HOW DEEPLY DO LLMS INTERNALIZE HUMAN CITATION PRACTICES? A GRAPH STRUCTURAL AND EMBEDDING-BASED EVALUATION GRAPH-STRUCTURAL

DATA ANALYTICS LABORATORY RESEARCH GROUP

Melika Mobini[1], Vincent Holst [1], Floriano Tori [1], Andres Algaba[1], Vincent Ginis[1,2]

[1]Vrije Universiteit Brussel , [2]Harvard University

CONTACT

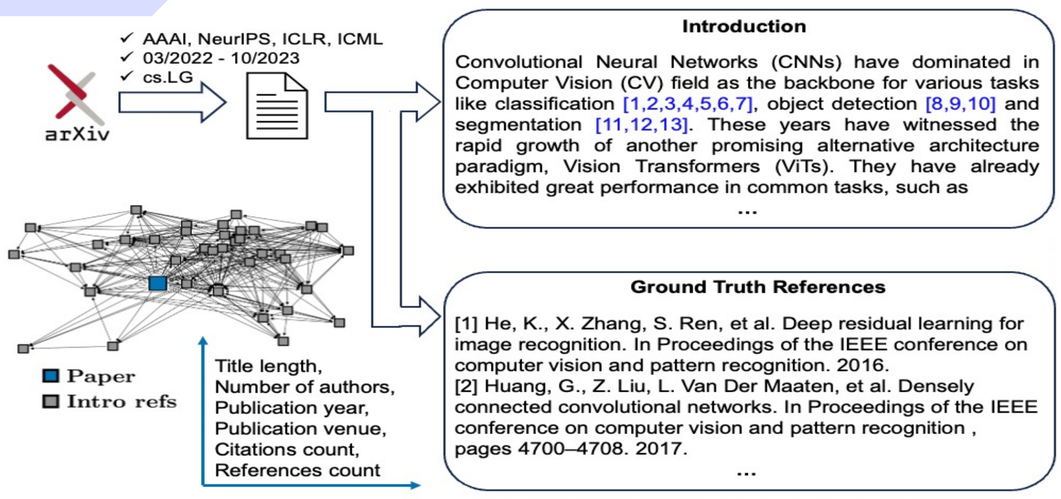
Melika.Mobini@VUB.be

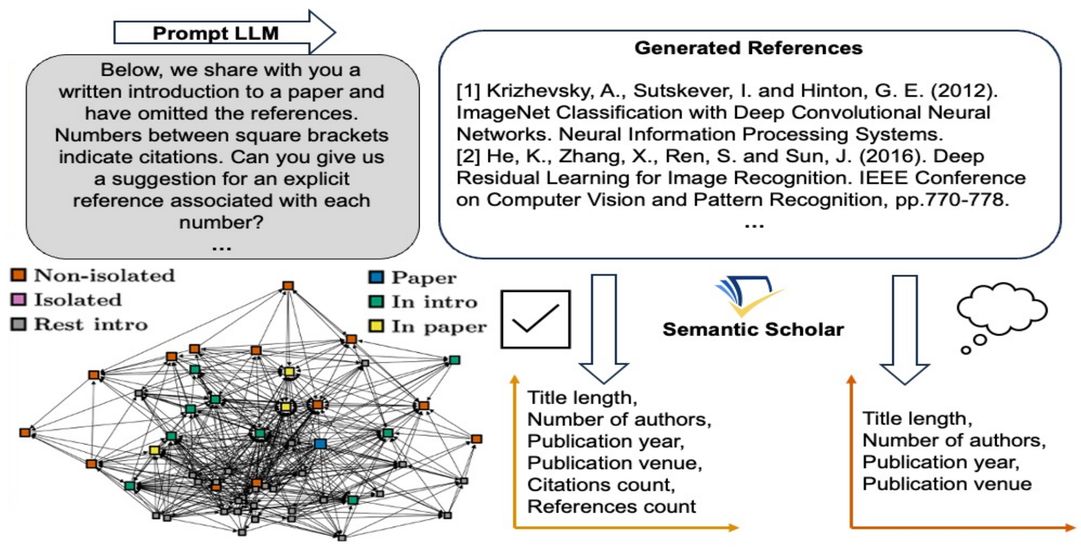
ICLR International Conference On Learning Representations

CONTEXT

- **LLMs in Scientific Research**: Assist in **literature synthesis** but may influence **citation practices**.
- **Key Concern**: Ensuring **integrity** in scientific communication and investigating **systemic biases**.
- **Human-AI Co-evolution**: AI-generated insights influence researchers, shaping future **LLM training**.
- **Beyond Citation Bias**: Investigate if LLMs internalize **citation structures meaningfully**.

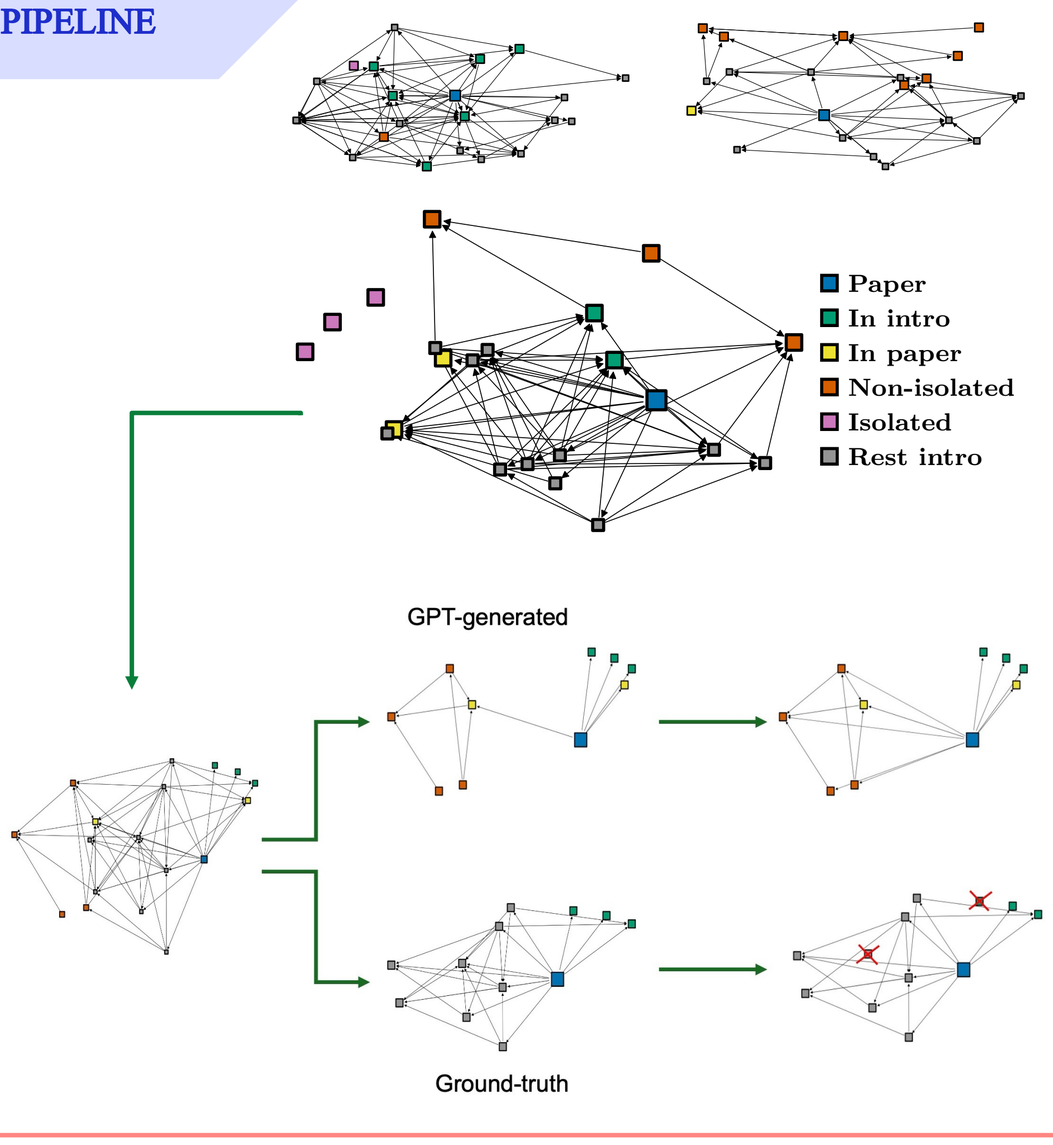
PIPELINE





- **Dataset:**
  - **166 papers** from *cs.LG.* (AAAI, NeurIPS, ICML, ICLR)
  - Papers **first appeared online after GPT-4's cut-off.** (March 2022 – Oct 2023)
  - **Extracted main content** separately from **ground truth references**.
- **LLM Citation Generation**:
  - **Suggested scholarly references** for anonymized in-text citations. (GPT-4, GPT-4o, and Claude 3.5)
  - **Existence check** via **Semantic Scholar**.

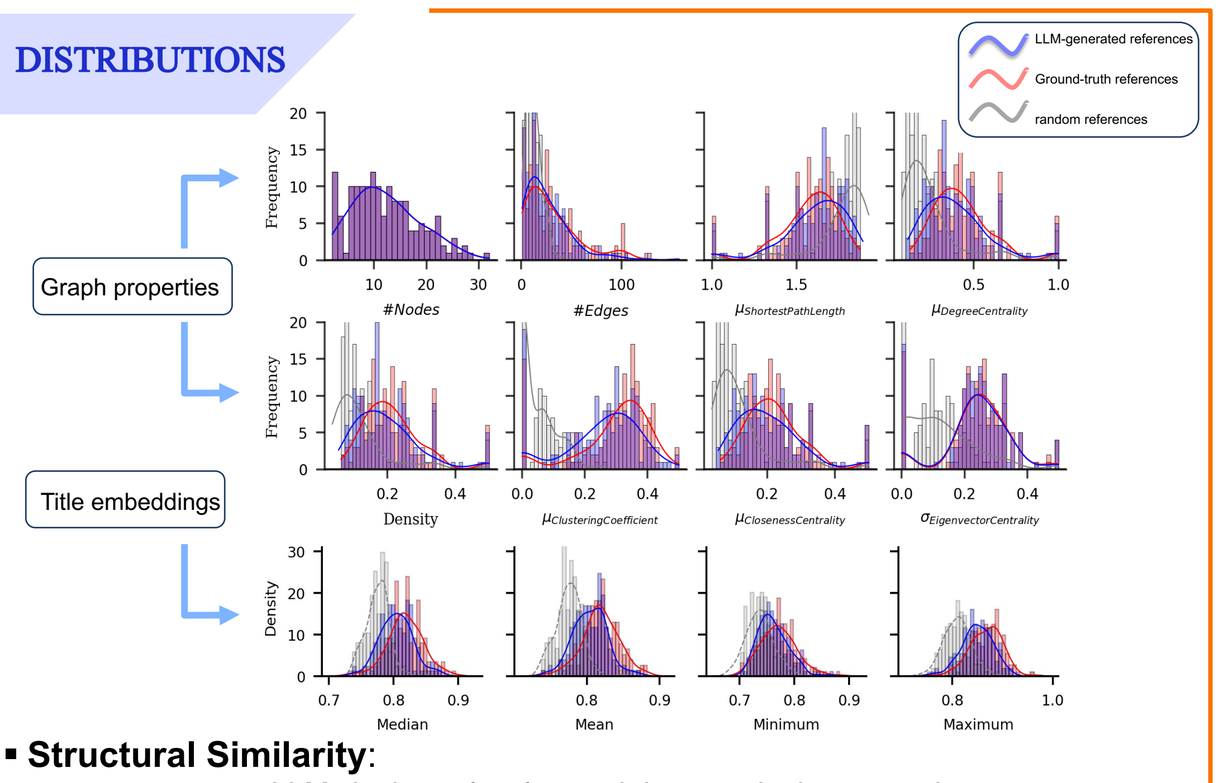
PIPELINE



- **Reference Categorization:**
  - **Focal paper → Blue**
  - **GPT-4 citations in the introduction → Green**
  - **GPT-4 citations appearing later in the paper → Yellow**
  - **GPT-4 citations linked to ground truth → Orange**
  - **GPT-4 citations linked to other generated references → Orange**
  - **Completely isolated GPT-4 citations → Purple**
  - **Ground-truth references not cited by GPT-4 → Grey**

- **Dataset**: 166 papers, each represented as a **citation network graph**.
- **Graph Construction**: Two distinct graphs per paper. **(GPT-generated vs. ground truth)**
- **Graph Count**: 332 graphs. (**166 GPT-generated + 166 ground-truth**)
- **Connectivity Check**:
  - Edges were added to ensure all references are linked to the **focal paper**.
- **Graph Simplification**:
  - Converted all graphs to **undirected format**.
- **Size Balancing**:
  - Randomly removed references from **ground-truth graphs.** (For a fair comparison)
- **Random baseline**
  - References **reshuffled** from papers in the same field

DISTRIBUTIONS



- **Structural Similarity**:
  LLM citations closely match human citation networks.
- **Cosine Similarity Analysis**:
  LLM citations align **closer** to human references than random ones.
- **Random Baseline**:
  Shows significant deviation from human and LLM citation structures.

RESULTS

Graph properties	Mean accuracy	Mean F1-score
Ground-truth vs. GPT	0.5167 ± 0.0224	0.5209 ± 0.0387
Ground-truth vs. Random	0.9271 ± 0.0264	0.9265 ± 0.0302
GPT vs. Random	0.9021 ± 0.0182	0.9066 ± 0.0168

Title embeddings	Mean accuracy	Mean F1-score
Ground-truth vs. GPT	0.6000 ± 0.0482	0.5998 ± 0.0653
Ground-truth vs. Random	0.8688 ± 0.0214	0.8720 ± 0.0187
GPT vs. Random	0.7396 ± 0.0132	0.7471 ± 0.0166

- **Random Forest Classifier performance description**:
  - **Evaluation Metrics:**
    Mean accuracy and F1-score from five independent runs.
  - **Features Used:**
    Graph-based properties & title embeddings.
  - **Dataset Split**:
    Training (70%) / Testing (30%)
    Using K-fold cross-validation.
- **Random Forest Classifier Results:**
  - **LLM-generated citations** and **human references:**
    Structurally and **semantically align closely**.
  - **LLM vs. Random** and **Ground Truth vs. Random**:
    Highly **distinguishable**.

LLMs **internalize citation behavior**, but risk **amplifying citation bias**.