Revisiting Text Simplification based on Complex Terms for Non-Experts

SimpleText@CLEF ‘25, Task 1.1

Nico Hofmann, Julian Dauenhauer, Nils Ole Dietzler,
Idehen Daniel Idahor, Christin Katharina Kreutz

2025-09-10 – Madrid, Spain

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

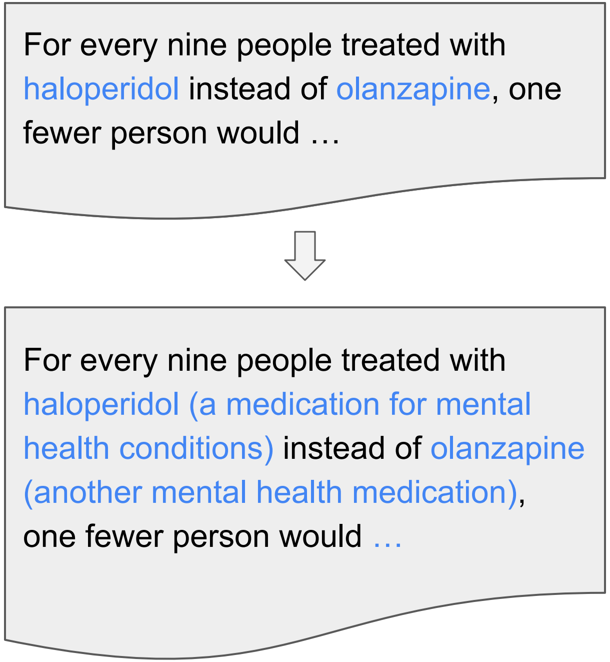
HERDER-INSTITUT
für historische Ostmitteleuropaforschung
INSTITUT DER LEIBNIZ-GEMEINSCHAFT

Motivation

- Simplification for non-experts
  → Readers have *good knowledge* of *language* but *not* of the *topic* at hand

- Low cost
- Mark complex terms indicating domain jargon, simplify these complex terms, keep the rest

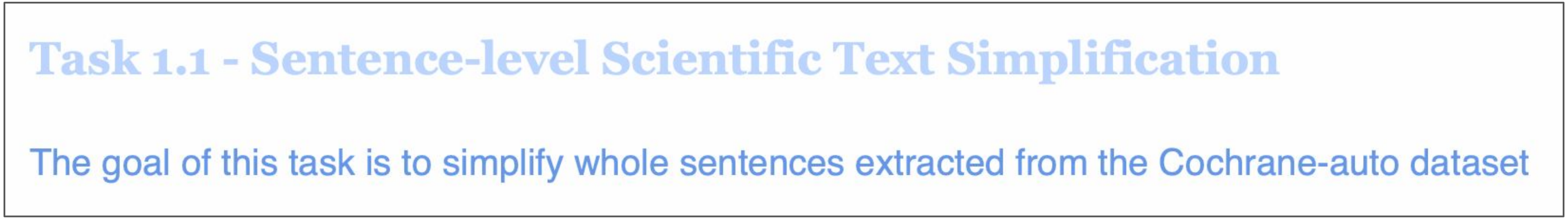
- Revisit approach from SimpleText@CLEF ’23



For every nine people treated with haloperidol instead of olanzapine, one fewer person would …

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

Task

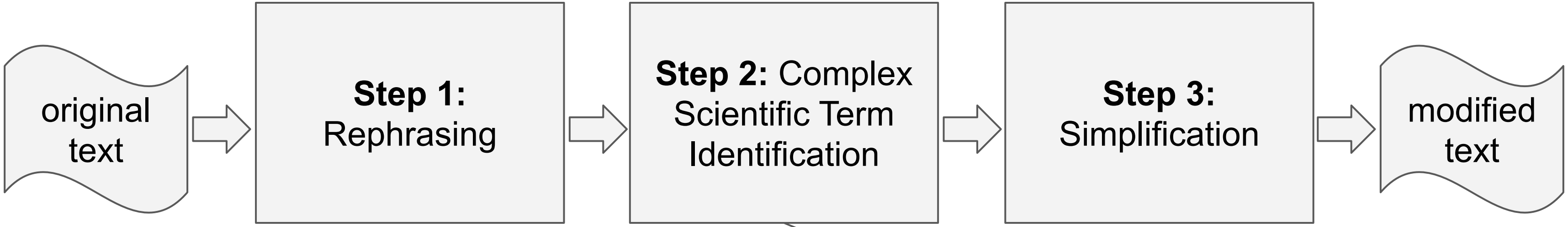


Task 1.1 - Sentence-level Scientific Text Simplification

The goal of this task is to simplify whole sentences extracted from the Cochrane-auto dataset

- Short texts
- Very little context
- Biomedical abstracts

Approach



original text

Step 1: Rephrasing

Step 2: Complex Scientific Term Identification

Step 3: Simplification

modified text

All steps are optional

Text Simplification of Scientific Texts for Non-Expert Readers

Notebook for the SimpleText Lab at CLEF 2023

Björn Engelmann[1], Fabian Haak[1], Christin Katharina Kreutz[1], Narjes Nikzad Khasmakhi[1] and Philipp Schaer[1]

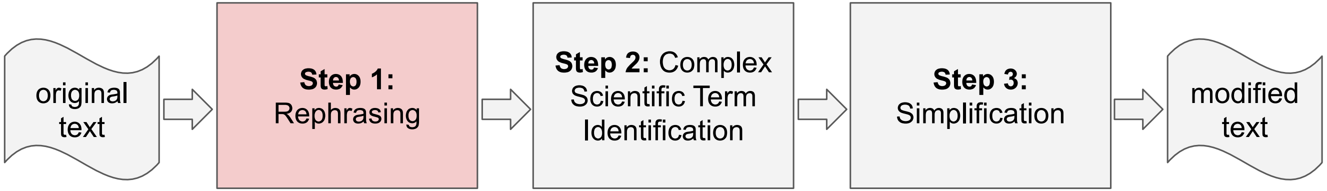
[1] *TH Köln – University of Applied Sciences, Cologne, Germany*

Abstract

Reading levels are highly individual and can depend on a text's language, a person's cognitive abilities, or knowledge on a topic. Text simplification is the task of rephrasing a text to better cater to the abilities of a specific target reader group. Simplification of scientific abstracts helps non-experts to access the

Idea: Cheap complex term identification, low-cost LLM for simplification

Pipeline

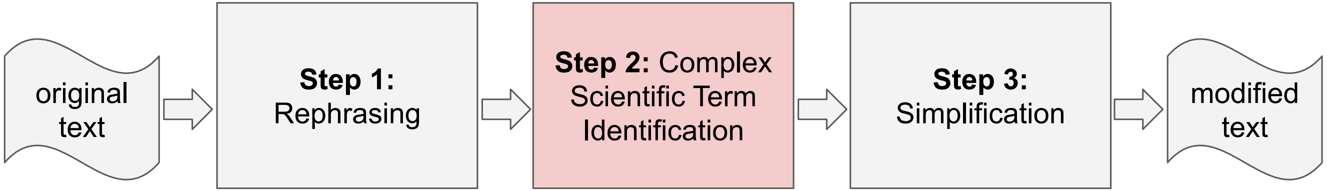

original text
Step 1: Rephrasing
Step 2: Complex Scientific Term Identification
Step 3: Simplification
modified text

You are a **text rephrasing system**. You will be given 10 texts, TREAT THEM SEPERATLY and also return 10 texts. Please rephrase the texts. Do not change their level of complexity (so do not make them more difficult in their structure) and do not make them considerably longer. [...]

You are a **text complication system** with good global and language knowledge but and expertise in specific domains. You will be given 10 texts, TREAT THEM SEPERATLY and also return 10 more complex texts. Please make the texts more complex (not in their structure) but not considerably longer. [..]

Should produce more complex texts and worse results

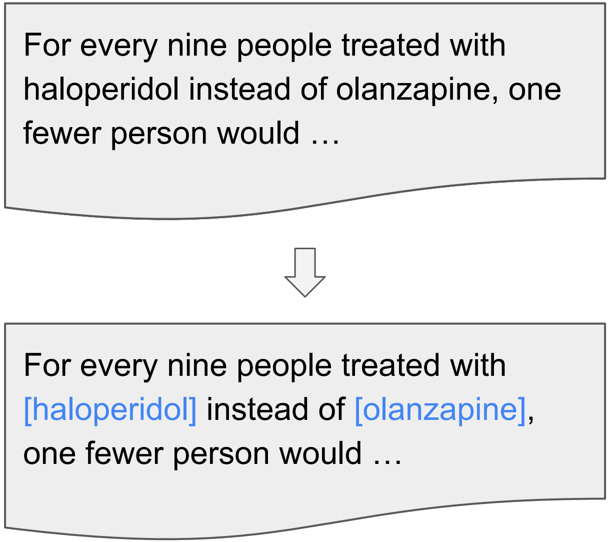
Pipeline


original text
Step 1: Rephrasing
Step 2: Complex Scientific Term Identification
Step 3: Simplification
modified text

- Identify keyphrases [Kulkarni et al., NAACL ‘22]

- Only keep keyphrases indicating domain jargon, usage of tf-idf-inspired measure [Engelmann et al., SimpleText@CLEF ‘23]

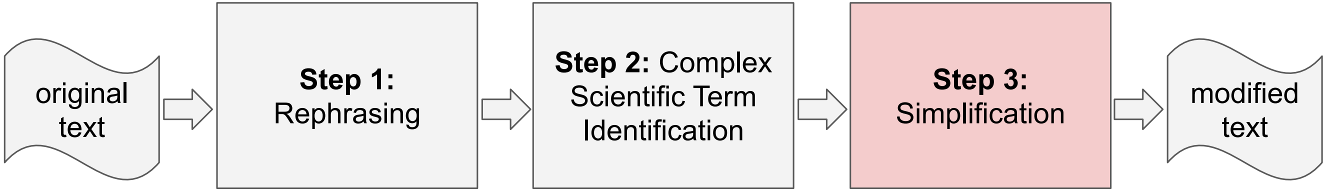
- Stricter decision if keyphrase is domain jargon, reduced amount of false positive keyphrases



For every nine people treated with haloperidol instead of olanzapine, one fewer person would …

For every nine people treated with [haloperidol] instead of [olanzapine], one fewer person would …

Pipeline

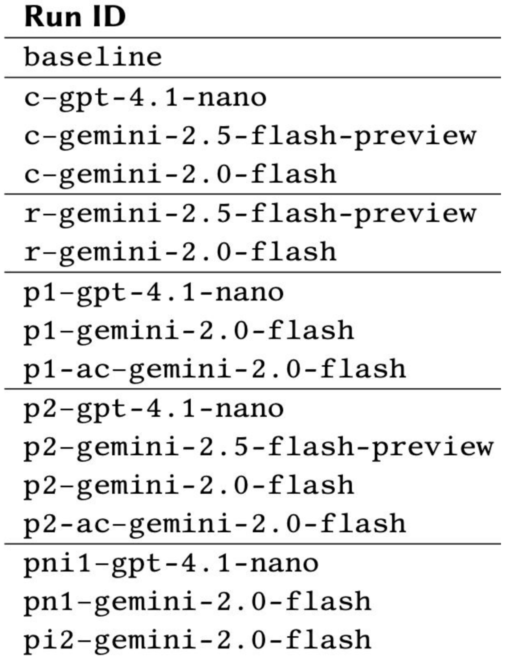

original text
Step 1: Rephrasing
Step 2: Complex Scientific Term Identification
Step 3: Simplification
modified text

- **P2:** Replace or explain marked terms [Engelmann et al., SimpleText@CLEF '23]
- **P1:** P2 + Persona as text simplification system, good global and language knowledge but no expertise in specific domains
- **PNI1**: P2 + More informal, implied persona ("the best and most reliable system for text simplification and language in general")
- **PN1**: P2 + Persona as someone attending conference
- **PI2**: P2 + Translator to translate complex technical language in everyday language

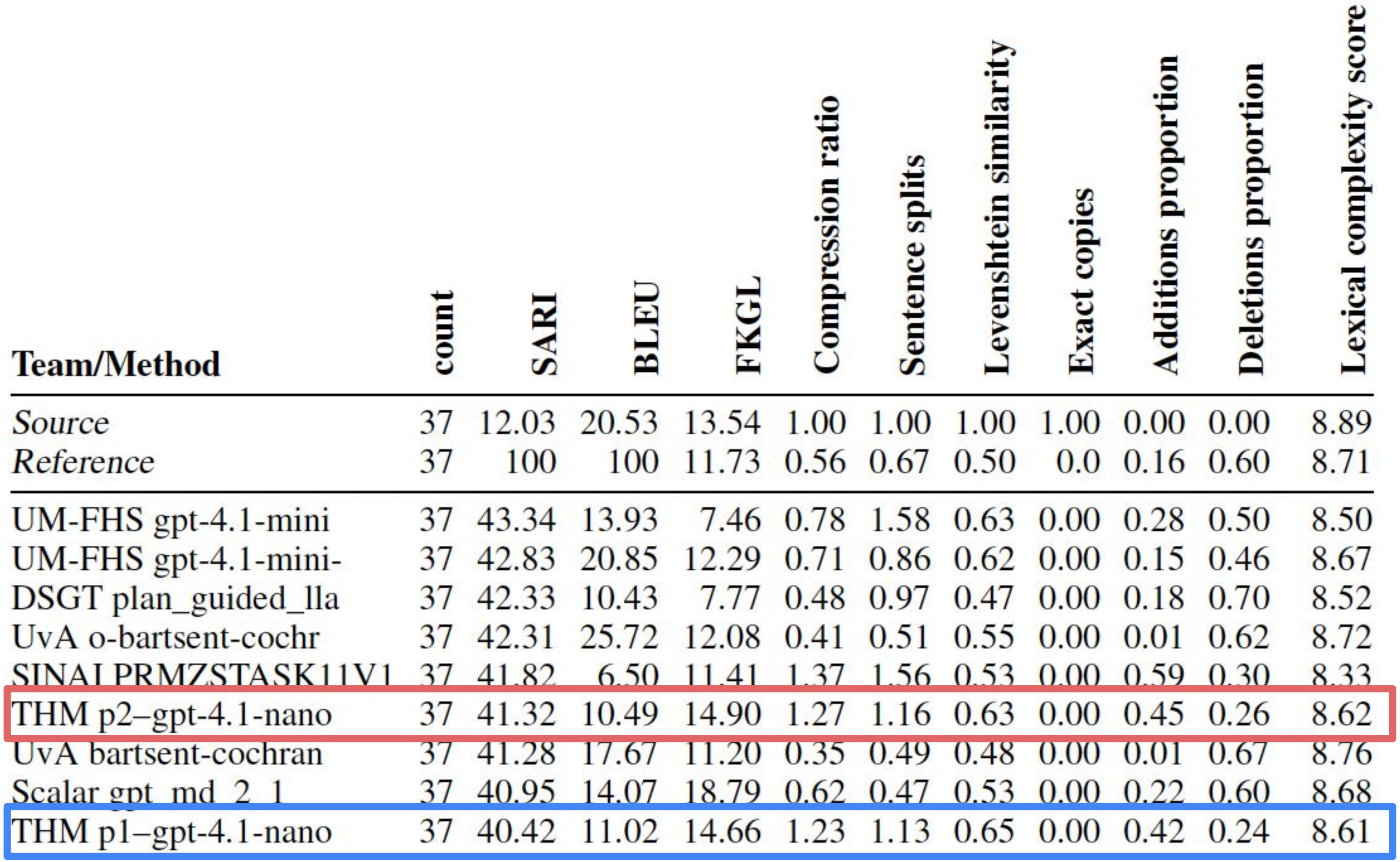
Experiments

- Used GPT-4.1-nano, Gemini-2.5-flash-preview, Gemini-2.0-flash, (GPT-3.5-turbo-0125)

- Types of runs:
  - Only rephrasing
  - Rephrasing, complex term identification, simplification
  - Complex term identification + simplification



Run ID
baseline
c-gpt-4.1-nano
c-gemini-2.5-flash-preview
c-gemini-2.0-flash
r-gemini-2.5-flash-preview
r-gemini-2.0-flash
p1-gpt-4.1-nano
p1-gemini-2.0-flash
p1-ac-gemini-2.0-flash
p2-gpt-4.1-nano
p2-gemini-2.5-flash-preview
p2-gemini-2.0-flash
p2-ac-gemini-2.0-flash
pni1-gpt-4.1-nano
pn1-gemini-2.0-flash
pi2-gemini-2.0-flash



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
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

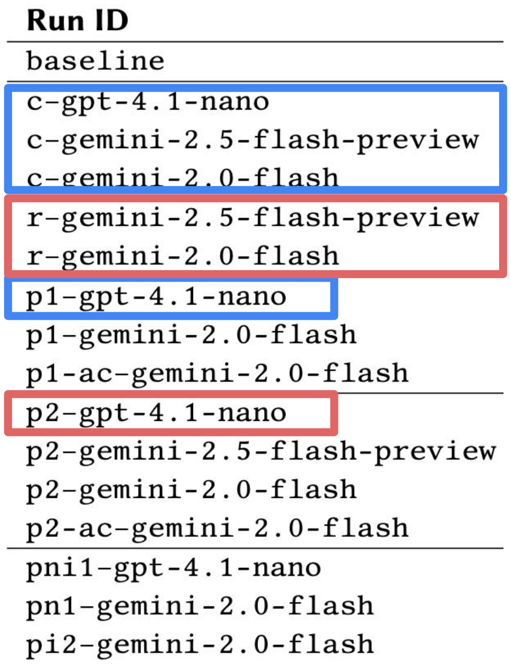
Results

- P1 & P2 + gpt-4.1-nano produced similar texts
- P2 produces shorter texts, stays closer to original
- P1 & P2 still lead to changes in non-marked parts

- Complexification not worse than rephrasing

- Higher failure rates in Gemini runs than GPT runs

- Total cost of participation < $12



Run ID
baseline
c-gpt-4.1-nano
c-gemini-2.5-flash-preview
c-gemini-2.0-flash
r-gemini-2.5-flash-preview
r-gemini-2.0-flash
p1-gpt-4.1-nano
p1-gemini-2.0-flash
p1-ac-gemini-2.0-flash
p2-gpt-4.1-nano
p2-gemini-2.5-flash-preview
p2-gemini-2.0-flash
p2-ac-gemini-2.0-flash
pni1-gpt-4.1-nano
pn1-gemini-2.0-flash
pi2-gemini-2.0-flash

Conclusion

- Complex term identification + low cost LLMs produce viable results
- Prompts producing texts similar to original texts perform better in evaluation than prompts producing more changed up wording
- Difficulty of text not sole determining factor for used evaluation measures → better evaluation measures needed

Big thanks to my student assistants and the SimpleText organisers!