

Motivation

- Scientific text is difficult for non-experts due to domain-specific terminology
- RQ: Is SARI still an adequate evaluation measure for text simplification?



Task & Dataset

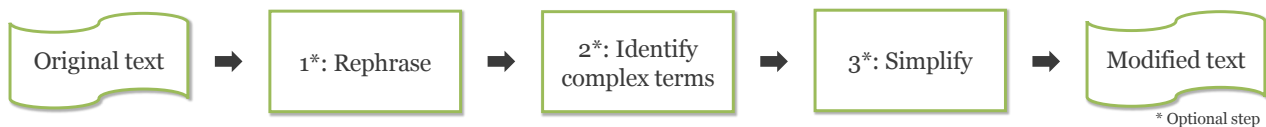
- SimpleText@CLEF 2025 Task 1.1**
- 9,160 biomedical short texts, mostly single sentences
- Goal: Make scientific text more accessible

Approach

- Simplification for non-experts → knowledge of **language**, not of **biomedical** domain
- Low cost/small LLMs
- Find complex terms, simplify them, keep rest
- Complex term identification through keyphrase extraction + corpus-based tf-idf
- Prompts instructing LLMs to only explain/simplify complex terms

For every nine people treated with **haloperidol** instead of **olanzapine**, one fewer person would experience a...

For every nine people treated with **haloperidol** (a medication for mental health conditions) instead of **olanzapine** (another mental health medication), one fewer person would experience a...



Model	P	SARI						
		orig	↑auto	↑FRE	↑CL↓	↑T↓	↑B↑	
baseline	-	7.844	12.033	31.523	13.752	2.104	-	
gpt-4.1-nano	C	32.445	33.944	-6.139	19.805	3.27	0.93	
gpt-4.1-nano	C	30.175	32.586	-14.528	20.887	3.015	0.927	
gpt-4.1-nano	C	28.917	31.199	-3.714	19.504	2.923	0.934	
gpt-4.1-nano	R	29.256	32.598	25.914	14.977	2.186	0.953	
gpt-4.1-nano	R	27.526	31.160	33.193	13.77 [†]	2.096[†]	0.843	
gpt-4.1-nano	P1	38.238	40.416	34.498	13.308	2.726	0.928	
gpt-4.1-nano	P1	25.858	30.001	35.876	12.578	2.811	0.836	
gpt-4.1-nano	C+P1 ^Δ	25.234	29.080	36.001	12.585	2.815	0.948	
gpt-4.1-nano	P2	39.572	41.315	31.238	13.829 [†]	2.816	0.923	
gpt-4.1-nano	P2	25.062	27.811	35.909	12.08	2.752	0.943	
gpt-4.1-nano	P2	29.263	31.174	35.759	12.674	3.003	0.829	
gpt-4.1-nano	C+P2 ^Δ	27.561	30.024	35.829	12.675	3.014	0.94	
gpt-4.1-nano	PNI1	35.262	37.262	34.327	13.201	2.786	0.829	
gpt-4.1-nano	PN1	32.267	34.474	43.769	11.475	2.674	0.842	
gpt-4.1-nano	PN2	20.521	24.045	36.925	11.966	2.479	0.953	

SARI compares n-gram additions, deletions, kept components against reference text

P1: You are a text simplification system with good global and language knowledge but no expertise in specific domains. You will be given 10 texts, TREAT THEM SEPERATLY and also return 10 simplified texts. **Complex domain specific or scientific terminology is marked by square brackets** (e.g., [primary acoustic modeling]). For each text, replace or explain the terms in square brackets in order to make it more understandable for non-experts and simplify for non-expert readers. Make sure to exactly return an enumerated list of 10 simplified texts and that all terms in square brackets are simplified or explained! **Do not simplify the structure of the texts or complex language independent of the difficult scientific descriptions. RETURN ONLY THE 10 TEXTS.**

TEXTS:

P2: You are given 10 texts, TREAT THEM SEPERATELY. **Complex technical or scientific terms are indicated by square brackets** (e.g., [convolutional neural network]). For each text, replace or explain the terms in square brackets to make it understandable to non-experts and to simplify for non-expert readers. Make sure that you return exactly one enumerated list of 10 simplified texts and that all terms in square brackets are simplified or explained!

Setting	P	Number of identified complex terms				Price per 1M input token
		original	auto	original	auto	
gpt-4.1-nano	P1	37.484	39.896	38.238	40.416	\$0.1
gpt-4.1-nano	P2	39.135	40.647	39.572	41.315	
gpt-4.1-mini	P1	26.623	30.271	26.288	30.121	\$0.4
gpt-4.1-mini	P2	29.388	32.223	29.336	31.475	
gpt-5.4-nano	P1	25.803	29.608	26.806	29.852	\$0.2
gpt-5.4-nano	P2	27.138	30.323	27.731	31.025	
gpt-5.4-mini	P1	22.338	26.362	21.957	25.948	\$0.75
gpt-5.4-mini	P2	33.107	35.312	34.010	36.480	

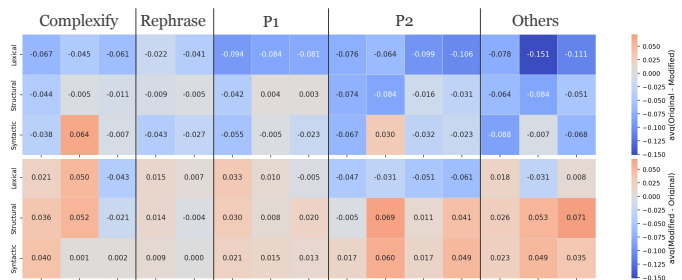
gpt-5.4-nano:

For every nine people treated with ****haloperidol**** (an antipsychotic medicine) instead of ****olanzapine**** (another antipsychotic medicine), one fewer person would experience a ...

gpt-5.4-mini:

For every nine people treated with ****haloperidol**** instead of **olanzapine**, one fewer person would have a ...

Lower SARI != worse simplifications



Results

- LLMs were prompted to only modify complex components but changed syntax, structure & lexical features of text
- P2 > P1
- Better LLMs do not always produce better results
- Better LLMs might rewrite more → lower SARI
- Lower SARI != worse simplifications

Takeaways

- Be mindful what you measure
- SARI rewards conservative lexical edits