Sentence-level Scientific Text Simplification With Just a Pinch of Data

Task1.1-SimpleText@CLEF2025

Marvin M. Agüero-Torales   Carlos Rodríguez Abellán   Carlos A. Castaño Moraga

Fujitsu, CoE, Data Intelligence
Madrid, Spain

September, 2025, Madrid, Spain

Motivation

- Scientific texts are dense and difficult to understand.
- Plain language is crucial for medicine, law, government, and education.
- Challenge: Lack of parallel corpora (complex ↔ simple pairs).

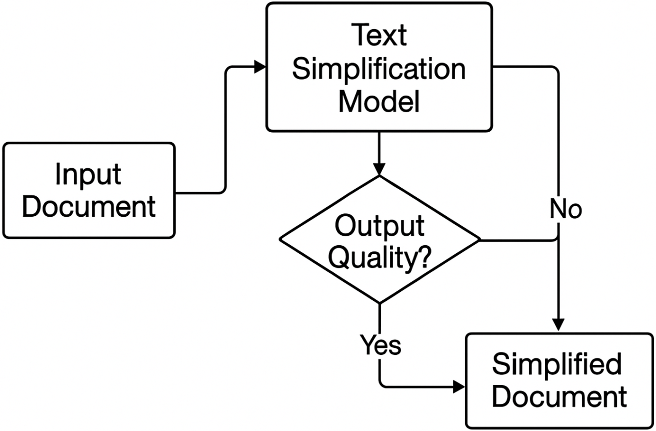
Research Question

Can we achieve competitive simplification with:

- Almost no training data (just 3 synthetic examples)?
- A mix of LLMs, rule-based, and ensemble methods?

Approach Overview

- Few-shot prompting with GPT-3.5-Turbo, o4-mini, T5-Efficient.
- Rule-based simplifier.
- Lightweight ensemble (shortest output).
- Unified LLM-as-a-Judge for evaluation and fallback.



Few-shot Prompting ("Pinch of Data")

- Only 3 synthetic sentence pairs used:
  1. Biomedical terminology
  2. Numerical information
  3. Discourse marker splitting
- Carefully curated examples outperform random sampling (+0.2 SARI).

Simplifiers

- Rule-based: remove parentheticals, split discourse markers...
- T5-Efficient: minimal prompting.
- GPT-3.5-Turbo / o4-mini: zero-shot and 3-shot.

Ensemble

- Select shortest non-empty output.
- Tie-breaking: GPT $>$ T5 $>$ Rule.
- Justification: brevity correlates with simplicity.

LLM-as-a-Judge

- Evaluates candidates on fluency, adequacy, simplicity (1–5).
- If score < 2.5: regenerate with GPT-3.5-Turbo.
- Provides automatic quality control.

Results

- Best: GPT-3.5-Turbo (3-shot) → SARI 38.84.
- Ensemble + Judge → SARI 38.55.
- Rule-based weaker (34.0).
- Optimal truncation: 45 characters.

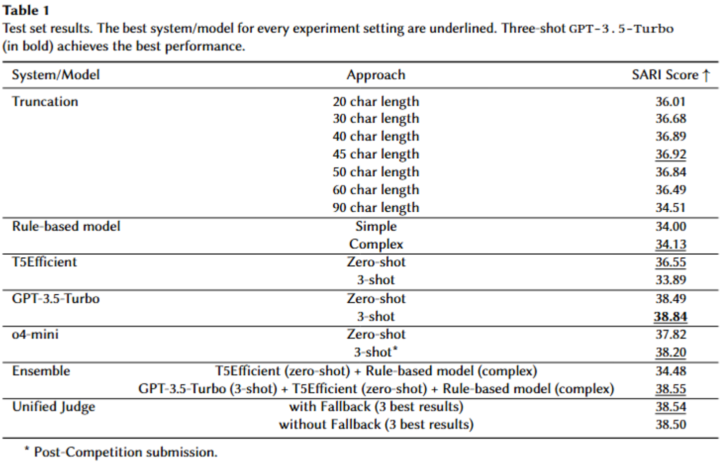


Table 1
Test set results. The best system/model for every experiment setting are underlined. Three-shot `GPT-3.5-Turbo` (in bold) achieves the best performance.

System/Model	Approach	SARI Score ↑
Truncation	20 char length	36.01
	30 char length	36.68
	40 char length	36.89
	45 char length	36.92
	50 char length	36.84
	60 char length	36.49
	90 char length	34.51
Rule-based model	Simple	34.00
	Complex	34.13
T5Efficient	Zero-shot	36.55
	3-shot	33.89
GPT-3.5-Turbo	Zero-shot	38.49
	3-shot	**38.84**
o4-mini	Zero-shot	37.82
	3-shot*	38.20
Ensemble	T5Efficient (zero-shot) + Rule-based model (complex)	34.48
	GPT-3.5-Turbo (3-shot) + T5Efficient (zero-shot) + Rule-based model (complex)	38.55
Unified Judge	with Fallback (3 best results)	38.54
	without Fallback (3 best results)	38.50

* Post-Competition submission.

Contributions

- High-quality simplification with minimal data.
- Novel unified judge for simplification.
- Establishes a new low-resource baseline.

Future Work

- Automated selection of examples.
- Explore non-synthetic few-shot data.
- Domain adaptation (medicine, law, government).
- Larger ensembles and multilingual settings.

Acknowledgments

- Supported by Fujitsu cloud credits (Microsoft Azure).
- Thanks to CLEF 2025 SimpleText organizers.