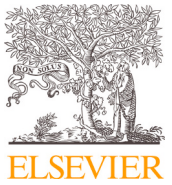
Transportation Research Part C 194 (2027) 106024



Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc



Measuring the state of open science in transportation using large language models

Junyi Ji [a,1,*], Ruth Lu [b,1], Linda Belkessa [c,2], Liming Wang [d,2], Silvia Varotto [e,2], Yongqi Dong [f], Nicolas Saunier [g], Mostafa Ameli [c], Gregory S. Macfarlane [h], Bahman Madadi [e], Cathy Wu [b,*]

[a] Vanderbilt University, Nashville, TN, United States
[b] Massachusetts Institute of Technology, Cambridge, MA, United States
[c] Université Gustave Eiffel, Champs-sur-Marne, France
[d] Portland State University, Portland, OR, United States
[e] École Nationale des Travaux Publics de l'État (ENTPE), Vaulx-en-Velin, France
[f] Delft University of Technology, Delft, The Netherlands
[g] Polytechnique Montréal, Montréal, QC, Canada
[h] Brigham Young University, Provo, UT, United States

ARTICLE INFO

Keywords:
Open science
Open science monitoring
Transportation research
Large language models
Open-source ecosystems

ABSTRACT

Open science initiatives have strengthened scientific integrity and accelerated research progress across many fields, but the state of their practice within transportation research remains under-investigated. Key features of open science, defined here as data and code availability, are difficult to extract due to the inherent complexity of the field. Previous work has either been limited to small-scale studies due to the labor-intensive nature of manual analysis or has relied on large-scale bibliometric approaches that sacrifice contextual richness. This paper introduces an automatic and scalable feature-extraction pipeline that combines full-text extraction and human agreement analysis in order to measure code and data availability in transportation research. We employ Large Language Models (LLMs) for this task and validate their performance against a manually curated dataset and through an inter-rater agreement analysis. We applied this pipeline to examine 10,724 research articles published in the Transportation Research Part series of journals between 2019 and 2024. Our analysis found that only 5% of quantitative papers shared a code repository, 4% of quantitative papers shared a data repository, and about 3% of papers shared both, with trends differing across journals, topics, and geographic regions. We found no significant difference in citation counts or review duration between papers that provided data and code and those that did not, suggesting a misalignment between open science efforts and traditional academic metrics. Consequently, encouraging these practices will likely require structural interventions from journals and funding agencies to supplement the lack of direct author incentives. The pipeline developed in this study can be readily scaled to other journals, representing a critical step toward the automated measurement and monitoring of open science practices in transportation research. An interactive explorer can be found on the project page https://www.rerite.org/MOST.

* Corresponding authors.
E-mail addresses: junyi.ji@vanderbilt.edu (J. Ji), cathywu@mit.edu (C. Wu).
[1] Equal contribution.
[2] Equal contribution.

https://doi.org/10.1016/j.trc.2026.106024
Received 20 January 2026; Received in revised form 28 July 2026; Accepted 7 September 2026
Available online 22 October 2026
0968-090X/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

1.1. Motivation and vision

Open science initiatives have accelerated research progress across many disciplines, including computational science (Peng, 2011), psychology (Open Science Collaboration, 2015), and statistics (Stodden, 2014). Recently, open science practices have gained increasing recognition in the transportation research community, as reflected by the introduction of data availability statements in Transportation Research journals (Transportation Research Part C, 2025a). Notably, *Transportation Research Part C* explicitly emphasizes and promotes open science in its aims and scope, stating that "Special emphasis is given in open science initiatives and promoting the opening of large-scale datasets that can support transferability and benchmarking of different approaches (Transportation Research Part C, 2025b)." However, despite these policy advancements, the extent to which open science practices have been adopted and implemented in transportation research remains largely unexamined.

The idea of open science (Woelfle et al., 2011) is to make scientific knowledge *openly available, accessible, and reusable*, as defined by the United Nations Educational, Scientific and Cultural Organization (UNESCO) (UNESCO, 2022). As open science has gained awareness (Winker et al., 2023; Gentemann, 2023), open science monitoring (OSM) has emerged as a research direction and has been endorsed by UNESCO. OSM is not merely bookkeeping: transparent, responsible measurement is a practical mechanism for guiding, evaluating, and accelerating the transformation of research practices. In 2025, core principles for OSM were established (Bobrov et al., 2025), emphasizing (i) relevance and significance, (ii) transparency and reproducibility, and (iii) self-assessment and responsible use.

Open science encompasses the ideas of reproducibility and replicability in science (National Academies of Sciences et al., 2019), where reproducibility refers to the ability to obtain the same results using the same data and methods, while replicability refers to the ability to obtain similar results using different data or methods. Making data and code available is the first and essential step towards achieving reproducibility and replicability (Crüwell et al., 2019), as it allows other researchers to verify and build upon the original work. A recent study within the artificial intelligence (AI) research community suggests that for the purpose of successful replication, the quality of shared code is a secondary consideration to its mere availability (Gundersen et al., 2025). This highlights the essential role of open code and data in promoting both reproducibility and replicability and motivates a systematic measurement of open science practices in transportation research, with a particular focus on the availability of data and code in papers published in Transportation Research journals. These motivate us to build a repeatable, low-cost system for monitoring open code and data practices in transportation research, helping the field make steady, evidence-based progress toward broader open science adoption.

1.2. Challenges

Measuring the state of open science practices is a complex and resource-intensive endeavor (Stagge et al., 2019; Yang et al., 2020; Youyou et al., 2023). Traditionally, this involves manually reviewing each paper to assess the availability of data and code (Stagge et al., 2019), as well as verifying the validity of provided resources. Such manual processes are labor-intensive, time-consuming (Stagge et al., 2019; Pineau et al., 2021; Seibold et al., 2021; Olszewski et al., 2023), and do not scale efficiently to large corpora of research articles. And without proper training and clear instructions, manual annotation, especially when involving multiple annotators, can easily produce inconsistent and inaccurate results. A recent study (Riehl et al., 2025) developed automated approaches based on keyword and link extraction using regular expressions to improve scalability. The proposed pipeline in Riehl et al. (2025) is reproducible and extensible, and we therefore benchmark it in this work as well. However, the regular expressions often fail to capture the surrounding context and thus generate false positives, for example, detecting benchmark repositories or GitHub links in citations as evidence of sharing. The trade-offs between accuracy and scalability are summarized in Table 1.

Furthermore, transportation research poses unique challenges for measuring open science practices because "data and code availability" is often conceptually under-specified and therefore difficult to operationalize. In particular, the boundary between using publicly accessible resources and making research artifacts available is frequently blurred. For example, many studies rely on widely used public datasets such as NGSIM (U. S. Department of Transportation Federal Highway Administration, 2016), yet do not release the specific processed or derived datasets they create for tasks such as calibrating car-following or lane-changing models. In such cases, the study benefits from open data, but the research-ready data needed to replicate the analysis may not be available. Conversely, even when raw data are proprietary or sensitive, researchers can often share post-processed or aggregated derivatives that enable partial or substantial reuse. These scenarios raise a measurement dilemma: should a paper be credited for "using publicly available data," for "making data available," for both, or for neither? This ambiguity implies that data and code availability in transportation research is not binary, but spans a spectrum from minimal reliance on existing open resources to proactive dissemination of reusable

Table 1
Accuracy and scalability trade-offs for code and data availability extraction from publications. We analyze LLM accuracy relative to human annotators in Section 5.4.2 through a multi-rater agreement analysis, which further confirms LLM capability for such feature extraction.

	Slow and labor-intensive	Fast and scalable
Accurate	Manual annotation (clear guidelines)	LLMs (this study)
Inaccurate	Manual annotation (with ambiguity)	Regular expressions

artifacts that sustain open science. These definitions need to be formalized upfront to ensure transparency and reproducibility before any measurement is conducted, regardless of the method used.

1.3. Problem statement and contributions

To address these challenges, this study first formalizes the operational definitions and decision rules required to distinguish between using or citing an existing public dataset, releasing processed research data, and releasing research code. We then introduce an automated framework that balances accuracy with scalability to measure open science practices in transportation research. Focusing on code and data availability as observable artifacts in full-text articles, we develop an extraction pipeline that using large language models (LLMs) to capture contextual signals that keyword searches and regular expressions can miss. The pipeline processes Elsevier `XML` files at a scale of 10,000 papers, identifies artifact links, classifies availability, and is evaluated against a human validation dataset.

Using these large-scale extracted features, we address the following research questions:

(i) What is the proportion of papers that make code and data available?
(ii) Which factors are associated with the availability of code and data?
(iii) Is code and data availability simply a matter of time, i.e., to what extent is open science behavior incentivized, or are external policies warranted?

The broader vision of this study is to establish a framework for monitoring open science practices in transportation research. Rather than emphasizing current gaps, we view the measurement of open science practices as a catalyst which enables the community to identify common challenges, recognize progress, and foster a culture of collaboration and reproducibility that advances scientific rigor and innovation in transportation research.

The contributions of this paper are summarized as follows:

(i) We develop and implement a scalable pipeline that extracts code and data availability and its contributing factors in full-text Transportation Research articles, enabled by LLMs.
(ii) We validate all LLM-extracted features with a manually annotated dataset and keep only those that meet established inter-rater agreement standards for downstream analysis.
(iii) We provide the first large-scale, field-wide analysis of the state of open science practices in transportation research, capturing the effect of temporal, topical, and geographical factors on code and data availability and outlining implications for accelerating open science in the community. An interactive explorer can be found on the project page https://www.rerite.org/MOST.

The scope of our work focuses on sharing code and data used in research via online repositories. A full glossary of open scholarship terms can be found at Parsons et al. (2022). There are many other dimensions of open science and reproducibility that are outside the scope of this paper, such as paywalls, institutional access, quality of documentation, computational reproducibility, and more. We hope that these can be further explored in future works.

1.4. Organization of the paper

The remainder of the paper is organized as follows. Section 2 situates the study within related work on open science in transportation and open science monitoring. Section 3 defines the terminology and operational criteria used for identifying data and code availability. Section 4 outlines the full-text dataset curation and the manual annotation procedure. Section 5 presents the automatic feature extraction pipeline, including meta-data parsing, topic modeling, LLM-based extraction, agreement analysis, and postprocessing. Section 6 reports descriptive findings and estimates choice models to examine factors associated with data and code availability. Section 7 discusses the results and performance. Section 8 provides implementation guidance and community-facing recommendations. Section 9 concludes the paper, summarizes limitations, and identifies directions for future work.

2. Related work

2.1. Open science in transportation research

2.1.1. Signs of momentum

Open science practices, including open-access publishing, open data, and open-source software, support reproducibility by making research materials transparent and accessible. Awareness of open science principles in transportation research, particularly regarding availability, transparency, and reproducibility, emerged in the 2010s. To address the “data deluge” in modern transportation research, comply with U.S. federal government policy changes, and maximize returns on research investments, Purdue University Libraries and its Joint Transportation Research Program (Newton et al., 2012; Zilinski et al., 2014) developed workflows for archiving transportation research outputs and data. These workflows covered creation, publication, and dissemination planning. These efforts highlighted a recurring issue: while many researchers were willing to share their data, they often lacked effective mechanisms to do so. Parallel developments occurred in Europe. Open science became a legal obligation under Horizon Europe (2021–2027) (Nielsen et al., 2023), the EU’s flagship program for supporting research and innovation. A common challenge is that research data is typically confined to internal use and often becomes unavailable after project completion. To address this, a data sharing framework (DSF) for naturalistic

driving studies (NDS) (Gellerman et al., 2016) was developed and demonstrated through the EU UDRIVE project (Eenink et al., 2014), the first EU project to explicitly address data-sharing prerequisites.

Similar reuse needs exist for code and tools in transportation research. As estimated in Tamminga et al. (2012), 30%-50% of the total effort required to develop a simulation case study is devoted to data-related tasks and network preparation, including calibration and validation that rely on extensive field data. Compounding this inefficiency, a significant portion of the effort is spent rebuilding basic functionality and repeatedly reinventing the wheel. Making tools and code openly available could therefore lead to substantial collective savings in time and effort. Motivated by this need, Tamminga et al. (2012) subsequently developed `OpenTraffic` (Tamminga et al., 2014), which provides standardized, multi-scale traffic problems and shared network settings to support researchers worldwide in developing and comparing methods.

More efforts have been conducted in specific transportation sub-fields, for example, car-following model calibration (Punzo et al., 2021), sensor-based mobility mode recognition (Wang et al., 2019), and discrete choice models for travel demand modeling (Wang et al., 2024). Additional community-driven initiatives have also emerged. The Zephyr Foundation, established in 2017, promotes open-source software and data sharing in transportation research, particularly in transportation and land-use decision-making. The International Association of Travel Behaviour Research (IATBR) has similarly advocated for a fully open science framework in which research outputs are freely accessible and reproducible (Caicedo et al., 2025). REproducible Research In Transportation Engineering (RERITE), a transportation research community the authors belong to, is also actively promoting open science principles (Wu et al., 2024a), developing best practices, and fostering community engagement. Note that transportation research is largely data driven, and widely used open datasets such as NGSIM (U. S. Department of Transportation Federal Highway Administration, 2016), PeMS (Choe et al., 2002), Transportation Networks (Transportation Networks for Research Core Team, 2025), and Swissmetro (Bierlaire et al., 2001) have played a major role in supporting empirical and methodological advances. Progress in adjacent fields like automated driving and computer vision has been strongly enabled by open science practices, where large-scale public datasets including `nuScenes` (Caesar et al., 2020) and the Waymo Open Dataset (Sun et al., 2020), Waymo Open Motion Dataset (Ettinger et al., 2021) have each accumulated thousands of citations and set expectations for transparency, benchmarking, and reproducibility. We also see a growing wave of bottom-up open data efforts from our own research community (Barmpounakis and Geroliminis, 2020; Seo et al., 2020; Gloudemans et al., 2023; Yabe et al., 2024a,b; Schicktanz et al., 2025). To help promote these initiatives, we maintain a curated list at `awesome-rerite`, collecting them to the best of our knowledge.

2.1.2. Persistent barriers and opportunities

Despite recent progress, substantial barriers to sharing data and code remain. Transportation research not only faces these general challenges but often amplifies them due to the nature of the data and the ecosystem in which it is produced. Transportation research operates under real-world constraints and complexities that make openness nontrivial. These constraints include privacy protections (Folla et al., 2021; Wood and van Schalkwyk, 2024; Shi et al., 2025), the large size and sensitivity of datasets (Nielsen et al., 2023), the heterogeneity of data sources (Welch and Widita, 2019), the diversity of regional characteristics (Sun and Kirtonia, 2020), and even geopolitical security concerns (Crampton, 2015; Nikander et al., 2024; Zhang, 2024), all of which complicate efforts to make data and code publicly available. They are further compounded by the need for active collaboration among diverse stakeholders (Nie, 2025), including academic institutions, government agencies, private companies, and the public. These intertwined constraints often limit the degree to which data and code can be openly shared, even when researchers are willing to do so.

From a research perspective, transportation research is highly interdisciplinary (Sun and Yin, 2017), spanning fields such as physics, economics, control theory, operations research, social science, psychology, computational science, and environmental science (Mahmassani, 2025). This diversity brings valuable perspectives but also creates inconsistencies in data formats, modeling approaches, and expectations for transparency and reproducibility. Researchers from different backgrounds often hold different views on what constitutes shareable data or reusable code, making it difficult to establish consistent open science practices across the field. Meanwhile, when data and code are made openly available, this diversity becomes a strength: researchers can reuse resources across domains, build on one another's methods, and accelerate collective progress. Ultimately, greater openness can bridge disciplinary divides, connect research with practice (Levin and Stern, 2024), and foster a more collaborative and transparent research culture.

2.2. Science of science in transportation research

The measurement of open science practices belongs to the broader field of *science of science* (SciSci), which uses large-scale literature data to characterize the structure and dynamics of research itself and to inform science policy and research management (Fortunato et al., 2018). Data and code availability is one such measurable dimension (Liu et al., 2023). SciSci methods have been applied extensively in transportation to understand the state of the field and its evolution. Bibliometric studies have examined research trends in individual journals (Jiang et al., 2020; Antoniou et al., 2025), the field as a whole (Sun and Yin, 2017; Modak et al., 2019; Sun and Kirtonia, 2020; Wandelt et al., 2025b,a; Zhao et al., 2026), and specific sub-fields (Tamakloe and Park, 2023; Ma et al., 2022), while co-authorship and collaboration networks have also been investigated (Sun and Rahwan, 2017). A related line of work analyzes the review process itself, including the reasons papers are rejected (Wu et al., 2024b) via 5000 rejected paper and the corresponding review comments and the governance of peer review in transportation research (Bae, 2026).

A common feature of these studies is their reliance on *structured* signals, for example titles, abstracts, keywords, author affiliations, and citations, that are readily extracted from bibliographic metadata. Far less attention has been paid to the *unstructured* full text of articles, which is precisely where evidence of open science practices, such as statements of code and data availability, typically resides. Our work targets this gap by extracting and analyzing these unstructured signals at scale from full text paper.

2.3. Monitoring and measuring open science practices

From a monitoring perspective, measurement of data and code availability falls into two regimes, depending on how journals and venues capture this information:

- Unstructured: publications may mention data and code availability, but do so in free text and inconsistent locations (e.g., acknowledgments, footnotes, methods, supplementary material notes, or informal "data available upon request" language). In this regime, reliable monitoring requires robust extraction methods from full-text documents.
- Structured: publications provide standardized, machine-readable metadata (e.g., required fields or standardized statements) for data and code availability. In this regime, monitoring is comparatively straightforward: extraction is largely a matter of parsing consistent metadata, and extensive post-processing is unnecessary.

At present, transportation research largely remains in the unstructured regime. As a result, the central challenge is not just defining what availability means, but also reliably extracting availability features at scale from full-text publications. This study therefore focuses on the unstructured case, while also motivating the structured case as a desirable endpoint: if structured data and code metadata became standard practice, the need for complex extraction tools would diminish substantially. We return to this distinction in the discussion of practices and future monitoring directions.

2.3.1. Unstructured

In many fields, measurement efforts start out unstructured, and clarity in definitions determines what should be measured. In signal processing, Vandewalle et al. (2009) proposed a six-level reproducibility scale, ranging from complete irreproducibility to effortless reproduction within 15 min using standard tools. More broadly, Goodman et al. (2016) note that the language and conceptual framework of "research reproducibility" remain nonstandard and unsettled across the sciences. To avoid misunderstandings that arise when reproducibility terminology is treated as a proxy for "truth," they consolidate prior taxonomies into three categories: methods reproducibility, results reproducibility, and inferential reproducibility. At a more operational level, reproducibility assessment can be staged from basic to stringent checks: whether code and data are available, whether the code runs, whether it runs with minor fixes, and whether it reproduces the reported results. Additional signals of artifact quality, such as README completeness and documentation, can further inform these assessments. In transportation research, Zheng (2021) proposed three requirements for reproducible publications: (i) sharing the scripts/codes for generating the entire paper (data analysis, figures, tables, models) using open-source programming languages, (ii) openly releasing the data, and (iii) clearly documenting the software environment.

Existing evaluation efforts are labor-intensive and therefore difficult to sustain beyond small scales. At the same time, simple pattern-matching approaches (Riehl et al., 2025) cannot reliably operationalize detailed definitions of reproducibility or extract the required features. This creates a dilemma between scale and accuracy: manual assessments do not scale, while automated keyword-based methods lack fidelity. As a result, neither approach consistently satisfies key open science monitoring principles, particularly transparency, reproducibility, and responsible use.

Emerging technologies, particularly large language models (LLMs), have demonstrated strong effectiveness in processing and analyzing complex textual data (Gilardi et al., 2023), creating new opportunities for open science monitoring. For example, `Reproscreener` (Bhaskar and Stodden, 2024) illustrates how LLMs can support post-publication monitoring by automating checks of machine learning papers for the presence of information and documentation needed to enable replication. However, the presence of such reporting elements does not necessarily guarantee computational reproducibility. Building on this line of work, Siegel et al. (2024) introduced the Computational Reproducibility Agent Benchmark (`CORE-Bench`), a benchmark suite designed to evaluate LLMs' capacity to perform end-to-end computational reproducibility checks. Beyond assessing whether reproducibility-related criteria are met, `CORE-Bench` also involves executing the provided code and data to verify whether computational results can be reproduced. The authors further use LLM embeddings to predict reproducibility scores, offering a computationally efficient alternative to labor-intensive manual verification. Together, these approaches demonstrate how LLMs can enable scalable monitoring of reproducibility-related practices within the broader open science ecosystem. More broadly, LLMs have been applied to adjacent scholarly workflows relevant to open science monitoring, including peer review automation (Lee et al., 2025), abstract screening (Huotala et al., 2024), and systematic review writing (Mahmoudi et al., 2025). Recent work has also explored using LLMs to extract funding information and associated targeted research questions from text at scale (Lai and Zhao, 2025). While these applications do not directly operationalize large-scale open science monitoring, they provide scalable components that can be integrated into monitoring systems, and their use for this purpose remains an early but promising direction.

2.3.2. Structured

In journals and peer-reviewed conferences such as Nature and its flagship journals, as well as leading AI and data science venues including NeurIPS (Neural Information Processing Systems), ICLR (International Conference on Learning Representations), ICML (International Conference on Machine Learning), and KDD (Knowledge Discovery in Databases), data and code availability are reported in dedicated, structured sections published alongside the main content. Measuring the data and code availability can be done just by reviewing the availability statement alone. One key takeaway is that introducing a structured statement into the submission, review, and camera-ready checks significantly improves code and data availability rates (Hutson, 2018; Pineau et al., 2021).

Moreover, the machine learning community is developing metadata frameworks, such as `Croissant` (Akhtar et al., 2024), to make datasets more readable for both machines and humans. If open science practices become widely accepted in transportation research,

availability features could be measured through structured data collected via self-reporting and peer review as part of the publication process. At that stage, this section could serve as a reference for the community to develop the structured checklist.

2.4. Large language models for feature extraction

As LLMs have grown more powerful, Natural Language Processing (NLP) benchmarks have also become more complex, allowing a better understanding of the possibilities and limitations of LLMs. Many works have assessed the ability of LLMs to recall, extract, and reason from information.

2.4.1. LLM capability testing on scientific papers

Papers often require inputs of tens of thousands of tokens, and relevant pieces of information may be scattered across the work. Because of the computational cost of processing more tokens, LLMs have a limited context window. But as the number of parameters has increased, LLMs have been able to handle in longer contexts (Hsieh et al., 2024), allowing them to be used for more tasks, such as reading scientific papers.

LLMs have been used in many ways to process papers, from performing literature reviews (Agarwal et al., 2025), answering factual questions about a work (Dasigi et al., 2021; Chen et al., 2026; Javaji et al., 2025), and even generating follow-up studies (Liu et al., 2026). Despite these advances, LLMs still face limitations. Long contexts still damage the question answering ability of LLMs (Du et al., 2025), and LLM capability works sometimes use an LLM-as-a-judge framework, where a more powerful model is used as ground truth (Chen et al., 2026). Because our full dataset is too large to be human-validated, we instead compared the LLM's performance on a small, human-validated sample of the whole set.

2.4.2. Prompting strategies

The performance of LLMs can be highly dependent on how the prompt is structured. Various techniques have been used to increase accuracy. Some of the most common include giving examples (Zhong et al., 2023), asking LLMs to first restate relevant information from long contexts (Du et al., 2025), inducing chain of thought reasoning (Zhong et al., 2023), or splitting long tasks into many smaller tasks (Zhong et al., 2023). Our result on the full dataset uses prompts that split the long tasks into smaller tasks, and we also evaluate the performance of other prompting methods on the small, manual validation dataset.

3. Definitions and terminology

Open science represents a comprehensive vision integrating a wide array of concepts (Vicente-Saez and Martinez-Fuentes, 2018; Parsons et al., 2022). However, using such broad terminology, which encompasses concepts ranging from open access publishing to general knowledge sharing, introduces unnecessary ambiguity to this study and the associated feature extraction. Therefore, we narrow the scope of *open science practices* to focus specifically on data and code availability connected with a paper. To align with established open science monitoring principles (Bobrov et al., 2025), we adopt clear, transparent decision rules as detailed below.

For data availability, we anchor our criteria to Elsevier's Research Data Guidelines for transportation research journals (Elsevier, 2025a):

"You are encouraged or required to:

- Deposit your research data in a relevant data repository;
- Cite and link to this dataset in your article;
- If this is not possible, make a statement explaining why research data cannot be shared."

Guided by these guidelines, we identify two distinct data availability features: (i) whether the dataset is deposited in an accessible repository, and (ii) whether the dataset is cited (and linked) in the paper. These features capture two different behaviors: depositing data represents a contribution to open science, whereas citing existing datasets indicates benefiting from it. We define "available data" as a publicly deposited dataset that includes the processed inputs. If neither depositing nor citing criteria are met, the paper is classified as having no data available.

Regarding code, while Elsevier guidelines broadly categorize it under research data, we analyze it as a distinct entity. We define code availability as the presence of a public repository or archival deposit containing source files.

Our scope in this study is restricted to artifact availability detectable within the full text and its links; we do not perform external web searches for resources not explicitly mentioned in the paper. Consequently, data and code availability, as measured here, serves as an early indicator and a reasonable proxy for computational reproducibility, rather than a guarantee of full reproducibility (Zheng, 2021; Seibold et al., 2021).

Fig. 1 presents the logic used in this paper for determining data and code availability. Conceptually, we avoid counting studies that do not rely on code or data, since including them would underestimate the actual rate of code and data availability. The full definitions of all features shown in the flowchart are provided below.

3.1. Data availability features

(i) Data repository. - Does the paper share a data repository? The repository must be created by the authors, but may contain data collected by others (i.e., processed data from NGSIM).

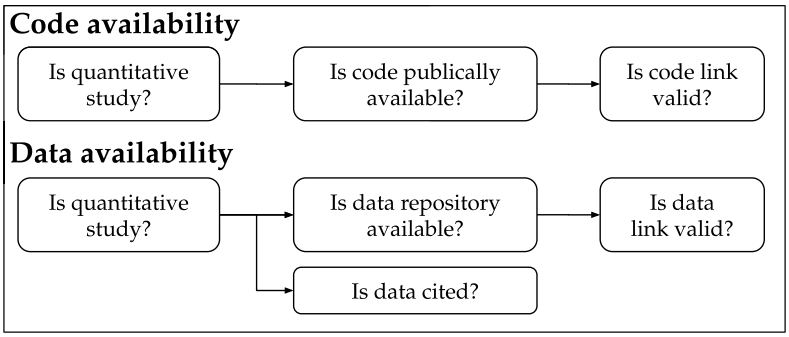


Fig. 1. Features considered for data and code availabilities.

(ii) Data citation or link. - Does the paper use and cite or link at least one publicly available dataset?
(iii) Data repository validity. - Is the repository actually live and does it contain data?

Based on the above definitions, we categorize data availability into four categories:

(i) The first category (NC, NR) represents papers that neither cite data nor provide a data repository.
(ii) The second category (NC, R) includes papers that provide a link to a data repository but do not cite any datasets. These papers may be fully reproducible if they may collect and use original data.
(iii) The third category (C, NR) consists of papers that cite datasets but do not provide a data repository.
(iv) The fourth category (C, R) encompasses papers that both cite datasets and provide a data repository.

This taxonomy allows for a nuanced understanding of data availability practices in transportation research.

3.2. Code availability features

(i) Code dependence. ("Is quantitative study?") - Did the paper do any quantitative calculations?
(ii) Code availability. - Does the paper share a link to code used in the text of the paper?
(iii) Code deposit link validity - Is the link actually live and does it contain code?

Similar to data availability, we categorize code availability into two categories: code available (CA) and code unavailable (CU).

We use these definitions and taxonomies consistently throughout this study for annotation, extraction, analysis and choice modeling. All decision rules and their corresponding illustrative examples are listed in Appendix A.

4. Data

4.1. Full-text paper dataset in transportation research parts

We curated a dataset of research articles published between 2019 and 2024 in Elsevier's Transportation Research journals, specifically Parts A through F (referred to as TR-A, TR-B to TR-F throughout the paper) and Transportation Research Interdisciplinary Perspectives (TR-IP).

In total, 10,990 publications were initially extracted. The dataset consists of full-text articles in `XML` format, automatically retrieved using the Elsevier API[3], which provides structured access to the content. For this study, we included only full-length research articles, excluding editorials, reviews and other non-research content. After filtering, the final dataset contained 10,724 papers. The distribution of articles by journal and year is illustrated in Fig. 2.

The full-text dataset curation pipeline builds on Elsevier's structured XML and stable Text-and-Data-Mining API, which makes full-text retrieval at scale straightforward and consistent across the Transportation Research journals. We focused on the TR Parts because they are flagship venues in the field. The pipeline can be transferred to other domain-specific transportation journals published by Elsevier such as Accident Analysis and Prevention, Travel Behaviour and Society, and the Journal of Transport Geography without major changes. Other transportation publishers (e.g., IEEE, Springer Nature, Wiley) provide text-mining APIs or bulk-download services (Riehl et al., 2025) that allow the pipeline to be adapted.

4.2. Manual annotation process

A manual validation dataset (MVD) is used to evaluate the automated extraction of the data- and code-availability features defined in Section 3. The MVD was constructed from the manual annotation process, using papers sampled from the full-text dataset. The

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

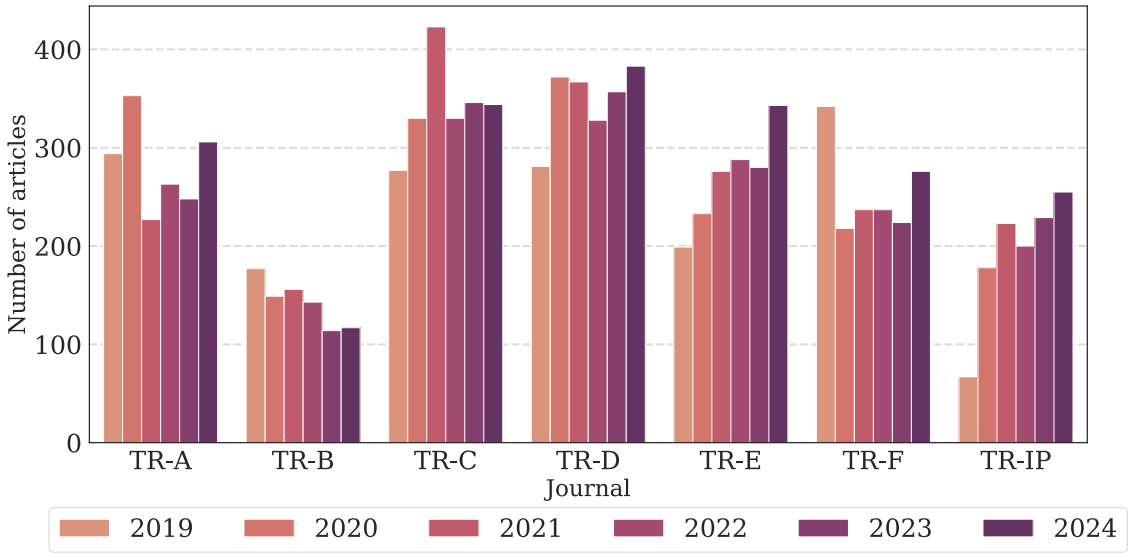


Fig. 2. Number of articles published in transportation research journals (TR-A, TR-B, TR-C, TR-D, TR-E, TR-F, and TR-IP) from 2019 to 2024 ($N = 10,724$).

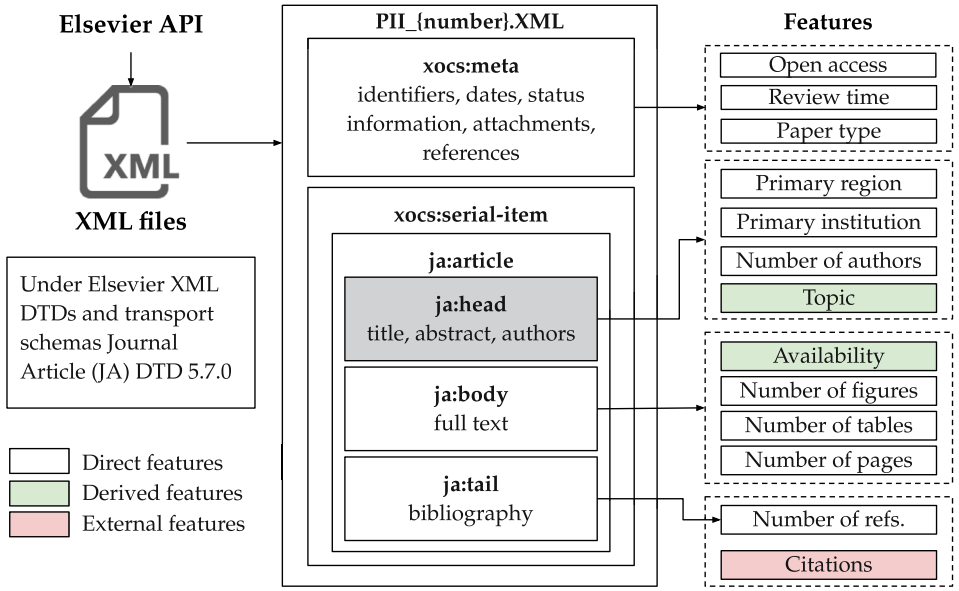


Fig. 3. Overview of the features extracted from our pipeline.

resulting MVD comprises 96 randomly selected full-length research articles from the set of 10,724 papers. Each paper in the MVD was independently reviewed by two annotators, who evaluated the availability of associated data and code. We created a set of definitions and examples for each feature that are included in Appendix A. The authors of this paper and other transportation researchers manually annotated the papers. Annotators received the same instructions and definitions for all annotation tasks.

5. Automatic feature extraction

As illustrated in Fig. 3, our pipeline processes Elsevier `XML` files to generate feature sets. We parsed structural tags (e.g., `xocs:meta`, `ja:head`) to obtain direct features, while employing statistical methods and LLMs to extract derived features. Additionally, we queried the Scopus Search API to obtain citation counts, which serve as our external features.

Table 2
Topics identified by LDA.

Topic Name	Top 5 Words	Count
Optimization & Routing Algorithms	problem, model, propose, time, algorithm	1629
Social & Policy Aspects of Mobility	study, use, mobility, policy, social	1064
Automated Driving & Human Factors	drive, driver, vehicle, pedestrian, automate	1004
Travel & Mode Choice Behavior	travel, choice, mode, car, trip	922
Data-Driven Modeling & Prediction	model, datum, base, method, use	913
Transportation Emissions & Policy	emission, transport, fuel, port, carbon	876
Supply Chain & Market Strategies	supply, chain, market, platform, price	788
Traffic Flow & Network Control	traffic, vehicle, control, flow, model	756
Driver Behavior & Safety Risk	driver, drive, behaviour, risk, safety	651
Urban Environment & Active Transport	urban, city, parking, environment, area	608
Public Transit Service & Demand	service, transit, passenger, bus, network	525
COVID-19 Impact on Travel & Activities	time, covid, travel, pandemic, activity	456
Electric Vehicles & Ride-Sharing	vehicle, electric, charge, ride, fleet	402
Air & Freight Logistics	delivery, truck, freight, airport, air	79
Road Infrastructure & Emergency Management	speed, road, evacuation, limit, zone	51

5.1. Article meta-data

The dataset obtained from the Elsevier API contains the full text of the papers in `XML` format. The `XML` files contain metadata about each paper (see the details of the codebook in Elsevier (2024). The metadata includes information such as the title, authors, abstract, keywords, submission history, funding agencies, and more. The meta-data appears as direct features in Fig. 3. From this information, we computed the review time from initial submission to paper acceptance in days.

5.2. Paper topic

To categorize papers into topics, the abstracts, keywords, titles, and journal names of all full articles were processed with Latent Dirichlet Allocation (LDA), which associates words with topics (Blei et al., 2003). Given a corpus of all documents, LDA identifies topics by learning which words tend to co-occur. It represents each document as a mixture of topics and returns a probability distribution over topics for each document. Then, the highest probability topic is assigned to the document. LDA has been used previously to study the transportation research trends (Sun and Yin, 2017; Sun and Kirtonia, 2020). More details can be found in Appendix F. An interactive t-SNE (van der and Hinton, 2008) visualization of the topic modeling results is available at https://www.rerite.org/MOST/explorer. This tool allows researchers to see the status of availability in different categories and find papers that make code or data repositories available in each category.

After grouping all papers into 15 topics, the top words identified from each topic were given to Google Gemini 2.5 Flash to create a list of preliminary topic names. These preliminary names were then carefully reviewed and refined by the authors. The number of papers in each topic are shown in Table 2.

5.3. Features extracted by LLMs

LLMs were used to extract features that are not explicitly provided in the paper meta-data. Although we initially considered text search methods, we found that text search performed worse, consistently overestimating features. A comparison of results found by LLMs and by text search can be found in Appendix C.

We use Google Gemini model `gemini-2.5-flash-lite` (Gemini Team, 2025) via its API for the feature extraction task. It was chosen for its large maximum context size (over 1 million tokens) and because it is one of the most capable LLMs as of 2025 at the time of the feature extraction (Cai et al., 2025).

To extract the features with LLMs, the contents of the `XML` files were extracted, processed into a markdown file, and uploaded using the Google Gemini API. Multiple prompts were then given to the API to extract features about the code and data in the paper. Recent studies, including (Son et al., 2024; Gozzi and Di Maio, 2024), demonstrate that when state-of-the-art LLMs are tasked with multiple heterogeneous subtasks in a single prompt, their accuracy declines compared to processing each instruction separately. As a result, we only extract a few related features in each prompt. Results were stored in a `JSON` file for each prompt, which are then combined. The prompt design is essential to the stability and scalability of the measurement framework. The prompts were designed to follow the operational definitions in Section 3, to separate heterogeneous decisions into smaller and more consistent subtasks, and to return structured `JSON` outputs that could be automatically postprocessed. This design reduces ambiguity in the extraction task, makes the pipeline easier to audit, and supports repeated application to large corpora. Prompts can be found in Appendix B as well as repository at https://github.com/RRinTransportation/OpenMOST. The accuracy of extracting the code and data features by LLMs is compared with human annotators in the next section.

Table 3
Percentage agreement and prevalence of true values by rater. Prev. = Prevalence; H1, H2 = human expert labelers 1, 2.

Features	All	H1 vs H2	H1 vs LLM	H2 vs LLM	Prev.(H1)	Prev. (H2)	Prev.(LLM)
`is_code_publicly_available`	0.9688	0.9688	0.9896	0.9792	0.0833	0.0521	0.0729
`is_data_cited`	0.6979	0.8542	0.7500	0.7917	0.3438	0.3021	0.2604
`is_data_repository_available`	0.9167	0.9271	0.9271	0.9792	0.0833	0.0312	0.0312
`is_quantitative_study`	0.8854	0.9062	0.9167	0.9479	0.8854	0.9167	0.9688

5.4. Human-LLM agreement evaluation

5.4.1. Evaluation criteria

To analyze the validity of the LLM pipeline, a three-rater inter-rater agreement (IRA) analysis (McHugh, 2012) was conducted on the 96 papers of the MVD. The raters included two independent human annotators (H1, H2) and the automated LLM pipeline. Selecting appropriate agreement metrics requires careful consideration of our dataset characteristics. Two complementary metrics are provided.

Our primary metric for overall agreement is Fleiss's Kappa (Fleiss, 1971). This statistic provides a robust measure of agreement among three or more raters that corrects for the probability that agreement occurred by chance. While Fleiss's Kappa scores tend to be more conservative in datasets with high class imbalance, they offer a more meaningful assessment of true consensus. We interpret these scores using the Landis & Koch (Landis and Koch, 1977) scale: 0.21-0.40 (Fair), 0.41-0.60 (Moderate), 0.61-0.80 (Substantial), and 0.81-1.00 (Almost Perfect).

We also used Percentage Agreement (PA), which represents the raw proportion of matching annotations across raters. While intuitive and easy to interpret, PA has an important limitation in the presence of class imbalance: when negative cases heavily outnumber positive cases (as shown by the "Prevalence" in Table 3), raters can achieve high PA simply by agreeing on the predominant negative class, without demonstrating true consensus on the more challenging positive cases. To address this, we also calculate Cohen's kappa, which measures agreement between two raters (Cohen, 1960). Cohen's kappa ranges from -1 to 1. Landis & Koch (Landis and Koch, 1977) suggested the same interpretation for Cohen's kappa.

5.4.2. Agreement analysis results

Agreement analysis across six features revealed Fleiss' Kappa values ranging from 0.399 to 0.839, with the LLM achieving agreement levels consistently comparable to inter-human agreement. Agreement varied systematically by feature objectivity (as shown in Table 4) .

The most objective feature, `is_code_publicly_available`, achieved *Almost Perfect* agreement (Fleiss' $\kappa = 0.839$, PA = 96.9%), with the LLM actually exceeding human-human agreement in pairwise comparisons (Cohen's $\kappa = 0.928$ for H1-LLM vs. 0.753 for H1-H2). This high agreement reflects the relatively unambiguous nature of code availability assessment. Data-related features showed more moderate results. The feature `is_data_cited` achieved *Moderate* agreement (Fleiss' $\kappa = 0.522$, PA = 69.8%), while `is_data_repository_available` ($\kappa = 0.399$, PA = 91.7%) reached *Fair* agreement. Notably, the *Fair* agreement score is comparable to the human-human baseline for these features (Cohen's $\kappa = 0.333$ for data repository). This reflects genuine ambiguity in the definitions and human assessments rather than LLM-specific limitations.

For `is_quantitative_study`, the LLM achieved *Moderate* agreement (Fleiss' $\kappa = 0.459$), again performing comparably to human-human agreement. The high prevalence values for certain features (e.g., 96.88% for `is_quantitative_study`) underscore why Kappa, rather than PA alone, provides a more meaningful assessment of consensus.

5.4.3. Implications

The agreement analysis demonstrates that our automated LLM pipeline is a valid tool for annotating artifact availability in transportation studies. The *Fair* to *Almost Perfect* Fleiss' Kappa scores indicate strong agreement between the LLM and two independent human experts, particularly when considering the high class imbalance in the dataset. Critically, pairwise comparisons (Table 4) show that the LLM's agreement with each human expert consistently matched the baseline human-human agreement, even on subjective features where the human annotators themselves showed imperfect concordance. This analysis provides confidence in using the LLM pipeline for large-scale artifact availability annotation.

5.4.4. Robustness analysis

Because LLMs are stochastic and our temperature was not set to zero, it is possible that repeating this pipeline run would lead to different results. To verify that this performance was not an outlier, the same pipeline was run ten times on the MVD. The results are shown in Table 5. All the Fleiss's Kappa mean values were within the same performance interval. The STDs are all under 5% for percentage agreement and 0.05 for Fleiss's Kappa. Given the sensitivity of Fleiss's Kappa to small deviations for heavily skewed variables, this indicates that the model is performing nearly identically across runs. This is likely due to the definitions for each variable used in the prompt that reduces ambiguity even when temperature is not set to zero.

To explore the performance of our pipeline with different models and prompting structures (Zhong et al., 2023), we also tested them using the same pipeline and found that none of them performed uniformly better or worse than our current settings. The results can be found in Appendix G.

Table 4
Kappa values for agreement on 96 papers. Cohen's κ is used to measure agreement between 2 raters, while Fleiss's κ can measure between multiple raters. Fleiss's κ indicates how much better agreement is between annotators than would be expected by chance. Interpretation guide with Fleiss's κ: 0.21-0.40 (Fair), 0.41-0.60 (Moderate), 0.61-0.80 (Substantial), and 0.81-1.00 (Almost Perfect) (Landis and Koch, 1977). The background highlighting is there to make the results easier to read and interpret.

Features	Cohen's κ (H1 vs H2)	Cohen's κ (H1 vs LLM)	Cohen's κ (H2 vs LLM)	Fleiss's κ (All)
`is_code_publicly_available`	0.7534	0.9277	0.8226	0.8388
`is_data_cited`	0.6672	0.4119	0.4858	0.5224
`is_data_repository_available`	0.3333	0.3333	0.6559	0.3994
`is_quantitative_study`	0.4757	0.3991	0.5238	0.4586

Table 5
Inter-rater agreement metrics across dataset variables (mean ± STD).

Variable	Percentage Agreement	Fleiss's Kappa
`is_code_publicly_available`	0.9677 ± 0.0033	0.8316 ± 0.0202
`is_data_cited`	0.6865 ± 0.0166	0.4916 ± 0.0308
`is_data_repository_available`	0.9156 ± 0.0033	0.3606 ± 0.0365
`is_quantitative_study`	0.8812 ± 0.0054	0.4635 ± 0.0196

5.5. Data postprocessing

To support reproducible downstream analyses, we postprocess all extracted outputs into an analysis-ready dataset. The postprocessing pipeline integrates three sources: (i) structured metadata parsed from the Elsevier `XML` files (Section 5.1); (ii) LLM-extracted features returned as `JSON` files (Section 5.3); and (iii) automated URL parsing and link-validation outputs derived from outbound links found in the full text and/or returned by the LLM prompts. Human annotations from the MVD (Section 4) are used exclusively to evaluate extraction accuracy and agreement (Section 5.4). Step-by-step postprocessing details can be found in Appendix D.

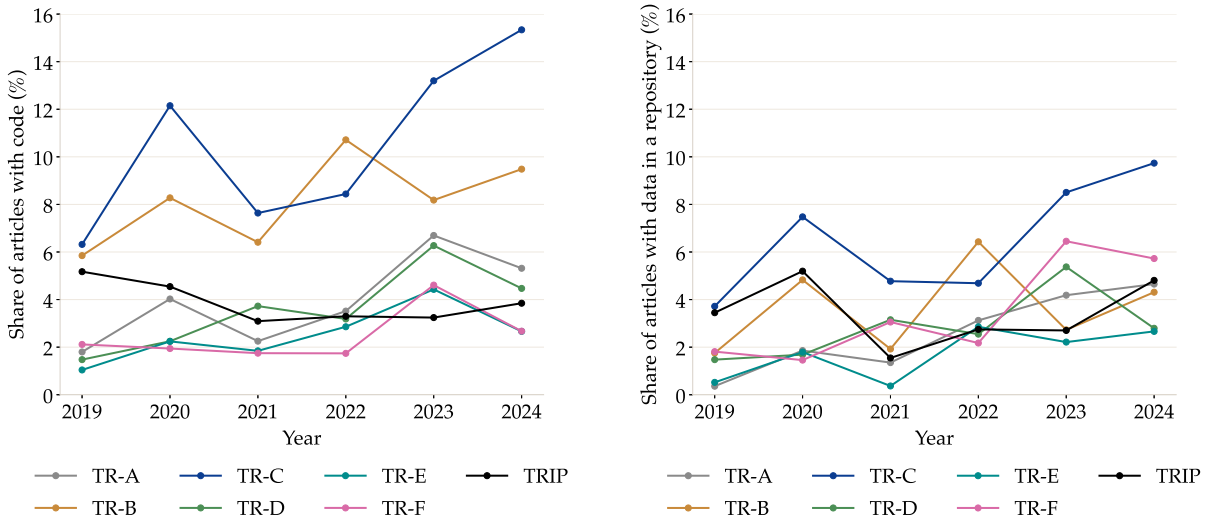
Feature inclusion for downstream analysis follows the agreement evaluation in Section 5.4. Features with low agreement that reflect ambiguity are excluded from modeling and replaced by higher-agreement proxies as described there. Features that remain central to the study but exhibit only Fair agreement under class imbalance (e.g., `is_data_repository_available`) are retained with caution and interpreted in light of this measurement uncertainty, as discussed in Section 7. Additional transformations required for estimation (e.g., one-hot encoding of categorical variables, collapsing sparse categories, and centering numeric covariates) are performed at the analysis stage and documented alongside the corresponding models in Section 6. The complete list of post-processed variables, their types, and operational definitions is provided in the data dictionary (Appendix E) and implementation details are documented in the accompanying repository.

6. Analysis

6.1. Descriptive statistics and bivariate tests

The analysis dataset contains the 10,480 papers that report quantitative studies (`is_quantitative_study`) among the full-length articles. Table 6 provides descriptive statistics for these papers, organized by code availability. The dataset contains 528 papers with code available, a little more than 5% of the dataset. In total, about 29% of the papers cite available data, but only 4% of the papers share a data repository; the remaining 69% neither cite data nor include a repository. Fig. 4 presents the temporal trends in data and code sharing within our research community over the years covered by this study.

We conducted a series of bivariate statistical tests on the dependent variables, specifically, whether the papers provide code or data, using t-tests on differences of means when the independent variable is numeric, and Pearson χ^2 tests of independence when the independent variable is categorical. The results of these tests revealed several statistically significant relationships that guided the analysis that follows in this section. To begin, papers that make code available are also more likely to share data. Newer papers make code available more frequently, and newer papers also tend to share data directly rather than simply cite existing data. Papers with corresponding authors in Europe and North America are more likely to make data available than papers with authors in Asia; the same trend holds for sharing code, with South American authors joining the group more likely to share than Asian authors. In absolute terms, authors based in Asia still contribute a substantial number ($N = 118$) of code repositories, but their share appears lower when expressed as a percentage of the total number of publications from the region ($N = 4516$). The usage of data varies widely by topic, with the topics "Urban Environment & Active Transport" and "COVID-19 Impact on Travel & Activities" having higher than expected rates of citing data from others (about 45% of papers cited or linked to external datasets in these topic categories). Papers in the topic "Data-driven modeling and prediction" share code at a much higher rate than other topics (about 15% of papers have code available). Other comparisons did not result in bivariate test statistics sufficient to reject a null hypothesis of no relationship or no difference in means. Note that the numbers reported in the analysis section are subject to estimation error arising from the methodology.



(a) Code repository by journal year

(b) Data repository by journal year

Fig. 4. Comparison of data repository patterns across journals and years.

These findings are informative, but they may be the result of correlated omitted variables. A complete list of statistics of all variables can be viewed here. In the next section we provide a choice model of data and code availability that can accommodate the simultaneous influence of multiple variables.

6.2. Choice model predicting the availability of code

The factors predicting the availability of code were analyzed using a logit model. This model allows one to investigate the effect of several explanatory variables while controlling for possible confounding factors. The utility functions for code available (CA) and code unavailable (CU) for each paper n are presented in Eqs. (1) and (2):

$$U_n^{CA} = \alpha^{CA} + \boldsymbol{\beta}^{CA}\mathbf{X}_n^{CA} + \varepsilon_n^{CA}, \tag{1}$$

$$U_n^{CU} = 0 + \varepsilon_n^{CU}, \tag{2}$$

where α^{CA} is the constant, $\boldsymbol{\beta}^{CA}$ is the vector of parameters associated with the explanatory variables $\mathbf{X}_n^{CA}$, and ε_n^{CA} and $\varepsilon_n^{CU}(t)$ are i.i.d. extreme value distributed error terms. Given the choice indicator Y_n, the probability of code being available is given by Eq. (3):

$$P_n(Y_n = 1) = \frac{\exp\left(\alpha^{CA} + \boldsymbol{\beta}^{CA}\mathbf{X}_n^{CA}\right)}{1 + \exp\left(\alpha^{CA} + \boldsymbol{\beta}^{CA}\mathbf{X}_n^{CA}\right)}. \tag{3}$$

The parameters were estimated using the Biogeme Python package (Bierlaire, 2023). The explanatory variables presented in the final model specification were chosen according to interpretation and statistical significance (p-value < 0.05). The numerical explanatory variables were centered on the means. Categorical variables were coded as multiple binary variables. The most frequent category was assumed to be the reference category and fixed to zero. When a variable was available for part of the papers, a binary variable was included into the equation to denote the missing values. The effect of each explanatory variable was tested using the likelihood ratio test between different model specifications. All variables related to the journal, the region of the corresponding author, the paper topic, and the paper characteristics were initially included in the utility function. The variables that showed low significance according to descriptive statistics were tested first. Binary variables extracted from the same categorical variable that had a similar impact on code availability were combined into a new variable. Variables that did not have a significant effect were excluded one by one.

Table 7 shows the goodness of fit measures for the final model and for a baseline model that includes only a constant. The adjusted likelihood ratio index shows that the inclusion of explanatory variables resulted in an improvement of 11.2% compared to the model that includes only a constant.

Table 8 shows the estimation results. All parameters are statistically significant at the 5% level. The alternative specific constant is negative, meaning that the papers are less likely to make code available, everything else being equal.

Certain journals had a significant effect on code availability. Papers published in TR-D were initially chosen as the reference category. Papers published in TR-B and TR-C were more likely to have code available. The effect was not significantly different between the two journals. Papers in the other journals (TR-A, TR-E, TR-F, and TR-IP) did not have a significant effect compared to TR-D.

Table 6
Descriptive statistics of variables by code availability.

		Code available (N = 528)		Not available (N = 9952)	
Numeric variables		Mean	Std. Dev.	Mean	Std. Dev.
Age of paper [years]		3.1	1.7	3.4	1.7
N. tables in paper		6.3	4.6	6.4	4.5
N. figures in paper		11.0	6.5	8.6	6.1
N. references		57.3	21.0	58.2	26.9
N. authors		3.8	1.7	3.8	1.7
Times cited		27.2	42.9	28.4	43.2
Review time [days]		268.6	141.4	254.3	148.9
Page count		19.8	6.6	18.4	6.3
Categorical variables		N	%	N	%
Data availability	Data is cited or linked	96	18.2	2810	28.2
	Data repository and citation	94	17.8	14	0.1
	Data repository is available	199	37.7	65	0.7
	No repository or citation	139	26.3	7063	71.0
Open Access	Not open access	324	61.4	7346	73.8
	Open access	204	38.6	2606	26.2
Journal	TR-IP	36	6.8	945	9.5
	TR-A	64	12.1	1555	15.6
	TR-B	67	12.7	771	7.7
	TR-C	212	40.2	1797	18.1
	TR-D	72	13.6	1911	19.2
	TR-E	41	7.8	1534	15.4
	TR-F	36	6.8	1439	14.5
Region of corr. author	Africa	5	0.9	69	0.7
	Asia	118	22.3	4398	44.2
	Europe	206	39.0	2699	27.1
	North America	156	29.5	2074	20.8
	Oceania	27	5.1	521	5.2
	South America	16	3.0	191	1.9
Availability statement	Availability statement is not present	139	26.3	7062	71.0
	Availability statement is present	389	73.7	2890	29.0
Paper topic	Air & Freight Logistics	2	0.4	77	0.8
	Automated Driving & Human Factors	34	6.4	957	9.6
	COVID-19 Impact on Travel & Activities	17	3.2	434	4.4
	Data-Driven Modeling & Prediction	142	26.9	772	7.8
	Driver Behavior & Safety Risk	10	1.9	631	6.3
	Electric Vehicles & Ride-Sharing	29	5.5	372	3.7
	Optimization & Routing Algorithms	114	21.6	1513	15.2
	Public Transit Service & Demand	29	5.5	489	4.9
	Road Infrastructure & Emergency Management	1	0.2	49	0.5
	Social & Policy Aspects of Mobility	26	4.9	862	8.7
	Supply Chain & Market Strategies	6	1.1	775	7.8
	Traffic flow & Network Control	45	8.5	710	7.1
	Transportation Emissions & Policy	20	3.8	848	8.5
	Travel & Mode Choice Behavior	28	5.3	883	8.9
	Urban Environment & Active Transport	25	4.7	580	5.8
Links to code repositories	Link to GitHub	329	62.3	4	0.0
	Link to other service	118	22.3	5	0.1
	No links	81	15.3	9943	99.9

Table 7
Statistics for the choice model predicting the availability of code.

Statistic	Value
Number of parameters associated with explanatory variables (K)	8
Number of observations	10,480
Constant log likelihood ($L(c)$)	-2,092
Final log likelihood ($L(\hat{\beta})$)	-1,850
Adjusted likelihood ratio index $\rho^2 = 1 - \frac{L(\hat{\beta})-K}{L(c)}$	0.112

Table 8
Choice model predicting the availability of code.

Variable	Description	Estimate (Robust SE)
α	Alternative-specific constant	-3.89 (0.124) ***
TRB, TRC	Binary variable equal to one when the journal is TR-B or TR-C	0.785 (0.104) ***
Europe, SouthAmerica, Africa	Binary variable equal to one when the region is Europe, South America, or Africa	1.37 (0.122) ***
NorthAmerica, Oceania	Binary variable equal to one when the region is North America or Oceania	1.11 (0.125) ***
DataDriven	Binary variable equal to one when the topic is data-driven methods	0.738 (0.123) ***
AutoDri, Cov19, SocPol, TraFlow, TraEm, TraMo	Binary variable equal to one when the topic is automated driving, COVID-19, transport policy, traffic flow, transport emissions, or travel mode	-0.613 (0.113) ***
DrivBeh, Log, RoInf, SuCha	Binary variable equal to one when the topic is driver behavior, logistics, road infrastructure, or supply chain	-1.32 (0.248) ***
NumFig	Number of figures in the paper	0.0194 (0.00738) **
PaperAge	Paper age in years	-0.178 (0.0284) ***

Notes: Robust standard errors in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$.

The regions in which the first author was located had a significant effect on code availability. Papers with authors in Europe, South America, Africa, North America, and Oceania were more likely to have code available than papers with authors in Asia. This effect was significantly higher for Europe, South America and Africa than for North America and Oceania. The effect was not significantly different within the group Europe, South America, and Africa and within the group North America and Oceania.

The topics of the papers had a significant impact on code availability. Papers on optimization were initially chosen as the reference category. Papers based on data-driven methods were the most likely to have available code. Papers focusing on automated driving, COVID-19, transport policy, traffic flow, transport emissions, and travel mode were less likely to have code available. The effect was not significantly different between these topics. Topics such as driver behavior, logistics, road infrastructure, and supply chain were the least likely to have code available. The effect was not significantly different between these topics. The other topics did not have a significant effect compared to optimization.

Certain characteristics of the paper had a significant effect on the availability of code. Papers with a high number of figures were more likely to have available code. Older papers were less likely to have code available. The number of authors, number of tables, reference count, and page count were tested but did not exhibit a significant effect and were therefore excluded from the final model specification.

6.3. Choice model predicting the availability of data

The factors predicting the availability of data were analyzed using a logit model. The utility functions of data citation and data repository (C, R), no data citation and data repository (NC, R), data citation and no data repository (C, NR), no data citation and no data repository (NC, NR) for each paper n are presented in Eqs. (4)–(7):

$$U_n^{C,R} = \alpha^{C,R} + \boldsymbol{\beta}^{C,R}\mathbf{X}_n^{C,R} + \varepsilon_n^{C,R}, \tag{4}$$

$$U_n^{NC,R} = \alpha^{NC,R} + \boldsymbol{\beta}^{NC,R}\mathbf{X}_n^{NC,R} + \varepsilon_n^{NC,R}, \tag{5}$$

$$U_n^{C,NR} = \alpha^{C,NR} + \boldsymbol{\beta}^{C,NR}\mathbf{X}_n^{C,NR} + \varepsilon_n^{C,NR}, \tag{6}$$

$$U_n^{NC,NR} = 0 + \varepsilon_n^{NC,NR}, \tag{7}$$

where $\alpha^{C,R}$, $\alpha^{NC,R}$, $\alpha^{C,NR}$ are the alternative specific constants, $\boldsymbol{\beta}^{C,R}$, $\boldsymbol{\beta}^{NC,R}$, $\boldsymbol{\beta}^{C,NR}$ are the vectors of parameters associated with the explanatory variables $\mathbf{X}_n^{C,R}$, $\mathbf{X}_n^{NC,R}$, $\mathbf{X}_n^{C,NR}$ and $\varepsilon_n^{C,R}$, $\varepsilon_n^{NC,R}$, $\varepsilon_n^{C,NR}$ are i.i.d. extreme value distributed error terms. Given the choice indicator Y_n, the probability that data is available $k \in C_D$ with C_D = {C,R; NC,R; C,NR; NC,NR} is given by Eq. (8):

$$P_n(Y_n = k) = \frac{\exp(\alpha^k + \boldsymbol{\beta}^k \mathbf{X}_n^k)}{\sum_{j \in C_D} \exp(\alpha^j + \boldsymbol{\beta}^j \mathbf{X}_n^j)}. \tag{8}$$

The parameters were estimated using the Biogeme Python package (Bierlaire, 2023). The explanatory variables presented in the final model specification were chosen according to interpretation and statistical significance (p-value < 0.05). The explanatory variables were processed as described in the previous section. All variables related to the journal, the region of the corresponding author, the paper topic, and the paper characteristics were initially included in the utility functions and assumed to have the same impact between the alternatives. Alternative specific parameters for each variable were tested using the likelihood ratio test. The variables that showed high significance based on descriptive statistics were first tested. The variables were combined and excluded as described in the previous section. Cross-nested and nested logit specifications were also tested to capture unobserved correlations between alternatives that include data citation and/or data repository. The estimation results did not produce a significant improvement in the

Table 9
Statistics for the choice model predicting the availability of data.

Statistic	Value
Number of parameters associated with explanatory variables (K)	26
Number of observations	10,480
Constant log likelihood ($L(c)$)	-7,895
Final log likelihood ($L(\hat{\beta})$)	-7,146
Adjusted likelihood ratio index $\rho^2 = 1 - \frac{L(\hat{\beta})-K}{L(c)}$	0.0916

final log likelihood. The nesting and membership parameters were not statistically significant. Therefore, the logit model was retained as the final specification.

Table 9 shows the goodness of fit measures for the final model and for a baseline model that includes only a constant. The adjusted likelihood ratio index shows that the inclusion of explanatory variables resulted in an improvement of 9.2% compared to the model that includes only a constant.

Table 10 shows the estimation results. Almost all parameters are statistically significant at the 5% level. One parameter that was significant at the 10% level was retained because it resulted in a significant improvement based on the likelihood ratio test. The alternative-specific constants are negative, meaning that the papers are less likely to provide a data citation or a data repository, everything else being equal. This effect is larger for providing a data repository than for providing a data citation.

Certain journals had a significant effect on data availability. Papers published in TR-D were initially chosen as the reference category. Papers in TR-B and TR-E were less likely to have a data citation and no data repository. The effect was not significantly different between the two journals. Papers in TR-C were more likely to have a data repository, regardless of data citation. Papers published in TR-F were less likely to have a data citation, regardless of the availability of a data repository. Papers published in TR-IP were less likely to have a data citation and a data repository. The other journals did not have a significant effect compared to TR-D.

The region in which the first author was located had a significant effect on data availability. Asia was initially chosen as the reference category. Papers with authors in Africa, North America, Europe, and Oceania were more likely to have a data citation and no data repository. This effect was significantly higher for Africa and North America than for Europe and Oceania. Articles with authors in Europe, North America, Oceania, and South America were more likely to provide a data repository regardless of data citation. This effect was significantly higher for Europe and North America than for Oceania and South America. The other regions did not have a significant effect compared to Asia.

The topic of the paper had a significant impact on data availability. Papers on optimization were initially chosen as the reference category. Papers based on data-driven methods were the most likely to provide a data citation and/or a data repository. The effect was higher when both a data citation and a data repository were provided. Papers focusing on COVID-19 or urban environment were more likely to have a data citation, and the effect was higher when a repository was also provided. The effect was not significant for a data repository without a data citation. The effect was not significantly different for these two topics. Papers on transport emissions were more likely to have a data citation without a data repository and less likely to provide a repository. Papers on driver behavior, transport policy, and traffic flow were less likely to provide a data citation and/or a data repository. The effect was higher for transport policy and traffic flow than for driver behavior. The effect did not differ significantly across a data citation and/or a data repository. Papers on automated driving were less likely to have a data citation and no data repository. Topics such as travel mode, logistics, road infrastructure, and supply chain were the least likely to provide a data citation and/or a data repository. The effect was higher for papers providing a data repository. The effect was not significantly different between these topics. The other topics did not have a significant effect compared to optimization.

Certain characteristics of the paper had a significant effect on the availability of data. Papers with a high number of figures were more likely to provide a data citation and/or a data repository. The effect did not differ significantly between a data citation and a data repository. Papers with a high number of tables were more likely to have either a data citation or a data repository, but not both. The effect did not differ significantly between a data citation and a data repository. Papers with a high number of pages were less likely to have data available. The effect did not differ significantly between a data citation and a data repository.

Papers with a high number of authors were less likely to have a data citation and no data repository. The effect of the number of authors on the other alternatives was not significant. Older papers were more likely to have a data citation and no data repository and less likely to provide a repository. The reference count did not have a significant effect.

6.4. Incentives for availability

Open science techniques often require additional investment and effort by researchers. If the discipline does not require these efforts, researchers may respond to various incentives. In our data, two incentives that we capture include a reduced review time prior to publication or a higher rate of citations on the resulting publication. The review time for a paper can be determined in the paper metadata by calculating the number of days from when the paper was first submitted to when the paper was accepted for publication, while the rate of citations is calculated as the number of citations in SCOPUS divided by the whole number of years since publication. Unfortunately, neither incentive is observable in the papers we studied.

In terms of the number of citations, Fig. 5 shows the share of citations per year for each paper ordered by paper acceptance date in SCOPUS. Fig. 5a groups the papers by their data availability in the categories described above, and Fig. 5b groups the papers by code

Table 10
Choice model predicting the data citation and the data repository.

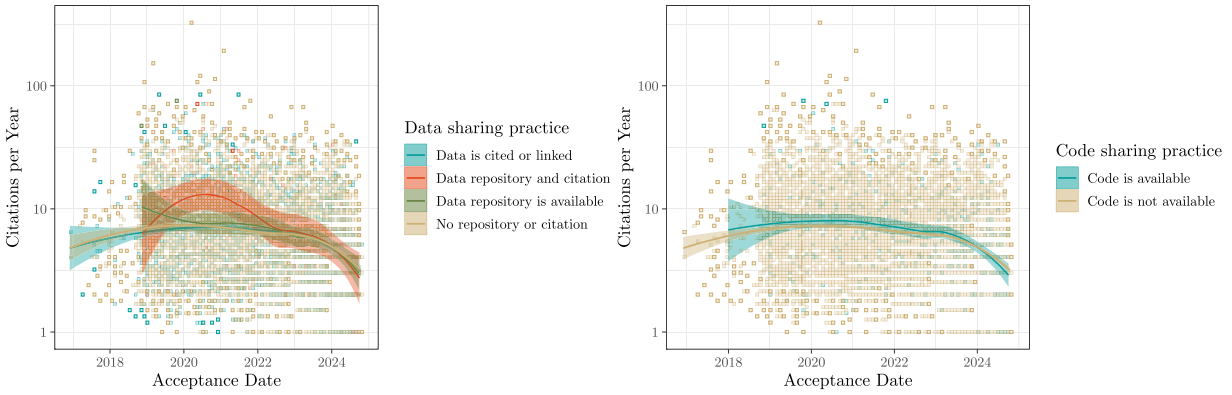
Variable	Description	Estimate (Robust SE)		
		$V_{C,NR}$	$V_{NC,R}$	$V_{C,R}$
α	Alternative-specific constants	-0.892 (0.0580) ***	-4.46 (0.149) ***	-5.60 (0.193) ***
TRB, TRE	Binary variable equal to one when the journal is TR-B or TR-E	-0.263 (0.0640) ***	–	–
TRC	Binary variable equal to one when the journal is TR-C	–	0.556 (0.128) ***	0.556 (0.128) ***
TRF	Binary variable equal to one when the journal is TR-F	-1.23 (0.117) ***	–	-1.23 (0.117) ***
TRIP	Binary variable equal to one when the journal is TR-IP	–	–	-0.859 (0.431) *
Africa, NorthAmerica	Binary variable equal to one when the region is Africa or North America	0.769 (0.0592) ***	–	–
Europe, Oceania	Binary variable equal to one when the region is Europe or Oceania	0.351 (0.0560) ***	–	–
Europe, NorthAmerica	Binary variable equal to one when the region is Europe or North America	–	1.68 (0.142) ***	1.68 (0.142) ***
Oceania, SouthAmerica	Binary variable equal to one when the region is Oceania or South America	–	1.17 (0.237) ***	1.17 (0.237) ***
DataDriven	Binary variable equal to one when the topic is data-driven methods	0.355 (0.0867) ***	0.879 (0.174) ***	1.83 (0.227) ***
Cov19, UrbEnv	Binary variable equal to one when the topic is COVID-19 or urban environment	0.443 (0.0812) ***	–	1.16 (0.284) ***
TransEmiss	Binary variable equal to one when the topic is transport emissions	0.335 (0.0862) ***	-0.525 (0.275)	-0.525 (0.275)
DrivBeh	Binary variable equal to one when the topic is driver behavior	-0.297 (0.125) *	-0.297 (0.125) *	-0.297 (0.125) *
SocPol, TraFlow	Binary variable for transport policy and traffic flow topics	-0.628 (0.0751) ***	-0.628 (0.0751) ***	-0.628 (0.0751) ***
AutoDri	Binary variable equal to one when the topic is automated driving	-1.05 (0.140) ***	–	–
TraMo, Log, RoInf, SuCha	Binary variable for travel mode, logistics, road infrastructure, or supply chain topics	-0.504 (0.0742) ***	-1.16 (0.238) ***	-1.16 (0.238) ***
NumFig	Number of figures in the paper	0.0238 (0.00507) ***	0.0238 (0.00507) ***	0.0238 (0.00507) ***
NumTab	Number of tables in the paper	0.0197 (0.00589) ***	0.0197 (0.00589) ***	–
NumPag	Number of pages	-0.0383 (0.00584) ***	-0.0383 (0.00584) ***	-0.0383 (0.00584) ***
NumAuth	Number of authors	-0.0340 (0.0142) *	–	–
PaperAge	Paper age in years	0.0740 (0.0139) ***	-0.224 (0.0337) ***	-0.224 (0.0337) ***

Notes: Robust standard errors in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$. '–' indicates that the variable is not included in the corresponding utility function. For variables with identical estimates across utilities, equality restrictions were imposed based on statistical tests.

availability. Both figures show a LOESS moving average rate of the mean citation rate per year, along with a 95% confidence interval of the mean. In neither case are the average citation rates different based on the availability of data or code. Fig. 5a shows a temporary increase in the citation rate for papers with a data repository and citation in 2020 and 2021, to the point where the uncertainty interval does not overlap the uncertainty interval of the other papers with different data sharing practices. Further examination of this effect showed that two papers with unusually high citations drive this effect. There was no such temporal variability in the citations or review time grouped by data sharing practice.

Table 11 presents estimates of log-linear regression models of the review time in days and citations per year for the papers in the dataset. These models include the journal, paper topic, and region of the corresponding author as dummy controls to account for differing practices between journals and reviewers in topics. Insignificant parameter estimates on these controls are suppressed in Table 11 for clarity but the controls by themselves explain virtually all of the reduction in variance in the models, with insignificant reductions in the model F-statistic when markers for data sharing and code availability are included (p-value of 0.44 for review time and 0.33 for citations per year).

There are many considerations in this analysis, and in contemplating the role of open science practices vis a vis researcher incentives. First, researchers might not necessarily expect to see reduced review times with open science practices. If a reviewer can follow and reproduce the analysis in the paper during the review, this may result in more thorough reviews and more extended resubmission processes. Improved and more comprehensive reviews would benefit science, but would not necessarily speed up the review process to the authors' benefit. Similarly, sharing data and code may lead to more citations from others who pursue subsequent research, but this effect is only likely if the original research question lends itself to continued study, and if the code and data are packaged in such a way that continued study is feasible. Further, the models presented in Table 11 only explain one-tenth to one-fifth of the variance in the data: important variables such as individual reviewer availability, the complexity of the specific research question and methods, and other sources of variance are not available to this analysis but likely play an important role.



(a) Citations per year by data sharing practice.

(b) Citations per year by code sharing practice.

Fig. 5. Citations per year of each paper over time by code and data-sharing practice, with a LOESS moving average line applied for mean comparison.

Table 11
Incentives for open science.

	Review time [days]		Citations per year	
	Controls	Open Science	Controls	Open Science
Intercept	5.189***	5.198***	1.262***	1.273***
	[5.010, 5.369]	[5.019, 5.378]	[1.022, 1.503]	[1.032, 1.514]
Code is available		0.013		-0.067
		[-0.051, 0.076]		[-0.154, 0.020]
Data repository is available (NC, R)		-0.018		0.020
		[-0.099, 0.064]		[-0.092, 0.132]
Data is cited or linked (C, NR)		-0.019		-0.012
		[-0.045, 0.007]		[-0.046, 0.022]
Data repository and citation (C, R)		0.052		0.138 +
		[-0.068, 0.171]		[-0.023, 0.299]
Journal: TR - A	0.755***	0.754***	0.415***	0.415***
	[0.710, 0.800]	[0.709, 0.800]	[0.353, 0.476]	[0.353, 0.476]
Journal: TR - B	0.619***	0.616***	0.523***	0.524***
	[0.563, 0.675]	[0.560, 0.672]	[0.446, 0.601]	[0.446, 0.602]
Journal: TR - C	0.415***	0.413***	0.722***	0.722***
	[0.368, 0.462]	[0.366, 0.460]	[0.656, 0.787]	[0.657, 0.788]
Journal: TR - D	0.033	0.032	0.600***	0.599***
	[-0.017, 0.083]	[-0.018, 0.081]	[0.538, 0.661]	[0.538, 0.661]
Journal: TR - E	0.305***	0.303***	0.682***	0.680***
	[0.253, 0.356]	[0.252, 0.355]	[0.610, 0.753]	[0.609, 0.752]
Journal: TR - F	0.223***	0.220***	0.231***	0.228***
	[0.171, 0.274]	[0.168, 0.271]	[0.159, 0.303]	[0.156, 0.300]
Topic: COVID-19 Impact on Travel & Activities	-0.213**	-0.213**	0.020	0.019
	[-0.347, -0.080]	[-0.347, -0.080]	[-0.160, 0.200]	[-0.162, 0.199]
Topic: Social & Policy Aspects of Mobility	-0.097	-0.100	0.265**	0.262**
	[-0.225, 0.030]	[-0.228, 0.028]	[0.091, 0.438]	[0.088, 0.435]
Topic: Electric Vehicles & Ride-Sharing	-0.035	-0.035	0.206*	0.208*
	[-0.170, 0.100]	[-0.170, 0.100]	[0.025, 0.388]	[0.026, 0.389]
Topic: Road Infrastructure & Emergency Management	0.197 +	0.195 +	-0.270*	-0.271*
	[-0.019, 0.412]	[-0.021, 0.411]	[-0.537, -0.003]	[-0.538, -0.005]
Region: South America	0.047	0.044	-0.209*	-0.212*
	[-0.105, 0.199]	[-0.108, 0.197]	[-0.409, -0.009]	[-0.412, -0.013]
Num.Obs.	9185	9185	10,480	10,480
R2	0.182	0.183	0.113	0.114
R2 Adj.	0.180	0.180	0.111	0.111
Log.Lik.	-7138.461	-7136.564	-11825.901	-11823.609
RMSE	0.53	0.53	0.75	0.75

95% confidence intervals are reported in brackets.
Additional controls: Journal, paper topic, and region of the corresponding author. Full controls included, only significant controls displayed.

7. Discussion

7.1. Main findings

The results of this study provide a field-level baseline for understanding open-science practices in transportation research. Rather than treating code and data availability as aggregate indicators alone, the analysis shows how these practices vary across publication venues, research areas, regions, and time. The findings allow identifying where open-science adoption is emerging, where barriers remain, and which parts of the research ecosystem may require targeted support.

Mean values. The code was available in 5% of the papers. Data repositories were available in 4% of the studies and open data were cited in 29% of the studies. These results show that data sharing and code sharing are still rare in the field of transportation. These results are slightly higher than previous findings showing that fewer than 2% of transportation simulation studies shared repositories with data and/or code (Riehl et al., 2025). The differences might be explained by the topics of the studies, the journals, the data extraction methods, and the availability definitions. The values are substantially lower than the rates reported in neighboring disciplines. For instance, previous studies found that code was available in 20% of artificial intelligence conference papers in 2010–2019 (Lin et al., 2022) and 18% of agent-based modeling studies in 2018 (Janssen et al., 2020).

Geographical effect. The availability of code and data are strongly influenced by the location of the authors. Papers with authors based in Europe, South America, Africa, North America, and Oceania are more likely to have code available than those of authors in Asia. Papers with authors based in Europe, North America, Oceania and South America are more likely to have a data repository than those of authors in Asia, with Europe and North America showing the strongest effect. Possible explanations are differences in open science experience, practices, institutional valorization and funding policies. For instance, studies in other domains found that researchers in Asia were more interested in requiring permission to access their data (Tenopir et al., 2015). Researchers in low- and middle-income countries were more concerned about data misuse and privacy violations and had less experience with data sharing outside trusted collaborations (Serwadda et al., 2018).

Temporal effect. The availability of code and data are also strongly influenced by the year of publication. Newer papers more frequently provide code and data repositories, whereas older papers tend to reference existing datasets and do not provide a data repository. Possible explanations are changes in institutional and funding policies that favor open science (Nielsen et al., 2023), the increasing availability of data sharing infrastructure such as repositories and platforms, the introduction of a data availability statement in TR journals, and the promotion of open science principles by transportation associations in recent years (Caicedo et al., 2025; Wu et al., 2024a). Studies in other domains have also shown that data and code sharing have gradually increased over time, influenced by journal policies (Roche et al., 2022; Fink and Marcus, 2025) and possibly technological advances (Fink and Marcus, 2025).

Influence of research topic. The research topic significantly influences the availability of both code and data. Transportation research can be influenced by practices in other domains due to its multidisciplinary nature. Studies on data-driven modeling and prediction exhibit the highest levels of code and data availability. Researchers working on these topics may be influenced by practices that are well established in the machine learning and computer science community, where major conferences encourage the submission of data and code (Pineau et al., 2021). At these conferences, most authors are willing to submit code, a high number of reviewers are interested in consulting the code, and an increasing number of participants are joining reproducibility challenges (Pineau et al., 2021). Studies on COVID-19 showed higher rates of citations to open data and data repositories than studies on optimization, but lower rates of code available. This result may be explained by the fact that several mobility datasets were released during the pandemic to support real time decision-making (Combs and Pardo, 2021; Li et al., 2021). In these emergency conditions, focusing on releasing data may have been considered more valuable than code. Studies on urban environment showed higher rates of citations to open data and repositories than studies on optimization. These studies may use map and land use datasets that are publicly available or can be directly extracted from open source tools (e.g., OpenStreetMap). The barriers to data sharing may therefore be lower compared to other topics. Studies on traffic flow and transport emissions showed lower rates of code and data repositories, while studies on automated driving showed a low rate of code available and citations to existing data than studies on optimization. These topics are dominated by simulation studies in which the code may be difficult to share or tied to closed simulation tools. Citations to existing data on automated driving may be particularly limited because it is an emerging technology. These results are in alignment with previous findings showing low rates of code and data availability for simulation studies in transportation (Riehl et al., 2025). The main reasons reported in surveys were high technical complexity, lack of resources, and lack of experience in sharing materials (Riehl et al., 2025). Studies on driver behavior, transport policy, and travel mode showed low levels of code and data availability than studies on optimization. These transportation topics may rely on proprietary data and involve human participants, raising security, ethical, and privacy concerns. Key data in transportation modeling (e.g., operator data, GPS data and smart-card) are often commercial or inaccessible (Mahajan et al., 2022). The results also mirror patterns observed in the social sciences. A previous study found that willingness to share data is higher in the hard sciences than in the humanities and social sciences (Enwald et al., 2022). Studies on logistics, road infrastructure, supply chain showed the lowest levels of code and data availability. These topics may rely on simulations and proprietary data that agencies are less likely to share due to organizational protocols, security concerns, and lack of resources. A previous study on civil infrastructure engineering found that data of competitive edges, private data, and limited support for data sharing were among the main barriers to data sharing (Wang et al., 2023).

Influence of journal. The journal paper significantly influences the availability of code and data. Even after controlling for the topic of the paper, differences in journal scope, editorial practices, cultural norms and target audience may influence these variations. Papers published in TR-C are the most likely to include data and code repositories. An explanation is that TR-C explicitly promotes open science initiatives and open datasets in its scope. Researchers who value open science may prefer to submit to this journal instead of others. Papers published in TR-B are more likely to include code than the other journals (except TR-C) and less likely to cite open dataset than papers in TR-D and TR-A. These results may indicate that methodological papers value more their code as a contribution. These papers may use synthetic data for validation more often than citing open datasets. In contrast, papers published on TR-A and TR-D may use publicly available data more often due to the emphasis on policy-relevant contributions. Papers published in TR-E are less likely to cite open datasets than papers published in TR-D and TR-A. The finding may be explained by the focus on applied logistics and supply chain, which was not fully captured by the topics. These studies often rely on proprietary data and simulations. Papers published in TR-IP are less likely to cite open data and have a data repository than papers in TR-D and TR-A. This result may be due to the different methodologies and data types that characterize interdisciplinary research and may hinder the release of open data. Papers published in TR-F are the least likely to cite open data regardless of the data repository. This result may be due to the focus on human factors and traffic psychology, which is not fully captured by the topics. In this field, qualitative or sensitive data involving human subjects may not be publicly shared due to privacy concerns.

Publication characteristics. The characteristics of the manuscript appear to influence open science practices, even after controlling for the topic and the journal. Papers with a high number of figures are more likely to have both code and data available. The high number of figures may be required to visualize computationally intensive workflows. Papers with a high number of tables are more likely to have citations to open data or a data repository. The high number of tables may be used to summarize quantitative datasets. By contrast, papers with a high number of pages are less likely to have citations to open data and data repositories. This result may be linked to papers that make a more theoretical contribution without requiring extensive data and code. Papers with a higher number of authors are less likely to have citations to open datasets. This result may reflect the reuse of data that are not publicly available with trusted collaborators.

The incentive gap. Despite increasing momentum toward open science in the transportation research community, the findings point to a persistent misalignment in academic incentives. Within the observed period, there is no evidence that sharing code or data is associated with higher citation rates or decreased review times. Taken together, these patterns suggest that current peer-review and reward structures offer limited professional returns for the additional effort required to make research towards reproducible.

7.2. Performance

Based on our analysis, the performance of the extraction pipeline is sufficient for analysis. Section 5.4.2 shows that, for most features, the agreement between LLMs and the expert annotator is similar to the agreement between expert annotators. Metrics such as Fleiss's Kappa or Cohen's Kappa show that the LLM can achieve acceptable results on all features. We expect that the extraction and analysis pipeline can be readily adapted to other transportation journals. Given the relatively high agreement across the features considered, we additionally believe that numerous other questions about open science practices can be considered and validated using the same methodology.

7.3. Limitations

7.3.1. Limitations in the sample for analysis

The number of papers that were annotated by two experts was fairly small compared to the number of papers used to extract features. Because we used data up until 2025, our work is a snapshot of the state of availability until then. This can serve to monitor community progress, but these conclusions may not hold for future work.

7.3.2. Limitations in the topic modeling

The LDA model used for topic modeling in this paper is also limited in its ability to classify papers. For instance, a number of papers classified under COVID-19 and pandemic were published before 2020 and refer to other emergency situations that limited mobility. We would like to clarify that this comes from the nature of the LDA model: it is a generative probabilistic (Bayesian) model built on the bag-of-words assumption. Thus, it is an unsupervised probabilistic topic model that groups documents by word co-occurrence. Its output is a distribution over topics for each paper (each paper is a mixture of topics) and a distribution over words for each topic. We then name each topic grouping using the top words from each topic. Although LDA represents each paper as a mixture, we assign each paper to its single highest-probability (via `argmax`) topic. For interdisciplinary papers, which are common in transportation, this discards mixed-membership information and can place a paper in a category an expert may not choose first.

We explored other categorizations based on domain taxonomies, for example the Transportation Research Board (TRB) committee structure, but found it difficult to reach consistent agreement on how to assign papers under such schemes. We therefore adopted LDA, as it is the most widely used approach for topic analysis in the transportation research literature (Sun and Yin, 2017). Topic categories enter the choice models as explanatory variables. A different categorization could therefore change which topics reach significance and the magnitude of their coefficients. Our choice models do not use all 15 categories individually. As shown in Tables 8 and 10, several topics turned out to have statistically similar effects, and these were therefore combined into grouped binary variables, yielding a coarser specification than the raw LDA output.

7.3.3. Limitations in the LLM extraction

While LLMs were the only solution to extract the necessary data with sufficient accuracy at the scale of the study, they are still imperfect. For example, LLMs could not perfectly recover metadata available in the XML file. Additionally, given that we did not set the temperature to 0 during extraction, our exact output is difficult to reproduce. Given that human annotations are also not perfectly reproducible, this is a limitation inherent to studies of this kind. However, our pipeline and the extracted data are openly available.

7.3.4. Limitations in the gap between availability and reproducibility

While this study helps measure open science practices, the proposed pipeline does not actually measure paper reproducibility, which would require significant additional human labor. However, we can conclude that most transportation papers are currently very difficult to reproduce, given that they do not meet the necessary condition to share the data and code they used.

8. Recommendations

We offer evidence-based recommendations from the lens of open science–not as an ends in itself–but as an enabler for accelerating research progress and strengthening scientific integrity. Thus, our recommendations focus on pragmatism and aligning desired behavior with researcher incentives. Overall, because a researcher's primary incentive is to publish papers and garner citations, we believe that publishers have the highest leverage to influence open science practices in transportation, so we focus our recommendations there.

This study confirms that the transportation research community is in the nascent stages of adopting open science practices. At the same time, it signifies that the community has a tremendous opportunity to shape the adoption of open science and the downstream impact. The following recommendations are informed by the findings and data produced from this study, the open science journeys of other research communities (both positive and cautionary), and by our collective decades of experience in the transportation research field.

8.1. Publishers

For publishers, we recommend a balanced approach: promote practices that build momentum in open science, while avoiding passing an undue burden of their implementation solely onto paper authors. We thus recommend explicitly rewarding behaviors that make it easier for authors to engage in open science.

Our study found no significant difference in citation counts or review duration between papers that provided data and code and those that did not, suggesting that traditional academic metrics do not incentivize researchers to practice open science. This is in contrast to the sentiment expressed by some journals; for example, Transportation Research Part C already conveys a positive stance on open science, with their guide for authors stating that sharing data "increases your exposure and may lead to new collaborations [and] provides readers with a better understanding of the described research". While soft signaling led to significant impact in the machine learning community (achieving as high as 79% code availability) (Pineau et al., 2021), it appears insufficient for the transportation research community. At the most extreme, journals such as Management Science require valid repository links (e.g., to GitHub, Zenodo, or Dataverse) or even full replicability prior to publication (Fišar et al., 2024; INFORMS Journal on Computing, 2025; Ferreira et al., 2010). A key advantage of compulsory open science policies is that they can shift researcher incentives and thus shift behavior 'overnight'. At the same time, these policies must be implemented carefully as it creates additional challenges, such as inhibiting publication of research involving sensitive or proprietary data or slowing the review process for researchers at critical career stages.

Rather than promote open science for the sake of open science, we suggest identifying policies or initiatives that better align researcher incentives with open science practices to accelerate transportation research. Success can be measured by highly and rapidly cited works that the community can build upon owing to its open data and code. At present, some of the most popular transportation datasets and benchmarks are not published in transportation journals (Caesar et al., 2020; Sun et al., 2020; Ettinger et al., 2021; Barnes et al., 2020; Bock et al., 2020; Wu et al., 2021; Jiang et al., 2021). Thus, a high-leverage behavior that publishers can use to bootstrap the practice of open science in transportation is to encourage the publication of high-quality open datasets and benchmarks. Not only will this ease the open sourcing of downstream artifacts, it will create a virtuous cycle where community members have access to exemplars to develop more datasets and benchmarks. Open datasets and benchmarks have the added benefits of improving the comparability of research and thus the measurement of progress, and attracting newcomers and talent from other fields. Journals can incentivize open datasets and benchmarks by hosting special issues (Buisson et al., 2025), explicitly including open data and benchmarks within the aims and scope of the journal, and hiring dedicated editors to handle such topics.

Another strategy to give credit to researchers who contribute to open science is to augment the CRediT (Contributor Roles Taxonomy) author statement (Allen et al., 2019) to explicitly list open science practices—such as creating a data archive or reproducing a study—as contributions to a research project.

There are additional low hanging fruits that publishers can implement. For instance, we suggest mandating structured data and code availability statements in submissions. Currently, none of the journals considered encourages or requires a code availability statement. TR-A, TR-B, TR-D, TR-E, encourage authors to share a data availability statement, but do not require it, while TR-C, TR-F, and TRIP require authors to include a data availability statement. Compared to TR-D, only papers from TR-C were more likely to have a data repository regardless of data citations. Compared to papers published in TR-D, papers published in TR-B and TR-C are more likely to have a code repository available, while papers published in TR-A, TR-E, TR-F, and TR-IP were not significantly more

likely to have a code repository available. In addition to encouraging open science practices, availability statements make it easier to measure the state of open science in the transportation community.

As numerous communities have done, peer-review guidelines could incorporate open science checklists without prolonging review times. Furthermore, in the era of LLMs, publishers could implement automated checklists using LLMs to process the full text of a submitted paper, extending the process implemented in this paper. However, we caution against implementing these checklists without incorporating them into the reviewer and editor workflow, lest they turn into a box-ticking exercise with limited accountability—automated or not. Additionally, drawing from successful initiatives in adjacent fields (ACM, 2020), journals could award open science "badges" for verified availability, signaling quality and encouraging compliance. These policies should address transportation-specific challenges, such as large dataset sizes and regional data heterogeneity, by allowing exceptions for proprietary data while promoting anonymized alternatives.

8.2. Conferences

Conferences can amplify the value of open science to the community. Conferences are the primary gathering place for researchers, and as such they play a critical role in conveying what is important and valued by the community. Conferences have numerous mechanisms for directing focus and attention, including but not limited to keynote talks, awards, and tutorials. For example, the machine learning community successfully drew attention to a reproducibility crisis in reinforcement learning by dedicating a keynote (Pineau, 2018) and several workshops to the topic. Similarly, conferences can issue paper awards associated with open science, such as "Best Open Paper". Conferences can signal the importance of sessions dedicated to open science by placing them in dedicated slots with no or few parallel sessions.

8.3. Funding agencies

The absence of robust open science practices imposes hidden costs on funding agencies such as the US DOT, NSF, and EU Horizon programs, which collectively fund redundant data collection and methodological development across multiple research teams addressing similar questions. As such, funding agencies can intervene by prioritizing funding for research teams that commit to open sharing; for instance, in the United States, "Funder requirement" is a "consistently leading motivating factor for data sharing" (Hahnel et al., 2024). Funding agencies can also support the creation of datasets or benchmarks for topics with low adoption like supply chain logistics or resilience and evacuation, targeted incentives. Discussed next, funding agencies can also help by supporting volunteer groups such as RERITE (RERITE, 2024), which commit to working across transportation research stakeholders to advance open science and scientific integrity.

8.4. Volunteer groups

In some fields such as computer science and physics, grassroots volunteer-based efforts lead to the development of open source tools and practices such as releasing preprints (Armeni et al., 2021). However, these efforts face resistance in transportation, motivating dedicated open science communities like RERITE (RERITE, 2024) that push the field at large forward to adopt these practices. Volunteer groups can engage in scientific research to support evidence-based open science practices (such as this article), and can institutionalize the effort in the form of an annual "transparency audit" to support open science decision makers, such as publishers, conferences, authors, etc. More broadly, these volunteer groups can bridge the gap where existing institutions are not fully equipped to advance open science. Examples include: rewarding open science contributions by promoting and disseminating open research; aiding in the creation of pedagogical materials for dissemination within educational institutions, conferences, and summer schools[4]; holding regular seminars and workshops to form an open science community within transportation; preparing reviewers, editors, and awards committees on how to evaluate open science practices; and, more generally, bootstrapping a culture of open science within the transportation community.

8.5. Research and educational institutions

Educational institutions can incorporate open science into their curricula. This can range from basic classroom modules on version control and data archiving to virtualization (e.g., Docker) and reproducibility workflows to intensive course projects that attempt to reproduce a transportation research paper.

Research institutions should recognize that research progress will be slow when researchers cannot build upon the work of others, as is the case when code and data availability are as low as they are currently in transportation research. They can explicitly factor this into hiring and promotion evaluations, and recognize the long-term scientific importance of dataset and benchmark contributions that lay the foundation for research progress. Universities can also implement open science and data sharing policies, with many of the highest-publishing universities around the world already doing so (Hahnel et al., 2024).

[4] Practical skills include how to anonymize sensitive data, use version control for code, and cite shared resources properly. Example include the reproducibility tutorials at ITSC 2024 (Wu et al., 2024a) and ITSC 2025 (Zheng et al., 2025). Our topic analysis and the extracted data availability statements could be mined to tailor educational material to specific community or sub-community needs.

8.6. Authors

Authors can contribute by learning, implementing, and sharing open science practices. Authors should recognize that they are the primary beneficiaries of practicing open science—by making their code and data available to the community, they inadvertently make it available to *themselves* and their collaborators six months down the line, when it comes time for a paper revision or a project extension. At the same time, it is more common that the implementation of open science is dictated by circumstance, for example based on the culture of the research group or the preferences of the research sponsor or data provider. Research group leaders can influence this by openly stating their support of open science. Data containing proprietary or personal information may be difficult to share, but this is likely not the limiting factor for most of the papers that do not share data. The majority of data availability statements gave "Data Available Upon Request" as the reason the data was not shared, suggesting that privacy is not the primary reason and that a lack of training in how to share data likely plays a bigger role.

We encourage authors to seek out creative ways to practice open science; for example, when working with proprietary data, authors can often make post-processed derived data available, or they can make their code available, along with some synthetic data. Even when open science is permitted, it is rarely encouraged, and there are numerous inhibitors along the way; for instance, some students have anecdotally told us that they sometimes feel embarrassed to make their code available because it's not "pretty." While we insist that available is better than pretty, some authors' concerns will not be assuaged by this. For those authors, thankfully, today's LLMs can assist authors in writing documentation and README files, and cleaning up the file and directory structure.

8.7. Researchers

We encourage researchers to continue conducting research to promote evidence-based decision making to advance transportation research. For example, interesting topics of further investigation could include more detailed analyses of transportation research topics that have notably high or low levels of open science practice, identifying and designing key missing benchmarks for transportation research, or exploring the use of LLMs for replicating key transportation research studies.

9. Conclusion

Our work highlights that the percentage of papers in transportation research using open science practices is critically low, with only 5% sharing a code repository and 4% sharing a data repository. Given the benefits of open science and the many other fields that have adopted such practices, it is both necessary and possible for the field of transportation research to improve. Given that there is not a significant link between open science practices and incentives like citations or reduced review time, encouraging more open science practices will likely require active changes made to be made by publishers, conferences, and funding bodies. Positive examples, such as TR-C, can serve as models for the rest of the community. Beyond the empirical results, this study provides a reusable measurement framework for transforming unstructured statements about research artifacts into structured and auditable open science indicators.

Furthering open science will also accelerate the field of transportation research itself. We found that 29% of papers cite or link to an open dataset, yet only 4% share a data repository. Given that some of these repositories contain repackaged datasets, less than 4% of the papers we reviewed contributed new data to the transportation community. This severe asymmetry suggests that data from a few works form the basis of a large proportion of the field. If more authors released new open datasets, they could have a significant impact on the field. We should also be aware of the risks of open science in the era of LLMs (Acion et al., 2023), that open science practices could facilitate the generation of fake research. However, we believe that the benefits of open science outweigh the risks, and that open science best practices can be designed to mitigate these risks. For example, sharing code and data can make it easier to detect and prevent fraud.

Measuring open science is not a one-shot effort, which is why we have developed an automated pipeline that can continue to be used to track the progress of code and data availability in the transportation research journals. The framework developed in this study establishes a baseline and can be repeatedly applied to track changes in code and data availability in journals and, with appropriate adaptation and validation, in other publication venues. We are committed to transparency and reproducibility in our research, and we also benefit from the reproducible pipeline introduced in recent related work (Riehl et al., 2025), which enables our study to build upon prior efforts and move the field forward. We hope to periodically check new papers and release their statistics to monitor and audit the progress.

Additionally, there are a number of directions this work could be expanded. Although we did a limited exploration of the repositories we found, we hope to further study the documentation of the code and data shared. Additional exploration into funding bodies and open science practices could also yield insights about what policies are effective in encouraging open science. The pipeline introduced in this study could be applied to additional journals from a broader range of publishers, although this depends on each publisher's policy regarding the use of full-text content for LLM-related applications.

Taken together, we envision this work as a first step toward accelerating open science in transportation, and we invite the community to reuse, challenge, and extend the pipeline and analysis presented here. We look forward to the new discoveries and the faster translation from research to practice that a sustainable open-source ecosystem, built on openly shared data, code, and research artifacts, will make possible.

CRediT authorship contribution statement

Junyi Ji: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Ruth Lu:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis; **Linda Belkessa:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation; **Liming Wang:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation; **Silvia Varotto:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization; **Yongqi Dong:** Writing – review & editing, Writing – original draft, Methodology, Investigation; **Nicolas Saunier:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation; **Mostafa Ameli:** Writing – review & editing, Writing – original draft, Supervision, Resources, Investigation, Funding acquisition, Formal analysis; **Gregory S. Macfarlane:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis; **Bahman Madadi:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis; **Cathy Wu:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of generative AI use

During the preparation of this work, the author used the Google Gemini 2.5 in order to extract the open science features from scientific papers. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

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 appreciate the feedback and support from our colleagues, including Zuduo Zheng at University of Queensland, Christine Buisson at École Nationale des Travaux Publics de l'État (ENTPE), Bidisha Ghosh at Trinity College Dublin, Joan Walker at UC Berkeley, Xinyu Chen at MIT, Gergely Zachár, Daniel B. Work and Jonathan Sprinkle at Vanderbilt University, Micah Altman at MIT Libraries. We gratefully acknowledge the volunteers who contributed to the manual validation dataset, including Khadidja Kadem, Cong Son Duong, Ali Naaman, Jincheng Ke, Maissa Mati, and Tuan Nguyen from Gustave Eiffel University COSYS/GRETTIA; Cameron Hickert, Jung-Hoon Cho, Shreyaa Raghavan, Vindula Jayawardana, Han Zheng, Hank Lin, Tianyue Zhou, and Edgar Ramirez Sanchez from MIT; Yiru Jiao from TU Delft; and Zhexian Li from USC. Their efforts were essential to the accuracy and reliability of our feature extraction process. Early versions of this work were presented at Transportation Research Symposium (TRS) 2025, Transportation Research Board Conference on Data and AI for Transportation Advancement (TRB DATA) 2025, International Symposium on Transportation Data and Modeling (ISTDM) 2025, Modeling Mobility Conference (MoMo) 2025, and ITS Irvine Symposium on Emerging Research in Transportation 2026. We appreciate the feedback we received from these presentations. Given her role as Editor, Cathy Wu had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to another journal editor. We thank the anonymous reviewers for their constructive comments, which have helped improve this work.

This research was partially supported by the National Science Foundation (NSF) CAREER award #2239566 (Wu), NSF award #2434399 (Wu), the Marie Sklodowska-Curie COFUND grant agreement no. 101,034,248 (Belkessa, Ameli), French ANR award #ANR-24-CE22-7264 (Ameli) and #ANR-24-CE22-2750 (Varotto). The views expressed herein are those of the authors and do not necessarily reflect the views of the funding agencies.

Data availability

We are committed to transparency and reproducibility in our research. The data used in this study include (i) raw full-text articles from Elsevier Transportation Research journals, (ii) features extracted using large language models (LLMs), and (iii) a manual labeled dataset for validation. The raw dataset for the full-text paper is not available due to the copyright policy of Elsevier (Elsevier, 2025b) for text and data mining, but the pipeline to extract the dataset using the API is available (see code availability). We therefore release three datasets: (i) a sample of full-text articles under the Creative Commons Attribution 4.0 license, (ii) the analysis dataset, and (iii) the manual validation dataset. The supporting data (Ji et al., 2026) is publicly avaiable at https://zenodo.org/records/23045963.

Code availability

Our code is publicly available at https://github.com/RRinTransportation/OpenMOST, covering full-text extraction, LLM-based feature extraction, and statistical analysis.

Appendix A. Feature definitions

These were the feature definitions provided to the manual annotators.

`is_quantitative_study`**:** Did the paper perform any quantitative calculations?

`is_code_publicly_available`**:** Does the paper claim to share the code used?

- **True** only if there is a link to a website, repository, or files containing code for the project.
- The link does not have to be active, it just needs to be stated in the paper. (i.e. "We share our code at nolongeractivelink.org." - link activeness will be checked by another feature)

`is_code_link_valid`**:** Is the link actually live and contains code files?

`is_data_repository_available`**:** Does the paper share a data repository? The repository must be created by the authors but does not need to contain data the authors created; i.e. it can contain processed data from NGSIM.

- **True** if the paper shares a link containing all data used for the experiment.
- **True** if a paper used a simulated dataset and shares code to fully reproduce the data.
- **False** if only certain sample instances of data are shared.
- **False** if the paper used data from NGSIM and links to NGSIM, but does not link to a repository where their preprocessed data is stored.
- **False** if the only data available is through supplementary materials (not accessible by everyone).

`is_data_link_valid`**:** Is the link actually live and contains data files?

`is_data_cited`**:** Does the paper use and cite at least 1 publicly available dataset? (This would include datasets released by governments as well as datasets like NGSIM, Sioux Falls, and Waymo dataset.)

- **True** if a paper uses Open Street Map and a private dataset from an e-bike company.
- **True** if Email/phone registration is required, but False if requires institutional Email/membership or payment.
- **False** if a paper uses a survey they conducted and private data from an airline.
- **False** if a paper uses data from a nonprofit or government that is not publicly available online.

Appendix B. Prompts used for LLM feature extraction

Extracting code-related features

You are an expert assistant specializing in academic paper computational reproducibility analysis. Your task is to extract specific metadata about availability from research papers. You must respond with only a valid `JSON` object, without any additional text comments, or markdown formatting. The papers are from Transportation Research journals. The JSON object must conform to the following schema:

```
{
  "title": "string", // The exact title of the paper.
  "is_code_used": "boolean", // Did the paper use code to reach its
     conclusions?
  "reason_code_is_used": "string", // **If code was used, you MUST
     extract the justification directly from the paper. This field
     MUST contain a VERBATIM quote. Do not summarize or paraphrase.**
  "reason_code_is_not_used": "string", // **If code was not used, you
     MUST extract the justification directly from the paper. This
     field MUST contain a VERBATIM quote. Do not summarize or
     paraphrase.**
  "is_code_publicly_available": "boolean", // If code was used, is is
     publicly available?
  "reason_code_available": "string", // **If code is available, you
     MUST extract the justification directly from the paper. This
     field MUST contain a VERBATIM quote. Do not summarize or
     paraphrase. For example, if the paper says, \'The code for this
     project can be found on GitHub\', you must return that exact
     phrase. If no reason is stated, return an empty string.**
  "reason_code_unavailable": "string", // **If code is not available,
     you MUST extract the reason directly from the paper. If no reason
      iss tated, return an empty string. This field MUST contain a
     VERBATIM quote. Do not summarize or paraphrase.**
}
...
```

Extracting data-related features

You are an expert assistant tasked with extracting metadata about research dataset availability — defined here as processed data by the team — from papers published in Transportation Research journals. The journal policy encourages research data deposit, citation and linking. Which means that there are 3 categories of data availability: (i) data repository (e.g. Zenodo, Figshare, etc.), (ii) data cited or linked to a public source (e.g. open data from a government agency or other research institutions), and (iii) data not available.
Reply with one valid JSON object only - no extra text, comments, or markdown—using the schema below:

```
    { "title": "string", // Exact paper title from the markdown file
       title

"is_data_used": "boolean", // Did the study explicitly rely on a
   dataset (not just a source/statistics from a report) to produce
   results (e.g. a paragraph describing the data used)?

"is_simulation_study": "boolean", // Is the study a simulation study (
   e.g. purely using a simulation tool to produce results)?

"reason_simulation_study": "string", // If yes, VERBATIM justification
    from the paper (no paraphrasing)

"reason_data_is_used": "string", // If yes, VERBATIM justification
   from the paper (no paraphrasing)

"is_data_repository_available": "boolean", // Is the data available in
    a repository (e.g. Zenodo, Figshare, Github, etc.)? Data is NOT
   the same as code or implementation or algorithms, but both can be
   in the same repository.

"reason_data_repository_available": "string", // If yes, VERBATIM
   justification from the paper (no paraphrasing)

}
...
```

Note: The rest of the prompts can be found on our repository.

Appendix C. Justification for the usage of LLMs

To justify our use of LLMs, we compare our process to the method developed and outlined by Riehl et al. (2025), which uses text search (TS) to find relevant links and keywords in the text (Table C.12, C.13).

The agreement percentage and Fleiss Kappa of text search is lower for all of the features because text search overestimated the prevalence (Table C.14). Most notably, `is_code_publicly_available` has moderate agreement (0.6364) using text search, while having substantial agreement with an LLM (0.8388). For `is_code_publicly_available`, we find that a simple text search for a github repository link overestimates code availability, leading to its decreased accuracy (Table C.14). Finally, text search cannot separate data and code repositories.

Appendix D. Data postprocessing

Postprocessing is fully deterministic and is designed to (i) separate paper-level attributes from artifact-level evidence and (ii) ensure that all downstream variables are computed from a consistent, auditable set of intermediate tables. We construct three relational

Table C.12
Kappa values for text matching feature extraction versus LLM extraction. Interpretation guide with Fleiss's κ: 0.21-0.40 (Fair), 0.41-0.60 (Moderate), 0.61-0.80 (Substantial), and 0.81-1.00 (Almost Perfect) (Landis and Koch, 1977).

Feature	Cohen Kappa(H1 vs H2)	Cohen Kappa(H1 vs TS)	Cohen Kappa(H2 vs TS)	Fleiss Kappa(H1 vs H2 vs TS)	Fleiss Kappa(H1 vs H2 vs LLM)
is_quantitative_study	0.4757	-0.0365	-0.0345	0.1782	0.4586
is_code_publicly_available	0.7534	0.5922	0.5961	0.6364	0.8388
is_data_repository_available	0.3333	0.3592	0.0986	0.2618	0.3994
is_data_cited	0.6672	0.0332	0.0275	0.0360	0.5224

Table C.13
Agreement percentages for text matching feature extraction versus LLM.

Feature	Percentage Agreement (H1 vs H2)	Percentage Agreement (H1 vs TS)	Percentage Agreement (H2 vs TS)	Percentage Agreement (H1 vs H2 vs TS)	Percentage Agreement (H1 vs H2 vs LLM)
is_quantitative_study	0.9062	0.8646	0.8958	0.8333	0.8854
is_code_publicly_available	0.9688	0.9271	0.9375	0.9167	0.9688
is_data_repository_available	0.9271	0.8854	0.8750	0.8438	0.9167
is_data_cited	0.8542	0.3750	0.3333	0.2812	0.6979

Table C.14
Prevalence of each feature found by H1, H2 and text search (TS).

Features	Prevalence (H1)	Prevalence (H2)	Prevalence (TS)
is_quantitative_study	0.8854	0.9167	0.9792
is_code_publicly_available	0.0833	0.0521	0.1146
is_data_repository_available	0.0833	0.0312	0.1146
is_data_cited	0.3438	0.3021	0.9688

tables keyed by DOI: a *paper-level* table (one row per DOI) containing bibliographic metadata, topic and region assignments, and final availability indicators; an *artifact-level* table (one row per extracted dataset or artifact associated with a DOI) containing repository attributes such as host category and link validity; and a *link-level* table (one row per DOI–URL pair) used to canonize and deduplicate URLs and to compute per-paper counts by host category. The main steps are as follows:

(i) **Parsing and flattening.** We parse the Elsevier `XML` metadata and merge them with the LLM `JSON` outputs. Nested structures (e.g., multiple dataset mentions or multiple links per paper) are flattened into the artifact- and link-level tables, while document-level fields (e.g., journal, year, authorship, and topic assignment) are stored in the paper-level table keyed by DOI.

(ii) **Identifier and type normalization.** We standardize identifiers and field types across sources. DOIs are lowercased, stripped of common prefixes (e.g., `doi:`), and checked for format validity to preserve referential integrity across the relational tables. Boolean fields returned by the LLM (e.g., `is_code_publicly_available` and `is_data_repository_available`) are coerced to consistent logical values; when the manuscript does not provide sufficient information, values are retained as missing rather than imputed.

(iii) **URL canonization, classification, and validation.** We collect all outbound URLs extracted from full text and LLM outputs, deduplicate them per DOI, and canonize them to a consistent form (e.g., standardizing schemes and removing parameters that do not change resource identity). Each URL is classified into host categories based on its domain (e.g., collaborative code hosts; DOI-minting general-purpose repositories; institutional repositories; and government open-data portals). We then perform automated validity checks to determine whether links are live and whether the landing content corresponds to the claimed artifact type (code versus data). Validity indicators are retained at the link/artifact level and later aggregated to paper-level variables.

(iv) **Aggregation and feature construction.** Artifact- and link-level indicators are aggregated to paper-level variables used in Section 6. In particular, we derive binary indicators for whether a paper provides any code link, whether a data repository is provided, whether at least one dataset is cited or linked, and whether at least one artifact link is valid. We also derive contextual covariates used in modeling (e.g., paper age, journal part, LDA topic, and corresponding-author region), along with summary link-count measures by host category.

(v) **Integrity checks and dataset filtering.** We apply consistency checks across tables, including DOI uniqueness, one-to-many coherence between paper- and link-level records, and logical consistency between availability flags and the presence of at least one URL in the corresponding host category. Records with missing or invalid DOIs, or with irreconcilable inconsistencies, are excluded. After these checks, 10,480 papers are retained from the 10,724 full-text articles.

Fig. D.6 illustrates the postprocessing pipeline as a flowchart.

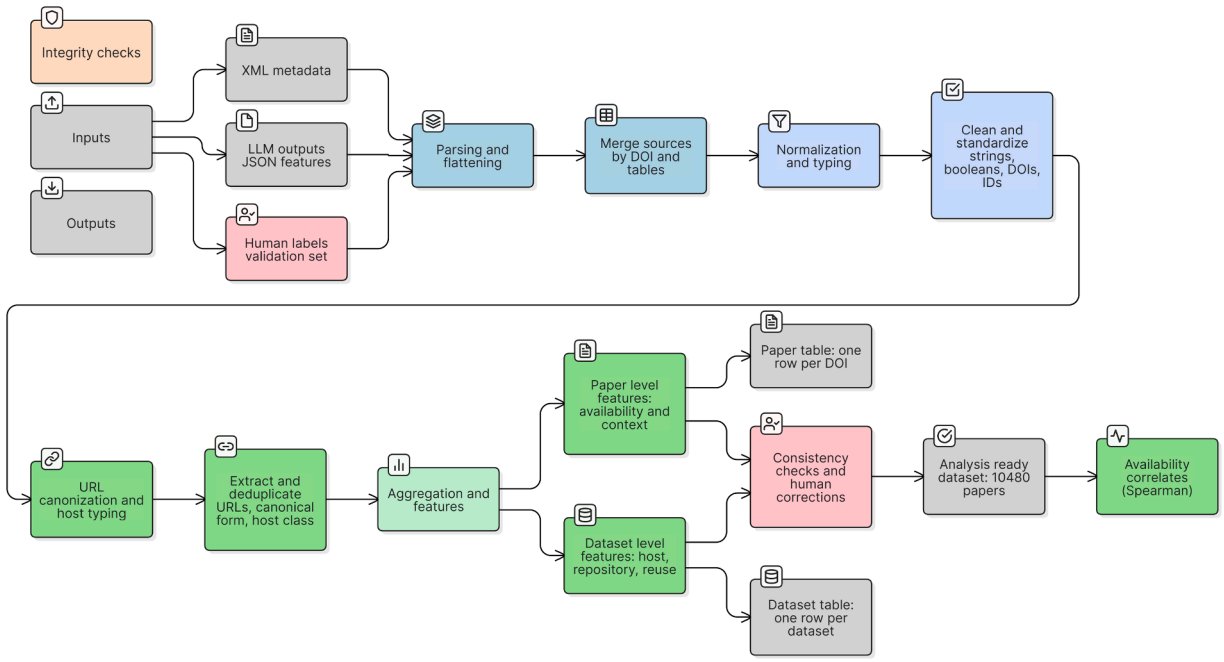


Fig. D.6. Flowchart of the postprocessing pipeline. Starting from raw XML metadata, LLM-extracted `JSON` features, and human annotations, the pipeline (i) parses and flattens inputs into paper- and dataset-level tables, (ii) normalizes types and encodings, (iii) canonizes and classifies URLs into code and data host categories, and (iv) aggregates and validates features to produce the final analysis-ready dataset.

Appendix E. Data dictionary and features

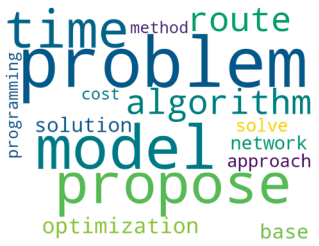
In addition to the metadata features, LDA extracted topics, and LLM extracted features, we also derived additional features. A data dictionary of all features in the final dataset are listed below with their type, origin, and description (Table E.15).

Table E.15
Post-processed and transformed features obtained from initial features in Appendix A.

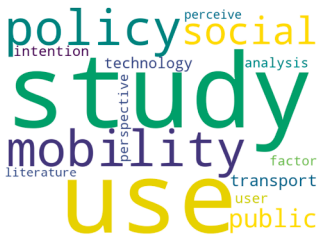
Name	Type	Stage / Origin	Description
code_link	object	parsing	Raw link to code, unprocessed or malformed
code_link_by_the_author	object	LLM extraction	Whether the code link was explicitly provided by the authors
programming_language	object	metadata parsing	Named programming language(s) used in the study (free text)
software_used	object	metadata parsing	Software tools or packages declared in method or appendix
cpu_used	object	metadata parsing	CPU model or specifications when computational config is reported
gpu_used	object	metadata parsing	GPU type used when deep learning or simulations are involved
os_used	object	metadata parsing	Declared operating system used in experiments
ram_used	object	metadata parsing	Amount of RAM reported (in GB)
lda_topic	int64	topic modeling	Raw LDA-inferred topic index before mapping to label
clean_primary_region	object	geographic parsing	Cleaned world region label based on author affiliation
is_data_cited_or_linked	bool	LLM extraction	Whether the data was explicitly cited or linked in the reference list
is_simulation_study	bool	LLM extraction	Flag indicating use of simulated or synthetic data in the analysis
links_to_the_data_repository	object	parsing	Raw list of extracted links to dataset repositories
primary_continent	object	geographic parsing	Redundant or preprocessed version of affiliation continent
n_open_repo	int64	aggregation	Count of dataset links hosted in open-science repositories (Zenodo, Figshare, etc.)
n_code_links	int64	aggregation	Total number of code links parsed (regardless of validity)
has_code_link_gitlab	bool	derived feature	Whether a GitLab URL is included in the code links
has_code_link_bitbucket	bool	derived feature	Whether a Bitbucket URL is included in the code links
availability_reasons	object	LLM extraction	LLM-provided rationale for dataset/code availability or absence
n_src_benchmark_hub	int64	source counts	Number of links coming from benchmark-hosting hubs (e.g., PapersWithCode)
n_src_code_host	int64	source counts	Number of code links from recognized code hosting platforms
n_src_gov_open_data	int64	source counts	Number of dataset links pointing to government open data portals
n_src_ngo_nonprofit_org	int64	source counts	Links to NGO or nonprofit-hosted datasets
n_src_open_repo_doi	int64	source counts	Number of datasets hosted on DOI-backed repositories (e.g., Zenodo)
n_src_other_web	int64	source counts	Links to generic web URLs not falling in known host categories
n_src_university_repo	int64	source counts	Number of links to university-hosted data repositories
uses_data	bool	derived feature	Final flag indicating actual dataset use after manual correction
paper_age_years	float64	derived metric	Years since publication (based on system crawl date)
region_normalized	object	derived feature	Encoded world region label after mapping
is_region_asia	bool	one-hot encoding	Whether primary affiliation is located in Asia
is_region_europe	bool	one-hot encoding	Whether primary affiliation is located in Europe
is_region_north_america	bool	one-hot encoding	Whether primary affiliation is located in North America
is_region_oceania	bool	one-hot encoding	Whether primary affiliation is located in Oceania
is_region_south_america	bool	one-hot encoding	Whether primary affiliation is located in South America
is_region_africa	bool	one-hot encoding	Whether primary affiliation is located in Africa
n_regions_listed	int64	derived feature	Number of distinct global regions covered by author affiliations
journal_rate_atleast	float64	journal aggregation	Journal-level share of papers with at least one dataset available
journal_rate_all	float64	journal aggregation	Journal-level share of papers with all datasets available
journal_n	int64	journal aggregation	Number of papers published in the journal in the dataset
papers_is_data_used_true	int64	journal aggregation	Raw count of papers that used data
papers_is_data_used_share	float64	journal aggregation	Share of papers in journal that used data
papers_is_data_cited_or _linked_share	float64	journal aggregation	Share of papers linking or citing datasets
papers_atleast_one_dataset _available_share	float64	journal aggregation	Share of papers with at least one dataset accessible
papers_all_datasets _available_share	float64	journal aggregation	Share of papers with all datasets accessible
papers_uses_data_share	float64	journal aggregation	Share of papers in journal that used data (final label)

Appendix F. Topic

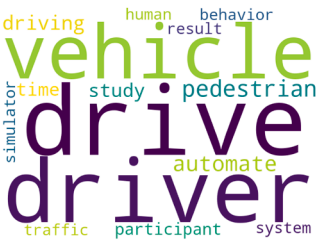
Topics were extracted using the gensim (Řehůřek and Sojka, 2010) package's LDA with 15 categories (decided by using coherence, with 15 categories having a coherence of 0.499). The LDA model was given the title, abstract, journal name, and keywords. The preprocessing step removed all non-alphabetic characters, excess spaces, emails, and made everything lowercase. The most relevant words for each topic are listed in Fig. F.7.



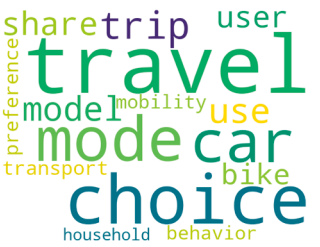
(a) Optimization



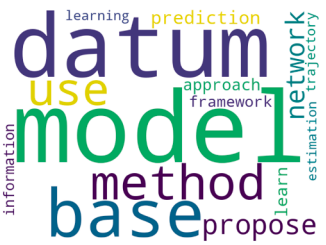
(b) Mobility policy



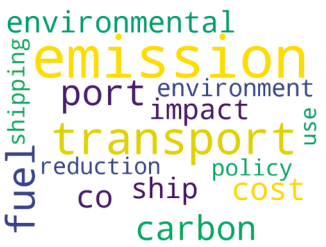
(c) Driving behavior in automated driving



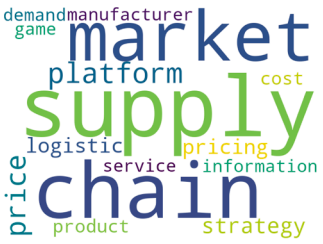
(d) Choice model



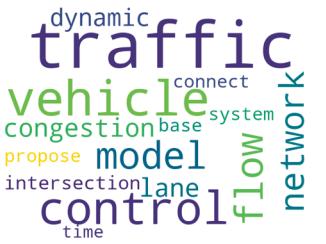
(e) Data-driven modeling and prediction



(f) Emissions



(g) Supply chain



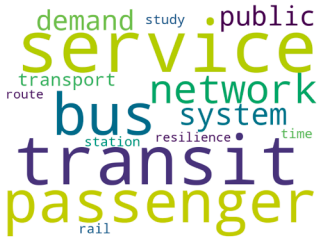
(h) Traffic flow and control



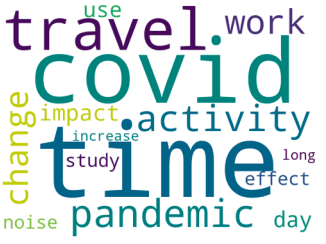
(i) Safety aspects of driving behavior



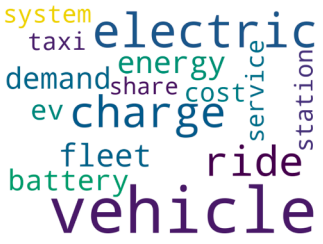
(j) Active transportation and land use



(k) Public transit



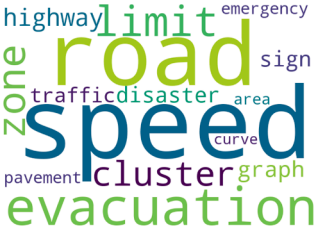
(l) COVID-19 and pandemic



(m) Electric vehicles and ride-sharing



(n) Air and freight logistics



(o) Resilience and evacuation

Fig. F.7. Word cloud visualization for each identified topic.

Appendix G. Other models and prompts

The exact same pipeline was used with a number of other models, some with more parameters, and some with thinking (Zhong et al., 2023). As shown in Table G.16, outside of `is_code_publicly_available`, most of the other models perform within the same interval as the one we used, Gemini-2.5-Flash-Lite. This confirms that our pipeline is not dependent on the exact model used, and can be updated with newer models if needed.

Table G.16
Comparison of agreement metrics across LLM configurations (PA: percent agreemend; FK: Fleiss Kappa).

	Gemini 2.5 Flash (w/ reason)		Qwen3.6-27B		GPT (reasoning)		GPT (no reasoning)		Gemini 2.5 Flash Lite (used in this study)	
Variable	PA	FK	PA	FK	PA	FK	PA	FK	PA	FK
is_code_publicly_available	0.9688	0.8310	0.9583	0.7945	0.9375	0.7273	0.9167	0.6618	0.9688	0.8388
is_data_repository_available	0.8542	0.3110	0.9062	0.4374	0.9062	0.3243	0.8854	0.2721	0.9167	0.3994
is_quantitative_study	0.8125	0.3997	0.8125	0.4146	0.8125	0.3840	0.8750	0.3836	0.8854	0.4586
is_data_cited	0.7292	0.5772	0.7812	0.6783	0.7604	0.6283	0.5833	0.4335	0.6979	0.5224

Note: PA = Percentage Agreement, FK = Fleiss Kappa.

To test the effect of changing prompts on the pipeline using Gemini-2.5-Flash-Lite (which does not have thinking enabled), we also used two different prompts to extract results: a prompt that consolidated all features (Oneshot), and one that told the model to manually use chain of thought (Manual CoT). As expected, Table G.17 shows that oneshot prompts performed worse overall, likely because many of the prompts were not similar, and the model struggled to handle them all at once. On the other hand, using manual CoT only improved results for `is_data_cited`. This may be because Gemini-2.5-Flash-Lite is not powerful enough to derive useful results from CoT, or the CoT prompting was not fine-tuned enough for this specific task.

Table G.17
Comparison of prompting methods.

	Manual CoT		Oneshot		*Multiple Prompts*	
Features	% Agr.	Fleiss κ	% Agr.	Fleiss κ	% Agr.	Fleiss κ
`is_code_publicly_available`	0.9688	0.8388	0.9688	0.8388	0.9688	0.8388
`is_data_repository_available`	0.9167	0.4706	0.9167	0.3043	0.9167	0.3994
`is_quantitative_study`	0.7917	0.3495	0.8958	0.5913	0.8854	0.4586
`is_data_cited`	0.6979	0.5091	0.7188	0.5397	0.6979	0.5224

We also test the many-shot in-context learning (ICL) (Agarwal et al., 2024). To characterize its effect, we report a focused many-shot ICL experiment on the manual validation set. To avoid contaminating the evaluation, we did not place all of our validated ground truth into the prompt, as this would leave no held-out samples for a fair comparison. We instead used 20 of our 96 validated samples with full human agreement as in-context examples and report results on the remaining 76 (Table G.18).

Table G.18
Results of many-shot in-context prompting.

	Results	
Features	% Agr.	Fleiss κ
`is_code_publicly_available`	0.7188	0.1219
`is_data_repository_available`	0.6979	-0.0195
`is_quantitative_study`	0.8646	0.3015
`is_data_cited`	0.6042	0.2190

As shown in Table G.18, performance under this configuration is lower than under our other prompting methods, particularly for `is_code_publicly_available` and `is_data_repository_available`. We do not read this as evidence that many-shot ICL is ineffective in general; rather, three factors specific to our setup likely contribute. First, only 20 examples in this study were available for the prompt, a small set from which to generalize. Second, including these examples expanded the prompt from under 20,000 tokens to roughly 300,000 tokens — well within the 1M-token limit of Gemini-2.5-flash-lite, but long enough that the result is consistent with the long-context degradation reported for models in this class. Finally, because we used human samples with full agreement in the prompt context, the remaining papers will have a lower agreement. Our design does not fully separate these effects, so we treat the result as empirical rather than conclusive. A stronger frontier model may behave differently, and we note this as a limitation and a direction for future study.

References

Acion, L., Rajngewerc, M., Randall, G., Etcheverry, L., 2023. Generative AI poses ethical challenges for open science. Nat. Hum. Behav. 7 (11), 1800–1801.

ACM, 2020. Artifact review and badging. Version 1.1. https://www.acm.org/publications/policies/artifact-review-badging

Agarwal, R., Singh, A., Zhang, L., Bohnet, B., Rosias, L., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., et al., 2024. Many-shot in-context learning. Adv. Neural Inf. Process. Syst. 37, 76930–76966.

Agarwal, S., Sahu, G., Puri, A., Laradji, I.H., Dvijotham, K. D.J., Stanley, J., Charlin, L., Pal, C., 2025. LitLLM: a toolkit for scientific literature review. arXiv:2402.01788.

Aims & scope - transportation research Part C: emerging technologies. https://www.sciencedirect.com/journal/transportation-research-part-c-emerging-technologies/about/aims-and-scope. Accessed: 2025-06-16.

Akhtar, M., Benjelloun, O., Conforti, C., Foschini, L., Giner-Miguelez, J., Gijsbers, P., Goswami, S., Jain, N., Karamousadakis, M., Kuchnik, M., et al., 2024. Croissant: a metadata format for ML-ready datasets. Adv. Neural Inf. Process. Syst. 37, 82133–82148.

Allen, L., O'Connell, A., Kiermer, V., 2019. How can we ensure visibility and diversity in research contributions? How the contributor role taxonomy (CRedit) is helping the shift from authorship to contributorship. Learn. Publ. 32 (1), 71–74.

Antoniou, C., Zhang, A., Abouelela, M., 2025. An overview of Transportation Research Part A: Policy and Practice in the past 45 years (1979–2025). Transportation Research Part A: Policy and Practice, 201, 104677.

Armeni, K., Brinkman, L., Carlsson, R., Eerland, A., Fijten, R., Fondberg, R., Heininga, V.E., Heunis, S., Koh, W.Q., Masselink, M., et al., 2021. Towards wide-scale adoption of open science practices: the role of open science communities. Sci. Public Policy 48 (5), 605–611.

Bae, A.J., 2026. Peer review governance in transportation research: a systematic audit of editorial practices across 108 major transportation journals. Transp. Res. Interdiscip. Perspect. 38, 102058.

Barmpounakis, E., Geroliminis, N., 2020. On the new era of urban traffic monitoring with massive drone data: the pNEUMA large-scale field experiment. Transp. Res. Part C: Emerg. Technol. 111, 50–71.

Barnes, D., Gadd, M., Murcutt, P., Newman, P., Posner, I., 2020. The oxford radar robotcar dataset: a radar extension to the oxford robotcar dataset. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 6433–6438.

Bhaskar, A., Stodden, V., 2024. Reproscreener: leveraging LLMs for assessing computational reproducibility of machine learning pipelines. In: Proceedings of the 2nd ACM Conference on Reproducibility and Replicability. Association for Computing Machinery, New York, NY, USA, p. 101-109. https://doi.org/10.1145/3641525.3663629

Bierlaire, M., 2023. A Short Introduction to Biogeme. Technical Report TRANSP-OR 230620. Transport and Mobility Laboratory, École Polytechnique Fédérale de Lausanne. Version prepared using Biogeme 3.2.11 (April 20, 2023). https://transp-or.epfl.ch/documents/technicalReports/Bier23.pdf

Bierlaire, M., Axhausen, K., Abay, G., 2001. The acceptance of modal innovation: the case of swissmetro. In: Swiss Transport Research Conference. Vol. 1.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3 (4–5), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Bobrov, E., Bracco, L., Dacos, M., Fressengeas, N., Hrynaszkiewicz, I., Iarkaeva, A., Peršić, A., Proudman, V., Romary, L., Sabo, R., 2025. The principles of open science monitoring. https://doi.org/10.5281/zenodo.17853227

Bock, J., Krajewski, R., Moers, T., Runde, S., Vater, L., Eckstein, L., 2020. The ind dataset: a drone dataset of naturalistic road user trajectories at german intersections. In: 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 1929–1934.

Buisson, C., Saunier, N., Varotto, S.F., Wu, C., Zheng, Z., 2025. Reproducible research in transportation. Publication collection focused on reproducible practices in transportation research. https://link.springer.com/collections/dgachijbgi

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11621–11631.

Cai, H., Cai, X., Chang, J., Li, S., Yao, L., Changxin, W., Gao, Z., Wang, H., Yongge, L., Lin, M., et al., 2025. SciAssess: benchmarking LLM proficiency in scientific literature analysis. In: Findings of the Association for Computational Linguistics: NAACL 2025, pp. 2335–2357.

Caicedo, J.D., Guirado, C., González, M.C., Walker, J.L., 2025. Sharing, collaborating, and benchmarking to advance travel demand research: a demonstration of short-term ridership prediction. Transp. Policy 171, 531–541.

Chen, Y., Zhang, F., Sun, S., Pang, Y., Wang, Y., Song, J., Li, X., Hou, L., Zhao, S., Tang, J., Li, J., 2026. RPC-Bench: a fine-grained benchmark for research paper comprehension. arXiv:2601.14289.

Choe, T., Skabardonis, A., Varaiya, P., 2002. Freeway performance measurement system: operational analysis tool. Transp. Res. Rec. 1811 (1), 67–75.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Combs, T.S., Pardo, C.F., 2021. Shifting streets COVID-19 mobility data: findings from a global dataset and a research agenda for transport planning and policy. Transp. Res. Interdiscip. Perspect. 9, 100322.

Crampton, J.W., 2015. Collect it all: national security, big data and governance. GeoJournal 80 (4), 519–531.

Crüwell, S., van Doorn, J., Etz, A., Makel, M.C., Moshontz, H., Niebaum, J.C., Orben, A., Parsons, S., Schulte-Mecklenbeck, M., 2019. Seven easy steps to open science: an annotated reading list. Z. f'ur Psychol 227, 237–248.

Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N.A., Gardner, M., 2021. A dataset of information-seeking questions and answers anchored in research papers. arXiv:2105.03011.

van der, M.L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.

Du, Y., Tian, M., Ronanki, S., Rongali, S., Bodapati, S., Galstyan, A., Wells, A., Schwartz, R., Huerta, E.A., Peng, H., 2025. Context length alone hurts LLM performance despite perfect retrieval. arXiv:2510.05381.

Eenink, R., Barnard, Y., Baumann, M., Augros, X., Utesch, F., 2014. UDRIVE: the European naturalistic driving study. In: Proceedings of Transport Research Arena. IFSTTAR.

Elsevier, 2024. Journals data XML. https://supportcontent.elsevier.com/Support%20Hub/DaaS/Journals_Data_XML.pdf Data as a Service documentation. Accessed 17 Jun 2025.

Elsevier, 2025a. Research data guidelines. Accessed: 2025-12-23. https://www.elsevier.com/researcher/author/tools-and-resources/research-data/data-guidelines

Elsevier, 2025b. Text and data mining license. https://www.elsevier.com/about/policies-and-standards/text-and-data-mining/license Accessed: 2025-10-22.

Enwald, H., Grigas, V., Jurgita, R., Kortelainen, T., 2022. Data sharing practices in open access mode: a study of the willingness to share data in different disciplines. Inf. Res. 27 (2), 1–18.

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C.R., Zhou, Y., et al., 2021. Large scale interactive motion forecasting for autonomous driving: the Waymo open motion dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9710–9719.

Ferreira, Jr, J., Diao, M., Zhu, Y., Li, W., Jiang, S., 2010. Information infrastructure for research collaboration in land use, transportation, and environmental planning. Transp. Res. Rec. 2183 (1), 85–93.

Fink, L., Marcus, J., 2025. Replication code availability over time and across fields: evidence from the german socio-economic panel. Econ. Inq. 63 (2), 357–386.

Fišar, M., Greiner, B., Huber, C., Katok, E., Ozkes, A.I., Collaboration, M. S.R., 2024. Reproducibility in management science. Manage. Sci. 70 (3), 1343–1356.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378–382.

Folla, K., Anagnostopoulou, A., Yannis, G., 2021. Main challenges and opportunities, constraints and barriers of open science in transport research. Int. J. Res. Humanit. Soc. Stud. 8 (5), 32–40.

Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., et al., 2018. Science of science. Science 359 (6379), eaao0185.

Gellerman, H., Svanberg, E., Barnard, Y., 2016. Data sharing of transport research data. Transp. Res. Procedia 14, 2227–2236.

Gemini Team, 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. Technical Report. Google. Accessed: 2025-06-20. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf

Gentemann, C., 2023. Why NASA and federal agencies are declaring this the year of open science. Nature 613 (7943), 217.

Gilardi, F., Alizadeh, M., Kubli, M., 2023. ChatGPT outperforms crowd workers for text-annotation tasks. Proc. Natl. Acad. Sci. 120 (30), e2305016120.
Gloudemans, D., Wang, Y., Ji, J., Zachar, G., Barbour, W., Hall, E., Cebelak, M., Smith, L., Work, D.B., 2023. I-24 motion: an instrument for freeway traffic science. Transp. Res. C: Emerg. Technol. 155, 104311.
Goodman, S.N., Fanelli, D., Ioannidis, J. P.A., 2016. What does research reproducibility mean? Sci. Transl. Med. 8 (341), 341ps12.
Gozzi, M., Di Maio, F., 2024. Comparative analysis of prompt strategies for large language models: single-task vs. multitask prompts. Electronics 13 (23), 4712.
Guide for authors - Transportation Research Part C: Emerging Technologies. https://www.sciencedirect.com/journal/transportation-research-part-c-emerging-technologies/publish/guide-for-authors Accessed: 2025-06-16.
Gundersen, O.E., Cappelen, O., Mølnå, M., Nilsen, N.G., 2025. The unreasonable effectiveness of open science in AI: a replication study. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39, pp. 26211–26219.
Hahnel, M., Smith, G., Campbell, A., 2024. The state of open data 2024: special report bridging policy and practice in data sharing. https://digitalscience.figshare.com/articles/report/_b_The_State_of_Open_Data_2024_Special_Report_b_Bridging_policy_and_practice_in_data_sharing/27337476, https://doi.org/10.6084/m9.figshare.27337476.v2
Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Ginsburg, B., 2024. RULER: what's the real context size of your long-context language models? In: First Conference on Language Modeling. https://openreview.net/forum?id=kIoBbc76Sy
Huotala, A., Kuutila, M., Ralph, P., Mäntylä, M., 2024. The promise and challenges of using LLMs to accelerate the screening process of systematic reviews. In: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, pp. 262–271.
Hutson, M., 2018. Artificial intelligence faces reproducibility crisis. Science 359, 725–726.
INFORMS Journal on Computing, 2025. Welcome to the IJOC software and data repositories. Accessed 2025-12-24. https://informsjoc.github.io/
Janssen, M.A., Pritchard, C., Lee, A., 2020. On code sharing and model documentation of published individual and agent-based models. Environ. Model. Softw. 134, 104873.
Javaji, S.R., Cao, Y., Li, H., Yu, Y., Muralidhar, N., Zhu, Z., 2025. Can AI validate science? Benchmarking LLMs for accurate scientific claim → evidence reasoning. arXiv:2506.08235.
Ji, J., Lu, R., Belkessa, L., Wang, L., Varotto, S., Dong, Y., Saunier, N., Ameli, M., Macfarlane, G.S., Madadi, B., Wu, C., 2026. Supporting data for measuring the state of open science in transportation using large language models. https://doi.org/10.5281/zenodo.23045963
Jiang, C., Bhat, C.R., Lam, W. H.K., 2020. A bibliometric overview of transportation research Part B: methodological in the past forty years (1979–2019). Transp. Res. B: Methodol. 138, 268–291.
Jiang, R., Yin, D., Wang, Z., Wang, Y., Deng, J., Liu, H., Cai, Z., Deng, J., Song, X., Shibasaki, R., 2021. DL-Traff: survey and benchmark of deep learning models for urban traffic prediction. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 4515–4525.
Lai, Y., Zhao, H., 2025. Comparative analysis of smart city scientific research trends in the USA and china. Nat. Cities , 2, 1–9.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.
Lee, J., Lee, J., Yoo, J.-J., 2025. The role of large language models in the peer-review process: opportunities and challenges for medical journal reviewers and editors. J. Educ. Eval. Health Prof. 22 (4). https://doi.org/10.3352/jeehp.2025.22.4
Levin, M., Stern, R., 2024. EAGER: How Well Does Traffic Control Theory and Optimization Research Work on Public Roads? Technical Report Award No. 2437781. National Science Foundation. Grant awarded to University of Minnesota-Twin Cities, effective 1 September 2024; expires 31 August 2026. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2437781.
Li, X., Xu, H., Huang, X., Guo, C., Kang, Y., Ye, X., 2021. Emerging geo-data sources to reveal human mobility dynamics during COVID-19 pandemic: opportunities and challenges. Comput. Urban Sci. 1 (1), 22.
Lin, J., Wang, Y., Yu, Y., Zhou, Y., Chen, Y., Shi, X., 2022. Automatic analysis of available source code of top artificial intelligence conference papers. Int. J. Softw. Eng. Knowl. Eng. 32 (07), 947–970.
Liu, L., Jones, B.F., Uzzi, B., Wang, D., 2023. Data, measurement and empirical methods in the science of science. Nat. Hum. Behav. 7 (7), 1046–1058.
Liu, Y., Yang, Z., Xie, T., Ni, J., Gao, B., Li, Y., Tang, S., Ouyang, W., Cambria, E., Zhou, D., 2026. ResearchBench: benchmarking LLMs in scientific discovery via inspiration-based task decomposition. arXiv:2503.21248.
Ma, W., Alimo, P.K., Wang, L., Abdel-Aty, M., 2022. Mapping pedestrian safety studies between 2010 and 2021: a scientometric analysis. Accid. Anal. Prev. 174, 106744.
Mahajan, V., Kuehnel, N., Intzevidou, A., Cantelmo, G., Moeckel, R., Antoniou, C., 2022. Data to the people: a review of public and proprietary data for transport models. Transp. Rev. 42 (4), 415–440.
Mahmassani, H.S., 2025. 2024 IATBR lifetime achievement lecture - travel behavior research: science, empirics, models and applications. YouTube, TBD National Center. Event recording. Accessed: 2025-10-13. https://www.youtube.com/watch?v=XTd0ezEo9OM
Mahmoudi, H., Chang, D., Lee, H., Ghaffarzadegan, N., Jalali, M.S., 2025. Critical assessment of large language models' (ChatGPT) performance in data extraction for systematic reviews: exploratory study. JMIR AI 4, e68097. https://doi.org/10.2196/68097.
McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. 22 (3), 276–282.
Modak, N.M., Merigó, J.M., Weber, R., Manzor, F., de Dios Ortúzar, J., 2019. Fifty years of transportation research journals: a bibliometric overview. Transp. Res. A: Policy Pract. 120, 188–223.
National Academies of Sciences, Medicine and Policy, Global Affairs, Board on Research Data, Information and Division on Engineering and Physical Sciences, Committee on Applied and Theoretical Statistics, et al., 2019. Reproducibility and Replicability in Science. National Academies Press.
Newton, M.P., Bullock, D.M., Watkinson, C., Bracke, P.J., Horton, D.K., 2012. Engaging new partners in transportation research: integrating the publishing, archiving, and indexing of technical literature into the research process. Transp. Res. Rec. 2291 (1), 111–123.
Nie, M., 2025. A brief history of travel forecasting. Transportation , 53, 3447–3472.
Nielsen, A.F., Michelmann, J., Akac, A., Palts, K., Zilles, A., Anagnostopoulou, A., Langeland, O., 2023. Using the future wheel methodology to assess the impact of open science in the transport sector. Sci. Rep. 13 (1), 6000.
Nikander, J., Jama, T., Tenkanen, H., 2024. Threats related to open geospatial data in the uncertain geopolitical environment. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 48 (4/W12-2024), 121–126.
Olszewski, D., Lu, A., Stillman, C., Warren, K., Kitroser, C., Pascual, A., Ukirde, D., Butler, K., Traynor, P., 2023. "Get in researchers; we're measuring reproducibility": a reproducibility study of machine learning papers in tier 1 security conferences. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. Association for Computing Machinery, New York, NY, USA, p. 3433-3459. https://doi.org/10.1145/3576915.3623130
Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251), aac4716.
Parsons, S., Azevedo, F., Elsherif, M.M., Guay, S., Shahim, O.N., Govaart, G.H., Norris, E., O'mahony, A., Parker, A.J., Todorovic, A., et al., 2022. A community-sourced glossary of open scholarship terms. Nat. Hum. Behav. 6 (3), 312–318.
Peng, R.D., 2011. Reproducible research in computational science. Science 334 (6060), 1226–1227.
Pineau, J., 2018. Reproducible, reusable, and robust reinforcement learning. Invited talk, Advances in Neural Information Processing Systems (NeurIPS), Montréal, Canada.
Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., Larochelle, H., 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). J. Mach. Learn. Res. 22 (164), 1–20.
Punzo, V., Zheng, Z., Montanino, M., 2021. About calibration of car-following dynamics of automated and human-driven vehicles: methodology, guidelines and codes. Transp. Res. C: Emerg. Technol. 128, 103165.
Řehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp. 45–50.
REproducible research in transportation engineering (rerite) working group. https://www.rerite.org/. Accessed: 2026-01-05.
Riehl, K., Kouvelas, A., Makridis, M.A., 2025. Revisiting reproducibility in transportation simulation studies. Eur. Transp. Res. Rev. 17 (1), 22.

Roche, D.G., Raby, G.D., Norin, T., Ern, R., Scheuffele, H., Skeeles, M., Morgan, R., Andreassen, A.H., Clements, J.C., Louissaint, S., et al., 2022. Paths towards greater consensus building in experimental biology. J. Exp. Biol. 225 (Suppl_1), jeb243559.
Schicktanz, C., Klitzke, L., Gimm, K., Lüdtke, R., Liesner, K., Mosebach, H.H., Heuer, F., Wodtke, A., Asbach, L., 2025. The DLR highway traffic dataset (DLR-HT): longest road user trajectories on a German highway. Authorea Prepr.
Seibold, H., Czerny, S., Decke, S., Dieterle, R., Eder, T., Fohr, S., Hahn, N., Hartmann, R., Heindl, C., Kopper, P., et al., 2021. A computational reproducibility study of PLOS One articles featuring longitudinal data analyses. PLoS One 16 (6), e0251194.
Seo, T., Tago, Y., Shinkai, N., Nakanishi, M., Tanabe, J., Ushirogochi, D., Kanamori, S., Abe, A., Kodama, T., Yoshimura, S., et al., 2020. Evaluation of large-scale complete vehicle trajectories dataset on two kilometers highway segment for one hour duration: zen traffic data. In: 2020 International Symposium on Transportation Data and Modelling.
Serwadda, D., Ndebele, P., Grabowski, M.K., Bajunirwe, F., Wanyenze, R.K., 2018. Open data sharing and the global south–who benefits? Science 359 (6376), 642–643.
Shi, L., Jiang, B., Yuan, Z., Perez, M.A., Guo, F., 2025. SynSHRP2: a synthetic multimodal benchmark for driving safety-critical events derived from real-world driving data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4586–4596.
Siegel, Z.S., Kapoor, S., Nadgir, N., Stroebl, B., Narayanan, A., 2024. CORE-Bench: fostering the credibility of published research through a computational reproducibility agent benchmark. Trans. Mach. Learn. Res.
Son, G., Baek, S., Nam, S., Jeong, I., Kim, S., 2024. Multi-task inference: can large language models follow multiple instructions at once? In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5606–5627.
Stagge, J.H., Rosenberg, D.E., Abdallah, A.M., Akbar, H., Attallah, N.A., James, R., 2019. Assessing data availability and research reproducibility in hydrology and water resources. Sci. Data 6 (1), 1–12.
Stodden, V., 2014. The reproducible research movement in statistics. Stat. J. IAOS 30 (2), 91–93.
Sun, L., Rahwan, I., 2017. Coauthorship network in transportation research. Transp. Res. A: Policy Pract. 100, 135–151.
Sun, L., Yin, Y., 2017. Discovering themes and trends in transportation research using topic modeling. Transp. Res. C: Emerg. Technol. 77, 49–66.
Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al., 2020. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2446–2454.
Sun, Y., Kirtonia, S., 2020. Identifying regional characteristics of transportation research with transport research international documentation (TRID) data. Transp. Res. A: Policy Pract. 137, 111–130.
Tamakloe, R., Park, D., 2023. Discovering latent topics and trends in autonomous vehicle-related research: a structural topic modelling approach. Transp. Policy 139, 1–20.
Tamminga, G., Knoppers, P., Van Lint, J., 2014. Open traffic: a toolbox for traffic research. Procedia Comput. Sci. 32, 788–795.
Tamminga, G., Miska, M., Santos, E., Van Lint, H., Nakasone, A., Prendinger, H., Hoogendoorn, S., 2012. Design of open source framework for traffic and travel simulation. Transp. Res. Rec. 2291 (1), 44–52.
Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., Dorsett, K., 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PloS one 10 (8), e0134826.
Transportation Networks for Research Core Team, 2025. Transportation networks for research. https://github.com/bstabler/TransportationNetworks. Accessed: November 13, 2025.
U. S. Department of Transportation Federal Highway Administration, 2016. Next generation simulation (NGSIM) vehicle trajectories and supporting data. [dataset]. provided by ITS datahub through data.transportation.gov. https://doi.org/10.21949/1504477
UNESCO, 2022. Understanding open science. https://doi.org/10.54677/UTCD9302. Document code: SC-PBS-STIP/2022/OST/1, 6 pages.
Vandewalle, P., Kovacevic, J., Vetterli, M., 2009. Reproducible research in signal processing. IEEE Signal Process. Mag. 26 (3), 37–47.
Vicente-Saez, R., Martinez-Fuentes, C., 2018. Open science now: a systematic literature review for an integrated definition. J. Bus. Res. 88, 428–436.
Wandelt, S., Zhang, A., Antoniou, C., Budd, L., Castro-Nuno, M., Ding, Y., Eliasson, J., Goessling, S., He, S.Y., Ison, S., et al., 2025a. The Hitchhiker's guide to challenges in transport policy research: towards ANSWERing questions regarding life, mobility, and everything. Transp. Res. A: Policy Pract. 200, 104650.
Wandelt, S., Zheng, C., Sun, X., 2025b. A review of trends and patterns in transportation journals from 2000 to 2024. Transp. Res. Interdiscip. Perspect. 31, 101406.
Wang, L., Gjoreski, H., Ciliberto, M., Mekki, S., Valentin, S., Roggen, D., 2019. Enabling reproducible research in sensor-based transportation mode recognition with the sussex-huawei dataset. IEEE Access 7, 10870–10891.
Wang, S., Mo, B., Zheng, Y., Hess, S., Zhao, J., 2024. Comparing hundreds of machine learning and discrete choice models for travel demand modeling: an empirical benchmark. Transp. Res. B: Methodol. 190, 103061.
Wang, Y., Tang, P., Liu, K., Cai, J., Ren, R., Lin, J.J., Cai, H., Zhang, J., El-Gohary, N., Berges, M., et al., 2023. Characterizing data sharing in civil infrastructure engineering: current practice, future vision, barriers, and promotion strategies. J. Comput. Civ. Eng. 37 (2), 04023001.
Welch, T.F., Widita, A., 2019. Big data in public transportation: a review of sources and methods. Transp. Rev. 39 (6), 795–818.
Winker, M.A., Bloom, T., Onie, S., Tumwine, J., 2023. Equity, transparency, and accountability: open science for the 21st century. Lancet 402 (10409), 1206-1209. https://doi.org/10.1016/s0140-6736(23)01575-1
Woelfle, M., Olliaro, P., Todd, M.H., 2011. Open science is a research accelerator. Nat. Chem. 3 (10), 745–748.
Wood, J.S., van Schalkwyk, I., 2024. Reproducibility in transportation research: importance, best practices, and dealing with protected and sensitive data. J. Transp. Technol. 15 (1), 179–202.
Wu, C., Ghosh, B., Zheng, Z., Martínez, I., 2024a. Reproducibility in transportation research: A hands-on tutorial. https://www.rerite.org/itsc24-rr-tutorial/
Wu, C., Kreidieh, A.R., Parvate, K., Vinitsky, E., Bayen, A.M., 2021. Flow: a modular learning framework for mixed autonomy traffic. IEEE Trans. Robot. 38 (2), 1270–1286. https://doi.org/10.1109/TRO.2021.3087314
Wu, J., Sanchez-Diaz, I., Yang, Y., Qu, X., 2024b. Why is your paper rejected? Lessons learned from over 5000 rejected transportation papers. Commun. Transp. Res. 4, 100129.
Yabe, T., Luca, M., Tsubouchi, K., Lepri, B., Gonzalez, M.C., Moro, E., 2024a. Enhancing human mobility research with open and standardized datasets. Nat. Comput. Sci. 4 (7), 469–472.
Yabe, T., Tsubouchi, K., Shimizu, T., Sekimoto, Y., Sezaki, K., Moro, E., Pentland, A., 2024b. YJMob100K: city-scale and longitudinal dataset of anonymized human mobility trajectories. Sci. Data 11 (1), 397.
Yang, Y., Youyou, W., Uzzi, B., 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. Proc. Natl. Acad. Sci. 117 (20), 10762–10768.
Youyou, W., Yang, Y., Uzzi, B., 2023. A discipline-wide investigation of the replicability of psychology papers over the past two decades. Proc. Natl. Acad. Sci. 120 (6), e2208863120.
Zhang, C., 2024. China's privacy protection strategy and its geopolitical implications. Asian Rev. Polit. Econ. 3 (1), 6.
Zhao, M., Chen, S., Sun, Y., 2026. Uncovering thematic correlations across transportation research journal series: pitting human expertise against machine intelligence. Transp. Res. A: Policy Pract. 206, 104901.
Zheng, Z., 2021. Reasons, challenges, and some tools for doing reproducible transportation research. Commun. Transp. Res. 1, 100004. https://doi.org/10.1016/j.commtr.2021.100004
Zheng, Z., Levinson, D., Mohammadian, S., Ghosh, B., Martínez, I., Wu, C., 2025. Reproducibility in transportation research: a hands-on tutorial 2.0 - reproducible manuscripts. https://zuduozheng.github.io/RERITE_TUTORIAL_2.0. Accessed: 2026-01-05.
Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D., 2023. Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. arXiv:2302.10198
Zilinski, L.D., Scherer, Jr, D.A., Bullock, D.M., Horton, D., Matthews, C.E., 2014. Evolution of data creation, management, publication, and curation in the research process. Transp. Res. Rec. 2414 (1), 9–19.