AIIRLab Systems for CLEF 2025 SimpleText: Cross-Encoders to Avoid Spurious Generation

Nicholas Largey, Deiby Wu & Behrooz Mansouri
Department of Computer Science, University of Southern Maine
SimpleText Lab, *CLEF 2025, Madrid, Spain*

AIIR Lab Participation

We participated in two main tasks:

- **Task 1 (Text Simplification):** at the Sentence & the Document Level
- **Task 2 (Controlled Creativity)**: Identify, Classify, and Avoid Hallucinations

The Core Approach to the Tasks: We utilized a mix of

- fine-tuned LLMs (Mistral, LLaMA)
- Cross-encoders
- Traditional machine learning models

Task 1: Simplifying Scientific Text

Models Used: Fine-tuned quantized LLMs from Unsloth

- Mistral-7b
- LLaMA-3.1-8b

Key Challenge & Solution

- **Problem:** Initial models produced hallucinations and extraneous, extremely noisy outputs
- **Solution**: Implemented explicit instructional prompts and output delimiters
  - e.g., "Start the Response with 'Simplification:'
  - Guiding the model and simplifying the parsing

Task 1: Selected Results

Subtask 1.1: Sentence Level

- Mistral-7b performed best with a SARI score of 36.08

Subtask 1.2: Document-Level

- Mistral-7b led with a SARI score of 42.4

Post-Competition Finding

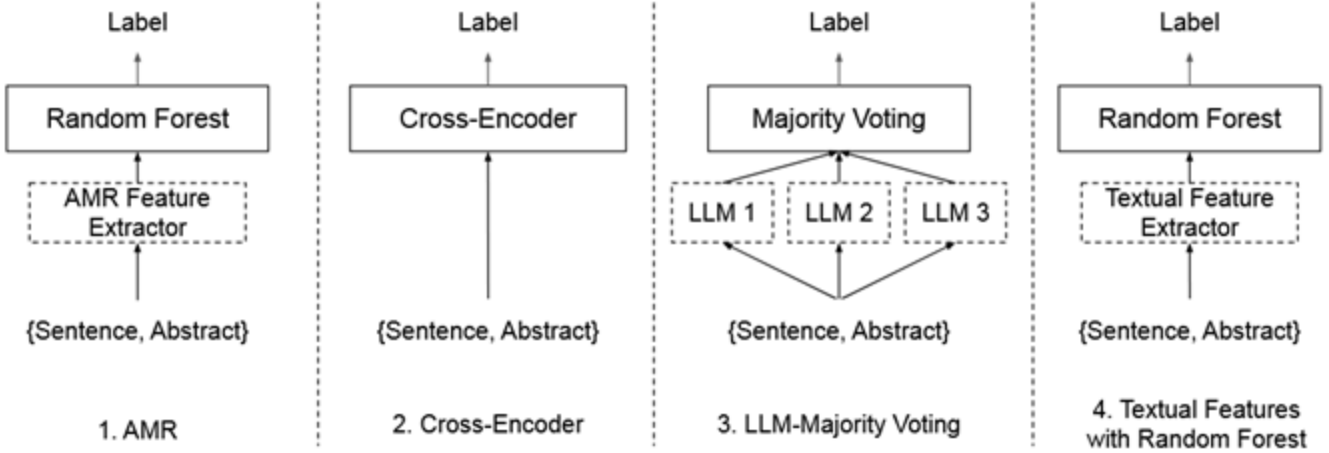
- The *base, non-quantized* versions of Mistral-7b and LLaMA-3.1-8b outperformed our fine-tuned models
- This suggests that for this specific task, the foundational models already possess strong simplification capabilities that our fine-tuning did not consistently improve upon

Task 2: Identify, Classify, and Avoid Hallucinations

Three Subtasks

- 2.1: Identify if a generated sentence is spurious
    - 4 systems
- 2.2: Detect and classify the *type* of error (e.g., factuality, redundancy)
    - 3 systems
- 2.3: Perform grounded simplification *by design*
    - 2 systems

SubTask 2.1: Spurious Generation Detection



1. AMR (Abstract Meaning Representation) + Random Forest
2. **Fine-tuned Cross-Encoder** (Best Performer, F1-score of 0.99)
3. LLM Majority-Voting
4. Textual Features + Random Forest

Subtask 2.2: Error Detection & Classification

- Text Classification with Fine-tuned BERT-based models
  - RoBERTa (threshold > 0.9)
  - Paraphrase-mpnet-base-v2 (threshold > 0.5)
- LLM-Majority Voting
  - Three large language models (LLaMA, Mistral, and Openchat)
  - All models used with few-shot prompting

System Prompt	You are a binary classifier for [Distortion Type]. [Brief definition of distortion]. Most simplified sentences do not contain this error. Only answer “Yes” if the error is clearly present. Respond only with “Yes” or “No”
User Prompt	Source sentence: [Test Source] Simplified sentence: [Test Simplified]

Subtask 2.2: Error Detection & Classification

- Text Classification with Fine-tuned BERT-based models
  - RoBERTa (threshold > 0.9)
  - Paraphrase-mpnet-base-v2 (threshold > 0.5)
- LLM-Majority Voting
  - Three large language models (LLaMA, Mistral, and Openchat)
  - All models used with few-shot prompting

The **Bi-encoder (paraphrase-mpnet)** performed best

- Especially at identifying "No Error" (F1 = 0.755)
- All models struggled with fine-grained error classification
- LLMs performed differently for each class, with no model outperforming all

Subtask 2.3: Perform Grounded Simplification by Design

Two Approaches

- **LLaMA Grounded:** Used a detailed system prompt instructing the LLM to produce grounded output.
- **LLaMA + Cross-encoder:** Generated a simplification with an LLM, then used our high-performing cross-encoder from Subtask 2.1 to check if it was spurious. If so, regenerate

Results

- Both grounded approaches **significantly outperformed the baseline**
  - SARI score of ~43 vs. 31
- They produced more concise and heavily modified simplifications
  - Maintaining quality, showing effective, controlled generation

Conclusion

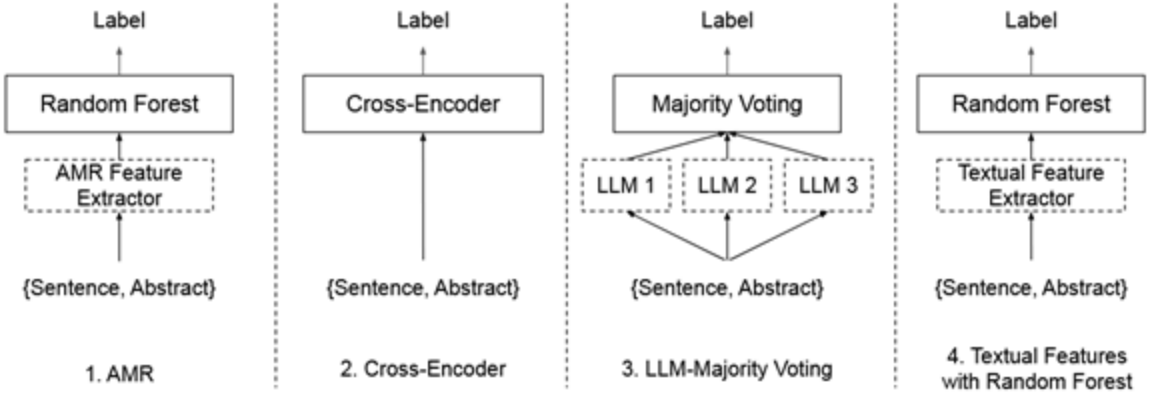
Key Takeaways

- **Explicit Prompts are Crucial:** Simple instructions and delimiters are effective for controlling LLM output
- **Cross-Encoders Excel at Grounding:** They are highly reliable for detecting spurious generation where LLMs might fail
- **Base LLMs Can Be Surprisingly Strong:** Fine-tuning is not always the answer and can sometimes degrade performance

Future Work

- Create an **integrated pipeline** that uses the cross-encoder to iteratively refine and verify simplifications generated by LLMs
- Improve fine-grained error classification with more sophisticated multi-label learning approaches

AIIRLab Systems for CLEF 2025 SimpleText: Cross-Encoders to Avoid Spurious Generation



Nicholas Largey, Deiby Wu, & Behrooz Mansouri
Contact: Nicholas.Largey@Maine.edu

AIR
Artificial Intelligence & Information Retrieval

UNIVERSITY OF SOUTHERN MAINE

MADRID CLEF 2025
Conference and Labs of the Evaluation Forum