AI vs. Human: Effectiveness of LLMs in Simplifying Italian Administrative Documents

Marco Russodivito[1,†], Vittorio Ganfi[1,*,†], Giuliana Fiorentino[1] and Rocco Oliveto[1]

[1]University of Molise, Italy

Abstract

This study investigates the effectiveness of Large Language Models (LLMs) in simplifying Italian administrative texts compared to human informants. This research evaluates the performance of several well-known LLMs, including *GPT-3.5-Turbo*, *GPT-4*, *LLaMA 3*, and *Phi 3*, in simplifying a corpus of Italian administrative documents (*s-ItaIst*), a representative corpus of Italian administrative texts. To accurately compare the simplification abilities of humans and LLMs, six parallel corpora of a subsection of *ItaIst* are collected. These parallel corpora were analyzed using both complexity and similarity metrics to assess the outcomes of LLMs and human participants. Our findings indicate that while LLMs perform comparably to humans in many aspects, there are notable differences in structural and semantic changes. The results of our study underscore the potential and limitations of using AI for administrative text simplification, highlighting areas where LLMs need improvement to achieve human-level proficiency.

Keywords

Automatic Text Simplification, Large Language Models, Italian Administrative language

1. Introduction

Due to the increasing popularity of generative Artificial Intelligence (AI) language tools [1, 2], significant attention has been devoted to the use of LLMs for text simplification [3]. Several studies have addressed the application of LLMs to simplify texts, particularly focusing on administrative documents, including those in Italian [4, 5, 6]. Italian administrative texts are often notably complex and obscure [7, 8, 9], which restricts a large segment of the population from fully accessing the content produced by the Italian public administration [10, 11].

This work aims to (a) evaluate the quality of automatic text simplification performed by several well-known LLMs, and (b) compare LLM-based simplification with human-based simplification. To address these research questions, the following procedures were undertaken:

1. From an *empirical perspective*, a large corpus of Italian administrative texts was collected (*i.e., ItaIst*). A parallel simplified counterpart of the corpus was created using different LLMs. Additionally, a shorter version of the administrative corpus was manually simplified by two annotators.
2. From an *analytical perspective*, several statistical analyses were conducted to measure the semantic and complexity closeness between human and LLM-generated data. The comparison of scores for both LLM and human datasets highlights significant differences and similarities in manual and AI-driven simplification.

The results concerning readability indexes (*e.g.,* Gulpease) and semantic and structural similarities (*e.g.,* edit distance) reveal that LLMs generally perform comparably to human informants. However, AI-simplified texts are slightly less similar to the original documents than those generated by human simplifiers. LLMs tend to introduce more changes in the simplified corpora than human annotators. The empirical study indicates that texts simplified by AI exhibit more structural and lexical dissimilarities from the original documents than those simplified by humans.

Replication package. All the codes and data are available on Figshare at https://figshare.com/s/4d927fe648c6f1cb4227.

2. Related Work

Several researchers have conducted research on evaluating the accountability of LLMs in text simplification and on assessing the metrics employed to measure the quality of LLM text simplification [12, 13, 14, 15, 16]. In particular, numerous studies have focused on assessing the use of LLMs to simplify Italian administrative texts, highlighting the potential of these models to enhance text readability. Some studies have specifically evaluated the readability of simplified administrative texts

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.
†These authors contributed equally.
marco.russodivito@unimol.it (M. Russodivito); vittorio.ganfi@unimol.it (V. Ganfi); giuliana.fiorentino@unimol.it (G. Fiorentino); rocco.oliveto@unimol.it (R. Oliveto)
0009-0004-8860-1739 (M. Russodivito); 0000-0002-0892-7287 (V. Ganfi); 0000-0002-0392-9056 (G. Fiorentino); 0000-0002-7995-8582 (R. Oliveto)

by comparing parallel corpora of simplified documents and adopting a qualitative interpretative approach [17]. Other contributions have assessed the outputs of LLMs in simplification tasks, particularly focusing on models partially trained on Italian [18].

Our paper analyzes the differences between LLM and human simplification of Italian administrative texts, following a quantitative approach. By examining these differences, our study aims to highlight the similarities and dissimilarities that emerge during the simplification of administrative documents by humans and AI.

3. Study Design

Our study aims to analyze the effectiveness of modern LLMs in simplifying administrative text. To achieve this, we address the following Research Question (RQ):

> *How effective are AI systems at simplifying administrative texts compared to humans?*

This question evaluates whether modern AI can achieve a level of quality comparable to human experts, our references, by analyzing how well LLMs can reduce complexity while preserving the original meaning of the texts.

The study has been conducted on a sub-corpus of *ItaIst*, utilizing several LLMs to support the text simplification process.

3.1. Corpus

The *ItaIst* corpus has been created as part of the VerbACxSS research project. It was composed by linguists and jurists to create a representative linguistic resource for contemporary administrative Italian [19, 20]. *ItaIst* was assembled by collecting recent official documents from local and regional public administration websites of eight Italian regions (Basilicata, Calabria, Campania, Lazio, Lombardy, Molise, Tuscany, and Veneto) covering topics such as *garbage*, *healthcare*, and *public services*. The corpus includes a variety of text types, such as *Tenders Notices*, *Planning Acts*, *Services Charters*.

The reliability of the corpus design was ensured by (a) linguists, who checked the corpus represents administrative Italian in terms of textual and diatopic features, and (b) jurists, who selected and validated each document included in *ItaIst*. The resulting corpus, comprising 208 documents, consists of around $2,000,000$ tokens and $45,000$ types[1]. More information about the *ItaIst* corpus can be found in Appendix A.

To make a fair comparison between humans and AI, a sub-corpus of *ItaIst* (hereinafter, *s-ItaIst*) was extracted. The *s-ItaIst* sub-corpus was composed by selecting representative documents from each region, balancing the topics and text types of the main corpus. Table 1 provides a summary of the *s-ItaIst*.

[1]https://huggingface.co/datasets/VerbACxSS/ItaIst

Table 1
An overview of the main metrics of the *s-ItaIst* corpus.

Metrics	Value
# documents	8
# sentences	1,314
# tokens	33,295
# types	5,622

3.2. LLMs

To investigate both open-source and commercial models, the *s-ItaIst* corpus was simplified using four distinct commercial LLMs, namely *GPT-3.5-Turbo* [21] and *GPT-4* [22] by OpenAI, *LLaMA 3* [23] by Meta, and *Phi 3* [23] by Microsoft. For open-source models, we used the *LLaMA 3* 8B[2] and *Phi 3* 3.8B[3] variants, both fine-tuned on large Italian corpora. This selection explores models of various sizes while ensuring optimal performance for Italian tasks.

A detailed prompt was formulated to instruct each model to perform the simplification task properly, avoiding summary and applying state-of-the-art simplification rules [9]. The full prompt can be found in Appendix B.

The OpenAI models were accessed via APIs[4], while the open-source models were hosted on an AWS EC2 G6[5] instance equipped with a single Nvidia L4 GPU with 24GB vRAM.

3.3. Experimental Procedure

To address our research question, we conducted an empirical study to compare automatic and manual simplifications. Our study, illustrated in Figure 1, can be summarized in three main steps: (i) constructing a corpus of administrative documents (*i.e., s-ItaIst*), (ii) simplifying this corpus using four LLMs and two human annotators, and (iii) comparing the LLM-simplified corpora with the human-simplified corpora.

It is worth noting that the *s-ItaIst* corpus was subdivided into small sections (2-6 sentences) to avoid exceeding the context windows of the LLMs and to facilitate human informants during simplification[6].

[2]https://huggingface.co/DeepMount00/Llama-3-8b-Ita (last seen 07-21-2024)
[3]https://huggingface.co/e-palmisano/Phi3-ITA-mini-4K-instruct (last seen 07-21-2024)
[4]https://openai.com/api/ (last seen 07-21-2024)
[5]https://aws.amazon.com/it/ec2/instance-types/g6/ (last seen 07-21-2024)
[6]*s-ItaIst* corpus was segmented into a total of 619 sections of text. Each section, then, was assigned to human annotators and LLMs for simplification.

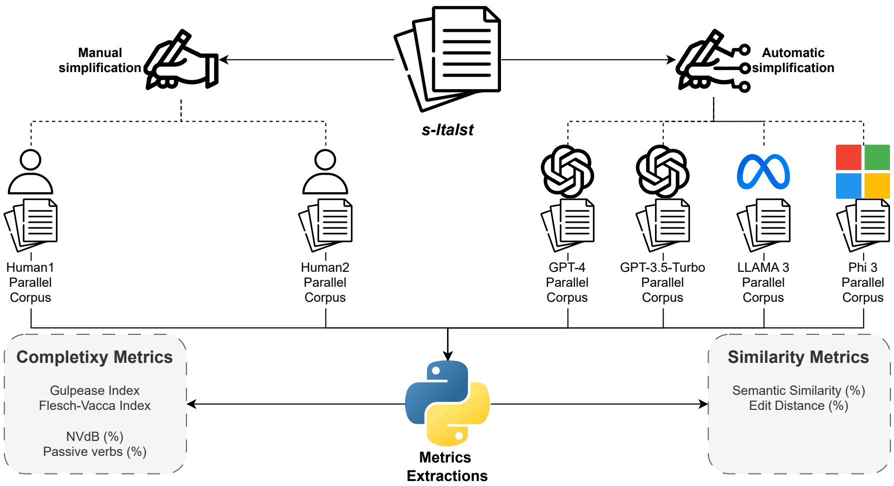

Figure 1: Experimental design schema: The *s-ItaIst* corpus was simplified both automatically and manually by two humans and four LLMs. The resulting parallel corpora were analyzed using complexity and similarity metrics.

Human annotators with strong backgrounds in linguistics and deep knowledge about administrative text simplification simplified the corpus following common simplification rules identified in the literature [24, 25, 8, 9]. They exploited a custom web application that (i) assigned sections of the document to simplify and (ii) tracked the time they spent during such an activity. Similarly, each LLM was instructed to automatically simplify every document in the corpus one section at a time.

This approach provided a comprehensive comparison dataset of six distinct parallel corpora. We analyzed these data to compare human and automatic simplifications by extracting features such as complexity and similarity metrics to measure the quality of the simplified texts and their relatedness to the original text. Furthermore, we computed the *Wilcoxon Signed-Rank Test* [26] to statistically evaluate the difference between LLMs and human metrics and *Cliff's Delta* [27, 28] to provide a measure of the effect size.

3.4. Metrics

To assess the quality of the simplifications, we employed both complexity and similarity metrics from the literature. Complexity metrics compare the ease of the original and simplified text, while similarity metrics measure the distance between them. We implemented these metrics according to the state-of-the-art, leveraging natural language processing (NLP) techniques (*e.g.,* tokenization, POS tagging[7]).

[7]The process of tokenization and tagging was conducted using the spaCy natural language processing tool: https://spacy.io (last seen 07-21-2024)

In literature several simplicity measures (for instance, SAMSA [29], and SARI [30]) are employed, although their results may vary depending on the level of analysis examined and, of course, on the design of the metrics. Therefore, SAMSA aims to measure structural simplicity through monitoring sentence splitting accuracy, and SARI was developed to measure the simplicity advantage when just lexical paraphrasing was evaluated. Furthermore, some study shows that when calculated using multi-operation manual references, both a generic metric like BLEU [31] and an operation-specific one like SARI have low associations with assessments of overall simplicity[32]. Thus, to measure the readability of investigated corpora we selected

1. *Flesch Vacca Index*, *Gulpease Index* and *READ-IT*, since they are advanced instruments designed to investigate the degree of simplicity of Italian texts, and
2. percentages of some lexical and structural features (*i.e.,* amount of most common lexical items and active verb forms) increasing the readability of texts.

Also for similarity metrics, computational literature offers several resources aiming to measure the structural or semantic proximity of texts. Some of these operate at the *n-gram* overlap (*e.g.,* BLEU [31] and METEOR [33]), while others consider other features. For this analysis, we select *Semantic Similarity* to quantify the degree of semantic closeness between corpora and *Edit distance* to measure structural similarities between investigated corpora.

To support future research, we have made our metrics implementation publicly available[8].

Details concerning considered complexity metrics herein are shown:

- ***Gulpease Index*** [34]: This metric evaluates the readability of an Italian text and assesses the education level required to fully comprehend it. It is calculated using the following formula:

$$89 + \frac{300 * (sentences) - 10 * (characters)}{tokens} \tag{1}$$

- ***Flesch Vacca Index*** [35]: This is an adaptation of the original *Flesch Reading Ease* formula for evaluating the readability of Italian texts, computed as follows:

$$217 - 130 * \frac{syllables}{tokens} - \frac{tokens}{sentences} \tag{2}$$

- ***READ-IT*** [36]: The tool is the first advanced readability evaluation instrument for Italian, combining traditional raw text features with lexical, morpho-syntactic, and syntactic information. Four different readability models are included in the tool: *READ-IT BASE* includes only raw features, calculating sentence length (average number of words per sentence) and word length (average number of characters per word); *READ-IT LEXICAL* combines raw (*e.g.,* word length) and lexical (*e.g.,* Type/Token Ratio) features; *READ-IT SYNTACTIC* employs raw text (*e.g.,* sentence length) and morpho-syntactic (*e.g.,* average number of clauses per sentence) properties; *READ-IT GLOBAL* includes all other features, combining raw text, lexical, morpho–syntactic and syntactic (*e.g.,* the depth of the whole parse tree) features [9].
- ***NVdB (%)***: *"Il Nuovo vocabolario di base della lingua italiana"* [37] consists of fundamental and commonly used words representing the essential lexicon of the Italian language. The ease of a text can be roughly estimated by the number of words listed in the basic vocabulary [38].
- ***Passive (%)***: Overuse of passive voice can lead to ambiguity and complexity, especially for readers who may struggle with comprehension [24, 25, 9]. It is calculated by identifying verbs with `aux:pass` occurring in the Dependency Parsing Tree.

Details concerning considered similarity metrics herein are shown:

- ***Semantic Similarity (%)*** [39]: This metric measures the distance between the semantic meanings of two documents. It can be computed exploiting relevant methodologies from the literature, such as *BERTscore*[40] and *SBERT*[41]. We opted for the latter approach, which leverages cosine similarity between contextual embeddings (obtained through `sentence-transformers` and an open-source multilingual model[10]) to evaluate similarity at the sentence level, encapsulating the overall contextual meaning [42].
- ***Edit distance (%)*** [43]: This metric measures the similarity between two strings based on the number of single-character edits (insertions, deletions, or substitutions) required to transform one text into the other. A value close to zero indicates a relatively minor difference between the two texts, while a high value indicates significant rephrasing.

3.5. Threats to validity

We analyze the validity of our study by examining construct, internal, and external validity. This evaluation helps us understand the strengths and limitations of our methodology and the generalizability of our findings.

Construct validity: The two linguistic experts involved in the manual simplification of the *s-ItaIst* corpus may have produced divergent variants due to their subjective approaches. Despite differences in seniority, both experts have strong linguistic backgrounds (holding PhDs) and several years of experience. Nevertheless, involving two human simplifiers allowed us to explore distinct simplification approaches and compare automatic simplification against two varied benchmarks.

Internal validity: The LLMs used for automatic text simplification, particularly those from HuggingFace, may have been trained on non-administrative texts, potentially introducing issues in the simplified text. However, we relied on state-of-the-art models tested against several benchmarks [44, 45, 46, 47]. Additionally, the *embeddings* for calculating *Semantic Similarity* were obtained through a multilingual model chosen for its high ranking on the MTEB leaderboard[11], particularly for its performance in the *STS22 benchmark (it)* [48].

External validity: Our study focuses on the sub-corpus *ItaIst*, consisting of eight administrative documents. Although the number of documents is relatively small, the corpus includes over $1,000$ sentences. Manual simplification of the corpus took *Human1* and *Human2* 15 and 23 hours respectively. Extending our study to the entire *ItaIst* corpus would have been infeasible. However, the documents of the *ItaIst* sub-corpus were not chosen randomly; they were selected to represent the variety of administrative texts.

[8] https://pypi.org/project/italian-ats-evaluator (last seen 07-21-2024)
[9] http://www.italianlp.it/demo/read-it (last seen 04-10-2024)
[10] https://huggingface.co/intfloat/multilingual-e5-base (last seen 07-21-2024)
[11] https://huggingface.co/spaces/mteb/leaderboard (last seen 07-21-2024)

Table 2
Metrics evaluated across the original corpus and the human and LLM simplified corpora.

	Original	*Human1*	*Human2*	*GPT-3.5-Turbo*	*GPT-4*	*LLaMA 3*	*Phi 3*
Tokens	33,295	34,135	29,755	30,032	31,722	36,035	36,056
Sentences	1,314	1,506	1,744	1,515	1,840	1,944	1,900
Tokens per Sentences	25.33	22.66	17.06	19.53	17.24	18.53	18.97
Sentences per Documents	164.25	188.25	218.00	189.37	230.00	243.00	237.50
Gulpease Index	44.31	49.72	50.64	48.49	**51.34**	50.26	50.16
Flesch Vacca Index	19.97	34.23	33.63	30.33	**36.75**	34.09	33.75
NVdB (%)	73.28	80.44	76.89	78.28	**81.07**	80.18	80.16
Passive (%)	20.87	15.78	17.71	13.99	**12.00**	15.81	15.72
READ-IT BASE (%)	75.91	68.62	51.00	66.61	**55.00**	58.37	57.69
READ-IT LEXICAL (%)	93.64	85.37	89.71	91.96	90.29	77.13	**75.74**
READ-IT SYNTACTIC (%)	63.72	53.14	40.09	38.42	**29.92**	40.97	41.24
READ-IT GLOBAL (%)	86.48	69.24	61.34	68.69	**54.60**	59.26	58.37
Semantic Similarity (%)	-	96.52	**97.26**	96.06	95.80	94.96	94.96
Edit distance (%)	-	35.84	29.20	49.21	52.14	55.48	55.44

4. Results and Discussion

A preliminary analysis of our results, summarized in Table 2, reveals several significant similarities and differences between the human and LLM datasets. For instance, the variation in the number of tokens is similar across both human and LLM corpora, although LLMs generally increase the number of sentences more prominently than human annotators.

Regarding complexity metrics, all the parallel corpora (both human and LLM) exhibit a general increase in readability compared to the original texts. For example, the majority of the corpora improve the *Gulpease Index* readability metric, shifting the difficulty level from *very difficult* to *difficult* for middle school reading levels [34] (except for *Human1* and *GPT-3.5-Turbo*). Additionally, complexity metrics vary similarly across both human and LLM groups, with differences between manual and AI simplifiers not significantly greater than those between *Human1* and *Human2* or among *GPT-3.5-Turbo*, *GPT-4*, *LLaMA 3*, and *Phi 3*.

The analysis of semantic and structural distance metrics from the original *s-ItaIst* shows more pronounced differences between human and LLM datasets. In terms of semantic similarity (*Semantic Similarity*), the *Human1* and *Human2* corpora are closer to the original meaning than the LLM-simplified corpora. These differences are even more pronounced when considering edit distance (*Edit distance*). The percentage of edit distance is higher in the LLM group, with each LLM corpus exceeding the human ones by at least 10%.

Higher degrees of *Semantic Similarity* and lower degrees of *Edit distance* in human corpora indicate that human annotators tend to make fewer changes to the original text compared to LLMs.

As reported in Table 2, *GPT-4* achieved the best results across the majority of metrics (except for *READ-IT LEXICAL*). To validate our outcomes, we performed the *Wilcoxon Signed-Rank Test* and calculated *Cliff's Delta* effect size to analyze the difference between *GPT-4* and human metrics. By examining the results in Table 3, we can assert that:

> GPT-4 *simplifications can be comparable to human simplifications.* GPT-4 *simplifications are negligibly better for complexity metrics, moderately worse for similarity, and largely rephrased compared to human simplifications.*

The results of the *Wilcoxon Signed-Rank Test* and *Cliff's Delta* Effect Size for the other models, though not fully significant, are listed in Appendix C.

A brief extract taken from Original, *Human1*, *Human2* and *GPT-4* parallel corpora, representing the same phrase simplified by the two human annotators and *GPT-4* is shown below [12]:

> **Original:** fatturato minimo annuo, per gli ultimi tre esercizi, pari o superiore al valore stimato del presente appalto
> ***Human1*:** Guadagno in un anno (fatturato minimo annuo) negli ultimi 3 anni di valore uguale o superiore al valore di questo bando
> ***Human2*:** l'ammontare di fatture emesse annualmente, per gli ultimi tre anni, deve essere pari o superiore al valore stimato del presente appalto
> ***GPT-4*:** un fatturato annuo minimo, negli ultimi tre anni, uguale o maggiore al valore stimato dell'appalto

[12] A more extensive example of data regarding human and LLM simplifications collected in the parallel corpora designed for this study can be found in Appendix D.

Table 3
Results of the *Wilcoxon Signed-Rank Test* and *Cliff's Delta* Effect Size performed on *GPT-4*, *Human1*, and *Human2* metrics.

	Metrics	*p-value*	Effect Size	
Human1	*Gulpease Index*	< 0.0001	negligible	↗
	Flesch Vacca Index	< 0.0001	negligible	↗
	NVdB	0.0108	negligible	↗
	Passive	0.0004	negligible	↘
	READ-IT BASE	< 0.0001	small	↘
	READ-IT LEXICAL	< 0.0001	negligible	↗
	READ-IT SYNTACTIC	< 0.0001	small	↘
	READ-IT GLOBAL	< 0.0001	small	↘
	Semantic Similarity	< 0.0001	small	↘
	Edit distance	< 0.0001	large	↗
Human2	*Gulpease Index*	0.0092	negligible	↗
	Flesch Vacca Index	< 0.0001	negligible	↗
	NVdB	< 0.0001	small	↗
	Passive	< 0.0001	negligible	↘
	READ-IT BASE	0.0292	negligible	↗
	READ-IT LEXICAL			
	READ-IT SYNTACTIC	< 0.0001	negligible	↘
	READ-IT GLOBAL	< 0.0001	negligible	↘
	Semantic Similarity	< 0.0001	medium	↘
	Edit distance	< 0.0001	large	↗

In the above syntagmas, the similarities between the simplifications are quite obvious: for example, the technical term *esercizio* or the more ambiguous word *pari* are replaced by the more common lexical equivalents *anno* or *uguale*, respectively.

5. Conclusion

In this study, we investigated the automatic simplification of Italian administrative documents. Our results demonstrate that LLMs can effectively simplify these texts, performing comparably to humans [13].

Among the models examined, *GPT-4* shows superior performance in text simplification, exhibiting significant improvements in complexity metrics. Nonetheless, it is noteworthy that humans tend to maintain a higher level of *Edit distance* and *Semantic Similarity*, ensuring the preservation of the original meaning and structure of the text. In other words, humans—aware of the importance of precise language for these documents—mostly preserved the original meaning and structure, whereas LLMs, while simplifying, tended to rephrase extensively. This rephrasing, although effective in reducing complexity, might inadvertently alter the legal nuances, which are critical in administrative texts.

Despite this limitation, LLMs can serve as valuable support tools for text simplification, significantly accelerating a process that typically requires hours of manual work. By generating initial drafts, LLMs can reduce the workload of human experts, who would then review and refine the AI-generated drafts, ensuring the preservation of the overall meaning and legal integrity of the text. The results achieved in our study indicated that modern LLMs can simplify administrative documents almost as effectively as humans. However, the achieved findings indicate that LLMs are not fully capable of preserving the semantic meaning of the text, tending to rephrase more extensively than humans. This could introduce legal issues into the simplified text. Further study could be conducted to evaluate the juridical equivalence of automatically simplified documents. A manual investigation of our parallel corpus, supervised by expert jurists, may reveal important implications in this sensitive context.

Another promising direction for future research is to investigate the impact of automatic simplification on text comprehension. An additional empirical study could be designed to evaluate whether automatically simplified documents are easier to understand than their original versions.

Additionally, it would be worthwhile to explore different prompting strategies to further improve simplification quality. For instance, few-shot prompting [50] with some manually simplified gold samples could better align LLMs with human style.

[13]Further evidence showing that LLM simplifications preserve the meaning of the original texts was obtained in a study, conducted on the same data. The unpublished research indicated that experienced evaluators, *i.e.,* jurists having administrative competence, agree that LLM simplifications of administrative texts maintain the legal integrity of the original documents [49].

Acknowledgments

This contribution is a result of the research conducted within the framework of the PRIN 2020 (Progetti di Rilevante Interesse Nazionale) "VerbACxSS: on analytic verbs, complexity, synthetic verbs, and simplification. For accessibility" (Prot. 2020BJKB9M), funded by the Italian Ministry of Universities and Research.
Giuliana Fiorentino and Rocco Oliveto are responsible for research question identification, study design, research supervision and data analysis. However, for academic reasons, Section 2, Section 3.1, Section 3.3, Section 4, and Section 5 are attributed to Vittorio Ganfi; and Section 1, Section 3, Section 3.2, Section 3.4 and Section 3.5 to Marco Russodivito.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NIPS), volume 30, 2017.

[2] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP), 2020, pp. 38–45.

[3] M. J. Ryan, T. Naous, W. Xu, Revisiting non-English text simplification: A unified multilingual benchmark, Association for Computational Linguistics (ACL) (2023).

[4] D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, Design and Annotation of the First Italian Corpus for Text Simplification, in: Linguistic Annotation Workshop (LAW), 2015, pp. 31–41.

[5] M. Miliani, S. Auriemma, F. Alva-Manchego, A. Lenci, Neural readability pairwise ranking for sentences in Italian administrative language, in: Asia-Pacific Chapter of the Association for Computational Linguistics(AACL) and International Joint Conference on Natural Language Processing (IJCNLP), 2022, pp. 849–866.

[6] M. Miliani, M. S. Senaldi, G. Lebani, A. Lenci, Understanding Italian Administrative Texts: A Reader-Oriented Study for Readability Assessment and Text Simplification, in: Workshop on AI for Public Administration (AIxPA), 2022, pp. 71–87.

[7] S. Lubello, La lingua del diritto e dell'amministrazione, Il mulino, Bologna, 2017.

[8] M. Cortelazzo, Il linguaggio amministrativo. Principi e pratiche di modernizzazione, Carocci, Roma, 2021.

[9] G. Fiorentino, V. Ganfi, Parametri per semplificare l'italiano istituzionale: Revisione della letteratura, Italiano LinguaDue 16 (2024) 220–237.

[10] E. Piemontese (Ed.), Il dovere costituzionale di farsi capire. A trent'anni dal Codice di stile, Carocci, Roma, 2023.

[11] S. Lubello, Da dembsher al codice di stile e oltre: un bilancio sul linguaggio burocratico, in: E. Piemontese (Ed.), Il dovere costituzionale di farsi capire A trent'anni dal Codice di stile, Carocci, Roma, 2023, pp. 54–70.

[12] G. Gonzalez Delgado, B. Navarro Colorado, The Simplification of the Language of Public Administration: The Case of Ombudsman Institutions, in: Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context, 2024, pp. 125–133.

[13] R. Doshi, K. Amin, P. Khosla, S. Bajaj, S. Chheang, H. P. Forman, Utilizing large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing, medRxiv (2023). doi:10.1101/2023.06.04.23290786.

[14] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, D. Kyriazis, Xai for all: Can large language models simplify explainable ai?, arXiv preprint arXiv:2401.13110 (2024).

[15] Y. Ma, S. Seneviratne, E. Daskalaki, Improving Text Simplification with Factuality Error Detection, in: Workshop on Text Simplification, Accessibility, and Readability (TSAR), 2022, pp. 173–178.

[16] F. Alva-Manchego, C. Scarton, L. Specia, Data-Driven Sentence Simplification: Survey and Benchmark, Computational Linguistics 46 (2020) 135–187.

[17] M. Miliani, F. Alva-Manchego, A. Lenci, Simplifying Administrative Texts for Italian L2 Readers with Controllable Transformers Models: A Data-driven Approach., in: CLiC-it, 2023.

[18] D. Nozza, G. Attanasio, et al., Is it really that simple? prompting language models for automatic text simplification in italian, in: CEUR Workshop Proceedings, 2023.

[19] D. Vellutino, et al., L'italiano istituzionale per la comunicazione pubblica, Il mulino, Bologna, 2018.

[20] D. Vellutino, N. Cirillo, Corpus «itaist»: Note per lo sviluppo di una risorsa linguistica per lo studio dell'italiano istituzionale per il diritto di accesso civico, Italiano LinguaDue 16 (2024) 238–250.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems (NIPS) 33 (2020) 1877–1901.

[22] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[23] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[24] E. Piemontese, Criteri e proposte di semplificazione, in: Codice di stile delle comunicazioni scritte a uso delle pubbliche amministrazioni, Istituto Poligrafico e Zecca dello Stato, Roma, 1994.

[25] A. Fioritto, Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche, Il mulino, Bologna, 1997.

[26] F. Wilcoxon, Probability tables for individual comparisons by ranking methods, Biometrics 3 (1947) 119–122.

[27] N. Cliff, Dominance statistics: Ordinal analyses to answer ordinal questions., Psychological bulletin 114 (1993) 494–509.

[28] N. Cliff, Ordinal methods for behavioral data analysis, Psychology Press, New York, 2014.

[29] E. Sulem, O. Abend, A. Rappoport, Semantic

structural evaluation for text simplification, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 685–696. URL: https://aclanthology.org/N18-1063. doi:10.18653/v1/N18-1063.

[30] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing Statistical Machine Translation for Text Simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://doi.org/10.1162/tacl_a_00107. doi:10.1162/tacl_a_00107.

[31] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[32] F. Alva-Manchego, C. Scarton, L. Specia, The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification, Computational Linguistics 47 (2021) 861–889. URL: https://doi.org/10.1162/coli_a_00418. doi:10.1162/coli_a_00418.

[33] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.

[34] P. Lucisano, M. E. Piemontese, Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana, Scuola e città (1988) 110–124.

[35] V. Franchina, R. Vacca, Adaptation of flesh readability index on a bilingual text written by the same author both in italian and english languages, Linguaggi 3 (1986) 47–49.

[36] F. Dell'Orletta, S. Montemagni, G. Venturi, Read–it: Assessing readability of italian texts with a view to text simplification, in: Proceedings of the second workshop on speech and language processing for assistive technologies, 2011, pp. 73–83.

[37] T. De Mauro, I. Chiari, Il nuovo vocabolario di base della lingua italiana (2016). URL: https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

[38] D. Brunato, F. Dell'Orletta, G. Venturi, Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian, Frontiers in Psychology 13 (2022). doi:10.3389/fpsyg.2022.707630.

[39] D. Chandrasekaran, V. Mago, Evolution of semantic similarity—A survey, ACM Computing Surveys (CSUR) 54 (2021) 1–37.

[40] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[41] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2019.

[42] A. Barayan, J. Camacho-Collados, F. Alva-Manchego, Analysing zero-shot readability-controlled sentence simplification, arXiv preprint arXiv:2409.20246 (2024).

[43] F. P. Miller, A. F. Vandome, J. McBrewster, Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? Levenshtein distance, spell checker, hamming distance, Alpha Press, Olando, 2009.

[44] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, International Conference on Learning Representations (ICLR) (2021).

[45] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, Hellaswag: Can a machine really finish your sentence?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, p. 4791–4800.

[46] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the ai2 reasoning challenge, arXiv preprint arXiv:1803.05457 (2018).

[47] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, M. Gardner, Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 2368–2378.

[48] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive text embedding benchmark, in: European Chapter of the Association for Computational Linguistics (EACL), 2023, pp. 2014–2037.

[49] G. Fiorentino, M. Russodivito, V. Ganfi, R. Oliveto, Validazione e confronto tra semplificazione automatica e semplificazione manuale di testi in italiano istituzionale ai fini dell'efficacia comunicativa, in: Automated texts In the ROMance languages and be-

yond" (AI-ROM-II), 2nd International Conference, To appear.

[50] J. Wang, K. Liu, Y. Zhang, B. Leng, J. Lu, Recent advances of few-shot learning methods and applications, Science China Technological Sciences 66 (2023) 920–944.

Table 5
Results of the *Wilcoxon Signed-Rank Test* and *Cliff's Delta* Effect Size performed on *GPT-3.5-Turbo*, *Human1*, and *Human2* metrics.

	Metrics	*p-value*	Effect Size	
Human1	*Gulpease Index*	< 0.0001	negligible	↘
	Flesch Vacca Index	< 0.0001	negligible	↘
	NVdB	< 0.0001	negligible	↘
	Passive			
	READ-IT BASE	0.0052	negligible	↘
	READ-IT LEXICAL	< 0.0001	negligible	↗
	READ-IT SYNTACTIC	< 0.0001	small	↘
	READ-IT GLOBAL			
	Semantic Similarity	< 0.0001	small	↘
	Edit distance	< 0.0001	medium	↗
Human2	*Gulpease Index*	< 0.0001	small	↘
	Flesch Vacca Index	< 0.0001	negligible	↘
	NVdB	< 0.0001	negligible	↗
	Passive	0.0072	negligible	↘
	READ-IT BASE	< 0.0001	small	↗
	READ-IT LEXICAL	0.0091	negligible	↗
	READ-IT SYNTACTIC			
	READ-IT GLOBAL	0.0003	negligible	↗
	Semantic Similarity	< 0.0001	medium	↘
	Edit distance	< 0.0001	large	↗

Table 6
Results of the *Wilcoxon Signed-Rank Test* and *Cliff's Delta* Effect Size performed on *LLaMA 3*, *Human1*, and *Human2* metrics.

	Metrics	*p-value*	Effect Size	
Human1	*Gulpease Index*	0.0077	negligible	↗
	Flesch Vacca Index			
	NVdB			
	Passive			
	READ-IT BASE	< 0.0001	small	↘
	READ-IT LEXICAL	< 0.0001	negligible	↘
	READ-IT SYNTACTIC	< 0.0001	small	↘
	READ-IT GLOBAL	< 0.0001	small	↘
	Semantic Similarity	< 0.0001	medium	↘
	Edit distance	< 0.0001	large	↗
Human2	*Gulpease Index*			
	Flesch Vacca Index			
	NVdB	< 0.0001	small	↗
	Passive			
	READ-IT BASE	< 0.0001	negligible	↗
	READ-IT LEXICAL	< 0.0001	small	↘
	READ-IT SYNTACTIC			
	READ-IT GLOBAL			
	Semantic Similarity	< 0.0001	large	↘
	Edit distance	< 0.0001	large	↗

Table 7
Results of the *Wilcoxon Signed-Rank Test* and *Cliff's Delta* Effect Size performed on *Phi 3*, *Human1*, and *Human2* metrics.

	Metrics	*p-value*	Effect Size	
Human1	*Gulpease Index*	0.0134	negligible	↗
	Flesch Vacca Index			
	NVdB			
	Passive			
	READ-IT BASE	< 0.0001	small	↘
	READ-IT LEXICAL	< 0.0001	negligible	↘
	READ-IT SYNTACTIC	< 0.0001	small	↘
	READ-IT GLOBAL	< 0.0001	small	↘
	Semantic Similarity	< 0.0001	medium	↘
	Edit distance	< 0.0001	large	↗
Human2	*Gulpease Index*			
	Flesch Vacca Index			
	NVdB	< 0.0001	small	↗
	Passive			
	READ-IT BASE	< 0.0001	negligible	↗
	READ-IT LEXICAL	< 0.0001	small	↘
	READ-IT SYNTACTIC			
	READ-IT GLOBAL			
	Semantic Similarity	< 0.0001	large	↘
	Edit distance	< 0.0001	large	↗

A. Corpus ItaIst

The *ItaIst* corpus is a comprehensive collection of Italian administrative documents. Table 4 provides an overview of the topics and regions from which these documents were collected. This corpus has been assembled to represent the diversity and complexity of contemporary administrative Italian, ensuring its relevance for linguistic and computational analysis.

Table 4
Topics and regions of documents collected in *ItaIst*

	Garbage	**Healthcare**	**Public services**
Basilicata	8	3	9
Calabria	11	5	9
Campania	14	7	9
Lazio	9	3	9
Lombardia	15	3	11
Molise	10	7	9
Toscana	19	4	12
Veneto	9	5	10

B. Prompt engineering

In the context of LLMs, the term *prompt* refers to the instructions provided to a language model to generate a specific response. *Prompt engineering* is the process of designing a clear and detailed *prompt* to instruct the model to generate a desired response. The prompt we used to ask the models to simplify administrative text is:

Sei un dipendente pubblico che deve scrivere dei documenti istituzionali italiani per renderli semplici e comprensibili per i cittadini. Ti verrà fornito un documento *pubblico e il tuo compito sarà quello di riscriverlo applicando regole di semplificazione senza però modificare il significato del documento originale. Ad esempio potresti rendere le frasi più brevi, eliminare le perifrasi, esplicitare sempre il soggetto, utilizzare parole più semplicii, trasformare i verbi passivi in verbi di forma attiva, spostare le frasi parentetiche alla fine del periodo.*

C. Tests

Table 5, Table 6, and Table 7 report the results of the statistical analyses conducted to compare the simplification performance of various LLMs against human experts. The *Wilcoxon Signed-Rank Test* and *Cliff's Delta* effect size were employed to evaluate the metrics of *GPT-3.5-Turbo*, *LLaMA 3*, and *Phi 3* models in comparison to two human simplifiers, labelled as *Human1* and *Human2*. These analyses provide insights into the relative effectiveness of AI-driven simplifications versus human efforts.

D. Examples

Table 8 provides several examples of text simplification. For each example, we present the original text alongside its simplified versions. The values of the complexity and similarity metrics are reported for each text.

Table 8
Examples of simplifications.

	Text	Gulpease Index	Flesch Vacca Index	NVdB (%)	Passive (%)	Semantic Similarity (%)	Edit distance (%)
Original	L'operatore di Polizia Locale, quindi, rappresenta un importante punto di riferimento per la collettività. Nell'ambito delle sue funzioni, esso svolge i propri compiti in maniera autorevole, dando prova di preparazione professionale e sensibilità nel contatto relazionale. La sua attività, inoltre, è caratterizzata dal costante sforzo teso alla migliore interpretazione delle situazioni e delle problematiche incontrate, applicando un approccio educativo e orientato alla più adeguata risposta ai bisogni della cittadinanza.	38	12	77 %	28 %	-	-
Human1	La Polizia Locale è un punto di riferimento per i cittadini. La Polizia Locale ha autorevolezza, professionalità e sensibilità nel contatto con i cittadini. La Polizia Locale cerca sempre di interpretare al meglio situazioni e problematiche incontrate. La Polizia Locale si comporta in modo da educare e rispondere adeguatamente ai bisogni dei cittadini.	55	33	67 %	0 %	93 %	56 %
Human2	L'operatore di Polizia Locale, quindi, è un importante punto di riferimento per la collettività. Quando è in servizio, esso svolge i propri compiti in maniera autorevole. Dimostra preparazione professionale e sensibilità nel contatto con gli utenti. Cerca sempre di interpretare le situazioni e i problemi nel migliore dei modi. Applica un approccio educativo e cerca di adeguarsi ai bisogni della cittadinanza.	58	42	83 %	0 %	98 %	35 %
GPT-4	L'operatore di Polizia Locale è un punto di riferimento importante per la comunità. Svolge i suoi compiti con autorità, mostrando competenza professionale e sensibilità nelle relazioni. Inoltre, lavora sempre per capire al meglio le situazioni e i problemi che incontra, usando un approccio educativo per rispondere adeguatamente ai bisogni dei cittadini.	48	32	84 %	0 %	97 %	48 %
GPT-3.5-Turbo	L'operatore di Polizia Locale è un punto di riferimento importante per la comunità. Nell'esercizio delle sue funzioni, svolge i compiti con autorevolezza, dimostrando professionalità e sensibilità nei rapporti con le persone. Si impegna costantemente a comprendere al meglio le situazioni e le problematiche affrontate, adottando un approccio educativo e orientato a rispondere in modo adeguato ai bisogni dei cittadini.	45	27	78 %	0 %	98 %	45 %
LLaMA 3	L'operatore di Polizia Locale è un punto di riferimento importante per la comunità. Esegue i suoi compiti con autorità, professionalità e sensibilità nel rapporto con la gente. La sua attività è caratterizzata dal costante impegno per comprendere meglio le situazioni e i problemi, e per rispondere in modo educativo ai bisogni dei cittadini, con un approccio orientato alla loro assistenza.	50	37	85 %	28 %	96 %	54 %
Phi 3	L'operatore di Polizia Locale è un punto di riferimento importante per la comunità. Esegue i suoi compiti con autorità, professionalità e sensibilità nel trattare le persone. La sua attività è guidata dal desiderio di capire meglio le situazioni e le problematiche, e di rispondere in modo appropriato ai bisogni dei cittadini, con un approccio educativo.	52	38	82 %	28 %	96 %	56 %