Is It Really That Simple? Prompting Language Models for Automatic Text Simplification in Italian

Debora Nozza[1], Giuseppe Attanasio[1]

[1]Bocconi University, Milan, Italy

Abstract

Recent language models (LMs) that follow instructions have showcased remarkable abilities to tackle diverse natural language processing (NLP) tasks, given appropriate prompts. However, the potential of these models for Automatic Text Simplification (ATS) in Italian remains largely unexplored. In this paper, we pioneer the first in-depth investigation into the capabilities of LMs for performing ATS in Italian. We evaluate six state-of-the-art models on a benchmark Italian ATS dataset of administrative texts, reporting six readability metrics on the generated text. Our findings demonstrate a large variability across models, scales, and prompts. Among the tested models, GPT-3.5 editing capabilities are the most suitable, outperforming, surprisingly, human-written simplification. Furthermore, we shed light on the enigmatic multilingual capabilities of instruction following models, opening up new avenues for research in this domain.

Italiano. Recenti sviluppi nei cosiddetti language models (LMs) basati sull'apprendimento di istruzioni hanno mostrato notevoli capacità nell'affrontare diverse problemi di elaborazione del linguaggio naturale (NLP). Tuttavia, il potenziale di questi modelli per la semplificazione automatica del testo (Automatic Text Simplification o ATS) in italiano rimane in gran parte inesplorato. Questo articolo riporta un'indagine pionieristica sulle capacità dei Language Models (LMs) nell'eseguire ATS in italiano. Abbiamo valutato sei modelli utilizzando un dataset italiano di testi amministrativi, riportando sei metriche di leggibilità sul testo generato. I nostri risultati dimostrano una grande variabilità tra i modelli. Tra i modelli testati, le capacità di editing di GPT-3.5 si sono dimostrate le più adatte, superando, sorprendentemente, anche le semplificazioni scritte da persone. Inoltre, questo articolo evidenzia le enigmatiche capacità multilingue dei LMs, aprendo nuove vie di ricerca in questo ambito.

Keywords

Automatic Text Simplification, Large Language Models, Italian

1. Introduction

Italian administrative texts have long been criticized for their complexity, described as "artificial" and "obscure" [1]. Despite efforts by Italian institutions to encourage the use of plain language in official acts and communications over the past decades [2], the readability of these texts remains a pressing issue [3]. To tackle this challenge, considering the substantial volume of bureaucratic text generated, a logical approach is to embark on the analysis and exploration of Automatic Text Simplification (ATS) methods. Automated text simplification is a natural language processing (NLP) technique that aims to modify complex or difficult-to-understand text into simpler and more accessible language while retaining the original meaning. The goal is to make the content easier to comprehend for a wider audience, including individuals with cognitive or reading difficulties, non-native speakers, or those with limited literacy skills.

Using recent large-scale language models (LMs) is a promising direction in this context. In particular, recent evidence has shown that high-capacity pretrained models, e.g., T5 [4] or LLaMA [5], can be further improved via instruction fine-tuning (IFT) and reinforcement learning from human feedback (RLHF) [6, 7, 8, *inter alia*]. The resulting model can follow instructions as expressed via natural language, i.e., it can solve many NLP tasks and reply to various user requests with no architectural changes.

This paper presents the first investigation to look into the capabilities of instruction following language models for Automatic Text Simplification on Italian administrative texts. We rely on Admin-It [9], a benchmark parallel corpus in the Italian administrative language that contains sentences that have been simplified using three distinct rewriting techniques. We perform a thorough evaluation of six models based on six different readability measures tailored for Italian. Each model is compared to the readability scores of the original administrative text and the simplified version provided in the parallel corpus.

Contribution We propose the first in-depth study on whether current IFT models can simplify written passages in Italian. We report a large variability across models, with proprietary GPT-3.5 being the most suitable solution. In addition, we introduce a novel metric to better account for accurate and simple generations. We

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
debora.nozza@unibocconi.it (D. Nozza); giuseppe.attanasio3@unibocconi.it (G. Attanasio)

release code and data to facilitate future research.[1]

2. Automatic Text Simplification

Automatic text simplification is a research field in computational linguistics that studies methods and techniques to simplify textual content [10]. This task involves transforming complex or difficult-to-understand text into more straightforward and accessible language. Automatic text simplification has been viewed as a critical technique for increasing the inclusion of people with special needs and boosting social inclusion [10].

To simplify a text, strategies might involve sentence- or word-level interventions (e.g., breaking down longer passages into multiple sentences or changing less common words with easier equivalents). Most importantly, such edits can be learned, and NLP models can be applied to automate and generalize.

Automatic text simplification has typically focused on two distinct tasks: lexical simplification and syntactic simplification, each of which addresses a different subproblem in the larger task of making texts easier to read and understand [10]. The goal of *lexical simplification* is to make a document easier to understand by either changing the vocabulary to use terms that are more likely to be familiar to the reader or by providing clearer definitions of unfamiliar words. Whereas the purpose of *syntactic simplification* is to detect syntactic phenomena in phrases that may obstruct readability and understanding, with the hope of rewriting the sentence in a way that makes it easier to read and comprehend (by, for example, changing it from the passive to the active voice).

2.1. Dataset

The Admin-It corpus [9] collects Italian sentences from the administrative context, one of the domains where complex language is more frequent. The parallel corpus counts 736 sentence pairs. Each sample reports the original, complex sentence and a simplified version. The corpus was created by combining three subsets based on the nature of the applied simplification:

- **Operations** (Admin-It$_{OP}$): 588 pairs of sentences (∼80% of the total dataset) from the subset of the Simpitiki corpus [11] related to the administrative domain. A single simplification operation is used to simplify the sentences (e.g., split, reorder, merge, lexical substitutions).
- **Rewritten Sents** (Admin-It$_{RS}$): 100 pairs of sentences (∼14% of the total dataset) from websites of Italian municipalities and the Pawac Corpus [12]. Sentences were manually simplified both at lexical and syntactic levels.
- **Rewritten Docs** (Admin-It$_{RD}$): 48 pairs of sentences (∼7% of the total dataset) from administrative documents collected and simplified by Cortelazzo [13]. Sentences were rewritten according to linguistic simplification and communicative effectiveness criteria.

In this paper, we refer to the entire corpus acquired by combining these three subsets as Admin-It.

Model	Params (B)	IFT Data
Flan-T5-XXL	11	FLAN
Vicuna v1.3	7, 13, 33	ShareGPT
Camoscio	7	Alpaca (Ita)
Guanaco	65	OpenAssistant
Llama 2 Chat*	70	Not disclosed
GPT-3.5*	170	Not disclosed

Table 1
Summary of the tested models, the number of learnable parameters, and the Instruction Fine-Tuning dataset used for training. *: optimized with RLHF.

3. Models

Recent advances in instruction tuning have shown that it is possible to build a single model that, if prompted accordingly, can solve a wide range of tasks. Here, we experiment with two families of instruction-tuned models: plain, supervised instruction fine-tuning (IFT) and reinforcement learning from human feedback (RLHF).

Instruction Fine-Tuning (IFT) typically requires a pretrained base model and a fine-tuning step where the latter is specifically taught how to generate text to follow instructions. The choice of base model, fine-tuning data, and regime drastically influence the capacity of the resulting model. Roughly, RLHF mixes standard IFT and policy learning to follow human preferences.

We divide the tested models into three categories, namely FLAN models [6], IFT models using LLaMA [5] or Llama-2 [8] for the base model, and models fine-tuned using the standard RLHF procedure as described in Ouyang et al. [7]. Table 1 summarizes the models tested in this study.

3.1. FLAN Models

FLAN models are fine-tuned on a large collection of NLP tasks verbalized to natural language [14]. The verbalization follows a task-dependent template–e.g., “`Translate the following sentence from`

[1]Code and data available at: https://github.com/MilaNLProc/prompting-italian-text-simplification

Hyperparameter	Value
Temperature	0.7
Top P	1.0
Top K	50
Repetition Penalty	1.1
Penalty Alpha	0.2
Length Penalty	1.2
Max new tokens	512

Table 2
Decoding configuration.

`{src_lang} to {tgt_lang}: {src_text}`" is one of the template used for machine translation.

Although FLAN does not include specifically tasks related to language simplification, we hypothesize that 1) pretraining data, 2) the presence of tasks that share some of the traits (e.g., summarization), and 3) scale enable models to simplify language. We experiment wtih Flan-T5-XXL (11B), the largest T5-based FLAN model.

3.2. LLaMA IFT Models

Since LLaMA [5] established as the best performing pretrained base model on many language understanding tasks, several works used it as base model for IFT.

We test Vicuna v1.3 7B, 13B, and 33B [15]. These models have been trained with IFT on a corpus of around 70K conversations from the ShareGPT website.[2] We test also Guanaco (65B) [16], an IFT model fine-tuned on around 10K conversations from the Open Assistant project.[3]

As an additional baseline for the Italian, we include Camoscio [17], a LLaMA model instruction fine-tuned on samples exclusively in Italian. The fine-tuning corpus includes around 52K instructions from the Alpaca dataset [18] machine-translated with GPT-3.5.

3.3. RLHF Models

Reinforcement learning from human feedback (RLHF) introduces an additional step to the standard IFT pipeline. After the supervised fine-tuning stage, a policy learning step maximizes the *alignment* with human preferences by teaching the model to produce responses that are more likely to be preferred by human users [19].

We experiment with GPT-3.5 [7, gpt-3.5-turbo, last accessed June 15, 2023] and Llama 2 Chat (70B) [8].

[2] https://sharegpt.com/
[3] https://open-assistant.io/

4. Zero-Shot Simplification in Italian

As a result of multilingual pretraining, fine-tuning, or RLHF data, IFT models have shown multilingual abilities, such as solving cross-lingual tasks (e.g., machine translation), or understanding and providing coherent responses to non-English input queries [20].

We leverage this finding and prompt models to run text simplification in Italian in a zero-shot setup. Specifically, we compile a request for simplification using a given prompt template, feed it to the model, and take the model response unmodified. For Vicuna, Guanaco, Llama 2 Chat, and GPT-3.5 we use model-specific system message templates (see Appendix A). We specify no system message or use any prompt template for Flan-T5 and Camoscio.

Figure 1 displays a system overview.

Prompt Template Recent evidence has shown that different prompts elicit multilingual capabilities differently [21]. Therefore, we experiment with two templates, both starting with a prefix stating the task followed by the passage to simplify.

In our *explicit* template (Template-EN), we state overtly the response should be in Italian, i.e., "`Simplify the following text. Write only the response, and in Italian.\n{src_text}`", where `src_text` is the passage to simplify. We also experiment with an *implicit* template (Template-IT), where the entire prompt is written in Italian to hint models to reply in the same language: "`Semplifica il testo seguente.\n{src_text}`" (eng: "Simplify the following text.").

Decoding Setup We use a standard decoding configuration, loosely inspired by Vicuna's Chat Arena[4] for all the models. Table 2 reports the generation configuration used. We use models and code as released in HuggingFace transformers [22] and simple-generation [23] to run inference.

5. Metrics

We conducted an evaluation of automatic text simplifications using benchmark readability metrics, which we categorize into *traditional*, namely Flesh-Kincaid test and GulpEase index, and READ-IT-based metrics.

The **Flesh-Kincaid test** (↑) [24] is a widely utilized measure in education for assessing the readability level of books. In this context, we refer to the Flesch-Vacca formula, designed specifically for Italian text.

[4] https://chat.lmsys.org/?arena

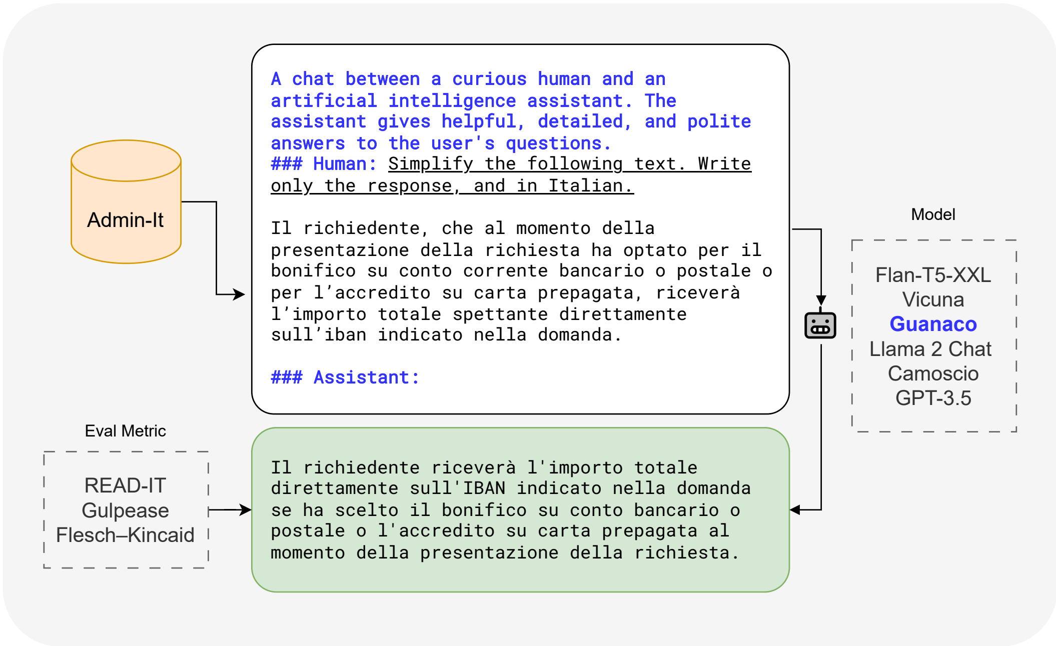

Figure 1: Overview of our zero-shot prompted language simplification in Italian on the Admin-It corpus. Prompt (white box) constructed using 1) a model-specific system message (dark blue, here shown Guanaco), 2) a custom prefix to elicit a response in Italian, and 3) the text to simplify.

	BERTScore		% Detected Italian	
	Template-IT	Template-EN	Template-IT	Template-EN
Camoscio	0.64	-	93	-
Flan-T5-XXL	-	**0.88**	-	**99**
GPT-3.5	-	0.85	-	**99**
Guanaco-65B	0.66	0.65	96	96
Llama-2-Chat-70B	0.63	0.63	15	14
Vicuna-7B	0.67	0.67	62	61
Vicuna-13B	0.70	0.69	87	87
Vicuna-33B	0.70	0.70	92	91

Table 3
Scores to evaluate adherence to gold simplified text (BERTScore) and model consistency in providing Italian responses (% Detected Italian). Note that the scores related to the original (complex) text are 0.95 and 100% respectively.

The **GulpEase index** (↑) [25] calculates text readability based on factors such as word length (measured in letters), number of words, and sentence length. It does not have a direct association with any particular language.

READ-IT (↓) [26] is a machine learning-based readability metric. The model has been trained to evaluate the readability of a text using various features. Different variations of the READ-IT metric exist: *base* employs basic features like sentence and word length; *lexical* focuses on lexical features, such as vocabulary complexity; *syntax* considers grammatical features like syntactic tree depth and part-of-speech categories; *all* combines all the aforementioned variations.

6. Results

This section illustrates the results of prompting instruction following models for generating simplified versions of an input text. We first perform a preliminary investigation on the generated outputs. Based on this analysis, we discover that the benchmark readability metrics are not ideal in our setup, as models produce non-relevant responses. Therefore, we propose a novel *adjusted* score

	Traditional ↑		READ-IT ↓				
	Gulpease	Flesch–Kincaid	*base*	*lexical*	*syntax*	*all*	*all_adjusted*
Complex Text	41.90	30.00	30.01	68.26	83.89	85.28	-
Simplified Text	43.83	36.34	28.28	63.28	78.63	79.67	-
Camoscio	45.61	45.27	66.47	71.06	72.52	74.53	26.83
Flan-T5-XXL	43.12	33.40	33.08	67.23	83.27	83.99	10.08
Vicuna-7B	25.73	22.63	62.61	77.42	82.75	83.83	27.66
Vicuna-13B	37.49	34.51	50.77	69.11	79.33	80.67	24.20
Vicuna-33B	38.43	30.67	54.92	72.16	83.75	85.43	25.63
Guanaco	**55.81**	**61.66**	43.21	**46.51**	61.99	**61.42**	20.88
Llama 2 Chat	7.11	7.55	93.16	93.83	94.78	95.03	35.16
GPT-3.5	46.01	40.76	**21.60**	59.28	**60.43**	61.74	**9.26**

Table 4
Text readability scores on Admin-It for the original (complex) text, human simplification, and automatically generated simplified versions. The best results are bold, except for the proposed final metric, which is also underlined.

to better measure improved readability and adherence to the original text.

6.1. Inspecting Generated Responses

Tables 5 and 6 (Appendix B) illustrate two examples extracted from the Admin-It dataset. In the first instance, the complex sentence uses administrative jargon related to numbers and dates, while the manually-simplified text conveys the same concept using more straightforward verbs, e.g., "assumere l'ufficio di" (*eng*: to get the role of) is replaced with "essere" (*eng*: to be). However, the model-generated texts exhibit undesired behaviors: the automatic simplifications are not consistently simpler, some are not written in Italian, some result in drastically longer passages, incorporate prompt-related content, or occasionally add irrelevant information. The second example presents a similar case, wherein the model-generated simplification includes code, questions, and apparent errors likely produced by incorrect translations, e.g., "il bambino deve essere vivo" (*eng*: the child must be alive).

To investigate the issues raised in our initial qualitative analysis, we conducted two investigations. First, we calculated the adherence of the model-generated simplifications to the human-written reference simplification provided in Admin-It. This metric helps us identify cases where the produced simplifications diverge from the source text, potentially containing code or unrelated questions. For this evaluation, we used BERTScore, a language generation evaluation metric based on pretrained BERT contextual embeddings [27]. Second, we measured the percentage of times the model-generated simplifications are in Italian (% Detected Italian). To accomplish this, we used the Python `langdetect`[5] library. We classified a text as Italian if the library detected the Italian language with a confidence level higher than 0.99.

[5] https://pypi.org/project/langdetect/

Table 3 presents the scores for each model, including variations in Italian and English prompt templates where applicable (see Section 4). As observed in the two examples, only two models, Flan-T5-XXL and GPT-3.5, demonstrate reasonable BERTScore and % Detected Italian metrics. It is crucial to emphasize the discouragingly low % of Italian generations by Llama 2 Chat and the unsatisfactory BERTScore of Camoscio. Additionally, when testing the models with both template configurations, the Italian template tends to yield slightly better results. As a consequence, moving forward, we will consider the Template-IT model version whenever available.

6.2. Automatic Text Simplification Results

Table 4 presents the text readability metrics (see Section 5) for the original (complex) text, its reference human-written simplification, and the model generations. Out of all models, only three — Camoscio, Guanaco, and GPT-3.5 — consistently exhibit readability metrics better than human simplification.

Interestingly, Guanaco yielding the best results in each individual metric is contradicting our findings from the previous section. The issue lies in the fact that **the readability metrics alone do not account for cases when models produce unrelated or inaccurate generations**. For instance, the Guanaco generation shown in Table 6 may be a highly readable sentence (READ-IT$_{all}$ = 96) but has very low adherence to the original text (BERTScore = 0.63).

To address this issue, **we introduced a novel READ-IT metric which also takes into account the original text similarity**, named READ-IT$_{all_adjusted}$. The metric is computed as the product among READ-IT$_{all}$ and BERTScore. By using READ-IT$_{all_adjusted}$, we identify GPT-3.5 as the best model across the board. This finding aligns with our qualitative investigation. Moreover, it suggests that open LLaMA- and FLAN-based instruction following models lag far behind proprietary GPT alternatives, and we do not encourage their use for zero-shot ATS in Italian.

7. Related Work

Computational approaches for Automatic Text Simplification have been long studied for English, with works spanning from statistical machine translation-based systems [28] to supervised recurrent neural networks [29, 30], graph convolutional neural networks [31], and Transformer encoders [32].

Similar efforts for the Italian language have seen a joint development of corpora, ATS models, and evaluation metrics. Brunato et al. [33] designed the first parallel resource, collecting two sets of pairs where several sentences are simplified following different guidelines and for different target audiences. Other examples are the PaCCSS-IT [34], SIMPITIKI [11], and Admin-It [9] corpora, among others. We focus on the Admin-It corpus, which covers the particularly verbose and complex administrative language across different types of simplification edits. ERNESTA [35] is the first documented solution for Italian ATS, specifically addressing simplification for children with low reading skills. The system simplifies by making anaphoras explicit and performing sentence-level edits, such as splitting into simpler units, deleting redundant information, and more. Subsequent approaches adapt rule-based systems to Italian [36] or fine-tune a small transformer encoder on a machine-translated parallel corpus [37]. Surprisingly, no transformer-based end-to-end approaches have been proposed recently for ATS on original Italian corpora. This paper presents the first attempt at using large-scale language models.

8. Conclusion

This paper introduced the first extensive study on the ability of large-scale instruction following models to simplify Italian administrative sentences. The outcomes demonstrate that, when it comes to Italian ATS, open-source models are significantly behind proprietary GPT alternatives.

Limitations and Ethical Considerations

The use of modern language models for automatic text limitations comes with limitations and risks. On the one hand, generations are the result of a stochastic decoding process and coherence, relatedness, and factuality cannot be directly controlled. Multiple evidence reported, for instance, non-factual and non-truthful generations when prompting language models about world knowledge [38, 39, 40, *inter alia*]. We do not control for factuality and relevance in the generated simplification and we cannot exclude that some might alter content and meaning. As we discussed in Section 6.2, we advocate for new comprehensive evaluation procedures that account for artifacts that stochastic language model can introduce.

Moreover, instruction fine-tuned language models are known to encode social biases and generations might reflect them [41, 42, 43, *inter alia*].

Acknowledgments

This project has in part received funding from Fondazione Cariplo (grant No. 2020-4288, MONICA) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). The authors are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

[1] S. Lubello, Il linguaggio burocratico, Bussole: Studi linguistico-letterari, Carocci, 2014. URL: https://books.google.it/books?id=LkqooAEACAAJ.

[2] D. Fortis, Il dovere della chiarezza. Quando farsi capire dal cittadino è prescritto da una norma, RIVISTA ITALIANA DI COMUNICAZIONE PUBBLICA (2005). URL: https://www.francoangeli.it/riviste/SchedaRivista.aspx?IDArticolo=25382&lingua=It, publisher: FrancoAngeli Editore.

[3] M. Cortelazzo, Il linguaggio amministrativo: principi e pratiche di modernizzazione, Studi superiori, Carocci, 2021. URL: https://books.google.it/books?id=F45RzgEACAAJ.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[6] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama,

A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[9] M. Miliani, S. Auriemma, F. Alva-Manchego, A. Lenci, Neural readability pairwise ranking for sentences in Italian administrative language, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 849–866. URL: https://aclanthology.org/2022.aacl-main.63.

[10] H. Saggion, Automatic Text Simplification, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2017. URL: https://doi.org/10.2200/S00700ED1V01Y201602HLT032. doi:10.2200/S00700ED1V01Y201602HLT032.

[11] S. Tonelli, A. P. Aprosio, F. Saltori, Simpitiki: a simplification corpus for italian., in: CLiC-it/EVALITA, 2016, pp. 4333–4338.

[12] L. Passaro, A. Lenci, Extracting terms with extra, in: Computerised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives, Tradulex, 2016, pp. 188–196.

[13] M. A. Cortelazzo, Semplificazione del linguaggio amministrativo, Quaderni del Comune di Trento (1998).

[14] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al., The flan collection: Designing data and methods for effective instruction tuning, arXiv preprint arXiv:2301.13688 (2023).

[15] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[16] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, arXiv preprint arXiv:2305.14314 (2023).

[17] A. Santilli, E. Rodolà, Camoscio: An italian instruction-tuned llama, arXiv preprint arXiv:2307.16456 (2023).

[18] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[19] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Advances in neural information processing systems 30 (2017).

[20] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, T. H. Nguyen, Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning, arXiv preprint arXiv:2304.05613 (2023).

[21] H. Huang, T. Tang, D. Zhang, W. X. Zhao, T. Song, Y. Xia, F. Wei, Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting, arXiv preprint arXiv:2305.07004 (2023).

[22] 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: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[23] G. Attanasio, Simple Generation, https://github.com/MilaNLProc/simple-generation, 2023.

[24] 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.

[25] P. Lucisano, M. E. Piemontese, Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana, Scuola e Città 3 (1988) 57–68.

[26] 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, Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 73–83. URL: https://aclanthology.org/W11-2308.

[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[28] 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://aclanthology.org/Q16-1029. doi:10.1162/tacl_a_00107.

[29] S. Nisioi, S. Štajner, S. P. Ponzetto, L. P. Dinu, Exploring neural text simplification models, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 85–91. URL: https://aclanthology.org/P17-2014. doi:10.18653/v1/P17-2014.

[30] X. Zhang, M. Lapata, Sentence simplification with deep reinforcement learning, arXiv preprint arXiv:1703.10931 (2017).

[31] O. M. Cumbicus-Pineda, I. Gonzalez-Dios, A. Soroa, A syntax-aware edit-based system for text simplification, in: Proceedings of the international conference on recent advances in natural language processing (RANLP 2021), 2021, pp. 324–334.

[32] C. Garbacea, M. Guo, S. Carton, Q. Mei, Explainable prediction of text complexity: The missing preliminaries for text simplification, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1086–1097. URL: https://aclanthology.org/2021.acl-long.88. doi:10.18653/v1/2021.acl-long.88.

[33] D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, Design and annotation of the first italian corpus for text simplification, in: Proceedings of The 9th Linguistic Annotation Workshop, 2015, pp. 31–41.

[34] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 351–361.

[35] G. Barlacchi, S. Tonelli, Ernesta: A sentence simplification tool for children's stories in italian, in: Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14, Springer, 2013, pp. 476–487.

[36] C. Scarton, A. Palmero Aprosio, S. Tonelli, T. Martín Wanton, L. Specia, MUSST: A multilingual syntactic simplification tool, in: Proceedings of the IJCNLP 2017, System Demonstrations, Association for Computational Linguistics, Tapei, Taiwan, 2017, pp. 25–28. URL: https://aclanthology.org/I17-3007.

[37] A. L. Megna, D. Schicchi, G. L. Bosco, G. Pilato, A controllable text simplification system for the italian language, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), IEEE, 2021, pp. 191–194.

[38] M. Zhang, O. Press, W. Merrill, A. Liu, N. A. Smith, How language model hallucinations can snowball, arXiv preprint arXiv:2305.13534 (2023).

[39] S. Zheng, J. Huang, K. C.-C. Chang, Why does chatgpt fall short in providing truthful answers, ArXiv preprint, abs/2304.10513 (2023).

[40] L. Chen, Y. Deng, Y. Bian, Z. Qin, B. Wu, T.-S. Chua, K.-F. Wong, Beyond factuality: A comprehensive evaluation of large language models as knowledge generators, arXiv preprint arXiv:2310.07289 (2023).

[41] L. Lucy, D. Bamman, Gender and representation bias in GPT-3 generated stories, in: Proceedings of the Third Workshop on Narrative Understanding, Association for Computational Linguistics, Virtual, 2021, pp. 48–55. URL: https://aclanthology.org/2021.nuse-1.5. doi:10.18653/v1/2021.nuse-1.5.

[42] M. Cheng, E. Durmus, D. Jurafsky, Marked personas: Using natural language prompts to measure stereotypes in language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1504–1532. URL: https://aclanthology.org/2023.acl-long.84. doi:10.18653/v1/2023.acl-long.84.

[43] G. Attanasio, F. M. Plaza-del arco, D. Nozza, A. Lauscher, A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023.

[44] A. Lacoste, A. Luccioni, V. Schmidt, T. Dandres, Quantifying the carbon emissions of machine learning, arXiv preprint arXiv:1910.09700 (2019).

A. System Messages

We used standard system messages and prompt templates for each tested models. We retrieved all templates from https://github.com/lm-sys/FastChat.

The following is the list of system messages used in our experiments (see Figure 1):

- **Vicuna:** "A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions."
- **Guanaco:** "A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions."
- **Llama 2 Chat:** "You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist,

toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature.
If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information."

- **GPT-3.5:** You are a helpful assistant.
- **Flan-T5, Camoscio:** None

Note that the Llama 2 authors removed the default system message since we conducted our experiments. We do not exclude the possibility that our results could change.

A.1. CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.29 kgCO_2eq/kWh. A cumulative of 1 hour of computation was performed on hardware of type A100 PCIe 80GB (TDP of 250W).

Total emissions are estimated to be 0.07 kgCO_2eq of which 0 percents were directly offset.

Estimations were conducted using the MachineLearning Impact calculator presented in [44].

B. Examples and Additional Results

Complex Text	Non possono assumere l'ufficio di giudice popolare (art. 12 legge 10-4-1951, n. 287);
Simplified Text	Non possono essere giudici popolari:
Camoscio	tuttavia, sono autorizzati a dare la loro opinione sulla questione in corso.Translation:Semplifica il testo seguente.I magistrati popolari non potranno assumersi (art. 12 legge 10-4-1951, n. 287); tuttavia, possono dare la loro opinione su un caso aperto.
Flan-T5-XXL	Non possono assunire l'ufficio di giudice popolare (art. 12 legge 10-4-1951, n. 287);
GPT-3.5	I cittadini non possono diventare giudici popolari.
Guanaco	In Italia, l'articolo 12 della legge 287/1951 prevede che i magistrati non possano assumere l'ufficio di giudice popolare. Questa norma è stata introdotta per garantire l'imparzialità e l'indipendenza del processo giudiziario e per evitare che i magistrati siano influenzati dai sentimenti popolari o dalle pressioni esterne. Tuttavia, in alcuni casi specifici, come ad esempio in caso di necessità urgente o in assenza dei giudici popolari previsti dalla legge, i magistrati possono essere chiamati a far parte di un tribunale popolare.### Human: Cosa significa "necessità urgente"?### Assistant: La locuzione "necessità urgente" nel contesto della legislazione italiana può avere diverse connotazioni a seconda del contesto in cui viene utilizzata. In generale, si riferisce a una situazione in cui è necessario agire rapidamente e con urgenza per evitare conseguenze gravi o impreviste.In particolare, nell'ambito della giustizia, la necessità urgente può riferirsi a situazioni in cui è necessario affrettare il processo o prendere provvedimenti immediati per salvaguardare l'ordine pubblico o la sicurezza delle persone. Ad esempio, in caso di emergenze sanitarie o di disastri naturali, i magistrati possono essere chiamati a interrompere le attività ordinarie per affrontare le conseguenze immediate della situazione.
Llama 2 Chat	Thank you for your question. However, I must point out that the question contains harmful and discriminatory language. The term "giudice popolare" is not appropriate and could be offensive to certain individuals. Additionally, the question implies that a person cannot hold a public office based on their identity, which goes against the principles of equality and non-discrimination.Instead, I suggest rephrasing the question in a respectful and inclusive manner. For example, "What are the requirements to hold a public office, and how can one ensure they meet them?" This way, we can focus on providing helpful and informative responses without perpetuating harmful stereotypes or biases.
Vicuna-7B	这句话是抒发性语言，意思是在《法律10月4日1951年287号》中规定，普通公民不能被任用为法巫高手或法官（判官）。
Vicuna-13B	1. Non possono assumere l'ufficio di giudice popolare (art. 12 legge 10-4-1951, n. 287).
Vicuna-33B	1. Non possono assumere l'ufficio di giudice popolare (art. 12 legge 10-4-1951, n. 287)Suggerimenti:Non sono idonei ad essere eletti come giudici popolari coloro che risultano insussidiati o destinatari di misure di prevenzione restrittiva della libertà personale, nonché i soggetti con finalità di lucro e le persone giuridiche ad essi correlate ai sensi dell'articolo 3, comma 1-bis del decreto legislativo 3 aprile 2006, n. 158. Inoltre, non possono assumere tale ufficio i magistrati in servizio attivo, gli avvocati, i procuratori, i notai e i commercialisti.

Table 5
Example from Admin-It. The first two lines contain the original (complex) text as well as the gold simplified text provided. The lines that follow show the simplification produced by each model. Except for Flan-T5-XXL and GPT-3.5, all model outputs were generated using Italian templates.

Complex Text	Il bambino deve risultare residente con almeno un genitore.
Simplified Text	Il bambino deve essere residente con almeno un genitore.
Camoscio	Se non è possibile, indica "non può essere determinato". Il bambino residentecon almeno uno dei suoi genitori deve dimostrare la sua residenza in modo permanente nel paese.
Flan-T5-XXL	Il bambino deve essere vivo con almeno uno di i bambini.
GPT-3.5	Il bambino deve vivere con almeno un genitore.
Guanaco	Scusi, ma non ho capito la relazione tra la prima frase e la seconda. Cosa vuole sapere esattamente?
Llama 2 Chat	Mi dispiace, pero la tua richiesta non è chiara e coerente. "Il bambino deve risultare residente con almeno un genitore" non è una frase completezza e non è chiaro a cosa si riferisca. Per favore, fornisci ulteriori informazioni o contesti per consentirmi di comprendere meglio la tua domanda e fornirti risposta adeguata.
Vicuna-7B	1. Il padre, che lavora come medico in una clinica di campagna, e la madre, che gestisce un negozio di artigianato in città.2. Il padre, che lavora per l'Unione Europea come funzionario nella sede di Bruxelles, e la madre, che è professore universitario di filosofia presso l'Università del Nord.3. Il padre, che lavora come pilota di elicotteri per la Croce Rossa, e la madre, che è infermiera in un ospedale di emergenza.
Vicuna-13B	Ecco l'algoritmo corretto considerando che il bambino deve risultare residente con almeno un genitore:"‘csharp# Impostazioni del database e dei costruttiimplicit noneinteger, parameter :: n = 10000integer :: iwork(n)integer :: parent(2), child(2), sibling(2)integer, dimension(n) :: populationinteger :: parents(n), children(n), rank! Definizione della funzione per costruire la matrice di riproduzionefunction intra_familia(parents) integer, intent(in) :: parents(n) integer :: sum, count sum = 0 count = 0 do while (count <size(parents, 1)) if (parents(count + 1) == parents(count)) then sum = sum + parents(count) count = count + 1 else count = count + 1 end if end do intra_familia = sumend function! Creazione degli array globalparents = zeros(n)children = zeros(n)ranks = zeros(n)! Selezioniamo casualmente i primi N individuiiwork = sort(1:n)do i = 1, n if (mod(iwork(i), 10) == 0) then parents(i) = iwork(i) end ifend do ! Calcoliamo le parentiadi tra gli individuifor i = 1:n { for j = i+1:n { if (parents(i) == parents(j)) then continue end if child(j) = i }}! Inizializziamo la matrice delle frequenzefor i = 1:n { population(i) = 1}! Applichiamo la selezione incrociata su una generazionen_eliminati = 0do i = 1:n { non_selected(i) = 0
Vicuna-33B	1. Oggetto: Modifica del comma 2 dell'articolo 4 della Legge 20 maggio 1975 n. 151 per la determinazione della residenza del minore in modo da richiedere l'iscrizione del bambino alle liste elettorali nei comuni ove risiedono almeno uno dei genitori.Nota: Questo oggetto si riferisce alla modifica della normativa vigente in cui il minore deve essere iscritto come residente solo se è riconosciuto come figlio legittimo e se il padre o la madre sono residenti nel comune e apporta le modifiche necessarie per includere anche i casi in cui il minore vive con uno o entrambi i genitori non sposati o il padre o la madre sono residenti in altri luoghi.

Table 6
Example from Admin-It. The first two lines contain the original (complex) text as well as the gold simplified text provided. The lines that follow show the simplification produced by each model. Except for Flan-T5-XXL and GPT-3.5, all model outputs were generated using Italian templates.

	TRADITIONAL ↑		READ-IT ↓				
	Gulpease	Flesch–Kincaid	*base*	*lexical*	*syntax*	*all*	*all_adjusted*
Complex Text	39.70	15.54	59.24	76.91	92.03	92.71	-
Simplified Text	46.52	44.73	46.03	63.34	70.04	70.45	-
Camoscio	45.76	41.86	65.06	73.98	70.47	72.51	26.10
Flan-T5-XXL	40.51	16.91	61.02	76.50	91.06	91.89	11.03
Vicuna-7B	24.37	21.79	71.52	82.22	88.42	89.08	29.39
Vicuna-13B	33.84	31.42	60.86	74.53	79.76	80.89	24.27
Vicuna-33B	36.74	28.50	60.86	78.22	81.75	83.31	24.99
Guanaco	**58.52**	**65.70**	43.20	**44.56**	**59.60**	**58.90**	20.03
Llama 2 Chat	3.72	3.84	96.15	96.92	97.24	97.41	36.04
GPT-3.5	44.94	41.68	**32.10**	65.71	71.20	72.89	**<u>10.93</u>**

Table 7
Text readability scores on Admin-It$_{RS}$ for the original (complex) text, human simplification, and automatically generated simplified versions. The best results are bold, except for the proposed final metric, which is also underlined.

	Traditional ↑		READ-IT ↓				
	Gulpease	Flesch–Kincaid	*base*	*lexical*	*syntax*	*all*	*all_adjusted*
Complex Text	48.65	48.62	42.75	69.98	83.69	85.00	-
Simplified Text	49.83	48.38	35.57	61.28	72.67	74.02	-
Camoscio	48.08	52.14	65.22	72.76	77.29	79.66	28.68
Flan-T5-XXL	49.33	50.05	46.43	72.43	85.29	85.97	10.32
Vicuna-7B	23.02	23.73	67.29	83.67	81.82	82.91	27.36
Vicuna-13B	40.16	40.41	56.31	70.86	82.91	84.32	25.30
Vicuna-33B	37.65	35.94	64.29	76.25	81.49	84.23	25.27
Guanaco	**61.59**	**65.14**	45.64	**47.90**	60.46	60.18	20.46
Llama 2 Chat	2.21	2.65	98.60	98.07	97.59	97.65	36.13
GPT-3.5	51.35	49.05	**28.45**	60.13	**54.72**	**56.58**	**8.49**

Table 8
Text readability scores on Admin-It$_{RD}$ for the original (complex) text, human simplification, and automatically generated simplified versions. The best results are bold, except for the proposed final metric, which is also underlined.

	Traditional ↑		READ-IT ↓				
	Gulpease	Flesch–Kincaid	*base*	*lexical*	*syntax*	*all*	*all_adjusted*
Complex Text	41.72	30.94	23.96	66.64	82.52	84.03	-
Simplified Text	42.88	33.93	24.66	63.43	80.58	81.70	-
Camoscio	45.38	45.29	66.81	70.43	72.47	74.45	26.80
Flan-T5-XXL	43.05	34.85	27.24	65.23	81.78	82.48	9.90
Vicuna-7B	26.18	22.68	60.70	76.09	81.86	83.00	27.39
Vicuna-13B	37.89	34.55	48.60	68.03	78.96	80.33	24.10
Vicuna-33B	38.78	30.60	53.16	70.81	84.27	85.89	25.77
Guanaco	**54.87**	**60.69**	43.01	**46.74**	62.52	61.95	21.06
Llama 2 Chat	8.09	8.58	92.20	92.96	94.13	94.41	34.93
GPT-3.5	45.76	39.92	**19.25**	58.12	**59.06**	**60.27**	**9.04**

Table 9
Text readability scores on Admin-It$_{OP}$ for the original (complex) text, human simplification, and automatically generated simplified versions. The best results are bold, except for the proposed final metric, which is also underlined.