# Automatic Transformation of Textual Content Using LLM-Technology: a Feasibility Study

Koen Smit, Sam Leewis,
Annemae van de Hoef, Duain Castro
HU University of Applied Science Utrecht, Utrecht, the Netherlands
koen.smit@hu.nl, sam.leewis@hu.nl, annemae.vandehoef@hu.nl. duain.castro@hu.nl

This study investigates the feasibility of using Large Language Model (LLM)-technology to classify and transform textual (educational) content according to CEFR proficiency levels in support of digital accessibility in higher education. We developed a browser-based Proof-of-Concept that performs on-the-fly CEFR classification and transform text to target levels (A1–C2) in Dutch and English. Using a corpus of 120 texts evenly distributed across CEFR levels, we evaluated classification accuracy and transformation effectiveness under multiple temperature settings. Results show modest zero-shot classification performance and systematic mid-level bias, particularly for Dutch. Transformation outcomes were stronger in English, especially for B1–B2 targets, but weak for extreme levels and prone to instability. The findings suggest that while LLMs show promise for automated readability adaptation, reliable deployment requires task-specific tuning, multilingual robustness testing, and human-centered evaluation to ensure meaningful accessibility gains.

**DOI**
https://doi.org/10.18690/um.fov.4.2026.35

**ISBN**
978-961-299-152-4

**Keywords:**
large language models, CEFR, digital accessiblity, text transformation, research paper

University of Maribor Press

## 1 Introduction

Providing universally accessible educational materials is a critical goal in higher education. Digital accessibility guidelines such as Web Content Accessibility Guidelines (WCAG) emphasize that content must be clear and understandable for diverse learners (World Wide Web Consortium (W3C), 2026), and initiatives like *Easy-to-Read* and *Plain Language* exist specifically to simplify complex texts (Martínez et al., 2024). In practice, however, achieving clarity remains challenging. WCAG's "Understandable" principle advises using straightforward language (short sentences, common words, simple tense) because users with cognitive or language impairments may struggle with complexity (World Wide Web Consortium (W3C), 2024). Similarly, for example, research shows that a large fraction of the Dutch-speaking population requires content at around Common European Framework of Reference for Languages (CEFR) B1 level to ensure comprehension (OECD, 2013; Raad voor Cultuur, 2019). If institutional websites or course materials are written at higher levels (CEFR B2–C2), significant portions of readers may be excluded. Yet many educational texts remain too complex, and manual simplification is labor-intensive and inconsistent (Martínez et al., 2024). Ensuring accessible language is therefore not only a technical or regulatory requirement but also reflects broader human values such as inclusion, equity, and equal participation in education.

Across higher education, digital accessibility has increasingly become embedded in binding regulatory frameworks. In the European context, the WCAG are formally codified in ISO/IEC 40500:2025 and operationalized through EN 301 549 and the European Accessibility Act, thereby extending accessibility requirements to websites, digital documents, and software systems used by public institutions (ETSI, 2025; European Commission, 2025; ISO, 2025; Ministerie van Algemene Zaken, 2017; Rijkswaterstaat, 2025). Although WCAG does not prescribe a fixed reading level, literacy research provides an empirical benchmark for what constitutes broadly accessible communication. Data from the Organisation for Economic Co-operation and Development (OECD)'s Programme for the International Assessment of Adult Competencies (PIAAC) indicate that a considerable proportion of adults perform at literacy levels comparable to CEFR B1 or below (OECD, 2025). In this regulatory and demographic context, the challenge is not merely drafting accessible text but systematically assessing and operationalizing language-level alignment within digital environments.

Recent advances in Large Language Models (LLMs) offer new solutions. These AI models can analyze and transform text, potentially automating readability adaptation. Several studies report dramatic improvements in accessibility when using LLMs to simplify content. For instance, one study found that LLM-generated summaries of medical articles lowered the average reading grade level from about 9.5 to roughly 5–7 while preserving factual accuracy (Romoff et al., 2025). Similarly, LLMs (ChatGPT, Gemini, Claude) reduced health education pamphlets from grade ~10 down to the recommended 5–6 level with no loss of understandability (Will et al., 2025). In educational settings, AI-tools have allowed students to select textbook passages and receive AI-simplified versions, significantly reducing complexity without altering meaning (Johnson et al., 2025). These findings suggest that LLMs can effectively generate "easy-to-read" versions of complex texts while maintaining meaning.

Building on this potential, our research answers the following question: To what extent can an LLM Proof-of-Concept (PoC) reliably classify the accessibility level of Dutch and English texts and transform them into more accessible versions while preserving meaning? We explore this by developing a browser plugin that (1) automatically classifies any text's reading level (in CEFR terms) and (2) transforms the text to different target levels. This design-science PoC allows content creators and users to personalize language complexity on the fly. By focusing on both classification and transformation in two languages, our work goes beyond prior studies to examine feasibility for inclusive, personalized content delivery in higher education.

## 2 Background and Related Work

Web accessibility standards underscore the importance of plain language. WCAG 2.0 (Guideline 3.1) and related materials advise using clear, simple wording and short sentences to make information comprehensible (World Wide Web Consortium (W3C), 2026). This aligns with broader "plain language" and Universal Design principles that call for well-organized, readable course materials (Martínez et al., 2024). In Europe, text complexity is often gauged by the CEFR levels (A1–C2), which describe learners' language proficiency (Council of Europe Publishing, 2020). For example, studies equate WCAG AAA readability targets to about CEFR B1 (roughly elementary schooling level), see e.g., (Velleman & Van Der Geest, 2014; World Wide Web Consortium (W3C), 2025). In practice, most native adults achieve

around B1–B2; writing above that level risks excluding those with limited literacy. Although CEFR was developed for language learners, it provides a useful scale for "accessibility leveling" of texts. For example, Dutch experts have also proposed native-language frameworks, such as the Referentiekader Nederlandse Taal, to more precisely define readability for native speakers.

Traditional readability metrics (e.g. Flesch-Kincaid (DuBay, 2004)) use surface features like sentence length and word syllables, but these often miss deeper aspects of difficulty. Machine learning approaches have been developed to predict CEFR levels more accurately. For instance, neural models (BERT, RoBERTa) trained on CEFR-annotated corpora can classify Dutch and English texts by reading level. These classifiers capture linguistic patterns beyond formulaic scores, and outperform simple heuristics in predicting text complexity. Large annotated datasets and corpora are now available: for example, the CEFR-based Sentence Profile corpus and multilingual resources like UniversalCEFR provide training material across languages (Zhang et al., 2026). Nonetheless, high-quality CEFR datasets are scarcer for Dutch than for English. Some authors have constructed Dutch readability corpora (e.g. via expert assessment tools, see e.g. a Dutch dataset by Breuker (2023) or an assessment tool by Velleman & Van Der Geest, (2014), but the volume remains limited. Hence, while automated scoring tools exist (and some web tools can estimate CEFR levels), their coverage and reliability vary.

Text simplification has been studied extensively in Natural Language Processing (NLP), which can be used to extract semantics from text (Nadkarni et al., 2011). Early methods used handcrafted rules or statistical models to split sentences and substitute complex words (Siddharthan, 2006). Neural approaches later emerged (encoder-decoder and transformer models) for end-to-end simplification. Recent research focuses on controlled simplification, where a complex sentence is rewritten to match a target proficiency level. Zhang et al. (2026) note that LLMs struggle with large leaps in difficulty but can improve when guided step-by-step. In practice, LLMs (especially GPT-style models) have demonstrated strong single-shot simplification capabilities: for example, one study showed GPT-3.5 Turbo greatly outperformed a traditional translation-based pipeline in simplifying Dutch municipal texts (Vlantis et al., 2024). LLMs have also been used to generate "easy-to-read" Spanish texts by fine-tuning on accessible corpora (Martínez et al., 2024). Furthermore, Leewis et al. (2025) evaluated nine LLMs for CEFR classification of English and Dutch

educational texts, finding that larger models perform best but results vary across models and temperature settings. In this context, temperature is a configurable parameter of an LLM (typically ranging from 0 to 1) that controls the variability of the model's output. Higher values lead to more diverse and less predictable responses, while lower values produce more consistent and deterministic outputs (Peeperkorn et al., 2024).

Empirical work consistently finds that LLM simplifications reduce grade levels and improve readability while largely preserving meaning. In healthcare, ChatGPT and similar models significantly lowered the reading difficulty of patient education materials (to around 6th-grade level) without introducing factual errors (Will et al., 2025). In orthopaedics, GPT-based tools reduced the average reading grade of spine health articles from ~9.5 to ~5–6, with reviewers noting "acceptable accuracy" (Romoff et al., 2025). In educational publishing, an AI-powered "simplifier" in a textbook e-reader reduced language complexity (as measured by syntax and lexicon) while maintaining semantic fidelity (Johnson et al., 2025). These results suggest LLMs can dynamically adapt content to different proficiency targets. Crucially, however, few studies have tested these tools in higher education contexts or in languages other than English.

Our work builds on the notion that both designers as well as learners should control content complexity. Related educational tools (e.g. adaptive learning platforms) tailor material to student proficiency, but typically focus on question difficulty or topic sequencing and less on personalization of the content itself for inclusiveness. Browser extensions like Brisk (Brisk Teaching, 2026) have begun allowing readers to request simplified versions of web text, but these are stand-alone and often not scientifically evaluated. Accessibility checkers (e.g. WAVE, axe) and readability scoring tools exist, yet they usually only flag issues rather than transform text. Our plugin is novel in combining on-the-fly language-level classification with LLM-driven rewriting. By embedding this into a web browser, we enable content creators and consumers to label and transform any page's text in real time (Johnson et al., 2025; Romoff et al., 2025; Will et al., 2025) , and that classifiers can assign CEFR levels based on learned language features. However, to the knowledge of the authors, no published studies have evaluated the reliability of this technology for higher-education materials or directly compared its impact on Dutch and English content.

Our study addresses this gap by prototyping an accessibility tool and assessing its effectiveness for multilingual, inclusive content.

## 3 Research Method

This study follows a design-science orientation: we develop an LLM-based working artifact (a Chrome browser plugin) to demonstrate feasibility, and we evaluate its technical performance and practical utility through controlled experiments and complementary stakeholder conversations (Hevner et al., 2004). The artifact implements two core functions: (1) automatic CEFR classification of on-page texts and (2) automatic transformation (rewriting) of a given text to any CEFR target (A1–C2) while aiming to preserve meaning.

### Experimental design and variables

Each experiment used the same local LLM runtime, i.e., a locally deployed Copilot variant. Due to policy restrictions of the case organization, Copilot is selected as the only fitting LLM. For classification, each input text was submitted with three decoding/temperature settings (0.2, 0.6, and 1.0) to probe stability. The LLM output was then saved for further evaluation. For transformation, for each source text the model was instructed to transform it to each CEFR level (A1–C2), producing one rewritten variant per target level. All prompts used a role-assignment pattern (e.g., "You are an expert English teacher…") and were delivered in the language of the input (English prompts for English texts, Dutch prompts for Dutch texts) to avoid language-mixing artefacts documented in preliminary testing. Non-conforming or empty model outputs were logged as "No Result (NR)". The design decisions (local Copilot LLM, three temperature settings) were chosen to balance reproducibility, privacy, and an exploration of model determinism vs. diversity (Holtzman et al., 2020).

We evaluated the PoC using a corpus of 120 short texts: 60 Dutch and 60 English, evenly distributed across CEFR levels (A1–C2; 10 texts per level). The texts were representative of higher-education web contexts (e.g., course descriptions, policy pages, informational texts). Gold-standard CEFR labels were available in the English dataset used in this study, which consisted of an excerpt from an existing dataset (Montgomerie, 2021). For Dutch, a separate dataset was created by classifying short

texts using an existing text classification tool (Velleman & Van Der Geest, 2014). Regarding the validation of rewriting results, a model trained for English CEFR classification was used to evaluate the rewritten English texts (Lee, 2025). Alternative approaches were not feasible, as they either introduced additional CEFR levels or were not freely accessible. For Dutch, the same classification tool was used to evaluate the rewritten texts.

For the classification experiment, each text was submitted to the locally deployed Copilot LLM. The model was instructed to classify the manually inputted text into one of the CEFR levels (A1–C2). Each text was processed three times using the predefined temperature settings described in the experimental design (0.2, 0.6, and 1.0). Outputs were constrained to a predefined JSON structure containing the assigned CEFR level. Empty or malformed outputs were logged as "No Result" (NR). Classification performance was assessed by comparing the model's assigned level with the known CEFR label of the input text. Results were summarized using overall accuracy and confusion matrices to identify patterns of over- or under-classification across levels. Differences between temperature settings were examined descriptively.

For the transformation experiment, each original text was rewritten to each CEFR target level (A1–C2) using structured prompts. The model output consisted of a rewritten version of the text in JSON format. The rewritten texts were subsequently re-classified to determine whether the intended target level had been achieved. The primary measure for transformation was therefore the proportion of rewritten texts that were classified at the requested CEFR level. In addition to quantitative comparison, outputs were qualitatively inspected to identify recurring patterns, such as systematic over- or under-simplification, instability across temperature settings, or deviations from the intended target level. This descriptive analysis was used to interpret performance trends and inform refinement of prompt design. All prompts and output logs were archived to ensure transparency and reproducibility of the experiments. The data and analysis is available upon request.

## 4 Results

An initial PoC was developed on the basis of several requirements gathered from stakeholders during exploratory discussions. These conversations were intentionally unstructured and open in nature, organised around thematic expectations concerning potential users of the plugin among the staff of the university of applied sciences.

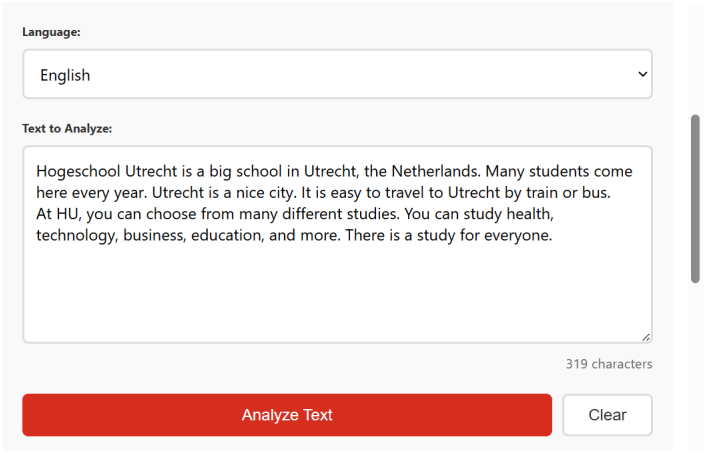


**Figure 1: PoC screenshot of the textual input**
Source: Own

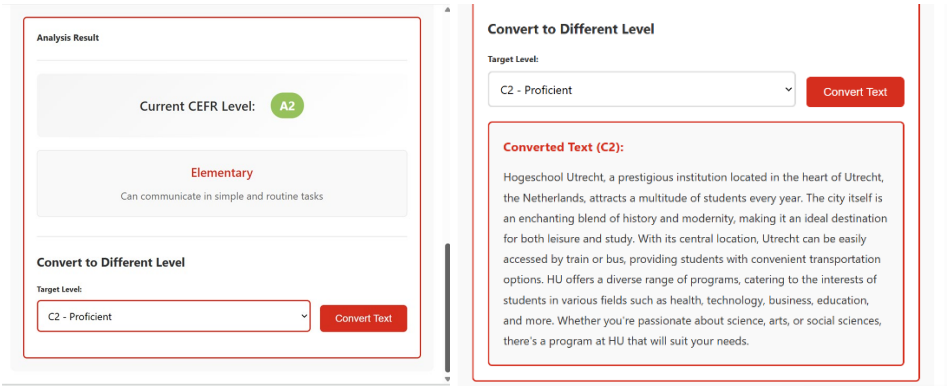


**Figure 2: CEFR-based language classification (left) and transformation (right) in the PoC**
Source: Own

Given that the PoC represents an early-stage implementation, the primary focus was testing the feasability of utilizing a LLM on classifying and transforming Dutch and English texts. In this early-stage the focus lies on establishing core functionality and testing the performance of the underlying LLM in relation to the front-end interface. Particular attention was paid to ensuring security and privacy, by running the LLM locally. Screenshots of the PoC are presented in Figure 1 (textual input) and Figure 2 (left: classification result, right: transformation functionality).

## Classification

Classification performance was modest for both languages. Using the Copilot LLM via the browser plugin, each text was classified at three temperature settings (0.2, 0.6, 1.0). For English, overall accuracy was 33.33% at temperatures 0.2 and 0.6, and 23.33% at 1.0 (Table 2). For Dutch, the best accuracy (35.0%) occurred at temperature 0.2, decreasing to 26.67% at 0.6 and 23.33% at 1.0 (see Table 1). Based on these results, a lower temperature setting (0.2), which produces more deterministic outputs, appears to yield slightly better performance for Dutch. For English, temperatures of 0.2 and 0.6 show comparable results, suggesting that performance is relatively stable across lower to mid-range temperature settings. However, overall differences between settings are modest, indicating that temperature may have only a limited impact on classification performance in this setup.

**Table 1: Accuracy metrics for CEFR text classification**

| Temperature | Total | Correct NL | Accuracy NL | Correct EN | Accuracy EN |
|---|---|---|---|---|---|
| 0.2 | 60 | 21 | 35% | 20 | 33,33% |
| 0.6 | 60 | 16 | 26,67% | 20 | 33,33% |
| 1.0 | 60 | 14 | 23,33% | 14 | 23,33% |

*English (Temp 0.2, Figure 3).* Classifications are relatively dispersed across levels, but systematic errors are evident. B2 texts are commonly recognised, whereas A2 examples are often mislabelled as A1 or B1. High-level texts (C1, C2) are frequently predicted as B2, indicating a mid-level convergence bias (tendency to cluster around B2). *Dutch (Temp 0.2, Figure 3).* The model shows a strong bimodal tendency towards A1 and B2. All A1 texts were correctly classified, but A2 examples were entirely predicted as A1; most higher-level texts (B1–C2) were classified as B2. This indicates difficulty distinguishing adjacent levels and particular confusion around the mid-

range levels (B1/B2). *English (Temp 0.6, Figure 4)*. The classifications closely mirrored those at temperature 0.2: A1 and B2 texts were consistently identified, while higher-level texts were often predicted as B2. *Dutch (Temp 0.6, Figure 4)*. Similarly, the results resembled those at temperature 0.2, again revealing a strong bimodal tendency toward A1 and B2 classifications. *English (Temp 1.0, Figure 5)*. At this temperature, the LLM demonstrated the greatest difficulty in classification, producing highly sporadic results. Notably, B1–C1 texts were not correctly classified once. *Dutch (Temp 1.0, Figure 5)*. Similarly, this temperature yielded more sporadic results. This is unsurprising, as classification is generally considered a deterministic task. A temperature of 1.0 is therefore too high for this purpose, as it reduces the model's determinism and leads to less consistent predictions.

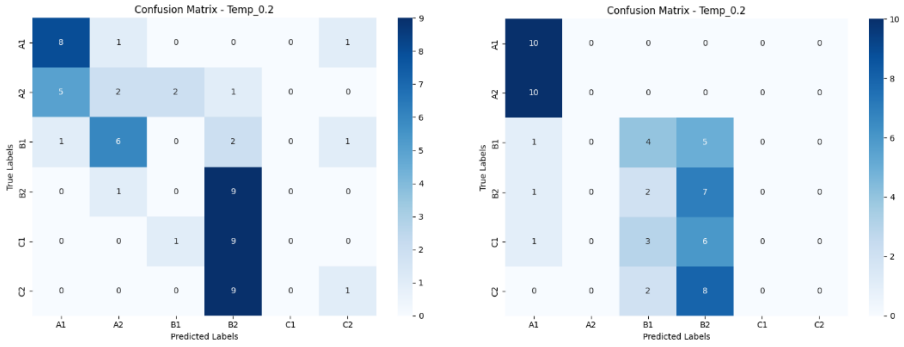


**Figure 3: Confusion patterns of English (left) and Dutch (right) CEFR text classification (T = 0.2)**

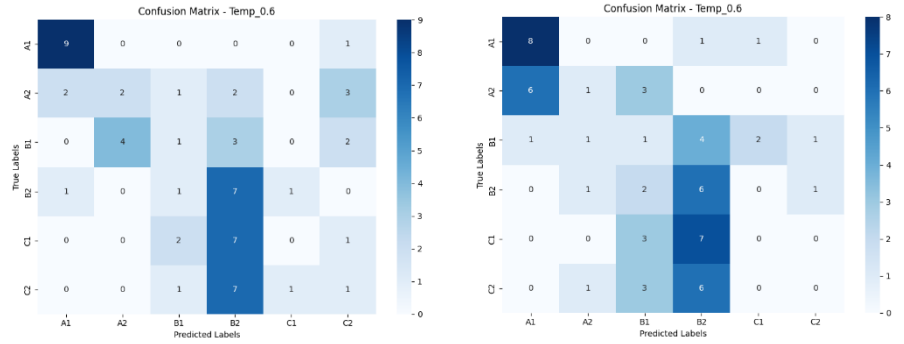


**Figure 4: Confusion patterns of English (left) and Dutch (right) CEFR text classification (T = 0.6)**

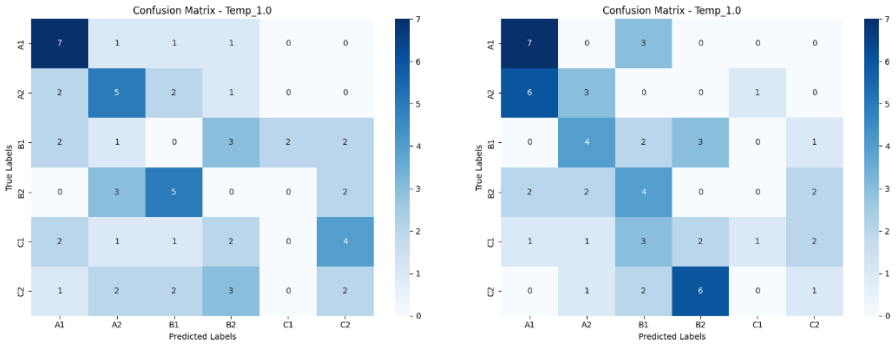


**Figure 5: Confusion patterns of English (left) and Dutch (right) CEFR text classification (T = 1.0).**

Overall accuracy is low in absolute terms and the confusion matrices show systematic biases rather than random noise. For English the model tends to collapse higher levels into B2; for Dutch it concentrates predictions at A1 or B2, with A2 collapsing to A1 and higher levels mapped to B2. The results suggest that, in the present zero-shot configuration, the Copilot LLM struggles to finely discriminate adjacent CEFR categories and that lower-temperature (more deterministic) decoding generally produced better classification accuracy.

## Transformation

| Language | Temperature | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|---|
| Dutch | 0.2 | 8/56 (14.29) | 13/56 (23.21) | 18/54 (33.33) | 14/54 (25.93) | 12/48 (25.00) | 10/55 (18.18) |
| Dutch | 0.6 | 4/57 (7.02) | 10/55 (18.18) | 18/53 (33.96) | 15/55 (27.27) | 11/54 (20.37) | 11/53 (20.75) |
| Dutch | 1.0 | 3/55 (5.45) | 12/53 (22.64) | 21/54 (38.89) | 16/52 (30.77) | 11/46 (23.91) | 9/50 (18.00) |
| English | 0.2 | 21/60 (35.00) | 24/60 (40.00) | 17/60 (28.33) | 36/60 (60.00) | 11/60 (18.33) | 3/60 (5.00) |
| English | 0.6 | 23/60 (38.33) | 29/60 (48.33) | 22/60 (36.67) | 33/60 (55.00) | 13/60 (21.67) | 6/60 (10.00) |
| English | 1.0 | 23/60 (38.33) | 26/60 (43.33) | 24/60 (40.00) | 37/60 (61.67) | 23/60 (38.33) | 5/60 (8.33) |

**Table 2: Transformation results of both Dutch and English languages. English totals per cell were 60 transformations in the reported experiment. Dutch totals vary because 77 transformations were produced in a different language and 43 transformations returned no output or an invalid text, reducing the evaluable Dutch sample**

The transformation experiments measured whether the LLM could transform each source text to specified CEFR target levels (A1–C2). For each language we report the percentage of transformations that were subsequently classified at the requested target level (i.e., target-match accuracy). The English results are substantially stronger than the Dutch ones; additionally, some Dutch transformations could not be evaluated because they were produced in the wrong language or the output was invalid (see note below).

- English. The LLM transforms most reliably to B2, achieving its highest target-match rates (≈55–62%). Mid-range targets (A1–B1) show moderate performance (≈28–48%), while the highest target (C2) has very low match rates (≈5–10%). The effect of temperature is not consistent across CEFR levels. In some cases (e.g., C1 at Temp 1.0), higher temperature leads to improved target-match accuracy, while in other cases no clear or systematic trend is observed.
- Dutch. Across all targets the accuracies are markedly lower. The best Dutch result is for B1 at Temp 1.0 (38.89%). Several target levels (notably A1) perform poorly and often decline as temperature increases. While lower temperature settings (0.2) sometimes yield slightly more stable results for certain levels such as A1, A2, and C1, higher temperatures occasionally produce better outcomes for intermediate levels such as B1 and B2. Overall, similar to English, no clear monotonic effect of temperature on performance can be observed. The Dutch evaluation is further complicated by missing / invalid outputs (see Table 2 note).

The transformation results indicate that the Copilot LLM, in this zero-shot plugin configuration, is better at producing transformations in English and is most effective for mid-level targets (especially B2). The model performs poorly for very advanced targets (C2) in both languages. For Dutch, the frequency of non-usable outputs and language-switching suggests limited robustness, which might be related to the model's training data distribution.

## 5 Discussion

The work of Martínez et al. (2024), Romoff M et al. (2025), Will et al. (2025), and Johnson et al. (2025) consistently report that LLMs can lower reading difficulty while largely preserving meaning, especially in health and textbook domains. These studies

emphasize human validation, report strong grade-level reductions, and validate AI simplification in single-language (mostly English) settings. Our work complements and extends this evidence by operationalizing CEFR levels (A1–C2) rather than grade metrics, combining both automatic CEFR classification and controlled transformation, and directly comparing Dutch and English performance. Unlike prior studies that focus on simplification outcomes, we reveal systematic zero-shot biases (mid-level convergence for English; bimodal clustering for Dutch) and practical failure modes.

This study has several limitations that we need to address. First, the PoC relied on a locally deployed Copilot LLM without task-specific fine-tuning for CEFR classification or controlled rewriting. The modest classification accuracy and uneven transformation results may therefore reflect limitations of zero-shot prompting rather than intrinsic limits of LLM technology. Also, although three temperatures were tested, temperature strategy exploration remained limited. Other sampling or prompting strategies were not systematically explored. Lastly, recent work shows that LLM-based approaches can achieve strong performance in text classification tasks, particularly when combined with fine-tuning or few-shot strategies, which often outperform zero-shot prompting approaches (Alizadeh et al., 2024; Bucher & Martini, 2024). This suggests that the relatively modest performance observed in this study may be partly attributable to the use of a locally deployed model without task-specific optimization. Another complexity is that the experimental corpus consisted of 120 short texts, which, while stratified by CEFR level, may not capture the full linguistic variability of higher-education materials. Furthermore, the study equates textual accessibility with CEFR level alignment, whereas accessibility as defined by WCAG encompasses additional structural (e.g., semantic markup, heading structure, navigability) and cognitive (e.g., predictability, clarity of instructions, cognitive load) dimensions not addressed in this evaluation. Another limitation is that we evaluated the classification of Dutch texts and the transformation of both languages' effectiveness through automated re-classification, which may inherit limitations of the underlying classification approach and introduce cyclical evaluation. Unfortunately, alternative approaches were not explored due to the limited availability of suitable open-source classification tools within the project constraints. In the end, we argue that this study shows that the LLM-technology, with proper finetuning and evolution (e.g., using guardrails), might be an adequate tool to support practitioners to increase the accessibility of their (digital) textual content.

## 6 Future Work

Future research should examine how task-specific model adaptation affects CEFR classification and controlled text transformation. This study relied on a general-purpose LLM in a zero-shot configuration, which limited stability when distinguishing adjacent CEFR levels. Experiments with fine-tuning or instruction tuning tailored to CEFR tasks may improve level discrimination. Comparative testing of prompt designs, few-shot versus zero-shot approaches, and alternative temperature settings could further clarify which configurations produce the most reliable outputs. An additional recommendation would be to conduct a benchmark study comparing different LLMs in order to determine which models perform best for this task. The empirical basis should also be expanded by including longer and more complex texts as well as discipline-specific materials such as STEM, legal, or social science content to assess robustness across domains. In addition, more rigorous human-centered evaluation is needed. Automated CEFR reclassification offers a scalable metric but does not fully capture meaning preservation or accessibility improvements. Future work should therefore include expert CEFR raters, stakeholder comprehension testing, and user-based evaluations. Field deployment studies could further explore how stakeholders use such tools and how automated adaptations are perceived in practice. Finally, the performance differences observed between English and Dutch highlight the need for systematic multilingual research, including analysis of cross-language robustness, language-switching outputs, and the effects of language-specific tuning.

## 7 Conclusion

This study contributes both theoretically and practically to the development on automated readability adaptation in higher education. Theoretically, it shows how CEFR can be operationalized as a scale for digital accessibility research by combining CEFR-based classification with controlled transformation in a design-science artifact. The results also nuance expectations about LLMs: in a zero-shot setting they exhibit systematic biases (mid-level convergence in English and bimodal clustering in Dutch) and limited ability to distinguish adjacent CEFR levels, highlighting limits to out-of-the-box robustness. Practically, the study presents a browser-based plugin for on-the-fly classification and rewriting and identifies key implementation lessons. While LLM-driven rewriting can produce useful adaptations, particularly for mid-

range English levels (B1–B2), issues such as language mixing, invalid outputs, and low reliability for extreme levels (A1, C2) still limit deployment readiness. The local deployment approach further illustrates a privacy-conscious integration strategy. These findings point to a pragmatic agenda for future work: targeted model tuning, richer multilingual corpora, structured human evaluation, and socio-technical assessment when moving from PoC to prototypes and real-world implementation.

**Acknowledgments**

We would like to thank the team of the Digital Learning Environment of HU University of Applied Sciences Utrecht in the Netherlands for their support in this project.

**References**

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2024). *Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning* (arXiv:2307.02179). arXiv. https://doi.org/10.48550/arXiv.2307.02179

Breuker, M. (2023). CEFR Labelling and Assessment Services. In G. Rehm (Ed.), *European Language Grid* (pp. 277–282). Springer International Publishing. https://doi.org/10.1007/978-3-031-17258-8_16

Brisk Teaching. (2026). *Personalize instruction with AI, right in your workflow*. Brisk Teaching. https://www.briskteaching.com/

Bucher, M. J. J., & Martini, M. (2024). *Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification* (arXiv:2406.08660). arXiv. https://doi.org/10.48550/arXiv.2406.08660

Council of Europe Publishing. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Companion volume*. Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages

DuBay, W. H. (2004). *The Principles of Readability*.

ETSI. (2025). *Draft EN 301 549* (Version 4.1.0). ETSI. https://www.etsi.org/deliver/etsi_EN/301500_301599/301549/04.01.00_20/en_301549v040100ev.pdf

European Commission. (2025). *The EAA comes into effect in June 2025. Are you ready?* European Commission. https://accessible-eu-centre.ec.europa.eu/content-corner/news/eaa-comes-effect-june-2025-are-you-ready-2025-01-31_en

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–106. https://doi.org/10.2307/25148625

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). *The Curious Case of Neural Text Degeneration* (arXiv:1904.09751). arXiv. https://doi.org/10.48550/arXiv.1904.09751

ISO. (2025). *ISO/IEC 40500:2025* (2nd edn). https://www.iso.org/standard/91029.html

Johnson, B. G., Jerome, B., Dittel, J. S., & Campenhout, R. V. (2025). Improving Textbook Accessibility through AI Simplification: Readability Improvements and Meaning Preservation. *Proceedings of the Sixth International Workshop on Intelligent Textbooks (iTextbooks 2025)*.

Lee, W. (2025). *CEFR Classifier* [Computer software]. https://huggingface.co/dksysd/cefr-classifier

Martínez, P., Ramos, A., & Moreno, L. (2024). Exploring Large Language Models to generate Easy to Read content. *Frontiers in Computer Science*, *6*, 1394705. https://doi.org/10.3389/fcomp.2024.1394705

Ministerie van Algemene Zaken. (2017, October 31). *Taalniveau B1* [Webpagina]. Dienst Publiek en Communicatie. https://www.communicatierijk.nl/vakkennis/rijkswebsites/aanbevolen-richtlijnen/taalniveau-b1

Montgomerie, A. (2021). *CEFR Levelled English Texts* [Data set]. https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

OECD. (2013). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. OECD. https://doi.org/10.1787/9789264204256-en

OECD. (2025). *Education at a Glance 2025: OECD Indicators*. OECD Publishing. https://doi.org/10.1787/1c0d9c79-en

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). *Is Temperature the Creativity Parameter of Large Language Models?* (arXiv:2405.00492). arXiv. https://doi.org/10.48550/arXiv.2405.00492

Raad voor Cultuur. (2019). *Lees! Een oproep tot een leesoffensief*. Den Haag: Raad Voor Cultuur. https://www.raadvoorcultuur.nl/documenten/2019/06/24/lees

Rijkswaterstaat. (2025). *Schrijftips voor taalniveau B1 in toepasbare regels*. Informatiepunt Leefomgeving. https://iplo.nl/digitaal-stelsel/toepasbare-regels/maken-testen/schrijfwijzer/schrijftips-taalniveau-b1/

Romoff, M., Brunette, M., Peterson, M. K., Hashmi, S. Z., & Kim, M. S. (2025). The role of large language models in improving the readability of orthopaedic spine patient educational material. *Journal of Orthopaedic Surgery and Research*, *20*(1), 531. https://doi.org/10.1186/s13018-025-05955-1

Siddharthan, A. (2006). *Syntactic simplification and text cohesion* (p. 195 pages) [PDF]. Computer Laboratory, University of Cambridge. https://doi.org/10.48456/TR-597

Velleman, E., & Van Der Geest, T. (2014). Online Test Tool to Determine the CEFR Reading Comprehension Level of Text. *Procedia Computer Science*, *27*, 350–358. https://doi.org/10.1016/j.procs.2014.02.039

Vlantis, D., Gornishka, I., & Wang, S. (2024). Benchmarking the Simplification of Dutch Municipal Text. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2217–2226.

Will, J., Gupta, M., Zaretsky, J., Dowlath, A., Testa, P., & Feldman, J. (2025). Enhancing the Readability of Online Patient Education Materials Using Large Language Models: Cross-Sectional Study. *Journal of Medical Internet Research*, *27*, e69955. https://doi.org/10.2196/69955

World Wide Web Consortium (W3C). (2024). *Web Content Accessibility Guidelines 2.2*. https://www.w3.org/TR/WCAG22/

World Wide Web Consortium (W3C). (2025). *Understanding Success Criterion 3.1.5: Reading Level*. World Wide Web Consortium (W3C). https://www.w3.org/WAI/WCAG21/Understanding/reading-level?utm_source=chatgpt.com

World Wide Web Consortium (W3C). (2026, March 12). *Use Clear and Understandable Content*. World Wide Web Consortium (W3C). https://www.w3.org/WAI/WCAG2/supplemental/objectives/o3-clear-content/

Zhang, J., Qiu, X. Y., Lu, L., Huang, Z., Hu, Y., Wu, Y., & Lu, J. (2026). *Let's Simplify Step by Step: Guiding LLM Towards Multilingual Unsupervised Proficiency-Controlled Sentence Simplification* (arXiv:2602.07499). arXiv. https://doi.org/10.48550/arXiv.2602.07499