

THM@SimpleText 2026 - Task 2: Ensemble-based Hallucination Classification in Text Simplification

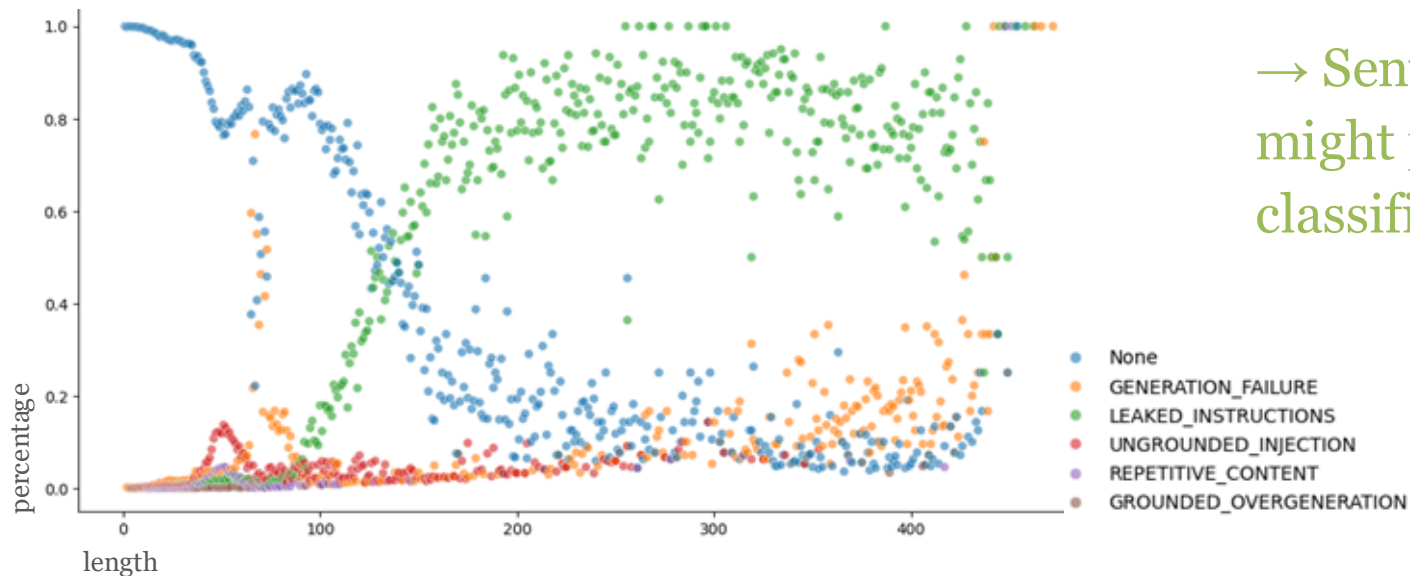
Julian Dauenhauer, Nico Hofmann, Christin Katharina
Kreutz
2026-09-22 – Jena, Germany

Idea + Approach

- Taking part in both Task 2.1 (binary) and Task 2.2 (multi-class)
- Inspiration from previous year's approaches (data analysis, BERT, ensemble, random forest)
- BERT ensemble + random forest classification
- LLM-as-a-judge for ambiguous cases

Dataset Analysis

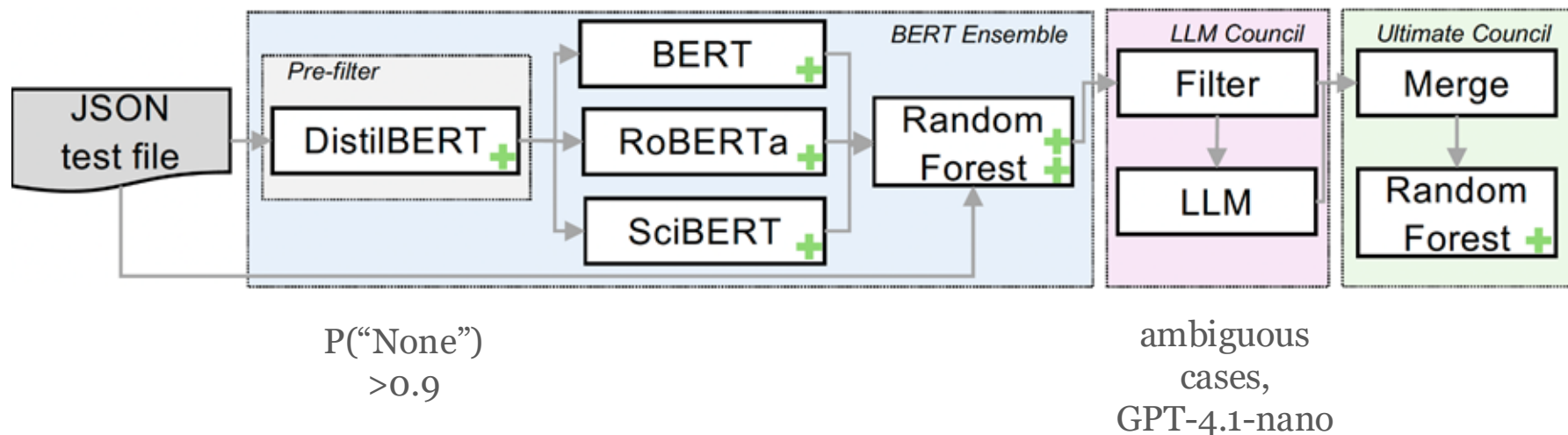
	Non-Spur	Spur	Failure	Leaked	Injection	Repetitive	Overgen
dev	91.58 (23.7)	8.42 (104.3)	3.61 (72.89)	2.54 (199.3)	1.77 (47.29)	0.50 (51.11)	0.02 (66.67)
train	92.33 (23.87)	7.67 (100.85)	3.23 (69.35)	2.35 (189.36)	1.64 (50.13)	0.42 (49.3)	0.03 (49.15)



→ Sentence length
might provide useful
classification signal

Pipeline

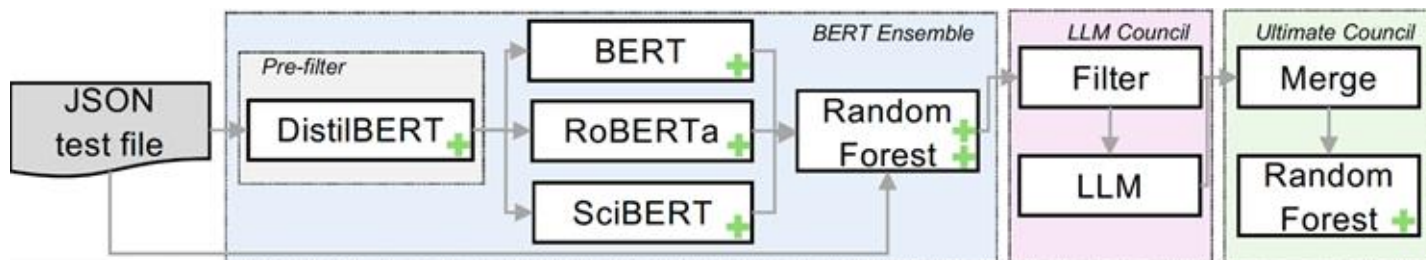
7 runs per task



ONLY multi-class classification

Results

- 95,6% of sentences had clear verdict before LLM Council
- 0.3% sentences required 2nd Random Forest pass
- 95.2% sentences classified as “None”
- 19.1% documents contained an error class
- 0 “Grounded overgeneration” labels assigned
- Best runs:
 - RoBERTa (0.7422 F1 document-level T 2.1)
 - DistilBERT (0.8151 F1 sentence-level T2.1, 0.7388 A T2.2)



Findings + Conclusion

- Full ensemble did not outperform simpler configurations
- DistilBERT pre-filter dominated pipeline
- False “None” predictions not corrected later
- Rare cases remained difficult, “Grounded overgeneration” never predicted

Big thanks to my student assistants
and the SimpleText organisers!