GPT-Based Lexical Simplification for Multi-Word Expressions Using Prompt Engineering

Sardar Khan Khayamkhani
Manchester Metropolitan University
sardarkhanksk@gmail.com

Matthew Shardlow
Manchester Metropolitan University
m.shardlow@mmu.ac.uk

Abstract

Multiword Lexical Simplification (MWLS) is the task of replacing a complex phrase in a sentence with a simpler alternative. Whereas previous approaches to MWLS made use of the BERT language model, we make use of the Generative Pre-trained Transformer architecture. Our approach employs Large Language Models in an auto-regressive format, making use of prompt engineering and few-shot learning to develop new strategies for the MWLS task. We experiment with several GPT-based models and differing experimental settings including varying the number of requested examples, changing the base model type, adapting the prompt and zero-shot, one-shot and k-shot in-context learning. We show that a GPT-4o model with k-shot in-context learning (k=6) demonstrates state-of-the-art performance for the MWLS1 dataset with NDCG=0.3143, PREC@5=0.1048, beating the previous BERT-based approach by a wide margin on several metrics and consistently across subsets. Our findings indicate that GPT-based models are superior to BERT-based models for the MWLS task.

1 Introduction

Lexical Simplification (LS) is a Natural Language Processing (NLP) task that enhances text accessibility by replacing complex words with simpler alternatives. It has been applied in various application settings, including non-native speakers (Paetzold and Specia, 2016b), multilingualism (Shardlow et al., 2024b), education (Uchida et al., 2018), and readability (Maddela and Xu, 2018). By simplifying lexical content, NLP systems aim to promote inclusivity and comprehension.

Multi-Word Lexical Simplification (MWLS) extends LS to multi-word expressions (MWEs), requiring models to preserve meaning while reducing complexity. The Plainifier system (Przybyła and Shardlow, 2020) was previously used to address these challenges and has been evaluated using the MWLS1 dataset[1].

This paper explores LLMs for MWLS using the MWLS1 dataset, which contains 1,462 instances and 7,059 human-provided simplifications, offering a strong benchmark. Recent models like GPT-3 (Brown et al., 2020), GPT-4 (Achiam et al., 2023), and Llama 3.1 (Touvron et al., 2023) have demonstrated effectiveness in NLP tasks, including text simplification, through prompt engineering. Studies show that task-specific prompts help guide models in simplifying MWEs while preserving meaning (Shin et al., 2020; Gao et al., 2021).

This work investigates prompt engineering (Liu et al., 2023) to enhance MWLS using LLMs. The proposed approach generates simplified alternatives while maintaining context, contributing to the development of accessible NLP tools.

Despite advancements in Lexical Simplification (LS), most research focuses on single-word simplification to aid non-native speakers and individuals with language impairments (Saggion et al., 2022; Shardlow et al., 2024a). However, existing LS techniques struggle with multi-word expressions (MWEs), which often carry context-dependent meanings. Traditional LS methods, designed for single-word substitutions, fail to handle these complexities effectively (Constant et al., 2017; Gooding and Kochmar, 2018). Whilst the Plainifier attempts to simplify MWEs while preserving meaning, it is based on Masked Language Modelling (Devlin et al., 2019), leaving room for advancement through the use of large Causal Language Models (Brown et al., 2020).

This paper explores AI-driven large language models (LLMs) for MWLS, focusing on GPT-based models. The goal is to assess whether LLMs

[1] https://github.com/piotrmp/mwls1

Proceedings of Recent Advances in Natural Language Processing,pages 546–556
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-064

can provide reliable, context-aware simplifications, advancing MWLS and improving text accessibility across diverse domains. The primary contribution is to investigate the potential of Generative Pre-trained Transformers (GPT) models for Multi-Word Lexical Simplification (MWLS). By designing customized prompts, this study explores how GPT models can be guided to generate simpler alternatives for complex multi-word expressions while preserving the original semantic content.

The work is structured as follows: We first overview related literature in Section 2. We then describe the resources and experimental methodology used in Section 3. Next, we present the results for each experiment in Section 4. We conclude with a discussion of the results in Section 5.

2 Related Work

2.1 Multi-Word Lexical Simplification

Lexical Simplification (LS) in Natural Language Processing (NLP) improves text accessibility by replacing complex words or phrases with simpler alternatives while preserving meaning. It is widely used in education, technology, and communication, helping non-native speakers and broader audiences (Siddharthan, 2014). Traditional LS methods, such as dictionary-based substitution and frequency analysis, often ignore context, leading to oversimplifications or errors (Zhu et al., 2010; Glavaš and Štajner, 2015). While effective for single-word simplification, these approaches struggle with complex structures, limiting their practical use.

Multi-Word Lexical Simplification (MWLS) extends LS to multi-word expressions (MWEs) like idioms, collocations, and technical terms. These expressions pose challenges due to their context-dependent meanings, which traditional LS models fail to simplify effectively (Constant et al., 2017). Systems like Plainifier (Paetzold and Specia, 2016b) address MWLS but lack flexibility and deep contextual awareness. This highlights the need for AI-driven solutions that can preserve meaning while adapting to linguistic complexity. By leveraging large language models (LLMs), this study explores improved MWLS approaches with greater accuracy, adaptability, and contextual understanding.

2.2 Approaches to LS

Lexical Simplification (LS) initially relied on rule-based, dictionary-driven methods, replacing complex words with simpler synonyms from predefined lists (Carroll et al., 1998). While effective for basic tasks, these methods lacked context awareness, often leading to inaccurate substitutions (Siddharthan, 2014). Early systems using Simple Wikipedia struggled with polysemy, idioms, and multi-word expressions (MWEs), limiting their applicability.

Modern LS approaches leverage machine learning, particularly supervised models trained on parallel corpora like Newsela, enabling better contextual preservation (Xu et al., 2015). LSBert, for example, applies BERT's masked language modelling to rank simplifications, improving results on datasets like LexMTurk (Qiang et al., 2020). However, even these models falter with MWEs, often generating overly simplified or inappropriate alternatives. MWEs (including idioms and collocations) remain challenging due to their context-dependent meanings. MWLS1 is a dedicated dataset, providing annotated examples to assess MWE simplification models (Przybyła and Shardlow, 2020).

Pretrained models like BERT and RoBERTa improved LS by using semantic embeddings, enhancing simplifications across domains (Liu et al., 2023). GPT models have recently shown promise for controlled simplification, particularly in technical fields like biomedical text (Li et al., 2024). However, LS still struggles with MWEs, fluency, and adaptability, limiting large-scale adoption (Lee and Yeung, 2018). A unified approach integrating advanced language models and multilingual datasets could bridge these gaps, pushing LS toward truly accessible text simplification.

2.3 LLMs in Natural Language Processing

Large Language Models (LLMs) have transformed Natural Language Processing (NLP) by generating contextually accurate and coherent text across various applications (Radford et al., 2019). Models like GPT-3.5 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023) excel in handling complex linguistic structures, making them effective for tasks such as text generation, translation (Kocmi et al., 2024), summarization (Zhang et al., 2023), and lexical simplification (Aumiller and Gertz, 2022). Their pre-training on vast datasets enables them to process polysemous terms, idioms, and multi-word expressions (MWEs) while maintaining semantic integrity (Liu et al., 2023).

Prompt engineering plays a key role in guiding LLMs for text simplification, ensuring fluency and contextual accuracy (Kew et al., 2023; Shin et al., 2020). However, challenges remain—LLMs can sometimes oversimplify, altering meaning or omitting important details. Additionally, their black-box nature limits interpretability, making it difficult to analyse decision-making, especially in fields like medical text simplification (Liu et al., 2019). Despite these issues, LLMs offer unprecedented adaptability, integrating readability, contextual relevance, and semantic depth, positioning them as powerful tools for advancing lexical simplification research.

3 Methodology

3.1 MWLS1 Dataset

The MWLS1 dataset, introduced by Przybyła and Shardlow (2020), is specifically designed for Multi-Word Lexical Simplification (MWLS). This dataset supports the evaluation and refinement of MWLS systems by providing annotated examples of multi-word expressions (MWEs) and their simpler alternatives. Publicly available, it facilitates collaboration and iterative improvements in the field of MWE simplification. By adopting MWLS1, this research aligns with established standards and contributes to advancing simplification methodologies. The dataset comprises 1,462 English sentences and 7,059 human-annotated simplifications, sourced from three domains:

1. BIBLE: Parallel corpus of the World English Bible (Christodouloupoulos and Steedman, 2015).

2. EUROPARL: Proceedings from the European Parliament (Koehn, 2005).

3. BIOMED: Text from the CRAFT biomedical corpus (Bada et al., 2012).

The dataset provides a categorisation of simplification targets into unigrams (e.g., *resistance*), bigrams (e.g. *indirect consequences*), and trigrams (e.g., *detect random integration*), enabling a detailed analysis of single-word and phrase-level simplification. Each instance in the MWLS corpus consists of a context, a target word or phrase and the potential replacements for that phrase. The target may be a single word, bigram or trigram and the replacements may also be n-grams where n = 1, 2 or 3. Examples taken from the dataset are shown in Table 1, which is reproduced from Przybyła and Shardlow (2020). The dataset is further divided into subsets such as bible1, bible2, europarl3, etc., ensuring balanced representation across n-gram sizes and linguistic complexities.

3.2 Use of GPT models

We used GPT-3.5-Turbo, GPT-4o and GPT-4o-Mini. All models were used via the OpenAI API and were accessed in a period between June 2023 and May 2025, with a final rerun of all models in May 2025. We did not experiment with models beyond the GPT family due to limited resources in the experiments. Although the GPT models are closed source and only accessible via an API, they do represent a high-performing set of models across many tasks and have been used successfully in previous lexical simplification approaches to gain state of the art performance (Aumiller and Gertz, 2022; Enomoto et al., 2024). We overview each model below with reference to the OpenAI model cards[2]:

GPT-3.5-Turbo: An updated version of GPT-3 (Brown et al., 2020) which is fine-tuned for dialogue interactions through supervised learning and reinforcement learning with human feedback. This is a text-only model. the model costs $0.50 per 1M input tokens.

GPT-4o: A more recent model than GPT3.5, this model is a flagship model provided by the OpenAI platform. GPT-4o is trained on text and images, although we only make use of text completions in this study (Hurst et al., 2024). Interaction with the model costs $2.50 per 1M input tokens.

GPT-4o-Mini: A smaller[3] version of the GPT-4o model. This model is cheaper to run ($0.16 per 1M input tokens) and has a faster inference speed than the GPT-4o model.

The experimental protocol we undertook for Multi-Word Lexical Simplification with GPT consists of loading the dataset, passing each complicated phrase and its respective sentence to the GPT model with prompt via an API request and storing the results for evaluation. The first author checked model outputs to identify failure cases (e.g., no output, boilerplate, user-facing messages) and reran all necessary cases.

[2] https://platform.openai.com/docs/models

[3] It is not specified by OpenAI how much smaller, or how the smaller model was created.

Case ID	Source	Sentence	Replacement
CASE_7739	BIOMED	The main difference in the two lines **essentially resides** in the strength of the promoter.	is basically
CASE_241	BIBLE	Thus says Yahweh of Armies, ”They **shall thoroughly glean** the remnant of Israel. Turn again your hand as a grape gatherer into the baskets.”	will gather
CASE_5327	EUROPARL	I support Ms Lulling's recommendations that the national systems should recognise the importance of protecting self-employed workers, and we should stand against all forms of **discrimination**, but I am still not convinced that this House is best placed to work on employment matters.	bias and unfairness
CASE_6461	BIOMED	Other **potentially biologically relevant** substrates include cholecystokinin and possibly other neuropeptides.	relevant
CASE_2260	BIBLE	A man's **foes** will be those of his own household.	enemies

Table 1: Five examples of sentences from the MWLS1 dataset, each shown with its identifier, source corpus, highlighted target and one replacement provided by annotators. Reproduced from Przybyła and Shardlow (2020).

Figure 1 shows the workflow that we used. The process begins with the OpenAI client being initialised. For each sentence in the dataset, the system finds the complicated phrase and a prompt is sent to the GPT model. The GPT model evaluates the prompt and generates a list of simplified alternatives, which are then saved in a new column in the dataset. This procedure is repeated for every instance in the dataset. A new session is initialised for each instance to prevent effects of catastrophic forgetting (Vu et al., 2022) in long prompting sessions.

We used the following system prompt in a zero-shot setting:

> *You are an AI that suggests simpler alternatives for complex words and phrases.*

And the corresponding system prompt for one- or few-shot learning. For one-shot learning we only specified a single example. For few-shot learning we specified $K = 6$ examples:

> *You are an AI that suggests simpler alternatives for complex words and phrases. Here are some examples of how to simplify multi-word expressions:*
> *<Example_1 Instance>*
> *<Example_1 Response>*
> ⋮
> *<Example_K Instance>*
> *<Example_K Response>*

For each instance across all experiments, we provided the initial prompt to the GPT model via the 'User' role:

> *Suggest* $\{n\}$ *simpler alternatives to the multi-word phrase* $\{complex\_word\}$ *in the following sentence. The alternatives should preserve the meaning and improve readability for non-native speakers. Here is the sentence:* $\{sentence\}$

Where the parameter n varies by experiment as shown in Table 2 and the *complex_word* and *sentence* are instance dependent. The model response was collected using the 'Assistant' role.

3.3 Experiments

We used the full MWLS corpus to evaluate the effectiveness of GPT models for Multi-Word Lexical Simplification (MWLS). We do not make use of training data as our approach is fully unsupervised and as such the full corpus is used for evaluation. The experiments we undertook modified key parameters of the workflow to assess their impact on simplification quality as shown in Table 2.

In Experiment 1, the number of alternatives was 30 and we introduced a prompt which focussed on readability for non-native speakers. This aimed to determine performance level and provide clear instructions for simplification quality. Experiment 2 reduced the number of alternatives to 15 and introduced a one-shot system prompt, offering a single

Parameters	Exp 1	Exp 2	Exp 3	Exp 4
Alternatives (n)	30	15	15	20
Model	GPT-3.5-Turbo	GPT-3.5-Turbo	GPT-4o	GPT-4o-Mini
K-Shot	Zero-Shot	One-Shot	Few-Shot	

Table 2: Varied experimental protocols used in our research

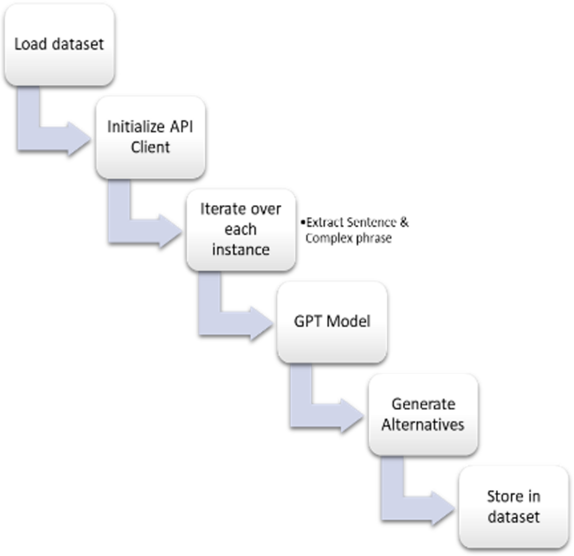

Figure 1: The experimental workflow we employed.

example to guide the model. This tested whether quality improved when fewer but more precise simplifications were requested. Experiment 3 modified the model to GPT-4o, maintaining 15 alternatives while also introducing a few-shot system prompt (6 examples) to assess how multiple examples influenced output relevance. Experiment 4 switched to GPT-4o-mini, increasing alternatives to 20 while retaining the same few-shot prompt. This evaluated whether a more compact model could maintain simplification quality while generating a broader range of alternatives.

3.4 Baseline Systems

We reproduce the results of the Plainifier system introduced by Przybyła and Shardlow (2020). The Plainifier system makes use of the Masked Language Model BERT (Devlin et al., 2019) which produces token-level probabilities across a vocabulary to replace a masked element in a sentence. The complex target (which may be a 1-, 2- or 3-gram) is masked and the language model is used to produce a set of replacements which are reordered according to language model probability, semantic similarity and familiarity.

We additionally reproduce the human baseline reported by Przybyła and Shardlow (2020). This was produced using the original annotations as a gold standard. The human baseline effectively demonstrates how well any one annotator can predict the suggested replacements of all other annotators and demonstrates that the task is inherently difficult for humans to perform.

3.5 Evaluation Metrics

We use the following metrics, which are the same as those used in the original work on MWLS and are used widely elsewhere in lexical simplification research:

Potential: A metric determining the average number of instances where at least one generated candidate is also in the gold-standard (Paetzold and Specia, 2016a). This is a permissive metric that demonstrates the potential of a substitution generation system to produce accurate simplifications.

Precision@K: The percentage of the first K generated candidates that are present in the gold standard (Štajner et al., 2022), averaged per-instance. In our setting we use K=5.

NDCG: Normalised discounted cumulative gain (Järvelin et al., 2008) considers the system-generated substitutions as a ranked list, casting the evaluation as an information retrieval problem (i.e., where the vocabulary is considered the information space and the gold standard is the items to retrieve).

All metrics are evaluated for each subset in the dataset, where a subset is a specific genre (bible, biomed and Europparl) for each N-gram (unigram, bigram and trigram). We present summary level statistics for unigrams, bigrams and trigrams (summarised across genre) and for all genres (summarised across n-gram size).

Metric	Subset	Plainifier	Human	Exp 1	Exp 2	Exp 3	Exp 4
NDCG	Bible_All	0.1528	0.1116	0.3022	0.2598	**0.3451**	0.3060
	Biomed_All	0.1460	0.1065	0.2554	0.2260	**0.2735**	0.2349
	Europarl_All	0.1575	0.0999	0.2679	0.2904	**0.3163**	0.2898
	All1	0.2578	0.2112	**0.4997**	0.5172	0.4851	0.3938
	All2	0.0936	0.0475	0.1725	0.1421	0.2512	**0.2567**
	All3	0.0458	0.0266	0.1191	0.0792	**0.1748**	0.1650
	All	0.1396	0.1015	0.2766	0.2610	**0.3143**	0.2798
Prec@5	Bible_All	0.0343	**0.2174**	0.1007	0.0906	0.1164	0.0976
	Biomed_All	0.0360	**0.2059**	0.0856	0.0785	0.0872	0.0692
	Europarl_All	0.0465	**0.1913**	0.0958	0.1024	0.1073	0.0909
	All1	0.0767	**0.3811**	0.1936	0.1970	0.1812	0.1440
	All2	0.0180	**0.1069**	0.0455	0.0390	0.0738	0.0668
	All3	0.0079	**0.0575**	0.0280	0.0201	0.0455	0.0397
	All	0.0365	**0.1929**	0.0946	0.0913	0.1048	0.0871
Potential	Bible_All	**0.9366**	0.2174	0.4398	0.3501	0.4639	0.4201
	Biomed_All	**0.8656**	0.2059	0.3787	0.2997	0.3651	0.3025
	Europarl_All	**0.8765**	0.1913	0.4040	0.3907	0.4459	0.4128
	All1	**0.9908**	0.3811	0.7137	0.6667	0.6261	0.5150
	All2	**0.8406**	0.1069	0.2854	0.2204	0.3735	0.3550
	All3	0.0458	0.0266	0.1746	0.1058	0.2487	**0.2540**
	All	**0.6451**	0.0575	0.4096	0.3500	0.4291	0.3837

Table 3: Experimental results. *Plainifier* and *Human* baselines are reproduced from Przybyła and Shardlow (2020).

4 Results

We present the results of our experiments in Table 3. We have reported three metrics (NDCG, Precision @5 and Potential) as defined previously. Each metric is evaluated for each subset of our corpus and the summary statistics are reported for each genre (*Bible_All*, *Biomed_All* and *Europarl_All*) and each N-gram size (*All1*, *All2* and *All3*). We also present summary statistics across the entire dataset as *All*.

Results are reported for the Plainifier system and the human baseline. These results are reproduced directly from the work of Przybyła and Shardlow (2020). We continue to report the results of the four experimental settings that we evaluated with our GPT-based MWLS system. We have highlighted the highest performing system in boldface type in each row (per-summary-statistic and per-metric) for the reader's convenience. We also highlight the key findings from our results in bold-face below.

In experiment 1 we deployed GPT-3.5 Turbo in a zero-shot setting, whereas in experiment 2 we deployed the same model in a one-shot setting. The results of experiment 2 are generally lower for all metrics than for experiment 1, with the exception of NDCG and Prec@5 for Europarl_All. We also changed the number of generated candidates in this setting from 30 to 15, which may have affected the potential metric, but should not have affected other metrics as the generated candidates beyond the 15th place were typically incorrect[4]. The drop in performance indicates that **the inclusion of a single in-context-learning example with 15 replacements did not lead to improved performance compared to zero-shot with 30 substitutions** in our experimental setting.

For experiment 3, we used the GPT-4o model instead of the GPT-3.5-Turbo model. Additionally, the experimental configuration used 6 in-context-learning examples instead of 1 as in experiment 2. The results of experiment 3 demonstrate improved performance on all metrics compared to all other experiments that we performed. The setting of experiment 3 had the highest performance across all summary subsets for NDCG, except for *All1* (unigrams) and *All2* (bigrams). Experiment 3 also showed improved performance over other experimental settings for Precision@5, where it attained the highest result of any GPT-based system for all subsets except for *All1* (unigrams). Experiment 3 also demonstrated the highest GPT-based perfor-

[4] Only the top 5 candidates are taken for prec@5 and lower order candidates have a minimal effect for NDCG

mance in the potential metric, yielding the highest result across the entire dataset (*All*). This indicates that **the use of GPT-4o with 6 examples of in-context learning gave the best machine-based performance for our dataset.**

We additionally experimented with the use of GPT-4o-Mini in the setting of Experiment 4. Here we are using a smaller but more efficient (time and token-cost) model, which is known to perform marginally worse than GPT-4o on baseline tasks. Our experiments demonstrate that GPT-4o-mini typically has the second-best performance on most evaluated subsets across all metrics. In 2 instances (NDCG All2 and Potential All3) GPT-4o-Mini outperforms GPT-4o. This indicates that **smaller language models can be competitive with larger language models for the MWLS task at lower cost and faster inference speed.**

Our results outperform those of the baseline system (the Plainifier model) for NDCG and Precision@5. Our best performing system based on GPT-4o represents a significant increase in performance for NDCG (Plainifier = 0.1396, GPT4o = 0.3143) and for Precision@5 (Plainifier = 0.0365, GPT4o = 0.1048). This improvement for NDCG and Precision@5 is also realised across every evaluated subset. **The plainifier outperforms all of our systems for the Potential metric**, due to the limited number of candidates generated by the GPT-based experiments.

Whereas the human baseline previously outperformed the Plainifier system, **we report an improvement in the NGCD metric over human performance. This is the first time that this has been reported for the MWLS1 dataset for NDCG.** Our best-performing system, based on GPT-4o outperforms the human baseline for all subsets on the NDCG metric by a large margin (Human-baseline = 0.1015, GPT-4o = 0.3143). Our system also outperforms the human baseline in terms of potential across all subsets, although humans typically only returned a few candidates each. **The human baseline outperforms our best-performing system for the precision@5 metric** (Human-baseline = 0.1929, GPT-4o = 0.1048), indicating that humans are generally closer to each others predictions for the top-5 candidates than GPT-based systems.

For the NDCG metric, we demonstrate a new state-of-the-art performance including outperforming the human baseline and previous state-of-the-art. Our system based on GPT-4o demonstrates the best performance of those systems evaluated, although we also report improved performance for unigrams with a system based on GPT-3.5 and bigrams with a system based on GPT-4o-mini. The improved NDCG results indicate that **our system consistently ranks relevant candidates higher in the returned list than previous attempts.**

For the precision@5 metric, we demonstrate that GPT-based systems produce a top-5 ranked candidate list that is more similar to a list produced by the MWLS1 annotators. Our GPT-based systems consistently outperform the Plainifier system, representing an improvement in automated approaches to MWLS for this metric.

Our GPT-based systems perform consistently higher than the human baselines for potential indicating that the system generally identified at least one relevant candidate. Although the Plainifier system performs very well on this metric for most subsets, our GPT-based system performs better on trigram generation (All3) where we return state of the art performance beating both the Plainifier and the human baseline.

5 Discussion

Our results demonstrate that GPT-based models produce state-of-the-art machine performance for the MWLS task. We have used 3 API based models, all of which outperformed the previous baseline system across two of the three evaluated metrics. Whereas the Plainifier system made use of the token-level probabilities arising from the BERT model, we instead used the GPT-based models in an auto-regressive mode through the ‘chat’ interface of the API. It is also possible to extract token-level probabilities from GPT-based models during generation and there may be potential future gains from examining the probability distributions of potential candidate substitutions. We also only experimented with GPT-based models and did not consider other similar open-source models for text generation.

The results for the Plainifier are low for NDCG and Prec@5, but exceptionally high for the potential metric. Potential is a metric that accepts any of the generated candidates of a system and the high results obtained by the Plainifier are due to the high number of candidates that it generates. In a theoretical setting, one could obtain a perfect result for potential by generating the entire vocabulary (and all possible n-grams thereof) and presenting this for evaluation. Whilst we could generate further candidates for each instance with the GPT-based model at no penalty to the Prec@5 metric, the barrier to accessibility is the linear cost associated with scaling the number of tokens returned by the model. This is a barrier to improved performance on the potential metric, however as this metric does not provide discounting for incorrect substitutions we do not consider Potential to be a reliable indicator of the performance of a deployed Multi-word lexical simplification system. In an application setting, we would likely set the number of returned candidates to be around that used in our experiments (15-30), or lower if the candidates are being used for vocabulary suggestions, where a reader may only be interested in the top 2 or 3 candidates.

Our results show a large increase over the previous state-of-the-art, however there is still much room for improvement across all metrics. The human baseline shows that the MWLS task is also difficult for annotators. The performance that we have reported is above human-level performance for both the NDCG and potential metrics.

In our experiments, we observed that the GPT-model frequently failed to generate complex multi-word expressions (MWEs), particularly trigrams and in domain-specific subsets such as Biomed3. These phrases often involve technical terminology or dense clinical expressions, where maintaining both semantic integrity and readability is significantly more challenging. In such cases, the model either generated overly generic simplifications or failed to produce contextually appropriate alternatives. Additionally, there were instances where the model returned an empty list of simplifications. The trigger for this behaviour is unclear. In this situation the instance is run again until successful.

6 Future Work

A significant area for future work on MWLS is the creation of new datasets. Whereas recent LS datasets have focussed on single-word replacements this is only limited in use to scenarios where a word can be directly replaced by another single-word. MWLS extends this paradigm to N-N replacement of words, but has not been widely adopted by LS dataset developers. Currently the data we have worked on is the only example of MWLS data that the authors are aware of. Further work to extend the volume of available data, contribute additional genres and non-English MWLS would be a positive step forward for the task.

Our experiments have demonstrated that GPT-based models are capable of performing well at the MWLS task for every genre and N-gram size that we evaluated. We used a single prompt to cover all generation and it may be the case that using multiple prompts for specific scenarios (i.e., genre-based prompting, or prompting for specific n-gram sizes) will yield more comprehensive results. We consider this to be a prompt engineering experiment and beyond the scope of the present work.

The MWLS dataset that we have considered only operates on 1-3 grams. It would be interesting to consider a dataset containing replacements that also allow for larger N-gram sizes. We see in the results that performance is typically higher for unigrams and bigrams and that performance dips for trigrams. We may expect that higher n-gram sizes will present an additional challenge for future MWLS systems. We may also consider the integration of linguistic knowledge at both the level of dataset creation (i.e., the selection of semantically coherent complex phrases) and at the level of system design. It may be the case that by providing additional linguistic informed knowledge through techniques such as Graph Neural Networks or Embeddings, we can further improve on MWLS.

7 Conclusion

We have reported a new GPT-based approach to the Multiword Lexical Simplification task. Whereas previous approaches made use of the BERT masked language model, we have made use of the Generative Pre-trained Transformer architecture. Our approach has employed Large Language Models in an auto-regressive format, and made use of prompt engineering and few-shot learning to develop new strategies for the MWLS task. We have shown experiments with several GPT-based models and differing experimental settings. We demonstrated that a GPT-4o-Mini model with 6-shot in-context learning gave state-of-the-art performance for the MWLS1 dataset with NDCG=0.3143, PREC@5=0.1048, beating the previous Bert-based approach by a wide margin on several metrics and consistently across subsets. Whilst human performance was high for Prec@5, our model outperformed the previous state-of-the-art system and represents a move towards human-level performance for this metric. Our findings indicate that GPT-based models are superior to BERT-based models for the MWLS task.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dennis Aumiller and Michael Gertz. 2022. UniHD at TSAR-2022 shared task: Is compute all we need for lexical simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 251–258, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the craft corpus. *BMC bioinformatics*, 13:1–20.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 workshop on integrating artificial intelligence and assistive technology*, pages 7–10. Madison, WI.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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)*, pages 3816–3830, Online. Association for Computational Linguistics.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings 30*, pages 4–15. Springer.

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

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór

Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

John Lee and Chak Yan Yeung. 2018. Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. Large language models for biomedical text simplification: Promising but not there yet. *arXiv preprint arXiv:2408.03871*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016a. Benchmarking lexical simplification systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3074–3080, Portorož, Slovenia. European Language Resources Association (ELRA).

Gustavo Paetzold and Lucia Specia. 2016b. Unsupervised lexical simplification for non-native speakers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Piotr Przybyła and Matthew Shardlow. 2020. Multi-word lexical simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1435–1446, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lsbert: A simple framework for lexical simplification. *arXiv preprint arXiv:2006.14939*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024a. An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 38–46, Torino, Italia. ELRA and ICCL.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024b. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in artificial intelligence*, 5:991242.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. CEFR-based lexical simplification dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative text summarization via ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.