

# Using and Evaluating LLMs in Academic Work

## Session 4: Network Diagnostics and the Hallucination Stress Test

Moses Boudourides

*Faculty, Graduate Program on Data Science  
Northwestern University*

[Moses.Boudourides@northwestern.edu](mailto: Moses.Boudourides@northwestern.edu)

[Moses.Boudourides@gmail.com](mailto: Moses.Boudourides@gmail.com)

**instats Seminar**

Thursday, February 26, 2026  
5:30 PM – 7:00 PM UTC

# Session 4: Network Diagnostics and the Hallucination Stress Test

- ① Network Diagnostics of LLM Outputs
- ② Hallucination Stress Test
- ③ Centrality Analysis
- ④ Modularity Analysis
- ⑤ Various Other Graph Measures

# From Representation to Diagnosis

- In Session 3 we translated text into two knowledge graphs:
  - $G_{LLM}$  — knowledge graph from LLM output
  - $G_{REF}$  — reference graph from trusted sources
- In this way, since a rich toolkit of network analysis techniques becomes available for the diagnosis of structural distortions in LLM-generated content, now we may ask:

*How can we detect structural distortion quantitatively?*

# The Hallucination Stress Test

- We designate a comparative, structure-based diagnostic procedure, providing quantitative metrics for the evaluation of the structural integrity of LLM-generated content, as a **hallucination stress test**, since it examines the stability of LLM-induced knowledge representations under comparison with trusted reference structures.
- In particular, we define a hallucination stress test as:

*A structured comparison of  $G_{LLM}$  against  $G_{REF}$  using network diagnostics to detect structural divergence.*

- The goal is not automation, but inspectable evidence, through
  - Identifying Core Concepts
  - Checking for Displacement
  - Investigating Upwardly Mobile Concepts
- These steps allow us to systematically identify and quantify structural hallucinations.

# Core Diagnostic Questions

- The hallucination stress test procedure involves and asks:
  - ① Identifying core concepts:  
*Are canonical concepts still central?*
  - ② Assessing hierarchical depth:  
*Has conceptual hierarchy been flattened?*
  - ③ Checking for displacement:  
*Have peripheral concepts become artificially dominant?*
  - ④ Investigating conceptual structure:  
*Are thematic communities preserved?*
- These steps allow us to systematically identify and quantify structural hallucinations.

# Step 1: Identifying Core Concepts

- From the reference graph, we identify the top-ranked concepts according to various centrality measures.
- Centrality measures estimate node importance.
- These are the canonical concepts of the field.

## Step 2: Checking for Displacement

- We then examine the centrality rankings in the LLM-generated graph.
- Are the canonical concepts still central?
- Have they been displaced by other concepts?
- Displacement in centrality ranking between  $G_{LLM}$  and  $G_{REF}$  signals structural distortion.

## Step 3: Investigating Upwardly Mobile Concepts

- We identify any concepts that have a significantly higher centrality rank in the LLM-generated graph than in the reference graph.
- These are potential “hallucinations.”
  - We must investigate their provenance.
  - Are they real but peripheral concepts that have been over-emphasized?
  - Or are they entirely fabricated by some sort of statistical autocompletion?

# Introduction to Centrality Measures

- Key network diagnostics of hallucination stress includes **centrality analysis**.
- Centrality measures offer a direct way to identify which concepts function as structural hubs within the graph.
- When applied comparatively, discrepancies in centrality rankings between an LLM-generated graph and a reference graph reveal cases in which peripheral or fabricated concepts are artificially inflated.

# Degree Centrality

- **Degree Centrality**  $C_D(v)$  is the number of connections a node  $v$  has:

$$C_D(v) = \deg(v)$$

- A node with high degree centrality is one that is connected to many other concepts, suggesting it plays a broad, integrative role.
- Compare:

$$C_D^{LLM}(v) \quad \text{vs} \quad C_D^{REF}(v)$$

# Degree Displacement

- Nodal **rank** in degree centrality = position of a node when all nodes are ordered by degree (most connected node has rank 1).
- Compute **rank correlation**:

$$\rho = \text{corr}(\text{rank}_{LLM}, \text{rank}_{REF})$$

- High  $\rho$  indicates structural alignment
- Low  $\rho$  indicates one of the following:
  - re-centering
  - distortion
  - misplaced centrality

# Betweenness Centrality and Distortion

- **Betweenness Centrality**  $C_B(v)$  of node  $v$  is a measure of how often a node  $v$  lies on the shortest path between other nodes. Formally:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where  $\sigma_{st}$  is the total number of shortest paths between nodes  $s$  and  $t$ , while  $\sigma_{st}(v)$  is the number of those shortest paths that pass through node  $v$ .

- It measures bridging role: A concept with high betweenness centrality acts as a bridge, connecting different clusters of ideas.
- Foundational theories often exhibit high betweenness.
- If a peripheral node in  $G_{REF}$  acquires high  $C_B$  in  $G_{LLM}$ , this indicates:
  - Artificial conceptual bridge or
  - Possible logical mis-structuring.

# Eigenvector Centrality and Influence Displacement

- The **eigenvector centrality** score  $x_i$  for node  $i$  is defined as:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^n A_{ij} x_j,$$

where  $A$  is the adjacency matrix of the graph and  $\lambda$  is a constant. By the Perron-Frobenius Theorem, the scores are given by the components of the principal eigenvector corresponding to the largest positive eigenvalue  $\lambda_{max}$  of the adjacency matrix  $A$ .

- The eigenvector centrality measures the importance of a node not only by how many connections it has, but by how important its neighbors are.
- Hence, nodes connected to influential nodes gain influence, something that captures recursive importance.
- Compare eigenvector rankings across graphs  $G_{LLM}$  and  $G_{REF}$ .
- Artificial elevation indicates structural hallucination of authority.

# Upward Mobility Metric

- When nodal rank is defined by eigenvector centrality, define **mobility** of node  $v$  by:

$$M(v) = \text{rank}_{REF}(v) - \text{rank}_{LLM}(v)$$

- Large positive  $M(v)$  indicates suspicious elevation such as
  - inflated peripheral nodes
  - re-centering of conceptual authority or
  - artificial prestige shifts

# Core Preservation Test

- For a given positive integer  $k$ , define the **core** as the top- $k$  nodes ranked by eigenvector centrality.
- Let

$$K_{REF}^{(k)} = \text{Top-}k \text{ nodes in } G_{REF}$$

$$K_{LLM}^{(k)} = \text{Top-}k \text{ nodes in } G_{LLM}$$

- Then the **core overlap score** between the two graphs is defined by:

$$C_k = \frac{|K_{REF}^{(k)} \cap K_{LLM}^{(k)}|}{k}$$

- $C_k \approx 1$  indicates core preserved.
- Low  $C_k$  indicates canonical omission or re-centering.

# Hierarchy Flattening

- Hierarchy is reflected in dispersion of centrality values.
- Let  $c(v)$  denote eigenvector centrality.
- Denote the  $c(v)$  dispersion:

$$\sigma^2 = \text{Var}(c(v))$$

- Flattening is indicated by:
  - Reduced variance  $\sigma^2$  (i.e., more uniform degree spread)
  - Compression of top-central nodes
  - Increased uniformity in centrality distribution
- Flattening is interpreted as loss of:
  - Epistemic depth or
  - Vertical differentiation

# Modularity

- **Modularity** provides a complementary measure by quantifying the strength of thematic separation within the graph.
- It measures the extent to which a network is divided into distinct communities, or modules.
- Low modularity in an LLM-generated knowledge graph suggests a breakdown of conceptual boundaries, often indicating generic or overly homogenized representations of complex intellectual landscapes.
- *This can be interpreted as structural hallucination, insofar as the LLM appears unable to differentiate clearly between distinct concepts or disciplinary domains.*

# Community Detection Algorithms

- **Community detection** algorithms further refine the assessment by examining how concepts cluster into thematic or methodological groups.
- In well-structured academic texts, such clusters correspond to recognizable subfields or lines of argumentation.
- Several algorithms are available for community detection, including Louvain, Leiden, Girvan-Newman among others.  
vspace0.2cm
- Significant deviations in clustering patterns may indicate that an LLM has conflated unrelated themes or constructed artificial bridges between conceptually distinct communities.
- *This can be a powerful way to identify structural hallucinations that would be difficult to detect through other means.*

# Graph Density

Graph **density** measures the proportion of realized connections  $E$  relative to all possible connections in the set of nodes  $V$ :

$$D = \frac{2|E|}{|V|(|V| - 1)}$$

- Captures the overall "tightness" or connectivity of the network.
- High density → many concepts are linked.
- Low density → more selective, structured connectivity.

Interpretation in LLM Evaluation:

- **Over-generalization:** LLM graphs may become artificially dense.
- Excessive connectivity can indicate blurred conceptual boundaries.
- Dense graphs may reflect associative fluency rather than structural discrimination.
- Healthy disciplinary structure typically exhibits clustered, modular patterns—not uniform saturation.

# Graph Isomorphism and Edit Distance

- Structural comparison can be evaluated at two levels:
- **Graph Isomorphism:**
  - Two graphs are isomorphic if there exists a bijection between their nodes preserving adjacency.
  - Perfect isomorphism → identical structural organization.
  - Rare in practice, but provides a theoretical upper bound for alignment.
- **Graph Edit Distance (GED):**
  - Minimum number of edits (node/edge insertions, deletions, substitutions) required to transform  $G_{LLM}$  into  $G_{REF}$ .
  - High GED → low structural integrity.
  - Low GED → high structural alignment.
- These metrics provide a global quantitative score of representational correspondence between the two knowledge structures.

# Local Clustering Coefficient

## Local Cohesion of Conceptual Neighborhoods

For node  $v$ :

$$C(v) = \frac{2e_v}{k_v(k_v - 1)}$$

where:

- $k_v$  = the degree of  $v$  (i.e., the number of neighbors of  $v$ )
- $e_v$  = number of edges among those neighbors
- High clustering:
  - Dense interconnection among related concepts
  - Indicates specialized subfields
- Low clustering:
  - Concepts loosely connected
  - Suggests surface-level aggregation
- LLM flattening often reduces meaningful clustering while increasing global density.

# The "Hub and Spoke" Distortion

## Artificial Centralization Around Generic Nodes

- LLM graphs often produce hubs centered on generic abstractions:
  - “The Study”
  - “Researchers”
  - “The Model”
  - “The Theory”
- This displaces domain-specific technical hubs:
  - “Neural Networks”
  - “Bayesian Inference”
  - “The SIS and SIR Epidemic Models”
- Result:
  - Artificial centralization
  - Loss of disciplinary specificity
  - Epistemic dilution
- Diagnostic: Rank top- $k$  nodes and inspect semantic specificity of hubs.

# Path Length Analysis

## Conceptual Distance and Logical Depth

- **Average shortest path length** (where  $d(i,j)$  denotes the length of the shortest path between nodes  $i$  and  $j$ ):

$$L = \frac{1}{n(n-1)} \sum_{i \neq j} d(i,j)$$

- Measures conceptual steps between ideas.
- In reference graphs:
  - Moderate path lengths reflect layered reasoning
  - Intellectual lineage preserved through intermediate nodes
- In LLM graphs:
  - Very short paths may indicate conceptual shortcuts
  - Excessive compression → loss of nuance
  - Excessive length → incoherent drift
- Path structure reveals whether reasoning is compressed, fragmented, or structurally faithful.

# Visualizing the Stress Test

## Multi-Metric Structural Profile

- Compare structural metrics of  $G_{LLM}$  and  $G_{REF}$ :
  - Degree distribution
  - Betweenness centrality
  - Eigenvector centrality
  - Graph density
  - Clustering coefficient
- Radar (spider) charts provide a geometric fingerprint of structural behavior.
- Distortions in the radar shape signal specific failure modes:
  - Inflated density → over-generalization
  - Collapsed centrality variance → hierarchy flattening
  - Betweenness loss → missing bridges
- The stress test translates qualitative judgment into a structural signature.

# Multiple Outputs from Similar Prompts

- We can also measure the **structural stability** of LLM-generated knowledge representations.
- High variance in the graph structure across multiple outputs from near-identical prompts is a signal of unreliable or opportunistic generation.

# Practical Thresholds and Benchmarks

- To make these diagnostics practical, it is necessary to establish thresholds and benchmarks for what constitutes a significant deviation.
- This can be done by analyzing a large corpus of human-written texts and LLM-generated texts that have been manually evaluated for quality.
- This allows for the development of a standardized and objective methodology for hallucination detection.

# Session 4 Summary

- The **hallucination stress test** provides a powerful framework for evaluating the structural integrity of LLM-generated text.
- It combines a variety of network analysis techniques, including **centrality analysis**, **modularity analysis**, and **community detection**.
- These techniques allow us to systematically identify and quantify structural hallucinations.
- By using these tools, we can move towards a more responsible and accountable use of LLMs in scholarly research.

# Looking Ahead to Session 5

- Structure is internal; Citations are external.
- Next Session: Citation Integrity and Bibliometric Grounding.
- Grounding the graph in the "real world" of published science.

# Questions and Discussion

Thank you!

Questions?

[Moyses.Boudourides@northwestern.edu](mailto:Moyses.Boudourides@northwestern.edu)

[Moyses.Boudourides@gmail.com](mailto:Moyses.Boudourides@gmail.com)