# Contemporary LLMs and Literary Abridgement: An Analytical Inquiry

Iglika Nikolova-Stoupak Gaël Lejeune Eva Schaeffer-Lacroix
Sens Texte Informatique Histoire, Sorbonne Université, Paris, France
iglika.nikolova-stoupak@etu.sorbonne-universite.fr,
{gael.lejeune, eva.lacroix}@sorbonne-universite.fr

## Abstract

Within the framework of this study, several contemporary Large Language Models (ChatGPT, Gemini Pro, Mistral-Instruct and BgGPT) are evaluated in relation to their ability to generate abridged versions of literary texts. The analysis is based on 'The Ugly Duckling' by H. C. Andersen as translated into English, French and Bulgarian. The different scenarios of abridgement experimented with include zero-shot, one-shot, division into chunks and crosslingual (including chain-of-thought) abridgement. The resulting texts are evaluated both automatically and via human evaluation. The automatic analysis includes ROUGE and BERTScore as well as the ratios of a selection of readability-related textual features (e.g. number of words, type-to-token ratio) as pertaining to the original versus automatically abridged texts. Professionally composed abridged versions are regarded as gold standard. Following the automatic analysis, six selected best candidate texts per language are then evaluated by volunteers with university education in terms of textual characteristics of a more qualitative nature, such as coherence, consistency and aesthetic appeal.

**Keywords:** LLMs, literary abridgement, multilingual text generation

## 1 Introduction

The current work seeks to provide an overview of the ability of contemporary Large Language Models (LLMs) to generate abridged versions of literary works. As per the Merriam-Webster dictionary, 'abridged' means 'shortened or condensed, especially by the omission of words or passages'. Notably, abridgement makes literary texts accessible to audiences that would find it hard to read or work with the corresponding original texts, such as young children, foreign language learners or people with learning disabilities. The term will not be used as synonymous to 'summary' due to both its particular relevance to the literary domain and its focus on overall simplification rather than merely reduction in size.

## 2 Background

Although literary abridgement by LLMs is not yet an established research topic, it implies several sets of abilities pertaining to the technology that are currently of marked academic interest, notably the use of long context, summarisation, and creative/literary writing.

### 2.1 LLMs and Long Context

A major limitation of contemporary LLMs is their imperfect ability to receive and make sense of large amounts of text. Through the tasks of multi-document question answering and key-value retrieval, Liu et al. (2023) evaluate LLM's general ability to use long contexts, revealing drawbacks such as position bias, i.e. the tendency of models to work better with information situated toward the beginning or end of a document (a.k.a the 'lost-in-the-middle' problem). They note that even extended-context models, such as LongChat-13B, are not characterised with better use of long context. In contrast, instruction fine-tuned models use contexts more efficiently. Different techniques to extend models' context window have been proposed, such as position interpolation, a type of minimal fine-tuning, in which position indices provided to transformer models are scaled down to accommodate the additional context (Chen et al., 2023). In relation to the task of machine translation (MT), Du et al. (2023) note that its quality diminishes significantly as document size increases, GPT-4 receiving the highest BLUE scores among contemporary LLMs when long context is involved.

### 2.2 LLMs and Summarisation

One of the Natural Language Processing (NLP) tasks that most directly benefit from the use of extensive context is document summarisation (in particular, abstractive summarisation as it pertains to a text's transformation rather than its mere reduction in size). Chang et al. (2024) divide a long document (over 100k tokens) into chunks and then merge them to derive full summaries. They experiment with merging the chunks hierarchically and incrementally and use textual coherence to evaluate the resulting summaries, thereby proposing an automatic metric of coherence. They attribute highest scores to GPT-4 and Claude 2 and to the practice of hierarchical merging. Wu et al. (2024) work around the aforementioned 'lost-in-the-middle' problem through an 'extract-then-evaluate' approach, in which they incrementally extract and concatenate key sentences from a document that result in the highest ROUGE score of the achieved summary.

Most state-of-the-art practices related to the task of summarisation pertain to the reduction in size of long, mostly news-based texts for the purpose of time efficiency whilst key information is preserved. Xiao and Chen (2023) focus on informativeness when applying evolutionary fine-tuning to news summarisation. Zhang et al. (2023) test ten LLMs' summarisation performance and compare it to that of humans, concluding that instruction tuning provides a significant benefit and that human summaries tend to be more abstractive in nature (i.e. use paraphrasing rather than direct extraction) than LLM-generated counterparts. Pu et al. (2023)'s bold statement that (human-based) "summarization is (almost) dead" is based on the results of five discrete summarisation tasks, including crosslingual summarisation (CLS). Pairwise human evaluation rates LLMs as markedly better at the tasks than both humans and fine-tuned neural models, and particularly strong in terms of fluency and coherence. In contrast, LLM's current performance in CLS is also tested by Wang et al. (2023a) and not left uncriticised. Based on CLS datasets and the ROUGE and BERTScore metrics, contemporary models such as GPT-4 and ChatGPT are evaluated as reaching competitive but still worse zero-shot performance compared to a BART model that has been fine-tuned for the task. Open-source models such as Vicuna-13B are judged to outright lack zero-shot CLS ability. Additional experiments reveal that a chain-on-thought method of asking the model to first translate and then summarise (and vice-versa) a text helps improve performance.

### 2.3 LLMs and Creative Writing/Literature

The involvement of LLMs in creative writing as much as gives rise to philosophical questions about the nature of creativity. Franceschelli and Musolesi (2023) apply Margaret Boden's theories of value, novelty and surprise to the function of LLMs, concluding that their creativity is by definition limited in nature and scope. Both due to ethical reasons and to a general opinion that LLMs' current abilities are still lacking, their role in creative writing is often limited to subtasks such as plot outlines or character development. Kreminski and Martens (2022) systematise the potential of current LLMs to provide support for writers, providing guidelines for their effective use in the overcoming of 'writer's block'. User-friendly tools like *Story Centaur* (Swanson et al., 2021), which is based on LLMs' few-shot abilities, have been developed to aid creative writers in their work by fulfilling narrowly framed tasks, such as the provision of a next sentence given the previous one and a 'magic word' to be incorporated. Also viewing LLMs as potential assistants in the creative writing process, Shanahan and Clarke (2023) use elaborate prompting strategies combined with fine-tuning of the temperature setting to collect textual samples from GPT-4 that they then evaluate qualitatively, basing themselves on an array of literary concepts including characterisation, imagery and use of idioms. They discuss the creativity of LLMs as analysable and multi-faceted albeit tightly dependent on the quality of underlying prompts. Other comprehensive studies on the topic include Gómez-Rodríguez and Williams (2023)'s evaluation of the creative writing abilities of a number of contemporary LLMs. The authors provide LLM models and several human writers with an identical creative writing task, and they apply to the issuing stories human evaluation based on established criteria in the domain, including coherence and the use of humour. They conclude that commercial LLMs perform comparably to human writers but do not match the latter in originality, and that the understanding of humour can be considered an emerging ability of LLMs.

Prior to the advancement to LLMs, translation of literary texts was seen as "the greatest challenge for MT" (Toral and Way, 2018) as it implies the reader's overall experience as opposed to a limited number of automatisable measures. Recently, Tencent AI Lab and China Literature Ltd. organised a shared task on discourse-level literary translation, thereby releasing a Chinese-English web novel corpus. Among the tested baseline systems, LLMs performed best by a significant margin based on both automatic and human evaluation (Wang et al., 2023b).

## 3 Methods

### 3.1 Texts

The utilised source texts are published translations of 'The Little Duckling' (H. C. Andersen) into English[1], French[2], and Bulgarian[3]. An original work written in a language that is not discussed (Danish) is deliberately opted for in order to avoid the presence of both original and translated texts in the following experiments. In addition, up to four published abridged versions per language are used in the context of automatic experiments in order to define reference ratios of textual features between an original and abridged version. For the one-shot scenario, an original and abridged version of 'The Little Match Girl' (H. C. Andersen) in each language are utilised[4]. In order for the relative impact of possible recognition of the text by LLMs to be tested, an alternative, non-published story, 'The Gift under the Bush' [5] is also used both in its original Bulgarian version and the author's own translations into the additional languages. Some models' context size restrictions did not allow for particular scenarios (typically, zero-shot) to be fulfilled on the respective full text. In this case, abridged versions were used as source texts (see Appendix C for details about the derivation of specific abridged versions by LLMs).

### 3.2 Models

The models experimented with are Mistral-Instruct, BgGPT, Gemini Pro and ChatGPT (as based on GPT-3.5). Mistral-Instruct (7B) is an open-source model, developed by Mistral AI as a fine-tuned version of the original Mistral model, whose main characteristics include high inference speed and a sliding window attention mechanism (Jiang et al., 2023). Its context window comes at 32k tokens. BgGPT-7B-Instruct by INSAIT is based on Mistral-7B and fine-tuned with large amounts of textual data for the purpose of better understanding and production of Bulgarian text (INSAIT, 2024). Gemini Pro (600B) is a user-friendly version of the state-of-the-art Gemini model by Google DeepMind, which is documented to outperform GPT-4 in 30 out of 32 language benchmarks (Anil et al., 2023). It has a context window of 128k tokens. OpenAI's GPT-3.5 is the model behind the free and most commonly used version of ChatGPT in the moment of writing of this article. For the purpose of this project, Mistral-Instruct was deployed through the LM Studio interface[6], Gemini through the Google AI Studio tool within the established free quota, and BgGPT and ChatGPT through their offi-

[1] Andersen, Hans Christian. The Ugly Duckling. 1843. https://pinkmonkey.com/dl/library1/tale120.pdf.

[2] Andersen, Hans Christian. Le vilain petit canard. 1843. https://touslescontes.com/biblio/conte.php?iDconte=158.

[3] Andersen, Hans Christian. Groznoto patentse. Translated by Svetoslav Minkov, Chitanka, 1977. https://chitanka.info/text/4819.

[4] English full: Andersen, Hans Christian. *The Little Match Girl.* Short Story America, 1845. https://shortstoryamerica.com/pdf_classics/andersen_little_match_girl.pdf.
English abridged: Andersen, Hans Christian. *The Little Match Girl.* https://fliphtml5.com/mcbeq/hrvp/basic.
French full: Andersen, Hans Christian. *La petite fille aux allumettes.* https://touslescontes.com/biblio/conte.php?iDconte=127.
French abridged: Andersen, Hans Christian. *La petite fille aux allumettes.* https://miladlh.com/wp-content/uploads/2020/11/La-Petite-Fille-aux-Allumettes.pdf.
Bulgarian full: Andersen, Hans Christian. *Malkata kibritoprodavachka.* https://chitanka.info/text/4826-malkata-kibritoprodavachka.
Bulgarian abridged: Andersen, Hans Christian. 'Malkata kibritoprodavachka.' In *Prikazki ot tsyal svyat*, transl. Vasil Velchev, 2009.

[5] Stoupak, Stefan. The Gift under the Bush. Unpublished manuscript, 2024.

[6] https://lmstudio.ai/

cial chatbot interfaces.

### 3.3 Abridgement Scenarios

This study seeks to test and compare the current ingrained capabilities of LLMs to generate abridged versions of literary texts. For the purpose, no extensive fine-tuning and prompt-engineering methods are applied. In addition, no definition of 'abridgement' is provided within prompts. The following discrete experimental settings are considered: zero-shot, chunking, one-shot and crosslingual. In the one-shot setting, an original and an abridged version of another text ('The Little Match Girl') are provided to the model as an example of the transformation it is expected to apply. In the chunking scenario, the original text is divided into several (typically, three) parts. Crosslingual experiments are conducted both in a zero-shot setting and via a simple chain-of-thought that asks the model to first translate and then provide an abridged version of the text (henceforth, 'chain-of-thought 1') and vice-versa (henceforth, 'chain-of-thought 2')[7]. Due to the possibility of the models having encountered 'The Little Duckling' during training, additional experiments are carried out using a text that has not been published before; which, however, is not coupled with a gold standard abridged version.

Experiments are carried out in English, French and Bulgarian. In the case of BgGPT, naturally only Bulgarian is used. The majority of Mistral experiments are discarded due to poor output quality[8].

### 3.4 Evaluation

#### 3.4.1 Automatic Evaluation

A selection of ten automatic measures is applied to the generated texts. For each language, the range of ratios between a full text and its human-made abridged versions is taken as gold standard that the abridged versions are compared against. For instance, if the ratios between the number of words in the original English text and the four human-made abridged English texts are between 2.0 and 10.0, the 'number of words' measure is marked positively for LLM-generated texts, for which it falls within this same range.

The ratios between the full and abridged versions of the text used in the one-shot setting are also included in the range. The same range of ratios is applied to the alternative text for the given language, as there is no professional abridged version of it. In the cases where abridged texts are derived from other abridged texts due to the models' context length restrictions, it is the ratios between the utilised human-made abridged text and the LLM-generated futher-abridged text that are taken into account.

As both ROUGE and BERTScore inherently compare two texts, it is directly the scores that compare abridged to original texts that are calculated. ROUGE is a standard measure for automatically-generated textual summaries that typically considers the overlap between a newly generated and a gold standard summary (Lin, 2004). For the purpose of this work, ROUGE-1 recall is used to calculate the portion of individual words in an abridged version that are present in the associated original text.

BERTscore, often used as an improved alternative to ROUGE, compares two texts based on the cosine similarity of token embeddings, thus capturing closeness of meaning (Zhang et al., 2020). F1 values of the BERTScore comparing original and abridged texts are calculated, thus providing a balanced measure of the inclusion of relevant information in an abridged text and its conciseness.

Readability is a notion that refers to the general complexity of a given text and, by extension, to its potential modification or simplification, especially in view of a particular reader profile (traditionally, defined by grade level). Most established readability formulas make use of shallow characteristics that have proven to be good proxies of complexity, such as the average number of syllables per word or the average number of words per sentence, used within the Flesch Reading Ease Formula (DuBay, 2007). Recent studies, such as Feng et al. (2010) have sought to further systematise the atomic features used in readability measurement as well as to determine their in-

[7] For the full prompts used, please consult the following repository: https://github.com/iglika88/Contemporary-LLMs-and-Literary-Abridgement/

[8] including 'one-shot' and all experiments involving non-English languages, with the sole exception of crosslingual abridgement from Bulgarian to English

terconnectedness. For the purpose of this study, a set of readability-related features is used that aims at informativeness as well as balance between different textual aspects: length (total number of words, number of words per sentence, number of letters per word), vocabulary (type-to-token ratio, concreteness as per Brysbaert et al. (2014)[9], words outside of a determined frequency list), syntax (ratio of content to function words) and discourse (presence of anaphora-denoting words).

### 3.4.2 Human Evaluation

Six LLM-generated texts per language are selected for the human evaluation survey. They are the texts rated most highly by the automatic evaluation process i.e. the ones with the highest number of characteristics that fall within the gold standard range. In cases of equal scores, a variety between models and generation scenarios is sought. Four versions of the survey per language were composed, each of them consisting of two texts to evaluate. One of the two texts was also present in another version, in order to allow for a calculation of agreement[10]. A minimum of one participant per version and per language (native or fluent speaker with a university background) completed the survey.

The general categories evaluated in the survey are: understandability, correctness, consistency, textual coherence and aesthetic appeal. The respondents were offered a scale of 4 ('no', 'mostly no', 'mostly yes' and 'yes') and also encouraged to leave comments in the form of free text. The protocol's overall form is adapted from Mousavi et al. (2022)[11].

Cohen's Kappa coefficient was calculated for the texts that are present within two versions of the survey. The response values were taken as categorical. In cases of more than two participants, Fleiss' Kappa (Fleiss et al., 1969) was also calculated.

## 4 Results

### 4.1 Automatic Evaluation

A model's performance is defined as the percentage of examined textual characteristics that fall within the range defined by the human-made abridged texts, as elaborated in Section 3.4.1. For instance, the study contains 24 texts generated by ChatGPT. In total, they are evaluated in terms of 226 characteristics, out of which 93 fall within the defined range, thus giving ChatGPT a score of 41%.

Observable tendencies related to the discussed atomic textual characteristics include too short length in relation to crosslingual and zero-shot generation scenarios (an exception being the Bulgarian language, for which zero-shot generation renders excessively long text). Also, the process of crosslingual generation results in a high percentage of words not appearing in the respective language's frequency list. Some characteristics, particularly ROUGE, BERTScore and type-to-token ratio, score particularly weakly in relation to the French language.

Results are further summarised in the following subsections. For the detailed results of the automatic evaluation, please refer to Appendix C.

#### 4.1.1 Performance by language

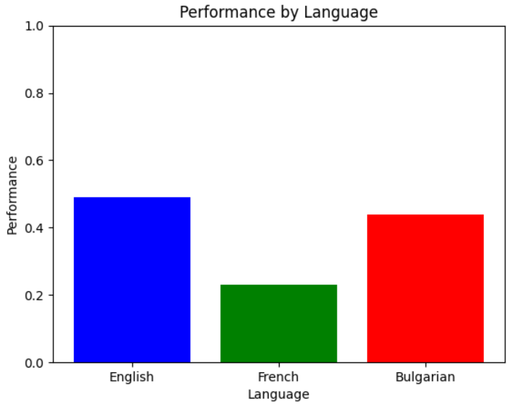

Figure 1: Performance by language

As shown in Figure 1, the highest performance is understandably attributable to English, somewhat surprisingly followed by the lower-resource language, Bulgarian.

[9] applicable only to English text

[10] For a breakdown of the texts, please refer to Appendix B. For the the full texts included in the survey, please consult https://github.com/iglika88/Contemporary-LLMs-and-Literary-Abridgement/

[11] For the entire protocol, please refer to https://github.com/iglika88/Contemporary-LLMs-and-Literary-Abridgement/

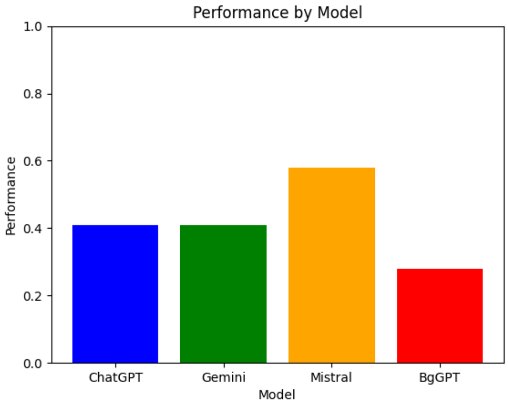

Figure 2: Performance by model

### 4.1.2 Performance by model

The model that scores highest is Mistral. However, it is to be kept it mind that only a limited number of experiments were carried out using this model, and that they were all in the strongest performing language, English. ChatGPT and Gemini demonstrate equal global performance, and BgGPT comes last (see Figure 2).

### 4.1.3 Performance by generation scenario

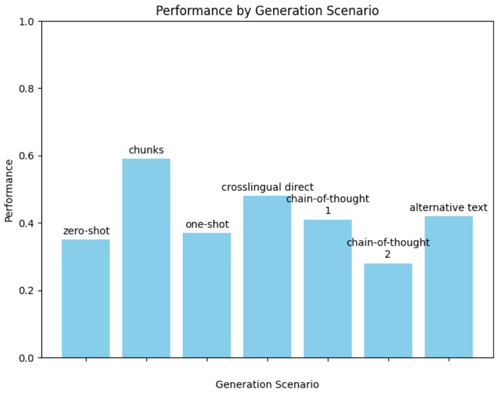

Figure 3: Performance by scenario

Top performance is exhibited by the 'chunks' and 'crosslingual: direct' abridgement scenarios (see Figure 3). In contrast, 'crosslingual: chain-of-thought 2' abridgement scores lowest. Zero-shot performance is in fact higher for the alternative text, showing that there is no significant influence of the text being present in training data on the models' performance.

### 4.1.4 Performance by model and language

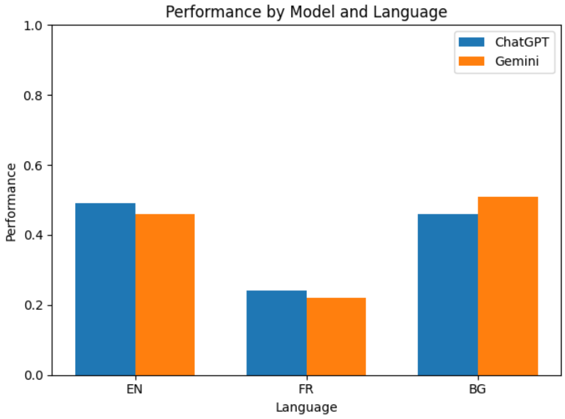

Figure 4: Performance by model and language

ChatGPT is observed to outperform Gemini in English and French, but not in Bulgarian (see Figure 4). The Mistral and bgGPT models are naturally excluded from this evaluation, as each of them addresses only a single language.

### 4.1.5 Performance by scenario and language

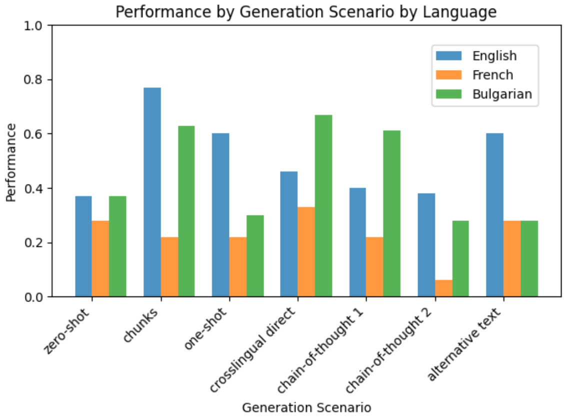

Figure 5: Performance by scenario and language

The one-shot setting in relation to both the primary and alternative text is high for the English language (see Figure 5). Crosslingual scenarios work best for Bulgarian, likely speaking of a benefit arising from use of the originally input English text.

### 4.1.6 Performance by scenario and model

As seen in Figure 6, ChatGPT outperforms Gemini in relation to the 'chunks', 'one-shot'

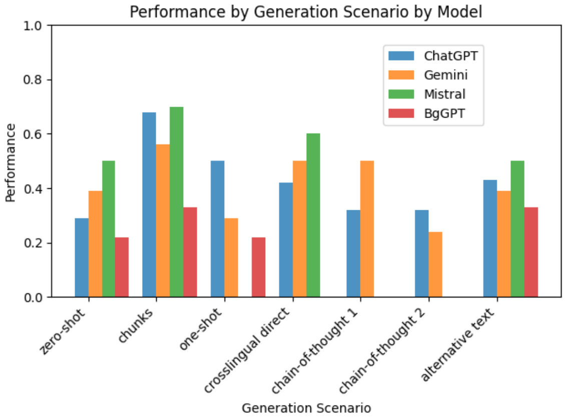

Figure 6: Performance by scenario and model

and 'crosslingual chain-of-thought 2' scenarios as well as with the alternative text. BgGPT's performance is the most uniform one between generation scenarios.

## 4.2 Human Evaluation

### 4.2.1 English Texts

Three participants responded to version 2, and one to each other version of the English survey. The first repeated text received low agreement per both Fleiss' Kappa and Cohen's Kappa, with the exception of the participant who responded to version one and the second participant who responded to version 2, who were in fair agreement. The second repeated text also received low Cohen's Kappa agreement.

Two texts received positive answers ('Yes' or 'Mostly Yes') for all categories of the survey: 'ChatGPT: crosslingual chain-of-thought 2' and 'ChatGPT: crosslingual direct'. 'Gemini: alternative' followed with 94.44%, 'Gemini: one-shot' with 81.48%, 'Mistral: chunks' with 55.56% and 'ChatGPT: chunks' with 50%.

Problems with understandability were noted in relation to 'Mistral: chunks' and 'ChatGPT: chunks'. The Mistral text was rated negatively for all aspects of correctness, whilst the 'ChatGPT: chunks' and 'Gemini: one-shot' ones were judged as having non-optimal structure. 'ChatGPT: crosslingual chain-of-thought 2' was seen as using awkward vocabulary (e.g. 'poultry yard') and unlikely parts of speech to render meaning. Inconsistencies in register and style were noted for the Mistral text and 'Gemini: one-shot'. In turn, 'ChatGPT: chunks' demonstrates inconsistency in the presented information (e.g. an action taking place twice) and use of pronouns (the duckling being referred to as 'it' and 'he' in different parts of the story). 'ChatGPT: chunks' also received a fully negative rating for coherence. Aesthetic characteristics (notably, textual length, pacing and engagement) were commonly marked negatively for all texts except ChatGPT's two crosslingual ones.

### 4.2.2 French Texts

Two participants responded to version 3 of the survey and one to each of the other three. The first repeated text is associated with low agreement as per Cohen's Kappa, and the second one, which was evaluated by three people, received low Fleiss' Kappa as well as low Cohen's Kappa with the exception of the second participant who responded to version 3 and the participant who responded to version 4 (fair agreement).

The most highly rated text was 'Gemini: crosslingual direct' (100% positive answers), followed by 'ChatGPT: zero-shot' (94.12%), 'Gemini: zero-shot' (90.74%), 'ChatGPT: chunks' (88.89%), 'ChatGPT: alternative' (82.86%) and 'Gemini: crosslingual chain-of-thought 2' (69.44%).

The texts' understandability was rated fully positively, with a mention of occasional complex vocabulary ('Gemini: zero-shot') and grammar ('ChatGPT: alternative'). Marked issues pertaining to correctness included the type of text not resembling an abridged story but rather a 'fable' ('Gemini: zero-shot') or just a 'short story' ('ChatGPT: chunks') and wrong use of tenses ('Gemini: zero-shot'). The structure of 'ChatGPT: one-shot' was the only one marked negatively, whilst it was explicitly noted that in 'Gemini: crosslingual direct', "all the [ugly duckling's] adventures are present". Consistency of style was marked negatively for 'ChatGPT: one-shot' and 'Gemini: crosslingual chain-of-thought 1'. Within the latter, grammar was perceived to be too simple as compared to vocabulary. For 'ChatGPT: alternative', information was also marked as inconsistent. Problems with transitions were noted in 'ChatGPT: alternative', 'Gemini: zero-shot' and 'Gemini: crosslingual chain-of-thought 1', and the last was also claimed to include problems with anaphora use. When it comes to aesthetic qualities, 'ChatGPT: alternative', 'Gemini: zero-shot', 'ChatGPT: chunks' and 'Gemini: crosslingual chain of thought 1' received negative scores for engagement, comments referring to the texts as 'non-fluid' and 'frustratingly' weakly developed. Problems of pacing and textual length (particularly, texts being too short) were also brought forward.

### 4.2.3 Bulgarian Texts

Four participants responded to version 1 of the survey and one person each for the other three versions. Cohen's Kappa for the repeated text in versions 1 and 2 is fair between participants 1 and 5 and low for the rest; Fleiss' Kappa is low. Cohen's Kappa for the other repeated text is fair.

The 'Gemini: chunks' text was rated most highly, with 90.28% positive answers, followed by 'Gemini: crosslingual train-of-thought 1' and 'ChatGPT: zero-shot' (88.89%), 'ChatGPT: one-shot' (69.45%), 'ChatGPT: chunks' (54.45%) and 'Gemini: crosslingual direct' (44.45%).

The texts' understandability was generally rated highly. In contrast, correctness received a high number of negative answers, particularly in relation to vocabulary, grammar, and structure. For instance, vocabulary in 'Gemini: chunks' was judged to often be wrong, wrongly used or seemingly translated, the verbs in 'ChatGPT: chunks' were said to often be wrongly interpreted in terms of transitiveness, and 'Gemini: chunks' felt as if it were 'mixed with other stories'. Consistency was marked negatively for the 'ChatGPT: chunks' and 'ChatGPT: one-shot' texts. Underlined problems of coherence included excessive repetition, wrong use of anaphora and, in the case of 'ChatGPT: chunks', confusing transitions. Aesthetics was mostly rated positively; the most common problem being 'length' ('Gemini: crosslingual chain-of-thought 1' was the only text referred to as 'too long' rather than 'too short'). The 'ChatGPT: chunks' text was noted to be lacking descriptions and character interaction.

## 5 Discussion

Although Bulgarian texts received comparatively lower scores in the conducted human evaluation, they were shown to be mostly competitive to counterparts in more highly resourced languages. Interestingly, they also tended to demonstrate different shortcomings compared with texts in English and French, such as excessive textual length.

Crosslingually derived texts were rated very highly by participants, notably occupying first place in the cases of French and English. Texts derived through the 'chunks' scenario were judged to have problems in relation to information and transitions, which leads us to hypothesise that an application of Chang et al. (2024)'s method of hierarchical merging would be of significant benefit.

The ChatGPT and Gemini models performed better than the smaller but instruction-tuned Mistral and BgGPT; however, the gap was not striking, Mistral-generated texts notably performing high in the conducted automatic evaluation.

The fact that agreement between participants in the survey is low speaks of high subjectivity, which in turn implies that the texts were mostly lacking obvious, objective drawbacks.

## 6 Conclusion and Future Directions

Four contemporary LLMs of different sizes and statuses of use were evaluated for their ability to provide abridged versions of a literary text. Three discrete languages were regarded: English, a relatively high-resourced language (French) and a relatively low-resourced language (Bulgarian).

Whilst English-language texts expectedly demonstrate superior quality, models such as ChatGPT and Gemini also perform competitively in other languages, whilst demonstrating different weaknesses in relation to different generation scenarios. Particularly, in a relatively low-resource language like Bulgarian, high quality text can be achieved if the models' limitations in terms of context length are overcome.

This study is an analytical inquiry into the current abilities of LLMs to generate abridged versions of literary texts on the basis of their original training data. These abilities are likely to be improved following additional training on relevant datasets as well as more elaborate prompting techniques.

A natural continuation of the presented study would be the exploration of abridgement by LLMs in relation to a variety of texts; this time with a focus on the models and abridgement scenarios that proved strongest.

## 7 Limitations

It is important to note that depth rather than width was opted for in the present study and its conclusions are mostly based on a single literary text. Therefore, key characteristics of original literary texts such as length and genre are disregarded as variables.

In addition, abridgement is considered as a general term and is not further broken down, such as based on targeted audience (e.g. children of a certain age). It should also be noted that application of the study's methods to additional texts is likely to necessitate refinement of the automatic evaluation metrics, such as type-to-token ratio, which is known to be highly dependent on a text's size. Finally, one respondent to the survey brought forward a text's similarity to a 'short summary' as a negative trait, whilst another one claimed that the text was a little 'too vivid' to be a 'summary'; which leads us to conclude that the term 'abridged version' is highly open to interpretation and that the survey would have benefited from a short definition of what is meant by it.

## References

R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, S. Petrov, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, and O. ... Vinyals. 2023. Gemini: A family of highly capable multimodal models. *arXiv*, 2312.11805.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation.

Zefeng Du, Wenxiang Jiao, Longyue Wang, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2023. On extrapolation of long-text translation with large language models.

William H. DuBay. 2007. *The Classic Readability Studies*. ERIC Clearinghouse.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *23rd International Conference on Computational Linguistics (COLING 2010), Poster Volume*, pages 276–284.

Joseph L Fleiss, Jacob Cohen, and Brian S Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327.

Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of llms on creative writing.

INSAIT. 2024. The latest advancements in llm: A comprehensive overview.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, and W. ... El Sayed. 2023. Mistral 7b. *arXiv*, 2310.06825.

Max Kreminski and Chris Martens. 2022. Unmet creativity support needs in computationally supported creative writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 74–82, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, and Giuseppe Riccardi. 2022. Evaluation of response generation models: Shouldn't it be shareable and replicable? In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 136–147, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.

Murray Shanahan and Catherine Clarke. 2023. Evaluating large language model creativity from a literary perspective.

Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256, Online. Association for Computational Linguistics.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text?

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Zero-shot cross-lingual summarization via large language models.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the wmt 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of llms.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by llms.

Le Xiao and Xiaolin Chen. 2023. Enhancing llm with evolutionary fine tuning for news summary generation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization.

## Appendix A  Automatically Evaluated Textual Features

| Textual feature | Definition |
|---|---|
| Total number of words | The number of words within the given text |
| Words per sentence | The average number of words per sentence in the text |
| Letters per word | The average number of letters per word in the text |
| Words not in frequency list | The number of words in the text that are not part of a defined frequency list. For English, the Dale-Chall list is considered. For Bulgarian and French, respectively, the top 3000 words from the Open Subtitles[1] and the Leeds Internet-FR Corpus[2] are taken. |
| Type-to-token ratio | The word-based (as opposed to lemma-based) ratio of types and tokens in the text |
| Concreteness | The average concreteness of the words found in Brysbaert's concreteness list |
| Anaphora-denoting words | The percentage of anaphora-related words in the text. For each language, these words are a defined set of definite articles, personal pronouns, demonstrative pronouns, relative pronouns, indefinite pronouns and adverbs of time and place |
| Ratio of content to function words | The approximate ratio of the words that carry semantical significance and the words that denote grammatical features in the text. For English, the CMU Pronouncing Dictionary for function words as available in Python's *nltk* library is used. For French and Bulgarian, part-of-speech tagging is applied to set apart the two kinds of words. Determiners, pronouns, conjunctions and adpositions are considered to be function words. |
| ROUGE | The ROUGE-1 recall value between an original and abridged text are taken. |
| BERTScore | The F1 BERTScore between an original and abridged text are considered. |

[1] https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Bulgarian_wordlist

[2] http://corpus.leeds.ac.uk/list.html

## Appendix B Human Evaluation: Evaluated Texts

English

| | Text 1 | Text 2 |
|---|---|---|
| Participant 1 | ChatGPT: crosslingual crain-of-thought 2 (BG) | ChatGPT: chunks |
| Participant 2 | ChatGPT: crosslingual crain-of-thought 2 (BG) | Gemini: one-shot |
| Participant 3 | Mistral: chunks | ChatGPT: crosslingual direct (BG) |
| Participant 4 | Mistral: chunks | Gemini: alternative |

French

| | Text 1 | Text 2 |
|---|---|---|
| Participant 1 | ChatGPT: alternative | ChatGPT: one-shot |
| Participant 2 | ChatGPT: alternative | Gemini: crosslingual direct (EN) |
| Participant 3 | Gemini: zero-shot | Gemini: crosslingual chain-of-thought 1 (EN) |
| Participant 4 | Gemini: zero-shot | ChatGPT: chunks |

Bulgarian

| | Text 1 | Text 2 |
|---|---|---|
| Participant 1 | ChatGPT: chunks | Gemini: chunks |
| Participant 2 | ChatGPT: chunks | Gemini: crosslingual direct (EN) |
| Participant 3 | ChatGPT: one-shot | Gemini: crosslingual chain-of-thought 1 (EN) |
| Participant 4 | ChatGPT: one-shot | ChatGPT: zero-shot |

## Appendix C Automatic Evaluation: Detailed Results

English

| | human-made 1[12] | human-made 2 [13] | human-made 3[14] | human-made 4[15] | 'Little Match Girl'[16] | ChatGPT: zero-shot | ChatGPT: chunks[17] |
|---|---|---|---|---|---|---|---|
| total words | 5.05 | 17.15 | 4.25 | 3.83 | 1.46 | 23.3 | **10.01**[18] |
| words per sentence | 2.89 | 3.15 | 1.91 | 2.77 | 1.65 | 1.64 | **1.67** |
| letters per word | 1.02 | 0.98 | 0.99 | 1.03 | 1.04 | 0.85 | 0.91 |
| words not in freq. list | 5.01 | 10.16 | 3.75 | 5.69 | 1.9 | 12.49 | **6.8** |
| TTR | 0.75 | 0.52 | 0.7 | 1.29 | 1.05 | 0.49 | **0.58** |
| concreteness | 0.94 | 0.91 | 0.97 | 0.91 | 0.91 | 0.99 | **0.97** |
| anaphora words | 0.8 | 1.23 | 1.01 | 1.57 | 0.91 | **1.15** | **1.05** |
| cont./funct. words | 1.04 | 0.53 | 0.86 | 0.74 | 1.63 | **1.58** | **1.48** |
| ROUGE | 0.67 | 0.58 | 0.58 | 0.64 | 0.4 | **0.6** | **0.61** |
| BERTScore | 0.84 | 0.82 | 0.83 | 0.82 | 0.82 | 0.81 | **0.82** |

| | ChatGPT: one-shot | ChatGPT: crossling. direct (FR)[19] | ChatGPT: crossling. direct (BG)[20] | ChatGPT: crossling. chain-of-thought 1 (FR)[21] [22] | ChatGPT: crossling. chain-of-thought 1 (BG)[23] | ChatGPT: crossling. chain-of-thought 2 (FR)[24] [25] | ChatGPT crossling. chain-of-thought 2 (BG)[26] |
|---|---|---|---|---|---|---|---|
| total words | 28.59 | 23.15 | **8.64** | 22.6 | 53.15 | 37 | **9.53** |

[12] Andersen, Hans Christian. The Ugly Duckling. Edited by Lynne Bradbury, Ladybird Books, adapted 1997.

[13] Andersen, Hans Christian. The Ugly Duckling. British Council. https://learnenglishkids.britishcouncil.org/sites/kids/files/attachment/story-time-the-ugly-duckling-transcript.pdf.

[14] Andersen, Hans Christian. The Ugly Duckling. https://www.joliet86.org/assets/1/6/The$_{U}gly_{D}uckling.pdf$.

[15] Andersen, Hans Christian. The Ugly Duckling. Edited by Maryann Dobeck, Parragon, 2009.

[16] Andersen, Hans Christian. The Little Match Girl. https://fliphtml5.com/mcbeq/hrvp/basic.

[17] The text was divided into 3 closely equal chunks. Depending on the language and model, some texts needed to be broken down into more chunks, in which case the number will be indicated.

[18] characteristics of the LLM-generated texts that fall within the gold standard range are marked in **bold**

[19] The text is directly abridged from the indicated language (here, French)

[20] The source text is not the full version but the abridged version 'human-made 1'

[21] The text is first translated from the source language (here, French) and then abridged using chain-of-thought prompts

[22] The source text is not the full version but the abridged version 'human-made 1'

[23] The source text is not the full version but the abridged version 'human-made 1'

[24] The source text is not the full version but the abridged version 'human-made 1'

[25] The text is first abridged in the source language (here, French) and then translated into the target language

[26] The source text is not the full version but the abridged version 'human-made 1'

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| words per sentence | **1.8** | 1.61 | **1.87** | **1.73** | 1.46 | 1.35 | **1.97** |
| letters per word | 0.94 | 0.81 | 0.94 | 0.92 | 0.83 | 0.89 | 0.92 |
| words not in freq. list | 20.59 | 12.49 | **5.82** | 16.21 | 26.28 | 29.31 | **6** |
| TTR | 0.47 | 0.46 | **0.57** | 0.5 | 0.43 | 0.48 | **0.53** |
| concreteness | **0.96** | 1.02 | 0.99 | **0.95** | 1.02 | 0.99 | 0.98 |
| anaphora words | **1.14** | **1.34** | **0.83** | **1.24** | **1.11** | **0.84** | **0.87** |
| cont./funct. words | **1.5** | **1.4** | **1.11** | **1.44** | **1.27** | 2.2 | **1** |
| ROUGE | 0.71 | **0.51** | **0.6** | **0.5** | **0.45** | **0.66** | **0.6** |
| BERTScore | **0.82** | 0.81 | 0.81 | 0.81 | 0.8 | 0.8 | 0.81 |

| | ChatGPT: alternative [27] | Gemini: zero-shot | Gemini: chunks | Gemini: one-shot | Gemini: crossling. direct (FR) | Gemini: crossling. direct (BG) | Gemini: crossling. chain-of-thought 1 (FR) |
|---|---|---|---|---|---|---|---|
| total words | **3.87** | 24.35 | **13.88** | **16.41** | **16.55** | 23.01 | 24.99 |
| words per sentence | 0.78 | 1.56 | **1.96** | **1.71** | 1.39 | 1.83 | 1.23 |
| letters per word | 0.92 | 0.85 | 0.87 | 0.94 | 0.84 | 0.81 | 0.93 |
| words not in freq. list | **3.28** | 13.61 | **7.86** | 11.21 | **10.03** | 12.49 | 16.21 |
| TTR | **0.79** | 0.47 | **0.53** | **0.55** | 0.51 | 0.51 | 0.49 |
| concreteness | 1.04 | 1.05 | 1.02 | **0.93** | 1.01 | 1.03 | **0.96** |
| anaphora words | 1.67 | **1.1** | **1.29** | **1.06** | **0.91** | **1.51** | **1.07** |
| cont./funct. words | **1.43** | **1.57** | **1.15** | **1.42** | **1.3** | 1.95 | **1.16** |
| ROUGE | **0.63** | **0.53** | **0.55** | **0.64** | **0.53** | **0.46** | **0.57** |
| BERTScore | 0.85 | 0.64 | 0.7 | 0.65 | 0.8 | 0.8 | **0.82** |

[27] The alternative text ('The Gift under the Bush') is abridged in a zero-shot setting

| | Gemini: crossling. chain-of-thought 1 (BG) | Gemini: crossling. chain-of-thought 2 (FR) | Gemini: crossling. chain-of-thought 2 (BG) | Gemini: alternative | Mistral: zero-shot | Mistral: chunks [28] | Mistral: crossling. direct (BG)[29] | Mistral: alternative |
|---|---|---|---|---|---|---|---|---|
| total words | 25.16 | 41.02 | 53.15 | **8.11** | **4.32** | **8.56** | **2.31** | **5.53** |
| words per sentence | 1.38 | 1.39 | 1.49 | 0.83 | 0.64 | 1.5 | 0.42 | 0.84 |
| letters per word | 0.87 | 0.88 | 0.93 | **0.99** | 0.97 | **1** | 0.89 | 0.95 |
| words not in freq. list | 13.61 | 20.05 | 38.1 | **7.29** | **2.92** | **6.4** | 1.88 | **5.02** |
| TTR | 0.47 | 0.44 | 0.42 | **0.78** | **0.68** | **0.59** | **0.75** | **0.7** |
| concreteness | 1.01 | **0.96** | 1.03 | 1 | 0.99 | 0.98 | 1.04 | 1 |
| anaphora words | **1.23** | **1.31** | **1.11** | **1.36** | 1.69 | **1** | **1.24** | **1.33** |
| cont./funct. words | **1.18** | **0.96** | **1.53** | **1.43** | **1.48** | **1.59** | **1.07** | **1.41** |
| ROUGE | **0.55** | **0.5** | 0.74 | **0.63** | **0.61** | **0.61** | **0.38** | |
| BERTScore | 0.81 | 0.8 | 0.81 | **0.83** | 0.86 | 0.81 | **0.83** | |

[28] The text was divided into 5 chunks.

[29] The source text is not the full version but the abridged version 'human-made 1'

| | |
|---|---|
| ROUGE | 0.69 |
| BERTScore | 0.85 |

French

| | human-made 1[30] | human-made 2[31] | human-made 3[32] | 'Little Match Girl' [33] | ChatGPT: zero-shot | ChatGPT: chunks | ChatGPT: one-shot |
|---|---|---|---|---|---|---|---|
| total words | 2.48 | 4.55 | 3.38 | 1.06 | 32 | 11.73 | 21.81 |
| words per sentence | 2.22 | 0.96 | 0.92 | 0.95 | 0.76 | 0.84 | **1.42** |
| letters per word | 1.1 | 0.96 | 0.96 | 1 | 0.95 | 0.95 | **0.97** |
| words not in freq. list | 2.34 | 3.83 | 2.67 | 1.17 | 23.42 | 8.6 | 14.05 |
| TTR | 0.87 | 0.71 | 0.74 | 1.03 | 0.54 | 0.63 | 0.6 |
| anaphora words | 1.08 | 1.13 | 1.09 | 0.92 | 1.2 | **1.03** | 0.81 |
| cont./funct. words | 0.8 | 0.72 | 1.04 | 1.05 | 0.44 | **0.91** | **0.73** |
| ROUGE | 0.55 | 0.48 | 0.53 | 0.59 | **0.58** | **0.54** | **0.59** |
| BERTScore | 0.74 | 0.71 | 0.73 | 0.8 | 0.62 | 0.63 | 0.66 |

| | ChatGPT: crossling. direct (EN) | ChatGPT: crossling. chain-of-thought 1 (EN) | ChatGPT: crossling. chain-of-thought 2 (EN) | ChatGPT: alternative | Gemini: zero-shot | Gemini: chunks | Gemini: one-shot |
|---|---|---|---|---|---|---|---|
| total words | 0.78 | 39.25 | 17.12 | 9.6 | 22.3 | 13.89 | 3.71 |
| words per sentence | 0.64 | 0.58 | 0.82 | 0.68 | **1.12** | 0.81 | 0.58 |
| letters per word | 1.04 | 0.93 | 0.87 | **0.97** | **0.98** | 0.87 | 0.93 |

[30] Andersen, Hans Christian. Le vilain petit canard. https://data.over-blog-kiwi.com/1/11/17/78/20210801/ob$_{8}$730$b8_{l}e - vilain - petit - canard - tapuscrit.pdf$.

[31] Andersen, Hans Christian. Le vilain petit canard. BIGBEN Kids. https://www.bigben.fr/wp-content/uploads/2021/10/Histoire$_{l}evilainpetitcanard.pdf$.

[32] Andersen, Hans Christian. Le vilain petit canard. https://bloc-note.ac-reunion.fr/9741309e/files/2020/03/0-conte-le-vilain-petit-canard.pdf.

[33] Andersen, Hans Christian. La petite fille aux allumettes. https://miladlh.com/wp-content/uploads/2020/11/La-Petite-Fille-aux-Allumettes.pdf.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| words not in freq. list | 0.91 | 25.55 | 10.81 | 7.2 | **2.93** | 10.81 | 25.54 |
| TTR | 1.18 | 0.47 | 0.54 | 0.66 | **0.79** | 0.54 | 0.47 |
| anaphora words | **1.11** | 1.34 | **1.07** | **1** | 1.33 | **1.07** | 1.34 |
| cont./funct. words | 1.5 | 0.56 | 0.71 | **1.03** | 1.33 | 0.71 | 0.56 |
| ROUGE | 0.36 | 0.63 | 0.45 | 0.65 | 0.34 | 0.45 | 0.63 |
| BERTScore | **0.77** | 0.61 | 0.63 | **0.71** | 0.69 | 0.63 | 0.61 |

| | Gemini: crossling. direct (EN) | Gemini: crossling. chain-of-thought 1 (EN) | Gemini: crossling. chain-of-thought 2 (EN) | Gemini: alternative |
|---|---|---|---|---|
| total words | 13.26 | 28.31 | 22.82 | 13.95 |
| words per sentence | **1.03** | **1.29** | 0.44 | 0.77 |
| letters per word | 0.94 | **0.98** | 0.86 | **1** |
| words not in freq. list | 9.47 | 16.53 | 14.79 | 11.52 |
| TTR | 0.61 | 0.53 | 0.56 | 0.62 |
| anaphora words | **1.1** | **0.99** | 1.23 | 1.72 |
| cont./funct. words | **0.75** | 0.56 | 0.64 | 0.47 |
| ROUGE | **0.58** | **0.52** | 0.44 | 0.69 |
| BERTScore | 0.63 | 0.68 | 0.62 | 0.68 |

Bulgarian

| | human-made 1 [34] | human-made 2 [35] | human-made 3 [36] | human-made 4 [37] | 'Little Match Girl' [38] | BgGPT: zero-shot[39] | BgGPT: chunks [40] |
|---|---|---|---|---|---|---|---|

[34] Andersen, Hans Christian. Groznoto patentse. https://roditel.bg/groznoto-patentse-prikazka-andersen/.

[35] Andersen, Hans Christian. Groznoto patentse. Edited by Tanya Petkova, adapted 2020. https://www.ourboox.com/books/грозното-патенце-2/.

[36] Andersen, Hans Christian. Groznoto patentse. Prikazki s Dji Dji. https://taleswithgigi.bg/the-ugly-duckling/.

[37] Andersen, Hans Christian. Groznoto pate. Zlatnoto pate, adapted 2007.

[38] Andersen, Hans Christian. Malkata kibritoprodavachka. Prikazki ot tsyal svyat, transl. Vasil Velchev, 2009.

[39] The source text is not the full version but the abridged version 'human-made 1'

[40] The text was divided into 4 chunks.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| total words | 8.71 | 5.21 | 9.1 | 3.45 | 1.76 | 1.62 | 1.39 |
| words per sentence | 1.01 | 0.99 | 0.94 | 1.16 | 2.4 | 0.89 | **1.02** |
| letters per word | 0.9 | 0.92 | 1 | 0.95 | 0.93 | 1.04 | **0.98** |
| words not in freq. list | 6.01 | 4.2 | 7.46 | 2.86 | 1.58 | **1.83** | 1.33 |
| TTR | 0.65 | 0.72 | 0.69 | 0.77 | 0.86 | 0.92 | 0.91 |
| anaphora words | 1.23 | 1.07 | 1.04 | 1.2 | 1.3 | 0.49 | 0.94 |
| cont./funct. words | 0.68 | 0.4 | 0.71 | 1.07 | 1.8 | **1.58** | **0.94** |
| ROUGE | 0.45 | 0.47 | 0.56 | 0.38 | 0.33 | 0.29 | 0.7 |
| BERTScore | 0.69 | 0.7 | 0.73 | 0.69 | 0.71 | 0.75 | 0.85 |

| | BgGPT: one-shot | BgGPT: alternative | ChatGPT: zero-shot[41] | ChatGPT: chunks [42] | ChatGPT: one-shot[43] | ChatGPT: crossling. direct (EN) | ChatGPT: crossling. chain-of-thought 1 (EN)[44] |
|---|---|---|---|---|---|---|---|
| total words | 1.14 | 1.04 | **2.22** | **2.1** | **2.03** | **1.87** | 32.65 |
| words per sentence | **1.03** | **0.98** | 0.75 | **1.11** | 0.93 | **1.16** | 0.54 |
| letters per word | 1.03 | **1** | 1.02 | **0.95** | **0.97** | **0.92** | **0.96** |
| words not in freq. list | 1.16 | 1.07 | **2.58** | **1.82** | **2.04** | 26.11 | **1.83** |
| TTR | 0.98 | 0.98 | 0.97 | **0.83** | 0.96 | 0.62 | 0.92 |
| anaphora words | 0.82 | 0.76 | 0.56 | **1.06** | 0.53 | 0.82 | **1.05** |
| cont./funct. words | **1.13** | **1.05** | 2.19 | **1.29** | 1.98 | 2.47 | 0.14 |
| ROUGE | 0.59 | 0.76 | **0.37** | 0.63 | **0.44** | **0.49** | **0.37** |
| BERTScore | 0.81 | 0.83 | **0.71** | 0.78 | **0.72** | 0.64 | 0.78 |

[41] The source text is not the full version but the abridged version 'human-made 1'

[42] The text was divided into 4 chunks.

[43] The source text is not the full version but the abridged version 'human-made 1'

[44] The source text is not the full version but the abridged version 'human-made 4'

| | ChatGPT: crossling. chain-of-thought 2 (EN) | ChatGPT: alternative | Gemini: zero-shot[45] | Gemini: chunks The text was divided into 4 chunks. | Gemini: one-shot | Gemini: crossling. direct (EN)[46] | Gemini: crossling. chain-of-thought 1 (EN) |
|---|---|---|---|---|---|---|---|
| total words | 24.49 | 1.03 | 1.4 | **2.39** | 1 | **5.4** | **3.95** |
| words per sentence | 0.92 | **0.97** | **1.23** | **1.34** | 0.87 | **1.07** | 0.83 |
| letters per word | 0.88 | **0.99** | **1** | **0.94** | **1** | **0.96** | **0.91** |
| words not in freq. list | 15.27 | 1.03 | 1.54 | **2.09** | 1.17 | **4.53** | **3.43** |
| TTR | 0.57 | 0.99 | 0.96 | **0.85** | 1.04 | **0.76** | **0.76** |
| anaphora words | 0.66 | 0.69 | 0.89 | 0.92 | 0.75 | **1.15** | **1.27** |
| cont./funct. words | **1.78** | **1.02** | **1.35** | **1.04** | **1.03** | 0.39 | 0.55 |
| ROUGE | **0.46** | 0.71 | **0.51** | **0.49** | 0.22 | **0.46** | **0.42** |
| BERTScore | 0.66 | 0.86 | 0.81 | 0.74 | 0.74 | **0.72** | 0.78 |

| | Gemini: crossling. chain-of-thought 2 (EN) | Gemini: alternative |
|---|---|---|
| total words | 17.53 | 1.07 |
| words per sentence | 0.67 | **1** |
| letters per word | 0.77 | 1.01 |
| words not in freq. list | 10.9 | 1.13 |
| TTR | **0.65** | 1.07 |
| anaphora words | **1.15** | 0.7 |
| cont./funct. words | **1.42** | **1.08** |
| ROUGE | 0.31 | 0.7 |
| BERTScore | 0.65 | 0.82 |

[45] The source text is not the full version but the abridged version 'human-made 1'

[46] The source text is not the full version but the abridged version 'human-made 1'