University of Amsterdam at the CLEF 2025 SimpleText Track

Taiki Papandreou, **Jan Bakker**, Jaap Kamps

Submissions

Task	Run	Description
1.1	bartsent-cochraneauto	Sentence-level BART model trained on Cochrane-auto
1.1	o-bartsent-cochraneauto	Plan-guided sentence-level BART model trained on Cochrane-auto
1.1	llama31	Jargon aware prompt with llama3.1-8B-instruct model (sentence level)
1.2	bartpara-cochraneauto	Paragraph-level BART model trained on Cochrane-auto
1.2	bartdoc-cochraneauto	Document-level BART model trained on Cochrane-auto
1.2	baseline-cochrane	Document-level BART baseline trained on Cochrane corpus
1.2	llama31	Jargon aware prompt with llama3.1-8B-instruct model (document level)

Jargon detection

MedReadMe dataset (Jiang & Xu, EMNLP 2024)

Contains **2,587** annotated sentences with a total of **5,207** jargon terms, sourced from **15** established medical simplification resources (PLOS, eLife, Cochrane, NIHR)

We trained a RoBERTa-large model for binary jargon span detection.

MedReadMe: A Systematic Study for Fine-grained Sentence Readability in Medical Domain (Jiang & Xu, EMNLP 2024)

Example

The long-acting bronchodilator tiotropium and single-inhaler combination therapy of inhaled corticosteroids and long-acting beta 2-agonists (ICS/LABA) are commonly used for maintenance treatment of patients with chronic obstructive pulmonary disease (COPD). Combining these treatments, which have different mechanisms of action, may be more effective than administering the individual components.

Jargon-aware prompting

STRICT_INSTRUCTIONS

Remember: Simplify language but keep ALL medical details accurate.

- Keep exact numbers

- Keep medical patterns (like 'myoclonic' if no exact simple equivalent exists)

Replace these terms ONLY if you can keep their exact medical meaning:

Detected jargon terms listed here

Text to simplify: *Original text*

Write one simplified sentence:

Quantitative results

Run			SARI (auto)	SARI (original)
UvA	task11	bartsent-cochraneauto	41.3	38.7
UvA	task11	o-bartsent-cochraneauto	42.3	38.5
UvA	task11	llama31	40.9	38.7
UvA	task12	bartpara-cochraneauto	35.0	37.9
UvA	task12	bartdoc-cochraneauto	37.3	37.1
UvA	task12	baseline-cochrane	42.1	41.8
UvA	task12	llama31	38.3	35.6

Qualitative results

Complex paragraph

There is some evidence that thrombolysis can probably improve outcomes after aneurysmal subarachnoid haemorrhage, without increasing the risk of haemorrhagic complications. Thrombolysis likely reduces the risk of poor functional outcome and cerebral artery vasospasm, and may reduce the risk of delayed cerebral ischaemia, but it likely makes little to no difference to case fatality or hydrocephalus, and may make little to no difference to the risk of cerebral infarction. However, the current evidence is still uncertain. The uncertainty is primarily due to the small total number of participants and outcome events. Data from further studies are required to confirm the efficacy of thrombolysis for improving outcomes after aneurysmal subarachnoid haemorrhage.

Simplification by jargon-aware Llama model

Some studies suggest that a treatment called thrombolytic therapy may help patients recover better from bleeding in the brain caused by a ruptured blood vessel, without causing further bleeding problems. This treatment might lower chances of: * Poor recovery * Blood vessel narrowing in the brain (after a stroke) * Brain damage from lack of blood flow (later on) It probably won't affect how many people die after getting this condition. It also doesn't seem to greatly change the chance of another blockage happening in the same area of the brain. The facts aren't clear yet. There aren't enough people in the study for a reliable answer. More research is needed to see if a treatment called thrombolysis can help people recover better from bleeding in their brain caused by a weak blood vessel bursting.

Spurious generation

Run	SARI	Source	Spurious Content	
	(217)	Number	Number	Fraction
UvA_Task11_bartsent-cochraneauto	38.7	9,160	6	0.00
UvA_Task11_o-bartsent-cochraneauto	38.5	9,160	1	0.00
UvA1_llama31	38.7	9,160	4226	0.46
UvA_Task12_baseline-cochrane	41.8	666	182	0.27
UvA_Task12_bartdoc-cochraneauto	37.1	666	103	0.15
UvA_Task12_bartpara-cochraneauto	37.9	666	44	0.07
taiki_task12_llama31	35.6	666	603	0.91

Conclusions

1) Our plan-guided BART model trained on Cochrane-auto achieved the highest SARI score, but did not significantly outperform the baseline.

2) Our jargon-aware LLaMA model performed competitively, demonstrating the **potential** of LLMs in zero-shot simplification when guided by jargon detection and structured prompts.

3) Maintaining coherence and factual accuracy across longer texts remains a **challenge**.

Thank you for listening.