Clearer Governmental Communication: Text Simplification with ChatGPT Evaluated through Quantitative and Qualitative Research

Nadine Beks van Raaij, Ksenia Podoynitsyna, Daan Kolkman
Jheronimus Academy of Data Science, Vlerick Business School, Utrecht University
Den Bosch - Eindhoven - Tilburg, The Netherlands, Vlerick, Belgium, Utrecht, The Netherlands
nadine.v.raaij (at) gmail.com, ksenia.podoynitsyna (at) vlerick.com, d.a.kolkman (at) uu.nl

Abstract

This study explores the use of ChatGPT for simplifying Dutch government letters to improve their comprehensibility while preserving legal accuracy. We employed a three-stage mixed-methods evaluation approach to assess the effectiveness of a naive baseline, RoBERTa, and ChatGPT in simplifying six of the most complex letters selected from a corpus of 200. The evaluation process involved comparing the outputs using four metrics (ROUGE, BLEU, BLEURT, and LiNT), followed by reviews from legal and linguistic experts, and culminating in a randomized controlled trial with 72 participants to test comprehension. Our results indicate that ChatGPT substantially enhances the comprehension of government letters, evidenced by more than a 20% increase in comprehensibility scores and a 19% improvement in participants' ability to correctly answer questions related to follow-up actions based on the simplified texts. Additionally, our study underscores the importance of a thorough evaluation framework and advises caution in solely depending on automated metrics for assessing text simplification.

Keywords: natural language generation, text simplification, ChatGPT 3.5, prompt engineering, legal documents, real-life task, human evaluation

1. Introduction

Text simplification (TS), a Natural Language Processing (NLP) task, aims to enhance readability and comprehensibility while retaining the essence of the text (Alva-Manchego and Shardlow, 2022; Al-Thanyyan and Azmi, 2021; Shardlow, 2014). TS can help diverse audiences, from people with disabilities (Carroll et al., 1998) and non-native speakers (Stajner, 2021) to those with limited literacy (Belder et al., 2010) by ensuring text accessibility and comprehension.

The value of TS is particularly apparent in government communication. Clear communication from government bodies is vital for promoting transparency, fostering civic engagement, and facilitating informed participation (Renkema, 2013; Lentz and Pander Maat, 2011; Sanders and Jansen, 2011; Kraf and Pander Maat, 2009). Yet, many governments, including that of the Netherlands, grapple with comprehensible communication (Pander Maat and van der Geest, 2021; Lentz et al., 2017). Recent episodes in the Netherlands underscore the challenge of government communications (Amnesty, 2021), with studies such as Pander Maat and van der Geest (2021) pinpointing issues in the comprehensibility of government letters.

Recognising these challenges, the Dutch government has taken proactive steps by enlisting communication experts to revise letters to citizens (Gebruiker-Centraal, 2022) and experimenting with NLP solutions (Rijksoverheid, 2023). Exploratory work by Feng et al. (2023) and Jeblick et al. (2022) demonstrates the potential of ChatGPT for TS on several benchmark datasets and radiology reports respectively. Motivated by these developments, our paper considers the question:

> *To what extent can large language models (LLMs) improve the comprehensibility of Dutch letters sent by governmental organisations?*

We answer this question by investigating empirically three approaches to TS: a naive token-substitution model, RoBERTa (Robustly Optimized BERT Pretraining Approach), and ChatGPT. We do so by a three-step mixed-method evaluation procedure which involves: 1. A comparison of evaluation metrics (ROUGE, BLEU, BLEURT, and LiNT); 2. Qualitative assessment by a legal and linguistic expert; 3. A randomized controlled trial with 72 participants. We demonstrate the importance of a robust evaluation procedure and find that TS using ChatGPT improves the comprehensibility of Dutch letters by 20%. Since ChatGPT 3.5 and 4 can handle multiple languages (Feng et al., 2023) our results have relevance for TS at large.

2. Related work

Although alternatives such as Bidirectional Encoder Representations from Transformers (BERT) exist, Generative Pre-trained Transformer (GPT) models typically outperform these alternatives (Tan and Kieuvongngam, 2020; Eisele, 2019), which is why we set out to explore GPT models in this study. The

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 152–178

value of this architecture has been demonstrated in relation to language learning (Young and Shishido, 2023; Luo et al., 2023) and TS of medical reports (Lyu et al., 2023; Holmes et al., 2023; Jeblick et al., 2022).
Suha and Azmi (2021) provide an overview of the past research for multiple languages in the field of TS and conclude that Data-driven simplifications outperform Rule-based simplifications. Furthermore, Suha and Azmi (2021) highlight the need for further research in developing new simplification techniques and reliable evaluation methods. Therefore, this research contributes to the research of performing a hybrid evaluation.

2.1. Prompt engineering ChatGPT

The quality of prompts provided to GPTs deeply impacts their outputs, which is why others have focused on prompt engineering for TS Feng et al. (2023); Holmes et al. (2023); Lyu et al. (2023); Engelmann et al. (2023). One recommendation of these studies is to process texts one by one instead of providing multiple texts at once as input for ChatGPT to avoid model hallucinations. Therefore, in this study, we chose to focus on one letter per prompt or a related set of prompts.
In addition, these studies use prompts that explicitly ask the model to "retain the content" and mention the original author's role or intended audience in the prompt to provide extra context. Often they also provide a dataset with example classifications of difficult/complex words/texts or offer example simplifications. These studies do not delve deeper into the methodology behind the generation of these few-shot/one-shot/zero-shot prompts or comparisons of different prompts that aim for the same audience and purpose. Holmes et al. (2023); Lyu et al. (2023) show the success of TS in a medical context for different audiences having differences in education level. Others have ventured to transform texts to particular readability levels in an effort to produce educational material for language students (Young and Shishido, 2023; Alkaldi and Inkpen, 2023). However, readability and comprehensibility are not the same[1] and without labeled texts, performing these simplifications is challenging.
This study employs prompt engineering for a single audience, citizens, who do not all have the same legal background or expert knowledge and should therefore receive plain language from governmental organizations. We follow up on the best practices of the above-named studies.
Our main focus is increasing the comprehensibility of the letters in practice. The prompts we used do not contain specifications about what is complex and what constitutes an example simplification. This is because there is a gap between what should be easy to comprehend and what actually is easy to comprehend for the majority of people. Therefore, we validate our results by focusing on the evaluation by the actual readers (through the randomized controlled trial) instead of prompting an automatic evaluation metric based on assigned examples that should be easy to comprehend or difficult to comprehend.

2.2. Automatic evaluation metrics

We use four quantitative evaluation metrics that align with established evaluation methods for automatic text summarization:

2.2.1. ROUGE

In a comprehensive review of automatic text summarization by Yadav et al. 2022 Recall-Oriented Understudy for Gisting Evaluation (ROUGE) was used. Additionally, Offerijns et al. 2020 and Gao et al. 2019 employed BLEU alongside ROUGE, enriching assessment with precision and recall considerations. Building on these foundations, this research also employs the ROUGE metric, which evaluates summarization and translation quality using scores ranging from 0 to 1, wherein higher values signify enhanced summarization or translation proficiency.

2.2.2. BLEU

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) is the second evaluation metric used in this research. BLEU is a popular automatic evaluation metric used to assess the quality of machine-translation output. It compares a machine-generated translation with one or more human reference translations and assigns a score based on how similar they are. The score ranges from 0 to 1, with 1 indicating a perfect match between the machine translation and the human reference translation.
This metric is not without problems for different text generation tasks. BLEU is not well suited, for example, for assessing simplicity from a lexical nor a structural point of view (Sulem et al., 2018). These findings indicated a weak or nonexistent correlation between BLEU and parameters related to grammaticality and meaning preservation in cases where sentence splitting is involved. Additionally, Sulem et al. (2018) found that BLEU tends to have a negative correlation with simplicity, which penalises

[1] Readability pertains to how easily a text can be read, often assessed through factors like sentence and word length (Dols, 2018; Lentz et al., 2017; Pander Maat and Dekker, 2016; Renkema, 2011). Comprehensibility relates to how well a reader can grasp a text's meaning, influenced by factors like idea complexity, text structure, and vocabulary difficulty. Comprehensibility ensures a text is not only easy to read but also easy to understand (Lentz et al., 2017; P., 2012; Renkema, 2011).

simpler sentences. They demonstrated this, via a created corpus for sentence splitting, containing multiple paraphrases, and compared it to human judgements. However, TS does not only rely on sentence splitting and other simplification studies (Xu et al., 2016; Stajner et al., 2014) have shown that it correlates with human judgements of grammaticality and meaning preservation. Therefore further research into BLEU was performed. Furthermore, BLEU was included in the end to create a benchmark for the automatic evaluation metrics. By comparing the scores of BLEU with BLEURT, the scores of the BLEURT become more valuable.

2.2.3. BLEURT

Incorporating BLEU-based Learned Evaluation for Text (BLEURT) enriches the evaluation strategy of this study. BLEURT evaluates text quality by gauging the correspondence between generated content and human assessments. Unlike BLEU, which primarily examines n-gram overlap, BLEURT delves into semantic alignment, enhancing its assessment accuracy. With a score range of -1 to 1, higher BLEURT scores signify superior performance. This approach is further validated by its alignment with human judgement, considering both surface-level and semantic similarity (Dipanjan and Parikh, 2020).

2.2.4. LiNT

Leesbaarheidsinstrument voor Nederlandse Teksten (LiNT) is the first evaluation metric that is used to evaluate the text on readability. LiNT was chosen as previous research proved this metric to be the most reliable Dutch metric to evaluate text on readability (Lentz, 2021; Pander Maat and Dekker, 2016; Kraf et al., 2011; Kraf and Pander Maat, 2009). LiNT makes calculations about the sentence structure characteristics and word characteristics and summarises this in a formula that is based on the T-scan (Pander Maat and Dekker, 2016) and outputs a LiNT score ranging from 1 to 100 (1 is the easiest, 100 the most difficult). Furthermore, LiNT categorises these scores into four levels: level one is the easiest, and level four is the most difficult. Level one holds for scores up to 36, between 36 and 51 the level is two, between 51 and 61.5 the level is three, and above 61.5 the level is four. The meaning of these levels according to Pander Maat and Ditewig 2017 is that with level one, 14% of the adult readers in the Netherlands and Flanders do not understand the text. For level two this is 30%, for level three this is 52% and for level four this is 80%.

2.3. Qualitative research with human evaluations

The cited studies underscore the importance of evaluating text summarization and simplification through a combination of quantitative and qualitative methods. Iskender et al. (2021) highlight the importance of exploring the reliability of human evaluations for text summarizations by analyzing the evaluators' characteristics. Furthermore, factors such as lexical and syntactic changes, and comprehensibility dimensions should be addressed. Notably, Nguyen et al. (2021) employ ROUGE for quantitative assessment and involve experts for qualitative evaluation, while Gosens (2008) conducted a qualitative study considering reader comprehension and analysed the results by means and standard deviation and made a comparison between the original and adjusted texts. Sikkema et al. (2017) explored comprehensibility dimensions in debt collection letters with education levels and letter volume as influential factors. Other related studies (Dols, 2018; Lentz et al., 2017; Renkema, 2011) also contribute insights into text evaluation methodologies. This research adapts best practices from previous studies, employing an expert review and a randomized controlled trial to evaluate three letters each and validate the results with regression analyses, aligning with established recommendations (Roobaea and Mayhew, 2014; Molich, 2010; Macefield, 2009; Hertzog, 2008; Faulkner, 2003).

3. Experiment

We conduct our experiment with three models: 1. a naive model that substitutes jargon with a simple explanation; 2. RoBERTa (Robustly Optimized BERT Pretraining Approach) that was finetuned with the same jargon-definition list as the naive model; 3. ChatGPT (based on GPT-3.5-turbo) with prompt engineering.

For ChatGPT, four different chats for every letter of the test data were used.[2] The simplified texts were pasted into a Word file and saved separately per letter. All the letters were manually checked for spelling and grammar mistakes. The generated letters by the naive model contained one spelling mistake which is explained further in section 4.1.2. The results of RoBERTa contain many grammatically incorrect sentences which are also further elaborated in section 4.1.2. No spelling or grammar mistakes were found in the letters simplified by ChatGPT. Furthermore, we checked if all the letters included the same contact details and, in case of a deviation, this has been adjusted to the original. Lastly, the layout of all the letters has been made equal meaning white spaces are added or deleted to comply with the layout of the original letters. This was done because previous research has shown that the layout influences the comprehension and interpretation of letters (Dols, 2018).[3]

[2]The questions asked in every chat for ChatGPT can be found in appendix B.

[3]The letters used for this research can be found in

Corsius et al. (2023) introduced a dataset of 200 letters (100 on Finance and 100 on Care) originating from multiple governmental organizations spread over the Netherlands. On average the length of these letters is 627 words. Corsius et al. (2023) identified six letters (3 on Finance and 3 on Care) that were hardest to comprehend. We use these six letters in the first stage of our evaluation procedure in which we compare the quantitative evaluation metrics (ROUGE, BLEU, BLEURT, and LiNT) for the three different TS approaches.

Given the often unreliable results of the evaluation metrics (Engelmann et al., 2023), the second stage of our evaluation procedure involves experts. More specifically, the outputs of the three models are evaluated by a legal and linguistic expert using the CCC-model (Lentz and Elling, 2003). The CCC-model is a framework for text evaluation that stands for Correspondence, Consistency, and Correctness, and that needs to be applied across five levels: text type, content, structure, formulation, and presentation. The experts discussed each simplified letter using this framework. Special consideration was given to the degree to which the simplified letters were equivalent from a legal perspective.

In the third and final stage of our evaluation procedure, we conducted a randomized controlled trial with 72 participants who all read three letters (in a random combination of original and simplified versions for each of these letters). As a result, we have 216 observations on the reader-letter level. This sample size is in line with recommendations by Fritz et al. (2012); Hertzog (2008).

Seventy-two participants were recruited online through convenience sampling. Participants were required to be at least 18 years old with a basic understanding of Dutch.[4]

As opposed to previous research on the comprehensibility of Dutch governmental texts (Corsius et al., 2023; Dols, 2018) this research distinguishes different reader' characteristics that influence the interpretation of comprehension. The participants were randomly divided into eight groups, each reading different combinations of letters in the same order. Figure 1 shows the distribution of the groups per education level. We follow a procedure where participants read three letters and answered questions about their content (understanding questions[5]), effectiveness (action questions[6]) and tone (tone questions[7]). The questionnaires were created by the linguist and legal expert and follow the guidelines of Grusky et al. (2018) and literature by Cox and Brayton (2008).[8]

4. Results

4.1. Automatic evaluation metrics

In this first stage of our evaluation, we find that both the naive approach and RoBERTa attain decent results (based on the automatic evaluation metrics[9]), while ChatGPT scores are less impressive.

4.1.1. Original letters

The LiNT score for the original letters was calculated to give an indication of the difficulty level. Five of the six original letters have a LiNT score between 36 and 51, indicating that 30% of adult readers in the Netherlands do not understand these letters. As these scores show, the letters in the theme Care are more difficult compared to the letters in the theme Finance. We will use these three letters as a critical case study in the randomized controlled trial.

4.1.2. Simplified letters

Interestingly the naive model has higher LiNT scores than the original letters, except for the $Regels\_pgb$ letter where the naive model scored 44 and the original 47. This indicates that the naive model decreased the readability. However, the LiNT scores did differ at most 4 points from the original letters and did have the same level categorisations, meaning that the difference is only minor. Looking at the other metrics, the naive model had high scores for both precision and recall. The BLEU and ROUGE scores are close to one for the naive model. This is to be expected from the fact that the ROUGE, BLEU, and BLEURT scores take the original letters as references and the naive model does not change any sentence structures or grammatical aspects. The BLEU scores decrease when the n-grams increase. This is logical as the naive model substitutes words or small parts of a sentence meaning that there is the smallest difference on the 1-gram level, and the biggest difference (lowest similarity) on the 4-gram level. However, these scores are still close to one, indicating a high similarity.

The RoBERTa model achieved the lowest LiNT scores and was able to get all letters categorised in level one. The lowest score was achieved for the letter $Betalen\_in\_delen$ with 26 points. The highest LiNT score of the RoBERTa model, being 32,

appendix A.

[4]Participants' characteristics such as education level and reading habits can be found in appendix F.

[5]Questions to test if the reader correctly understood what was meant with certain terms and statements.

[6]Questions to test if the reader knew which steps to take or what actions to do in certain situations or when encountering problems.

[7]Questions regarding the interpretation and tone of the text.

[8]The full questionnaires of the randomized controlled trial can be found in appendix D.

[9]The results of the automatic evaluation metrics of these models can be found in appendix E.

was achieved for the letter $Regels\_pgb$. These two letters are also marked as the simplest and most difficult letters based on the scores of the original letters. RoBERTa model thus scores considerably lower for the LiNT metric compared to the original letters. For the BLEU and ROUGE metrics we can see lower scores compared to the naive model. Regarding the BLEURT score, RoBERTa model achieved the lowest scores and seems to have only limited similarity with the original letters.

For four of the six letters, ChatGPT scored significantly higher compared to the original letter for the LiNT metric. This indicates that ChatGPT transformed the original letters to letters that are harder to read. For the other two letters ($Regels\_pgb$ and $Gemeentelijke\_belastingen$), ChatGPT scored lower compared to the original letter. Remarkable for these two letters is that they have the highest (BLEU_1 = 0.79, ROUGE_1 = 0.76) and lowest (BLEU_1 = 0.32, ROUGE_1 = 0.58) BLEU and ROUGE scores. This could imply that the LiNT metric encountered difficulties in evaluating these letters with the result that the scores differ from the others.

Regarding the BLEURT metric, ChatGPT scores range from 0.46 to 0.75. From these results, it seems that ChatGPT is able to simplify the letters while retaining the structure of the original letters. Taking this evidence together with the results of the expert review and the randomized controlled trial results, we can conclude that the technical metrics results should be treated with caution when evaluating the results of the TS task.

4.2. Expert review

The recommendations of the research of Cramwinckel 2014 together with the juridical background of the legal expert have been used as a guideline for the evaluation of the simplified letters in terms of juridical correctness. Below a summary of the experts' review is given.

The experts observe the letters simplified by the naive model are almost identical to the original letters. This is a result of too few words occurring in the original letters that were in the definition list of the naive model. Therefore the naive model did not find enough words to replace, which resulted in identical letters except 3 words per letter on average. Furthermore, the naive model replaced subwords which are part of a longer word. In instance where the full word is not included in the definition list, replacement of subwords results in linguistically incorrect sentences. An example is the original word "mogelijk" (possible) where "gelijk" (equal) was found in the definition list and had a definition of "nu" now). The original word was replaced by "monu", which is not a Dutch word. Therefore it was concluded that the naive model did not give the aimed simplifications and was not evaluated further.

The experts also evaluated the simplifications of the RoBERTa model. It was concluded that this model simplified the letters too much, with the result that the meaning of the text was gone. An example of an oversimplification is the original word "besluit" (decision) which was simplified to "antwoord" (response) by RoBERTa. This is neither linguistically nor juridically correct. Therefore we decided not to evaluate this model any further.

From the simplifications of ChatGPT, the linguistic expert observed that they have a shorter syntactic dependency length (SDL) compared to the original letters, which makes them easier to read. This is in line with Kleijn et al. 2016 who proved in their research that shorter SDL results in shorter processing times and positively affects the understanding of texts. Furthermore, the linguist expert concluded that the simplified texts were linguistically correct. The legal expert concluded that the important juridical information of the original letters was present in the simplified letters too. In sum, the simplifications of ChatGPT were considered sufficient in terms of linguistic and juridical correctness and were further evaluated with the randomized controlled trial.

4.3. Prompt engineering ChatGPT

Based on the results of the first two stages of our evaluation procedure, we refined the prompts.

In the first attempt, ChatGPT's ability to simplify text to Common European Framework of Reference for Languages (CEFR) levels was explored. The output was then classified by "Klinkende taal" (as Kraf et al. (2011) concluded this software performed the best for this classification task) and the experts to a CEFR language level. However, due to the contextual complexity of governmental letters, accurately determining the language level proved challenging. This aligns with prior research (Suha and Azmi, 2021) suggesting that CEFR may not be suitable for texts with specialized content, leading to the exclusion of this approach.

An effort was made to enrich ChatGPT's vocabulary and improve simplification quality by providing additional input based on a jargon definition list by Gebruiker-Centraal (2022). The jargon of this definition list did occur only limited in the tested letters and had very general explanations according to the linguistic expert. Although ChatGPT didn't directly utilize these definitions for simplification, it aided in detecting and avoiding difficult jargon. As this approach didn't contribute significantly, it wasn't included in the final prompt version.

Furthermore, a comparison was made between simple prompts and combinations of prompts consistent with the "chain-of-thought" approach, with various questioning approaches tested. All comply-

ing with the best practices of Madaan et al. (2023); Yu et al. (2023). Asking for "comprehension" rather than "simplification" yielded better results, avoiding over-simplification and information loss. The choice between "simple" and "easy" phrasing did not substantially impact outcomes. Among the different questions, using a few-shot approach consistently produced improved simplifications. Based on this, the final version of the ChatGPT prompt utilized the one-shot approach for enhanced simplification.

4.3.1. Final version: from bullet points to an easy text

From the few-shot attempts, it became clear that when was asked to rewrite the text to bullet points, all the important information was included. Since this was one of the problems with the earlier simplifications, we gave ChatGPT a prompt to first rewrite the text in bullet points and then make an easy text from these bullet points.[10]

4.4. Randomized controlled trial

Seventy-two participants (thirty-six men and thirty-six women) were recruited online between February and March 2023 for this study through convenience sampling. Participants were asked to fill in their availability and contact details. Prior to conducting the reading comprehension experiment, ethical approval from our Ethical Review Board was sought and obtained. Participants were required to be at least 18 years of age and have at least a basic understanding of Dutch. No other demographic characteristics were considered in the recruitment process.

The experiment has followed the guidelines of the ISO framework (Bevan et al., 2016). Participants were randomly divided into eight groups, with each group reading a different combination of the three letters. Figure 1 shows the number of participants per group and education level. The abbreviation "O" represents the Original version whereas "G" represents the Generated simple version by ChatGPT.

The letters were presented to participants in a predetermined order. This was done to control for order effects and to reduce potential biases. Before reading the letters, participants were given a brief introduction to the study and provided with a short scenario introduction for every letter. They were instructed to read the letters carefully and take notes if they wished. After reading a paragraph of the letter, participants were asked to answer questions about the letter.

The questionnaire consists of both closed and open-ended questions and took approximately 10-15 minutes to complete in total per letter. After reading all three letters and completing the questionnaires, participants were debriefed on the purpose of the study and thanked for their participation.

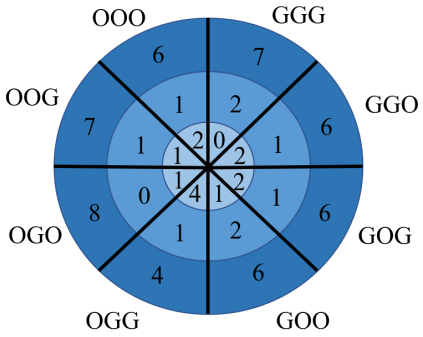

Figure 1: **Participants distribution:** Every chart represents a subgroup that reads the same letters in the same order. "O" represents the original letter and "G" represents the generated simplified letter. Every area has a number representing the number of participants. The inner circle represents the low-educated participants, the middle ring represents the middle-educated participants and the outside ring represents the high-educated participants. No distinction was made in this graph between men and women since there was an equal division within the subgroups.

4.4.1. Means comparison of original and simplified letters

Figure 2 shows the scores for the original and simplified letters. The first original letter saw notable enhancement when simplified, showing increased correct answers. This pattern persisted across subsequent letters, confirming improved comprehension. Aggregated results further confirm this, with participants scoring above 90% for both understanding and action question types for the simplified letters.

4.4.2. Regression analyses

We further investigate the difference in the performance of participants using (generalized) linear regression analyses and Multivariate Analysis of Variance (MANOVA).[11] This approach was chosen in order to make valid conclusions and investigate possible correlations that might influence the percentages and averages as seen in previous research (Dols, 2018).

Both analyses confirm that the simplified versions of the letters were better understood having significant scores for the simplified type of letter influencing the percentages of correctly answered questions for all three letters. Additionally, the age of participants

[10] This resulted in the final prompts which are shown in chat four of appendix B

[11] The full outcome of these analyses can be found in appendix G and H - values for the dummies $l1_g$,$l2_g$, and $l3_g$ represent simplified letters.

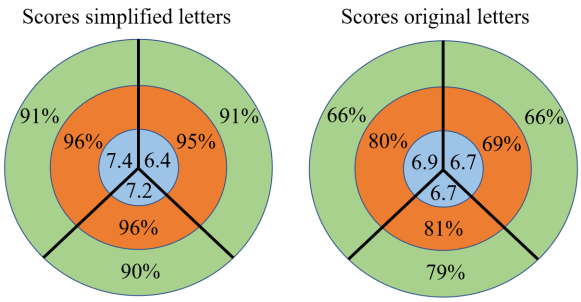

Figure 2: **Scores of the letters** The left diagram represents the scores of the simplified letters and the right diagram represents the scores of the original letters. Every chart represents a letter. The top-right represents the first letter, the top-left represents the second letter and the lowest chart represents the third letter. The surface of the chart represents the number of words that the letter contained (meaning that a bigger surface relates to a longer letter). The colours represent the type of questions. Green: understanding questions, orange: action questions and blue: tone questions. For the understanding and action questions, the percentage of correct answers is shown. For the tone questions, the average grade that participants assigned to the letters is shown.

emerged as a significant variable, demonstrating that higher ages negatively influenced the percentage of correctly answered action questions.
The MANOVA results further confirm our findings, providing an additional layer of confirmation for the positive impact of the simplified letters. The notable consistency across these types of analyses proves the robustness of our findings.

5. Discussion

This study investigated the extent to which a naive model, RoBERTa, and ChatGPT can improve the comprehensibility of formal texts written by Dutch governmental organisations. The challenge of TS in such a context is that the result needs to retain essential information allowing citizens to take actions while making the text easier to understand and act upon.
The results from multiple attempts at prompt engineering showed that it is possible to develop a one-shot learning approach (Kojima et al., 2022) to achieve excellent results, which makes scaling up this TS task easier. Initially writing the text in bullet points, followed by transforming these into easy-to-read text, proved to be the most effective prompt for this research.
Despite the evaluation metrics suggesting otherwise, the expert analysis determined that only the ChatGPT model's generated letters fulfilled the simplification criteria, maintaining all crucial information and terminology. Consequently, we proceeded with this model exclusively for the randomized controlled trial.
The results of the randomized controlled trial show that the ChatGPT model excelled in terms of enhancing the comprehensibility of the letters. An average increase of more than 20% was achieved for the percentage of correctly answered understanding questions. For the percentage of correctly answered action questions, there was an increase of 19% on average. Additionally, the grade for the tone was higher for the second and third letters, namely 0.5 on a 10-point scale. Only the first letter received on average 0.3 less compared to the tone grade for the original letter. However, the results regarding the tone were not significant because of inconsistent grading by the participants and too limited data because only three grades were solicited.
The randomized controlled trial results were further analysed with regression analyses to examine how correctly answering understanding and action questions was influenced by the simplified versions of the letters, controlling for various other variables. Three models were used to perform the analyses: Generalised Linear models, Linear Models, and MANOVA models. Across all models, the dummy variables indicating the simplified version were consistently significant, validating the results presented in Figure 2. Regarding our main research question, the machine learning model ChatGPT has demonstrated a substantial improvement in terms of the comprehensibility of the letters.

5.1. Future work and limitations

Future work could focus on improving the prompts, as initial exploration of tailoring prompts to particular audiences shows promise. Tailored prompts can make calls to action clearer and more compelling for specific audiences, thereby increasing the likelihood of the desired response, whether it's complying with regulations, or participating in civic activities.
Therefore, future work in this area could focus on developing more sophisticated techniques for audience analysis and prompt customization, thereby maximizing the impact of simplified texts for diverse audiences.

5.1.1. Evaluation methodology

TS for the purpose of general readership is a task that requires human evaluation to validate the results of the task (Engelmann et al., 2023). Numerous automatic evaluation metrics are developed to help alleviate this resource-intensive task. This study joins Young and Shishido (2023) in raising concerns with regard to the reliability of the automatic evaluation metrics for assessing simplification tasks performance of LLMs in general and ChatGPT in particular. We find that ROUGE, BLEU, and BLEURT poorly capture the quality of TS of governmental texts for general audiences. Hence they should be used with caution and ideally refined. Consensus is lacking on such metrics' suitability for TS assessment (Engelmann et al., 2023). Our results demonstrate that the models with the highest BLEU and ROUGE scores did not necessarily yield the best simplifications. We observe that BLEURT scores were not consistently 1.0 when evaluating identical reference text due to dataset limitations and model constraints such as syntactic structure: BLEURT may not fully account for changes in syntactic structure introduced by the simplification process. A simplified sentence may have a different sentence structure compared to the original, which could affect readability and clarity in a positive way but result in lower scores for the automatic evaluation metrics as the structure changes compared to the reference.

Furthermore, automatic evaluation metrics may struggle to evaluate how well the simplified text captures the intended meaning within the broader context. They primarily focus on local similarity measures and may not capture broader contextual information. However, our experience is that the original letters are not very well-structured neither coherent. Changes to both sentences structure and paragraphs placement to make it in a broader context coherent, is advisable in such cases.

For proper evaluation, combining BLEURT scores with other metrics and expert assessments is advised. Future research could consider expanding the reference texts to improve the performance of automatic evaluation metrics. The qualitative interviews and randomized controlled trials, though valuable, have limitations. Future studies should involve a wider range of experts and include for example the original letters' authors. Moreover, including more people with lower education levels and those with reading disabilities as participants could yield potentially even greater results for the TS impact in the randomized controlled trials.

In conclusion, this research offers insights into the efficacy of simplifying Dutch formal texts with ChatGPT, while at the same time underpinning the need for refinement and further exploration of evaluation methodologies.

5.1.2. Scaling up text simplification tools

The scope of this study was limited to testing multiple text simplification models and their evaluation. However, for future deployment, research needs to be performed with the stakeholders who are going to use the envisioned tool in their work as authors of letters. Therefore, it is recommended to conduct interviews with these stakeholders and find a form of implementation that suits them. An example format of implementation could be a web-based interface such as "Simpel" (Rijksoverheid, 2023) has for citizens (but then designed for personal computer use instead of smartphones) that allows the authors of the letters to input the original text and receive the simplified version straight away. The interface could also provide options for customising the level of simplification based on the target audience or purpose of the letter, as it can be fully focused on supporting the authors of the letters. By showing the original input and the generated simplified text on one screen the authors can rate the level of simplification and extent to which the essence of the text is retained, being important from the legal perspective. Correlating these ratings with new and existing TS evaluation metrics will allow the researchers to refine them further.

Bringing the results of this research into deployment requires several steps. First, a suitable model must be chosen to be able to simplify large letters at once. As alternatives for ChatGPT are popping up (Harnish, 2023), a comparison of these models should be made whereas the best model should be chosen for implementation. Furthermore, a way to check automatically for missing information must be implemented and/or a disclaimer must be provided that the author must check this him- or herself.

Once the model has been successfully deployed and proven effective for the authors, it could have a significant impact on improving the readability and comprehensibility of governmental texts. This could lead to better communication and engagement with citizens, as well as more efficient and effective use of resources by governmental organisations.

Considering the ongoing efforts to make LLMs in general and ChatGPT in particular more responsible, the performance of the next generation ChatGPT (e.g. ChatGPT 4.0) is not necessarily better than ChatGPT 3.5 (Chen et al., 2023) hence performance of TS tasks also requires a continuous re-assessment as new LLMs emerge. Scaling to different types of letters and languages also requires further investigation.

5.2. Ethical considerations

The deployment of LLMs for the simplification of formal texts from governmental organizations to citizens introduces a novel approach to enhancing accessibility and comprehension. While this technology promises significant benefits in making government communications more understandable to a broader audience, it also raises ethical considerations that must be addressed to ensure its responsible use. This section outlines the primary ethical concerns related to potential biases and harms that could arise from such automated systems and the measures taken to mitigate these risks.

5.2.1. Potential biases and harms

The use of LLMs comes with specific ethical concerns, which we tried to address using the following strategies:

1. Controlled Input Information: Unlike typical LLM applications that generate content based on provided information, our approach strictly limits the model's role to simplifying the text without altering the content. This significantly reduces the risk of introducing new biases or errors in the message content, as the original information remains intact.

2. User Oversight and Control: We emphasize the importance of human oversight in the text simplification process. By ensuring that users (government officials or designated communicators) retain full control over the output, we can mitigate risks associated with automated generation. This approach allows for the careful review and adjustment of simplified texts to ensure they accurately and effectively convey the intended message without unintended biases or simplifications that could distort the meaning.

3. Transparency and Accountability: We tried to be transparent in the use of LLMs for text simplification. Specifically, by documenting and communicating the processes involved, including how the models were trained and the criteria used for simplification.

Overall, we feel that by maintaining strict control over the input information, ensuring user oversight, promoting transparency, and committing to continuous improvement, we can leverage the benefits of this technology for TS while minimizing risks.

6. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification.

Wejdan Alkaldi and Diana Inkpen. 2023. Text simplification to specific readability levels. *Mathematics*, 11.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification.

Fernando Alva-Manchego and Matthew Shardlow. 2022. Towards readability-controlled machine translation of COVID-19 texts. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 287–288, Ghent, Belgium. European Association for Machine Translation.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

International Amnesty. 2021. Xenofobe machines: Discriminatie door ongereguleerd gebruik van algoritmen in het nederlandse toeslagenschandaal.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

R. Anthony, Artino Jr., Jeffrey S. La Rochelle, Kent J. Dezee, and Hunter Gehlback. 2014. Developing questionnaires for educational research: Amee guide. *Medical Teacher*, 36(87):463–474.

Jan De Belder, Koen Deschacht, and Marie-Francine Moens. 2010. Lexical simplification.

Nigel Bevan, Jim Carter, Jonathan Earthy, Thomas Geis, and Susan Harker. 2016. New iso standards for usability, usability reports and usability measures. *HCI*, pages 268–278.

Eva Boontje. 2011. Een onderzoek naar de begrijpelijkheid van hypotheekvoorlichting. Master's thesis, Bacheloropleiding Communicatiestudies Faculteit Geesteswetenschappen Universiteit Utrecht, April.

Bram Bulte, Leen Sevens, and Vincent Vandeghinste. 2018. Automating lexical simplification in dutch. Master's thesis, Centre for Computational Linguistics, KU Leuven, Belgium, June.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Wuwei Lan Yang Zhong Wei Xu Chao Jiang, Mounica Maddela. 2021. Neural crf model for sentence alignment in text simplification. Master's thesis, Department of Computer Science and Engineering The Ohio State University, August.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time?

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation.

Mischa Corsius, Vera Lange, Yvette Linders, Henk Pander Maat, Els van der Pool, Nina Sangers, Keun Sliedrecht, Wouter Sluis-Thiescheffer, and Charlotte Swart. 2023. Monitor begrijpelijkheid overheidsteksten.

James Cox and Keni Brayton. 2008. How to build the best questionnaires in the field of eduction. pages 2–13.

T. A. Cramwinckel. 2014. De belastingdienst als vertaler: van wettekst naar webtekst. een casestudy. *MBB*, (7-8):299–312.

Pieter Delobelle, Thomas Winters, and Nettina Berendt. 2020. Robbert: a dutch roberta-based language model. Master's thesis, Department of Computer Science, KU Leuven, Faculty of Electrical Engineering and Computer Science, TU Berlin, September.

Alice Delorme Benites and Caroline Lehr. 2021. Neural machine translation and language teaching – possible implications for the cefr. Master's thesis, Zürcher Hochschule für Angewandte Wissenschaften Institut für Übersetzen und Dolmetschen.

Ashwin Devaraj, William Sheffield, Byron C. Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. *Association for Computational Linguistics*, 1(60):7331–7345.

Thibault Sellam Dipanjan and Das Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. (58):7881–7892.

Marc Dols. 2018. Brieven aan burgers: Analyse en evaluatie. Master's thesis, Communicatie- en Informatiewetenschappen Specialisatie: Bedrijfscommunicatie en Digitale Media (BDM) Faculteit Geesteswetenschappen Universiteit van Tilburg, May.

Michael Eisele. 2019. On automatic summarization of dutch legal cases. Master's thesis, Hamburg Universität Facultät Für Mathematik, Informatik und Naturwissenschaften, October.

ELSA-Lab. 2022.

Björn Engelmann, Fabian Haak, Christin Katharina Kreutz, Narjes Nikzad Khasmakhi, and Philipp Schaer. 2023. Text simplification of scientific texts for non-expert readers.

Laura Faulkner. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, I& Computers*, 3:379–383.

Hiske Feenstra, Karen Keune, Henk Pander Maat, Theo Eggen, and Ted Sanders. 2015. Geautomatiseerde beoordeling van schrijfvaardigheid. Master's thesis.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models.

Ellie Fossey, Carol Harvey, Fiona McDermott, and Larry Davidson. 2002. Understanding and evaluating qualitative research. *Australian and New Zealand Journal of Psychiatry*, 36:717–732.

Fritz, Morris, and Richler. 2012. Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141(1), 2–18.*

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. *Association for the Advancement of Artificial Intelligence*, pages 6423–6430.

Gebruiker-Centraal. 2022.

Dianne Gosens. 2008. Het effect van lexicale en syntactische wijzigingen op het begrip en de waardering van een autoverzekeringspolis. Master's thesis, Masteropleiding Communicatiestudies Faculteit Geesteswetenschappen Universiteit Utrecht, June.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. Master's thesis, Department of Computer Science, Cornell Tech Cornell University, New York, NY 10044, June.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Bart Haak. 2020. Information extraction from homicide-related dutch texts using bert. Master's thesis, Jheronimus Academy of Data Science, July.

Brian Harnish. 2023. Chatgpt alternatives you can try in 2023.

Melody A. Hertzog. 2008. Considerations in determining sample size for pilot studies.

Jason Holmes, Zhengliang Liu, Lian Zhang, Yuzhen Ding, Terence T. Sio, Lisa A. McGee, Jonathan B. Ashman, Xiang Li, Tianming Liu, Jiajian Shen, and Wei Liu. 2023. Evaluating large language models on a highly-specialized topic, radiation oncology physics.

Neslihan Iskender, Tim Polzehl, and Sebastian Moller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. Master's thesis, Technische Universitat Berlin, Quality and Usability Lab, April.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, and Michael Ingrisch. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports.

J. Joshua. 2023. Data controls faq.

Suzanne Kleijn, Henk Pander Maat, and Ted Sanders. 2016. Effects of dependency length on the processing and understanding of texts. Master's thesis, Communicatie- en informatiewetenschappen Universiteit Utrecht.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid Google Research, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners.

Rogier Kraf, Leo Lentz, and Henk Pander Maat. 2011. Drie nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid. *Tijdschrift voor Taalbeheersing*, (3):249–265.

Rogier Kraf and Henk Pander Maat. 2009. Leesbaarheidsonderzoek: oude problemen, nieuwe kansen. *Tijdschrift voor Taalbeheersing*, (2):97–123.

Yogesh Kumar, Komalpreet Kaur, and Sukhpreet Kaur. 2021. Study of automatic text summarization approaches in diferent languages. Master's thesis, Department of Computer Science Engineering, Chandigarh Group of Colleges, Landran, Mohali, India, January.

Naomi Langstraat. 2019. Creating a classroom-mt: Connecting simplification methods to language learner levels in monolingual machine translation. Master's thesis, Utrecht University, Faculty of Humanities, Bachelor of Science - Artificial Intelligence, July.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw @ ami and haspeede2: Fine-tuning a pre-trained comment-domain bert model.

Leo Lentz. 2021. Wat zijn tekstbegrijpelijkheids voorspellingen waard? een vergelijkend onderzoek. *Handboek Didactiek Nederlands*, (4).

Leo Lentz and Sanne Elling. 2003. De voorspellende kracht van het ccc-model. *Tijdschrift voor Taalbeheersing*, (3):221–235.

Leo Lentz, Louise Nell, and Henk Pander Maat. 2017. Begrijpelijkheid van pensioencommunicatie: effecten van wetgeving, geletterdheid en revisies.

Leo Lentz and Henk Pander Maat. 2011. Een leesbare bijsluiter. *Tijdschrift voor Taalbeheersing*, (2):128–151.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Master's thesis, Facebook AI, Oktober.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. Master's thesis, Stanford University, January.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

van de Nick Luijtgaarden, Daniël Prijs, Marijn Schraagen, and Floris Bex. 2022. Abstractive summarization of dutch court verdicts using sequence-to-sequence models. Master's thesis, Utrecht University, The Netherlands, December.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Qing Lyu, Josh Tan, Michael E. Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J. Myers, Ge Wang, and Christopher T. Whitlow. 2023. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: Promising results, limitations, and potential.

Ritch Macefield. 2009. How to specify the participant group size for usability studies: A practitioner's guide. *Journal of usability studies*, 5:34–35.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore. Association for Computational Linguistics.

Paulo R. A. Margarido, Thiago A. S. Pardo, Gabriel M. Antonio, Vinícius B. Fuentes, Rachel Aires, Sandra M. Aluísio, and Renata P. M. Fortes. 2008. Automatic summarization for text simplification: Evaluating text understanding by poor readers. Master's thesis, HAN university of applied sciences.

Louis Martin, Angela Fan, Éric de la Clergerie, and Antoine Bordes Benoît Sagot. 2021. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. Master's thesis, Facebook AI Research, Paris, France, April.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of gpt-3 and advanced neural language models. Master's thesis, Center on

Terrorism, Extremism, and Counterterrorism Middlebury Institute of International Studies, September.

Rolf Molich. 2010. A critique of "how to specify the participant group size for usability studies: A practitioner's guide". *Journal of usability studies*, 5:124–128.

Elisa Nguyen, Daphne Theodorakopoulos, Shreyasi Pathak, Jeroen Geerdink, Onno Vijlbrief andMaurice van Keulen, and Christin Seifert. 2021. A hybrid text classification and language generation model for automated summarization of dutch breast cancer radiology reports. Master's thesis, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands, May.

Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. 2020. Better distractions: Transformer-based distractor generation and multiple choice question filtering. Master's thesis, Leiden Institute of Advanced Computer Science, Leiden University, October.

OpenAI. 2023. Gpt-4 technical report. Technical report.

Gavora P. 2012. Text comprehension and text readability. *Faculty of Humanistic studies, Tomas Bata University, Czech Republic*, pages 9–10.

Gustavo H. Paetzold and Lucia Specia. 2017. A survey on lexical simplification. Master's thesis, The University of Sheffield Western Bank Sheffield United Kingdom, November.

Kendeou Panayiota, Krista R. Muis, and Sandra Fulton. 2011. Reader and text factors in reading comprehension processes.

Henk Pander Maat and Nick Dekker. 2016. Tekstgenres analyseren op lexicale complexiteit met t-scan. *Tijdschrift voor Taalbeheersing*, 38(3):263–304.

Henk Pander Maat and Sanne Ditewig. 2017. Hoe worden onderwijsteksten vereenvoudigd, en helpt dat? *Handboek Didactiek Nederlands*, (39):245–263.

Henk Pander Maat, Leo Lentz, and Raynor D. K. 2015. How to test mandatory text templates: The european patient information leaflet. Technical report, PLOS one.

Henk Pander Maat and Thea van der Geest. 2021. Monitor begrijpelijkheid overheidsteksten.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Association for Computational Linguistics*, 1(40):311–318.

Wouter Peer. 2023.

Daniël Prijs. 2022. On automatic summarization of dutch legal cases. Master's thesis, Utrecht University Graduate School of Natural Sciences Business Informatics, July.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Jan Renkema. 2011. Een kwalitatief onderzoek naar de begrijpelijkheid van digitale informatie van de belastingdienst.

Jan Renkema. 2013. Een kwalitatief verkennend onderzoek naar de kwaliteit van antwoordbrieven van de belastingdienst.

M. Heerkens K.R. Leuverink I. Veringa Renkema J. 2012. Minder belasting door meer begrip. *Universiteit van Tilburg Faculteit Geesteswetenschappen*, (4).

Rijksoverheid. 2023. Terugblik demo donderdag: Lees simpel app versimpelt overheidsinformatie. *Rijksprogramma voor Duurzaam Digitale Informatiehuishouding*.

Samuel Ronnqvist, Jenna Kanerva, and Tapio Salakoski Filip Ginter. 2020. Is multilingual bert fluent in language generation? Master's thesis, TurkuNLP, Department of Future Technologies University of Turku, Finland.

Alroobaea Roobaea and Pam J. Mayhew. 2014. How many participants are really enough for usability studies? Technical report, IEEE.

Alessandra Rossetti. 2019. Simplifying, reading, and machine translating health content: An empirical investigation of usability. Master's thesis, School of Applied Language and Intercultural Studies Dublin City University, April.

Muhammad Salman, Armin Haller, and Sergio J. Rodríguez Méndez. 2023. Syntactic complexity identification, measurement, and reduction through controlled syntactic simplification.

Ted Sanders and Carel Jansen. 2011. Begrijpelijke taal – fundamenten en toepassingen van effectieve communicatie. *Tijschrift voor Taalbeheersing*, (33):201–207.

Victor Sanh and Thomas Wolf Lysandre Debut, Julien Chaumond. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Juliën Schoonbrood. 2013. Onderzoek naar de invloed van financiële geletterdheid op pensioenkennis en de vind- en begripsprestaties van de startbrief. Master's thesis, Communicatie- en informatiewetenschappen Universiteit Utrecht, November.

Matthew Shardlow. 2014. A survey of automated text simplification.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable.

A. F. Siddiqi. 2014. An observatory note on tests for normality assumptions. *Journal of modelling in management*, (3):290–305.

Tialda Sikkema, Leo Lentz, Henk Pander Maat, and Nadja Jungmann. 2017. De taakgerichtheid van de aanmaning en de dagvaarding in incassozaken. *Tijdschrift voor Taalbeheersing*, 39(3):273–295.

Lynn Snyder, Caccamise Donna, and Wise Barbara. 2005. The assessment of reading comprehension. *Journal of modelling in management*.

S. Soberón and Winfried Stute. 2017. Assessing skewness, kurtosis and normality in linear mixed models. *Journal of Multivariate Analysis*, pages 123–140.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. pages 2637–2652.

Sanja Stajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. pages 1–10.

S. Suha and Aqil M. Azmi. 2021. Automated text simplification: A survey. 54(2):36.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. Master's thesis, Department of Computer Science, The Hebrew University of Jerusalem, October.

Bowen Tan and Virapat Kieuvongngam. 2020. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. Master's thesis, Laboratory of Molecular Genetics Rockefeller University New York, June.

(VNG) Vereniging Nederlandse Gemeenten. 2021. Regionaal verbeteren brieven omgevingswet. *Vereniging van Nederlandse Gemeenten (VNG)*.

Dr. S. Vijayarani, Ms. J. Ilamathi, and Ms. Nithya. 2020. Preprocessing techniques for text mining - an overview.

de Wietse Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert mode. Master's thesis, CLCG, University of Groningen, The Netherlands, December.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. Master's thesis, University of California, Santa Barbara, September.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques methods. 34(4):1029–1046.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. Master's thesis, School of Engineering and Applied Sciences Harvard University Cambridge, MA, USA, July.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. (7):401–415.

Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. 2022. Automatic text summarization methods: A comprehensive review. Master's thesis.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. 2022. What gpt knows about who is who. Master's thesis, Institute for Applied Computational Science Harvard University, May.

Julio Christian Young and Makoto Shishido. 2023. Evaluation of the potential usage of chatgpt for providing easier reading materials for efl students.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020. Retrospective reader for machine reading comprehension. Master's thesis, Department of Computer Science and Engineering, Shanghai Jiao Tong University.

7. Appendices

A. Letters used for randomized controlled trial

The letters used for the randomized controlled trial can be found on this github page: https://anonymous.4open.science/r/COLING-24-93E6/.
The letters were in Dutch and are translated here for comprehension of this research.

B. Questions prompt-engineering ChatGPT

The prompts for ChatGPT were in Dutch and are translated here for comprehension of this research. For all these prompts "de volgende tekst" refers here to the text of the letter.

Chat one: CEFR levels

- Kun je het niveau van de volgende tekst bepalen volgens de CEFR-classificaties?
- Kun je de volgende tekst vereenvoudigen naar CEFR-niveau A1?
- Kun je de volgende tekst tekst vereenvoudigen naar CEFR-niveau A2?
- Kun je de volgende tekst vereenvoudigen naar CEFR-niveau B1?
- Kun je de volgende tekst vereenvoudigen naar CEFR-niveau B2?
- Kun je de volgende tekst herschrijven naar CEFR-niveau A1?
- Kun je de volgende tekst herschrijven naar CEFR-niveau A2?
- Kun je de volgende tekst herschrijven naar CEFR-niveau B1?
- Kun je de volgende tekst herschrijven naar CEFR-niveau B2?
- Kun je de volgende tekst herschrijven naar CEFR-niveau C1?
- Kun je de volgende tekst herschrijven naar CEFR-niveau C2?

Translations:

- Can you determine the level of this text according to the CEFR classifications?
- Can you simplify this text to CEFR level A1?
- Can you simplify this text to CEFR level A2?
- Can you simplify this text to CEFR level B1?
- Can you simplify this text to CEFR level B2?
- Can you rewrite this text to CEFR level A1?
- Can you rewrite this text to CEFR level A2?
- Can you rewrite this text to CEFR level B1?
- Can you rewrite this text to CEFR level B2?
- Can you rewrite this text to CEFR level C1?
- Can you rewrite this text to CEFR level C2?

Chat two: Additional input

- Kunt u deze tekst vereenvoudigen met behulp van deze definitielijst waarbij in de eerste kolom het moeilijke woord staat en in de tweede kolom de eenvoudige definitie?
- Kunt u deze definitielijst gebruiken om deze tekst te vereenvoudigen?
- Kunt u deze definitielijst gebruiken om deze tekst te herschrijven?

Translations:

- Can you simplify this text with the use of this definition list having in the first column the difficult word and in the second column the simple definition?
- Can you use this definition list to simplify this text?
- Can you use this definition list to rewrite this text?

Chat three: Zero-shot vs Few-shot

1. Kun je de volgende tekst begrijpbaarder schrijven?
2. Kun je de volgende tekst versimpelen?
3. Kun je de volgende tekst opschrijven in bullet points?
4. Kun je deze bullet points in een eenvoudige tekst opschrijven?
5. Kun je hiervan een makkelijke tekst schrijven?

Translations:

1. Can you make the following text more comprehensible?
2. Can you simplify the following text?
3. Can you write the following text in bullet points?
4. Can you write a simple text based on these bullet points?
5. Can you write an easy text from this?

Chat four: Final version

1. Kun je de volgende tekst opschrijven in bullet points?
2. Kun je deze bullet points in een makkelijke tekst schrijven?

Translations:

1. Can you rewrite this text into bullet points?
2. Can you write these bullet points into an easy text?

C. Introduction and scenario description randomized controlled trial

The introduction and scenario descriptions were in Dutch but are translated for the comprehension of this research.

Introductie onderzoek

Beste lezer, Wat fijn dat je meedoet aan het lezersonderzoek van mijn thesis. Je krijgt zo drie teksten te zien over het thema zorg. Om de privacy van de gemeenten en de geadresseerden te bewaken, zijn de teksten geanonimiseerd en gesitueerd in de denkbeeldige gemeente Zilverdam. Ik wil je vragen om de brieven één voor één te lezen en daarbij te zeggen wat je denkt. Het is belangrijk om aan te geven als je iets niet begrijpt of onduidelijk vindt. Ik wil je vragen om deze stukken te markeren. Daarnaast worden er vragen gesteld door mij over de inhoud van de brieven tijdens het onderzoek en naderhand over de toon van brief. Deze vragen geven inzicht in hoe makkelijk:

- Je begrijpt wat er staat;
- Je begrijpt wat er gedaan moet worden;
- Je de toon gepast vindt.

Scenario script thema zorg:

Brief 1: WMO voorzieningen

Jouw tante Janny woont samen met haar man in de gemeente Zilverdam. Ze zijn beide met pensioen. Janny heeft steeds meer moeite met haar evenwicht. Ze loopt nu met een stok. Traplopen vindt ze erg lastig. De slaapkamer is boven en daarom wil Janny een traplift. Jij bent Janny's mantelzorger. Ze vraagt jou of je wilt kijken of wat er geregeld kan worden bij de gemeente.

Brief 2: Regels PGB

Janny en jij hebben een gesprek gehad met iemand van de gemeente. De traplift die Janny wil, zit niet in het aanbod van de gemeente. De gemeente zegt dat ze moet kijken naar een pgb. Lees de tekst om te kijken of een pgb iets voor Janny is.

Brief 3: Besluit PGB

Janny kon niet langer wachten en heeft alvast een traplift besteld. Met de gemeente maakte ze ondertussen een plan en deed de aanvraag voor een pgb. Lees de tekst om uit te leggen wat er is besloten.

Research Introduction

Dear reader, thank you for participating in the reader survey for my thesis research. You will now be presented with three texts on the topic of healthcare. To protect the privacy of municipalities and recipients, the texts have been anonymized and are situated in the imaginary municipality of Zilverdam. I kindly request you to read each of the letters one by one and share your thoughts as you do so. It's important to indicate if there is anything you do not understand or find unclear. Please mark these sections. Additionally, I will ask questions during and after the research about the content of the letters and the tone used in them.
These questions will provide insight into how easily you:

- Understand the content;
- Comprehend what needs to be done;
- Find the tone appropriate.

Scenario Script: Theme Care

Letter 1: WMO Facilities

Your aunt Janny lives with her husband in the municipality of Zilverdam. They are both retired. Janny is experiencing increasing balance issues and now uses a cane. Climbing stairs is challenging for her. The bedroom is upstairs, so Janny wants a stairlift. You are Janny's caregiver, and she has asked you to see if anything can be arranged with the municipality.

Letter 2: PGB Regulations

You and Janny had a conversation with someone from the municipality. The stairlift Janny wants is not part of the municipality's offerings. The municipality suggests she explore a Personal Budget (PGB). Please read the text to determine if a PGB is suitable for Janny.

Letter 3: PGB Decision

Janny couldn't wait any longer and has already ordered a stairlift. In the meantime, she worked with the municipality to create a plan and applied for a PGB. Please read the text to understand what decision has been made.

D. Questionnaires for randomized controlled trial

The questions and answers were in Dutch but are translated for the comprehension of this research.

Vraag	Juiste antwoord
Wat moet je doen als je hulp nodig hebt om zelfstandig te kunnen blijven wonen?	1. Melden hulpvraag
	2. Formulier in vullen
Heb je een DigiD nodig voor het invullen van het formulier?	Ja
Wat kun je doen als je geen DigiD hebt?	Telefonisch contact opnemen
Hoe meld je een hulpvraag?	Via de knop "Verzoek voor Wmo-voorziening"
Hoe heet het meldingsformulier?	Sociale dienstverlening
Binnen hoeveel tijd wordt de aanvraag beoordeeld?	8 weken
Waar wordt ernaar gekeken bij het beoordelen van de nodige zorg?	1. Hulp buren/familie
	2. Algemene voorzieningen
	3. Maatwerkvoorzieningen
Zijn maatwerkvoorzieningen persoonlijk?	Ja
Wat is een voorbeeld van een algemene voorziening?	Maaltijdservice
	Maatjesproject
Via wat kun je maatwerkvoorzieningen krijgen?	Zorg in natura (ZIN)
	Persoonsgebonden budget (PGB)
Wat houdt zorg in natura in?	De gemeente regelt alles
Hoe kun je opzoeken welke zorg de gemeente inkoopt?	Via de website Sociale kaart Zilverdam
Is de eigen bijdrage voor ZIN en PGB hetzelfde?	Ja
Wat houdt een persoonsgebonden budget in?	Zelf verantwoordelijk om de zorg te regelen
	(inkopen zorg, administratie, opstellen zorgovereenkomst)
Voor welke zorg en ondersteuning betaal je een eigen bijdrage?	Zie opsomming
Hoe hoog is de maximale eigen bijdrage?	19 Euro
Waarvan is de maximale eigen bijdrage afhankelijk?	leeftijd
	wel/geen partner
Wie bepaalt de hoogte van de eigen bijdrage?	CAK
Tot wanneer betaal je de eigen bijdrage?	Zolang je ondersteuning nodig heeft / tot de kostprijs bereikt is
Wat is het abonnementstarief?	De eigen bijdrage Wmo
Klopt het dat de eigen bijdrage af hangt van de hoeveelheid zorg?	Nee
Klopt het dat je tijdens een vakantie geen eigen bijdrage betaalt?	Nee
Waar vindt je meer informatie over de hoogte van de eigen bijdrage?	Website CAK
Aan wie betaal je de eigen bijdrage?	CAK
Hoevaak betaal je de eigen bijdrage?	Elke maand
Hoe betaal je de eigen bijdrage?	Automatische incasso / acceptgiro
Wat kun je doen als je meer informatie wilt over betalen?	Website CAK bezoeken / CAK bellen op 0800 1925
Wat moet je doen als je indicatie afloopt?	Contact opnemen zorgaanbieder
Wanneer moet je contact opnemen met de zorgaanbieder als je indicatie afloopt?	2 maanden voor het aflopen
Moet je voor het aflopen van de indicatie het aanvraagformulier invullen	Nee
Wat moet je doen voor vragen over sociale dienstverlening?	Bellen met 0900 1234
Waarvoor is de casemanager?	Overige vragen (als je er via andere manieren niet uitkomt)
time:	
Hoe duidelijk vind je de brief in het algemeen? (1-10)	
Vind je dat er overbodige informatie in de brief staat? Zo ja, waar?	
Wat vind je van de toon van de tekst? (bijvoorbeeld streng of vriendelijk)	
Hoe erg gepast vind je de toon van de brief in het algemeen? (1-10)	
Wat zou de schrijver van de tekst als eerste moeten veranderen?	

Figure 3: Questions with answers for letter one: WMO voorzieningen (Dutch).

Question	Correct answer
What should you do if you need help to continue living independently?	1. Report your request for help
	2. Fill out a form
Do you need a DigiD to fill out the form?	Yes
What can you do if you don't have a DigiD?	Contact by phone
How do you report a request for help?	Through the "Request for Wmo-provision" button
What is the name of the reporting form?	Social services
Within how much time is the request assessed?	8 weeks
What is considered when assessing the necessary care?	1. Help from neighbors/family
	2. General provisions
	3. Customized provisions
Are customized provisions personal?	Yes
What is an example of a general provision?	Meal service
	Buddy project
How can you receive customized provisions?	Nature of care (ZIN)
	Personal budget (PGB)
What does nature of care (ZIN) entail?	The municipality handles everything
How can you find out which care the municipality procures?	Through the website Social Map Zilverdam
Is the own contribution for ZIN and PGB the same?	Yes
What does a personal budget (PGB) entail?	Personally responsible for arranging care
	(purchasing care, administration, establishing care agreement)
For which care and support do you pay an own contribution?	See list
What is the maximum own contribution amount?	19 Euros
What is the maximum own contribution amount dependent on?	age
	marital status
Who determines the amount of the own contribution?	CAK
Until when do you pay the own contribution?	As long as you require support / until the cost threshold is reached
What is the subscription fee?	The Wmo own contribution
Is it true that the own contribution depends on the amount of care?	No
Is it true that you do not pay an own contribution during a vacation?	No
Where can you find more information about the amount of the own contribution?	CAK website
To whom do you pay the own contribution?	CAK
How often do you pay the own contribution?	Every month
How do you pay the own contribution?	Automatic debit / payment slip
What can you do if you want more information about payments?	Visit the CAK website / Call CAK at 0800 1925
What should you do when your indication expires?	Contact the care provider
When should you contact the care provider when your indication expires?	2 months before it expires
Do you need to fill out the application form before the indication expires?	No
What should you do for questions about social services?	Call 0900 1234
What is the casemanager for?	Other questions (if you cannot resolve them through other means)
time:	
How clear do you find the letter overall? (1-10)	
Do you think there is any unnecessary information in the letter? If yes, where?	
What do you think of the tone of the text? (e.g., strict or friendly)	
How appropriate do you find the tone of the letter overall? (1-10)	
What should the writer of the text change first?	

Figure 4: Questions with answers for letter one: WMO Facilities (translated).

Vraag	Juiste antwoord
Waarvoor wordt een pgb gebruikt?	Het regelen van ondersteuning of hulp
Klopt het dat bij een pgb je zelf de zorgverlener kiest?	Ja
Klopt het dat bij een pgb je zelf bepaalt wanneer je hulp krijgt?	Ja
Klopt het dat bij een pgb je zelf bepaalt hoe je hulp krijgt?	Ja
Heeft de gemeente heeft met alle (zorg)aanbieders een contract afgesloten?	Nee
Kun je een pgb gebruiken voor zorgaanbieders waarmee de gemeente geen contract mee h	Ja
Wat koop je met een pgb zelf in?	Jeugdzorg/ondersteuning/hulpmiddelen
Wat betekend de regie voeren in deze brief?	Zie opsomming
Wanneer mag je geen pgb gebruiken?	Zie opsomming
Wat moet er in het uitvoeringsplan staan?	Zie opsomming
Is een zorgovereenkomst hetzelfde als een uitvoeringsplan?	Nee
Wat moet er in de zorgovereenkomst staan?	Afspraken met de zorgverlener
Wie houdt in de gaten wanneer het pgb stopt?	Uzelf/de zorgvrager
Wat moet je doen als het pgb stopt en je nog zorg nodig hebt?	Nieuw gesprek bij het wijkteam aanvragen
Hoever voor het stoppen van het pgb moet je dit (gesprek wijkteam) aanvragen?	8 weken voor de einddatum van het pgb
Waarop controleert het wijkteam?	1. Onjuist gebruik pgb's
	2. Fraude
Hoe vraag je een pgb aan?	Volgen van het stappenplan/brochure/folder
Wat voor vragen kun je stellen aan de Sociale Verzekeringsbank?	Zie opsomming
Hoe en aan wie moet je vragen stellen over de inhoud van de zorgovereenkomst?	Gemeente, telefonisch op 14033
Waarover kan de contactpersoon bij het wijkteam meer informatie geven?	pgb-bedragen, toekenning pgb, rekeninstrument, hulpvraag
Hoe vraag je een pgb aan voor meerdere gezinsleden?	Via 1 plan
Wat zijn eisen van het plan voor een pgb voor meerdere gezinsleden?	1. Zorgbehoefte alle gezinsleden (die hulp nodig hebben) overzichtelijk
	2. Verband zorgbehoefte alle gezinsleden (die hulp nodig hebben) duidelijk
Wat is een alternatief van een pgb?	Zorg in natura (zin)
Wat is het verschil tussen een pgb en zin?	Pgb heeft andere zorgaanbieders dan zin
Wat is de eigenbijdrage voor een pgb?	19 Euro
Wat kun je doen als je hulp nodig hebt en je weet niet bij wie je moet zijn?	1. Bellen gemeente 14033
	2. Mailen gemeente info@zilverdam.nl
Waar moet je naartoe als je persoonlijk iemand wilt spreken?	Informatiewinkel
Hoe vind je een informatiewinkel in de buurt?	Indebuurt033
Met wat voor vragen helpt de gemeente u niet?	Specifieke vragen over de gezondheid
time:	
Hoe duidelijk vind je de brief in het algemeen? (1-10)	
Vind je dat er overbodige informatie in de brief staat? Zo ja, waar?	
Wat vind je van de toon van de tekst? (bijvoorbeeld streng of vriendelijk)	
Hoe erg gepast vind je de toon van de brief in het algemeen? (1-10)	
Wat zou de schrijver van de tekst als eerste moeten veranderen?	

Figure 5: Questions with answers for letter two: Regels PGB (Dutch).

Question	Correct answer
What is a personal budget (PGB) used for?	Managing support or assistance
Is it true that with a PGB, you can choose your own care provider?	Yes
Is it true that with a PGB, you decide when you receive assistance?	Yes
Is it true that with a PGB, you decide how you receive assistance?	Yes
Has the municipality entered into contracts with all (care) providers?	No
Can you use a PGB for care providers with whom the municipality has no contract?	Yes
What do you purchase with a PGB?	Youth care/support/aids
What does "taking control" mean in this letter?	See list
When are you not allowed to use a PGB?	See list
What should be included in the implementation plan?	See list
Is a care agreement the same as an implementation plan?	No
What should be included in the care agreement?	Agreements with the care provider
Who monitors when the PGB stops?	Yourself/the care recipient
What should you do if the PGB stops, and you still need care?	Request a new conversation with the neighborhood team
How long before the PGB expiration date should you request this (neighborhood team conversation)?	8 weeks before the PGB's end date
What does the neighborhood team check?	1. Improper use of PGBs
	2. Fraud
How do you apply for a PGB?	Follow the step-by-step guide/brochure/folder
What kind of questions can you ask the Social Insurance Bank?	See list
How and to whom should you ask questions about the content of the care agreement?	Municipality, by phone at 14033
What additional information can the contact person at the neighborhood team provide?	PGB amounts, PGB allocation, calculation tool, assistance request
How do you apply for a PGB for multiple family members?	Through one plan
What are the plan requirements for a PGB for multiple family members?	1. Clear overview of the care needs of all family members (needing assistance)
	2. Clear connection between the care needs of all family members (needing assistance)
What is an alternative to a PGB?	Care in kind (ZIN)
What is the difference between a PGB and ZIN?	PGB has different care providers than ZIN
What is the own contribution for a PGB?	19 Euros
What can you do if you need assistance and don't know who to contact?	1. Call the municipality at 14033
	2. Email the municipality at info@zilverdam.nl
Where should you go if you want to speak with someone in person?	Information desk
How can you find an information desk in the area?	Indebuurt033
What kind of questions does the municipality not help with?	Specific health-related questions
time:	
How clear do you find the letter overall? (1-10)	
Do you think there is any unnecessary information in the letter? If yes, where?	
What do you think of the tone of the text? (e.g., strict or friendly)	
How appropriate do you find the tone of the letter overall? (1-10)	
What should the writer of the text change first?	

Figure 6: Questions with answers for letter two: PGB Regulations (translated).

Vraag	Juiste antwoord
Wat is er op 13-09-2022 gebeurd?	Mevrouw (Brons) gemeld bij de gemeente met een hulpvraag
Wat is er op 18-09-2022 gebeurd?	Huisbezoek bij mevrouw (Brons)
Wat is de hulpvraag?	Een traplift
Wat is er besloten?	1. Mevrouw (Brons) krijgt een persoonsgebonden budgt (pgb)
	2. Mevrouw (Brons) krijgt een onderhoudsbedrag voor de traplift
Wat is een voorbeeld van een wijziging volgens de tekst?	Verzuizen/samenwonen/afname/toename beperking
Wat moet je doen bij een verandering van de persoonlijke situatie?	Doorgeven aan de gemeente
Hoe moet je een verandering doorgeven aan de gemeente?	Bellen naar 012-3456789 en vragen naar team Zorg
Wanneer moet je bezwaar maken?	Als je het niet eens bent met het besluit
Wat moet er in het bezwaar staan?	Zie opsomming
Tot wanneer kun je bezwaar indienen?	Tot 6 weken na de dag waarop het besluit is verzonden
Hoe maak je bezwaar?	1. Brief sturen naar adres van de gemeente
	2. Via www.zilverdam.nl -> loketten
Kun je bezwaar maken zonder een DigiD?	Ja
Wat is clientenondersteuning?	Meer hulp via een maatschappelijk werker
Hoeveel kost de clientenondersteuning?	Niks/Gratis
Hoe kun je clientenondersteuning aanvragen?	1. Bellen naar 012-3456789
	2. Mailen naar info@indebuurt.nl
	3. Binnenlopen informatiewinkel in de buurt
Wat moet je doen bij andere vragen/opmerkingen?	Contact opnemen Mevrouw Oudklomp
Kun je mevrouw Oudklomp op vrijdag bellen?	Nee
time:	
Hoe duidelijk vind je de brief in het algemeen? (1-10)	
Vind je dat er overbodige informatie in de brief staat? Zo ja, waar?	
Wat vind je van de toon van de tekst? (bijvoorbeeld streng of vriendelijk)	
Hoe erg gepast vind je de toon van de brief in het algemeen? (1-10)	
Wat zou de schrijver van de tekst als eerste moeten veranderen?	

Figure 7: Questions with answers for letter three: Besluit PGB (Dutch).

Question	Correct answer
What happened on 13-09-2022?	Mrs. (Brons) reported to the municipality with a request for assistance
What happened on 18-09-2022?	Home visit to Mrs. (Brons)
What is the request for assistance?	A stairlift
What was decided?	1. Mrs. (Brons) will receive a personal budget (PGB)
	2. Mrs. (Brons) will receive a maintenance amount for the stairlift
What is an example of a change according to the text?	Moving/cohabiting/decrease/increase in limitations
What should you do in case of a change in your personal situation?	Report it to the municipality
How should you report a change to the municipality?	Call 012-3456789 and ask for the Care team
When should you file an objection?	If you disagree with the decision
What should be included in the objection?	See list
Until when can you file an objection?	Up to 6 weeks after the day the decision was sent
How do you file an objection?	1. Send a letter to the municipality's address
	2. Via www.zilverdam.nl -> service counters
Can you file an objection without a DigiD?	Yes
What is client support?	Additional assistance through a social worker
How much does client support cost?	Nothing/Free
How can you request client support?	1. Call 012-3456789
	2. Email info@indebuurt.nl
	3. Visit an information desk in the neighborhood
What should you do for other questions/comments?	Contact Mrs. Oudklomp
Can you call Mrs. Oudklomp on Fridays?	No
time:	
How clear do you find the letter overall? (1-10)	
Do you think there is any unnecessary information in the letter? If yes, where?	
What do you think of the tone of the text? (e.g., strict or friendly)	
How appropriate do you find the tone of the letter overall? (1-10)	
What should the writer of the text change first?	

Figure 8: Questions with answers for letter three: PGB Decision (translated).

E. Results automatic evaluation metrics

Finance

Betalen_in_delen			ChatGPT	Naive	BERT	Original
LiNT			38	31	23	30
Rouge_1	rouge_1	recal	0.58839	0.94204	0.85232	-
	rouge_2	recal	0.34656	0.9018	0.73573	-
	rouge_L	recal	0.44591	0.94204	0.85232	-
BLEURT			0.60454	0.76511	0.37346	-
BLEU	blue_1	score	0.35816	0.87371	0.79755	-
		precision	0.8	0.9234	0.83544	-
	blue_2	score	0.27154	0.83336	0.71883	-
		precision	0.45985	0.84009	0.67865	-
	blue_3	score	0.21218	0.79659	0.64936	-
		precision	0.28938	0.76923	0.55508	-
	blue_4	score	0.16575	0.76533	0.59179	-
		precision	0.17647	0.71734	0.46921	-

Gemeentelijke_belastingen			ChatGPT	Naive	BERT	Original
LiNT			36	42	27	40
Rouge_1	rouge_1	recal	0.57904	0.89091	0.76554	-
	rouge_2	recal	0.30203	0.81092	0.60113	-
	rouge_L	recal	0.37684	0.88112	0.75424	-
BLEURT			0.45893	0.68226	0.3884	-
BLEU	blue_1	score	0.32691	0.87615	0.74517	-
		precision	0.8445	0.88595	0.76879	-
	blue_2	score	0.23403	0.8301	0.65409	-
		precision	0.4328	0.79526	0.59233	-
	blue_3	score	0.17002	0.78757	0.57542	-
		precision	0.23181	0.71688	0.45946	-
	blue_4	score	0.13084	0.74721	0.5059	-
		precision	0.15405	0.64525	0.3547	-

Kwijtschelding			ChatGPT	Naive	BERT	Original
LiNT			47	42	26	38
Rouge_1	rouge_1	recal	0.7171	0.90206	0.79377	-
	rouge_2	recal	0.50076	0.83204	0.62419	-
	rouge_L	recal	0.60817	0.90206	0.78859	-
BLEURT			0.74787	0.73536	0.35939	-
BLEU	blue_1	score	0.5406	0.87316	0.76936	-
		precision	0.8764	0.90576	0.80688	-
	blue_2	score	0.44505	0.83157	0.67622	-
		precision	0.59398	0.82152	0.62334	-
	blue_3	score	0.37568	0.79089	0.59755	-
		precision	0.43396	0.74211	0.48936	-
	blue_4	score	0.32404	0.75189	0.53231	-
		precision	0.33712	0.67018	0.39467	-

Care

Wmo_voorzieningen			ChatGPT	Naive	BERT	Original
LiNT			47	45	33	43
Rouge_1	rouge_1	recal	0.61753	0.90183	0.80489	-
	rouge_2	recal	0.39362	0.83122	0.65976	-
	rouge_L	recal	0.43692	0.89995	0.80019	-
BLEURT			0.62513	0.68564	0.52028	-
BLEU	blue_1	score	0.38767	0.86475	0.77323	-
		precision	0.85267	0.90684	0.81007	-
	blue_2	score	0.3055	0.82143	0.69577	-
		precision	0.52952	0.81827	0.65589	-
	blue_3	score	0.24258	0.78087	0.62825	-
		precision	0.33641	0.74	0.53663	-
	blue_4	score	0.19843	0.74165	0.56777	-
		precision	0.23889	0.66635	0.43905	-

Regels_pgb			ChatGPT	Naive	BERT	Original
LiNT			43	44	32	47
Rouge_1	rouge_1	recal	0.75641	0.90377	0.75617	-
	rouge_2	recal	0.51204	0.83886	0.57032	-
	rouge_L	recal	0.59615	0.90223	0.75	-
BLEURT			0.66788	0.77866	0.29067	-
BLEU	blue_1	score	0.67914	0.85129	0.7146	-
		precision	0.78803	0.90236	0.75748	-
	blue_2	score	0.55015	0.80455	0.61266	-
		precision	0.51712	0.80599	0.55678	-
	blue_3	score	0.45461	0.75998	0.5259	-
		precision	0.36021	0.7188	0.41074	-
	blue_4	score	0.38566	0.71817	0.45419	-
		precision	0.2732	0.64241	0.31013	-

Besluit_pgb			ChatGPT	Naive	BERT	Original
LiNT			48	43	30	40
Rouge_1	rouge_1	recal	0.60591	0.90746	0.80238	-
	rouge_2	recal	0.42469	0.84347	0.65672	-
	rouge_L	recal	0.52463	0.90547	0.79841	-
BLEURT			0.66899	0.63879	0.3776	-
BLEU	blue_1	score	0.3972	0.87275	0.77832	-
		precision	0.8255	0.92229	0.81377	-
	blue_2	score	0.33269	0.834	0.70161	-
		precision	0.57912	0.84221	0.66126	-
	blue_3	score	0.28377	0.798	0.62895	-
		precision	0.42905	0.77207	0.52846	-
	blue_4	score	0.24584	0.76326	0.56211	-
		precision	0.3322	0.70576	0.41955	-

Figure 9: Automatic evaluation metric results of the simplification models.

F. Table with descriptive statistics of variables from the randomized controlled trial

variable	mean	sd	median	min	max	range	skew	kurtosis	se
group	4.5	2.307367258	4.5	1	8	7	0	-1.286697163	0.271925839
age	41.93055556	17.57104403	43	18	83	65	0.371331742	-1.072921957	2.070767398
gender	0.5	0.503508815	0.5	0	1	1	0	-2.027584877	0.059339083
language	0.902777778	0.298339169	1	0	1	1	-2.66263155	5.161875509	0.035159608
edu	2.513888889	0.787097638	3	1	3	2	-1.155913305	-0.405255457	0.092760346
reading_work	3.708333333	2.497533995	4	0	8	8	-0.056788253	-1.28511434	0.294337204
reading_spare	2.152777778	1.61122578	2	0	12	12	3.174404923	16.90846477	0.189884779
disability	0.180555556	0.38734884	0	0	1	1	1.626480752	0.655114473	0.045649499
letters	13.51388889	10.268224	10	2	50	48	1.437533288	1.521199667	1.210121803
grade_clarity	6.736111111	1.861341751	7	2	10	8	-0.61367629	-0.469347613	0.219361229
grade_tone	6.916666667	1.535954078	7	3	10	7	-0.414277684	-0.172643999	0.181013924
l1_g	0.5	0.503508815	0.5	0	1	1	0	-2.027584877	0.059339083
l2_g	0.5	0.503508815	0.5	0	1	1	0	-2.027584877	0.059339083
l3_g	0.5	0.503508815	0.5	0	1	1	0	-2.027584877	0.059339083
l1_action_good	82.06018519	17.83793556	83.33333333	41.66666667	100	58.33333333	-0.636727324	-0.820435435	2.102220866
l1_understanding_good	78.62103175	15.64798818	82.14285714	35.71428571	96.42857143	60.71428572	-0.696801138	-0.452457676	1.844133093
l1_tone	7.013888889	1.605387545	7	2	10	8	-0.465337643	0.10820614	0.189196737
l1_u_grade	5.972222222	1.887484519	6.75	2	8	6	-0.605014758	-0.974974131	0.222442184
t1	16.77777778	6.426665887	17	6	35	29	0.642435057	0.325082181	0.757389838
l2_action_good	3.930555556	2.844943074	2.5	1	8	7	0.221212221	-1.720206341	0.335279757
l2_understanding_good	12.08333333	5.343813061	13.5	1	18	17	-0.420209405	-1.23082262	0.629774409
l2_tone	6.652777778	1.548827109	7	1	9	8	-1.033864164	1.592069942	0.182531025
l2_u_grade	5.854166667	1.372356614	6	2	8	6	-0.53682276	0.296717762	0.161733778
t2	14.94444444	4.515665655	15	6	30	24	0.167020914	0.465562666	0.532176301
l3_action_good	3.5	2.455232986	4	1	7	6	0.160466908	-1.707572903	0.289351982
l3_understanding_good	8.763888889	3.151092167	10	1	12	11	-0.531208236	-0.965990445	0.371359773
l3_tone	6.923611111	1.66923218	7	1	10	9	-0.849591298	1.093499008	0.196720899
l3_u_grade	6.909722222	1.655818569	7	1	10	9	-0.93570467	1.211396159	0.19514009
t3	8.819444444	4.193645054	8	3	20	17	0.958874797	0.306674783	0.494225809
t_totaal	40.54166667	11.34595231	42.5	22	70	48	0.295444261	-0.442937443	1.337133303
average_g_tone	6.863425926	1.44819957	7	1.333333333	9.333333333	8	-0.808477719	1.465542263	0.170671956
average_g_understanding	6.24537037	1.310536398	6.5	1.666666667	8.333333333	6.666666666	-0.833036412	0.709394273	0.154448196
t_action_good	85.87655644	8.915723494	9.413333335	67.45888889	100	32.54111111	-0.284391643	-0.72255892	1.05072809
t_understanding_good	79.95641992	8.514992439	8.080876007	57.87545788	93.95604396	36.08058608	-0.60070304	0.157339839	1.003501483

Table 1: Descriptive statistics of the variables from the randomized controlled trial.

G. Results (Generalised) Linear Models analyses

	(g)lm_letter_1		(g)lm_letter_2		(g)lm_letter_3		(g)lm_totaal	
	l1_action_good	l1_understaning_good	l2_action_good	l2_understanding_good	l3_action_good	l3_understanding_good	t_action_good	t_understanding_good
l1_g	**9.26** ***	**10.94** ***	-	-	-	-	**4.96** ***	**7.36** ***
l2_g	-	-	**6.12** ***	**11.80** ***	-	-	**3.82** ***	**4.86** ***
l3_g	-	-	-	-	**5.58** ***	**7.81** ***	**2.46** *	**3.40** **
age	**-2.96** **	-0.71	**-2.17** *	-1.34	-1.19	-0.51	**-3.77** ***	-1.50
gender	-0.46	0.74	-0.11	0.43	-0.04	-0.31	-0.33	0.48
language	-0.1	0.70	**3.60** ***	-0.53	0.36	1.11	**2.06** *	0.69
edu	-0.25	1.30	1.09	-0.34	-0.06	**-1.80** .	0.45	-0.35
reading_work	-0.73	0.41	-0.13	-0.15	-0.79	0.69	-0.97	0.13
reading_spare	**2.03** *	**2.20** *	-0.60	-0.51	1.02	-0.43	1.60	0.74
disability	-0.86	0.41	0.41	-0.03	**2.00** *	1.00	0.90	0.44
letters	0.13	0.51	1.61	**2.37** *	0.79	-0.20	1.26	1.30
grade_clarity	**2.50** *	**1.98** .	1.40	**2.53** *	1.49	0.48	**3.23** ***	**2.73** **
grade_tone	**-2.34** *	-1.62	-0.15	-0.62	-0.15	0.59	-1.63	-1.14
N	72	72	72	72	72	72	226	226
AIC (glm)	573.35	536.86	578.90	526.83	577.33	520.66	493.25	468.91
Multiple R-squared (lm)	0.63	0.71	0.51	0.74	0.42	0.58	0.53	0.63
Adjusted R-squared (lm)	0.56	0.65	0.42	0.69	0.32	0.50	0.43	0.55
F-statistic (lm)	9.15	13.19	5.70	15.66	4.00	7.43	5.13	7.81

Figure 10: Results of both the Generalised Linear Models analyses and Linear Models with significant values in **bold**. The codes for significance are: **.** p<0.10, * p<0.05, ** p<0.01 and ***p<0.001. *l1_g,l2_g, l3_g* present the dummy variables for the three letters respectively. The model descriptives are defined per model on the bottom lines of the table.

H. Results MANOVA analyses

	manova_letter_1	manova_letter_2	manova_letter_3	manova_totaal
l1_g	**0.75** ***	-	-	**0.53** ***
l2_g	-	**0.76** ***	-	**0.38** ***
l3_g	-	-	**0.68** ***	**0.28** **
age	**0.32** ***	**0.26** **	**0.31** ***	**0.40** ***
gender	0.03	0.08	0.05	0.09
language	0.11	**0.27** **	0.09	**0.20** *
edu	0.10	**0.21** *	0.16	**0.16** .
reading_work	0.12	0.03	0.10	0.05
reading_spare	0.11	0.07	0.09	0.09
disability	0.08	0.02	0.11	0.03
letters	0.04	0.09	0.02	0.05
grade_clarity	0.04	**0.17** .	0.08	**0.16** .
grade_tone	**0.21** *	**0.17** .	0.09	**0.22** *
N	72	72	72	226

Figure 11: Results MANOVA analyses: Pillai's trace values with significant values in **bold**. The codes for significance are: **.** p<0.10, * p<0.05, ** p<0.01 and ***p<0.001. *l1_g,l2_g, l3_g* present the dummy variables for the three letters respectively.