Automatic Text Simplification for French Medical Documents with LLMs: The Role of Target Audience and Genre

Rémi Cardon[♦*], **A. Seza Doğruöz**[♥]
♦Computer Science and Engineering Department, UC3M, Spain
♥LT3, IDLab, Universiteit Gent, Belgium
rcardon@inf.uc3m.es, as.dogruoz@ugent.be

Abstract

Medical information is hard for non-specialists to understand, despite its importance for treatment success. Automatic text simplification (ATS) rewrites complex documents into simpler versions, with effectiveness measured through ATS evaluation metrics and readability metrics. A key challenge in ATS is calibrating simplification to match the reading abilities of specific target audiences, as different populations have different comprehension needs. Since socio-demographic factors such as education level and health literacy are known to correlate with reading abilities, we hypothesize that large language models (LLMs) may be able to adjust their simplification strategies when provided with descriptions of target audiences. In this study, we investigate how LLMs simplify French medical documents when prompted with socio-demographic characteristics of target patients. We compare this approach with prompts based on language proficiency levels (CEFR) to determine whether LLMs respond differently to explicit proficiency levels versus implicit audience descriptions. Our experiments with five LLMs on three types of French medical documents show that CEFR prompts produce greater readability variation (particularly for Llama-3.1-8B), while socio-demographic factors yield more homogeneous outputs. Text genre also considerably impacts LLM outputs for ATS.

1. Introduction

The ability to understand medical information is correlated with the chances of success for a medical treatment (Berkman et al., 2011). However, the growing availability of online medical information does not yield an improvement in understanding of medical content by the general public (Parker et al., 1999).

Automatic text simplification (ATS) is a line of research in natural language processing (NLP) that develops tools and resources to make written texts easier to read for target audiences (Saggion, 2017; Alva-Manchego et al., 2020), for example by shortening sentences or replacing technical terms. To assess whether ATS systems reduce comprehension difficulty, researchers rely on both ATS-specific evaluation metrics (e.g., comparing n-gram operations and measuring meaning preservation) and readability metrics that quantify textual characteristics as proxies for reading ease. However, sociodemographic characteristics (e.g., age, education level, health literacy) of the target audience are rarely taken into account in ATS research (Gooding, 2022).

In terms of implementation, ATS methods have recently been explored with large language models (LLMs) (Kew et al., 2023), which use prompts to simplify texts and increase their readability. ATS researchers have explored adding readability information to prompts (Barayan et al., 2025). To build upon this research, we investigate how LLMs simplify medical texts in French when the prompt includes information about the target audience. More precisely, we focus on the effect of target audience description on readability when simplifying medical texts in French.

Our hypothesis is that LLMs encode representations linking socio-demographic factors to language skills, particularly reading abilities. If validated, this would enable more targeted simplification strategies tailored to specific patient populations through audience-specific prompts. Our contributions are as follows: (i) We conduct, to our knowledge, the first study to investigate how socio-demographic descriptions of target audiences as patients influence the simplified medical text output of LLMs for French documents; (ii) we present, to our knowledge, the first study on document-level ATS for French medical texts.

*Research done while employed at UGent.

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 164–180
11 May 2026.

2. Related Work

2.1. Automatic Text Simplification

ATS has traditionally been performed at the sentence level (Chandrasekar et al., 1996; Alva-Manchego et al., 2020), taking a sentence as input and simplifying it to convey the same meaning in an easier-to-read form. For example, "*Medication inhibiting the peristalsis are counter-indicated in this situation*" can be simplified to "*In this case, do not take medication for stopping or decreasing the intestinal transit*" (Grabar and Saggion, 2022).

For English, efforts have extended beyond the sentence level to paragraphs (Devaraj et al., 2021; Lu et al., 2023) and full documents (Cripwell et al., 2023a; Mo and Hu, 2024; Nagai et al., 2024). However, medical ATS in French remains limited and has only been explored at the sentence level (Cardon and Grabar, 2020; Todirascu et al., 2022).

Regarding methodologies, the field has transitioned from rule-based systems to generative models over time, incorporating statistical approaches and other machine learning techniques (Cardon and Bibal, 2023).

ATS and Target Audience Most ATS work approaches simplification as a one-size-fits-all task (Gooding, 2022), despite the fact that different target audiences have different needs (Rennes et al., 2022). This monolithic approach is reflected in evaluation protocols. Automatic metrics require annotated data, either as references (Xu et al., 2016; Papineni et al., 2002) or as training data (Maddela et al., 2023; Cripwell et al., 2023b). Such corpora are costly to create, and available corpora are typically not associated with specific target audiences.

Regarding human evaluation of ATS, it is mostly carried out by researchers, their colleagues, or crowd workers, which is problematic in terms of representativeness of the target population (Doğruöz et al., 2023). A few studies have directly involved members of target audiences in evaluation (Alonzo et al., 2020, 2022), but this remains rare.

Readability-controlled ATS Readability-controlled ATS focuses on leveraging readability information to adjust LLM outputs to target readability levels. Common readability measures include (1) the Flesch-Kincaid Grade Level (Kincaid et al., 1975, FKGL), a traditional formula for English readability that uses the total number of words, sentences, and syllables to output a grade corresponding to estimated readability; and (2) the Common European Framework of Reference for Languages (Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001, CEFR), which assesses language proficiency for second language learners using six levels (A1, A2, B1, B2, C1, C2) with descriptions of expected skills at each level.

Recent work has incorporated FKGL or CEFR levels in prompts for simplification (Imperial and Tayyar Madabushi, 2023; Barayan et al., 2025), though only for English and sentence-level simplification. Moreover, these approaches consider readability as a textual property, leaving aside the socio-demographic characteristics of the reader.

Evaluation To evaluate ATS performance for document-level simplification, Sun et al. (2021) introduce D-SARI, an adaptation of SARI (Xu et al., 2016, System output Against References and Input), originally developed for sentence-level ATS. SARI compares token n-grams that are added, deleted, and kept when simplifying from input to reference text against the same operations from input to output texts. D-SARI follows the same principles with adjustments to penalize large discrepancies in document length and sentence repetitions.

Mo and Hu (2024) assess documents using readability features including traditional metrics (e.g., FKGL) and other aspects such as type-token ratio and syntactic complexity. ATS research commonly uses BERTScore (Zhang et al., 2020) to evaluate meaning preservation (Alva-Manchego et al., 2021). BERTScore is a metric that relies on BERT embeddings to measure semantic similarity between texts.

Regarding human evaluation, sentence simplification is typically judged in terms of grammaticality or fluency (*is the output grammatical?*), simplicity (*is the output simpler than the input?*), and meaning preservation or adequacy (*is the original meaning preserved?*)

using a 5-point Likert scale. This evaluation approach is also used for document-level simplification (Sun et al., 2021), though implementation and interpretation details are not yet stabilized (Stodden, 2021).

2.2. LLMs and Target Audiences

There is a lack of research on whether LLMs can effectively take socio-demographic factors into account. When addressing socio-demographic groups and LLMs, current NLP research mostly focuses on biases. For example, Navigli et al. (2023); Nozza et al. (2022); Kotek et al. (2023); Gallegos et al. (2024) identify bias through sentence completion tasks, prompting LLMs with the beginning of a group definition and studying the representation the model outputs. This research focuses on semantic content rather than linguistic form.

We found only one study focusing on the links between target groups and writing style (Malik et al., 2024). They incorporate socio-demographic factors (location, occupation, political affiliation, and age) in prompts and analyze the writing style of models when instructed to impersonate someone with those factors. In this section, we have reviewed studies on how LLMs behave when producing text *about* or *as* members of given social groups. We have found no study on how LLMs produce text *for* specific social groups, which is the focus of our question.

3. Methodology

In this section, we describe the data we use (Section 3.1), the models we selected for our experiments (Section 3.2), the socio-demographic factors we incorporate into our study (Section 3.3), and the metrics we use to analyze the outputs (Section 3.4).

3.1. Data

For our experiments, we use the CLEAR corpus (Grabar and Cardon, 2018), a freely available French medical corpus compiled for ATS research. It includes document pairs on the same topic targeting different audiences, with three different document types:

1. **Cochrane Summaries**: These are summaries written by the Cochrane Foundation[1] for medical practitioners and manually simplified into plain language summaries (PLS). Each summary addresses a specific medical research question. These summaries have also been used in biomedical ATS for English (Devaraj et al., 2021). The summaries are available in 20 different languages, including French.[2]

2. **Drug Information**: For every drug introduced to the French market, manufacturers must publicly release information[3] in two forms: one for medical practitioners (RCP, *résumé des caractéristiques du produit*, summary of product characteristics) and one for patients (the paper leaflet provided in each medicine box).

3. **Encyclopedia Articles**: Articles from the medical section of French Wikipedia[4] and corresponding articles from French Vikidia,[5] a collaborative online encyclopedia written for children aged 8 to 13.

For our experiments, we randomly sample 100 document pairs (complex and simple versions) from each sub-corpus. Examples for each dataset are available in Appendix B.

3.2. Models

We experiment with five open-source LLMs of comparable size (7-9B parameters), which all support French and have instruction-following capabilities: `Llama-3.1-8B-Instruct`[6] (henceforth `Llama`), `DeepSeek-R1-Distill-Llama-8B`[7] (henceforth `LlamaDS`), a version of Llama

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

[2] https://cochranesupport.wiley.com/s/article/languages-supported-in-cochrane-library

[3] https://base-donnees-publique.medicaments.gouv.fr/

[4] https://fr.wikipedia.org/wiki/Portail:Médecine

[5] https://fr.vikidia.org/wiki/Portail:Médecine

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

3.1 fine-tuned on DeepSeek's outputs, `Qwen2.5-7B-Instruct`[8] (henceforth `Qwen`), `DeepSeek-R1-Distill-Qwen-7B`[9] (henceforth `QwenDS`), a version of Qwen 2.5 fine-tuned on DeepSeek's outputs, and `BioMistral-7B`[10] (Labrak et al., 2024), a biomedical model.

These models were selected based on several criteria: (1) comparable size to ensure fair comparison, (2) availability and reproducibility as open-source models, (3) documented performance on multilingual or French tasks, and (4) for BioMistral, specialization in the biomedical domain. For all experiments, we set the temperature to 0.01, ensuring that the same prompt yields consistent outputs. On average, each model took approximately 16 hours per corpus (approximately 48 hours total across all three corpora) to process 100 documents with the 13 different prompts (6 CEFR levels, 5 IPS-based factors, and 2 health literacy conditions) on an NVIDIA GeForce RTX 4090.

3.3. Socio-demographic Factors

The socio-demographic factors in our study are selected based on their documented relationship with the ability to access and process medical information, which we use as an approximation of reading abilities. We focus on the level of education and health literacy.

Level of Education To indicate the level of education (for the reader), we rely on data released by the French government. For each elementary school (*école primaire*), middle school (*collège*), and high school (*lycée*) in France, the French Ministry of Education publishes an indicator called IPS (*indice de position sociale*, social position index), which measures the social status of students (Dauphant et al., 2023). IPS is calculated using PCS (*profession et catégorie sociale*[11]), which maps 40 parental occupations to numeric codes. IPS is calculated for each student based on the PCS value assigned to their parent(s).[12] IPS values range from 45 (a student with an unqualified working mother and a student father) to 185 (a student with an engineer mother and a professor/scientist father). PCS codes have different weights depending on whether the parent is a mother or a father (Rocher, 2016). For example, a student with a single mother who is a teacher has an IPS of 154, while a student with a single father with the same occupation has an IPS of 146.[13] The publicly released values are average IPS per school type.

IPS is used to inform policy decisions regarding education in France. Schools with low average IPS receive additional public funding (e.g., higher teacher-to-student ratios, educational assistants). IPS values correlate with diploma success rates (e.g., baccalauréat) and grade progression, strengthening IPS as a relevant indicator for approximating students' abilities to access and process information. IPS values also correlate with high school type ("secteur" in French): public schools (free registration, only public funding) or private schools (paid registration, partial public funding).

Health Literacy Health literacy (HL) in France refers to patients' ability to access, understand, evaluate, and use information needed for healthcare (Rey et al., 2023). According to this study, 10% of the French population has low HL, rising to 33% among people with self-declared 'poor' or 'very poor' health. Additionally, HL increases with education level (i.e., the level of official diploma obtained). This evidence reinforces the relationship between reading abilities and socio-demographic factors of patient populations. The connection between HL and reading abilities is particularly relevant for ATS: patients with lower HL struggle not only with medical-specific content but also with the linguistic complexity of health documents. Self-reported health status serves as a proxy for HL in our study because the Rey et al. study demonstrates a strong correlation between poor health and low HL (33% vs. 10% in the general population).

[8] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[9] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[10] https://huggingface.co/BioMistral/BioMistral-7B

[11] occupation and social category

[12] https://www.education.gouv.fr/l-indice-de-position-sociale-ips-357755

[13] https://www.education.gouv.fr/media/158757/download

Factors Used in Prompts Building upon these studies on French education and health, we include the following social factors in our prompts:

- **High school students** (based on IPS studies):
  - High school type: *lycée public / privé* (public, private)
  - Household income level: *milieu défavorisé / moyen / favorisé* (low, average, high)
- **Adults** (based on HL studies):
  - Health condition: *mauvais / bon état de santé* (poor, good)

In addition, to facilitate comparison with recent ATS approaches (see Section 2), we include the six CEFR levels (A1, A2, B1, B2, C1, C2), a scale of language proficiency levels.

3.4. Analysis Metrics

To analyze the readability of texts in our corpus and the models' outputs, we use the following metrics.

Reading Ease Level (REL) We compute the Reading Ease Level (Kandel and Moles, 1958, REL), an adaptation of Flesch Reading Ease (Flesch, 1948) for French. A higher score indicates a more readable text. The formula is:

$$207 - 1.015 \times \frac{\text{Number of words}}{\text{Number of sentences}} - 73.6 \times \frac{\text{Number of syllables}}{\text{Number of words}} \quad (1)$$

Automatic Evaluation Metrics We use two automatic metrics: D-SARI[14], which measures simplification quality by comparing n-gram operations between input, output, and reference texts (Sun et al., 2021), and BERTScore[15] (using the `bert-base-multilingual-cased` model), which measures meaning preservation through semantic similarity (Zhang et al., 2020).

[14]Implementation: https://github.com/RLSNLP/Document-level-text-simplification

[15]Implementation: https://github.com/Tiiiger/bert_score

4. Results

We begin by examining the original CLEAR corpus texts in terms of readability characteristics and automatic metrics. We then analyze the outputs of our selected models, first providing an overview of their performance, then examining the impact of text genre and prompt factors on simplification outcomes.

4.1. Original Texts

We first examine the simplification strategies present in the human-written simplifications of the CLEAR corpus. Figure 1 shows the distribution of REL score differences between complex and simple text pairs, broken down by dataset.[16] Higher values indicate greater simplification according to REL.

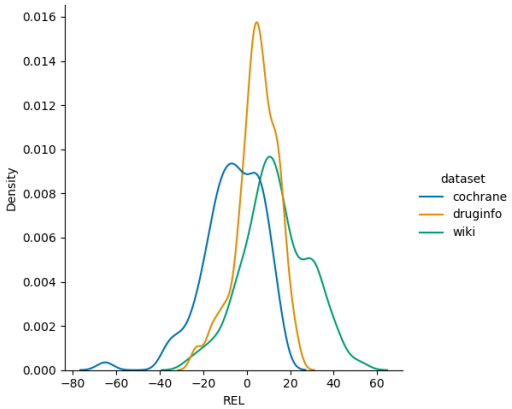


Figure 1: Difference in REL values between complex and simple versions of CLEAR subcorpora. Higher values indicate greater simplicity gain.

We observe that simplification strategies differ substantially between datasets. The Drug Information (Druginfo) dataset shows the least variation, with differences centered near zero. The encyclopedia dataset (Wiki) shows an increase in REL values (indicating successful simplification), whereas the Cochrane reviews show a decrease (indicating that the simplified versions have lower REL scores than the original texts). This suggests that different text

[16]All figures in this paper use `seaborn`'s `colorblind` color palette (https://seaborn.pydata.org/generated/seaborn.color_palette.html).

genres employ different simplification strategies, and that REL alone may not capture all aspects of simplification, particularly for specialized medical literature.

4.2. Model Simplifications

4.2.1. Overview and Data Cleaning

We plot the REL values by model across all datasets and prompts (Figure 2). While the lowest REL value in human-written texts is 4.95 (a Cochrane review on antibiotics for outpatient treatment), the models sometimes output texts with negative REL values. There are 160 such texts across the 19,500 generated texts: 97 produced by `BioMistral`, 51 by `QwenDS`, and 12 by `Llama`. These anomalous outputs are spread equally across all prompts.

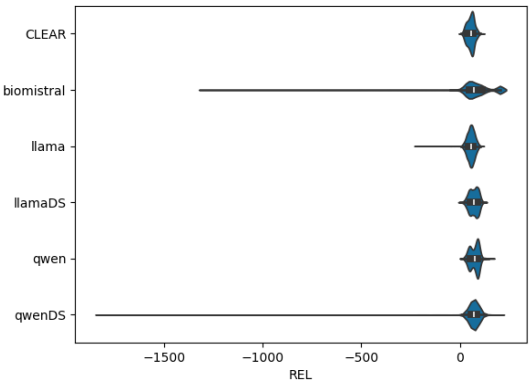


Figure 2: REL values by model and for the original simple documents (CLEAR), across all datasets and prompts.

The lowest REL value (-1,826.45) is produced by `QwenDS` on a simplification of the “Ectopia" Wikipedia article with the *poor health* prompt. The output ends with 5,496 consecutive characters composed of repetitions of non-French character sequences. Manual inspection reveals that texts with REL values below 4 contain similar anomalies (repetitions, formatting errors, or truncated outputs).

On the other hand, the highest REL value for human-written texts is 115.41 (a Vikidia article on the urinary meatus). The maximum value for LLM-generated texts (205.96) is obtained by `BioMistral` on 367 outputs, all consisting of a single number (digit 1 followed by zeros). Among the 847 texts with REL above 130, 799 are produced by `BioMistral` (20.49%), 28 by `QwenDS` (0.72%), and 20 by `Qwen` (0.51%). Notably, the `Qwen` models' high-REL outputs are partially or entirely in Chinese despite being prompted in French.

Based on these observations, we filter out texts with REL 4 or REL 130. The percentage of texts retained after filtering: `BioMistral` 76.84%, `Llama` 99.67%, `LlamaDS` 99.95%, `Qwen` 99.49%, and `QwenDS` 97.97%. As we identified almost 25% of anomalous texts from `BioMistral`, we exclude this model from subsequent analyses. The lower bound of 4 corresponds to the minimum REL observed in human-written texts; the upper bound of 130 was chosen with a conservative margin above the maximum human REL (115.41). Future work should investigate more principled threshold-selection methods.

4.2.2. Impact of Text Genre

Figure 3 shows the distribution of REL values after filtering, broken down by dataset. For all models, the range of REL values is wider than for the original simple documents. However, the genre of the original text clearly impacts the output. All models produce texts with lower REL values for the Cochrane dataset than for the other two. `Llama` shows the greatest proximity to human-written texts, while the REL distributions for the three other models appear more similar to each other.

Table 1 shows D-SARI and BERTScore by model and dataset. D-SARI scores align with those reported in recent document-level ATS work on English datasets (Fang et al., 2025; Bahrainian et al., 2024). Given that we operate in a zero-shot setting and that simplification strategies vary considerably across the three corpora (as shown in Figure 1), we interpret these scores as evidence that model outputs warrant further investigation despite the absence of corpus-specific fine-tuning.

In terms of meaning preservation (BERTScore), `LlamaDS` produces texts closest to the original across all datasets, even exceeding human simplifications for Cochrane and Wiki. D-SARI is not reported for human simplifications (CLEAR) because the simple documents in CLEAR serve as the reference texts in this metric, making

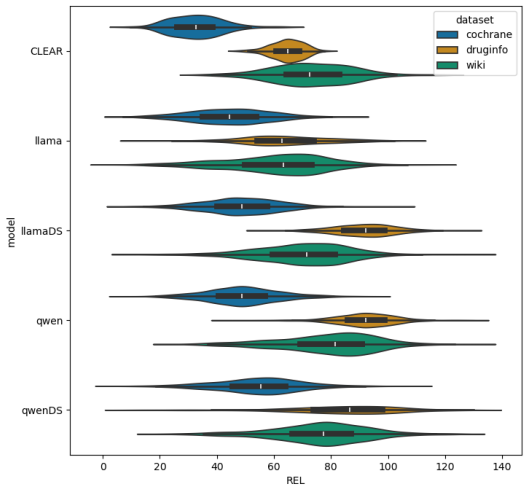


Figure 3: REL values between 4 and 130, by model and for the original simple documents (CLEAR), across all datasets and prompts.

a comparison of reference against itself undefined. Conversely, `Qwen` obtains the lowest BERTScore values. The Druginfo dataset proves most challenging for all models, though notably, human simplifications for this dataset retain meaning better than for the other two.

4.2.3. Impact of Factors in the Prompt

To examine the impact of prompt factors on outputs, we analyze REL and BERTScore values. Figure 4 (Appendix A) shows these values for each model across all datasets, with factors grouped by type: high school students (based on IPS studies), adults (based on health literacy studies), and language proficiency (CEFR levels).

CEFR levels have the most substantial influence on outputs. `Llama` shows the most pronounced progression from A1 to C2 for both REL and BERTScore, with A1 prompts producing the most readable texts (highest REL) and C2 prompts producing the least readable texts (lowest REL), consistent with the scale's intended progression. `QwenDS` is the only model that does not appear to have encoded the CEFR scale, showing no systematic variation across levels.

All models show minimal sensitivity to factors related to health condition (good vs. poor health). Similarly, income-related prompts (IPS) show little effect, though `Llama` produces texts with notably higher REL values for the *low income* factor compared to *average* and *high income*.

	D-SARI	BERTScore
CLEAR (human)		
Cochrane	-	73.75
Druginfo	-	76.60
Wiki	-	69.70
Llama		
Cochrane	**35.57**	73.14
Druginfo	29.51	60.84
Wiki	32.68	72.30
LlamaDS		
Cochrane	31.78	**79.83**
Druginfo	**31.01**	**65.14**
Wiki	30.49	**74.79**
Qwen		
Cochrane	34.07	68.40
Druginfo	28.51	60.91
Wiki	**33.92**	69.43
QwenDS		
Cochrane	30.74	77.38
Druginfo	29.91	61.73
Wiki	28.61	70.91

Table 1: D-SARI and BERTScore values for each model by dataset, across all prompts. Bold values indicate highest scores for each dataset.

5. Discussion

Our results highlight substantial variation in how models respond to prompt factors. `Llama` shows the greatest sensitivity, with outputs complying with traditional readability definitions when prompted with CEFR levels. Other models behave differently: their CEFR outputs show lower REL variation, and the DeepSeek-distilled models (`LlamaDS` and `QwenDS`) do not strictly follow the scale (e.g., *A1* produces lower REL values than *A2*).

Socio-demographic factors impact readability differently than CEFR levels. Models produce limited variation across factors like high school type, income level, health condition. However, there is one notable exception. The

low income prompt yields the highest REL values among socio-demographic factors for `Llama`. It is the only factor that correlates with both lower education and low health literacy (Dauphant et al., 2023; Rey et al., 2023). This finding suggests that the model has captured this relationship.

Conversely, *poor health* does not trigger highly readable outputs, instead producing the second-lowest REL values after *good health*. This suggests models' representations of socio-demographic factors and reading abilities may be incomplete or inconsistent.

5.1. Impact of Text Genre

Our results demonstrate that text genre has a considerable impact on model outputs. All models produce texts with lower REL values (indicating less readability according to this metric) for the Cochrane dataset compared to the other two datasets. This finding is particularly significant given that the human-written simplifications in the CLEAR corpus also show genre-specific patterns (Figure 1).

This genre effect carries important implications for ATS system design, suggesting that simplification strategies should adapt to both target audiences and source text genres. Even when targeting the same audience, different document types—medical literature reviews (Cochrane), drug information leaflets (Druginfo), and encyclopedia articles (Wiki)—may benefit from distinct simplification approaches. Future work should investigate whether explicitly incorporating genre information in prompts enhances simplification quality.

5.2. Model Comparison and Fine-tuning Effects

The differences between `Llama` and `Qwen`, and their respective DeepSeek-distilled counterparts (`LlamaDS` and `QwenDS`), highlight the significant impact of fine-tuning strategies on model capabilities for ATS. `QwenDS` shows no systematic variation when prompted with different factors, suggesting that the distillation process may have reduced the model's sensitivity to prompt nuances. This finding underscores the importance of careful evaluation when adopting fine-tuned or distilled models for tasks requiring prompt-based control.

`LlamaDS` achieves the highest BERTScore values across all datasets, indicating superior meaning preservation compared to other models and even surpassing human simplifications for some datasets. However, this comes at the cost of reduced readability variation in response to prompt factors. This trade-off between meaning preservation and controllable simplification is an important consideration for practical ATS systems.

6. Conclusion

We investigated how text genre and socio-demographic factors in prompts affect outputs when prompting LLMs for document-level ATS of French medical texts. `Llama-3.1-8B` shows the greatest sensitivity to prompt factors, while other models show less variation or inconsistent patterns.

For `Llama`, CEFR levels in prompts produce outputs that comply with traditional definitions of text complexity, following the expected progression from A1 (most readable) to C2 (least readable). Socio-demographic factors trigger different patterns, with the *low income* factor producing notably more readable outputs, consistent with documented correlations between socioeconomic status and reading abilities. Other socio-demographic factors (health condition, school type) show minimal effect.

We also demonstrate that text genre has a considerable impact on model outputs, with all models producing different readability characteristics for Cochrane reviews, drug information, and encyclopedia articles. This finding suggests that effective ATS systems should account for both target audience characteristics and source text genre.

Our findings represent a first step toward systematically incorporating user descriptions in LLM prompts for ATS. Future work should investigate additional socio-demographic factors, test interactions between multiple factors, diversify the array of models evaluated, and most importantly conduct validation studies with members of target populations to assess whether the observed readability variations translate to improved comprehension and usability.

7. Limitations

As our work explores novel research questions, the scope of our experiments is necessarily limited. We selected five models of comparable size to enable fair comparison, but focused detailed analysis on four models after excluding `BioMistral` due to a high rate of anomalous outputs. The rapid evolution of LLMs means that our findings may not generalize to newer or larger models.

Our study shares limitations common to ATS research. Little is known about the concrete effectiveness of different simplification strategies for specific target audiences. While we found readability variation depending on sociodemographic factors in prompts, we cannot confirm whether these variations are actually beneficial for the corresponding populations. This limitation is particularly significant given the unexpected behavior for some factors (e.g., *poor health* producing less readable outputs than *good health*).

We did not conduct human evaluation with members of target populations. Automatic metrics (REL, D-SARI, BERTScore) provide useful quantitative measures but do not capture all aspects of text comprehension and usability. For instance, REL is based primarily on sentence and word length, which may not fully reflect the complexity of medical terminology or conceptual difficulty. Conducting studies with actual target audiences represents important future work that will determine whether our quantified readability variations correspond to meaningful improvements in understanding.

Additionally, our analysis focused on readability metrics and did not examine other aspects of simplification quality, such as the preservation of medical accuracy, the appropriateness of added elaborations, or the completeness of information. The exclusion of linguistic criteria analysis (e.g., passive voice, sentence complexity patterns) limits our understanding of specific simplification strategies employed by different models.

8. Plain Language Summary

Understanding medical information can be difficult for people who are not medical experts. This is a problem because patients need to understand their health information to make good decisions about their care. One way to help is by using computer programs that rewrite complicated medical texts into simpler language. These programs are called automatic text simplification (ATS) systems.

In our study, we wanted to see if artificial intelligence (AI) models could make French medical documents easier to read for different groups of people. We tested whether these AI models could change the way they simplify text based on who the reader is. For example, we gave the AI information about the reader's education level or health background, or we told it what level of French the reader understands.

We used five different AI models and three types of French medical documents. We found that when we told the AI the reader's language skill level, the simplified texts varied more in how easy they were to read. When we described the reader's background (like education or health), the AI's simplified texts were more similar to each other. The type of medical document also made a difference in how the AI rewrote the text.

In summary, our research shows that AI can help make medical information easier to understand, and that the way we describe the intended reader affects how the AI rewrites the text. This could help create better health information for different groups of people.

9. Bibliographical References

Oliver Alonzo, Sooyeon Lee, Mounica Maddela, Wei Xu, and Matt Huenerfauth. 2022. A dataset of word-complexity judgements from deaf and hard-of-hearing adults for text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 119–124, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Oliver Alonzo, Matthew Seita, Abraham Glasser, and Matt Huenerfauth. 2020. Automatic text simplification tools for deaf and hard of hearing adults: Benefits of lexical simplification and providing users with au-

tonomy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Seyed Ali Bahrainian, Jonathan Dou, and Carsten Eickhoff. 2024. Text simplification via adaptive teaching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6574–6584, Bangkok, Thailand. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, and Karen Crotty. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*, 155(2):97–107.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Council of Europe. Council for Cultural Cooperation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023b. Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore. Association for Computational Linguistics.

Fannie Dauphant, Franck Evain, Marine Guillerm, Catherine Simon, and Thierry Rocher. 2023. L'indice de position sociale (ips): Un outil statistique pour décrire les inégalités sociales entre établissements. *Note d'information de la DEPP*, pages pp–1.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, and Zheng Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5751–5767, Singapore. Association for Computational Linguistics.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Liliane Kandel and Abraham Moles. 1958. Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, 19(1958):253–274.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated reliability index, fog count and flesch reading ease formula) for navy enlisted personnel (research branch report 8-75). memphis, tn: Naval air station; 1975. *Naval Technical Training, US Naval Air Station: Millington, TN*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains.

Junru Lu, Jiazheng Li, Byron Wallace, Yulan He, and Gabriele Pergola. 2023. NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1079–1091, Dubrovnik, Croatia. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Manuj Malik, Jing Jiang, and Kian Ming A. Chai. 2024. An empirical analysis of the writing styles of persona-assigned LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19369–19388, Miami, Florida, USA. Association for Computational Linguistics.

Kaijie Mo and Renfen Hu. 2024. ExpertEase: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.

Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. 2024. A document-level text simplification dataset for Japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 459–476, Torino, Italia. ELRA and ICCL.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ruth M Parker, Mark V Williams, Barry D Weiss, David W Baker, Terry C Davis, Cecilia C Doak, Leonard G Doak, Karen Hein, Cathy D Meade, Joann Nurss, et al. 1999. Health literacy-report of the council on scientific affairs. *Jama-Journal of the American Medical Association*, 281(6):552–557.

Evelina Rennes, Marina Santini, and Arne Jonsson. 2022. The Swedish simplification toolkit: – designed with target audiences in mind. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 31–38, Marseille, France. European Language Resources Association.

Sylvie Rey, Aude Leduc, Xavier Debussche, Laurent Rigal, V Ringa, V Costemalle, et al. 2023. Une personne sur dix éprouve des difficultés de compréhension de l'information médicale. *Études et résultats*, 1269(mai):8.

Thierry Rocher. 2016. Construction d'un indice de position sociale des élèves. *Éducation & formations*, (90):5–27.

H. Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Regina Stodden. 2021. When the scale is unclear–analysis of the interpretation of rating scales in human evaluation of text simplification. *Proceedings of the 1st Workshop on Current Trends in Text Simplification (CTTS 2021, co-located with SEPLN 2021)*.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

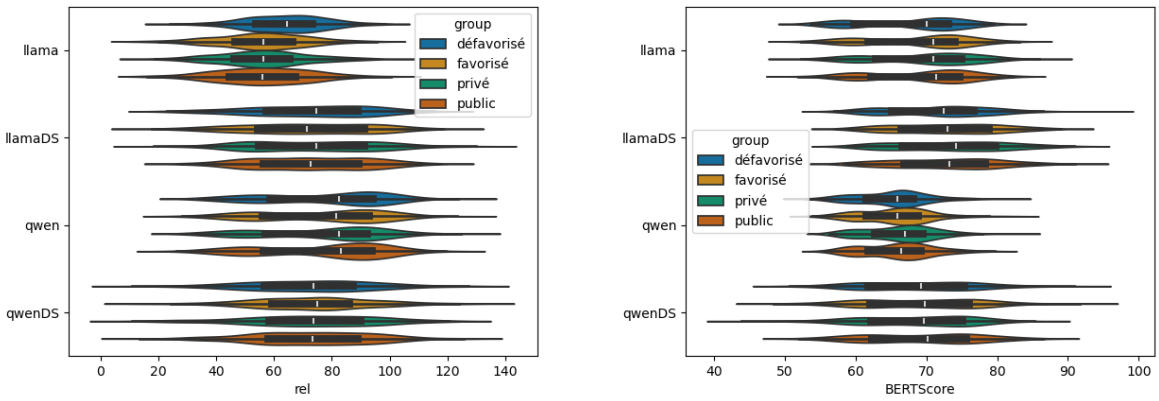
Amalia Todirascu, Rodrigo Wilkens, Eva Rolin, Thomas François, Delphine Bernhard, and Núria Gala. 2022. HECTOR: A hybrid TExt SimplifiCation TOol for raw texts in French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4620–4630, Marseille,

France. European Language Resources Association.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

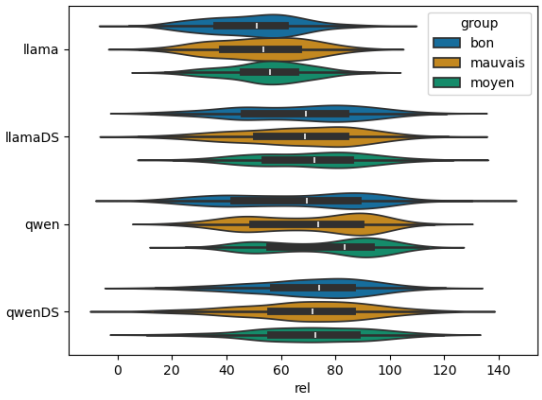
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A. Results by Factor Group

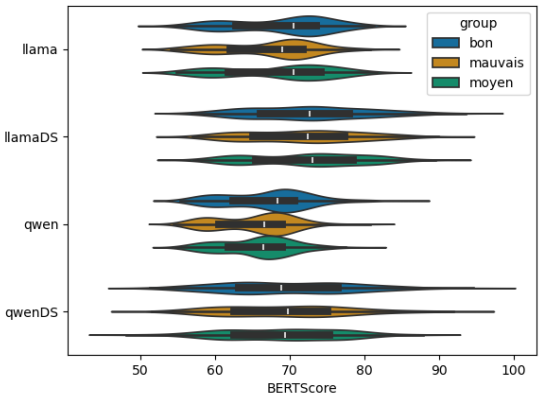


(a) IPS REL

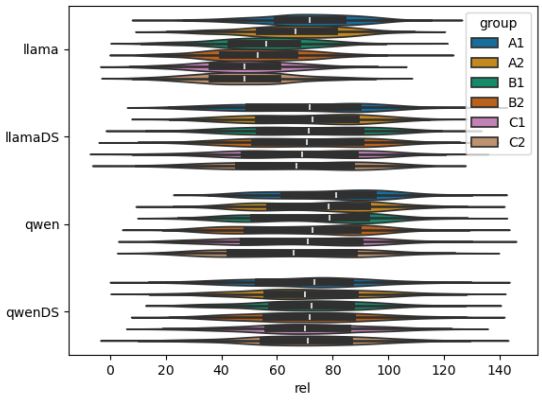
(b) IPS BERTScore



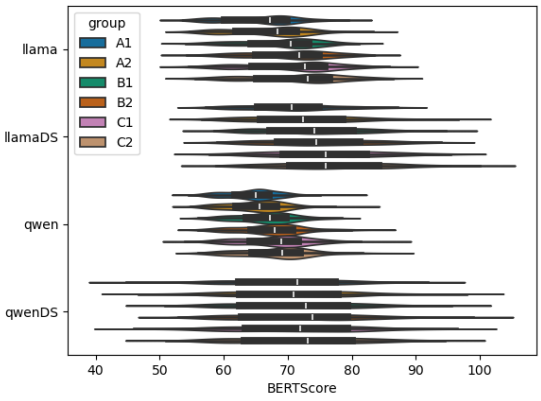
(c) HL REL



(d) HL BERTScore



(e) CEFR REL



(f) CEFR BERTScore

Figure 4: REL and BERTScore values for the different factors used in the prompts, by model.

B. Examples from CLEAR Corpus

Examples from each dataset type are provided in Tables 2, 3, and 4.

Cochrane review	
Literie en plume versus literie synthétique pour les patients asthmatiques. Contexte. Deux récentes études épidémiologiques rapportaient que les épisodes de respiration sifflante étaient plus fréquents chez les enfants utilisant des oreillers synthétiques que chez ceux utilisant des oreillers en plume. Objectifs. Évaluer l'efficacité de la literie en plume pour contrôler les symptômes de l'asthme. Stratégie de recherche documentaire. Le registre spécialisé du groupe Cochrane sur les voies respiratoires a été consulté en utilisant des termes prédéfinis. Les recherches étaient à jour en février 2009. Critères de sélection. Seuls les essais randomisés et les essais cliniques comparatifs étaient éligibles. Recueil et analyse des données. Aucun essai ne remplissait les critères d'inclusion dans la revue. Résultats principaux. La consultation de la littérature électronique a permis d'identifier 15 études en vue de l'examen de l'intégralité des articles. Après examen, aucune de ces études ne remplissait les critères d'inclusion dans la revue. Conclusions des auteurs. Bien que de récentes études épidémiologiques suggèrent que la literie en plume est associée à une respiration sifflante moins fréquente qu'avec des fibres synthétiques, les preuves actuellement disponibles sont insuffisantes pour évaluer les bénéfices cliniques de la literie en plume dans la prise en charge de l'asthme.	Feather versus synthetic bedding for patients with asthma. Background. Two recent epidemiological studies reported that wheezing episodes were more common in children using synthetic pillows than in those using feather pillows. Objectives. To assess the effectiveness of feather bedding in controlling asthma symptoms. Search strategy. The Cochrane Airways Group Specialised Register was searched using predefined terms. The searches were current to February 2009. Selection criteria. Only randomised trials and controlled clinical trials were eligible. Data collection and analysis. No trials met the inclusion criteria for the review. Main results. Electronic literature search identified 15 studies for full-text review. After review, none of these studies met the inclusion criteria for the review. Authors' conclusions. Although recent epidemiological studies suggest that feather bedding is associated with less frequent wheezing than synthetic fibres, the evidence currently available is insufficient to assess the clinical benefits of feather bedding in the management of asthma.
PLS	
Literie en plume versus literie synthétique pour les patients asthmatiques. Un allergène est une substance qui déclenche une réaction allergique chez les personnes qui y sont sensibles. Les acariens sont l'un des principaux allergènes de l'asthme. On pense que les oreillers et la literie contenant des fibres artificielles (fabriquées par l'homme) sont moins susceptibles d'accumuler des allergènes que les oreillers et les édredons en plume. Néanmoins, certaines preuves indiquent que la literie en plume pourrait, au contraire, être moins susceptible de causer de l'asthme. Cette revue n'a identifié aucun essai comparant les plumes aux fibres synthétiques et des recherches sont nécessaires afin d'établir le type de literie le mieux adapté aux patients asthmatiques.	Feather versus synthetic bedding for asthma patients. An allergen is a substance that triggers an allergic reaction in people who are sensitive to it. Dust mites are one of the main allergens in asthma. Pillows and bedding containing artificial (man-made) fibres are thought to be less likely to accumulate allergens than feather pillows and duvets. However, there is some evidence that feather bedding may be less likely to cause asthma. This review did not identify any trials comparing feathers with synthetic fibres and research is needed to establish which type of bedding is best for asthma patients.

Table 2: Example of a Cochrane review and its corresponding PLS, with English translations (done with Google Translate).

RCP	
4.3. Contre-indications Ce médicament est contre-indiqué en cas d'hypersensibilité à l'un des constituants. 4.4. Mises en garde spéciales et précautions d'emploi Le traitement par cet élément minéral trace ne dispense pas d'un traitement spécifique éventuel. Ce médicament contient du lactose. Son utilisation est déconseillée chez les patients présentant une intolérance au galactose, un déficit en lactase de Lapp ou un syndrome de malabsorption du glucose ou du galactose (maladies héréditaires rares). 4.5. Interactions avec d'autres médicaments et autres formes d'interactions Les données disponibles à ce jour ne laissent pas supposer l'existence d'interactions cliniquement significatives. 4.6. Grossesse et allaitement En l'absence de données expérimentales et cliniques et par mesure de précaution, l'utilisation de ce médicament est à éviter pendant la grossesse et l'allaitement.	4.3. Contraindications This medication is contraindicated in cases of hypersensitivity to any of the ingredients. 4.4. Special warnings and precautions for use Treatment with this trace mineral does not replace the need for specific treatment. This medication contains lactose. Its use is not recommended in patients with galactose intolerance, the Lapp lactase deficiency, or glucose-galactose malabsorption (rare hereditary diseases). 4.5. Interactions with other medicinal products and other forms of interaction The data available to date do not suggest the existence of clinically significant interactions. 4.6. Pregnancy and breastfeeding In the absence of experimental and clinical data, and as a precautionary measure, the use of this medication should be avoided during pregnancy and breastfeeding.
Leaflet	
Contre-indications Ne prenez jamais OLIGOSTIM COBALT, comprimé dans le cas suivant: · antécédent d'allergie à l'un des constituants. EN CAS DE DOUTE, IL EST INDISPENSABLE DE DEMANDER L'AVIS DE VOTRE MEDECIN OU DE VOTRE PHARMACIEN. Précautions d'emploi ; mises en garde spéciales Faites attention avec OLIGOSTIM COBALT, comprimé: Mises en garde spéciales Le traitement par cet élément minéral trace ne dispense pas d'un traitement spécifique éventuel. L'utilisation de ce médicament est déconseillée chez les patients présentant une intolérance au galactose, un déficit en lactase de Lapp ou un syndrome de malabsorption du glucose ou du galactose (maladies héréditaires rares). Précautions d'emploi EN CAS DE DOUTE NE PAS HESITER A DEMANDER L'AVIS DE VOTRE MEDECIN OU DE VOTRE PHARMACIEN. Interactions avec d'autres médicaments Prise ou utilisation d'autres médicaments: Si vous prenez ou avez pris récemment un autre médicament, y compris un médicament obtenu sans ordonnance, parlez-en à votre médecin ou à votre pharmacien.	Contraindications Never take OLIGOSTIM COBALT tablets in the following cases: · a history of allergy to any of the ingredients. IF IN DOUBT, IT IS ESSENTIAL TO ASK YOUR DOCTOR OR PHARMACIST FOR ADVICE. Precautions for use; special warnings Take special care with OLIGOSTIM COBALT tablets: Special warnings Treatment with this trace mineral does not replace the need for specific treatment. The use of this medicine is not recommended for patients with galactose intolerance, the Lapp lactase deficiency, or glucose-galactose malabsorption (rare hereditary diseases). Precautions for use IF IN DOUBT, DO NOT HESITATE TO ASK YOUR DOCTOR OR PHARMACIST FOR ADVICE. Interactions with other medications Taking or using other medications: If you are taking or have recently taken any other medication, including medication obtained without a prescription, talk to your doctor or pharmacist.

Table 3: Extracts of an RCP and its corresponding leaflet, with English translations (done with Google Translate).

Wikipedia article	
Les maladies cardio-vasculaires (ou maladies cardiovasculaires) sont les maladies qui concernent le cœur et la circulation sanguine. Dans les pays occidentaux, l'expression la plus courante est la maladie coronarienne, responsable de l'angine de poitrine ou encore des infarctus. Ces maladies touchent plus certaines catégories de population (ouvriers, personnes exposées à certaines pollutions, victimes d'obésité, etc.) et leur prévalence régionale est marquée (par exemple en France, à la fin du XXe siècle dans le Nord-Pas-de-Calais et en Alsace, deux régions nettement plus touchées que les autres régions et la moyenne nationale, comme pour plusieurs types de cancers)[1]. Elles comptent souvent parmi les facteurs qui diminuent le plus l'espérance de vie d'une population et semblent être un facteur de risque de dépression (chez les jeunes filles au moins[2]).	Cardiovascular diseases (or cardiovascular diseases) are diseases that affect the heart and blood circulation. In Western countries, the most common term is coronary artery disease, responsible for angina or heart attacks. These diseases affect certain population groups more (workers, people exposed to certain types of pollution, victims of obesity, etc.), and their regional prevalence is marked (for example, in France, at the end of the 20th century in Nord-Pas-de-Calais and Alsace, two regions significantly more affected than other regions and the national average, as is the case for several types of cancer)[1]. They are often among the factors that most reduce a population's life expectancy and appear to be a risk factor for depression (at least among young women[2]).
Vikidia article	
Une maladie cardio-vasculaire (ou cardiovasculaire) est une maladie qui touche le fonctionnement du cœur et de la circulation sanguine. Selon l'OMS, les maladies cardio-vasculaires représentent 29 % de la mortalité totale, ce qui en fait la première cause de mortalité dans le monde1. Les principales causes des maladies cardio-vasculaires sont le tabagisme, une mauvaise alimentation pas assez variée, et un manque d'activité sportive1. Faire suffisamment d'exercice chaque jour statistiquement divise le risque d'AVC et d'accident cardiaque par deux à tout âge.	Cardiovascular disease (or cardiovascular disease) is a condition that affects the functioning of the heart and blood circulation. According to the WHO, cardiovascular disease accounts for 29% of all deaths, making it the leading cause of death worldwide. The main causes of cardiovascular disease are smoking, an unhealthy and insufficiently varied diet, and a lack of physical activity. Getting enough exercise every day statistically halves the risk of stroke and cardiac arrest at any age.

Table 4: Extract of a Wikipedia article and its corresponding Vikidia article, with English translations (done with Google Translate).