Fine-Grained Error Annotations for Sentence Simplification by Large Language Models

Xuanxin WU[1] Yuki Arase[2]
[1]Graduate School of Information Science and Technology, Osaka University
[2]School of Computing, Institute of Science Tokyo
xuanxin.wu@ist.osaka-u.ac.jp
arase@c.titech.ac.jp

Abstract

Large language models (LLMs) demonstrate strong performance in text simplification, yet current metrics lack the informativeness of more detailed schemes that annotate individual errors. Clearly stating these limitations is essential to understand the simplification quality of LLMs. Building on our previous work [1], which introduced an error-based human annotation framework to assess GPT-4's simplification capabilities, this study expands the scope by including two additional LLMs, Qwen2.5-72B and Llama-3.2-3B, along with more datasets. Our human-annotated corpus comprises fine-grained error analyses for 4,500 complex-simple sentence pairs and Likert-scale ratings for 10,471 pairs, one of the largest scales to date. Results show that LLMs generally generate fewer erroneous simplification outputs than the previous state-of-the-art (SOTA). However, LLMs have their limitations, as seen in larger LLMs struggle with lexical paraphrasing.

1 Introduction

Sentence simplification automatically rewrites sentences to make them easier to read and understand by modifying their wording and structures, without changing their meanings. It helps people with reading difficulties, such as non-native speakers and individuals with aphasia [2, 3].

With the rise of LLMs demonstrating exceptional abilities, some studies [4, 5] evaluated their performance in sentence simplification, including both automatic scoring and conventional human evaluations where annotators assess the levels of fluency, meaning preservation, and simplicity [6, 7, 8, 9]. However, given the general high performance of LLMs, these approaches may be too superficial to capture the subtle yet critical aspects of simplification quality. In contrast, Heineman et al. [10] proposed a detailed human evaluation framework for LLMs, categorizing 21 linguistically based success and failure types. However, their approach appears to be excessively intricate and complex, resulting in low consistency among annotators, thus raising concerns about the reliability of the evaluation. The trade-off between interpretability and reliability underscores the necessity for a more balanced approach. To bridge the gap, we designed an **error-based human evaluation framework** to identify key failures in important aspects of sentence simplification, such as inadvertently increasing complexity or altering the original meaning [1]. This approach aligns closely with human intuition by focusing on outcome-based assessments rather than linguistic details, making the annotation easy without necessitating a background in linguistics.

We apply our human evaluation framework to evaluate the performance of GPT-4[1)] [11], Qwen2.5-72B [12], and Llama-3.2-3B[2)] [14] in English sentence simplification. We believe that these models offer a representative selection across large, medium, and small sizes of LLMs. We use prompt engineering and evaluate models on four representative datasets on sentence simplification: Turk [15], ASSET [16], Newsela [17], and SimPA [18]. We benchmark LLMs against Control-T5 [19], the previous SOTA supervised simplification model. In total, we collect human assessments of 4500 simplifications for error identification and 10,471 simplifications for Likert-scale ratings. To the best of our knowledge, our corpus is the largest to date for fine-grained annotations evaluating simplification quality.

1) We used the 'gpt-4-0613' and accessed it via OpenAI's APIs.
2) We used the 'Qwen2.5-72B-Instruct' and 'Llama-3.2-3B-Instruct'. We runned the two models using Transformers library[13].

Table 1: Definitions and Examples of Errors

Error		Definition	Source	Simplification
Lack of Simplicity	Lexical	More intricate lexical expression(s).	...it shows Harry's bravery...	...it portrays Harry's courage...
	Structural	More difficult grammatical structure.	The other incorporated cities on the Palos Verdes Peninsula include...	Other cities on the Palos Verdes Peninsula include..., which are also incorporated.
Altered Meaning	Lexical	Significant deviation in the meaning due to lexical substitution(s).	The Britannica was primarily a Scottish enterprise.	The Britannica was mainly a Scottish endeavor.
	Structural	Significant deviation in the meaning due to structural changes.	...first famed Colombian trainer Francisco Maturana, and then Julio César Falcioni.	...two famous Colombian trainers, Francisco Maturana and Julio César Falcioni.
Coreference		A named entity critical to understanding the main idea is replaced with a vague description.	Sea slugs dubbed sacoglossans are some of the most...	These are some of the most...
Repetition		Unnecessary duplication of sentence fragments	The report emphasizes the importance...	The report emphasizes the importance, the significance, and the necessity...
Hallucination		Inclusion of incorrect or unrelated information not present in the original sentence.	...Fray is not done, Fray is coming back.	...Fray will return, although the story is not yet finished.

With these annotations, we conduct a large-scale analysis of models. Our key findings are summarized as follows:

- **LLMs generally surpass the previous SOTA in performance**; LLMs tend to generate fewer erroneous simplifications and preserve the original meaning better, while maintaining comparable levels of fluency and simplicity.
- Among the LLMs, GPT-4 and Qwen2.5-72B surpass Llama-3.2-3B, with Qwen2.5-72B generating fewer errors than GPT-4. This implies the **strong potential of medium-sized LLMs in simplification tasks**.
- Larger LLMs have their limitations, as seen in GPT-4 and Qwen2.5-72B' s struggles with **lexical paraphrasing**.

2 Error Annotation Schemes

2.1 Datasets

We used test sets from four representative datasets on English sentence simplification. These datasets have distinctive features due to differences in simplification strategies and domains as summarized below.

- **Turk [15]:** This dataset comprises 359 sentences from English Wikipedia, each paired with eight simplified references written by crowd-workers. It is created primarily focusing on **lexical paraphrasing**.
- **ASSET [16]:** This dataset uses the same 359 source sentences as the Turk dataset. It differs from Turk by aiming at rewriting sentences with more **diverse transformations**, i.e., paraphrasing, deleting phrases, and splitting a sentence, and provides 10 simplified references written by crowd-workers.
- **Newsela [17, 20]:** This dataset originates from a collection of news articles accompanied by simplified versions written by professional editors. The test split contains 1,077 sentence pairs. After careful observation, we found that **deletions of words, phrases, and clauses** predominantly characterize the Newsela dataset.
- **SimPA [18]:** This dataset originated from the public administration domain. It contains 1,100 original sentences with two versions of simplified sentences: (1) lexical simplifications (2) lexical and syntactic simplifications. We selected the second version for its **diverse transformations**.

2.2 Models

For the three LLMs, we reused the prompts from our previous work, which were developed through prompt engineering and proven effective (see Appendix A). Given the similarity between SimPA and ASSET in emphasizing diverse transformations as outlined in their annotation guidelines, we did not include SimPA in the prompt en-

gineering process. Instead, we directly applied the instruction from ASSET, accompanied by 3-shot examples with single references from SimPA itself. We replicated Control-T5 model following the approach used in the original study [19]. Note that we did not evaluate Control-T5 on SimPA since the training dataset is not available.

2.3 Human Annotation Procedure

We conducted two annotation tasks:

- **Error Identification**: Following the error-based human annotation framework from our previous work [1], we analyzed model-generated simplifications to identify key failures in critical aspects of sentence simplification. Table 1 provides definitions and examples of the target errors. In this task, we sampled 300 source sentences from each test set, along with simplification outputs generated by models, resulting in a total of 4500 complex-simple sentence pairs.
- **Likert Scale Rating**: We evaluated fluency, meaning preservation, and simplicity using a 1–3 Likert scale to evaluate overall quality. In this task, we examined on all 10, 471 model-generated simplification outputs.

Our annotators were graduate students and alumni (second-language learners with advanced English proficiency) affiliated with our organization, and native speakers with English teaching experience. To ensure quality control, annotators had to pass a qualification test before participating in the task. In both tasks, seven annotators participated, and each simplification was evaluated by three annotators. To resolve annotator disagreements on error labels, all annotators involved in error identification participated in discussion sessions to collectively review their labels until reaching the consensus.

3 Result Analysis

3.1 Error Identification

This section presents a comparative analysis of erroneous simplification outputs generated by models, and reports additional observations during the annotation.

3.1.1 Characteristic Errors in Models

We quantitatively analyze the frequency of different error types in the simplifications generated by the models.

Table 2: Comparison of error types across models

Error Type	GPT-4	Qwen	Llama	T5
Lack of Sim-L	144 (100)	99 (77)	61 (51)	(4)
Lack of Sim-S	11 (8)	26 (24)	17 (14)	(15)
Altered Meaning-L	94 (74)	59 (55)	149 (110)	(176)
Altered Meaning-S	12 (8)	10 (10)	12 (11)	(15)
Coreference	15 (14)	3 (2)	51 (35)	(104)
Repetition	0 (0)	1 (0)	53 (53)	(7)
Hallucination	9 (7)	5 (5)	62 (52)	(29)
T&A&N	(211)	(173)	(326)	(350)
TOTAL	285	203	405	—

The results are summarized in Table 2. We also report the results after excluding SimPA for fair comparison with Control-T5, that is, only on Turk, ASSET, and Newsela (denoted as 'T&A&N') and with those values indicated in round brackets. The best-performing values (fewer occurrences, better performance) are highlighted in green.

LLMs Outperform Control-T5 **Control-T5** generated more errors overall (350 occurrences) than the LLM group (211 for GPT-4, 173 for Qwen, and 326 for Llama after excluding SimPA). Among the LLMs, **Qwen2.5-72B** produced the fewest errors (203), followed by GPT-4 (285), and Llama-3.2-3B (405). Notably, Qwen performs best in four out of seven error categories, suggesting that while larger LLMs generally perform better, performance may not always scale directly with model size in simplification.

Lexical Paraphrasing is the Biggest Challenge Both **GPT-4** and **Qwen2.5-72B** show similar tendencies, with errors predominantly from `Lack of Simplicity-Lexical` (144 for GPT-4 and 99 for Qwen) and `Altered Meaning-Lexical` (94 for GPT-4 and 59 for Qwen). This reflects their propensity to employ complex lexical expressions or misinterpret meanings through lexical choices. **Control-T5** shows notably high frequencies in `Altered Meaning-Lexical` (176) and `Coreference` (104). This indicates difficulties with preserving original lexical meanings and ensuring referential clarity. Across **all models**, errors in lexical aspect (`Lack of Simplicity-Lexical`, `Altered Meaning-Lexical`, `Coreference`, `Repetition`) surpass the occurrences of errors in structural aspect (`Lack of Simplicity-Structural`, `Altered Meaning-Structural`) as a general tendency.

Llama-3 is Prone to Repetition Error Remarkably, for **Llama-3.2-3B**, while paraphrasing remains a significant issue, errors such as `Repetition` and `Hallucination`, are notably more frequent than in other models. Llama-3.2-3B appears to combine multiple simplifications into a single output, leading to repetitive content. Below is an example:

Source: But landowner Gene Pfeifer refused to give up his 3-acre riverfront property in the middle of the proposed library site.

Llama: Gene Pfeifer didn't want to sell his 3-acre land. Gene Pfeifer refused to sell his land. Gene Pfeifer didn't want to give up his 3-acre property.

3.1.2 Additional Observations

During the error identification annotation process, we observed two nuanced phenomena in LLMs' simplifications that were difficult to fit into specific error categories. `Change of focus` fails to meet the satisfactory criteria, and `Factual Information Not Inferable from the Source Sentence` could be controversial. This section outlines the models where these phenomena were observed and provides examples for each category.

Change of Focus Simplifications that inappropriately alter the original sentence's focus, leading to misleading interpretations. This was only observed in Control-T5 with four reported cases and Llama-3.2-3B with six. In the following example, Llama-3.2-3B redirects attention from the agreement and actions of other judges to the federal court's decision itself.

Source: Other judges agreed with the federal court's decision and started marrying same-sex couples in the morning.

Llama: The federal court ruled that same-sex couples could get married.

Factual Information Not Inferable from the Source Sentence We found cases where information not explicitly present in the source sentence was added to the simplifications. This was observed in all models, with four reported cases in GPT-4, 12 in Qwen2.5-72B, five in Llama-3.2-3B, and 12 in Control-T5. These additions were generally factual and, although not inferable from the source sentence, were verified to be factual using online sources. This type of information can be controversial, as it does not strictly adhere to the input. However, it may facilitate the reader's understanding of the source sentence. For example, in the case below, "Lincoln's assassination" cannot be inferred directly from the source sentence. However, Qwen2.5-72B included this detail, likely drawing on its internal knowledge by linking the provided date and named entities.

Table 3: Average Ratings

Dimension	GPT-4	Qwen	Llama	T5
Fluency	2.99	2.99	3.00	2.98
Meaning	2.80	2.80	2.22	1.66
Simplicity	2.84	2.93	2.82	2.94

Source: For example, there's a letter of sympathy from Queen Victoria to Mary Todd Lincoln on April 29, 1865, calling his assassination "so terrible a calamity".

Qwen: Queen Victoria wrote a letter of sympathy to Mary Todd Lincoln about Lincoln's assassination.

3.2 Likert Scale Rating

In this section, we compared model performances by averaging annotators' ratings. Results are summarized in Table 3. For **fluency**, all models demonstrate high fluency levels, indicated by the average ratings approach three. This suggests that these models generated grammatically correct simplifications without significant differences in fluency. For **meaning preservation**, GPT-4 (2.80) and Qwen2.5-72B (2.80) outperform Llama-3.2-3B (2.22) and Control-T5 (1.66) . Conversely, for **simplicity**, GPT-4 (2.84) and Qwen2.5-72B (2.93)'s ratings are comparable with those of Llama-3.2-3B (2.82) and Control-T5 (2.94). This contrast suggests that Llama-3.2-3B and Control-T5 may be comparably good at generating simpler outputs but at the cost of losing the original meaning.

4 Discussions

Our human error annotation revealed that LLMs struggle to handle lexical paraphrasing while their simplification quality surpasses that of the previous SOTA model. Our investigation opens up multiple directions for future research. The corpus we created can support studies exploring strategies like instruction tuning to automate error annotations. Furthermore, future studies could investigate how to mitigate lexical paraphrasing issues.

Acknoledgement

This work was supported by JSPS KAKENHI Grant Number JP21H03564.

References

[1] Xuanxin Wu and Yuki Arase. An in-depth evaluation of gpt-4 in sentence simplification with error-based human assessment, 2024.

[2] Gustavo Henrique Paetzold. **Lexical Simplification for Non-Native English Speakers**. PhD thesis, University of Sheffield, September 2016. Publisher: University of Sheffield.

[3] John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. Simplifying text for language-impaired readers. In **Ninth Conference of the European Chapter of the Association for Computational Linguistics**, pp. 269–270, Bergen, Norway, June 1999. Association for Computational Linguistics.

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

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

[6] Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. Complexity-weighted loss and diverse reranking for sentence simplification. 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)**, pp. 3137–3147, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. Neural CRF model for sentence alignment in text simplification. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7943–7960, Online, July 2020. Association for Computational Linguistics.

[8] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. The (un)suitability of automatic evaluation metrics for text simplification. **Computational Linguistics**, Vol. 47, No. 4, pp. 861–889, December 2021.

[9] Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. Controllable text simplification with explicit paraphrasing. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3536–3553, Online, June 2021. Association for Computational Linguistics.

[10] David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3466–3495, Singapore, December 2023. Association for Computational Linguistics.

[11] OpenAI. Gpt-4 technical report, 2023.

[12] Jinze Bai et al. Qwen technical report, 2023.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. **CoRR**, Vol. abs/1910.03771, , 2019.

[14] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[15] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 401–415, 2016.

[16] Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4668–4679, Online, July 2020. Association for Computational Linguistics.

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

[18] Carolina Scarton, Gustavo Paetzold, and Lucia Specia. SimPA: A sentence-level simplification corpus for the public administration domain. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[19] Kim Cheng Sheang and Horacio Saggion. Controllable sentence simplification with a unified text-to-text transfer transformer. In **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 341–352, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.

[20] Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 584–594, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

A Best Prompts in GPT-4's prompt engineering

Figure 1 illustrates the best prompts that achieved the highest SARI scores during GPT-4's prompt engineering in our previous work [1]. Each prompt comprises: instructions, examples of original to simplification(s) transformation, and a source sentence.

You are required to simplify the original sentence by using simpler concepts, words, or phrases. Please keep the meaning the same. Only provide one result.

Original sentence: San Francisco Bay is located in the U.S. state of California, surrounded by a contiguous region known as the San Francisco Bay Area, dominated by the large cities San Francisco, Oakland and San Jose.
Simplified sentence: San Francisco Bay is located in the U.S. state of California, surrounded by a contiguous region known as the San Francisco Bay Area, influenced by the large cities, San Francisco, Oakland and San Jose.

Original sentence: The book chronicles events which take place in the fictional space colony of Windhaven.
Simplified sentence: The book chronicles events which take place in the space colony of Windhaven.

Original sentence: Some academic journals do refer to Wikipedia articles, but are not elevating it to the same level as traditional references.
Simplified sentence: Some academic journals do refer to Wikipedia articles, but are not using it to the same level as common references.

Original sentence: {input}

(a) Turk style + Few-shot + Single ref

You are required to simplify the original sentence by applying different transformations. Please keep the meaning the same. Only provide one result.

Original sentence: Rollins retired in 1962 and opted to become a coach.
Simplified sentence: Rollins retired in 1962. He then chose to become a coach.

Original sentence: Tourism is concentrated in the mountains, particularly around the towns of Davos / Arosa, Laax and St. Moritz / Pontresina.
Simplified sentence: Tourism takes place in the mountains around the towns of Davos / Arosa, Laax and St. Moritz / Pontresina.

Original sentence: First Fleet is the name given to the 11 ships which sailed from Great Britain on 13 May 1787 with about 1,487 people to establish the first European colony in New South Wales.
Simplified sentence: 11 ships sailed from Great Britain on 13 May 1787 carrying about 1,487 people. These ships aimed to establish the first European colony in New South Wales. These 11 ships were named First Fleet.

Original sentence: {input}

(b) ASSET style + Few-shot + Single ref

You are required to simplify the original sentence. You can delete information that makes the sentence difficult to understand. Only provide one result.

Original sentence: Becker was trailing an underwater camera that will help him and the other scientists figure out how to wrench out an extensive network of oyster racks held up by some 4,700 wooden posts sunk into the Estero 's sandy bottom.
Simplified sentences:
The camera will help scientists figure out how to remove the oyster racks.
The posts are sunk into the Estero 's sandy bottom.
The racks are held up by about 4,700 wooden posts.

Original sentence: He also announced a 15 percent increase in the minimum wage, effective next month, and an increase in scholarships for high school and college students.
Simplified sentences:
He said the minimum wage for workers will go up.
President Maduro said he would fix some things.
The minimum wage is the least amount of money someone can get paid to work.

Original sentence: The monitoring site, more than 5,000 feet above sea level on a pine-studded overlook above the lowest layer of the atmosphere, gives Faloona access to undisturbed air from across the Pacific before it is fouled by U.S. pollution sources.
Simplified sentences:
The spot is more than 5,000 feet above sea level.
His measuring instruments are located on Chews Ridge in the Santa Lucia Mountains.
There he can test the air blowing in from across the Pacific.

Original sentence: {input}

(c) Newsela style + Few-shot + Multi refs

Figure 1: Best prompts in GPT-4's prompt engineering.