Exploring Artifacts Availability in Transportation Research Using Large Language Models

Junyi Ji[a,1], Ruth Lu[b,1], Linda Belkessa[c], Yongqi Dong[d],
Liming Wang[e], Bahman Madadi[f], Silvia Varotto[f], Nicolas Saunier[g], Greg MacFarlane[h],
Cathy Wu[b,*]

[a] Vanderbilt University, Nashville TN, United States
[b] Massachusetts Institute of Technology, Boston MA, United States
[c] Université Gustave Eiffel, Paris, France
[d] RWTH Aachen University, Aachen, Germany
[e] Portland State University, Portland OR, United States
[f] École Nationale des Travaux Publics de l'État (ENTPE), Vaulx-en-Velin, France
[g] Polytechnique Montréal, Montréal, Canada
[h] Brigham Young University, Provo UT, United States
[1] Equal contribution * Corresponding author

Extended abstract submitted for presentation at the ISTDM 2025
Montréal, Canada, September 3rd to 5th, 2025

February 28, 2025

Keywords: Publication Data; Metascience; Transportation Research Trends; Large Language Models

1 Introduction

The fields of engineering and science are increasingly integrating the availability of artifacts (i.e., code and data) as a fundamental research element in academic publications to enhance research reproducibility (National Academies of Sciences, Engineering, and Medicine, 2019). The availability of artifacts can contribute to the development of open datasets and benchmarks within the research community to accelerate scientific discovery. Transportation is an active research community (Sun & Yin, 2017), and the availability of artifacts could further enhance its momentum. This article targets the research question: *What is the current state of artifact availability in transportation research?*

Previous text mining technique, mostly using pattern matching for extracting the related context and keywords. Or rely on human interpretation of a small sample of articles (n=360) (Stagge *et al.*, 2019). It is a process that can either be labor-intensive, capturing detailed information, or semi-automated, providing rough insights for large-scale data. Large language models (LLMs) have the potential of context understanding (Zhu *et al.*, 2024) to strike a balance between resolution and scale in understanding the artifact availability in transportation research.

To tackle these challenges, we develop pipeline based on large language models that automates extraction of artifact availability features from full text articles and apply the pipeline to over 10,000 articles published in Transportation Research Series between 2019 and 2024. We first retrieve full-text via the Elsevier API[1] for the articles published in flagship transportation journals. After manual validation on a sample dataset and improve on the prompts, we deploy the large language models (LLMs) to extract artifacts availability features on the full text articles. In this abstract, we demonstrate the performance of the pipeline on a well-labeled dataset and show

[1] Text and data mining via Elsevier API, https://www.elsevier.com/about/policies-and-standards/text-and-data-mining

the result on the testing dataset[2]. Future work will focus on rigorously validating the feature extraction process and explore the trend and pattern in artifact availability in articles published in Transportation Research series. Our method and data will be made available publicly and can be used by researchers in other fields to explore artifacts availability in their discipline.

2 Data

2.1 Articles retrieval and processing

Previous research on mining publications typically use metadata for analysis. The Elsevier API offers an opportunity to explore full-text information, enabling a deeper understanding of publication details related to artifacts availability. The articles are selected from 2019 and 2024 in flagship transportation journals, in particular Transportation Research Parts A-F and Interdisciplinary Perspectives. Figure 1 shows the number of publication trend in the selected journals. Each article is stored in XML[3] format and identified by its Digital Object Identifier (DOI).

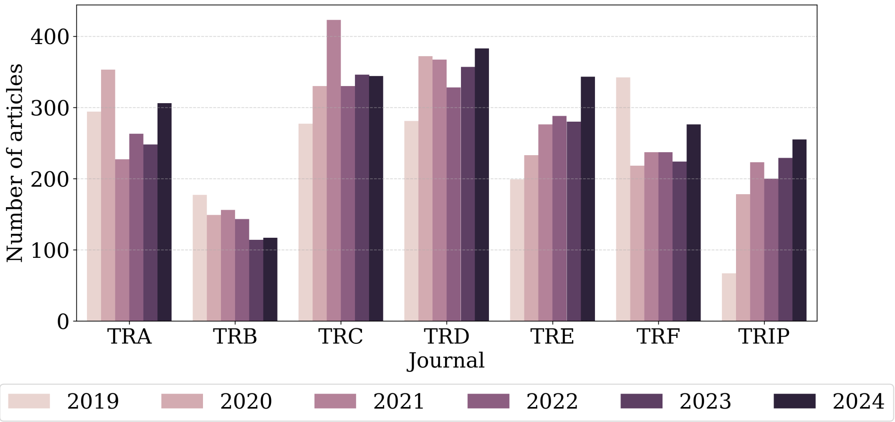

Figure 1: Number of articles published in various journals (TRA, TRB, TRC, TRD, TRE, TRF, and TRIP are abbreviations for the different parts of Transportation Research) from 2019 to 2024 (n=10,990).

2.2 Dataset split: Manual validation, testing and full dataset

(i) Manual validation dataset (MVD): The data is manually labeled by researchers in transportation based on predefined guidelines. It contains 30 articles, randomly drawn from the broader dataset of 10,000 articles. Each article is reviewed by two independent reviewers, and if there is a disagreement, a third reviewer is consulted to resolve the conflict. Our proposed pipeline is evaluated and fine-tuned on this dataset.

(ii) Testing dataset: The data is a random sample of 1,000 articles from the full dataset. It is labeled with the proposed pipeline and used to check for data quality issues.

(iii) Full dataset: The 10,000 articles, which are supposed to be processed with the proposed pipeline to deliver the results.

3 Method

3.1 Features design

To quantify the artifacts availability, A complete list of the designed features is presented in Figure 2. In this extended abstract, we primarily explore the following features (listed below) and evaluate the performance of the proposed pipeline on MVD.

[2] Note that the extended abstract is a work-in-progress submission, and the results may be revised as the methodology is further refined.

[3] Extensible Markup Language (XML), with a sample schema available at here.

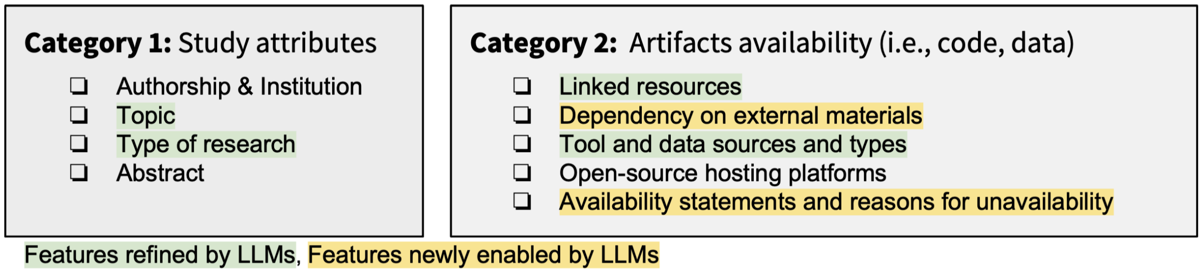
Category 1: Study attributes
❑ Authorship & Institution
❑ Topic
❑ Type of research
❑ Abstract
Category 2: Artifacts availability (i.e., code, data)
❑ Linked resources
❑ Dependency on external materials
❑ Tool and data sources and types
❑ Open-source hosting platforms
❑ Availability statements and reasons for unavailability
Features refined by LLMs, Features newly enabled by LLMs

Figure 2: Features extracted related to study attributes and artifact availability

(i) *Is the article dependent on data for reproducibility?* If the article used data in research, is the data needed for reproducing the results by external researchers?

(ii) *Data type.* Data type includes collected, synthetic, both or neither. *Collected data* is obtained directly from real-world sources such as sensors, experiments, or surveys. *Synthetic data* is artificially generated using simulations or models.

(iii) *Is data and code publicly available?* Does the paper explicitly link the code and data of this paper as an open-source resource?

3.2 Feature extraction: Prompt engineering

The current pipeline is based on the prompt interacting with Google Gemini 2.0 (Google, 2024) is used for feature extraction. Our current pipeline is to leverage Gemini's long-token processing capabilities to extract features directly.

4 Results: Evaluation on MVD

Figure 3 visualized the distribution of data dependencies, data types, and availability within MVD.

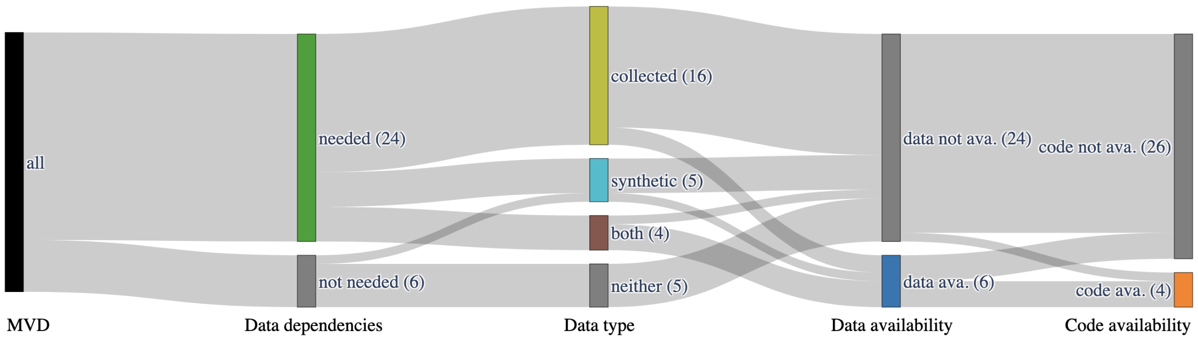

Figure 3: Sankey diagram on MVD (n=30): Note that the manual label can also be incorrect in some features.

Due to the sparsity of the data and code availability, we evaluate the performance of the proposed pipeline based on the false positive Rate (FPR), false negative rate (FNR), and accuracy from the confusion matrix on MVD.

Metric	**data dep.**	**code**	**data**	**both**	**collected**	**neither**	**synthetic**
FPR	0.143	0.000	0.148	0.038	**0.286**	0.040	0.040
FNR	0.000	0.200	0.333	**0.750**	0.125	0.200	0.200
Accuracy	0.967	0.967	0.833	0.867	0.800	0.933	0.933

Table 1: Performance evaluation metrics (False Positive Rate, False Negative Rate, and Accuracy) for different feature categories.

The *data dependencies* and *code availability* features yield the most effective results, while the data type, especially the *both* and *collected* category, struggle with high false negatives and false positives, respectively. We showcase the scalability of the proposed pipeline by applying it to the testing dataset, as illustrated in Figure 4.

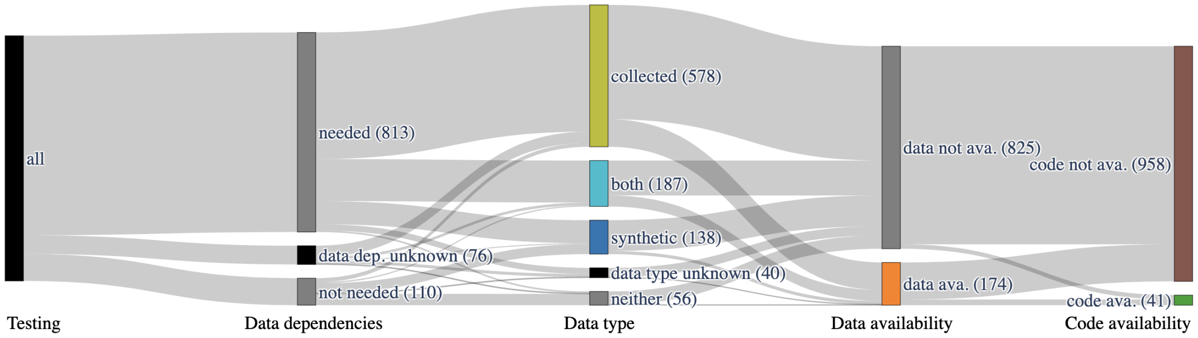

Figure 4: Sankey diagram on testing dataset (n=1000): Note that this demonstrates the scalability of the proposed pipeline, but the current accuracy is not final; it is presented here as a proof of concept.

5 Summary and future work

This extended abstract serves as a proof of concept for the proposed LLM-integrated pipeline for measuring artifacts availability. The current prompts and pipeline have limitations in accurately interpreting certain features but serve as a benchmark for refinement. With the pipeline to be improved, it could enable authors, reviewers, and editors of journals to verify the availability of artifacts before submission and publication. For example, we use the pipeline to extract the features for this abstract submission, the output can be accessed here. We intent to release the tool and data analysis code presented in this article to support reuse of artifacts and research reproducibility.

Data and code availability statements

The data supporting this study is currently unavailable as it is still being processed. Code and prompts used in this abstract is available at `https://github.com/RRinTransportation/ISTDM25-artifacts`.

Acknowledgment

This research utilized data accessed through Elsevier's APIs in compliance with Elsevier's Text and Data Mining policy. We are grateful for the Reproducible Research in Transportation Engineering (RERITE) Working Group for providing us with the opportunity to build and strengthen this collaborative team. We also appreciate the insightful feedback from Dr. Micah Altman at MIT Libraries.

References

Google. 2024. *Gemini 2.0: Our latest, most capable AI model yet.* Accessed: 2025-02-28.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and replicability in science.* National Academies Press.

Stagge, James H, Rosenberg, David E, Abdallah, Adel M, Akbar, Hadia, Attallah, Nour A, & James, Ryan. 2019. Assessing data availability and research reproducibility in hydrology and water resources. *Scientific data*, **6**(1), 1–12.

Sun, Lijun, & Yin, Yafeng. 2017. Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies*, **77**, 49–66.

Zhu, Yutao, Zhang, Peitian, Zhang, Chenghao, Chen, Yifei, Xie, Binyu, Liu, Zheng, Wen, Ji-Rong, & Dou, Zhicheng. 2024. INTERS: Unlocking the Power of Large Language Models in Search with Instruction Tuning. *Pages 2782–2809 of: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Bangkok, Thailand: Association for Computational Linguistics.