Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut Data Science
Fachgebiet Data Science und Digital Libraries

An automatic reproducibility checking pipeline for the Learning Analytics Academic Community

Bachelorarbeit

im Studiengang Informatik

von

Domenik Kern

Prüfer: Prof. Dr. Sören Auer
Zweitprüfer: Dr. Mohammadreza Tavakoli
Betreuer: Dr. Gábor Kismihók

Hannover, 12.09.2025

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 12.09.2025

Domenik Kern

Zusammenfassung

Kurze Zusammenfassung der Arbeit in ca. 200 Wörtern

Abstract

Reproducibility in research is essential for ensuring its trustworthiness and transparency. When research cannot be reproduced, its findings cannot be fully trusted. Many studies in the past have failed to be reproduced, leading to what is commonly referred to as the reproducibility crisis. Over the years, various solutions have been proposed, including the adoption of open science practices to improve the situation. However, within the Learning Analytics community, there remains considerable room for improvement, as recent studies have shown.

To help ensure that research adheres to reproducibility guidelines and thereby directly improve its rigor, this thesis proposes an automatic reproducibility checking pipeline. The pipeline evaluates research papers using a reproducibility checklist, aided by the Large Language Model Gemma3. This approach also enabled an overview of the current state of the Learning Analytics & Knowledge Conference, allowing comparison with data from previous years. Additionally, the pipeline was evaluated through a human study to verify its reliability.

The results indicated a generally positive reception of the tool. However, it turned out to not be a perfect evaluator on its own, but rather a tool to provide a general overview of a research paper's reproducibility, which should then be followed by manual evaluation to ensure completeness.

Contents

Chapter 1

1 Introduction

Learning Analytics, a prominent subfield of educational technology, has experienced significant growth since its formal emergence in 2011. The field focuses on leveraging data to improve both teaching and learning processes through methods such as visualization, recommendations, and other data-driven innovations [22]. This growth is reflected in the increasing interest from the research community, as seen in the record-breaking number of full paper submissions at this year's Learning Analytics and Knowledge (LAK) conference. However, as the field matures, so too does the need for ensuring the trustworthiness, transparency, and scientific validity of its research outputs. All of these aspects are encompassed by reproducibility, a fundamental cornerstone of high-quality research.

1.1 Problem Statement

More than a decade ago, the term reproducibility crisis emerged in response to concerning findings across various scientific disciplines showing that many research results could not be reliably reproduced [9]. As a countermeasure, the Open Science movement proposed and promoted practices aimed at increasing transparency, accountability, and replicability in research [9]. The degree to which these practices have been adopted in the Learning Analytics and Knowledge (LAK) community was examined in a study from 2023, which analyzed papers from LAK'21 and LAK'22 [11]. The findings were disappointing, indicating that reproducibility and adherence to Open Science principles remain limited within the field [11]. Verifying reproducibility can serve as an important quality check and lead to methodological improvements, which in turn can support more reliable decision-making. Despite its importance, there is currently no automated way to assess reproducibility at scale, making it difficult for the community to identify and address shortcomings efficiently.

1.2 Research Objective

This thesis aims to explore how reproducibility in Learning Analytics research can be assessed more effectively using emerging technologies. Specifically, it investigates the potential of Large Language Models (LLMs) to support the automated evaluation of reproducibility in research. LLMs have shown promise in various text analysis tasks and may significantly reduce the time and effort needed to assess complex academic texts. To this end, I propose an automated reproducibility checking pipeline that utilizes an LLM to assess the extent to which a research paper addresses key aspects of open science and reproducibility. Additionally, the pipeline is applied to provide an updated overview of the current state of reproducibility, enabling a comparison to findings from previous years and potentially revealing progress over time.

1.3 Results of the Thesis

Using the pipeline, an overview of the current reproducibility state within the LAK conference proceedings was generated, revealing a generally positive trend. Similarly, the pipeline itself proved effective in producing a quick overview of research papers, which can significantly support manual evaluations. This potential was also reflected in a small-scale human evaluation of the tool, where it received positive feedback, including from members of the Learning Analytics research community.

1.4 Structure of the Thesis

The structure of this thesis is as follows: Chapter 2 provides the necessary background, introducing the concept of reproducibility, its perceived crisis in research, and the relevance of open science principles. It also presents the Large Language Model (LLM) used in the pipeline. Chapter 3 details the implementation of the pipeline, including the reproducibility checklist and the setup of the human evaluation study. Chapter 4 presents and interprets the results of both the human evaluation and the large-scale analysis of current LAK proceedings. Chapter 5 discusses the findings and outlines the limitations of the pipeline. Finally, Chapter 6 concludes the thesis and highlights directions for future work

Chapter 2

Background

This chapter introduces the key concepts underpinning the later proposed pipeline, including reproducibility, open science, and open access, as well as education technology and learning analytics. It also provides a brief overview of natural language processing and large language models, concluding with a description of the Gemma 3 model.

2.1 Reproducibility

Reproducibility is a cornerstone of scientific integrity. Although its definition varies across fields and remains somewhat unclear, its core attribute is the ability of independent researchers to draw the same conclusions from an experiment by following the original documentation [8]. It is closely related to but distinct from replication, which involves the ability to obtain consistent results across studies aimed at answering the same scientific question but using new data or other new computational methods [17]. Reproducibility ensures that research findings are reliable and not a result of undisclosed conditions.

In computational fields, reproducibility often focuses on sharing and annotating data and code, while in psychology and other empirical sciences, it may involve redoing experiments under similar conditions [9].

Strategies to enhance reproducibility include adopting Open Science practices, such as detailed documentation, version control, and preregistration. The Transparency and Openness Promotion (TOP) guidelines and tools like the Open Science Framework aim to address these challenges by promoting transparency and accountability in research[9].

The Open Science Handbook[4] provided a clear instruction on how to incorporate reproducibility into research:

- Preregister important study design and analysis information to increase transparency and counter publication bias of negative results

- Track changes to your files, especially your analysis code, using version control
- Document everything done by hand in a README file. Create a data dictionary to describe important information about your data
- Consider using Jupyter Notebooks, KnitR, Sweave, or other approaches to literate programming to integrate your code with your narrative and documentation
- Share your used Data and Materials
- Report and publish your methods and interventions explicitly and transparently and fully to allow replication

Improving reproducibility leads to increased rigor and quality of scientific outputs, and thus to greater trust in science [4].

Despite its importance, reproducibility faces significant challenges. For example, a study of 133 papers in learning analytics found that none could be reproduced within a 15-minute time-frame, primarily due to missing libraries, undocumented randomness, or platform-specific dependencies [11]. Issues related to code and software dependencies also pose substantial barriers [11]. A common reason for the failure of reproducibility is the lack of information about specific libraries or their version used in the research [11]. Without this information, replicating the computational environment can be challenging or impossible. The use of platform-dependent scripts, such as Bash scripts specific to certain operating systems, can further restrict reproducibility [11]. Inconsistent package version management in programming languages can also contribute to these issues [11]. Additionally non-deterministic randomness in analytical procedures, particularly in machine learning models, can also decrease reproducibility [11]. Without setting a fixed seed for random processes, it is impossible to obtain the exact same outcomes [11]. Another Challenge is the insufficient documentation of materials, datasets and analytical procedures [11]. Often, insufficient documentation makes it difficult to understand the data processing steps, the rationale behind analytical choices, or even to locate the necessary materials [11].

This lack of transparency, combined with inconsistent implementation of open science and reproducibility principles, has significantly hindered reproducibility efforts and contributed to what is now widely recognized as a reproducibility crisis in scientific research [9].

2.1.1 Reproducibility Crisis

Despite widespread recognition of the importance of reproducibility and the availability of strategies to support it, many scientific disciplines continue to struggle with consistently implementing these practices. This ongoing difficulty in replicating research findings has led to what is now commonly referred to as the *Reproducibility Crisis* [9]. A 2016 poll conducted by the journal Nature found that over half (52%) of surveyed scientists believed science was indeed facing a reproducibility crisis [2]. Others have a more optimistic view on the situation. For example, Vazire [26] instead refers to it as a *credibility revolution*. Spellman [23] calls it *Revolution 2.0*, a time where research, with the help of technology, improves by allowing full descriptions of methods, making data sets and analysis code available, preregistration of studies, and fostering a culture of transparency and accountability in scientific practices. The early creation of the Open Science Framework has provided an existence proof that these things can easily be implemented [23]. Again, underlining the importance of transparent and open research.

Despite these advances, concerns about reproducibility remain widespread. A recent 2025 survey of 452 professors across India and the USA reported that approximately 80% of Indian researchers and over 90% of U.S. researchers were familiar with the reproducibility crisis [5]. However, nearly 20% of respondents believed their peers remained completely unaware of the issue [5]. The survey also identified key barriers to reproducibility, with the majority of engineering researchers in both countries citing the unavailability of raw data as the primary obstacle, followed closely by the lack of access to code [5]. Other prominent challenges included selective reporting and publication pressure [5]. When it came to replication attempts, only 15% of Indian researchers who tried to reproduce others' work reported affirmative results, compared to 33% in the USA [5]. A common reason for unsuccessful reproduction was the lack of adequate methodological information provided in published studies [5]. However, better community practices—such as sharing code and maintaining detailed project pages on platforms like GitHub—have made reproducibility easier over time [5]. Proper documentation, along with accessible data and code, is now recognized as critical to successful reproduction [5]. While reproducing others' work was significantly more challenging in the past, the situation has improved considerably in step with the open science movement [5].Nevertheless, the adoption of open science principles remains inconsistent across research communities. For instance, Haim et al. [11] found that the Learning Analytics community still has substantial room for improvement in this area. Continued efforts are needed to incentivize data and code sharing, enhance methodological transparency, and more fully integrate open science practices into standard research workflows across disciplines.

2.2 Education Technology and Learning Analytics

Education Technology is a field focused on developing and evaluating tools and processes to enhance learning and teaching. It operates in a cyclical manner: Researchers create new technologies, educators and learners provide feedback, and researchers refine their innovations based on these data[11]. A prominent subfield within Education Technology is Learning Analytics, which involves collecting and analyzing data from learners to improve educational outcomes. It emerged as a distinct discipline with the establishment of the International Conference on Learning Analytics and Knowledge (LAK) in 2011. Over the years the field has grown and diversified and was recently redefined as follows:

> *Learning analytics is the collection, analysis, interpretation and communication of data about learners and their learning that provides theoretically relevant and actionable insights to enhance learning and teaching. (SoLAR2025)*

Learning Analytics is a human-centered, multidisciplinary field focused on the intersection of data and learning [22]. The LA community conducts theory-driven investigations into learning processes that are relevant to various stakeholders, such as students, teachers, learning designers and advisors [22]. It informs feedback and offers actionable insights. The goal is to improve learning, learners' well being and the quality of education [22].

Rooted in practical application, LA takes a holistic approach to data: encompassing data collection, analysis, communication, and feedback, all to provide insights back to stakeholders [22]. These insights inform actions such as implementing educational interventions. The field is guided by responsible, sustainable, and ethical data use [22]. By leveraging evidence-based and contextually-aware data from educational technologies, strategic and operational decisions can be made to support learners, teachers, and other educational stakeholders [22]. Alongside quantitative methods—such as statistical approaches, machine learning, and other AI techniques—qualitative methods, mixed approaches, and insights from the learning sciences are also employed to enhance learning [22].

Because research in Learning Analytics builds upon previous findings, it is crucial to trust the reliability and validity of existing work before using it as a foundation for new insights. This makes reproducibility and adherence to open science practices essential. Ensuring that research can be independently verified strengthens the rigor and transparency of the field.

2.3 Open Science

Open Science is a movement aimed at making scientific research, data, and dissemination accessible to all levels of society. It encompasses principles such as transparency, re-use, participation, cooperation, accountability, and reproducibility, with the goal of improving the quality and reliability of research[4]. Open Science practices include open access to publications, data-sharing, open notebooks, transparent research evaluation, reproducible research, and open-source software [4] and many more as seen in the Open Science Taxonomy by Pontika et al.2.1.

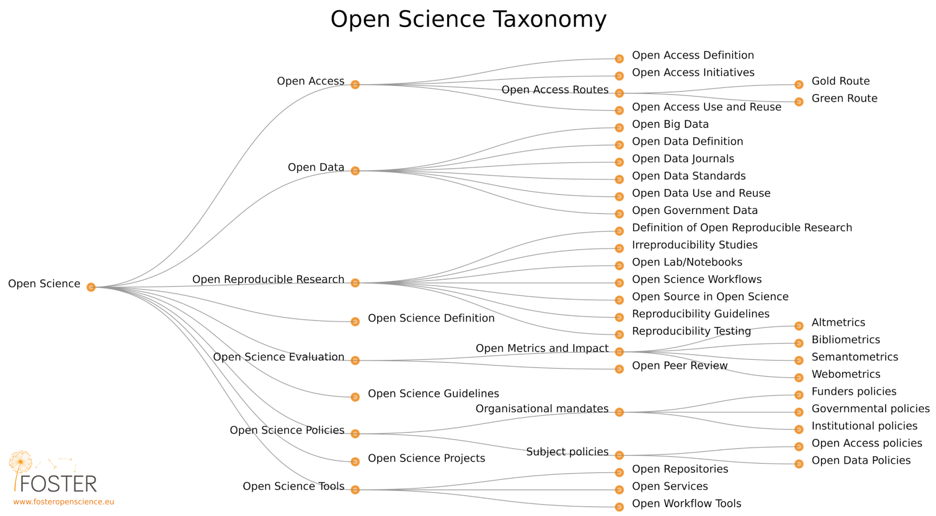

Figure 2.1: The Open Science Taxonomy, Pontika et al. [19].

The adoption of Open Science varies across disciplines. For instance, in learning analytics, less than half of the research papers examined at the 13th International Conference on Learning Analytics and Knowledge adhered to standard Open Science principles, with only about 5% fully meeting these criteria [11]. Common barriers include a lack of knowledge among researchers and logistical challenges in making datasets, code, and materials openly available [11]. Despite these challenges, Open Science is seen as a catalyst for improving research rigor and fostering cumulative scientific progress.

As stated by Haim et al. [11], key components of Open Science include:

- Open Methodology: Details of the methods and evaluation used by the authors, including but not limited to the setup, logic flow, aggregations of results, etc.

- Open Data: Data that can be freely accessed, reused, and redistributed, often following the FAIR principles.

- Open Materials: Sharing research tools, software, and other resources to enable replication and collaboration.

- Preregistration: Documenting research plans in advance to reduce bias and enhance transparency.

The Open Science movement also addresses cultural and systemic challenges, such as aligning scientific norms with values like communality and disinterestedness, as opposed to secrecy and self-interest [9]. By promoting these practices, Open Science aims to create a more inclusive, equitable, and trustworthy scientific ecosystem.

2.3.1 Open Access

Open Access (OA) refers to the free, online availability of scholarly literature with minimal copyright and licensing restrictions. It is a response to the *access revolution* made possible by digital technology and the internet, allowing authors to share perfect copies of their work globally at little to no cost[24].

Traditionally, most authors sold their work to earn royalties [24]. However in contrast scholars write research articles not for profit but for impact [24]. These authors are typically supported by institutions or salaries and benefit professionally when their work is widely read and cited [24]. For them, open access aligns better with their goals than commercial publishing, which can limit readership and impact[24].

Two major types of barriers are removed through Open Access:

- Price Barriers: OA literature is free to read, unlike other content that requires payment or subscription. This broadens access, especially for readers without institutional support[24]
- Permission Barriers: OA allows for uses beyond reading, such as redistributing, translating, data mining, and reformatting, all with proper attribution to the author.[24]

The BBB Definition, from the Budapest Open Access Initiative [13], Bethesda Statement on Open Access Publishing [12], and Berlin Declaration on Open Access [21] statements form the foundation of OA. These declarations emphasize that OA should eliminate both financial and legal barriers to allow wide reuse of scholarly work, while still requiring attribution to authors, all made possible by the Internet. The goal of OA is to enable the free flow of research and scholarly information, improve its usefulness, and accelerate innovation and discovery [24].

2.3.2 FAIR Principles

The FAIR Principles: Findable, Accessible, Interoperable, and Reusable were introduced to support long-term stewardship, management, and meaningful reuse of scientific data in a digital age. Developed by a coalition of scholars, publishers, and funding bodies, the FAIR framework ensures that both humans and machines can discover, access, and use research outputs effectively [27]. Importantly, FAIR is a framework, not a rigid standard. It allows incremental adoption and domain-specific adaptation, making it practical across diverse scientific disciplines[27].

Findable: Data and associated metadata must be findable. This includes assigning a globally unique and persistent identifier (e.g., DOI), enriching metadata to describe the data accurately, and ensuring that both are indexed in searchable repositories.

Accessible: Once located, data should be retrievable using standardized, open protocols. Even when access to the data itself is restricted, the metadata should remain openly available to ensure transparency. This allows controlled access where necessary, without undermining the broader principle of openness.

Interoperable: Data and metadata should use standardized formats and shared vocabularies that allow integration with other datasets and tools. Interoperability ensures that data can be combined and analyzed across platforms and disciplines, enhancing cross-disciplinary research and reproducibility.

Reusable: Ultimately, FAIR aims at optimizing the reuse of data. Metadata must be richly described with clear provenance and usage licenses. They should be based on domain-specific standards, allowing others to replicate, adapt and build on the research.

FAIR and Open Science: The FAIR Principles are not synonymous with Open Access, but they are essential to realizing the goals of Open Science. While Open Access removes price and permission barriers, FAIR ensures that openly shared research is also structured, discoverable, and reusable. FAIRness is increasingly being mandated by funding agencies and institutions as a prerequisite for good data stewardship [27].

2.4 Natural Language Processing

Natural Language Processing (NLP) is a subfield of Artificial Intelligence that focuses on enabling machines to interpret, generate, and interact using human language [15]. Because of its ability to process and understand large volumes of unstructured text, NLP has enabled a broad range of practical applications. Some of the most frequent applications are text classification, machine translation, information retrieval, and question answering , but NLP also encompasses a wide range of other tasks [16].

2.4.1 Large Language Model

A Large Language Model (LLM) is a type of AI system designed to understand and generate human language. Built upon deep learning architectures, particularly the Transformer [25], LLMs represent one of the most advanced developments in NLP. They are trained on massive text corpora to learn linguistic patterns, grammar, facts, and reasoning abilities. These models are capable of performing a wide range of NLP tasks, such as machine translation, text summarization, question answering, and sentiment analysis. Some of the currently best-performing models are Grok-3-Preview-02-24, GPT-4.5-Preview, Gemini-2.0-Flash-Thinking-Exp-01-21, and DeepSeek-R1 [10].

LLMs operate within a defined context window, which limits how much text the model can consider at one time. The model's input, known as a prompt, guides its output. Prompt engineering involves crafting inputs to elicit desired behaviors or responses from the model [20].

The temperature parameter controls the randomness of the model's output. Lower values (e.g., 0) make responses more deterministic and focused, while higher values (e.g., 0.8) encourage diversity and creativity [18]. This allows users to fine-tune how conservative or imaginative the model's output should be, depending on the task at hand. However, it should be noted that setting the temperature to 0 does not guarantee a fully deterministic output [18]. In the context of reproducibility, the temperature should be put to 0 to make sure the output is as reproducible as possible.

When used together, these methods allow for controlled generation of outputs targeted at specific applications. Increasingly, such capabilities are being leveraged in academic research workflows, where LLMs can support or automate tasks traditionally performed by human researchers.

2.4.2 Use-cases of LLM in research

Keya et al. [14] leveraged the use of LLMs for scientific abstract summarization in 2025. In their study, each abstract was condensed into a single sentence by an LLM and subsequently evaluated both automatically and by human reviewers. They compared multiple models across various prompting techniques and analyzed performance across several scientific domains, including computer science. Their key findings highlight, first, that automated evaluation results aligned closely with human assessments, and second, that the choice of prompting technique plays a significant role in optimizing LLM performance, with different models benefiting from different approaches.

In another study from Bermejo et al. [3], the effectiveness of LLMs in extracting complex information from textual data was evaluated . In their work, both human coders and LLMs were tasked with performing increasingly complex information retrieval tasks using newspaper articles that required substantial contextual understanding [3]. Across all tasks, the LLMs consistently outperformed human participants, even when dealing with lengthy and challenging texts, indicating their superior performance

[3]. These results highlight the significant advancements in NLP technology and demonstrate that sophisticated textual analyses can now be readily implemented by researchers, even without extensive prior expertise in the field [3].

2.5 Gemma 3

Gemma 3 is the latest lightweight open language model developed by Google DeepMind[10]. One of its key features is the support for long context lengths of up to 128K tokens, enabling it to process multi-page documents or lengthy articles in a single pass [10]. The model incorporates advanced filtering techniques to reduce the risk of generating unwanted or unsafe content and to remove personal or sensitive information from its output [10].

Despite its relatively modest size, the instruction-tuned Gemma 3-27B-IT variant ranks among the top 10 models in the Chatbot Arena benchmark [6], outperforming significantly larger models such as DeepSeek-V3 and Qwen2.5-70B [10].

Designed for efficiency, the model runs on standard consumer-grade hardware, including phones, laptops, and high-end GPUs, while maintaining strong performance across a range of tasks [10]. It is also openly available: users can interact with it via an API (with free and unlimited access, provided token limits are respected) or run it locally using platforms like Hugging Face or Ollama.

In this thesis, Gemma 3 is used in the pipeline to automatically evaluate the reproducibility of research papers, selected for its strong instruction-following and long-context capabilities, as well as its open-access nature and freely available API, which make it both practical and accessible for academic use.

Chapter 3

Implementation

This chapter outlines the approach used for the reproducibility checking pipeline. The goal is to automatically assess to what extent a research paper satisfies a predefined checklist of reproducibility criteria. This is achieved through automated text extraction and analysis in Python and supported by the large language model Gemma 3. The pipeline processes academic papers in PDF format, extracts relevant sections, and evaluates them against a reproducibility checklist, producing structured outputs. The abstract workflow can be seen in Figure 3.1.

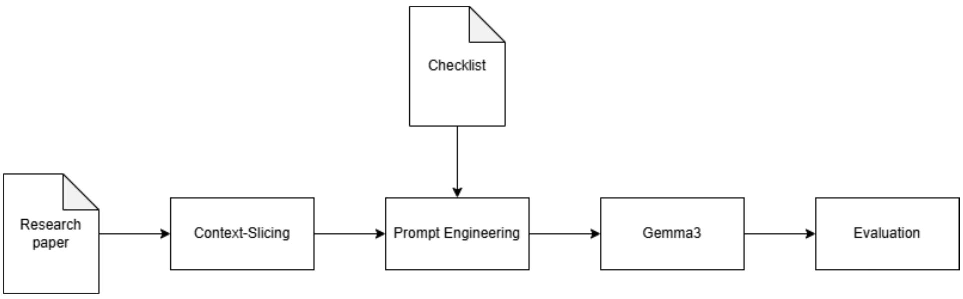

Figure 3.1: The Pipeline Workflow

3.1 Reproducibility Checklist

The core of the pipeline relies on a reproducibility checklist designed to capture key elements necessary for a research paper to be considered reproducible. The checklist includes categories such as detailed methodology description, availability of data, access to code, clarity of results and preregistration. These categories were selected based on open science and reproducibility criteria mentioned earlier and reflect essential components.

The checklist is briefly summarized below in Table 3.1 and can be found in Appendix A:

Category	Summary of Key Criteria
Open Methodology & Documentation	Clear explanation of methodology and a step-by-step description of experiment.
Data Accessibility & Transparency	Datasets openly available, well-documented, and collected through transparent methods.
Code & Software Availability	Code is shared in a public repository and includes scripts for data preparation and analysis.
Type of Analysis	Quantitative or qualitative methods are described clearly, including models, assumptions, and coding practices.
Results & Interpretation	Transparent reporting of outcomes and discussion of limitations and potential biases.
Preregistration	Evaluates whether the study was preregistered.

Table 3.1: Summary of the reproducibility checklist categories.

3.1.1 Category: Open Methodology & Documentation

This criterion addresses the transparency and detailed description of the research process. Clear and comprehensive documentation of the methodology is essential, as it enables others to replicate the study accurately. Without such documentation, reproduction becomes challenging or impossible.

3.1.2 Category: Data Accessibility & Transparency

This criterion focuses on the openness, documentation, and collection methods of the data used. Reproducibility requires applying the same methods to identical data, making data accessibility and understanding critical. To enhance transparency, the data collection process should be explicitly described.

3.1.3 Category: Code & Software Availability

To replicate a study, access to the code and software used is necessary. These should be shared through a publicly accessible repository, such as GitHub, and include code for data preparation, such as preprocessing, to ensure consistency across multiple runs.

3.1.4 Category: Type of Analysis

The research should provide clear descriptions of the analysis type, whether quantitative or qualitative. For quantitative methods, all statistical tests, methods, and assumptions must be explicitly stated, along with effect sizes, confidence intervals, and p-values. This allows for comparison and validation of results upon reproduction. For qualitative methods, the subjective context and interpretations of participants and data should be detailed [7]. Additionally, data coding and analysis processes should be shared to enhance transparency and reproducibility in qualitative research [7].

3.1.5 Category: Results & Interpretation

Results must be thoroughly documented to enable comparison and validation after reproduction. To increase transparency, limitations and potential biases in the results should be discussed. This minimizes subjective interpretations and strengthens the reliability of the findings.

3.1.6 Category: Preregistration

While preregistration may seem unnecessary if methods, data, code, and results are well-documented, it significantly enhances transparency and mitigates publication bias, particularly for negative results. As highlighted in the Open Science Handbook, preregistration helps address the reproducibility crisis and promotes open science practices. Therefore, studies should be preregistered, with a link provided to the preregistration (e.g., on the Open Science Framework).

3.2 Python and Packages

All code was written in Python 3.13 and is available in the following GitHub repository: `https://github.com/domkka/bachelorthesis/tree/main/pythonprototype`

To install all required packages, the following command should be executed:

pip install -r requirements.txt

The **tkinter** package was used to implement the user interface, which can be seen in Figure 3.2. There are four buttons with following functions: **Select PDF File**; **Load Checklist**; **Change API Key**; **Run Evaluation**. When the **Select PDF File** button is pressed, the function `load_pdf()` is called. A file dialog window opens, allowing the user to choose a research paper in PDF format. The selected document is then processed using the **pypdf** package. Text extraction and sectioning are performed by the function `extract_sections_using_bookmarks()`, which searches the

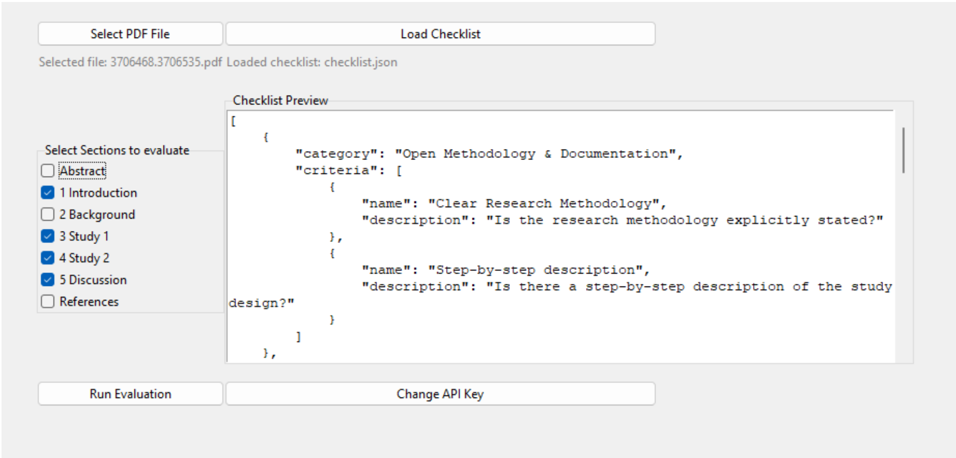

Figure 3.2: Reproducibility Checking Pipeline GUI

complete text for headings corresponding to PDF bookmarks and splits the document accordingly, storing each section separately.

After loading the paper, a new interface frame appears, listing all detected sections. Users can choose specific sections for evaluation by selecting the corresponding checkboxes. Prior to evaluation, the text is filtered to include only the selected sections, thereby reducing the token length and ensuring that the language model focuses on the most relevant content. For example, uninformative parts such as the *References* section can be excluded if not selected, as they do not contribute meaningfully to checklist-based reproducibility assessment.

By clicking the **Load Checklist** button, the `load_checklist()` function is triggered. This function allows the user to select a JSON file, ideally the `checklist.json`; however, the checklist can also be modified as long as the same structure is maintained. Providing a correctly formatted checklist is essential to ensure that the evaluation pipeline functions as intended and produces valid output. When the checklist is loaded, it is also previewed in a separate frame within the interface.

To make requests to the Google Gemini API, an API key is required. A key can be generated at: `https://aistudio.google.com/apikey`. To integrate the key with the pipeline, it must either be saved as an environment variable named `GEMINI_API_KEY`, or entered manually into the corresponding field in the user interface after clicking the **Change API Key** Button.

After loading the checklist and selecting the desired sections from the PDF, the evaluation is performed by clicking the **Start Evaluation** button, which triggers the `run_evaluation()` function. In this step, the selected sections and the checklist are combined into a prompt, which is then sent to the Gemma 3 model via an API request in the `generate()` function . The **google-genai** package is used to handle these API calls. Upon receiving a successful response, the output must be parsed back into valid JSON, since it often includes backticks and formatting prefixes despite explicit instructions not to. Finally, the completed checklist is saved in JSON format and displayed in the preview frame.

With the help of the **auto-py-to-exe** tool, the python code was compiled to an executable, making it easily shared to other users.

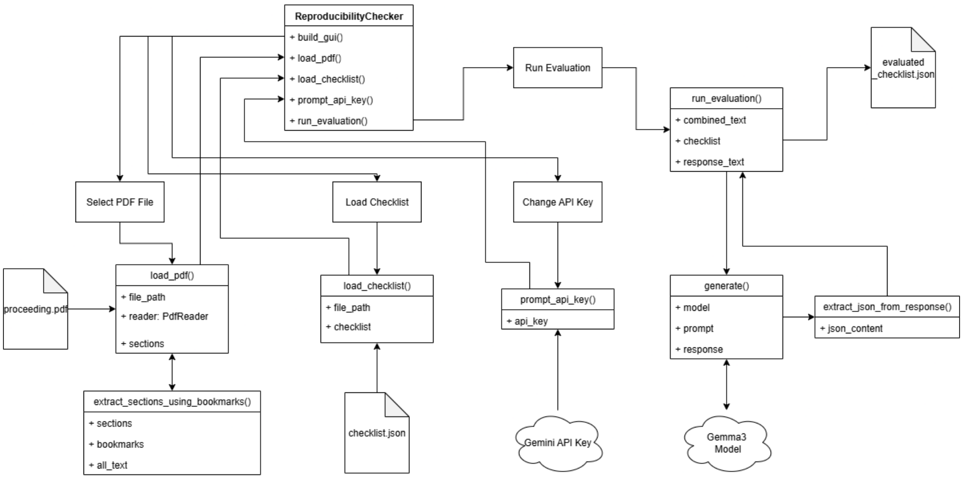

Figure 3.3: Detailed view of the Pipeline's Workflow

3.3 Checklist Verification with Language Models

The text extracted from each paper was analyzed using natural language processing techniques to automatically identify whether the checklist criteria were met or not. For this task, a large language model (LLM), the open-source Gemma 3 model, was used via its free API calls. For Prompting, a zero-shot-approach was taken, which includes the evaluation task, the sliced research paper and the reproducibility checklist. Also it was instructed to only fill out the checklist and return it as a JSON String. After refining the prompt to comply with the instructions, the final version is presented below:

```
Evaluate the following document with respect to the checklist criteria below:

Document:
"""{pdf_text}"""

Checklist:
{checklist}

For each checklist item, provide:
```

```
- ‘criterion‘: the name of the checklist item
- ‘status‘: one of "Met", "Not Met"
- ‘justification‘: a short explanation of where or why the criterion is (not) met
(e.g., section title, paragraph context, or quote)

Output format:
Return only valid JSON, structured like this:
[
    {
        "category": "Open Methodology & Documentation",
        "results": [
            {
                "criterion": "...",
                "status": "Met" | "Not Met",
                "justification": "..."
            }
        ]
    },
    ...
]
Do not include any explanation or commentary outside the JSON.
```

3.3.1 Possible Problems

Despite using a deterministic temperature setting of 0 for all API calls, there could be some reproducibility issues when trying to evaluate the same paper again. In some rare special cases, the model gave papers neither the status "Met" nor "Not Met", but rather "Partially Met" or "N/A" even though it was not prompted to do so. While the language model provides short explanations for each classification, the underlying decision-making process remains largely opaque due to the inherent nature of LLMs. However, to increase transparency, the model also returns the specific sections or text snippets from which the relevant information was extracted, or provides links that support the assessment. This helps to ground each evaluation in the original context and provides traceability for further manual verification.

3.4 Human Evaluation

The Pipeline is evaluated through a user study. Each participant is assigned two randomly selected papers from the LAK'25. Participants include experienced researchers from the Learning Analytics academic community as well as less experienced researchers currently in their last semester in Computer Science.

Each paper is evaluated twice: first manually using the provided Reproducibility Checklist, and then automatically using the Prototype Pipeline. After both evaluations, participants complete a survey regarding the usefulness and completeness of the Checklist as well as the accuracy and usability of the Pipeline. This process enables a thorough assessment of the Checklist's coverage and the effectiveness of the Pipeline, while also helping to identify potential areas for improvement.

By directly comparing the manually completed Checklists with those automatically generated by the Pipeline, we can evaluate the accuracy of the language model.

The survey includes Likert scale questions and open-ended questions to gather quantitative and qualitative feedback. The study process and the survey can be found in Appendix B and C.

3.5 Scoring for Proceedings Overview

The LLM's output for each paper is parsed as a JSON object making it easily readable and usable for further processing steps. Each paper is assessed according to the checklist and each category is rated as either "Met" or "Not Met". To achieve an overall status of "reproducible," a paper must fully meet all criteria.

The rationale for this strict scoring system is that the absence of a single component could compromise the ability of other researchers to reproduce a study's findings. Thus, one missing criterion is sufficient for a paper to be classified as not reproducible. This scoring scheme enforces a high standard and ensures that only papers meeting all fundamental reproducibility conditions are acknowledged as such.

Preregistration is treated as a bonus category and does not influence the overall reproducibility rating. While preregistration is not strictly necessary for a study to be reproducible, it plays a critical role in addressing publication bias and enhancing the credibility of research by making a prior hypotheses and methods transparent.

The scoring was performed using a separate Python script, `jsonscorer.py`, which processes the generated JSON files and applies the defined evaluation rules. Although not part of the main pipeline, this script is an essential post-processing step in the reproducibility assessment for both the Study and the overview of the reproducibility state of LAK'25.

Chapter 4

Results & Interpretation

This chapter presents the results of the study conducted to evaluate the reproducibility checking pipeline. First, the outcome of the human evaluation is described and compared to the pipelines. Subsequently, the findings are interpreted and discussed with respect to the pipeline's effectiveness, while also highlighting potential areas for improvement. Finally, the results of the evaluation of all proceedings from the 2025 Learning Analytics & Knowledge Conference are presented to provide an overview of the current state of reproducibility in the field, as well as to identify potential improvements and insights for future work.

4.1 Study results

The study included a total of seven participants, four experienced researchers from the Learning Analytics academic community and three less experienced researchers currently in their last semester. Consequently, 14 papers were evaluated, where each one was assessed twice: once manually with the Reproducibility Checklist and once using the pipeline prototype. The evaluation results are presented and compared in the following figures where each Bar represents the Manually or Prototype evaluated papers. Finally, the Likert-scale responses from the survey, as well as the answers to the open-ended question, are also presented.

4.1.1 Reproducibility Scores

Figure 4.1 presents the distribution of reproducibility scores across both evaluation methods. In the manual evaluations, scores ranged from 1 to 5: one paper received a score of 1, one a 2, six papers were rated 3, three received a 4, and three were rated 5. In contrast, the pipeline evaluation produced scores ranging only from 3 to 5, with eight papers receiving a 3, three a 4, and three a 5.

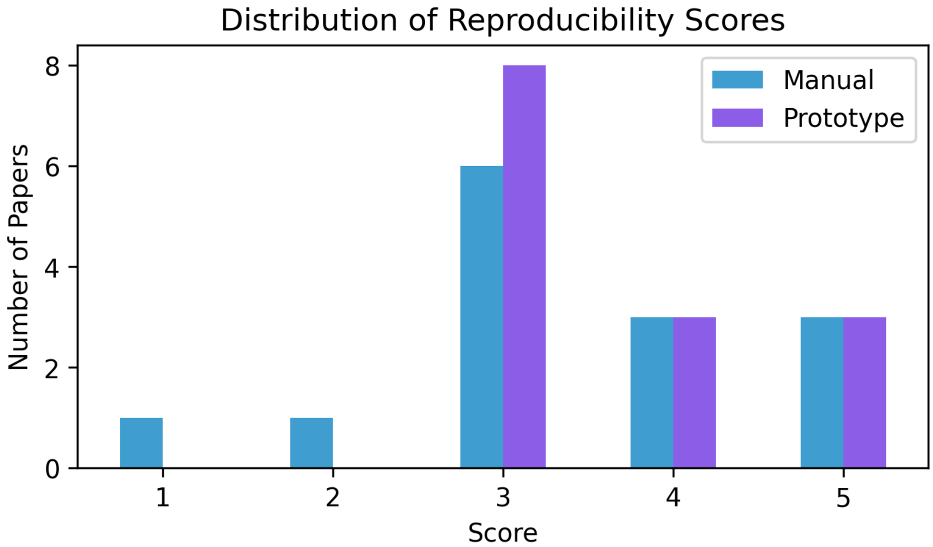

Figure 4.1: Distribution of Reproducibility Scores Manual vs Prototype Evaluation

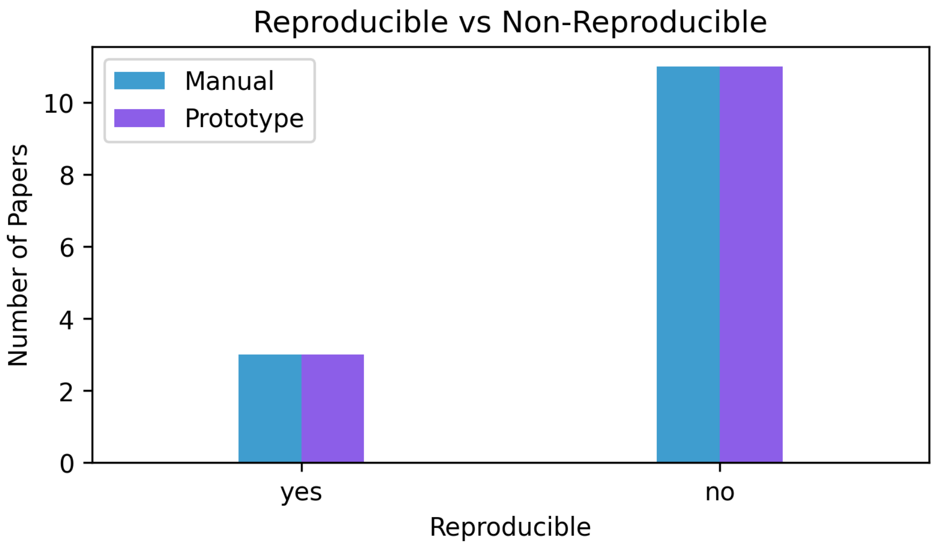

Figure 4.2: Number of papers deemed Reproducible or Not

Based on these scores, Figure 4.2 shows the absolute number of papers classified as reproducible or not. In this context, only papers that received the maximum score of 5 were considered reproducible. According to this criterion, three out of the 14 papers were deemed reproducible in both evaluations.

4.1.2 Reproducibility Criteria

To provide a more detailed view, the next figures break down which specific reproducibility criteria were met or not.

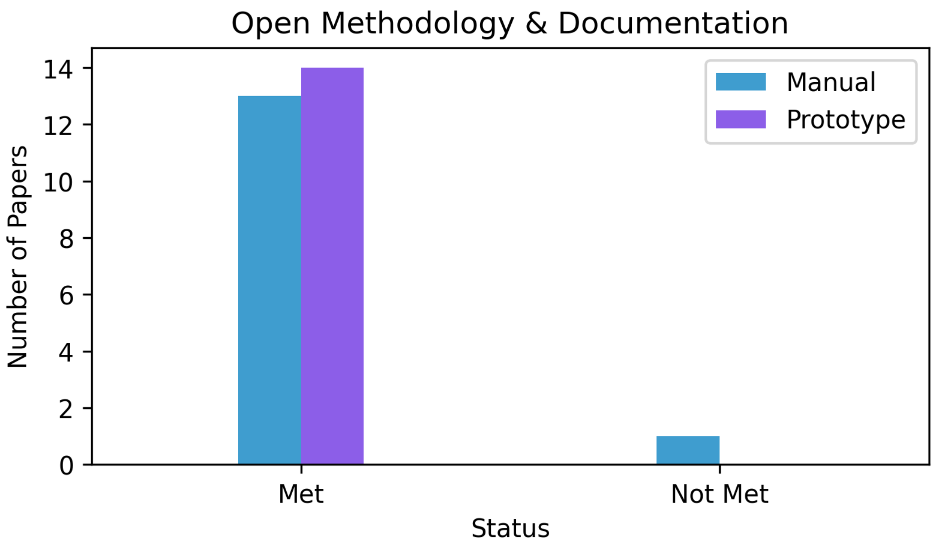

Figure 4.3: Number of papers meeting the Open Methodology & Documentation criterion

Figure 4.3 shows the number of papers that met the Open Methodology & Documentation criterion from the checklist. In the manual evaluation, only one paper did not meet this criterion. This was justified by the observation that the methodological steps were only briefly described and lacked a detailed explanation. In contrast, the prototype deemed the same paper compliant. As shown in Figure 4.4, the results for the Data Accessibility & Transparency criterion reveal a larger discrepancy between the evaluation methods. In the manual evaluation, 7 papers were deemed to share data, whereas the prototype identified only 5 as containing shared data. In one instance, the prototype even indicated where the data could be found but still did not assign a “Met” status. In another case, the data was presented within a table but was neither externally published nor linked for the prototype to find.

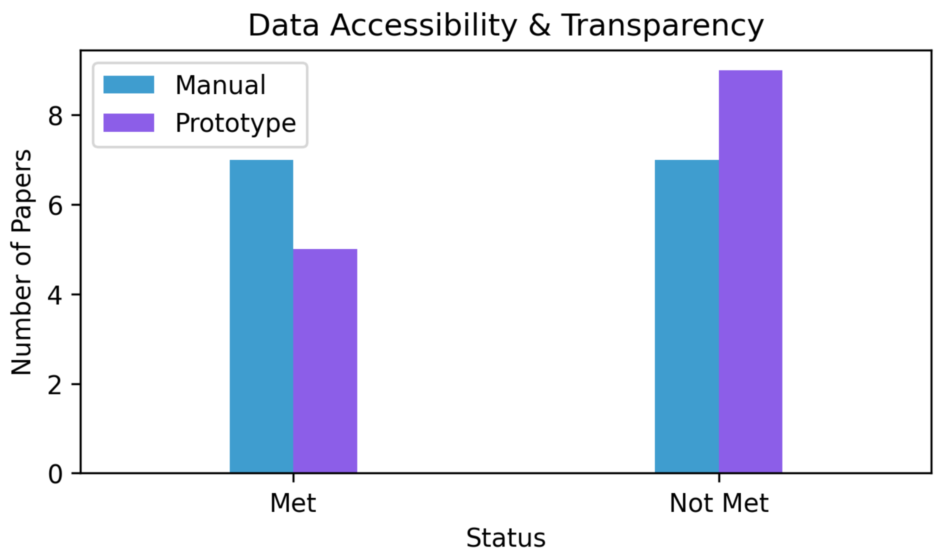

Figure 4.4: Number of papers meeting the Data Accessibility & Transparency criterion

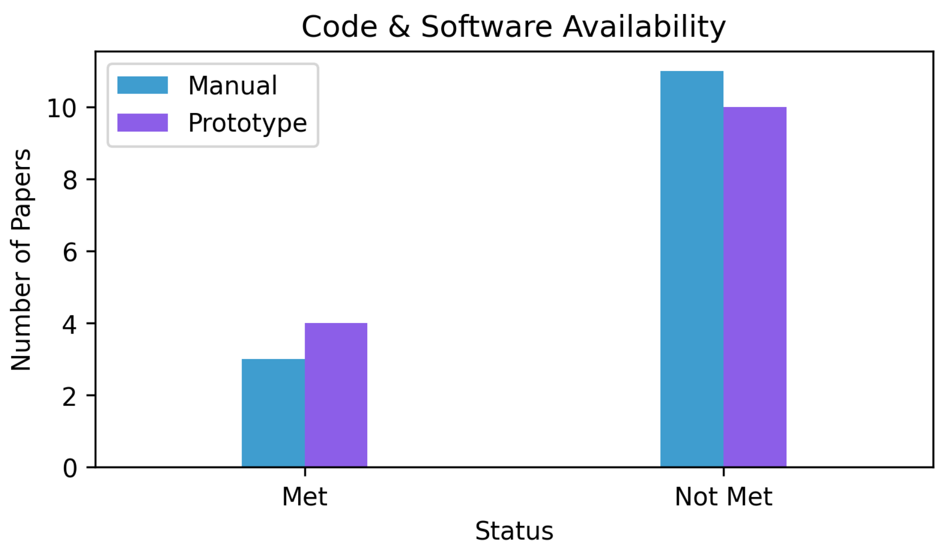

Figure 4.5: Number of papers meeting the Code & Software Availability criterion

For the Code & Software Availability criterion, Figure 4.5 shows that three papers met the criterion in the manual evaluation. In contrast, four papers were classified as meeting the criterion in the pipeline-based evaluation. For one paper flagged by the pipeline as meeting the criterion, a GitHub link was provided; however, upon manual inspection, no code was found and only a PowerPoint slide deck was available. The number of papers

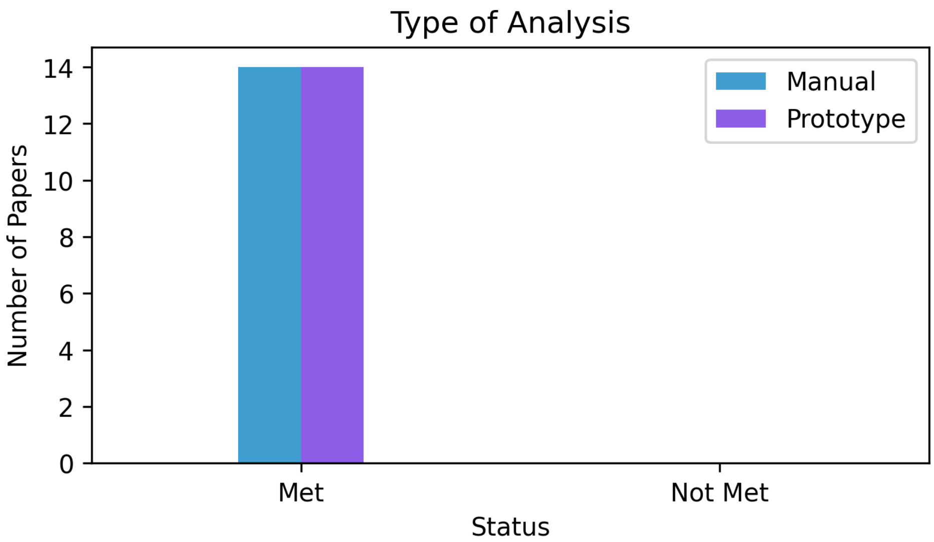

Figure 4.6: Number of papers meeting the Type of Analysis criterion

that meet the Type of Analysis criterion is shown in Figure 4.6. Here, the 14 papers met this criterion in both evaluation methods. Figure 4.7 shows that, in the manual evaluation, three out of 14 papers did not meet the Results & Interpretation criterion. This was again justified by the fact that the interpretation sections in those papers were not deeply enough discussed. In contrast, the prototype did not identify any issues, and consequently, all 14 papers were marked as meeting the criterion.

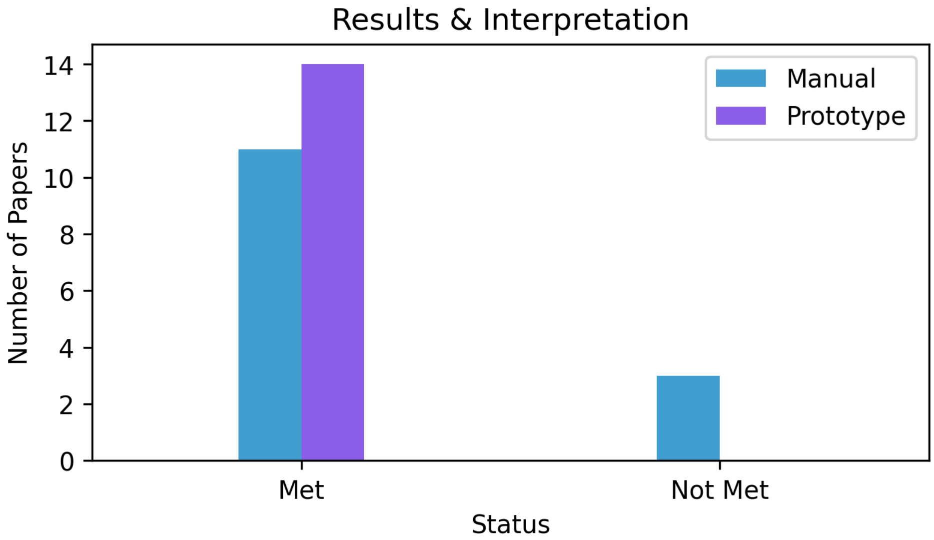

Figure 4.7: Number of papers meeting the Results & Interpretation criterion

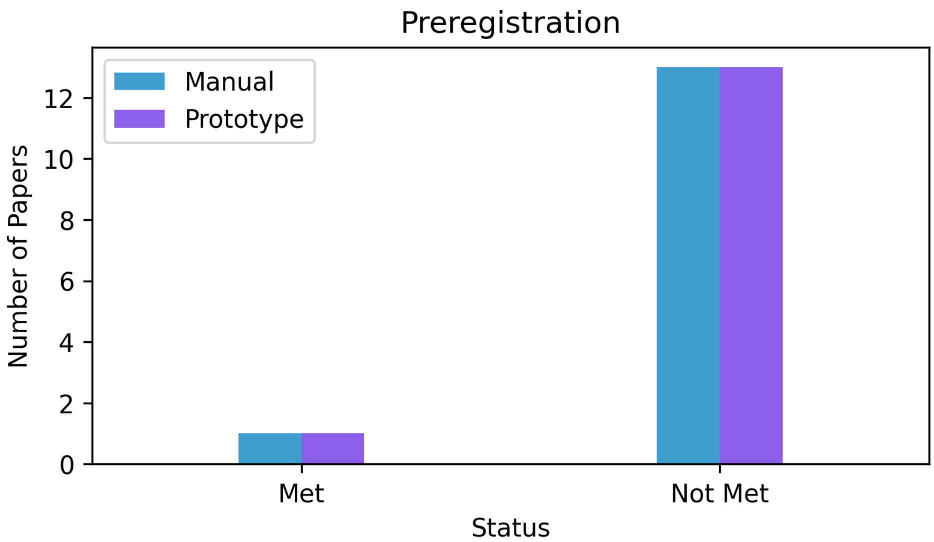

Figure 4.8: Number of papers meeting the Preregistration criterion

Lastly, the Preregistration criterion was evaluated. As shown in Figure 4.8, only one paper included a link to a preregistration, and this was identified by both evaluation methods.

4.1.3 Survey Results

The survey consisted of four questions related to the Reproducibility Checklist, five questions focused on the pipeline prototype, and three open-ended questions for additional feedback.

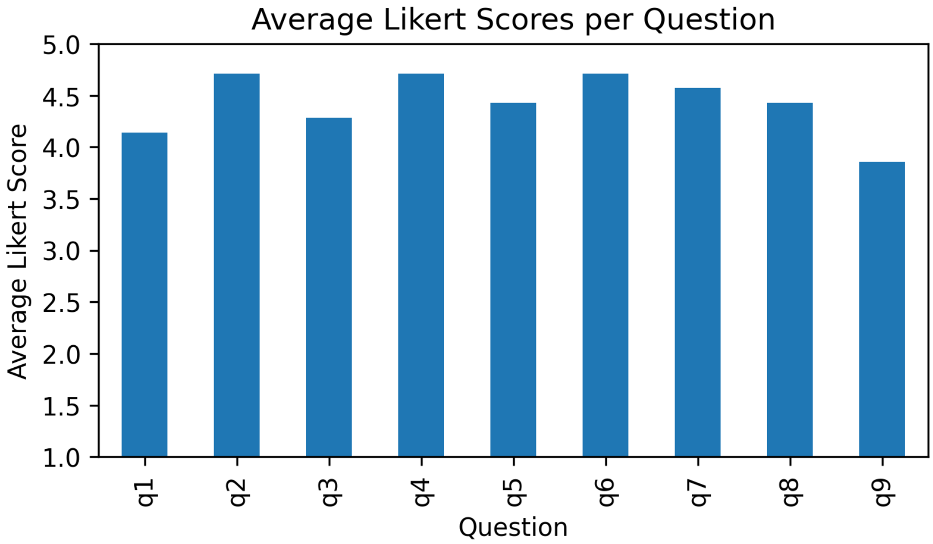

Figure 4.9: Average Likert Scores per Question

Figure 4.9 shows the average Likert scores for each question on a 5-point scale. Overall, the responses indicate a positive perception of both the checklist and the pipeline prototype. For instance, Question 2 ("The checklist captures key elements of reproducibility") and Question 4 ("The checklist helped identify reproducibility aspects") both received an average score of 4.71. Similarly, Question 6 ("The prototype has a clear use case") was also rated highly. Questions related to the pipeline's usability and usefulness received consistently strong ratings as well. The lowest-rated item was Question 9 ("I trust the pipeline's output when assessing reproducibility"), with an average score of 3.8, reflecting that the pipeline's outputs did not always align with the manual evaluations. The open-ended responses highlighted that the prototype improved the workflow by offering a fast, automated assessment process. Participants noted that it saved time by reducing the need to read entire papers in detail and supported the creation of a mental framework when assessing research.

4.2 Pipeline Results for LAK'25

To provide an overview of the reproducibility state at LAK'25, all available proceedings were evaluated using the reproducibility checking pipeline. The proceedings are open access and can be found here [1]. One proceeding was not compatible with the pipeline, resulting in a total of 69 evaluated papers. The following figures present the evaluation results.

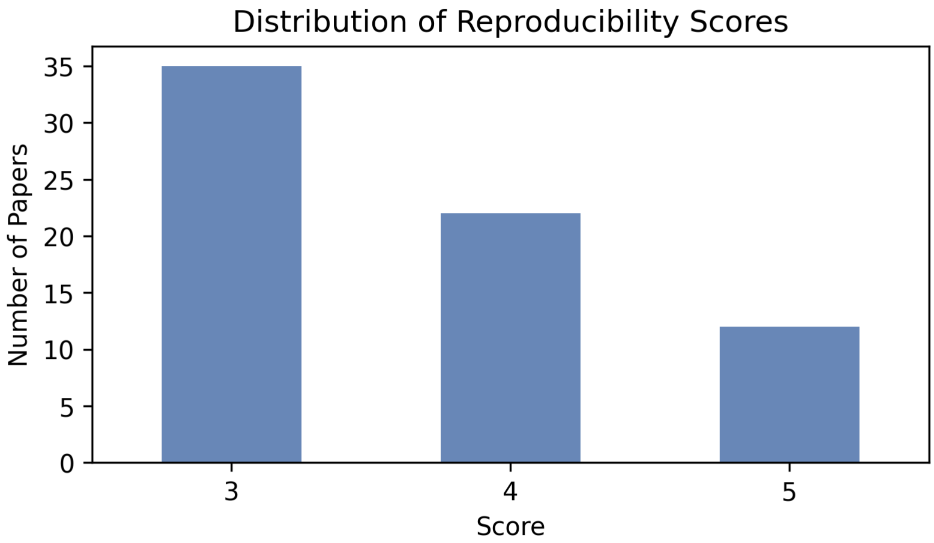

Figure 4.10: Distribution of Reproducibility Scores

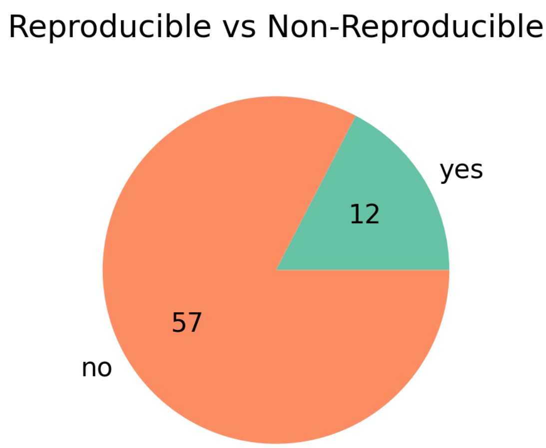

Figure 4.11: Reproducible vs Non-Reproducible

Figure 4.10 shows the distribution of reproducibility scores across the evaluated proceedings. Of the 69 papers, 35 received a score of 3, 22 received a score of 4, and the remaining 12 received the maximum score of 5. Based on this scoring, 12 papers are considered reproducible, as illustrated in Figure 4.11.

Figure 4.12 shows the compliance levels by criterion. All 69 papers received a "Met" status for the following criteria: Open Methodology & Documentation, Type of Analysis, and Results & Interpretation. For the criterion Data Accessibility & Transparency, 23 papers provided open data. It should be noted that, of the 46 articles that did not include open data, 16 were evaluated as having a justified reason for this: privacy concerns or reliance on unpublished datasets. Similarly, 23 proceedings met the Code & Software Availability criterion. However, for those that did not share their code, no justification was provided. Lastly, only one paper was found to be preregistered.

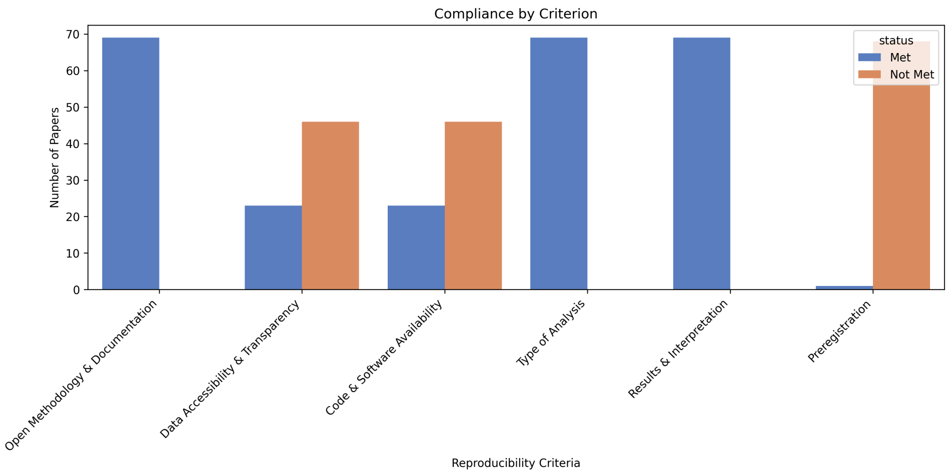

Figure 4.12: Compliance by Criterion

4.2.1 Comparison to Open Science Principle and Reproducibility Review of LAK'21/22

When comparing the results to the Reproducibility Review of LAK'21/22 [11], similar patterns emerge two years later. All papers met the Open Methodology criterion, which aligns with expectations given the structured submission requirements for LAK papers. However, it is worth noting that in the manual evaluation, some papers were marked as not meeting this criterion. This suggests that, while most papers describe the methodology to some extent, there is still room for improvement in terms of depth and clarity.

In terms of Open Data, there appears to be a positive development. Whereas approximately 25% of papers in LAK'21/22 shared open data, this share has now increased to around one-third. Taking into account that 16 out of the 69 evaluated papers cited valid reasons for not sharing data, such as privacy concerns, the proportion of papers to share and actually doing so approaches 50%. This upward trend may be linked to the updated LAK guidelines, under which all proceedings are now Open Access, unlike earlier classifications that included Available, Open Access, and Public Access.

Similarly to Open Data, Open Materials/Code also appears to have improved. While only approximately 13% of papers openly shared their code in 2021 and 2022, this number has increased to around one-third in the current evaluation.

Regarding preregistration, only one paper was found to be preregistered, same as it was in the findings from LAK'21/22. This indicates that preregistration remains a low priority among researchers in this field.

In this review, 17% of papers were evaluated as reproducible based on full compliance with all specified criteria. In comparison, only 5% of papers in 2021 and 2022 fully met the open science principles defined by Haim et al. [11].

Their review also included attempts to actually reproduce the research within a 15-minute time frame, which was unsuccessful for all papers. This highlights an important distinction: even if a paper is evaluated as reproducible based on documentation and transparency criteria, it may not be practically reproducible in a real-world attempt. Nonetheless, the findings suggest a positive development in open science practices and a gradual improvement in reproducibility within the LAK community.

Chapter 5

Discussion and Limitations

The results of this study highlight both the strengths and limitations of the automated reproducibility checking pipeline. One of the key limitations lies in the pipeline's differentiation of superficial and detailed content. For example, in cases where methodological or interpretative descriptions were relatively vague or lacking in depth, the pipeline always marked these criteria as "Met," while human evaluators did not. This indicates that the pipeline may currently lack the sensitivity to assess qualitative depth in academic writing.

Additionally, there were instances where the pipeline failed to detect data that had been shared, possibly due to format, placement, or phrasing within the paper. This limitation suggests that the pipeline may require further optimization to accurately identify and interpret links or references to supplementary material. It is also worth noting that the current implementation uses the Gemma 3 model, which is a less costly but still capable large language model. Replacing it with an even more powerful model could improve performance by enabling larger context windows and more in-depth textual analysis, potentially enhancing the accuracy of the evaluation.

When examining the reproducibility checklist results more closely, three criteria from it were consistently marked as "Met" by the pipeline across all papers. On one hand, this could indicate that the papers were genuinely compliant with these aspects. On the other hand, it may suggest that the criteria of the reproducibility checklist and the way they are phrased, lack sufficient depth to allow meaningful differentiation from the LLM. This raises the possibility that these criteria should be revised or divided into more granular sub-questions. Doing so could help the language model better distinguish between superficial and thorough chapters.

Currently the pipeline has some compatibility issues regarding its execution on different operating systems. Since it was developed in Python on a Windows operating system and compiled into an executable using the **auto-py-to-exe** tool, it cannot be executed on non-Windows systems. This restricts cross-platform usability and may limit adoption by users on macOS or Linux. Furthermore, the prototype currently supports only structured PDF files in order to reduce the input context size. This limits its ability to process unstructured or other types of documents, potentially excluding a subset of papers from evaluation. One proceeding could not be evaluated at all, likely due to formatting issues or structural incompatibility, which prevented the pipeline from generating any output. Especially for papers that do not follow the same structure as the LAK proceedings, splitting chapters using regex searches can fail due to conflicts. For example, when trying to evaluate this thesis with the pipeline, it would not work because bookmarks are already defined in the contents section. Therefore, an alternative approach is needed to extract the chapters. One method was tested and can be found in the GitHub repository under `/bachelorthesis/pythonprototype/altchecker`, where chapters are extracted using page numbers instead of regex searches.

Due to rate limitations imposed by the Gemini API, only a limited number of papers can be processed per minute, which can significantly increase evaluation time when analyzing multiple papers in parallel. In terms of output format, one participant noted that the current JSON structure may hinder readability. This suggests that converting the output into a more user-friendly format, such as a structured text report or table, could improve accessibility.

Despite these issues, the pipeline proved highly valuable in creating a structured overview of each paper. This capability significantly accelerates the evaluation process, especially when dealing with a large number of publications. In many cases, the pipeline's assessments were entirely accurate, providing a strong foundation for reproducibility screening. While the pipeline is not yet a full replacement for manual review, it demonstrates considerable potential as a time-saving support tool. What requires several hours of manual analysis can now be performed within minutes, making it a useful aid in large-scale reproducibility evaluations.

The overall reproducibility state of the Learning Analytics & Knowledge Conference appears to have improved. Although this study did not attempt to directly reproduce the research findings, improvements were observed in adherence to open science principles. Particularly, more code and data were shared compared to previous years, indicating progress in transparency. However, a persistent barrier remains: the use of personal or sensitive data, which often cannot be made publicly available. As a result, the reproducibility of qualitative research in particular continues to be constrained within the field.

The sample size for this study may be somewhat limited, particularly considering that three of the seven participants were less experienced researchers. This could have affected their ability to accurately assess qualitative depth in academic writing, potentially leading to an overestimation of the pipeline's performance.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Trust in research is foundational to scientific progress, and one of the most effective ways to build this trust is by ensuring that research is reproducible. However, the academic community continues to face a reproducibility crisis, where many published studies cannot be reliably reproduced. A key strategy to address this challenge lies in the adoption and enforcement of open science principles.

To contribute to this effort, this thesis introduced an automated reproducibility checking pipeline tailored to the Learning Analytics research community. The pipeline evaluates key aspects of reproducibility by analyzing published papers using a Large Language Model. It was applied to the latest LAK proceedings to assess the current state of reproducibility in the field and revealed an improvement in open science practices compared to prior years, particularly in the sharing of code and data. Nevertheless, persistent challenges remain, such as restrictions on sensitive datasets and the limited uptake of preregistration.

The study also identified key limitations of the pipeline: reduced sensitivity to qualitative depth in text, occasional failures in detecting shared materials, and technical constraints such as operating system compatibility, which ironically restricts its reproducibility. Despite this, it proved to be a useful tool for generating quick and informative overviews of research papers, reducing the time for evaluation. With further refinement and more powerful models, it has the potential to support researchers in conducting even faster and more consistent reproducibility assessments at scale.

6.2 Future Work

Future research could explore whether using more powerful Large Language Models would enhance the accuracy and nuance of evaluations, particularly for qualitative aspects such as methodology and interpretation. Alternatively, revising and expanding the current checklist into more granular items may enhance the depth and reliability of the pipeline's assessments. Additionally, to further validate the effectiveness of the pipeline, it would be valuable to attempt reproducing a selection of high-scoring papers. This could help determine whether the pipeline's output truly reflects the actual reproducibility of the research.

Appendix A

Reproducibility Checklist

Below the full Reproducibility Checklist can be found

Reproducibility Checklist DOI:		Met
Open Methodology & Documentation		
Clear Research Methodology	Is the research methodology explicitly stated?	☐
Step-by-step description	Is there a step-by-step description of the study design?	☐
Data Accessibility & Transparency		
Open Data	Are the raw or processed datasets publicly available or has it been mentioned where the data can be found?	☐
	If data cannot be shared, is there a clear justification (e.g., privacy concerns)?	☐
Data Documentation	Is there a data dictionary or codebook explaining variables, formats, and preprocessing steps?	☐
Data Collection Methods	Are data collection methods fully described?	☐
Code & Software Availability		
Open Source Code	Is the analysis code (Python, R, etc.) shared in an open repository (GitHub, GitLab) or somewhere else publicly available?	☐
	If code cannot be shared, is there a clear justification?	☐

Code for Data preparation	Is code for data preparation (e.g. pre-processing) shared/mentioned?	☐
Type of Analysis		
Quantitative Research Methods	Are statistical tests, models, and assumptions clearly described?	☐
	Are effect sizes, confidence intervals, and p-values reported?	☐
Qualitative Research Methods	Is subjective context and interpretation given?	☐
	Is data coding and analysis clearly described and shared?	☐
Results & Interpretation		
Clearly presented Results	Have the results been clearly documented?	☐
Limitations and potential biases	Are limitations and potential biases discussed?	☐
Preregistration		
Preregistration	Does the Paper contain a link to the preregistration (e.g the Open Science Foundation)?	☐

Appendix B

Study process

Below the full Study process can be found

Study Process for my Bachelorthesis on
“An Automatic Reproducibility Checking Pipeline for the Learning Analytics Academic Community”

You are given two randomly selected proceedings from the Learning Analytics & Knowledge Conference 2025, a Checklist (“Reproducibility Checklist.pdf”) and a Prototype (“reproducibilitychecker.exe”).

For each of the two proceedings fill out the Checklist:

- Write DOI at the top.
- Mark question as Met if the requirement is fulfilled.
- give a brief justification where in the paper the information is found (e.g, section, supplementary link, etc.)

After doing this for both proceedings, use the prototype to automatically evaluate the proceedings.

The prototype works as follows:

- Launch “reproducibilitychecker.exe” (Note: It might be flagged as an unrecognized app - click “More Info” -> “Run anyway”)
- Click “Select PDF File” and choose the proceeding
- Select the sections most likely to contain relevant information (Note: if everything is selected it might exceed the Token limit, e.g References should not be selected)
- Click “Load Checklist” and select the “checklist.json”
- Click “Change API Key” and copy-and-paste the API-Key found in “super-confidential-secret.txt“
- Click “Run Evaluation”

The automatically evaluated checklist can then be found in the preview window or in the directory: generatedjson/{doi}.json

Once you have evaluated both papers, please complete the small survey (Survey.pdf) and send me the filled checklists and survey to domenik.kern@stud.uni-hannover.de

Appendix C

Survey

Below the full Survey can be found

SURVEY

Thank you for participating in the study for my Bachelorthesis "An Automatic Reproducibility Checking Pipeline for the Learning Analytics Academic Community" . Your feedback is important as I continuously strive to improve my work. Please take a few minutes to complete this Survey.

PERSONAL INFORMATION:

Name: ☐

Email Address: ☐

Note: The information collected will be used solely for the purpose of this survey and will be kept confidential.

CHECKLIST EVALUATION:

Please rate the following aspects of the checklist on a scale of 1 to 5, where 1 stands for 'Strongly Disagree' and 5 stands for 'Strongly Agree'.

	1	2	3	4	5
The checklist consists of clear defined questions .	☐	☐	☐	☐	☐
The checklist captures key elements of reproducibility.	☐	☐	☐	☐	☐
The checklist is scalable to allow for more detailed assessment	☐	☐	☐	☐	☐
The checklist helped identify reproducibility aspects.	☐	☐	☐	☐	☐

PROTOTYPE EVALUATION:

Please rate the following aspects of the prototype on a scale of 1 to 5, where 1 stands for 'Strongly Disagree' and 5 stands for 'Strongly Agree'.

	1	2	3	4	5
The prototype is easy to use.	☐	☐	☐	☐	☐
The prototype has a clear use-case.	☐	☐	☐	☐	☐
The prototype can help identify reproducibility.	☐	☐	☐	☐	☐
The feedback was helpful when reading and evaluating the paper	☐	☐	☐	☐	☐
I trust the pipelines’s output when assessing reproducibility.	☐	☐	☐	☐	☐

OPEN FEEDBACK:

Are there any important aspects of reproducibility you think the checklist missed?

In what ways did the prototype support or simplify your work?

Do you have any other comments or suggestions regarding the checklist and/or prototype?

Thank you for taking the time to complete the study and survey. Your feedback is invaluable in helping me enhance the pipeline. If you have any further comments or concerns, please feel free to reach out to me at:

domenik.kern@stud.uni-hannover.de

I appreciate your participation.

Sincerely,
Domenik Kern

Bibliography

[1] *LAK '25: Proceedings of the 15th International Learning Analytics and Knowledge Conference*, New York, NY, USA, 2025. Association for Computing Machinery.

[2] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 2016.

[3] V. J. Bermejo, A. Gago, R. H. Gálvez, and N. Harari. Llms outperform outsourced human coders on complex textual analysis, November 13 2024. Available at SSRN: `https://ssrn.com/abstract=5020034` or `http://dx.doi.org/10.2139/ssrn.5020034`.

[4] S. Bezjak et al. *Open Science Training Handbook*. Zenodo, 2018.

[5] T. Chakravorti, S. Koneru, and S. Rajtmajer. Reproducibility and replicability in research: What 452 professors think in universities across the usa and india. *PLOS ONE*, 20(3):1–19, 03 2025.

[6] W.-L. Chiang et al. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[7] N. L. Cole et al. Reproducibility and replicability of qualitative research: an integrative review of concepts, barriers and enablers, Dec 2024.

[8] G. O. Erik. The fundamenta principles of reproducibility. *Royal Society*, 2021.

[9] F. Fidler and J. Wilcox. Reproducibility of scientific results. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

[10] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[11] A. Haim, S. Shaw, and N. Heffernan. How to open science: A principle and reproducibility review of the learning analytics and knowledge conference. In *LAK2023: LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 156–164, 2023.

[12] R. JLIS.it. Bethesda statement on open access publishing (2003). *JLIS.it*, 3(2), Dec. 2012.

[13] R. JLIS.it. Budapest open access initiative (2002). *JLIS.it*, 2012.

[14] F. Keya, M. Y. Jaradeh, and S. Auer. *Leveraging LLMs for Scientific Abstract Summarization: Unearthing the Essence of Research in a Single Sentence*. Association for Computing Machinery, New York, NY, USA, 2025.

[15] E. D. Liddy. Natural language processing. *Encyclopedia of Library and Information Science*, 2001.

[16] A. Montejo-Ráez et al. Current approaches and applications in natural language processing. *Applied Sciences*, 12(10):4859, 2022.

[17] E. National Academies of Sciences and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC, 2019.

[18] S. Ouyang et al. An empirical study of the non-determinism of chatgpt in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–28, 2025.

[19] N. Pontika, P. Knoth, M. Cancellieri, and S. Pearce. Fostering open science to research using a taxonomy and an elearning portal. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business*, i-KNOW '15. Association for Computing Machinery, 2015.

[20] P. Sahoo et al. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[21] M. P. Society. Berlin declaration on open access to knowledge in the sciences and humanities. `https://openaccess.mpg.de/Berlin-Declaration`, 2003. Accessed: 29.06.2025.

[22] Society for Learning Analytics Research. What is learning analytics? `https://www.solaresearch.org/about/what-is-learning-analytics/`. Accessed: 09.06.2025.

[23] B. A. Spellman. A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, 10(6):886–899, 2015. PMID: 26581743.

[24] P. Suber. *Open Access*. The MIT Press, 2012.

[25] A. Vaswani et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[26] S. Vazire. Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13(4):411–417, 2018. PMID: 29961410.

[27] M. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Sci Data*, 3(160018), 2016.