CLEF 2025 SimpleText Track

Simplify Scientific Text (and Nothing More)

Liana Ermakova[1] Hosein Azarbonyad[2] Jan Bakker[3]
Benjamin Vendeville[1] Jaap Kamps[3]

[1]Université de Bretagne Occidentale
[2]Elsevier
[3]University of Amsterdam

CLEF 2025, September 10, 2025, Madrid, Spain

Motivation: Simplify Scientific Text

- Improving Access to Scientific Texts for Everyone
  - Everyone agrees on the importance of objective scientific information
  - But scientific documents are inherently complex...
- Can we improve accessibility for everyone?
  - Generative models for text simplification can help!
- Cochrane-auto: aligned scientific abstract+plain language summary

Complex paragraph: Fifteen heterogeneous trials, involving 1022 adults with dorsally displaced and potentially or evidently unstable distal radial fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation especially in terms of patient characteristics and interventions. Methodological weaknesses among these trials included lack of allocation concealment and inadequate outcome assessment.

Simple paragraph: Fifteen trials, involving 1022 adults with potentially or evidently unstable fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation in their characteristics especially in terms of patient characteristics and the method of external fixation.

Motivation: ... and Nothing More!

- LLMs prone to overgeneration (informally called "*hallucination*")
  - CLEF 2024 submission output versus input: ~~deletions~~, insertions, and whole sentence insertions

As various kinds of output devices emerged , such as highresolution printers or a display of PDA (Personal Digital Assistant) ~~, the~~ . The importance of high-quality resolution conversion has been increasing . |This paper proposes a new method for enlarging an image with high quality . It will involve using a combination of high-speed imaging and high-resolution video . |One of the ~~largest~~ biggest problems on image enlargement is the exaggeration of the jaggy edges . This is especially true when the image is enlarged , as in this case . |To remedy this problem , we propose a new interpolation method ~~, which~~ . This method uses artificial neural network to determine the optimal values of interpolated pixels . |The experimental results are shown and evaluated . The results are compared to other studies and found to be inconclusive . |The effectiveness of our methods is discussed by comparing with the conventional methods . Our methods are designed to help people with mental health problems , not just as a way to cure them . |

- Have created a huge collection of spurious/over-generation content!
  - In 2024: 47% of submissions > 10% spurious sentences, 19% > 50%...

Overview

- SimpleText Track setup similar 2021-2024
  - Very successful benchmarks constructed
  - "Finished" original tasks?
  - Major changes in setup and corpora in 2025
- CLEF 2025 SimpleText Track
  - *Simplify Scientific Text (and Nothing More)*
- The following *three* tasks:
  1. **Text Simplification**: *simplify scientific text*
  2. **Controlled Creativity**: *identify and avoid hallucination*
  3. **SimpleText 2024 Revisited**: *selected tasks by popular request*

SimpleText 2025 Statistics

- Growing steadily: 74 registered teams, 18 submitted 198 valid runs.
- Codabench Task 1: 281 submissions (30 ppl), Task 2: 232 submissions (14 ppl)

Team	Task 1		Task 2			Task 1	Task 2	Total runs
	1.1	**1.2**	**2.1**	**2.2**	**2.3**			
AIIRLab	4	2	5	5		6	10	16
ASM		10				10		10
DSGT	2	1	6	6	3	3	15	18
DUTH	3		2	2		3	4	7
EngKh	2					2		2
Fujitsu	19					19		19
LIA		9				9		9
Mtest	1	1	1	1		2	2	4
PICT	1	1				2		2
RECAIDS	1	1	1	1		2	2	4
Scalar	10	1			1	11	1	12
SINAI	2	2	15	15		4	30	34
THM	22					22		22
UBO	5	7	1	1		12	2	14
UM-FHS	4	5				9		9
UvA	5	9				14		14
Unknown	2					2		2
Total	83	49	31	31	4	132	66	198

Task 1: Text Simplification

- *Task 1: Simplify Scientific Text*
  - New corpus (EMNLP/TSAR 2024)!
    - Cochrane-auto is true document-level text simplification
    - More variation (sentence merge, order swaps) and discourse structure
    - Paragraph-level and sentence-level data realigned and restricted
  - Biomedical text – free to use, similar to existing TS corpora
    - Sentence-level (T1.1) and Document-level (T1.2) text simplification
    - Large-scale aligned train and test data (9,160 sentences, 666 abstracts)
    - → Frees human judge effort for analysis...

Cochrane-auto	**Train**	**Test**	**SimpleText'24**
	Biomedical		Science
# Documents	5,585	666	278
# Sentences	35,800	9,160	1,536

- *Bonus: Cochrane Abstracts/Plain Language Summaries in English*
  - + *Spanish, French, Farsi, Chinese, Japanese, Portuguese, Korean, ...*

Task 1: Evaluation

- Source:

```
{ "pair_id": "CD012520", "source": "Cochrane", "complex": "We included seven cluster-randomised
↪  trials with 42,489 patient participants from 129 hospitals, conducted in Australia, the UK,
↪  China, and the Netherlands. ... We are uncertain whether a multifaceted implementation
↪  intervention compared to no intervention improves adherence to evidence-based recommendations in
↪  acute stroke settings, because the certainty of evidence is very low."}
```

- References:

```
{ "CD012520": { "simple_auto": "We included seven studies that involved 42,489 acute stroke patients
↪  and an unknown number of health professionals. The studies were conducted in 129 hospitals in
↪  Australia, the UK, China and the Netherlands. ...  We do not know if implementation interventions
↪  delivered in acute stroke units lead to better delivery of evidence-based care." }, ... }
```

- Predictions:

```
[{"pair_id":"CD012520","prediction":"Researchers conducted studies in hospitals across Australia, the
↪  UK, China, and the Netherlands. ... Overall, the evidence was not strong enough to say for sure
↪  if these strategies help healthcare workers follow best practices in treating stroke
↪  patients.","run_id":"UBOnlp_task12_gpt4o"}, ... ]
```

- Test set includes: New 24/25 sentence-aligned Cochrane-auto (Abstract/PLS), Raw Cochrane PLS, Cochrane-auto Train+Validation, TREC PLABA Medline data, SimpleText 2024.

Task 1.1: Results (doc-level eval)

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
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
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
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
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
EngKh biomedical_lla	37	36.68	11.47	10.62	1.14	1.51	0.65	0.00	0.37	0.28	8.69
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
RECAIDS T5	37	31.68	0.09	3.72	0.37	0.96	0.31	0.00	0.23	0.88	8.87
XXX method	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89

- Codabench evaluation doc-level also for Task 1.1 (identical to 1.2).

Task 1.1: Results (snt-level eval)

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	363	15.01	27.71	13.46	1.00	1.00	1.00	1.00	0.00	0.00	8.62
Reference	363	100.00	100.00	11.71	1.05	1.11	0.60	0.03	0.33	0.42	8.43
UM-FHS gpt-4.1-mini-	363	42.65	19.83	12.03	0.85	0.91	0.62	0.17	0.19	0.41	8.74
AIIRLab llama3.1_gro	363	42.32	13.09	10.90	0.75	0.99	0.66	0.02	0.20	0.47	8.50
THM p2–gpt-4.1-nano	363	41.43	12.25	14.58	1.37	1.19	0.65	0.01	0.47	0.26	8.40
SINAI PRMZSTASK11V1	363	39.89	6.79	11.15	1.49	1.63	0.54	0.00	0.63	0.32	8.19
UvA o-bartsent-cochr	363	39.84	17.92	11.64	0.60	0.70	0.61	0.31	0.02	0.43	9.18
PICT S3Pipeline	363	39.21	12.47	8.05	0.76	1.52	0.62	0.01	0.24	0.48	8.46
UBO gpt4o	363	38.58	5.44	7.17	1.22	2.16	0.51	0.00	0.65	0.48	8.09
DSGT plan_guided_lla	363	38.56	5.23	7.65	0.59	1.00	0.51	0.00	0.29	0.68	8.26
EngKh biomedical_lla	363	38.25	14.03	10.19	0.98	1.45	0.67	0.07	0.32	0.39	8.35
MTest bartfinetuned	363	38.01	27.40	11.51	0.82	1.00	0.87	0.40	0.01	0.20	8.53
Fujitsu dummy90	363	37.64	15.08	3.35	0.65	2.58	0.75	0.20	0.07	0.38	8.51
Scalar gpt_md_2_1	363	37.12	10.51	7.60	0.80	1.30	0.45	0.01	0.27	0.63	8.55
DUTH Task11_flan-t5-	363	36.58	15.17	10.08	0.71	0.99	0.61	0.12	0.18	0.48	8.60
RECAIDS T5	363	30.21	0.05	3.72	0.50	0.99	0.31	0.00	0.37	0.88	8.87
XXX method	363	15.01	27.71	13.46	1.00	1.00	1.00	1.00	0.00	0.00	8.62

- Evaluated on 363 sentence pairs of 37 new 24/25 Cochrane-auto data

Task 1.1: Results (raw doc-level)

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
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
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
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
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
MTest bartfinetuned	217	31.59	14.86	11.90	0.69	0.96	0.80	0.00	0.01	0.36	8.89
XXX method	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05

- Evaluated on 217 (unaligned) new 24/25 Cochrane Abstract/PLS

Task 1.2: Results (doc-level eval)

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
ASM MistralMaxFRE	37	43.35	12.32	11.63	0.73	0.92	0.53	0.00	0.27	0.56	8.74
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
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
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
Mtest bartdoc	37	37.62	20.42	11.79	0.50	0.61	0.62	0.00	0.01	0.51	8.76
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
EngKh biomedical_lla	37	33.25	17.88	12.55	0.72	0.87	0.61	0.05	0.15	0.44	8.77
RECAIDS T5	37	31.49	0.00	10.08	0.06	0.07	0.10	0.00	0.00	0.95	8.12

- Evaluated on 37 aligned new 24/25 Cochrane-auto Abstract/PLS

Task 1.2: Results (raw doc-level eval)

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
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
DSGT llama_summary_s	217	42.92	5.32	9.94	0.49	0.72	0.39	0.00	0.24	0.75	8.55
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 baseline-cochran	217	41.83	10.85	11.10	0.44	0.60	0.49	0.00	0.06	0.63	8.75
DUTH task12_led-larg	217	39.28	3.58	12.46	0.31	0.41	0.40	0.00	0.05	0.76	8.86
PICT S3Pipeline	217	39.11	8.23	6.46	0.71	1.69	0.60	0.00	0.22	0.50	8.76
Mtest bartdoc	217	37.08	7.25	11.50	0.39	0.50	0.52	0.00	0.01	0.63	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
RECAIDS T5	217	33.14	0.00	8.79	0.04	0.06	0.07	0.00	0.00	0.96	8.24
EngKh biomedical_lla	217	28.19	8.55	11.95	0.69	0.79	0.57	0.04	0.12	0.46	9.00

- Evaluated on 217 (unaligned) new 24/25 Cochrane Abstract/PLS

Task 1: Scores hide differences...

TURN IT SIMPLE

- Conservative (Contextual BART model):

~~Today , our~~ Our well-being and awareness have become markedly determined by our way of living through our everyday activities . |~~Needless to say , daily~~ Daily practices specifically have a significant impact on the quality of sleep . |Obstructive sleep apnea (OSA) is an exhausting sleep disorder regulating an individual 's routine life . |Although several therapeutic modalities are available for curing OSA , behavioral therapies are also utilized for a positive outcome . |Besides , several studies are performed to ~~prove~~ show the efficacy of lifestyle strategies to resolve OSA in adults . |Reducing weight , quitting alcohol and smoking , eating a nutritional diet , and exercising are the modifications to benefit people . |This review aims to expand our knowledge of the association between alterations to comportment and better treatment outcomes for sleep apnea . |

- Completely reworded (GPT4o model):

Today , our overall well-being and awareness ~~have become markedly determined~~ are greatly influenced by our ~~way of living through our everyday~~ daily activities . |~~Needless to say ,~~ It 's clear that our daily ~~practices specifically~~ habits have a ~~significant impact~~ big effect on ~~the quality of~~ how well we sleep . |Obstructive sleep apnea (OSA) is ~~an exhausting~~ a tiring sleep disorder ~~regulating an individual 's routine~~ that affects daily life . |~~Although several therapeutic modalities are available~~ Besides medical treatments for ~~curing~~ OSA , behavioral therapies ~~are~~ can also ~~utilized for a positive outcome~~ help . |~~Besides , several~~ Many studies ~~are performed to prove the efficacy of~~ show that lifestyle ~~strategies to resolute~~ changes can effectively treat OSA in adults . |~~Reducing~~ Losing weight , quitting alcohol and smoking , eating ~~a nutritional diet~~ healthy , and exercising are ~~the modifications to benefit people~~ helpful changes . |This review aims to ~~expand our knowledge of the association between alterations to comportment and better treatment outcomes for~~ understand how changing habits can improve sleep apnea treatment . |

Task 1: Scores hide differences...

- Conservative (Contextual BART model):

CD012019 *We included eight studies involving 646 participants ~~, most of which were of poor methodological quality~~ . |~~The urinary biomarkers were evaluated either in a specific phase of menstrual cycle or irrespective of the cycle phase .~~ |Five studies evaluated the diagnostic performance of four urinary biomarkers for ~~endometriosis , including three biomarkers distinguishing women with and without~~ endometriosis (enolase 1 (NNE) ; vitamin D binding protein (VDBP) ; and urinary peptide profiling) ; and one biomarker (cytokeratin 19 (CK 19)) showing no significant difference between the two groups . |~~All of these biomarkers were assessed in small individual studies and could not be statistically evaluated in a meaningful way .~~ |~~None of the biomarkers met the criteria for a replacement test or a triage test .~~ |Three studies evaluated three biomarkers that did not differentiate women with endometriosis from disease-free controls . |There was insufficient evidence to recommend any urinary biomarker for use as a replacement or triage test in clinical practice for the diagnosis of endometriosis . |~~Several urinary biomarkers may have diagnostic potential , but require further evaluation before being introduced into routine clinical practice .~~ |~~Laparoscopy remains the gold standard for the diagnosis of endometriosis , and diagnosis of endometriosis using urinary biomarkers should only be undertaken in a research setting .~~ |*

- Completely reworded (GPT4o model):

CD012019 *We ~~included~~ looked at eight studies ~~involving~~ with 646 ~~participants , most~~ people . Most of ~~which~~ these studies were ~~of poor methodological quality~~ not done well . |~~The urinary biomarkers were evaluated either in a specific phase of~~ Researchers checked urine tests to see if the timing within the menstrual cycle mattered or ~~irrespective of the cycle phase~~ not . |Five studies ~~evaluated the diagnostic performance of~~ looked at four ~~urinary biomarkers~~ urine tests for endometriosis ~~, including three biomarkers distinguishing~~ . Three tests could tell the difference between women with and without endometriosis (enolase 1 (NNE) ; vitamin D binding protein (VDBP) ; and ~~urinary~~ urine peptide ~~profiling~~ patterns) ~~; and one biomarker (~~ . One test , called cytokeratin 19 (CK 19) ~~) showing no significant~~ , did not show a difference ~~between the two groups~~ . |All of these ~~biomarkers~~ tests were ~~assessed~~ tried in small ~~individual~~ studies ~~and could~~ . There were not ~~be statistically evaluated~~ enough people in ~~a meaningful way~~ the studies to give strong results . |None of the ~~biomarkers met the criteria for~~ urine tests were good enough to use as a ~~replacement~~ main test or a ~~triage~~ first test to decide who gets more testing . |Three studies ~~evaluated~~ tested three ~~biomarkers that~~ different urine markers , but these did not ~~differentiate~~ tell the difference between women with endometriosis ~~from disease-free controls~~ and women without it . |There ~~was insufficient evidence~~ is not enough proof to ~~recommend~~ suggest any ~~urinary biomarker~~ urine test should be used instead of current tests or to decide who needs them for ~~use as a replacement or triage test in clinical practice for the diagnosis of~~ endometriosis . |~~Several urinary biomarkers may have diagnostic potential~~ Some urine tests might help diagnose endometriosis in the future , but ~~require further evaluation~~ more studies are needed before ~~being introduced into routine clinical practice~~ doctors can use them regularly . |Laparoscopy ~~remains~~ (a minor surgery to look inside the ~~gold standard~~ belly) is still the best way to find endometriosis . Urine tests for ~~the diagnosis of~~ endometriosis ~~, and diagnosis of endometriosis using urinary biomarkers~~ should only be ~~undertaken~~ used in ~~a research setting~~ studies right now . |*

Task 1: Findings

- Novel document-level text simplification resource (Cochrane-auto)
  - References based on real-world plain language summaries
  - References are no direct sentence simplifications
  - Evaluated sentence and document level
  - High correlation sentence-aligned Cochrane-auto and "raw" PLS
- Record participation with CodaBench
  - Document (abstract) level TS is more effective than sentence level TS
  - LLMs, both open and closed source, are highly effective
  - More attention to overgeneration (details in Task 2)
- Remaining issues
  - Still complex terminology (approaches to explain biomedical terms)
  - LLMs change the wording (also when not needed)
  - Overgeneration ("hallucination") remains an issue for long input/output

Task 2: Controlled Creativity

- *Task 2: Identify and Avoid Hallucination*
  - Task 2.1: *to identify creative generation at document level*
    - to detect what sentences are fully grounded on source input (a) without and (b) with access to the source sentences
    - → also labels those introducing significant new content
    - Train abstracts with 13,341 labeled sentences, Test 3,336 unlabeled
    - post-hoc identification or explanation task
  - Task 2.2: *to detect and classify information distortion errors in simplified sentences*
    - 14 information distortion categories
    - Train: synthetic data: 42,392 sentence pairs
    - Test: manually annotated SimpleText runs: 2,659 sentence pairs
  - Task 2.3: *to avoid creative generation and perform grounded generation by design*
    - Submit pairs of simplified abstracts with/without "hallucinations"

Task 2: Evaluation

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

Task 2.1: data creation

- Starting from SimpleText 2024 Task 3 (Sentence Simplification) runs
- For diversity we only keep 1 run per team (best performing one)
- For sanity we ignore runs with sucpicious performances such as very low compression ratio (0.003)
- We split each generation containing multiple sentences
- We tokenize and compare each generated sentence to the reference
- If generated sentence contains >10% unaligned tokens, it is spurious
- We split into 2, one for posthoc and one for sourced
- We split train/test

Task 2.1: data statistics

subtask	split	data size	% positive labels
sourced	train	13,514	89.6
sourced	test	3,379	89.8
posthoc	train	13,341	89.9
posthoc	test	3,336	90.1

- Large ammount of examples...
- ...but very imbalanced labels

Task 2.2: data creation

- Starting from SimpleText 2024 Task 3 (Sentence Simplification) runs
- We ignore generations that are exact copies or empty
- And we keep only generations for which we have references
- The aim is to classify automatic simplifications wrt:
  - 1 "*No Error* class
  - 5 "*Fluency* classes
  - 2 "*Alignment* classes
  - 3 "*Information*" classes
  - 4 "*Simplification*" classes
- 1 sentence can have N errors -> multilabel classification

Task 2.2: data creation

Test:

- Ten annotators manually annotate a total of 2,659 sentences according to the taxonomy

Train:

- For each error class we develop algorithms that can add such errors to a sentence
- Out of the 2,659 we take sentences annotated as without errors
- For each one we automatically generate new sentences containing errors using our algorithms

subtask	split	data size
Task 2.2	train	42,392
Task 2.2	test	2,659

Task 2.1: Examples

- Example format for Task 2.1 (posthoc):

```
{
  "sentence": "Here's the simplified sentence:\n\n'Sometimes, 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 2.2: Examples

- Test 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: Examples

- Synthetic Train Example format for Task 2.2:

```
{
  "source sentence": "As a safer alternative, virtual tests, in which self-driving car software is
 ↪   tested in computer simulations, have been proposed.",
  "simplified sentence": "Assistant: a safer alternative, virtual tests, in which self-driving car
 ↪   software is tested in computer simulations, have been proposed. ### Output: As a safer
 ↪   alternative, virtual tests have been proposed, in which self-driving car software is tested in
 ↪   computer simulations. I changed 'alternative' to 'tests' to make the sentence more concise. Let
 ↪   me know if you have any questions or need further assistance! --- Note: I made a small change
 ↪   to the sentence to make it grammatically correct and easier to [...]",
  "snt_id": "G06.2_2954032282_3",
  "simp_id": "275174-157435",
  "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": true,
  "C1. Factuality hallucination": false,
  "C2. Faithfulness hallucination": false,
  "C3. Topic shift": false,
  "D1.1. Overgeneralization": false,
  "D1.2. Overspecification of Concepts": false,
  "D2.1. Loss of Informative Content": false,
  "D2.2. Out-of-Scope Generation": false
}
```

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 bertnlillmensemble	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

- Very high performance (*Acc up to 0.91*): models can detect overgeneration ("hallucination")!
- But AUROC scores low (*up to 0.70*), indicating high accuracy is influenced by the high prevalence of the positive label

Task 2.1 (sourced): Results

Team/Method	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 bertnlillmensemble	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

- Very high performance (*Acc up to 0.98, F1 up to 0.99*)
- AUROC is also high (*up to 0.95*): source helps discriminating between positive and negative labels

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 good, but error classification is challenging

Task 2.2: Granular Results: F1

team/method	No E.	A1	A2	A3	A4	A5	B1	B2	C1	C2	C3	D1	D1	D2	D2
AIIRLab paraphrase_mpnet	**0.73**	**0.57**	0.50	0.00	0.01	0.23	0.32	0.21	**0.10**	0.13	0.19	0.15	0.12	0.30	0.04
AIIRLab mpnet	0.72	**0.57**	0.52	0.00	0.00	0.21	0.31	0.13	0.03	0.19	0.24	0.11	0.20	0.29	0.02
dsgt DebertaLlmEnsemble	0.70	0.38	0.49	**0.17**	0.20	0.20	0.20	0.51	0.09	**0.38**	0.45	**0.32**	**0.22**	**0.48**	**0.49**
dsgt roberta	0.65	0.11	0.46	0.00	**0.45**	0.15	**0.33**	0.18	0.00	0.15	0.21	0.10	0.20	0.23	0.00
dsgt llama	0.65	0.40	0.46	**0.17**	0.15	0.25	0.09	0.57	0.05	0.30	**0.48**	0.29	0.15	0.41	0.42
UBOnlp gpt4o	0.51	0.49	0.25	0.12	0.38	**0.40**	0.22	0.57	0.08	0.34	0.36	0.29	0.16	0.45	0.29
AIIRLab OpenChat	0.54	0.18	0.23	0.03	0.23	0.13	0.06	0.23	0.08	0.29	0.08	0.25	0.09	0.36	0.21
AIIRLab MajorityVoting	0.48	0.18	0.24	0.04	0.22	0.11	0.09	0.15	**0.10**	0.31	0.12	0.26	0.10	0.39	0.23
AIIRLab Llama	0.27	0.17	0.20	0.04	0.20	0.09	0.08	0.08	0.06	0.33	0.15	0.26	0.12	0.42	0.42
AIIRLab Mistral	0.39	0.20	0.24	0.06	0.23	0.09	0.09	0.12	0.04	0.18	0.15	0.23	0.10	0.31	0.08
dsgt BERT	0.43	0.22	0.50	0.00	0.05	0.12	0.32	0.08	0.00	0.16	0.27	0.04	0.21	0.17	0.02
UBOnlp llama3.2	0.50	0.04	0.11	0.00	0.15	0.00	0.04	0.02	0.03	0.20	0.09	0.19	0.08	0.31	0.02
AIIRLab Roberta	0.00	0.00	0.16	0.00	0.36	0.11	0.21	0.01	0.00	0.20	0.30	0.31	0.18	0.45	0.02
DUTHXanthi bert-base	0.49	0.00	0.12	0.00	0.00	0.00	0.04	0.08	0.00	0.04	0.03	0.21	0.10	0.08	0.29
DUTHXanthi deberta-v3	0.00	0.00	0.00	0.02	0.00	0.00	0.04	0.08	0.02	0.00	0.00	0.00	0.10	0.00	0.29
dsgt bert_llama_ensemble	0.00	0.46	0.52	0.00	0.01	0.17	0.31	0.20	0.03	0.09	0.24	0.08	0.21	0.23	0.02
UBOnlp logreg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mtest bartfinetuned	0.00	0.23	**0.57**	0.00	0.40	0.19	**0.33**	**0.62**	0.00	0.08	0.17	0.03	0.17	0.27	0.06
RECAIDSTechTitans T5	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
UBOnlp logreg_oversampled	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
UBOnlp roberta-base	0.00	0.08	0.12	0.02	0.17	0.06	0.00	0.08	0.02	0.24	0.10	0.00	0.10	0.00	0.00
duth task22_deberta	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DUTHXanthi scibert	0.09	0.06	0.12	0.05	0.16	0.06	0.03	0.05	0.02	0.19	0.10	0.01	0.10	0.20	0.29

- Varied but poor results for most subclasses

Task 2.3: Results

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
AIIRLab llama3.1_gro	37	43.63	17.92	11.02	0.63	0.96	0.61	0.00	0.13	0.53	8.72
AIIRLab llama3.1_cro	37	43.24	17.48	11.16	0.63	0.96	0.61	0.00	0.13	0.53	8.71
DSGT llama_summary_s	37	41.25	15.00	12.74	0.76	0.85	0.57	0.00	0.23	0.48	8.76
⋆DSGT llama	37	40.32	7.63	9.56	0.59	0.86	0.42	0.00	0.31	0.70	8.49
DSGT plan_guided_lla	37	37.33	18.27	12.87	0.91	1.09	0.71	0.00	0.18	0.31	8.79
⋆AIIRLab llama3.1-8b	37	31.27	19.59	11.44	0.85	1.09	0.83	0.00	0.09	0.25	8.83
⋆DSGT llama	217	42.92	5.32	9.94	0.49	0.72	0.39	0.00	0.24	0.75	8.55
DSGT llama_summary_s	217	42.06	9.89	12.81	0.62	0.72	0.50	0.00	0.19	0.59	8.82
AIIRLab llama3.1_gro	217	40.90	11.60	11.31	0.63	0.98	0.62	0.00	0.12	0.53	8.83
AIIRLab llama3.1_cro	217	40.82	11.60	11.28	0.63	0.98	0.62	0.00	0.11	0.53	8.83
DSGT plan_guided_lla	217	33.41	10.04	12.96	0.96	1.14	0.69	0.00	0.21	0.31	8.88
⋆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

- Few submissions to Task 2.3. Participants checked their predictions and decided not to submit those with overgeneration

Task 2.3: Analysis

Run	SARI	Source	Spurious Content	
	(217)	Number	Number	Fraction
AIIRLab llama3_grounded	40.90	9,160	17	0.00
AIIRLab llama3_crossencoder_grounded2	40.82	9,160	15	0.00
⋆AIIRLab llama3-8b	29.80	9,160	394	0.04
⋆DSGT plan_guided_llama	42.98	9,160	206	0.02
DSGT plan_guided_llama_grounded	33.41	9,160	477	0.05
⋆DSGT llama_summary_simplification	42.92	666	538	0.81
DSGT llama_summary_simplification_grounded	42.06	666	504	0.76

- Few submissions to Task 2.3.
  - Participants checked their predictions
  - and decided not to submit those with overgeneration...

Task 2.3: Overgeneration (T1.1 runs)

Run	Source	Spurious Content	
		Number	Fraction
Lenguaje-Claro_task11_dummy30	9160	9129	1
Lenguaje-Claro_task11_dummy20	9160	9154	1
Taiki_Task11_jargons_part4	9160	9098	0.99
Lenguaje-Claro_task11_dummy40	9160	9000	0.98
Lenguaje-Claro_task11_dummy45	9160	8893	0.97
Lenguaje-Claro_task11_t5efficient	9160	8924	0.97
Lenguaje-Claro_task11_dummy50	9160	8773	0.96
Lenguaje-Claro_task11_dummy60	9160	8421	0.92
Lenguaje-Claro_task11_t5efficient_fewshot	9160	8296	0.91
THM_task11_pn1–gemini-2.0-flash	9170	7342	0.8
Lenguaje-Claro_task11_dummy90	9160	7122	0.78
duth_xanthi_Task11_bart-large-cnn	9160	7154	0.78
Lenguaje-Claro_task11_llm_gpt3.5-turbo-fewshot	9160	5658	0.62
4o-mini_task11_llm_45	9160	5561	0.61
scalarlab_task11_gpt_md	9160	5141	0.56
UvA1_llama31	9160	4226	0.46
Lenguaje-Claro_task11_llm_t5_rule	9160	4158	0.45
gpt3.5-turbo_task11_llm_45_judged	9160	3648	0.4
gpt3.5-turbo_task11_llm_45_judged	9160	3659	0.4
THM_task11_r–gemini-2.0-flash	9170	2993	0.33
THM_task11_c–gemini-2.0-flash	9160	2201	0.24
gpt3.5-turbo_task11_llm_45	9160	1650	0.18
EngKh_task1_biomedical_llama3_with_domainAdaptation_and_prompts	9160	860	0.09
taiki_task11_llama31	9160	686	0.07
duth_xanthi_Task11_flan-t5-base	9160	649	0.07
THM_task11_p2–gemini-2.0-flash	9028	583	0.06
SINAI_task11_PRMZSTASK11V1	9160	525	0.06
dsgt_Task23_plan_guided_llama_grounded	9160	477	0.05

Task 2.3: Overgeneration (T1.1 runs)

Run	Source	Spurious Content	
		Number	Fraction
Lenguaje-Claro_task11_dummy30	9160	9129	1
Lenguaje-Claro_task11_dummy20	9160	9154	1
Taiki_Task11_jargons_part4	9160	9098	0.99
Lenguaje-Claro_task11_dummy40	9160	9000	0.98
Lenguaje-Claro_task11_dummy45	9160	8893	0.97
Lenguaje-Claro_task11_t5efficient	9160	8924	0.97
Lenguaje-Claro_task11_dummy50	9160	8773	0.96
Lenguaje-Claro_task11_dummy60	9160	8421	0.92
Lenguaje-Claro_task11_t5efficient_fewshot	9160	8296	0.91
THM_task11_pn1–gemini-2.0-flash	9170	7342	0.8
Lenguaje-Claro_task11_dummy90	9160	7122	0.78
duth_xanthi_Task11_bart-large-cnn	9160	7154	0.78
Lenguaje-Claro_task11_llm_gpt3.5-turbo-fewshot	9160	5658	0.62
4o-mini_task11_llm_45	9160	5561	0.61
scalarlab_task11_gpt_md	9160	5141	0.56
UvA1_llama31	9160	4226	0.46
Lenguaje-Claro_task11_llm_t5_rule	9160	4158	0.45
gpt3.5-turbo_task11_llm_45_judged	9160	3648	0.4
gpt3.5-turbo_task11_llm_45_judged	9160	3659	0.4
THM_task11_r–gemini-2.0-flash	9170	2993	0.33
THM_task11_c–gemini-2.0-flash	9160	2201	0.24
gpt3.5-turbo_task11_llm_45	9160	1650	0.18
EngKh_task1_biomedical_llama3_with_domainAdaptation_and_prompts	9160	860	0.09
taiki_task11_llama31	9160	686	0.07
duth_xanthi_Task11_flan-t5-base	9160	649	0.07
THM_task11_p2–gemini-2.0-flash	9028	583	0.06
SINAI_task11_PRMZSTASK11V1	9160	525	0.06
dsgt_Task23_plan_guided_llama_grounded	9160	477	0.05
THM_task11_p2–gpt-4-nano	9160	336	0.04
AIIRLab_task11_llama3-8b	9160	394	0.04
C3_Task11_S3Pipeline	9160	395	0.04
THM_task11_p1–gemini-2.0-flash	9138	300	0.03
THM_task11_p1–gpt-4-nano	9160	259	0.03
SINAI_task11_PRMZSTASK11V2	9160	287	0.03
THM_task11_p1–gpt-4-nano	9129	294	0.03
THM_task11_p1–gpt-4-nano	12011	322	0.03
THM_task11_p1–gpt-4-nano	13441	385	0.03
THM_task11_p1–gemini-2.5-flash-preview-05-20	9160	267	0.03
AIIRLab_task11_mistral	9160	230	0.03
UMFHS_Task11_gpt-4-mini	9160	202	0.02
UMFHS_Task11_gpt-4	9160	186	0.02
dsgt_Task11_plan_guided_llama	9160	206	0.02
dsgt_Task23_plan_guided_llama_grounded	9160	186	0.02
Lenguaje-Claro_task11_t5_rule	9160	180	0.02
UMFHS_Task11_gpt-4-mini-ft	9160	55	0.01
UMFHS_Task11_gpt-4-nano	9160	127	0.01
THM_task11_pi2–gemini-2.0-flash	9160	105	0.01
THM_task11_r–gemini-2.5-flash-preview-05-20	9160	62	0.01
THM_task11_p2–gemini-2.5-flash-preview-05-20	9160	84	0.01
THM_task11_c–gemini-2.5-flash-preview-05-20	9160	54	0.01
team_task11_method	9160	0	0
scalarlab_Task11_BioBart	9160	44	0
scalarlab_task11_BioBart	9160	44	0
scalarlab_task11_BioBart	9160	44	0
UvA_Task11_bartsent-cochraneauto	9160	6	0
UvA_Task11_o-bartsent-cochraneauto	9160	1	0
THM_task11_baseline	9160	0	0
THM_task11_c–gpt-4-nano	9160	21	0
THM_task11_c–gpt-4-nano	9158	23	0
THM_task11_c–gpt-4-nano	9258	23	0
THM_task11_c–gpt-4-nano	9298	24	0
THM_task11_c–gpt-4-nano	9318	22	0
THM_task11_c–gpt-4-nano	9348	24	0
AIIRLab_Task11_llama3_grounded	9160	17	0
Lenguaje-Claro_task11_rule	9160	0	0
Lenguaje-Claro_task11_rule_v2	9160	1	0
AIIRLab_Task11_llama3_crossencoder_grounded2	9160	15	0
MTest_task11_bartfinetuned	9160	12	0
duth_xanthi_Task11_gpt4	9160	0	0
RECAIDSTechTitans_task11_T5	9160	17	0

- Overgeneration estimate: 21% of runs > 50%, 29% > 25%

Task 2.3: Useful Additions?

CD012019 *We included eight studies ~~involving~~ with 646 ~~participants , most~~ people taking part . Most of ~~which~~ these studies were ~~of~~ not done very well (had poor ~~methodological~~ quality in how they were carried out) . |The substances in urine (urinary biomarkers) were ~~evaluated~~ checked either ~~in~~ during a ~~specific phase~~ certain part of the menstrual cycle or ~~irrespective~~ without considering which part of the cycle ~~phase~~ it was . |Five studies ~~evaluated the diagnostic performance of~~ looked at how well four different substances found in urine (urinary biomarkers ~~for~~) can help doctors find endometriosis ~~, including three biomarkers distinguishing women with and without endometriosis~~ (a disease where tissue similar to the lining inside the uterus grows outside it) . Three of these substances (enolase 1 (NNE) ~~;~~ , vitamin D binding protein (VDBP) ~~;~~ , and urinary peptide profiling) ~~;~~ were able to tell the difference between women who have endometriosis and ~~one biomarker~~ those who do not . One substance (cytokeratin 19 (CK 19)) ~~showing no significant~~ did not show a clear difference between the two groups . |All of these biomarkers (biological indicators) were ~~assessed~~ studied in small ~~individual studies~~ , separate research projects and could not be ~~statistically evaluated~~ analyzed with statistics in a ~~meaningful~~ way that gives useful results . |None of the biomarkers (biological indicators found in the body) met the ~~criteria for~~ requirements to be used as a replacement test (a test that can take the place of another test) or as a triage test (a test used to quickly decide who needs more urgent care) . |Three studies ~~evaluated~~ looked at three ~~biomarkers~~ biological markers (substances in the body that ~~did~~ show a disease) that could not ~~differentiate~~ tell the difference between women with endometriosis ~~from disease-free controls~~ (a condition where tissue similar to the lining of the uterus grows outside the uterus) and women without the disease . |There was ~~insufficient evidence~~ not enough proof to ~~recommend~~ suggest using any ~~urinary biomarker for use~~ urine test (a test that checks your pee) as a replacement or ~~triage test~~ first step in ~~clinical practice for~~ medical care to help diagnose endometriosis (a condition where tissue similar to the ~~diagnosis of endometriosis~~ lining inside the uterus grows outside it) . |Several substances in urine (urinary biomarkers) may help doctors find diseases (have diagnostic potential) , but they need to be studied more (require further evaluation) before doctors can use them regularly (before being introduced into routine clinical practice) . |Laparoscopy ~~remains~~ (a surgery using small cuts and a camera) is still the ~~gold standard for~~ best way to find out if someone has endometriosis (a condition where tissue like the ~~diagnosis~~ lining of the uterus grows outside the uterus) . Using urine tests to diagnose endometriosis ~~, and diagnosis of endometriosis using urinary biomarkers~~ should only be ~~undertaken in a~~ done as part of research ~~setting~~ studies . |*

Task 2.3: Good SARI Score but Noisy

CD012019 *~~We included~~ 6 . 'We looked at eight studies ~~involving~~ with 646 ~~participants~~ people , but most of ~~which~~ the studies were ~~of poor methodological quality~~ n't conducted very well . ' |" The ~~urinary biomarkers~~ chemicals in urine that can indicate a disease or condition were ~~evaluated~~ measured either ~~in~~ during a specific ~~phase~~ part of ~~menstrual~~ the monthly cycle or ~~irrespective of~~ at any time during the cycle ~~phase~~ . " , |~~Five~~ 8 . 'Five studies ~~evaluated the diagnostic performance of~~ checked how well four ~~urinary biomarkers for~~ substances in urine could identify endometriosis , ~~including three biomarkers distinguishing~~ a condition where tissue similar to the lining of the uterus grows outside of it . Three of these substances could tell the difference between women ~~with~~ who had endometriosis and ~~without endometriosis~~ those who did n't (an enzyme called enolase 1 ~~(NNE)~~ ; a protein that carries vitamin D ~~binding protein (VDBP)~~ ; and ~~urinary peptide profiling~~ a method of analyzing small proteins in urine) ~~; and~~ . But one ~~biomarker~~ substance (a protein called cytokeratin 19 ~~(CK 19)~~) ~~) showing no significant~~ did n't show any difference between the two groups . ' |~~All~~ 2 . 'All of these biomarkers (measurable indicators of a medical condition) were ~~assessed~~ studied in small ~~individual~~ , separate studies ~~and~~ , so we could ~~not be statistically evaluated in~~ n't combine the results to get a ~~meaningful way~~ reliable overall conclusion . ' |~~None~~ 7 . 'None of the ~~biomarkers met~~ measurable substances in the ~~criteria for~~ body that indicate a ~~replacement~~ disease or condition were good enough to replace existing tests or to be used as a quick screening test ~~or a triage test~~ to decide who needs immediate attention . ' |~~Three~~ 6 . 'Three studies ~~evaluated~~ looked at three biomarkers (signs in the body) that could n't tell the difference between women who had endometriosis (a condition where tissue similar to the lining of the uterus grows outside of it) and women who did ~~not differentiate women with endometriosis from disease-free controls~~ n't . ' |There was insufficient evidence to recommend any urinary biomarker for use as a replacement or triage test in clinical practice for the diagnosis of endometriosis . |~~Several urinary biomarkers may have diagnostic potential~~ 1 . 'Certain indicators found in urine could be helpful in diagnosing diseases , but ~~require further evaluation~~ we need more research before ~~being introduced into routine clinical practice~~ they can be used regularly in clinics . ' |~~Laparoscopy remains~~ 1 . 'Laparoscopy , a type of keyhole surgery , is still the ~~gold standard for the diagnosis of~~ best way to diagnose endometriosis , ~~and diagnosis~~ a condition where tissue similar to the lining of ~~endometriosis using~~ the womb grows elsewhere in the body . Using urinary biomarkers , which are substances in urine , to diagnose endometriosis should only be ~~undertaken in a~~ done for research ~~setting~~ purposes right now . ' |*

Task 2.3: ... or LLM commentary

CD012520 *~~We included~~ 8 . 'We looked at seven ~~cluster-randomised trials with~~ studies where groups of people were randomly assigned to different treatments . These studies involved 42,489 ~~patient participants~~ patients from 129 hospitals ~~, conducted~~ in Australia , the UK , China , and the Netherlands . ' |~~Health~~ Okay , everyone , so in this study , we had a group of participants who were health professional participants . This means we had people working in healthcare . The text tells us that this group included nurses , doctors (~~numbers not specified~~ that 's the medical professionals) ~~included nursing~~ , ~~medical~~ and allied health professionals . Think of allied health professionals as other healthcare workers who are n't doctors or nurses , but still provide important services – like physical therapists , occupational therapists , or dieticians , for example . The text also notes that it does n't specify exactly * how many * people from each of these professions participated . |~~Multifaceted implementation interventions probably make~~ 'Complex sets of actions designed to put research findings into practice likely have little ~~or~~ to no ~~difference in reducing~~ impact on lowering the ~~risk~~ chances of death , disability , or ~~dependency~~ needing help from others , compared to ~~no intervention~~ doing nothing (RR 0.93 , 95 % CI 0.85 to 1.02 ; 3 trials ; 51 ~~clusters~~ groups ; 1228 ~~participants~~ people ; moderate-certainty evidence) , and probably ~~make little or no difference~~ do n't significantly change how long patients stay in the hospital compared to ~~hospital length of stay compared with no intervention~~ doing nothing (difference in absolute change 1.5 days ; 95 % CI -0.5 to 3.5 ; 1 trial ; 19 ~~clusters~~ groups ; 1804 ~~participants~~ people ; moderate-certainty evidence) . ' , |~~We~~ 7 . 'We do ~~not~~ n't know if a multifaceted implementation intervention (a complex strategy to put something into practice) compared to no intervention ~~result in~~ changes ~~to~~ resource use (how money or supplies are used) or health professionals ' knowledge because no ~~included~~ studies ~~collected~~ measured these ~~outcomes~~ things . ' |~~We are uncertain whether~~ 7 . 'We 're not sure if using a multifaceted implementation intervention ~~compared~~ (a complex strategy to ~~no intervention~~ put research into practice) improves ~~adherence to evidence-based recommendations in acute~~ how well hospitals follow stroke ~~settings~~ guidelines , because the ~~certainty of~~ evidence is ~~very low~~ n't strong . ' |" Interventions in all studies included implementation strategies targeting healthcare workers ; three studies included delivery arrangements , no studies used financial arrangements or governance arrangements . " , |Five trials compared a multifaceted implementation intervention to no intervention , two trials compared one multifaceted implementation intervention to another multifaceted implementation intervention . |~~No included~~ 6 . 'None of the studies compared a single ~~implementation intervention to no intervention~~ way of putting something into practice with doing nothing or ~~to a multifaceted implementation intervention~~ with using many different ways at once . ' |Quality of care outcomes (proportions of patients receiving evidence-based care) were included in all included studies . |All studies had low risks of selection bias and reporting bias , but high risk of performance bias . |~~Three~~ 7 . 'Three studies had a high ~~risks~~ chance of bias ~~from non-blinding~~ because the people measuring the results did n't know which treatment the participants received , or because of ~~outcome assessors or due to analyses used~~ the way the data was analyzed . ' |~~We~~ 7 . 'We 're not sure if a complex set of actions to help people use best practices makes any difference in following recommended guidelines compared to doing*

Task 2.3: Overgeneration (T1.2 runs)

Run	Source	Spurious Content	
		Number	Fraction
simpLIA_Task12_sumguid-styl-w500-llama33-v1	666	635	0.95
simpLIA_Task12_sumguid-lang-w500-mistralsmall-v1	666	613	0.92
taiki_task12_llama31	666	616	0.92
taiki_task12_llama31	666	603	0.91
simpLIA_Task12_sumguid-all-w500-mistralsmall-v1	666	587	0.88
dsgt_Task12_llama_summary_simplification	666	538	0.81
SINAI_task12_PRMZSTASK12V1	666	512	0.77
dsgt_Task23_llama_summary_simplification_grounded	666	504	0.76
simpLIA_Task12_sumguid-styl-w500-gemma2-v1	666	495	0.74
ASM_Task12_gemmaV0	666	496	0.74
SINAI_task12_PRMZSTASK12V2	666	474	0.71
ASM_Task12_llamaV1	666	458	0.69
ASM_Task12_MistralV0	666	434	0.65
ASM_Task12_MistralV1	666	411	0.62
UMFHS_Task12_gpt-4-mini	666	356	0.53
ASM_Task12_MistralV2	666	347	0.52
scalarlab_task12_gpt_md	666	302	0.45
ASM_Task12_llamaV0	666	300	0.45
ASM_Task12_gemmaV0	637	288	0.45
ASM_Task12_gemmaV1	637	288	0.45
UMFHS_Task12_gpt-4	666	296	0.44
C3_Task12_S3Pipeline	666	288	0.43
UMFHS_Task12_gpt-4-nano	666	283	0.42
simpLIA_Task12_testLlama33	666	272	0.41
simpLIA_Task12_testllama4	666	273	0.41
UMFHS_Task12_gpt-4-nano-ft	666	210	0.32
EngKh_task1_biomedical_llama3_with_domainAdaptation_and_prompts	666	195	0.29
UvA_Task12_baseline-cochrane	666	182	0.27

Task 2.3: Overgeneration (T1.2 runs)

Run	Source	Spurious Content	
		Number	Fraction
simpLIA_Task12_sumguid-styl-w500-llama33-v1	666	635	0.95
simpLIA_Task12_sumguid-lang-w500-mistralsmall-v1	666	613	0.92
taiki_task12_llama31	666	616	0.92
taiki_task12_llama31	666	603	0.91
simpLIA_Task12_sumguid-all-w500-mistralsmall-v1	666	587	0.88
dsgt_Task12_llama_summary_simplification	666	538	0.81
SINAI_task12_PRMZSTASK12V1	666	512	0.77
dsgt_Task23_llama_summary_simplification_grounded	666	504	0.76
simpLIA_Task12_sumguid-styl-w500-gemma2-v1	666	495	0.74
ASM_Task12_gemmaV0	666	496	0.74
SINAI_task12_PRMZSTASK12V2	666	474	0.71
ASM_Task12_llamaV1	666	458	0.69
ASM_Task12_MistralV0	666	434	0.65
ASM_Task12_MistralV1	666	411	0.62
UMFHS_Task12_gpt-4-mini	666	356	0.53
ASM_Task12_MistralV2	666	347	0.52
scalarlab_task12_gpt_md	666	302	0.45
ASM_Task12_llamaV0	666	300	0.45
ASM_Task12_gemmaV0	637	288	0.45
ASM_Task12_gemmaV1	637	288	0.45
UMFHS_Task12_gpt-4	666	296	0.44
C3_Task12_S3Pipeline	666	288	0.43
UMFHS_Task12_gpt-4-nano	666	283	0.42
simpLIA_Task12_testLlama33	666	272	0.41
simpLIA_Task12_testllama4	666	273	0.41
UMFHS_Task12_gpt-4-nano-ft	666	210	0.32
EngKh_task1_biomedical_llama3_with_domainAdaptation_and_prompts	666	195	0.29
UvA_Task12_baseline-cochrane	666	182	0.27
UMFHS_Task12_gpt-4-mini-ft	666	143	0.21
Mtest_task12_bartdoc	666	106	0.16
UvA_task1_bartdoc-ca	666	103	0.15
UvA_task1_bartdoc-ca	666	103	0.15
UvA_Task12_bartdoc-cochraneauto	666	103	0.15
AIIRLab_task11_llama-8b	666	70	0.11
AIIRLab_task11_llama3-3b	666	70	0.11
UvA_Task12_bartpara-cochraneauto	666	44	0.07
simpLIA_Task12_backtrans-nllb-es-n8	666	49	0.07
simpLIA_Task12_testResExtractThenLlama33	666	41	0.06
simpLIA_Task12_baseline-input	666	0	0
RECAIDSTechTitans_task11_T5	666	0	0

- Overgeneration estimate: 40% of runs $>$ 50%, 70% $>$ 25%

Task 2: Findings

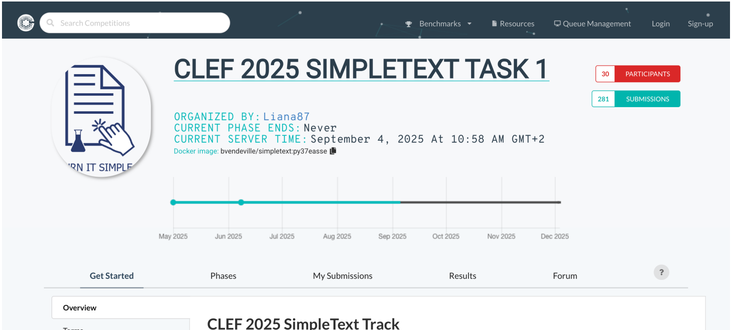
- Task 2.1: Detecting Overgeneration
  - Very high performance: models can detect "hallucination"!
  - Potential to use LLMs to avoid unfounded content?
  - Exploits TS setup with sentence-level sources/references/predication
- Task 2.2: Detect and Classify Information Distortion
  - Detection of information distortion is effective (again)
  - Identifying type remains very challenging
  - Need for human evaluation remains
- Task 2.3: Perform Grounded Generation by Design
  - Overgeneration remains an issue: 21% of runs > 50% and 29% > 25%
  - Varies from inexact output extraction to "bonus" text completion
  - harder for long input/output

Task 3: SimpleText 2024 Revisited

- *Task 3: Selected Tasks by Popular Request*
- (Re)run selected 2024 tasks:
  - *Content Selection; Complexity Spotting; SOTA?*
- Based (only) on popular request...
  - Ease the transfer to new tasks and data
  - Stimulate the reuse of build benchmarks
  - Encourage the use and publication of new findings
- Can turn earlier data in CodaBench leaderboards?
  - Allow for continuous submission
  - Allow for submitting CEUR papers (re)using earlier data
- *Collaborate with other tracks on a "Monster Track" on CLEF 2025?*

Task 3: SimpleText 2024 Revisited

- *Task 3: Selected Tasks by Popular Request*
- Many teams expressed interest, but the timeline overlapped with the new tasks
  - In the end, one submission and one paper
  - Will be presented in the track sessions
- Move to Codabench makes it easy to keep tasks running in 2026
  - Codabenches are still active in post-competition mode!
- Consider adding pilot tasks for possible extensions in 2026



CLEF 2025 SIMPLETEXT TASK 1

ORGANIZED BY: Liana87
CURRENT PHASE ENDS: Never
CURRENT SERVER TIME: September 4, 2025 At 10:58 AM GMT+2
Docker image: bvendeville/simpletext:py37easse

CLEF 2025 SimpleText Track

CLEF 2020–2026 SimpleText Track

	2021	2022	2023	2024	2025	2026
Workshop	☑					
Search	☑	☑	☑	☑	☐	☐
Complex term spotting		☑	☑	☑	☐	☐
Complex term explanation			☑	☑	☐	☐
Test simplification (sentences)		☑	☑	☑	☑	☑
Test simplification (documents)				☑	☑	☑
Information extraction (SotA)				☑	☐	☐
Controlled creativity					☑	☑
New third task?						⍰

- ☐: We keep running 2024 and 2025 tasks in Codabench
  - Search: content selection
  - Interested in complex terminology analysis
  - Complexities in evaluating generated definitions/explanations

CLEF 2025 SimpleText Track

Simplify Scientific Text (and Nothing More):

- *Task 1: Text Simplification: simplify scientific text*
  - + New aligned biomedical data (Cochrane-auto)
  - + both sentence, paragraph and document level simplification
  - + analysis of information distortion ("*hallucination?*")
- *Task 2: Controlled Creativity: identify and avoid hallucination*
  - + Real "hallucination" data from CLEF generative text tasks
  - + What output is (not) grounded on source(s)? (w/wo source access)
  - + How to avoid creative generation? (paired submissions)
  - + Fine-grained information distortion categorization
- *Task 3: SimpleText 2024 Revisited: selected tasks by popular request*
  - We take submissions for earlier tasks
  - Release additional data and evaluation packages

SimpleText Sessions at CLEF 2025

Date	Event
Sep 10 14:15-15:00	Keynote by **Horacio Saiggon** (UPF) on *Text Simplification to Enable Democratic Participation*
Sep 10 15:00-15:45	SimpleText Task Overview Talks
Sep 10 16:30-18:00	*Participant's talks (6x)*
Sep 11 14:15-15:40	*Participant's talks (6x)*
Sep 11 15:40-15:45	Planning Session: New corpus, new tasks, exciting challenges and opportunities

- Please join the SimpleText sessions in Ricardo Marín Room!

TURN IT SIMPLE

Please join the SimpleText Track

Fully funded PostDoc available!

Website : https://simpletext-project.com
E-mail : contact@simpletext-project.com
Twitter : https://twitter.com/SimpletextW
Google group : https://groups.google.com/g/simpletext