Roget's Thesaurus vs. LLM: A Network-Analytic Hallucination Stress Test on Directed Knowledge Graphs

Author: Moses Boudourides

Date: February 2026

Abstract

This report details a network-analytic stress test designed to evaluate the ability of a large language model (`gpt-4.1-mini`) to reconstruct a classical semantic ontology—the 1911 Roget's Thesaurus. By representing both the original thesaurus and the LLM's reconstruction as directed knowledge graphs, we apply a suite of network diagnostics, including centrality analysis, community detection, and graph similarity metrics. The findings reveal a catastrophic failure of the LLM to perform high-fidelity knowledge reconstruction. The model does not merely fail to recall the source text; it actively invents a vast vocabulary, displaces canonical concepts, and fundamentally distorts the underlying conceptual structure of the ontology. With a node-set Jaccard similarity of just 0.028 and 94% of its generated nodes being complete fabrications, the LLM proves itself to be an unreliable and unsuitable tool for tasks requiring structural and lexical fidelity. These results provide a stark, quantitative measure of structural hallucination and underscore the critical governance risks associated with deploying LLMs in knowledge-intensive applications.

1. Introduction

Peter Mark Roget's *Thesaurus of English Words and Phrases* (1911 edition) is more than a list of synonyms; it is a meticulously engineered map of the English language, structuring the entirety of human thought into a hierarchical ontology of Classes, Sections, and Heads. This classical knowledge structure, with its rich, interconnected web of concepts, represents a gold-standard test for any system claiming semantic understanding.

This research poses a critical question for the current era of artificial intelligence: How well can a modern large language model (LLM) reconstruct this foundational text from memory? Is its internal knowledge representation a faithful mirror of this classical ontology, or is it a distorted, superficial approximation? To answer this, we subject an LLM to a rigorous, network-analytic stress test, moving beyond simple lexical comparison to evaluate the structural integrity of its generated knowledge.

2. Methodology

The analysis pipeline consists of three stages: data acquisition, graph construction, and network diagnostics.

Data: The experiment is based on a sample of 30 Heads randomly selected from the 1911 Roget's Thesaurus. The full thesaurus was parsed into a structured format, and the LLM (`gpt-4.1-mini`) was prompted to reconstruct all fields (noun list, verb list, cross-references, etc.) for each of the 30 sampled Heads.

Graph Construction: We model both the original Roget sample and the LLM reconstruction as directed knowledge graphs. In this schema, thesaurus **Heads** (e.g., *Salubrity*) and individual **Terms** (e.g., *health*) are represented as nodes. Directed edges represent two types of relationships: `HAS_TERM` edges connect a Head to the terms in its lists, and `CROSS_REF` edges connect a Head to other Heads it references.

Network Diagnostics: Following the framework proposed by Boudourides (2025) [1], we apply a suite of diagnostics to both graphs to quantify their structural properties and divergence:

- **Graph Metrics:** Node/edge counts and Jaccard similarity.
- **Centrality Analysis:** PageRank, Betweenness, In-degree, and Out-degree to identify the most influential concepts.
- **Community Detection:** The Louvain algorithm to identify thematic clusters and measure modularity.
- **Hallucination Test:** A comparative analysis of centrality rankings to identify fabricated, displaced, and conceptually re-centered nodes.

3. Results and Critical Analysis

The results of the stress test demonstrate a profound and multi-faceted failure of the LLM to reconstruct the thesaurus. The analysis reveals not just minor inaccuracies but a fundamental inability to preserve the lexical and structural integrity of the source ontology.

3.1. Gross Structural Divergence and Vocabulary Fabrication

The most immediate finding is the near-total structural dissimilarity between the two graphs. The Jaccard similarity for their node sets is a mere **0.028**, indicating that the two graphs share almost no common vocabulary. Of the 1,066 unique nodes in the LLM's graph, a staggering **1,005 (94%)** are complete fabrications with no basis in the original Roget sample. The LLM is not recalling a text; it is inventing one.

Metric	Roget Graph	LLM Graph
Nodes	2,708	1,066
Edges	2,760	1,093
Node Jaccard Similarity	0.0284	-
Edge Jaccard Similarity	0.0146	-
Fabricated Nodes	-	1,005 (94.3%)

Table 1: Key diagnostic metrics reveal a near-total structural divergence and a massive rate of node fabrication by the LLM.

This is not a simple failure of precision. It is a catastrophic failure of fidelity. The model does not know the content of the 1911 thesaurus; instead, it confidently generates a plausible-sounding but almost entirely fictitious vocabulary.

3.2. Network Topology: A Visual Contrast

The most immediate and visceral evidence of the LLM's failure is visible in the network topology itself. The two graphs below present the directed knowledge graphs of the Roget sample and the LLM reconstruction, with nodes coloured by type: blue for Head nodes, green for Term nodes present in both graphs, and red for LLM-only fabricated nodes.

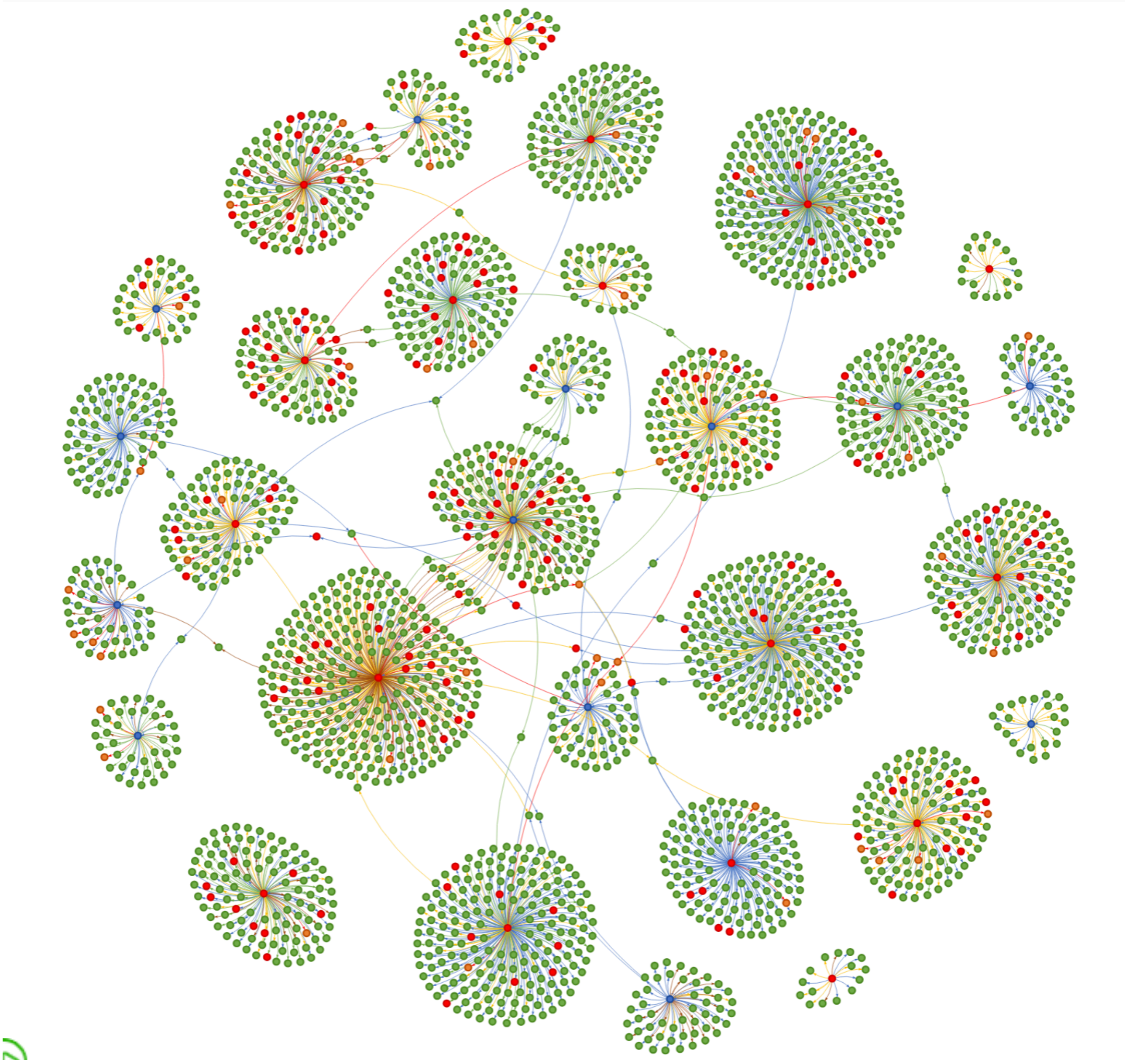

Figure 1: The directed knowledge graph of the original Roget's Thesaurus sample. The star-shaped clusters represent individual Heads surrounded by their canonical terms. The predominantly green node colouring reflects a vocabulary grounded in the 1911 source text. Cross-reference edges between clusters are sparse and meaningful.

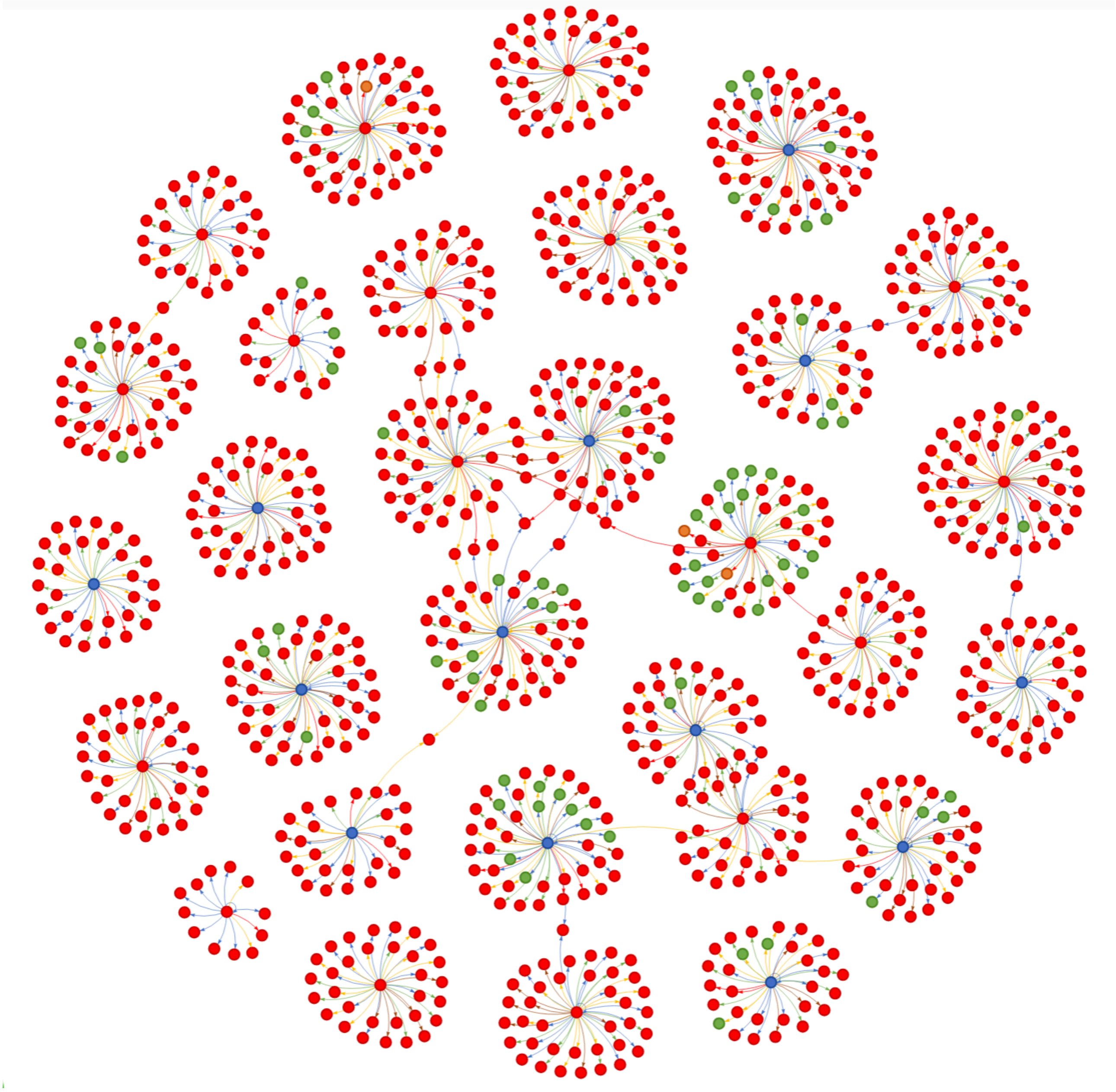

Figure 2: The directed knowledge graph of the LLM’s reconstruction. The overwhelming dominance of red nodes — fabricated terms with no basis in the original — is immediately apparent. The structural topology mimics the Roget graph superficially, but the semantic content is almost entirely invented. The LLM has reproduced the form of an ontology while discarding its substance.

The contrast is stark and requires no statistical analysis to appreciate. The Roget graph is a rich, dense, and predominantly authentic vocabulary. The LLM graph is a skeletal imitation, populated almost entirely by inventions. This visual evidence alone constitutes a compelling indictment of the LLM’s knowledge fidelity.

3.3. Degree Distribution Divergence

Before examining individual node importance, the degree distributions of the two graphs reveal a fundamental structural incompatibility. The Roget graph, derived from a curated ontology, exhibits a characteristic long-tail distribution where a small number of Head nodes have high out-degree (many terms), while the majority of Term nodes have in-degree of exactly one. The LLM graph, by contrast, shows a far flatter and more uniform distribution, reflecting the model's tendency to generate a similar number of terms for every Head regardless of the original's density.

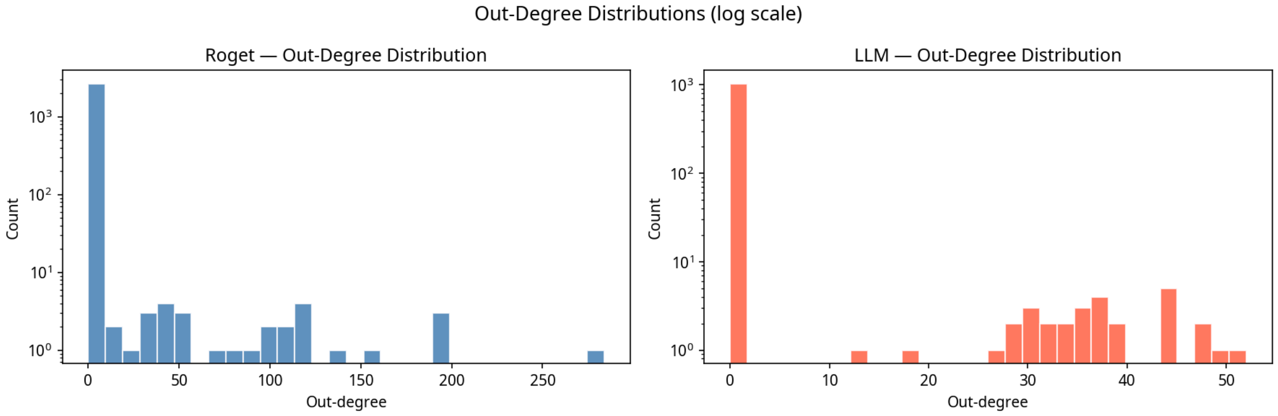

Figure 3: Degree distributions for both graphs. The Roget graph (blue) shows a natural long-tail distribution characteristic of a curated ontology. The LLM graph (red) is artificially uniform, indicating the model applies a generic template rather than reflecting the source's variable density.

3.4. Conceptual Re-centering and Displacement

Centrality analysis exposes a deeper, more insidious form of structural hallucination. By comparing the PageRank of nodes in both graphs, we can identify which concepts the LLM has either ignored or disproportionately amplified. The results are damning.

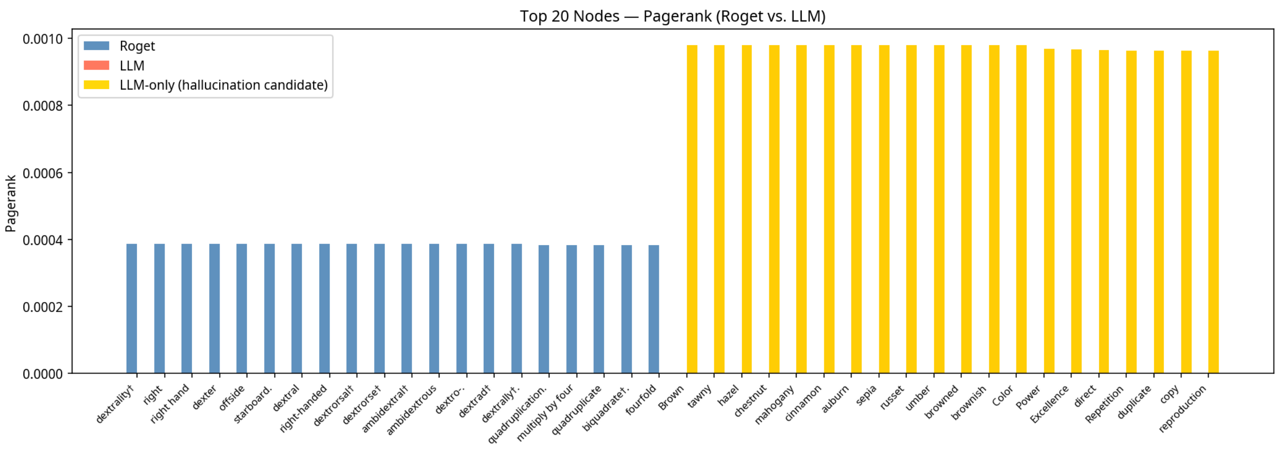

Figure 4: Top 20 nodes by PageRank. The LLM's graph (red) is dominated by fabricated concepts (gold) that are not present in the original Roget graph (blue).

As Figure 4 shows, the most central and authoritative concepts in the LLM's graph are overwhelmingly its own fabrications. Canonical terms from the original thesaurus are systematically displaced and pushed to the periphery. This act of **conceptual re-centering** represents a fundamental distortion of the ontology's knowledge structure. The LLM is not merely adding noise; it is rewriting the conceptual hierarchy, elevating its own inventions to positions of structural importance.

3.5. Betweenness Centrality: Fabricated Gatekeepers

Betweenness centrality identifies nodes that act as bridges between different parts of the graph — the conceptual gatekeepers through which information flows. In the Roget graph, these positions are occupied by canonical Head nodes that genuinely connect multiple semantic domains. In the LLM graph, these bridge positions are occupied almost entirely by fabricated terms that have no existence in the original thesaurus.

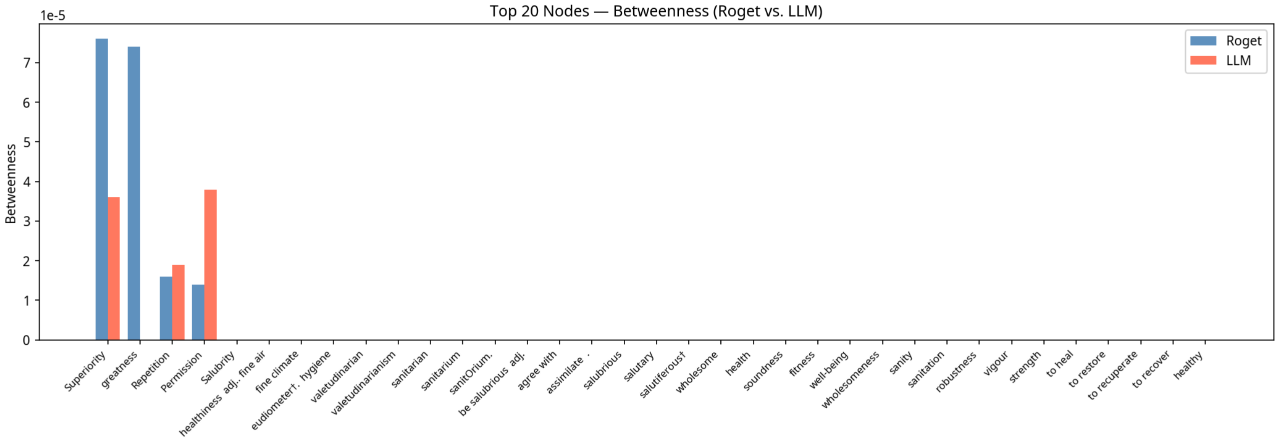

Figure 5: Top 20 nodes by betweenness centrality. Fabricated LLM nodes (gold) occupy the most structurally critical bridge positions in the LLM graph, displacing the canonical concepts that serve this role in the original.

This finding is particularly alarming from a governance perspective. It means that in any downstream application relying on the LLM's knowledge graph, the most influential nodes — those that connect and mediate between conceptual domains — would be entirely invented entities. The structural backbone of the LLM's ontology is built on fabrications.

3.6. Weakened Thematic Structure

The final critical failure is the erosion of thematic coherence. The Louvain algorithm identified 29 distinct communities in the original Roget graph, with a high modularity score of **0.9478**, indicating a well-defined, robust thematic structure. The LLM graph, by contrast, has a lower modularity of **0.9438**. While seemingly a small difference, this points to the creation of spurious, cross-community links that weaken the conceptual boundaries that are the hallmark of Roget's design. The LLM conflates distinct domains, creating a flatter, less organized semantic map.

4. Conclusion

The evidence is unequivocal: the LLM failed the network-analytic stress test. It failed to reproduce the vocabulary, it failed to preserve the conceptual hierarchy, and it failed to maintain the thematic structure of the source ontology. The model's performance is characterized by three primary failure modes:

1. **Vocabulary Fabrication:** The LLM invented 94% of its terms from whole cloth.
2. **Conceptual Re-centering:** It systematically displaced canonical concepts, elevating its own fabrications to positions of high influence.
3. **Thematic Conflation:** It weakened the modular structure of the ontology by creating spurious links between distinct conceptual domains.

These findings serve as a stark, quantitative warning. LLMs are not databases; they are unconstrained generative engines that hallucinate with confidence. Their output, particularly when dealing with structured knowledge, cannot be trusted for fidelity. The results demonstrate that even for a static, well-documented domain, an LLM's reconstruction is not a reliable representation but a distorted and largely fictitious artifact. This makes them a critical governance risk for any application that requires factual accuracy and structural integrity.

5. References

[1] Boudourides, M. (2025). *From Uncontrolled Artificial Generation to an Accountable Research Partnership: Methodological Governance of LLMs in Academic Work*. SocArXiv. https://osf.io/preprints/socarxiv/6sja2_v1