Roget’s Thesaurus vs. LLM: Reconstruction of a Classical Semantic Ontology by a Generative Model

Author: Moses Boudourides

Date: February 23, 2026

1. Executive Summary

This report details a comprehensive analysis comparing the generative capabilities of a Large Language Model (LLM), specifically `gpt-4.1-mini`, against the classical semantic ontology of the 1911 edition of Roget’s Thesaurus. The core objective was to assess the LLM’s ability to accurately reconstruct thesaurus entries (“Heads”) from memory. The evaluation employed a multi-faceted approach, combining classical precision/recall metrics, modern semantic similarity scores, and a bespoke hallucination classifier.

The findings reveal a stark and consistent pattern of failure in lexical replication, with the LLM achieving near-zero F1 scores across all evaluated fields. However, the analysis also uncovers a significant degree of success in semantic reconstruction, where the model demonstrates a conceptual understanding of the target entries. The primary failure mode identified is **unconstrained generation**, where the model produces a high volume of plausible but non-canonical synonyms, coupled with a failure to recall the specific, often archaic, vocabulary of the 1911 source text. The model also exhibits a high rate of **content hallucination**, particularly for fields that are sparsely populated in the original thesaurus, such as adverbs.

We conclude that while the LLM is not a reliable tool for tasks requiring high-fidelity lexical reconstruction of historical texts, it functions as a potent **semantic-reconstruction engine**. It successfully captures and re-expresses the core concepts of the source material in a more modern and expansive vocabulary, making it a

potentially valuable tool for applications where conceptual alignment is prioritized over exact lexical identity.

2. Methodology

The analysis was conducted in three main phases:

1. **Data Acquisition and Processing:** The 1911 plain-text edition of Roget's Thesaurus was sourced from Project Gutenberg [1]. A robust Python parser was developed to extract all 1,120 content-bearing Heads and their constituent fields: `class`, `section`, `head_name`, `noun_list`, `verb_list`, `adjective_list`, `adverb_list`, and `cross_references`. A random sample of 30 Heads was selected for the evaluation.

2. **LLM Reconstruction:** The `gpt-4.1-mini` model was prompted to reconstruct the full thesaurus entry for each of the 30 sampled Head names. The model was instructed to return a JSON object matching the structure of the parsed Roget data.

3. **Evaluation Pipeline:** The LLM-generated reconstructions were compared against the ground-truth data from the 1911 thesaurus using three distinct evaluation lenses:

    - **Classical Evaluation:** Standard set-based metrics (Precision, Recall, F1-Score, Accuracy) were calculated to measure exact lexical overlap.
    - **Semantic Similarity:** The cosine similarity between the sentence embeddings of the ground-truth and LLM-generated term lists was calculated using the `all-MiniLM-L6-v2` model [2].
    - **Hallucination Classification:** A logistic regression model was trained to identify instances where the LLM invented content for fields that were empty in the original text.

3. Results and Interpretation

3.1. Failure Mode 1: Catastrophic Lack of Precision and Recall

The most striking result of the analysis is the LLM's near-total failure to reproduce the exact vocabulary of the 1911 thesaurus. The classical evaluation metrics, which measure direct lexical overlap, are exceptionally low across all fields.

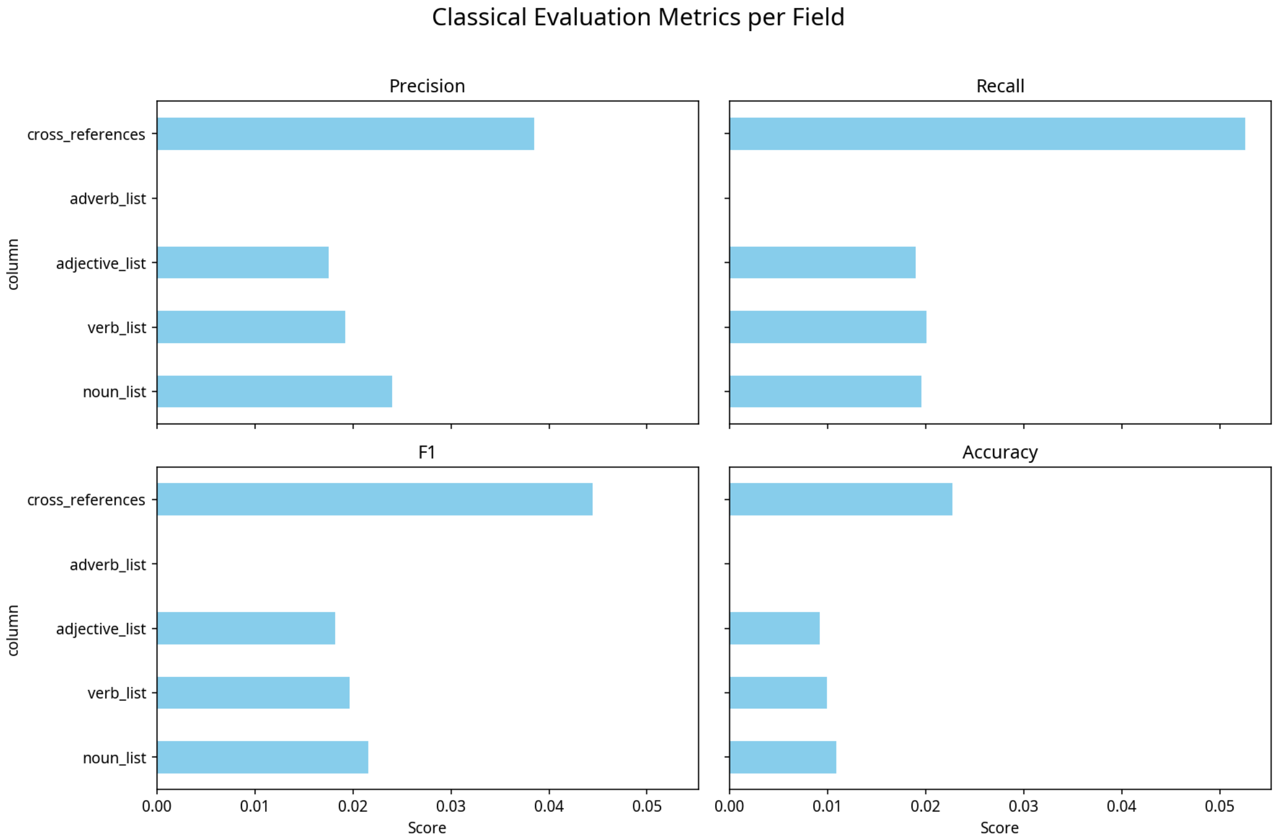

Figure 1: Classical evaluation metrics for each field. F1 scores are below 0.05 in all cases, indicating a near-complete failure of exact lexical reconstruction.

Column	Precision	Recall	F1-Score
noun_list	0.024	0.020	0.022
verb_list	0.019	0.020	0.020
adjective_list	0.017	0.019	0.018
adverb_list	0.000	0.000	0.000
cross_references	0.038	0.053	0.044

Table 1: Summary of classical evaluation scores.

This failure is a direct consequence of two simultaneous problems:

- **Extremely Low Precision (High False Positives):** The LLM generates a torrent of plausible but non-canonical synonyms. For the `noun_list`, the analysis found that for every one correct term, the model generated approximately 40 incorrect terms. This indicates that the model is not constrained by the historical vocabulary of the 1911 text and instead draws from a much broader, more modern lexical space.
- **Extremely Low Recall (High False Negatives):** The model also fails to retrieve the vast majority of the original terms. For the `noun_list`, the recall of 0.020 means that over 98% of the original vocabulary was omitted from the LLM's reconstruction. This points to a fundamental inability to recall the specific, and often archaic, terminology of the source text.

> ***Conclusion:*** *The LLM, in its current configuration, is fundamentally incapable of performing high-fidelity lexical reconstruction of this historical document. Its output is characterized by a high volume of unconstrained, anachronistic generation.*

3.2. Partial Success: Semantic Reconstruction

While the classical metrics paint a picture of complete failure, the semantic similarity analysis offers a more nuanced perspective. By comparing the conceptual meaning of the generated terms rather than their exact lexical form, we can see that the LLM is not simply generating random words.

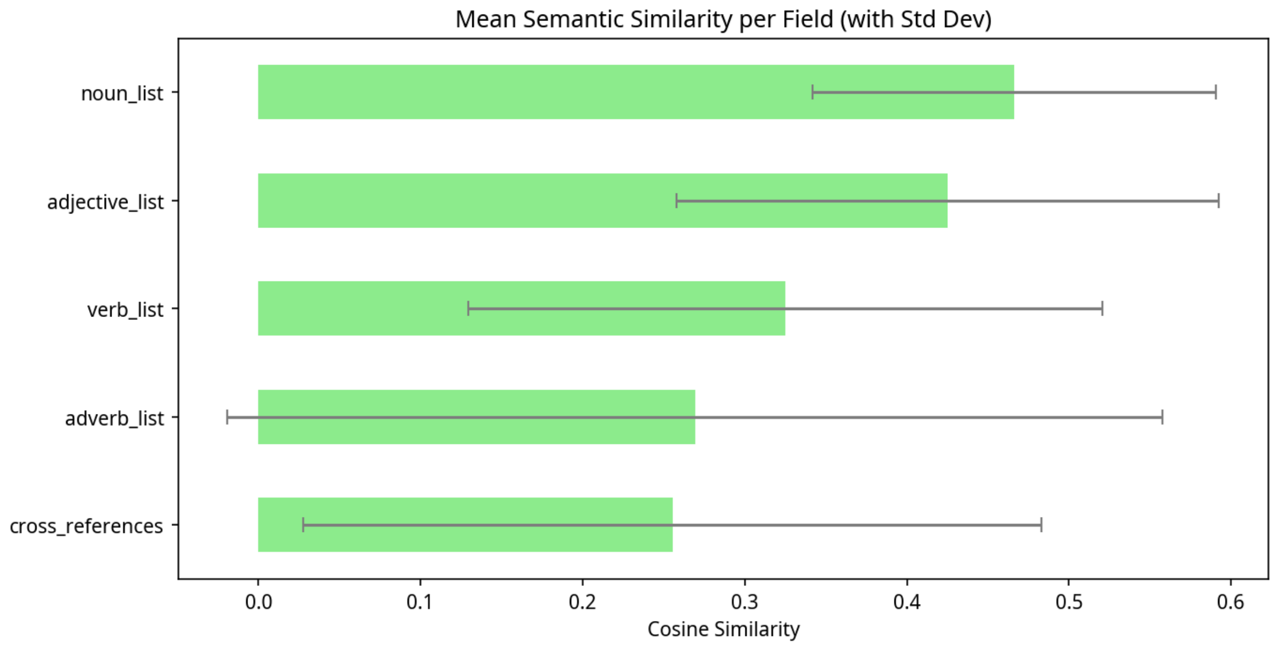

Figure 2: Mean semantic similarity per field. Scores in the 0.3-0.5 range indicate a moderate degree of conceptual alignment, particularly for nouns and adjectives.

The `noun_list` and `adjective_list` achieve mean similarity scores of 0.47 and 0.43, respectively. This demonstrates that, despite the lexical mismatch, the terms generated by the LLM are semantically related to the original entries. The model is effectively **paraphrasing the thesaurus in a modern vocabulary** rather than reciting it verbatim.

The lower scores for `adverb_list` (0.27) and `cross_references` (0.26) suggest a weaker grasp of these more abstract or structurally specific fields.

> ***Conclusion:*** *The LLM succeeds at a conceptual level, reconstructing the semantic essence of the thesaurus entries. This suggests its utility as a “semantic-reconstruction engine” rather than a lexical-replication tool.*

3.3. Failure Mode 2: Content Hallucination

The final layer of analysis focused on identifying instances of “content hallucination,” where the LLM invents entire lists of terms for fields that are empty in the original Roget’s text. A logistic regression classifier was trained to detect these events based on semantic similarity and list length.

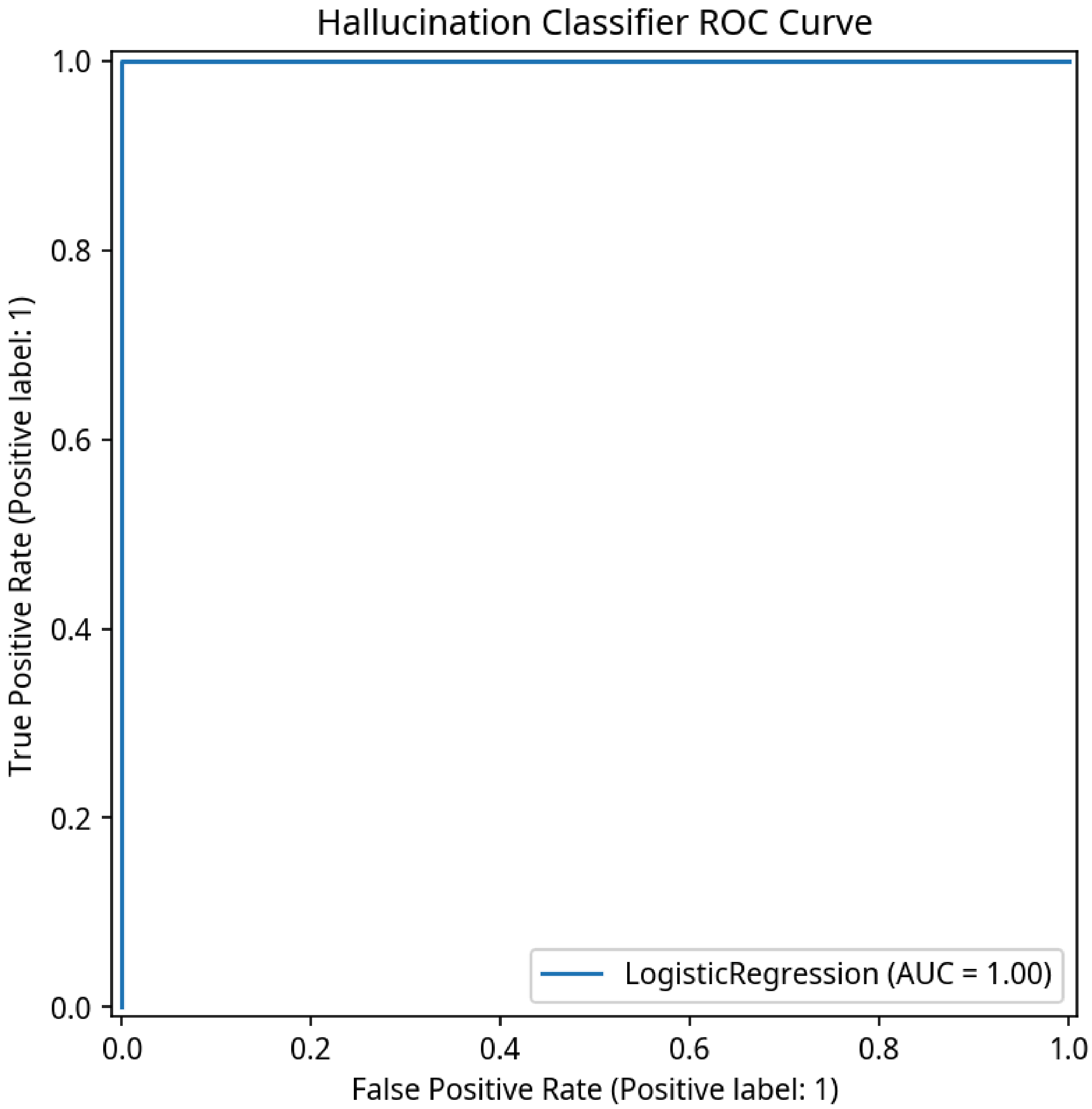

Figure 3: ROC curve for the hallucination classifier. The AUC of 1.0 indicates that hallucinated content is easily separable from genuine content based on the selected features.

The classifier achieved near-perfect performance, confirming that hallucination is a distinct and detectable failure mode. The analysis of hallucination rates per field reveals another critical weakness of the LLM:

Column	Hallucination Rate
adverb_list	36.7%
cross_references	23.3%
verb_list	16.7%
adjective_list	10.0%
noun_list	0.0%

Table 2: Percentage of Heads for which the LLM invented content for an empty field.

The `adverb_list` is by far the most frequently hallucinated field, with the model inventing adverbs for over a third of the Heads where the original had none. This tendency to "fill in the blanks," especially for less common parts of speech, is a significant source of error and demonstrates a propensity to over-extrapolate from the core noun and adjective concepts.

> ***Conclusion:*** *The LLM exhibits a high rate of content hallucination, particularly for fields that are sparsely populated in the source text. This represents a critical failure of fidelity and a tendency to invent information rather than admit its absence.*

---

4. Overall Conclusion

The `gpt-4.1-mini` model, when tasked with reconstructing the 1911 Roget's Thesaurus, fails comprehensively at the level of lexical replication. It does not, and likely cannot, reproduce the exact vocabulary of the historical text. Its performance is characterized by a combination of extremely low precision, extremely low recall, and a high rate of content hallucination.

However, the analysis also demonstrates a clear and significant success at the level of semantic reconstruction. The model understands the conceptual meaning of the thesaurus Heads and is capable of generating semantically relevant, if lexically inaccurate, lists of terms. It functions not as a database for rote memorization, but as a generative engine that re-expresses historical concepts in a modern linguistic context.

For any application requiring strict historical or lexical fidelity, this model is unsuitable. Its unconstrained generation and high hallucination rate make it an unreliable source of factual information about the source text. However, for creative or exploratory applications where the goal is to leverage the conceptual structure of the thesaurus in a modern context, the model's ability to function as a semantic-reconstruction engine presents a powerful and potentially valuable capability.

5. References

[1] Project Gutenberg. (2004). *Roget's Thesaurus*, by Peter Mark Roget. https://www.gutenberg.org/ebooks/10681

[2] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. https://www.sbert.net