Lexical Simplification for Scientific Texts: Revisiting SARI in the Era of Modern LLMs

Nico Hofmann, Julian Dauenhauer, Nils Ole Dietzler,
Idehen Daniel Idahor, Christin Katharina Kreutz
2026-09-22 – Jena, Germany

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

HERDER-INSTITUT
für historische Ostmitteleuropaforschung
INSTITUT DER LEIBNIZ-GEMEINSCHAFT

CLEF 2026
JENA

TURN IT SIMPLE

Motivation

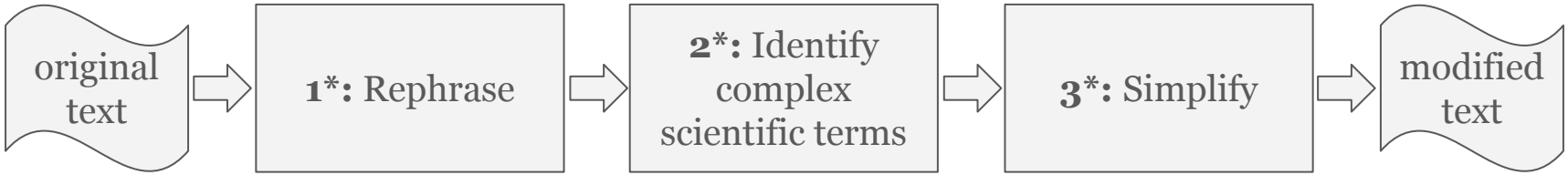
- Scientific text difficult for non-experts due to domain jargon
- SimpleText@CLEF ‘25 Task 1.1: Make biomedical scientific text more accessible

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...

Idea

- Goal: Simplification for non-experts
  → knowledge of **language**, not of **biomedical** domain
- Low cost/small LLMs
- Find complex terms, simplify them, keep rest
- Prompts instructing LLMs to only explain/simplify complex terms

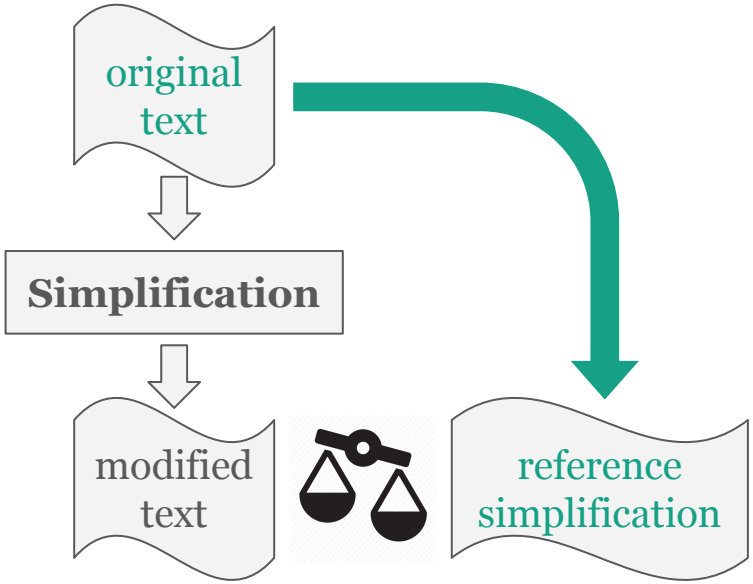


* steps are optional

Submit + Results

- Submitted runs with different prompts, used cheap LLMs
- SimpleText ’25 Task 1.1 mainly evaluated with SARI

SARI considers added, deleted, kept n-grams of produced against reference simplification



Model	P	f10	f1	orig	auto
baseline	-	-	-	7.844	12.033
gpt-4.1-nano	C	8	1	32.445	33.944
gemini-2.5	C	18	7	30.175	32.586
gemini-2.0	C	10	1	28.917	31.199
gemini-2.5	R	1	0	29.256	32.598
gemini-2.0	R	47	1	27.526	31.160
gpt-4.1-nano	P1	67	22	38.238	40.416
gemini-2.0	P1	42	22	25.858	30.001
gemini-2.0	C+P1$^{\triangle}$	44	22	25.234	29.080
gpt-4.1-nano	P2	75	21	**39.572**	**41.315**
gemini-2.5	P2	4275	1550	25.062	27.811
gemini-2.0	P2	87	132	29.263	31.174
gemini-2.0	C+P2$^{\triangle}$	91	132	27.561	30.024
gpt-4.1-nano	PNI1	32	$^{*}<320$	35.262	37.262
gemini-2.0	PN1	174	1268	32.267	34.474
gemini-2.0	PI2	199	174	20.521	24.045

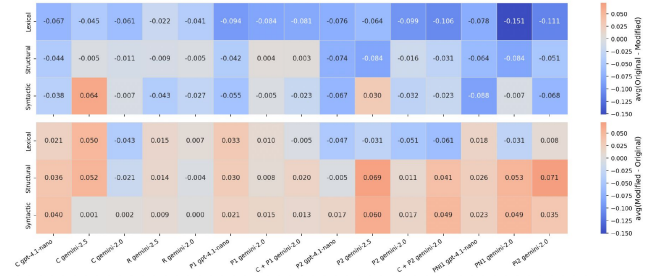
Runs performed okay

...but why?

Is SARI still an adequate evaluation measure for LLM-based text simplification?

We considered different

- Numbers of complex terms
- LLMs
- Simplicity signals



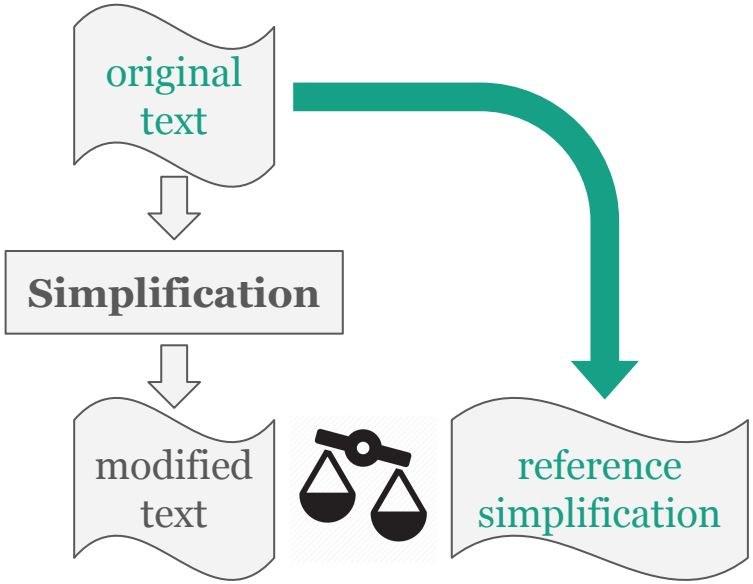
Setting		Comprehensive				Targeted			
Model	P	f10	f1	original	auto	f10	f1	original	auto
gpt-4.1-nano	P1	72	17	37.484	39.896	67	22	38.238	40.416
gpt-4.1-nano	P2	60	24	39.135	40.647	75	21	**39.572**	**41.315**
gpt-4.1-mini	P1	95	20	26.623	30.271	129	33	26.288	30.121
gpt-4.1-mini	P2	109	30	29.388	32.223	126	39	29.336	31.475
gpt-5.4-nano	P1	51	19	25.803	29.608	62	15	26.806	29.852
gpt-5.4-nano	P2	38	13	27.138	30.323	44	14	27.731	31.025
gpt-5.4-mini	P1	66	20	22.338	26.362	63	20	21.957	25.948
gpt-5.4-mini	P2	65	16	33.107	35.312	79	15	34.010	36.480

θ	.01	.025	.05	.075	.1	.2	.3	.4	.5	1
# texts with CTI	7400	7285	6901	6500	5687	3824	2775	2566	2539	2524

Model	P	f10	f1	orig	auto	FRE ↑	CL↓	T↓	B↑
baseline	-	-	-	7.844	12.033	31.523	13.752	2.104	-
gpt-4.1-nano	C	8	1	32.445	33.944	-6.139	19.805	3.27	0.93
gemini-2.5	C	18	7	30.175	32.586	-14.528	20.887	3.015	0.927
gemini-2.0	C	10	1	28.917	31.199	-3.714	19.504	2.923	0.934
gemini-2.5	R	1	0	29.256	32.598	25.914	14.977	2.186	**0.953**
gemini-2.0	R	47	1	27.526	31.160	33.193	13.77†	**2.096**†	0.843
gpt-4.1-nano	P1	67	22	38.238	40.416	34.498	13.308	2.726	0.928
gemini-2.0	P1	42	22	25.858	30.001	35.876	12.578	2.811	0.836
gemini-2.0	C+P1△	44	22	25.234	29.080	36.001	12.585	2.815	0.948
gpt-4.1-nano	P2	75	21	**39.572**	**41.315**	31.238	13.829†	2.816	0.923
gemini-2.5	P2	4275	1550	25.062	27.811	35.909	12.08	2.752	0.943
gemini-2.0	P2	87	132	29.263	31.174	35.759	12.674	3.003	0.829
gemini-2.0	C+P2△	91	132	27.561	30.024	35.829	12.675	3.014	0.94
gpt-4.1-nano	PNI1	32	*<320	35.262	37.262	34.327	13.201	2.786	0.829
gemini-2.0	PN1	174	1268	32.267	34.474	**43.769**	**11.475**	2.674	0.842
gemini-2.0	PI2	199	174	20.521	24.045	36.925	11.966	2.479	**0.953**

Findings

- LLMs prompted to only modify complex components but changed syntax, structure & lexical features of text
- More powerful LLMs do not always achieve better SARI scores



More powerful LLMs
→ more rewriting
→ lower n-gram overlap
→ worse SARI

Worse SARI ≠ worse simplifications

Takeaways

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

Big thanks to my student assistants and the SimpleText organisers!