What do Large Language Models Actually Know?

Jenelle Bankas, sc23jeb@leeds.ac.uk

University of Leeds

12 September 2025

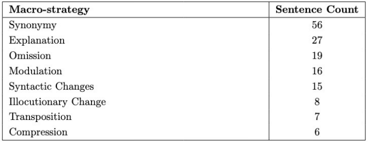
N8 CIR
COMPUTATIONALLY INTENSIVE RESEARCH

UNIVERSITY OF LEEDS

Project Overview

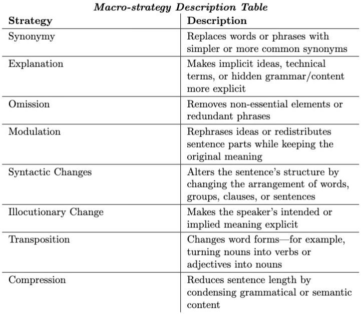
- What is eXplainable AI (XAI) and related concepts?
- How does this relate to text simplification?

Data & Methods



Macro-strategy	Sentence Count
Synonymy	56
Explanation	27
Omission	19
Modulation	16
Syntactic Changes	15
Illocutionary Change	8
Transposition	7
Compression	6

Figure: Strategy Sentence Distribution



Macro-strategy Description Table

Strategy	Description
Synonymy	Replaces words or phrases with simpler or more common synonyms
Explanation	Makes implicit ideas, technical terms, or hidden grammar/content more explicit
Omission	Removes non-essential elements or redundant phrases
Modulation	Rephrases ideas or redistributes sentence parts while keeping the original meaning
Syntactic Changes	Alters the sentence's structure by changing the arrangement of words, groups, clauses, or sentences
Illocutionary Change	Makes the speaker's intended or implied meaning explicit
Transposition	Changes word forms—for example, turning nouns into verbs or adjectives into nouns
Compression	Reduces sentence length by condensing grammatical or semantic content

Figure: Strategy Description

Tools

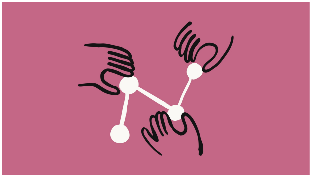


Figure: Circuit Tracer

Figure: Captum

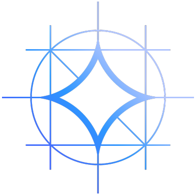


Figure: Gemma 3.2-1B Instruct

Results I

- **What were the main findings?**

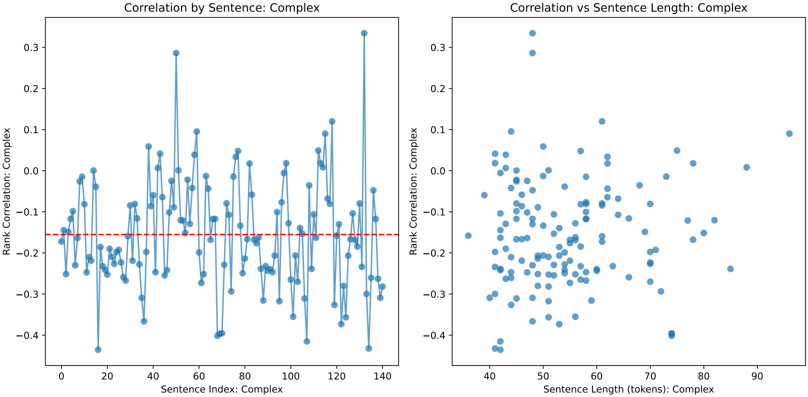


Figure: Complex Sentence Correlation Analysis

Results II

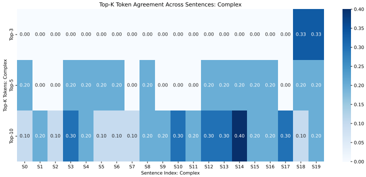


Figure: Complex Sentence Heat Map

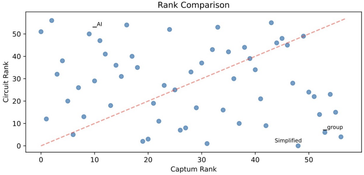


Figure: Complex Sentence Token Rank Scatter Graph

Discussion

- **What were the key takeaways from this project?**