Final Year Project

IDENTIFICATION OF DYSLEXIA STUDENTS BY THEIR HANDWRITING AMONG STUDENTS USING MULTIMODAL LLM

Detailed Project Plan

Chung Cheuk Man
3035928079

Lo Lok Fung
3035930228

Sin Chung Hang
3035928067

Tiquia Joezen S.
3035930096

Supervisor: Prof. Luo Ping

TABLE OF CONTENTS

1. Project Background

The concern with SEN students is an increasingly important issue in Hong Kong. In 2023/24, over 30000 primary school students were diagnosed with at least one type of SEN, with around 10% of the total number of primary school students that year. According to the guideline from the Education Bureau of Hong Kong (n.d.), nine kinds of special educational needs can attain support from the authority.

These nine types include (Education Bureau, n.d.):

1. Intellectual Disability
2. Autism Spectrum Disorder
3. Attention Deficit/Hyperactivity Disorder
4. Mental Illness
5. Specific Learning Difficulties
6. Physical Disability
7. Visual Impairment
8. Hearing Impairment
9. Speech and Language Impairment

This project outlines the procedure for the identification of specific learning disabilities, with dyslexia in particular. Such learning difficulties seriously jeopardize a student's capability to read and write, not to mention academic performance. Theoretically speaking, early identification and intervention will ensure all the differences needed to provide proper support for such youngsters embarking on their academic journey.

Diagnosis is usually done through the current methodology that involves teacher observation, standardized testing by an education psychologist and clinical psychologist, and student's family feedback. These procedures are limited in many aspects, such as the fact that traditional methods often fail to detect more minor manifestations of difficulties in accuracy and consistency.

To resolve this problem, large language models are an innovative solution for handling image classification, scene text understanding, and object character recognition. Advancements in AI these years have made it possible to identify handwriting and words with more consistent accuracy, unlike the current practice being done manually. It is now possible to combine textual and visual data to provide a comprehensive picture of student reading-writing ability by using InternVL.

InternVL is a pre-trained open-source multimodal large language model. Compared to the more advanced and existing large language models such as ChatGPT-4 and Claude 3.5, it works well in various tasks, like document and chart comprehension, infographics QA, scene text understanding, and OCR. (InternVL, 2024).

Given InternVL's capabilities, the project creates an integrated system that enhances the identification of students with dyslexia in Hong Kong. Also, it will help to generate a corresponding report to show the handwriting weaknesses of the students. As a result, it would provide teachers with specific insights into the situations, enabling them to facilitate prompt responses that meet each student's unique learning needs.

2. Project Objectives

It will implement an LLM model for better identification and supporting students with dyslexia in Hong Kong. The key objectives listed are as follows:

- Preparation of fine-tuning dataset

  Collect the students' Chinese handwriting dataset in labelled form, dyslexia, and non-dyslexia, respectively.

- Identifying dyslexia and analyzing handwriting issues by fine-tuning LLM:

  The process for improving model performance, based on fine-tuning the InternVL model will be reflected based on the detection of students' handwriting scans.

- Generating reports on Dyslexia student's handwriting:
  Providing meaningful insights, such as the type of writing problems, about the students so that teachers and parents can understand students' handwriting difficulties better.

- Evaluation of fine-tuned LLM model:

  Measuring the fine-tuned LLM model accuracy.

3. Project Methodology

This project focuses on the features of dyslexia and analyzes handwriting samples to identify dyslexia among students. The project will be investigated through different aspects, including data collection, the procedure of fine-tuning using InternVL, and lastly, the evaluation of the InternVL model.

1. Participants

A sample size of about 1000 students in primary schools and kindergarten will be randomly selected. Those who would participate shall include:

- 50% of students with dyslexia were previously diagnosed by educational psychologists and clinical psychologists.
- 50% of neurotypical students will serve as the control group, regardless of age and socio-economic background.

Recruitment will be further facilitated in cooperation with Special Educational Needs Coordinators and Social Workers at School; parental consent will be required before the children participate.

2. Data Collection Methods

2.1 Samples of Handwriting

There will be a request to undertake a standardized writing task designed to elicit natural handwriting. This includes copying a paragraph of text. Writing samples shall be collected on standardized A4 paper for consistency in analysis. A total size of about 100,000 Chinese characters in handwriting among all student samples would be collected to ensure sufficient data for fine-tuning.

2.2 Demographic and Background Information

A questionnaire will be sent to retrieve demographic data, like age and gender, and other relevant background information concerning diagnosed learning difficulties.

2.3 Labeling data and preprocessing

Before fine-tuning the model, the data should be labelled with three scores (orthographical, phonological, and semantic) and the student's situation (Lee et al., 2022).

3. Fine-tuning procedures:

3.1 Prepare customized SFT (Supervised Fine-Tuning) data

```
{
  "id": 0,
  "image": "path/to/image.jpg",
  "width": 111,
  "height": 222,
  "conversations": [
    {"from": "human", "value": "<image>\nuser input"},
    {"from": "gpt", "value": "assistant output"},
    {"from": "human", "value": "user input"},
    {"from": "gpt", "value": "assistant output"}
  ]
}
```

The image data entry will be a JSONL file (InternVL, 2024). The above is the sample image data structure in the JSONL file (InternVL, 2024). Those data entries will be incorporated into the SFT data for future fine-tuning processes.

According to an academic paper (Tong, 2020), the sampling metrics will be of three dimensions: orthographical, phonological, and semantic. These three dimensions will be used to classify whether the student belongs to dyslexia. Therefore, they will be labelled as the expected values in the conversation. from_gpt field.

3.2 Fine-tuning process

Fine-tuning the model by using different hyperparameters such as learning rate, number of training epochs, weight decay, batch size, and so on. Therefore, the fine-tuned model will be able to provide the value of orthographical, phonological, and semantic errors, and whether the student's writing belongs to dyslexia.

4. Evaluation:

4.1 Prepare data

Organize the data in a suitable format (a JSONL file). Original data will be split into data for fine-tuning and data for evaluation in a suitable portion.

4.2 Evaluation Score

For every student, there are three aspect rubrics to be graded, which are orthographical, phonological, and semantic errors (Tong et al., 2020). After taking these metrics into account, some benchmarks provided on InternVL will be used to assess students' performance.

4.3 Decision on Evaluation Benchmarks

Suitable benchmarks will be selected to evaluate the model performance and the likelihood of a student being diagnosed with dyslexia (InternVL, 2024). Initially, OCRBench, CCBench, and SEED are considered appropriate ones for this project because they feature image-character processing abilities and Chinese character handling capabilities.

4. PROJECT SCHEDULE AND MILESTONES

EVENT	START	DATE DONE
Deliverables of Phase 1 (Inception)	Mid-Sept	Oct 1, 2024
1. Detailed project plan	Mid-Sept	Oct 1, 2024
2. Project web page (set up)	Mid-Sept	Oct 1, 2024
3. Parties contacts	Mid-Sept	End of Project
Research on InternVL	Oct	Dec
First Presentation	Nov 1, 2024	Jan 13,2025
Deliverables of Phase 2 (Elaboration)	Oct 1, 2024	Jan 26, 2025
1. Preliminary implementation	Oct 1, 2024	Jan 26, 2025
2. Detailed interim report	Oct 1, 2024	Jan 26, 2025
Deliverables of Phase 3 (Construction)	Jan 26, 2025	Apr 21, 2025
1. Finalized tested implementation	Jan 26, 2025	Apr 21, 2025
2. Final report	Jan 26, 2025	Apr 21, 2025
Final presentation	Jan 26, 2025	Apr 22-26, 2025
Project exhibition	Apr 21, 2025	Apr 30, 2025
• 3-min video	Apr 21, 2025	Apr 30, 2025

5. Reference

Bureau, E. (n.d.). SENSE. https://sense.edb.gov.hk/en/types-of-special-educational-needs/specific-learning-difficulties/introduction.html

Education Bureau. (n.d.). SENSE - Introduction. SENSE - Integrated Education and Special Education Information Online. https://sense.edb.gov.hk/en/types-of-special-educational-needs/specific-learning-difficulties/introduction.html

Education Bureau. (n.d.). SENSE - Types of Special Educational Needs. SENSE - Integrated Education and Special Education Information Online. https://sense.edb.gov.hk/en/types-of-special-educational-needs/

Education Bureau. (n.d.). SENSE – Statistics of primary school education. https://www.edb.gov.hk/attachment/tc/about-edb/publication-stat/figures/pri_tc.xlsx

Education Bureau. (n.d.). SENSE – Statistics of primary Special education. https://www.edb.gov.hk/attachment/tc/about-edb/publication-stat/figures/special_tc.xlsx

M, A. (n.d.). How I trained an algorithm to have its own handwritten style. Medium. https://medium.com/@adityam/how-i-trained-an-algorithm-to-have-its-own-handwritten-style-d4b2fde8f0d1

Handwritten Recognition Techniques: A Comprehensive Review. (n.d.). https://www.researchgate.net/publication/228959667_Handwritten_Recognition_Techniques_A_Comprehensive_Review

Tong, X. S., Zhang, P., & He, X. (2020). Statistical learning of orthographic regularities in Chinese children with and without dyslexia. *Child Development*, *91*(6), 1953–1969. https://doi.org/10.1111/cdev.13384

Lee, M. K. S., & Tong, X. S. (2020). Spelling in developmental dyslexia in Chinese: Evidence of deficits in statistical learning and over-reliance on phonology. *Cognitive Neuropsychology*, *37*(7–8), 494–510. https://doi.org/10.1080/02643294.2020.1765754

Lee, M. K. S., Liu, H. W., & Tong, X. S. (2022). Identifying Chinese children with dyslexia using machine learning with character dictation. *Scientific Studies of Reading*, *27*(1), 82–100. https://doi.org/10.1080/10888438.2022.2088373

internvl. (2024, September 18). Evaluation of InternVL2 Series. internvl. https://internvl.readthedocs.io/en/latest/internvl2.0/evaluation.html#

internvl. (2024, September 18). Fine-tune on a custom dataset. internvl. https://internvl.readthedocs.io/en/latest/internvl2.0/finetune.html