

Evaluating the Impact of Automated Labeling on Retrieval Instability in Neural IR

William A. Ingram
Virginia Tech
Blacksburg, VA, USA
waingram@vt.edu

Abstract

Effective information retrieval (IR) depends on accurate relevance classification. But when the criteria are subjective or underspecified, small variations in classification can cause consequential shifts in retrieval results. The potential for such variability becomes critical for institutions when they use IR for research assessment. Retrieval instability can lead to relevant literature being overlooked, hindering a comprehensive understanding of the research landscape, and potentially undermining the validity of subsequent analyses and decisions.

We investigate this problem within the context of the United Nations Sustainable Development Goals (SDGs), a global framework for addressing environmental, social, and economic challenges. Scholarly research is vital for understanding, implementing, and monitoring SDG progress. Universities report SDG-related research to demonstrate impact, and international rankings incorporate SDG alignment into evaluations, influencing funding, policy, and institutional strategy. However, the nuanced nature of the SDGs makes it difficult to define what constitutes an SDG contribution [1]. Commonly used Boolean queries and controlled vocabularies for SDG retrieval cannot reliably differentiate substantive contributions (based on semantic relevance) from mere term occurrences.

In prior work, Large Language Models (LLMs) have been used to filter Boolean search results in systematic reviews by scoring documents for relevance to a specific information need [2]. Other studies demonstrate that LLMs can generate high-quality relevance labels for IR evaluation [4]. This prompted an investigation into using LLMs to judge SDG contribution through relevance filtering, which revealed variability in the judgments made by different LLMs on the same set of documents [3]. This observation suggests that the classification behavior of LLMs are sensitive to the specific parameters inherent to each model.

In this study, we prompt multiple LLMs to judge the SDG relevance of abstracts retrieved using Boolean queries. Abstracts judged relevant are used as positive training examples for fine-tuning multi-label SDG classifiers. We use these classifiers to simulate retrieval, applying fixed scoring functions to isolate fluctuations in ranking stability attributable to the different LLM relevance judgments. Our goal is to analyze how the structured signal of upstream inconsistencies in LLM-derived relevance judgments manifests as variations in

retrieval outcomes, providing a novel lens for investigating ranking stability under classification uncertainty. This research centers on three key questions:

- **RQ1:** How do different LLMs diverge in their filtering decisions, and what effect does this have on ranking stability in retrieval systems trained on filtered data?
- **RQ2:** Can divergence in labeling decisions be systematically explained or predicted from document content?
- **RQ3:** What distinguishes documents where LLMs disagree on relevance, and can these differences be predicted from lexical or surface-level features?

Using SDG classification as a case study of subjective relevance, we evaluate retrieval stability under classification uncertainty and address broader concerns regarding the reproducibility of LLM-based classification pipelines and their downstream effects.

CCS Concepts

• **Information systems** → **Information retrieval**; *Digital libraries and archives*; • **Computing methodologies** → *Natural language processing*.

Keywords

LLM disagreement, classification uncertainty, ranking stability

ACM Reference Format:

William A. Ingram. 2025. Evaluating the Impact of Automated Labeling on Retrieval Instability in Neural IR. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, July 13–18, 2025, Padua, Italy. ACM, New York, NY, USA, 1 page. <https://doi.org/10.1145/3726302.3730128>

References

- [1] Caroline S. Armitage, Marta Lorenz, and Susanne Mikki. 2020. Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results? *Quantitative Science Studies* 1, 3 (Aug 2020), 1092–1108. doi:10.1162/qss_a_00071
- [2] Fernando M. Delgado-Chaves, Matthew J. Jennings, Antonio Atalaia, Justus Wolff, Rita Horvath, Zeinab M. Mamdouh, Jan Baumbach, and Linda Baumbach. 2025. Transforming Literature Screening: The Emerging Role of Large Language Models in Systematic Reviews. *Proceedings of the National Academy of Sciences* 122, 2 (Jan. 2025), e2411962122. doi:10.1073/pnas.2411962122
- [3] William A. Ingram, Bipasha Banerjee, and Edward A. Fox. 2024. Agentic AI for Improving Precision in Identifying Contributions to Sustainable Development Goals. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData)*. IEEE, Washington, DC, USA, 8677–8679. doi:10.1109/BigData62323.2024.10825072
- [4] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models Can Accurately Predict Searcher Preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1930–1940. doi:10.1145/3626772.3657707

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGIR '25, Padua, Italy

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1592-1/2025/07

<https://doi.org/10.1145/3726302.3730128>