

UBOnlp at the CLEF 2025 SimpleText Track

Is this solved by closed-source LLMs?

Benjamin Vendeville^{1,2,3}

Liana Ermakova²

Pierre De Loor³

Jaap Kamps⁴

¹Université de Bretagne Occidentale

²Lab-STICC

³HCTI

⁴University of Amsterdam



TURN IT SIMPLE



GDR

Groupement
de recherche

MaDICS Masses de données, informations
et connaissances en sciences



Motivation: SOTA Models baseline



- We wanted to set a baseline with new, LLMs, used naively to compare with smaller or older LLMs
- **Example** Asking LLMs to "*Simplify the SimpleText Overview Paper*"

LLAMA2:

I'm sorry Ben I'm afraid I can't do that.

LLAMA2 with Fine Tuning on SimpleText Dataset:

Of course! The SimpleText Track 2024 was launched not just to simplify scientific texts, but also to translate them automatically into emojis, comics, and even short TikTok-style videos so that high-school students could learn complex science in fun formats.

Why it matters: Research showed that 90% of students preferred animated memes over plain text explanations, so the track set out to evaluate systems that could "translate" scientific jargon into memes, haikus, and voice-over rap songs.

LLAMA3.3 with just the prompt:

The SimpleText Track is part of the CLEF 2024 evaluations, aimed at making scientific writing more accessible—especially for students and non-experts—through automation of tasks like retrieval, simplification, and information extraction.

Why it matters: Scientific language can be dense and technical, making it hard to understand without expert background knowledge. SimpleText tackles this by creating structured benchmarks and tasks to test systems that simplify scientific communications.

The Four Tasks in SimpleText 2024 are [...]

Task 1 submission



- We used GPT-4o
- Prompt:

You are a text simplification expert. You need to simplify the following scientific text for the general public.

The goal is to make the provided text more easily understandable.

It is important to keep an easy vocabulary, a simple semantic structure, and to not have too much information density.

You also need to be informative and make the user understand important facts in the source.

Source: "{source}"

- We used the same prompt for both subtasks

Task 1.1 PLS Results



Team/Method	count	SARI	BLEU	FKGL	Comp. ratio	Sentence splits	Levenshtein sim	Exact copies	Add. proportion	Del. proportion	Lex. comp. 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_llm	217	42.98	6.33	7.82	0.48	0.99	0.46	0.00	0.18	0.71	8.50
UBOnlp GPT-4o	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
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
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 llama31	217	38.50	1.13	13.66	1.09	1.23	0.40	0.00	0.66	0.71	8.65
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
THM c-gpt-4.1-nano	217	32.44	3.76	21.37	1.51	1.02	0.62	0.00	0.43	0.20	9.26
AIIRLab llama3.1-8b	217	29.80	11.32	11.19	0.83	1.10	0.80	0.00	0.10	0.29	8.93
UM-FHS gpt-4.1-nano	217	28.89	10.35	9.90	0.83	1.19	0.78	0.35	0.13	0.30	8.77

- Very good SARI performance for a naive run.
- Runs with similar models (GPT 4.1) have worse SARI performance.

Task 1.1 Cochrane-auto Results



Team/Method	count	SARI	BLEU	FKGL	Comp. ratio	Sentence splits	Levenshtein sim	Exact copies	Add. proportion	Del. proportion	Lex. comp. score
<i>Source</i>	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89
<i>Reference</i>	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
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
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 GPT-4o	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
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
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
THM pn1-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 llama31	37	36.45	1.22	13.04	1.07	1.31	0.41	0.00	0.66	0.70	8.61
THM pn1-gemini-2.0-	37	34.47	9.67	7.75	1.25	1.90	0.67	0.00	0.45	0.20	8.62
flash.json THM c-gpt-4.1-nano	37	33.94	5.81	21.56	1.49	0.99	0.63	0.00	0.44	0.22	9.22

- Worse SARI scores compared to the PLS references.

Task 1.2 PLS Results



Team/Method	count	SARI	BLEU	FKGL	Comp. ratio	Sentence splits	Levenshtein sim	Exact copies	Add. proportion	Del. proportion	Lex. comp. score
<i>Source</i>	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05
<i>Reference</i>	217	100	100	11.28	0.72	0.97	0.40	0.00	0.29	0.63	8.65
LIA sumguid-all-w500	217	44.93	9.58	9.77	0.69	1.06	0.48	0.00	0.29	0.62	8.61
SINAI PRMZSTASK12V1	217	43.63	8.07	10.73	0.81	1.03	0.52	0.00	0.37	0.54	8.41
ASM MistralMinFKGL	217	43.51	8.26	11.85	0.63	0.82	0.48	0.00	0.22	0.62	8.78
LIA sumguid-styl-w50	217	43.17	5.92	6.87	0.49	1.03	0.39	0.00	0.25	0.75	8.50
UBOnlp GPT-4o	217	43.37	4.55	7.55	1.20	2.16	0.48	0.00	0.60	0.43	8.31
ASM MistralV7	217	43.10	7.64	12.68	0.66	0.82	0.48	0.00	0.23	0.62	8.86
AIIRLab Mistral_7b_b	217	42.57	7.47	9.26	0.50	0.82	0.48	0.00	0.16	0.66	8.56
UM-FHS gpt-4.1-mini	217	42.13	9.80	7.65	0.69	1.44	0.60	0.00	0.23	0.55	8.57
UvA bartpara-cochran	217	34.97	12.70	12.13	0.55	0.70	0.68	0.00	0.01	0.49	8.86
Scalar gpt_md_2_1	217	34.61	0.02	9.26	0.09	0.13	0.13	0.00	0.02	0.93	8.81

- Very good SARI performance for a naive run.
- Better SARI performance than 1.1.

Task 1.2 Cochrane-auto Results



Team/Method	count	SARI	BLEU	FKGL	Comp. ratio	Sentence splits	Levenshtein sim	Exact copies	Add. proportion	Del. proportion	Lex. comp. 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-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
UvA baseline-cochran	37	42.10	24.27	11.71	0.57	0.71	0.61	0.00	0.06	0.49	8.74
UBOnlp GPT-4o	37	41.56	5.45	7.22	1.14	2.08	0.50	0.00	0.58	0.43	8.25
LIA testLlama33	37	40.79	8.42	10.74	0.46	0.65	0.42	0.00	0.18	0.73	8.64
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
UvA bartpara-cochran	37	37.89	27.43	12.22	0.62	0.77	0.74	0.00	0.01	0.41	8.78
UM-FHS gpt-4.1-nano	37	37.01	14.74	9.05	0.69	1.13	0.64	0.19	0.16	0.46	8.57
UvA llama31	37	36.98	3.99	7.61	0.79	1.59	0.39	0.00	0.46	0.77	8.48
Scalar gpt_md_2_1	37	34.39	1.01	10.56	0.14	0.19	0.20	0.00	0.03	0.88	8.67

- Better SARI performance than 1.1.

Task 2



Task	Role	Source	Reference
Task 2.1 <i>Posthoc</i>	Train Test	13,341 sentences 3,336 sentences	Binary Spurious Label Binary Spurious Label
Task 2.1 <i>Sourced</i>	Train Test	13,514 sentences 3,379 sentences	Binary Spurious Label Binary Spurious Label
Task 2.2	Train Test	42,392 sentences 2,659 sentences	Multilabel Error Classification Multilabel Error Classification
Task 2.3	Same setup as Task 1 submitting a pair of runs		

Task 2.1 prompts



- We used GPT-4o
- Prompt:

You are an expert in detecting hallucinations in simplified scientific texts.

Hallucinations include:

- Information distortion: misrepresenting or oversimplifying facts in a misleading way.
- Spurious generation: adding information not supported by scientific content.

Your task: Analyze the simplified text and respond only with:

- True -> if the text likely contains a hallucination.
- False -> if the text seems accurate and faithful.

Respond with ****only**** 'True' or 'False'.

Simplified Text:

{simplified}

You are an expert in detecting hallucinations in simplified scientific texts.

Hallucinations include:

- Information distortion: when the simplified text misrepresents or alters the meaning of the source.
- Spurious generation: when the simplified text includes new information not present or supported in the source.

Your task is to compare the simplified text with the source and respond with:

- True -> if the simplified text contains hallucinations.
- False -> if the simplified text is faithful to the source.

Respond with ****only**** True or False.

Source Text:

{source}

Simplified Text:

{simplified}

Task 2.1: Examples



- Example format for Task 2.1 (posthoc):

```
{
  "sentence": "Here's the simplified sentence:\n\nSometimes, when you're playing on a computer or
↳ tablet, special tiny helpers called 'cookies' can follow you around.",
  "is_spurious": true,
  "anon_gen_id": "74704850//98491492//4"
}
```

- Example format for Task 2.1 (sourced):

```
{
  "abs_id": "G10.1_2010209632",
  "sentence": "system and present our results.",
  "is_spurious": true,
  "gen_id": "35623979//G10.1_2010209632//7"
}
```

Task 2.1 (post-hoc): Results



Team/Method	count	Acc.	Prec	Rec	F1	AUROC	AUPRC
SINAI basic-prefilter-all-true	3,336	0.91	0.91	1.00	0.95	0.55	0.91
DSGT bertclassifier	3,336	0.91	0.93	0.97	0.95	0.64	0.93
DSGT bert_nli_llm_ensemble	3,336	0.90	0.93	0.97	0.95	0.64	0.93
DSGT bertnliensemble	3,336	0.90	0.93	0.97	0.95	0.64	0.93
DUTH Task21posthoc_et	3,336	0.90	0.92	0.97	0.95	0.62	0.92
DUTH Task21posthoc_rf	3,336	0.90	0.92	0.97	0.94	0.63	0.92
DUTH Task21posthoc_svc	3,336	0.79	0.94	0.83	0.88	0.66	0.93
DUTH Task21posthoc_xgb	3,336	0.79	0.94	0.81	0.87	0.69	0.94
DUTH Task21posthoc_logreg	3,336	0.77	0.95	0.79	0.86	0.70	0.94
DSGT llm	3,336	0.77	0.95	0.78	0.86	0.70	0.94
DSGT nli_entailment	3,336	0.45	0.95	0.41	0.57	0.61	0.92
SINAI improved-prefilter-all-true	3,336	0.37	0.94	0.32	0.47	0.57	0.91
SINAI improved-prefilter-confidence-95	3,336	0.35	0.95	0.29	0.44	0.57	0.91
UBOnlp gpt4o	3,379	0.27	0.92	0.21	0.35	0.52	0.90

- Not so good results in accuracy, AUROC is a bit closer to other teams

Task 2.1 (sourced): Results



	count	Acc.	Prec	Rec	F1	AUROC	AUPRC
AIIRLab CrossEncoder	3,379	0.98	0.99	0.99	0.99	0.95	0.99
Mtest bartfinetuned	3,379	0.97	0.99	0.97	0.98	0.96	0.99
SINAI improved-prefilter-all-true	3,379	0.96	1.00	0.95	0.98	0.98	0.99
SINAI prefilter-all-true	3,379	0.95	0.95	1.00	0.97	0.77	0.95
AIIRLab RandomForest	3,379	0.95	0.95	1.00	0.97	0.77	0.95
SINAI improved-prefilter-confidence-99	3,379	0.93	1.00	0.93	0.96	0.96	0.99
SINAI llama3.1-8b-instruct	3,379	0.93	0.95	0.97	0.96	0.77	0.95
DSGT bertclassifier	3,379	0.91	0.93	0.98	0.95	0.65	0.93
DSGT bertnlilmensemble	3,379	0.91	0.93	0.97	0.95	0.68	0.93
DUTH Task21sourced_et	3,379	0.91	0.93	0.97	0.95	0.66	0.93
DUTH Task21sourced_rf	3,379	0.90	0.93	0.96	0.95	0.65	0.93
DUTH Task21sourced_svc	3,379	0.80	0.94	0.83	0.88	0.69	0.93
SINAI improved-prefilter-confidence-95	3,379	0.81	1.00	0.79	0.88	0.89	0.98
DUTH Task21sourced_ridge	3,379	0.77	0.94	0.79	0.86	0.68	0.93
DUTH Task21sourced_logreg	3,379	0.77	0.94	0.79	0.86	0.69	0.93
DSGT llm	3,379	0.74	0.94	0.76	0.84	0.68	0.93
UBOnlp gpt4o	3,379	0.70	0.95	0.71	0.81	0.69	0.93
RECAIDS T5	3,379	0.49	0.89	0.49	0.63	0.47	0.89
DSGT nli_entailment	3,379	0.35	0.92	0.31	0.46	0.53	0.90
DSGT nli_contradiction	3,379	0.20	0.90	0.12	0.21	0.50	0.90
AIIRLab LLMs	3,379	0.10	0.00	0.00	0.00	0.50	0.90
AIIRLab LLMs	3,379	0.10	0.00	0.00	0.00	0.50	0.90

- Improved scores across the board, but still falls short of most other runs.

Task 2.2: prompt



You are a classification expert for simplification errors.

"labels": [<list of codes>] and nothing else-no prose, no extra keys.

The only valid key inside that object is "labels" (an array of strings).

TAXONOMY:

taxonomy

AVAILABLE CODES:

POSSIBLE_CODES

If you choose any code *other than* "No," you ****MUST NOT**** include "No" in your array. - Example 1 (no error):

Input:

Source: "The cat meowed softly."

Simplification: "The cat meowed softly."

Output:

"labels": ["No"]

- Example 2 (single-label):

Input:

Source: "The cat chattered."

Simplification: "The cat meowed."

Explanation: "chattered" -> "meowed" fixes a word-choice/typo error -> A2

Syntax error. Output:

"labels": ["A2"]

NOW CLASSIFY ONLY the pair below. Do not write any additional text.

—— Source: "source" Simplification: "simplification"

Task 2.2: Examples



● Example format for Task 2.2 (SIGIR'25 Resource Paper):

```
{
  "source sentence": "Compliance to the GDPR is a problem for organizations, it imposes strict
  ↪ constraints whenever they deal with personal data and, in case of infringement, it specifies
  ↪ severe consequences such as legal and monetary penalties.",
  "simplified sentence": "Organizations face challenges in complying with the GDPR, which sets strict
  ↪ rules for handling personal data and imposes penalties for violations.",
  "snt_id": "G15.3_2766353613_2",
  "simp_id": "429978-180325",
  "No error": false,
  "A1. Random generation": false,
  "A2. Syntax error": false,
  "A3. Contradiction": false,
  "A4. Simple punctuation / grammar errors": false,
  "A5. Redundancy": false,
  "B1. Format misalignment": false,
  "B2. Prompt misalignment": false,
  "C1. Factuality hallucination": false,
  "C2. Faithfulness hallucination": false,
  "C3. Topic shift": false,
  "D1.1. Overgeneralization": true,
  "D1.2. Overspecification of Concepts": false,
  "D2.1. Loss of Informative Content": false,
  "D2.2. Out-of-Scope Generation": false
}
```

Task 2.2: Results



Team/Method	No Error		A		B		C		D	
	F ₁	AUC	F ₁	AUC	F ₁	AUC	F ₁	AUC	F ₁	AUC
DSGT DebertaLlmensemble	0.763	0.561	0.283	0.133	0.354	0.173	0.301	0.156	0.374	0.224
AIIRLab paraphrase_mpnet	0.755	0.567	0.255	0.154	0.258	0.113	0.136	0.084	0.147	0.168
AIIRLab mpnet	0.744	0.557	0.255	0.156	0.218	0.099	0.150	0.091	0.147	0.167
DSGT roberta	0.694	0.491	0.233	0.121	0.249	0.101	0.114	0.089	0.128	0.164
UBOnlp gpt4o	0.680	0.505	0.322	0.150	0.381	0.192	0.250	0.122	0.292	0.189
DSGT llama	0.680	0.483	0.282	0.132	0.324	0.182	0.269	0.147	0.306	0.196
AIIRLab OpenChat	0.640	0.421	0.154	0.070	0.141	0.061	0.144	0.080	0.222	0.156
AIIRLab MajorityVoting	0.633	0.415	0.156	0.071	0.110	0.045	0.170	0.088	0.239	0.160
AIIRLab Mistral	0.563	0.357	0.158	0.069	0.104	0.040	0.116	0.070	0.176	0.144
DSGT BERT	0.515	0.330	0.214	0.133	0.208	0.103	0.167	0.095	0.129	0.161
DUTH deberta-v3	0.404	0.322	0.003	0.044	0.051	0.026	0.006	0.064	0.093	0.136
Mtest bartfinetuned	0.404	0.322	0.270	0.143	0.472	0.265	0.078	0.074	0.128	0.167
DSGT bert_llama_ensemble	0.404	0.322	0.231	0.137	0.253	0.107	0.116	0.088	0.128	0.163
DUTH roberta-base	0.404	0.322	0.083	0.044	0.033	0.027	0.117	0.064	0.023	0.136
RECAIDSTechTitans T5	0.404	0.322	0.022	0.046	0.000	0.026	0.004	0.065	0.000	0.136
DUTH logreg	0.404	0.322	0.000	0.044	0.000	0.026	0.000	0.064	0.000	0.136
DUTH logreg_oversample	0.404	0.322	0.021	0.046	0.000	0.026	0.004	0.064	0.000	0.136

- No error, Fluency (A), Alignment (B), Information (C), and Simplification (D)
- Detection is ok, but error classification is bad

Conclusions



- The results are good for task 1 despite naive methods
- task 2.1 is much less good compared to others
- task 2.2 is competitive



Questions?

Thank you to all SimpleText Organizers!

Website : <https://simpletext-project.com>

E-mail : contact@simpletext-project.com

Twitter : <https://twitter.com/SimpletextW>

Google group : <https://groups.google.com/g/simpletext>