Plan Overview

A Data Management Plan created using DMPonline

Title: Identify students’ common errors in machine learning through a Language Model Tutor

Creator: Boyun Zhang

Affiliation: Delft University of Technology

Template: TU Delft Data Management Plan template (2025)

Project abstract:

The study explores the effectiveness of small language models in assisting with machine learning education. 40 volunteers with a basic background in linear regression and using language model will be recruited, primarily from CSE Bachelor students following the course CSE1210. Firstly, each participant will complete a pre-survey to identify their attitude to AI, and a test which contains some questions about linear regression concepts. Then they will be provided with a carefully designed linear regression assignment. They are expected to use a small language model as a tutor to solve the assignment problems. All interactions between the participants and the language model which include the questions posed and the responses received, will be automatically recorded for analysis. After completing the assignment, participants need to complete a questionnaire to assess their satisfaction with the system’s performance, and a test about linear regression concept to check whether their understanding has deepened after using the system.

ID: 173511

Start date: 15-01-2025

End date: 30-08-2025

Last modified: 02-05-2025

Identify students’ common errors in machine learning through a Language Model Tutor

0. Adminstrative questions

1. Provide the name of the data management support staff consulted during the preparation of this plan and the date of consultation. Please also mention if you consulted any other support staff.

Richard Grimes, Data Steward at the Faculty of Electrical Engineering, Mathematics, and Computer Science, has reviewed this DMP on 30th April 2025.

2. Is TU Delft the lead institution for this project?

- Yes, the only institution involved

I. Data/code description and collection or re-use

3. Provide a general description of the types of data/code you will be working with, including any re-used data/code.

Type of data/code	**File format(s)**	**How will data/code be collected/generated?** *For re-used data/code: what are the sources and terms of use?*	**Purpose of processing**	**Storage location**	**Who will have access to the data/code?**
Research Code	Mostly .py files	Re-used code from Manuel Valle Torre (https://github.com/mvallet91/JELAI). The research code is hosted on a TU Delft facuty managed server.	Provide user with a system that allow them to interact with language model and work in jupyter environment.	Github	the TUD project team
Answers to A linear regression assignment	Jupyter files	Participants finish the assignment and submit via the research code.	To grade assignment.	TUD OneDrive	the TUD project team
Anonymised data on questions sent to language models, and the responses generated	.csv files	Collected by python package Jupyterlab Pioneer.	To identify knowledge gaps of users and assess the quality of responses	TUD OneDrive	the TUD project team
Quantitative questionnaires	Excel files	Questionnaires filled digitally by using MS Forms.	To measure overall satisfaction, and perceived helpfulness	TUD OneDrive	the TUD project team
Knowledge Test	PDF files	Test filled digitally by using MS Forms.	To measure students' understanding about linear regression	TUD OneDrive	the TUD project team

II. Storage and backup during the research process

4. How much data/code storage will you require during the project lifetime?

- < 250 GB

5. Where will the data/code be stored and backed-up during the project lifetime? (Select all that apply.)

- TU Delft OneDrive

III. Data/code documentation

6. What documentation will accompany data/code? (Select all that apply.)

- Software – Usage documentation (README file, docstrings, and in-line comments)
- Procedure – A description of data processing procedure(s) (such as laboratory setup, simulation workflows).
- Data – README file or other documentation explaining how data are organised
- Data – Methodology of data collection

IV. Legal and ethical requirements, code of conducts

7. Does your research involve human subjects or third-party datasets collected from human participants?

If you are working with a human subject(s), you will need to obtain the HREC approval for your research project.

- Yes – please provide details in the additional information box below

I intend to apply for ethical approval from the Human Research Ethics Committee, but have not yet done so.

8. Will you work with personal data? (This is information about an identified or identifiable natural person, either for research or project administration purposes.)

- No

9. Will you work with any other types of confidential or classified data or code as listed below? (Select all that apply and provide additional details below.)

If you are not sure which option to select, ask your Faculty Data Steward for advice.

- No, I will not work with any other types of confidential or classified data/code

10. How will ownership of the data and intellectual property rights to the data be managed?

For projects involving commercially-sensitive research or research involving third parties, seek advice of your Faculty Contract Manager when answering this question.

This is an internal TUD MSc thesis project.

11. Which personal data or data from human participants do you work with? (Select all that apply.)

- Free text fields (for instance, in questionnaires) in which participants could unintentionally share personal data

12. Please list the categories of data subjects and their geographical location.

Survey participants are CSE Bachelor students following the course Probability Theory and Statistics (CSE1210) in Tu delft.

V. Data sharing and long term preservation

26. What data/code will be publicly shared?

Please provide a list of data/code you are going to share under 'Additional Information'.

- All data/code produced in the project

28. How will you share your research data/code?

- I am a Bachelor's/Master's student at TU Delft and I will share the data/code in the body and/or appendices of my thesis/report in the Education Repository

31. When will the data/code be shared?

- At the end of the research project

VI. Data management responsibilities and resources

33. If you leave TU Delft (or are unavailable), who is going to be responsible for the data/code resulting from this project?

My supervisor [Gosia, Migut, Assistant Professor, Department of Intelligent Systems], with email address [m.a.migut@tudelft.nl].

34. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

This is not applicable.

35. Which faculty do you belong to?

- Faculty of Electrical Engineering, Mathematics, and Computer Science (EEMCS)