# SINAI in SimpleText CLEF 2025: Simplifying Biomedical Scientific Texts and Identifying Hallucinations Using GPT-4.1 and Pattern Detection

**Jaime Collado-Montañez**
Jenny Alexandra Ortiz-Zambrano
César Espin-Riofrio
Arturo Montejo-Ráez

CLEF 2025 - Madrid, Spain
9-12 September 2025

# Tasks 1.1 & 1.2:
# Simplyfing Biomedical Texts

- Simplify scientific texts for non-specialist readers
- Sentence-level and document-level
- SARI, BLEU, compression, readability

# Zero-shot prompting

Using GPT-4.1

| **Aspect** | **Prompt 1** | **Prompt 2** |
|---|---|---|
| **Instruction style** | Direct: *“I want you to...”* | Suggestive: *“It should replace...”* |
| **Complex words** | All complex words in simplified text get explanations (always). | Only *relatively* complex words get explanations, given immediately after simplification. |
| **Acronym handling** | Replace acronyms fully with their meaning. | Keep acronym, add full meaning in parentheses at least the first time it appears. |
| **Overall style** | Simpler, stricter, rigid rules. | Flexible and ambiguous. |

# Results

| Prompt | Task | SARI | BLEU | Compression ratio | FKGL | Lexical complexity |
|---|---|---|---|---|---|---|
| source | - | 12.03 | 20.53 | 1.00 | 13.54 | 8.89 |
| v1 | sentence | 41.82 | 6.50 | 1.37 | 11.41 | 8.33 |
| v2 | sentence | 37.84 | 5.93 | 1.64 | 12.97 | 8.47 |
| v1 | document | 43.93 | 10.81 | 0.86 | 10.45 | 8.33 |
| v2 | document | 38.50 | 10.30 | 1.09 | 11.55 | 8.44 |

# Results

| Prompt | Task | SARI | BLEU | Compression ratio | FKGL | Lexical complexity |
|---|---|---|---|---|---|---|
| source | - | 12.03 | 20.53 | 1.00 | 13.54 | 8.89 |
| v1 | sentence | 41.82 | 6.50 | 1.37 | 11.41 | 8.33 |
| v2 | sentence | 37.84 | 5.93 | 1.64 | 12.97 | 8.47 |
| v1 | document | 43.93 | 10.81 | 0.86 | 10.45 | 8.33 |
| v2 | document | 38.50 | 10.30 | 1.09 | 11.55 | 8.44 |

# Results

| Prompt | Task | SARI | BLEU | Compression ratio | FKGL | Lexical complexity |
|---|---|---|---|---|---|---|
| source | - | 12.03 | 20.53 | 1.00 | 13.54 | 8.89 |
| v1 | sentence | 41.82 | 6.50 | 1.37 | 11.41 | 8.33 |
| v2 | sentence | 37.84 | 5.93 | 1.64 | 12.97 | 8.47 |
| v1 | document | 43.93 | 10.81 | 0.86 | 10.45 | 8.33 |
| v2 | document | 38.50 | 10.30 | 1.09 | 11.55 | 8.44 |

# Results

| Prompt | Task | SARI | BLEU | Compression ratio | FKGL | Lexical complexity |
|---|---|---|---|---|---|---|
| source | - | 12.03 | 20.53 | 1.00 | 13.54 | 8.89 |
| v1 | sentence | 41.82 | 6.50 | 1.37 | 11.41 | 8.33 |
| v2 | sentence | 37.84 | 5.93 | 1.64 | 12.97 | 8.47 |
| v1 | document | 43.93 | 10.81 | 0.86 | 10.45 | 8.33 |
| v2 | document | 38.50 | 10.30 | 1.09 | 11.55 | 8.44 |

# Results

| Prompt | Task | SARI | BLEU | Compression ratio | FKGL | Lexical complexity |
|---|---|---|---|---|---|---|
| source | - | 12.03 | 20.53 | 1.00 | 13.54 | 8.89 |
| v1 | sentence | 41.82 | 6.50 | 1.37 | 11.41 | 8.33 |
| v2 | sentence | 37.84 | 5.93 | 1.64 | 12.97 | 8.47 |
| v1 | document | 43.93 | 10.81 | 0.86 | 10.45 | 8.33 |
| v2 | document | 38.50 | 10.30 | 1.09 | 11.55 | 8.44 |

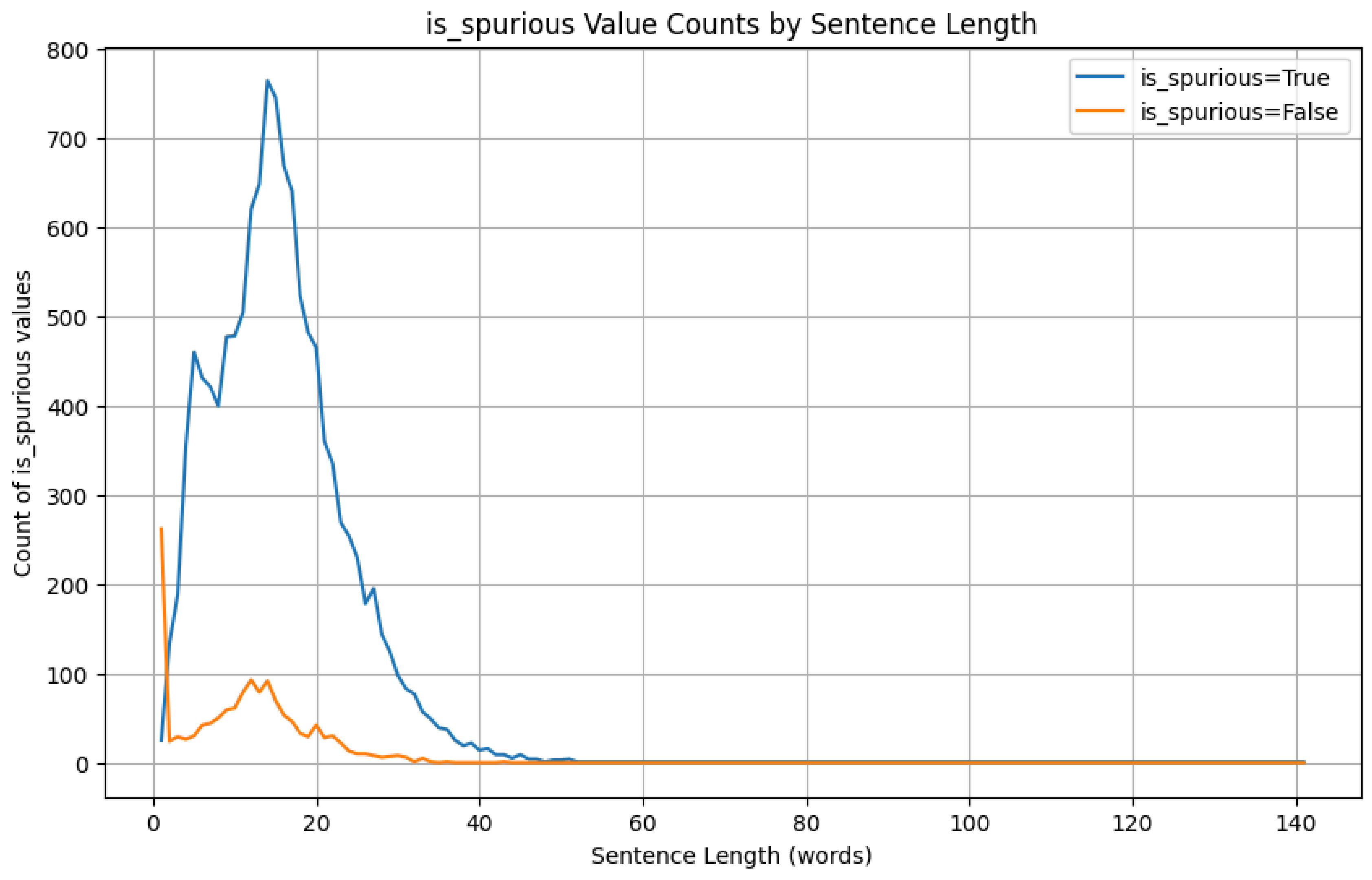
# Task 2.1:
# Hallucination detection

- Detect creative generation at the abstract or document level
- Sourced and post-hoc
- Precision, Recall and F1-score

# Exploratory Analysis

## Sourced training set

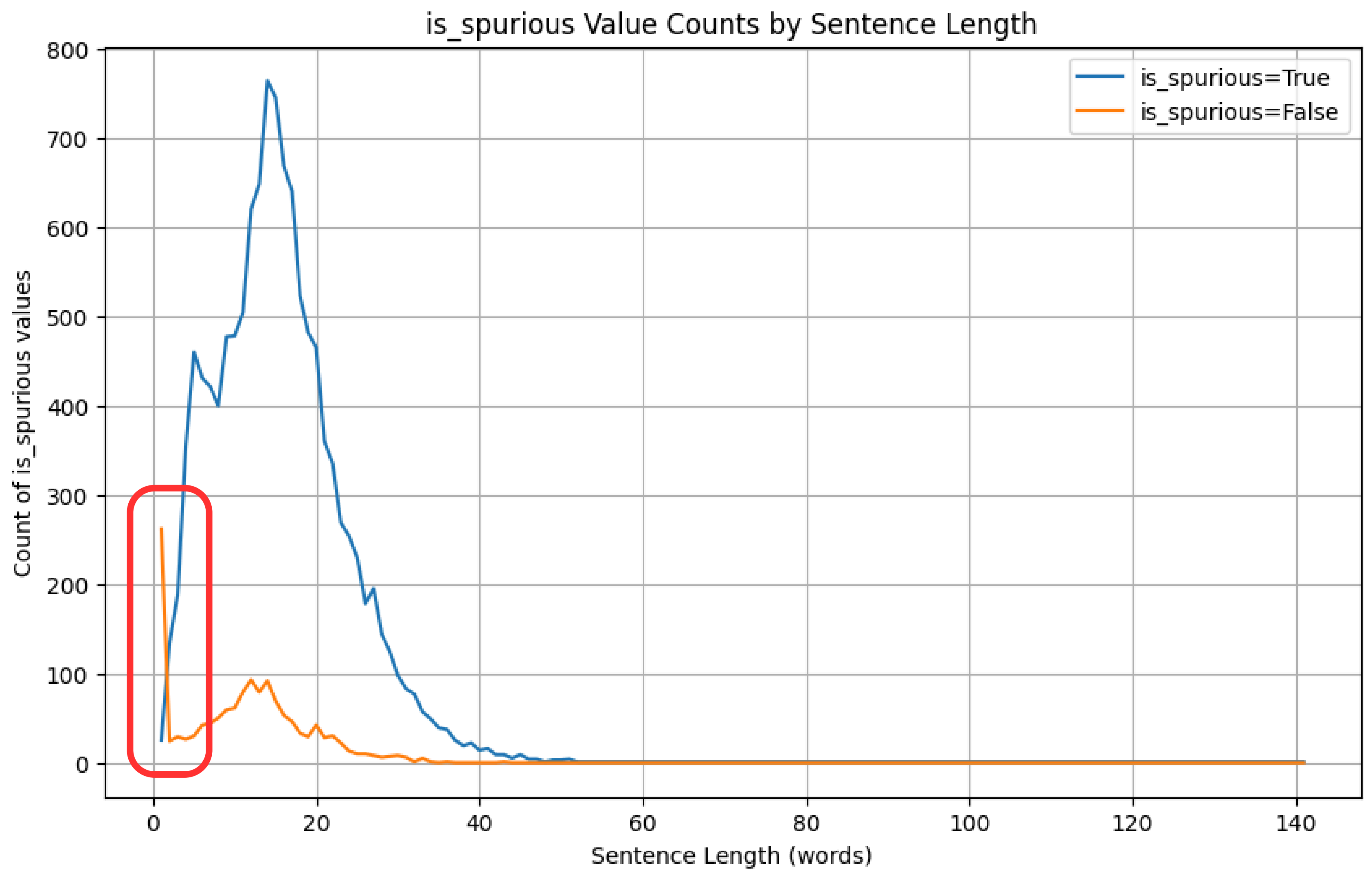
| | Spurious | Not Spurious |
|---|---|---|
| **# Examples** | **12115** | **1399** |
| **Average sentence length** | **15.386** | **11.152** |
| **# One word sentences** | **25** | **262** |



# Exploratory Analysis

## Sourced training set

| | Spurious | Not Spurious |
|---|---|---|
| **# Examples** | **12115** | **1399** |
| **Average sentence length** | **15.386** | **11.152** |
| **# One word sentences** | **25** | **262** |



# Exploratory Analysis

Patterns everywhere

| Pattern | # Spurious (12115) | # Not Spurious (1399) |
|---|---|---|
| One-word sentence (“#.”) | 14 | 0 |
| One-word sentence (“.”) | 0 | 244 |
| Sentence almost literally in source | 19 | 790 |
| Double space trailing sentences (“  “) | 1241 | 0 |

# Exploratory Analysis

Patterns everywhere

| Pattern | # Spurious (12115) | # Not Spurious (1399) |
|---|---|---|
| One-word sentence (“#.”) | 14 | 0 |
| One-word sentence (“.”) | 0 | 244 |
| Sentence almost literally in source | 19 | 790 |
| Double space trailing sentences (“  “) | 1241 | 0 |

# Exploratory Analysis

Patterns everywhere

| Pattern | # Spurious (12115) | # Not Spurious (1399) |
|---|---|---|
| One-word sentence ("#.") | 14 | 0 |
| One-word sentence (".") | 0 | 244 |
| Sentence almost literally in source | 19 | 790 |
| Double space trailing sentences ("  ") | 1241 | 0 |

Almost 75% of the examples!

# Our Approach

01 **Artificial source generation**: Create sources for the post-hoc dataset using llama-3.1-8b-instruct

02 **Rule-based filter:** Pre-anotate all sentences matching one of the patterns

03 **LLM prompting:** Ask llama-3.1-8b-instruct to answer whether the remaining sentences are spurious based on the source context (Yes or No)

04 **Confidence threshold:** Consider the sentence Not Spurious only if the LLM probability of generating the No token is higher than a given threshold (95% or 99%)

# Results

| | Sourced | | | Post-hoc | | |
|---|---|---|---|---|---|---|
| **Run** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| **Filters: One-word and double space**<br>**Confidence threshold: All Spurious** | 0.912 | **1** | 0.954 | 0.911 | **1** | **0.953** |
| **Filters: One-word, double space and literal match**<br>**Confidence threshold: 95%** | **1** | 0.786 | 0.88 | **0.957** | 0.222 | 0.36 |
| **Filters: One-word, double space and literal match**<br>**Confidence threshold: 99%** | **1** | 0.926 | 0.961 | 0.948 | 0.289 | 0.443 |
| **Filters: One-word, double space and literal match**<br>**Confidence threshold: All Spurious** | **1** | 0.953 | **0.976** | 0.942 | 0.317 | 0.474 |

# Conclusions

✓

### Rule-based filtering

The data shows hard-to-predict patterns for LLMs which should be filtered out

✗

### Artificial generation

Artificial generation of sources was not helpful for post-hoc examples

# THANK YOU!

**Jaime Collado-Montañez**
Jenny Alexandra Ortiz-Zambrano
César Espin-Riofrio
Arturo Montejo-Ráez

SINAI

UJA.
Universidad de Jaén

MADRID
CLEF 2025
Conference and Labs of the Evaluation Forum