Meaning Preservation

Score	Description	Example
n/a	Original text	Aida (sometimes spelled Aïda), is a tragic opera in four acts. Antonio Ghislanzoni wrote the words. Giuseppe Verdi composed the music. Aida was first performed at the Khedivial Opera House in Cairo, Egypt on 24 December 1871. Since then, it has been performed around the world. Favorite musical numbers include "Heavenly Aida" (Celeste Aida) and the "Triumphal March". The opera was made into a movie in 1953 starring Sophia Loren.
5	All key information from the original is present. No content is omitted or added. The simplified sentence conveys exactly the same meaning.	Aida is a sad opera in four parts. Antonio Ghislanzoni wrote the words. Giuseppe Verdi made the music. Aida was first shown at the Khedivial Opera House in Cairo, Egypt on 24 December 1871. Since then, it has been shown in many places around the world. Popular songs from the opera include "Heavenly Aida" and the "Triumphal March." The opera was also made into a movie in 1953 with Sophia Loren.
4	One or two minor omissions or shifts in wording. No critical meaning is lost. A reader can reconstruct the full meaning without effort.	'Aida (sometimes spelled Aïda) is a sad opera, which means a theater play where characters sing and not speak. It has four parts. The words were written by Antonio Ghislanzoni, and the music was composed by Giuseppe Verdi. Aida was first performed at the Khedivial Opera House in Cairo, Egypt, on 24 December 1871. Since then, it has been performed all over the world. Some well-known musical pieces from the opera include Heavenly Aida (Celeste Aida) and the **Big Parade**. The opera was turned into a movie in 1953, starring Sophia Loren.'
3	One or two important losses or shifts in meaning. The overall idea is still clear, but nuance or details are missing and original meaning reconstruction requires some effort or contextual knowledge.	'Aida (sometimes spelled Aïda) is a sad **story** in four parts. Antonio Ghislanzoni wrote the words. Giuseppe Verdi made the music. Aida was first shown at the Khedivial Opera House in Cairo, Egypt on 24 December 1871. Since then, it has been shown all over the world. Popular songs include "Heavenly Aida" and the "Triumphal March." The story was made into a film in 1953 with Sophia Loren.'
2	Multiple (n < 5) important elements are missing or misleading. The reader cannot fully recover the intended message.	'Aida (sometimes spelled Aïda) is a sad **story** in four parts. Antonio Ghislanzoni wrote the words. Giuseppe Verdi created the music. Aida was first shown at the Khedivial Opera House in Cairo, Egypt, on 24 December 1871. Since then, it has been shown around the world. Popular songs include **"Beautiful Aida"** (Celeste Aida) and the "**Big Parade."** The story was made into a movie in 1953 starring Sophia Loren.'
1	Meaning is severely distorted or completely unrelated. Multiple (n >= 5) factual elements are missing or mistaken, or the simplified sentence contradicts or replaces the original message.	Aida (in some cases with dots on i) is an unhappy **tale in four times**. Antonio **Ghislanzi** made the saying part. Music was done by Giuseppe **Verde**. Aida had its first show at the Egyptian Khedive's music palace on the **24th of Christmas**, 1871. After then, it has been shown in many earth places. Songs loved by ears are "**Sky Aida**" and "**March for the Win**." In 1953, this art changed into a big screen happening with famous woman Sophia Loren.