UM_FHS at the CLEF 2025 SimpleText Track

Comparing No-Context and Fine-Tune Approaches for GPT-4.1 Models in Sentence and Document-Level Text Simplification

Primoz Kocbek[1,2], Gregor Stiglic[1,3]

[1]*Faculty of Health Sciences, University of Maribor, Slovenia*
[2]Faculty of Medicine, University of Ljubljana, Slovenia
[2]University of Edinburgh, Usher Institute, Edinburgh, UK

Madrid, Spain

September 10th, 2025

Background and Motivation

- continuation of our previous work from TREC 2024 PLABA track 1 (end-to-end biomedical abstracts adaptations)
- using NIH guidelines for written health materials (average literacy level <K8 - students 13–14 years old
- European Health Literacy Survey (HLS-EU) - at least 1 in 10 (12%) respondents showed insufficient health literacy and almost 1 in 2 (47%) had limited (insufficient or problematic) health literacy

Task and training data

- Task 1: Text Simplification: Simplify scientific text
  - Task 1.1 - Sentence-level Scientific Text Simplification
  - Task 1.2 - Document-level Scientific Text Simplification
- Training data
  - Cochrane-auto corpus, derived from biomedical literature abstracts and lay summaries from Cochrane systematic reviews
  - Data realigned at the paragraph, sentence, and document levels

Models and methods used

- Focused on gpt-4.1 family of models (version 2025-04-14)
  - gpt-4.1
  - gpt-4.1-mini
  - gpt-4.1-nano

- Methods
  - Prompt template on all models

  - Fine-tuned gpt-4.1-mini and gpt-4.1-nano (not on gpt-4.1 due to cost)

Evaluation Metrics

- SARI, BLEU, FKGL, Compression ratio, Levenshtein similarity, Lexical complexity score,…
- ...in the biomedical domain human assessment is still the gold standard

Prompt template (Task 1.1)

- decided that we used prompts at the document level
  - The sentences were supplied as a list of sentences to the LLM
  - Special emphasis on the number of sentences or if a sentence is omitted
  - Note: if only sentences were provided, the text adaptations might be worse, since the context in the surrounding context would not be present

Prompt template (Task 1.1)

- used OpenAI gpt-4.1 prompting guide
- started with general prompt structure
- optimized with ChatGPT

```
# Role and Objective
# Instructions
## Sub-categories for more detailed instructions
# Reasoning Steps
# Output Format
# Examples
## Example 1
# Context
# Final instructions and prompt to think step by step
```

https://cookbook.openai.com/examples/gpt4-1_prompting_guide

Prompt template (Task 1.1)

- User prompt
- Input: list of sentences (strings)
- Output: list of sentences (strings), same length
- Define the rules according to NIH guidelines

```
TASK – Plain-language sentence adaptation (based on NIH guidelines for written health materials)
```

```
INPUT =['SENTENCE_1', 'SENTENCE_2', . . . , 'SENTENCE_N']

OUTPUT FORMAT → ['ADAPTATION_1', 'ADAPTATION_2', . . . , 'ADAPTATION_N']
```

```
ESSENTIAL RULES
• Audience Write for readers at about a US 8th-grade level (K8 or smart 13-14 year old student).
• Workflow (1) Carry over each sentence exactly as written, (2) decide if it should be adapted or omitted, (3) review the whole list for coherence while keeping every '' placeholder.
• Splitting If a sentence contains more than one idea, split it into shorter sentences inside the same pair of single quotes; never merge content from different source items.
• Omission If a sentence is irrelevant to lay readers (for example, detailed measurement methods), output the empty string '' for that element.
• Jargon Replace professional terms with common words. If no plain synonym exists, keep the term once and add a brief parenthetical gloss.
• Statistics Remove p-values, confidence intervals, and similar numbers unless they are essential for understanding.
• Voice Use active voice when possible.
• Pronouns Resolve ambiguous pronouns or other references.
• Subheadings Remove IMRAD labels, such as 'Background:', 'Introduction:', 'METHODS:', 'Results:', 'Discussion:' or integrate them into a full sentence.
• Output Return one **Python list with N elements**—exactly the same number of elements as the input list—and nothing else. Double check this.
```

Prompt template (Task 1.1)

- Give detailed instructions
- Provide an example (jargon can be Substituted, Explained, Generalized, Exemplified, Omitted)
- Some instructions are stated multiple times
- Also included NIH adapted guidelines (from PLABA)

```
INSTRUCTIONS
1 Produce one list with N elements in the original order.
2 For each element follow this three-step process:
   • First: Carry the sentence over unchanged.  SENTENCE 1 → ADAPTATION 1,
   ..., SENTENCE_N → ADAPTATION_N
   • Second - decide and modify ADAPTATIONS as needed:
     – If it is already plain → leave it as is.
     – If it is irrelevant → replace with ''.
     – Otherwise → simplify it (you may split it).
   • Third: After processing all items, review the entire list for flow and
   pronoun clarity. Also keep every '' element in place.
3 Double-check (again) that the output list contains N elements and that no
facts have been added or lost. If the number DO NOT match return to point 1
and re-do all the steps. Repeat until the number MATCH.
Return **only** the final list.
```

```
QUICK EXAMPLES
• Simplify 'Myocardial infarction is a leading cause of mortality worldwide.
' → 'A heart attack is a major cause of death worldwide.'
• Carry over 'Metabolism is essential for life.' → 'Metabolism is essential
for life.'
• Omit 'Blood pressure was measured with a sphygmomanometer.' → ''
• Split 'Cardiovascular disease is the leading cause of mortality, and it is
influenced by genetics as well as lifestyle.' → 'Heart disease is the leading
cause of death. Genetics and lifestyle also influence it.'
```

Prompt template (Task 1.2)

- Similar as for Task 1.1
- No explicit instructions, the output is just a string

```
TASK - Plain-language sentence adaptation (based on NIH guidelines for
written health materials)

ESSENTIAL RULES
• Audience Write for readers at about a US 8th-grade level (K8 or smart
13-14 year old student).
• Splitting If a sentence contains more than one idea, split it into shorter
sentences inside the same pair of single quotes; never merge content from
different source items.
• Omission If a sentence is irrelevant to lay readers (for example, detailed
measurement methods), output the empty string ''.
• Jargon Replace professional terms with common words. If no plain synonym
exists, keep the term once and add a brief parenthetical gloss.
• Statistics Remove p-values, confidence intervals, and similar numbers
unless they are essential for understanding.
• Voice Use active voice when possible.
• Pronouns Resolve ambiguous pronouns or other references.
• Subheadings Remove IMRAD labels, such as 'Background:', 'Introduction:',
'METHODS:', 'Results:', 'Discussion:' or integrate them into a full sentence.
• Output Return only the final simplified sentence as string.

QUICK EXAMPLES
• Simplify 'Myocardial infarction is a leading cause of mortality worldwide.' →
'A heart attack is a major cause of death worldwide.'
• Carry over 'Metabolism is essential for life.' → 'Metabolism is essential
for life.'
• Omit 'Blood pressure was measured with a sphygmomanometer.' → ''
• Split 'Cardiovascular disease is the leading cause of mortality, and it
is influenced by genetics as well as lifestyle.' → 'Heart disease is the
leading cause of death. Genetics and lifestyle also influence it.'
```

Fine-tuning (Task 1.1 and 1.2)

Supervised fine-tuning (SFT) lets you train an OpenAI model with examples for your specific use case. The result is a customized model that more reliably produces your desired style and content.

https://platform.openai.com/docs/guides/supervised-fine-tuning

Fine-tuning (cost estimate)

- $25 for gpt-4.1, $5 for gpt-4.1-mini, $1.5 for gpt-4.1-nano per million tokens (time of writing)
- training data: 4.8 million tokens sentence level, 2.1 million document level
- gpt-4.1 it would be ~$172 (**just a projection**), gpt-4.1-mini ~$34.5 and gpt-4.1-nano $10.3

Results

- First test dataset - 37 Cochrane abstracts aligned with their plain language summaries via Cochrane-auto, comprising of 587 sentence pairs

- Second test dataset - 217 unaligned abstract-summary pairs (only used for task 1.2)

Results

Model	SARI	BLEU	FKGL	Compression ratio
Source	12.03	20.53	13.54	1.00
Reference	100.00	100.00	11.73	0.56
gpt-4.1-nano	29.47	18.46	11.10	0.86
gpt-4.1-nano-ft	/	/	/	/
gpt-4.1-mini	43.34	13.93	7.46	0.78
gpt-4.1-mini-ft	42.83	20.85	12.29	0.71
gpt-4.1	38.84	14.04	8.51	0.79

Comments (first test set):

- best-performing model was gpt-4.1-mini (SARI 43.34)
- - FKGL, was below grade 8, aligning well with NIH guidelines for plain language adaptations, reference FKGL was above

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89
Reference	37	100	100	11.73	0.56	0.67	0.50	0.0	0.16	0.60	8.71
UM-FHS gpt-4.1-mini	37	43.34	13.93	7.46	0.78	1.58	0.63	0.00	0.28	0.50	8.50
UM-FHS gpt-4.1-mini-	37	42.83	20.85	12.29	0.71	0.86	0.62	0.00	0.15	0.46	8.67
DSGT plan_guided_lla	37	42.33	10.43	7.77	0.48	0.97	0.47	0.00	0.18	0.70	8.52
UvA o-bartsent-cochr	37	42.31	25.72	12.08	0.41	0.51	0.55	0.00	0.01	0.62	8.72
SINAI PRMZSTASK11V1	37	41.82	6.50	11.41	1.37	1.56	0.53	0.00	0.59	0.30	8.33
THM p2–gpt-4.1-nano	37	41.32	10.49	14.90	1.27	1.16	0.63	0.00	0.45	0.26	8.62
UvA bartsent-cochran	37	41.28	17.67	11.20	0.35	0.49	0.48	0.00	0.01	0.67	8.76
Scalar gpt_md_2_1	37	40.95	14.07	18.79	0.62	0.47	0.53	0.00	0.22	0.60	8.68
UBOnlp gpt4o	37	40.74	7.53	7.39	0.46	0.80	0.41	0.00	0.23	0.73	8.31
THM p1–gpt-4.1-nano	37	40.42	11.02	14.66	1.23	1.13	0.65	0.00	0.42	0.24	8.61
PICT S3Pipeline	37	40.15	12.96	7.61	0.71	1.53	0.62	0.00	0.21	0.49	8.84
Fujitsu llm_t5_rule	37	39.04	6.70	6.79	0.31	0.71	0.42	0.00	0.08	0.76	8.85
UM-FHS gpt-4.1	37	38.84	14.04	8.51	0.79	1.26	0.68	0.30	0.22	0.41	8.49
UvA llama31	37	38.76	2.83	8.30	0.93	1.58	0.46	0.00	0.60	0.66	8.34
DUTH Task11_flan-t5-	37	38.73	18.84	11.95	0.61	0.78	0.66	0.00	0.10	0.50	8.96
Fujitsu t5efficient	37	38.60	4.28	5.58	1.79	3.63	0.43	0.00	0.77	0.29	10.31
Fujitsu llm_gpt3.5-t	37	38.53	6.30	5.18	0.36	0.99	0.45	0.00	0.11	0.74	8.89
Fujitsu llm_45_judge	37	38.41	5.45	5.26	0.32	0.89	0.42	0.00	0.09	0.77	8.87
Fujitsu dummy60	37	38.37	14.50	1.19	0.37	2.74	0.52	0.00	0.08	0.67	8.74
SINAI PRMZSTASK11V2	37	37.84	5.93	12.97	1.64	1.63	0.56	0.00	0.59	0.17	8.47
THM pni1–gpt-4.1-na	37	37.60	8.24	15.21	1.84	1.63	0.56	0.00	0.57	0.12	8.61
UvA bartdoc-ca	37	37.25	19.54	11.97	0.51	0.61	0.62	0.00	0.02	0.52	8.77
EngKh biomedical_lla	37	36.68	11.47	10.62	1.14	1.51	0.65	0.00	0.37	0.28	8.69
UvA llama31	37	36.45	1.22	13.04	1.07	1.31	0.41	0.00	0.66	0.70	8.61
AIIRLab mistral	37	36.08	18.41	12.78	0.94	1.06	0.76	0.00	0.19	0.28	8.81
MTest bartfinetuned	37	34.98	26.52	11.94	0.74	0.98	0.83	0.00	0.01	0.30	8.78

https://www.dei.unipd.it/~faggioli/temp/clef2025/paper_344.pdf

Results

Model	SARI	BLEU	FKGL	Compression ratio
Source	12.03	20.53	13.54	1.00
Reference	100.00	100.00	11.73	0.56
gpt-4.1-nano	37.01	14.74	9.05	0.69
gpt-4.1-nano-ft	43.61	16.00	10.63	0.50
gpt-4.1-mini	43.53	14.11	7.48	0.72
gpt-4.1-mini-ft	42.82	22.94	11.93	0.60
gpt-4.1	43.83	18.12	8.80	0.67

Comments (first test set):

- gpt-4.1 achieved the highest SARI score (43.83), closely followed by gpt-4.1-nano-ft (43.61).
- gpt-4.1 better adhered to NIH guidelines with an FKGL of 8.80, compared to 10.63 for gpt-4.1-nano-ft.

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89
Reference	37	100	100	11.73	0.56	0.67	0.50	0.0	0.16	0.60	8.71
LIA sumguid-all-w500	37	44.55	12.18	9.71	0.84	1.26	0.50	0.00	0.35	0.54	8.56
SINAI PRMZSTASK12V1	37	43.93	10.81	10.45	0.86	1.07	0.55	0.00	0.39	0.49	8.33
UM-FHS gpt-4.1	37	43.83	18.12	8.80	0.67	1.10	0.58	0.14	0.21	0.53	8.44
UM-FHS gpt-4.1-nano-	37	43.61	16.00	10.63	0.50	0.69	0.45	0.00	0.16	0.65	8.55
LIA sumguid-lang-w50	37	43.61	10.55	10.50	0.83	1.18	0.47	0.00	0.37	0.57	8.52
UM-FHS gpt-4.1-mini	37	43.53	14.11	7.48	0.72	1.49	0.62	0.00	0.25	0.52	8.52
ASM MistralMaxFRE	37	43.35	12.32	11.63	0.73	0.92	0.53	0.00	0.27	0.56	8.74
ASM MistralV0	37	43.31	12.41	11.65	0.73	0.92	0.53	0.00	0.27	0.55	8.74
ASM MistralMinFKGL	37	43.24	12.27	11.63	0.73	0.93	0.53	0.00	0.27	0.56	8.75
ASM MistralV7	37	42.95	11.34	12.53	0.78	0.94	0.51	0.00	0.30	0.55	8.80
ASM MistralV7CleanLi	37	42.93	11.38	13.77	0.78	0.84	0.51	0.00	0.29	0.56	8.80
UM-FHS gpt-4.1-mini-	37	42.82	22.94	11.93	0.60	0.76	0.60	0.03	0.10	0.52	8.73
AIIRLab Mistral_7b_b	37	42.40	12.98	8.82	0.58	0.94	0.52	0.00	0.21	0.61	8.48
UvA baseline-cochran	37	42.10	24.27	11.71	0.57	0.71	0.61	0.00	0.06	0.49	8.74
LIA sumguid-styl-w50	37	41.98	10.38	10.09	0.63	1.00	0.46	0.00	0.27	0.66	8.65
UBOnlp gpt4o	37	41.56	5.45	7.22	1.14	2.08	0.50	0.00	0.58	0.43	8.25
LIA sumguid-styl-w50	37	41.11	8.73	6.35	0.61	1.30	0.42	0.00	0.33	0.68	8.44
AIIRLab llama_3.1-8b	37	41.07	8.61	9.22	0.46	0.70	0.43	0.00	0.20	0.72	8.44
LIA testLlama33	37	40.79	8.42	10.74	0.46	0.65	0.42	0.00	0.18	0.73	8.64
DSGT llama_summary_s	37	40.32	7.63	9.56	0.59	0.86	0.42	0.00	0.31	0.70	8.49
PICT S3Pipeline	37	40.29	13.43	7.77	0.74	1.55	0.63	0.00	0.21	0.47	8.77
AIIRLab llama-8b	37	39.14	5.62	8.88	0.34	0.62	0.35	0.00	0.15	0.80	8.43
AIIRLab llama3.2-3b	37	39.14	5.62	8.88	0.34	0.62	0.35	0.00	0.15	0.80	8.43
DUTH task12_led-larg	37	39.11	9.83	12.41	0.37	0.47	0.45	0.00	0.06	0.70	8.80
SINAI PRMZSTASK12V2	37	38.50	10.30	11.55	1.09	1.16	0.63	0.00	0.43	0.29	8.44

https://www.dei.unipd.it/~faggioli/temp/clef2025/paper_344.pdf

Results

Model	SARI	BLEU	FKGL	Compression ratio
Source	7.84	10.55	13.29	1.00
Reference	100.00	100.00	11.28	0.72
gpt-4.1-nano	28.89	10.35	9.90	0.83
gpt-4.1-nano-ft	/	/	/	/
gpt-4.1-mini	42.13	9.52	7.56	0.74
gpt-4. 1-mini-ft	39.16	11.95	12.23	0.67
gpt-4.1	37.93	9.46	8.82	0.76

Comments (second test set):

- gpt-4.1-mini best performance (SARI 42.13 and a FKGL of 7.56)
- -gpt-4.1-nano-ft generated no usable output.

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05
Reference	217	100	100	11.28	0.72	0.97	0.40	0.00	0.29	0.63	8.65
DSGT plan_guided_lla	217	42.98	6.33	7.82	0.48	0.99	0.46	0.00	0.18	0.71	8.50
UBOnlp gpt4o	217	42.20	4.05	7.49	0.38	0.68	0.37	0.00	0.18	0.78	8.37
UM-FHS gpt-4.1-mini	217	42.13	9.52	7.56	0.74	1.52	0.61	0.00	0.26	0.53	8.54
SINAI PRMZSTASK11V1	217	41.25	4.59	12.39	1.44	1.56	0.51	0.00	0.61	0.30	8.44
UvA llama31	217	40.92	2.62	8.63	1.00	1.64	0.45	0.00	0.62	0.64	8.35
THM p2–gpt-4.1-nano	217	39.57	6.50	15.40	1.32	1.20	0.60	0.00	0.47	0.27	8.68
UM-FHS gpt-4.1-mini-	217	39.16	11.95	12.23	0.67	0.82	0.60	0.00	0.14	0.50	8.76
PICT S3Pipeline	217	39.11	8.30	6.52	0.69	1.65	0.60	0.00	0.21	0.52	8.85
Scalar gpt_md_2_1	217	38.96	8.25	19.45	0.62	0.43	0.52	0.00	0.23	0.60	8.77
Fujitsu llm_gpt3.5-t	217	38.84	3.05	5.04	0.35	1.02	0.44	0.00	0.11	0.75	8.96
UvA bartsent-cochran	217	38.71	6.01	11.34	0.31	0.46	0.45	0.00	0.00	0.72	8.81
Fujitsu llm_t5_rule	217	38.55	2.75	6.60	0.31	0.77	0.42	0.00	0.08	0.77	8.95
Fujitsu llm_45_judge	217	38.54	2.34	5.19	0.31	0.93	0.41	0.00	0.09	0.78	8.95
UvA o-bartsent-cochr	217	38.53	8.57	11.99	0.37	0.49	0.51	0.00	0.01	0.67	8.78
UvA llama31	217	38.50	1.13	13.66	1.09	1.23	0.40	0.00	0.66	0.71	8.65
Fujitsu llm_45	217	38.49	2.06	5.32	0.31	1.00	0.40	0.00	0.09	0.79	8.90
THM p1–gpt-4.1-nano	217	38.24	6.59	15.03	1.28	1.18	0.63	0.00	0.45	0.25	8.69
Fujitsu llm_45fewSho	217	38.20	1.87	3.51	0.28	0.88	0.37	0.00	0.12	0.81	8.82
UM-FHS gpt-4.1	217	37.93	9.46	8.82	0.76	1.22	0.64	0.23	0.22	0.46	8.54
UvA bartdoc-ca	217	37.14	7.23	11.43	0.39	0.49	0.52	0.00	0.01	0.63	8.85
SINAI PRMZSTASK11V2	217	35.95	4.03	14.00	1.76	1.64	0.54	0.00	0.61	0.15	8.56
DUTH Task11_flan-t5-	217	35.35	10.07	11.21	0.60	0.80	0.65	0.00	0.09	0.51	9.00
THM pni1–gpt-4.1-na	217	35.26	5.23	15.49	1.94	1.72	0.54	0.00	0.59	0.12	8.68
AIIRLab mistral	217	33.95	10.30	13.26	0.93	1.04	0.72	0.00	0.21	0.32	8.86
RECAIDS T5	217	33.89	0.03	3.72	0.37	0.98	0.31	0.00	0.23	0.89	8.87
EngKh biomedical_lla	217	33.16	7.30	10.76	1.18	1.53	0.65	0.00	0.37	0.25	8.75

https://www.dei.unipd.it/~faggioli/temp/clef2025/paper_344.pdf

Discussion and future work

- gpt-4.1-mini outperformed gpt-4.1 in this task
- Focused on OpenAI models, for private healthcare data a locally deployed LLM should be considered/used;
- FT did not show improvement, except for a specific case
- (Future work) Expansion to other domains might be feasible, since the best performing approach just uses in-prompt context
- (Future work) LLM-as-a-judge with CoT, such as G-Eval, that can evaluate an output on any criteria, might be an addition for human evaluation

Thank you for your attention!

primoz.kocbek@um.si