Using and Evaluating LLMs in Academic Work

Session 5: Citation Integrity and Bibliometric Grounding

Moses Boudourides

Faculty, Graduate Program on Data Science
Northwestern University

Moses.Boudourides@northwestern.edu

Moses.Boudourides@gmail.com

instats Seminar

Friday, February 27, 2026
4:00 PM – 5:30 PM UTC

Northwestern | SCHOOL OF PROFESSIONAL STUDIES

Session 5: Citation Integrity and Bibliometric Grounding

1. Bibliographic Grounding
2. Bibliographic Failure
3. Citation Validation
4. Bibliometric Analysis

Beyond Conceptual Structure

Semantic–structural coherence is necessary but not sufficient.

A text may exhibit:

- Internal logical consistency
- Plausible conceptual architecture
- Apparent structural alignment

Yet still fail in the scholarly discourse.

Academic legitimacy requires:

- Proper citation practices
- Accurate attribution of intellectual labor
- Verifiable bibliographic grounding
- Transparent positioning within an existing corpus

We now shift from conceptual graphs to the **bibliographic layer** of evaluation.

The Role of Citation in the Knowledge System

Citations are structural devices, not ornamental additions.

They:

- Situate claims within an existing intellectual field
- Attribute intellectual labor and prevent epistemic appropriation
- Anchor arguments historically within lineages of thought
- Signal disciplinary alignment and methodological commitments
- Create traceable pathways for verification and critique

A structurally coherent argument without bibliographic grounding lacks epistemic accountability.

Bibliographic Failure Modes

LLM-generated texts may exhibit distinct forms of bibliographic distortion:

- **Fabricated references** (non-existent works)
- **Misattributed works** (incorrect author–work pairings)
- **Implausible citation contexts** (irrelevant or misleading linkage)
- **Distorted intellectual lineages** (false genealogies of ideas)

These failures occur at the level of:

- Existence
- Attribution
- Relevance
- Historical positioning

They undermine not fluency, but epistemic integrity.

Fabricated References

A fabricated reference is a bibliographic artifact that:

- Does not exist in authoritative databases (e.g., Web of Science, Scopus, Dimensions, OpenAlex etc.)
- Is syntactically and stylistically plausible
- Contains realistic author names, venues, and publication years
- Mimics legitimate citation formatting conventions

This form of hallucination is particularly dangerous because:

- It passes superficial credibility checks
- It may be propagated uncritically

Detection method:

- Programmatic verification via bibliographic database APIs
- DOI resolution verification

Misattribution

Misattribution occurs when bibliographic elements are real, but incorrectly linked.

- A real author is linked to a non-existent work
- A real work is credited to the wrong author
- A valid publication year is mismatched
- A real work is cited for claims it does not contain

Consequences:

- Erosion of intellectual credit
- Corruption of citation networks
- Distortion of scholarly authority

Misattribution represents structural misalignment within the bibliographic graph.

Implausible Citation

The most subtle distortion is the implausible citation.

- The cited work exists.
- The authors are correctly identified.
- The formatting is accurate.

Yet:

- The cited work does not substantively support the claim.
- The disciplinary context is mismatched.
- The citation inflates conceptual authority.

This produces:

- Artificial alignment
- Epistemic camouflage
- Bibliographic noise within the knowledge graph

The reference may be real, but its connection to the claim is inappropriate or misleading: The distortion lies in how the citation is used, not in whether it exists.

Citation Validation Protocol

Step 1: Reference Extraction

- Parse all cited references from the LLM-generated text.
- Normalize metadata (authors, title, year, venue, DOI).
- Remove formatting artifacts and duplicates.

Step 2: Programmatic Verification

Query authoritative bibliographic APIs:

- CrossRef
- OpenAlex
- Dimensions
- Web of Science etc.

Goal: establish existence, metadata integrity, and disciplinary placement.

Existence Verification

For each citation c:

$$\text{Verify}(c) = \begin{cases} 1, & \text{if record exists in authoritative database} \\ 0, & \text{otherwise} \end{cases}$$

Define hallucination rate:

$$H = 1 - \frac{1}{|C|} \sum_{c \in C} \text{Verify}(c)$$

- $H = 0$ → no fabricated references.
- $H > 0$ → presence of fabricated citations.

Existence failure signals **factual hallucination**.

Metadata Consistency Check

Beyond existence, verify metadata alignment:

- Author names (ordering and spelling)
- Publication year
- Journal or venue
- DOI resolution consistency
- Title similarity (fuzzy matching threshold)

Define consistency indicator:

$$\mathrm{Match}(c) \in [0,1]$$

Low match score → bibliographic distortion.

Metadata mismatch signals **structural bibliographic misalignment**.

Citation Network Construction

From validated references construct:

$$G_{CITE} = (V_{papers}, E_{citations})$$

- Nodes: cited papers (or authors)
- Edges: citation relationships among them
- Induced subgraph of the broader scholarly network

This graph represents the bibliographic footprint of the LLM text.

We now move from isolated references to relational structure.

Reference Baseline Network

Construct a baseline citation network G_{REF} from:

- Canonical literature corpus
- Structured bibliometric queries
- Field-defining review articles
- Authoritative disciplinary sources

G_{REF} serves as the structural benchmark for:

- Canonical centrality indices
- Intellectual lineage
- Citation density patterns

Canonical Work Detection

Identify high-centrality nodes in G_{REF}:

$$\text{Core}(G_{REF}) = \{p \mid \text{centrality}(p) \geq \tau\}$$

These represent:

- Foundational works
- Core methodological contributions
- Seminal theoretical advances
- High-impact review papers

Canonical nodes define the epistemic backbone of the field.

Omission Detection

Let K be canonical papers from G_{REF}.
Compute overlap:

$$O = \frac{|K \cap V_{LLM}|}{|K|}$$

- High $O \rightarrow$ foundational grounding preserved.
- Low $O \rightarrow$ canonical omission.

Omission signals **bibliographic shallowness**.

Citation Neighborhood Analysis

For each cited paper p:

- Examine its citation neighborhood in G_{REF} (**citation ego-centered graphs**).
- Analyze thematic clustering.
- Compare neighborhood structure with its placement in G_{LLM}.

Discrepancies indicate:

- Contextual misalignment
- Artificial cross-field linkage
- Spurious intellectual proximity

Author-Topic Consistency

For author a cited on topic T:

$$\text{Consistency}(a, T) = \mathbf{1}\left(\text{Publications}(a) \cap T \neq \emptyset\right)$$

- Empty intersection → implausible author-topic pairing.
- Weak intersection → marginal relevance.
- Strong intersection → disciplinary alignment.

This detects contextual hallucination without fabrication.

Bibliometric Benchmarking

Compare structural properties of G_{LLM} against the field baseline G_{REF}:

- Citation frequency distribution
- Journal / venue representation
- Author prominence and centrality
- Temporal citation patterns

Goal:

- Detect deviations from established field structure
- Quantify bibliographic alignment
- Identify systemic distortions rather than isolated errors

Benchmarking transforms citation evaluation into structural comparison.

Degree Distribution in Citation Graph

Citation networks typically exhibit:

- Heavy-tailed (power-law-like) distributions
- Core–periphery organization
- Few highly cited canonical works
- Many low-degree peripheral works

Let $P(k)$ denote degree distribution.
Flattening is indicated by:

- Reduced variance in $P(k)$
- Disappearance of high-degree canonical hubs
- Excess uniformity in citation counts

Deviation from heavy-tailed structure signals bibliographic distortion.

Journal Distribution Analysis

Examine venue representation:

- Are leading, field-defining journals present?
- Are high-impact venues appropriately represented?
- Is there over-concentration in marginal or generic outlets?
- Are venues topically aligned with the claim?

Journal distribution reveals:

- Disciplinary positioning
- Methodological alignment
- Field awareness

Venue imbalance may indicate shallow or opportunistic grounding.

Venue Concentration Metric

Let J denote the set of cited journals.
Define venue entropy:

$$H = -\sum_{j \in J} p_j \log p_j$$

where p_j is the proportion of citations from journal j.

Interpretation:

- High $H \rightarrow$ diversified disciplinary sourcing.
- Low $H \rightarrow$ concentration in few venues.

Abnormally low entropy suggests thematic narrowing or bibliographic bias.

Citation Age Distribution

Define citation age:

$$\Delta t = t_{\text{current}} - t_{\text{publication}}$$

Analyze distribution of Δt across cited works.

Indicators:

- Excess recency bias → shallow historical grounding.
- Absence of foundational older works → omission.
- Unrealistic clustering of publication years → synthetic generation.

Balanced age distribution reflects intellectual continuity.

Author Centrality in Citation Network

Construct author-level graph from G_{CITE}.
Compute:

- Degree centrality (citation frequency)
- Betweenness (brokerage between clusters)
- Eigenvector centrality (intellectual prestige)

Artificial elevation of marginal authors may indicate:

- Hub distortion
- Conceptual miscentering
- Authority inflation

Author prominence must reflect disciplinary reality.

Temporal Coherence

Assess chronological structure of citations:

- Are foundational works temporally prior?
- Is theoretical progression historically plausible?
- Are recent works building upon older ones?

Temporal incoherence may reveal:

- Reversed intellectual genealogy
- Artificial recency bias
- Flattened historical depth

Chronology is a structural constraint on scholarly integrity.

Cluster Structure in Citation Network

Apply community detection to the citation graph G_{CITE}.

Communities typically correspond to:

- Subfields within the discipline
- Methodological clusters
- Intellectual schools or paradigms
- Competing theoretical traditions

Compare cluster structure with baseline G_{REF}:

- Are core disciplinary clusters present?
- Are boundaries between clusters preserved?
- Are unrelated domains artificially merged?

Misaligned or collapsed clusters signal bibliographic misplacement and conceptual boundary erosion.

Cross-Prompt Citation Stability

Generate multiple LLM outputs under comparable prompts.

For each output i, construct a citation graph:

$$G_i = (V_i, E_i), \quad i = 1, \ldots, k$$

Evaluate:

- Persistence of canonical works across runs
- Stability of core citation hubs
- Consistency of venue representation

If citations fluctuate substantially across prompts, grounding may be stochastic rather than knowledge-driven.

Citation Stability Index

Define citation stability across k runs:

$$S_c = \frac{\left|\bigcap_{i=1}^{k} V_i\right|}{\left|\bigcup_{i=1}^{k} V_i\right|}$$

Interpretation:

- $S_c \approx 1 \rightarrow$ strong canonical persistence.
- Moderate $S_c \rightarrow$ partial stability.
- Low $S_c \rightarrow$ high citation volatility.

Low stability suggests opportunistic or prompt-sensitive citation generation, rather than robust disciplinary grounding.

Bibliographic Divergence Score

Rather than evaluating isolated indicators, we aggregate multiple signals into a composite index.

Combine:

- Canonical omission rate
- Degree distribution divergence
- Venue entropy deviation
- Citation stability index

Define composite score:

$$B = w_1 O + w_2 D_{\mathrm{deg}} + w_3 |H - H_{REF}| + w_4 (1 - S_c)$$

Higher B indicates greater bibliographic divergence from disciplinary structure.

Integration with Structural Diagnostics

Bibliographic distortion rarely occurs in isolation.
It often correlates with:

- Concept displacement (miscentered core concepts)
- Artificial bridges (spurious cross-domain links)
- Hierarchical flattening (loss of canonical prominence)
- Hub inflation (generic abstraction dominance)

True evaluation requires joint assessment of:

$$\text{Conceptual Structure} \quad + \quad \text{Bibliographic Grounding}$$

Fluency without citation integrity is structurally incomplete.

Documentation Protocol

Evaluation must be reproducible.
Researchers may append:

- Citation validation tables (existence + metadata checks)
- Canonical inclusion metrics
- Bibliometric divergence scores
- Stability indices across prompts
- API query logs

Transparency transforms AI evaluation from impressionistic critique to auditable methodology.

Governance as Ecosystem Protection

Unchecked bibliographic hallucination:

- Pollutes the scholarly record
- Propagates false citations into downstream research
- Distorts intellectual genealogies
- Inflates artificial authority structures
- Undermines cumulative knowledge formation

Governance is not censorship — it is protection of the epistemic ecosystem.

Session 5 Summary

- **Citation integrity** and **bibliometric grounding** are essential for responsible LLM use in scholarly practice.
- We have identified several forms of bibliographic failure, including fabricated references, implausible citations, and distorted citation neighborhoods.
- We have also outlined a set of protocols for validating the citation integrity of LLM-generated content and for grounding it within real scholarly ecosystems using bibliometric benchmarking.

Looking Ahead to Session 6

- How do we bring all this together in the classroom and the lab?
- Next Session: Attribution, Assessment, and Accountable Partnerships.
- Final conclusions on Methodological Governance.

Questions and Discussion

Thank you!

Questions?

Moses.Boudourides@northwestern.edu
Moses.Boudourides@gmail.com