Combining Log Data and Collaborative Dialogue Features to Predict Project Quality in Middle School AI Education




Conrad Borchers* — Carnegie Mellon University

Xiaoyi Tian* — NC State University

Kristy Elizabeth Boyer — UF University of Florida

Maya Israel — UF University of Florida

*Equal contribution

Paper link: https://tinyurl.com/csedm-amby

Introduction

- Project-based learning (PBL) is crucial in computing
- Predicting project quality during learning processes
  - inform adaptive modules
  - insights on effective student collaboration

This study: predict the quality of student chatbot projects in an collaborative, AI learning context

Research Questions

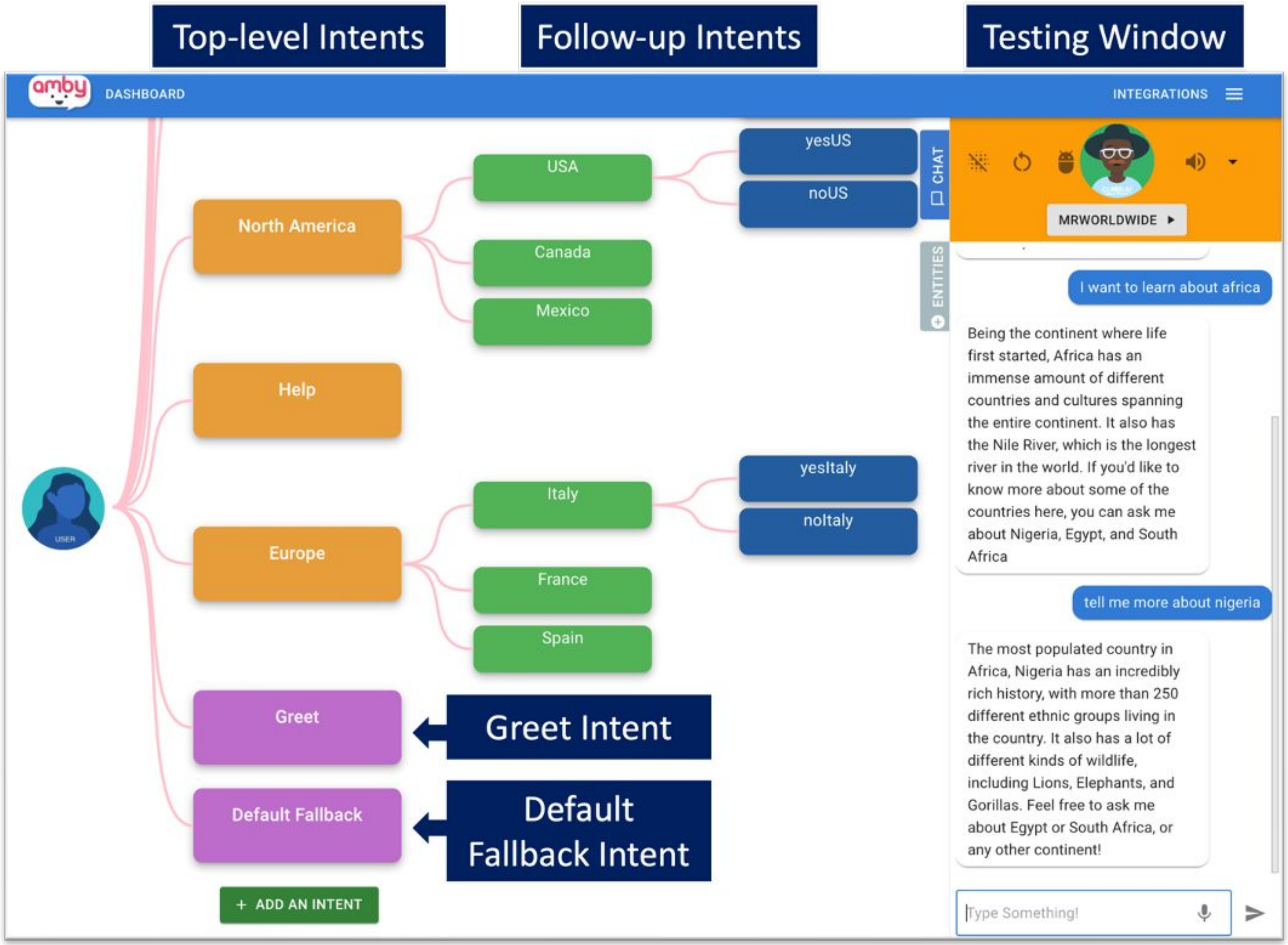
- RQ1: How well can student project quality be predicted from single modalities (dialogue, log data)?
- RQ2: To what extent does the multimodal fusion of these data sources enhance predictive accuracy?

Context: Pair Programming on AI Chatbots

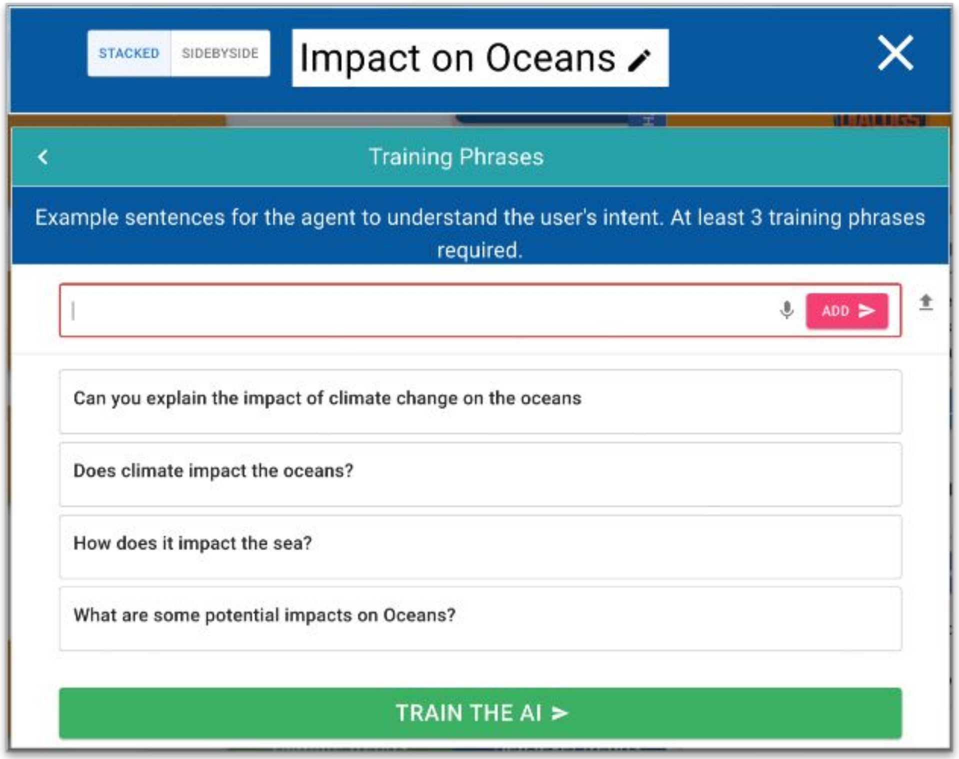
- Middle school students (average age 11.7 years) in science class
- Pair Programming for chatbots over three 40-min class sessions
- 47 student pairs (94 individuals)

Learning Platform:

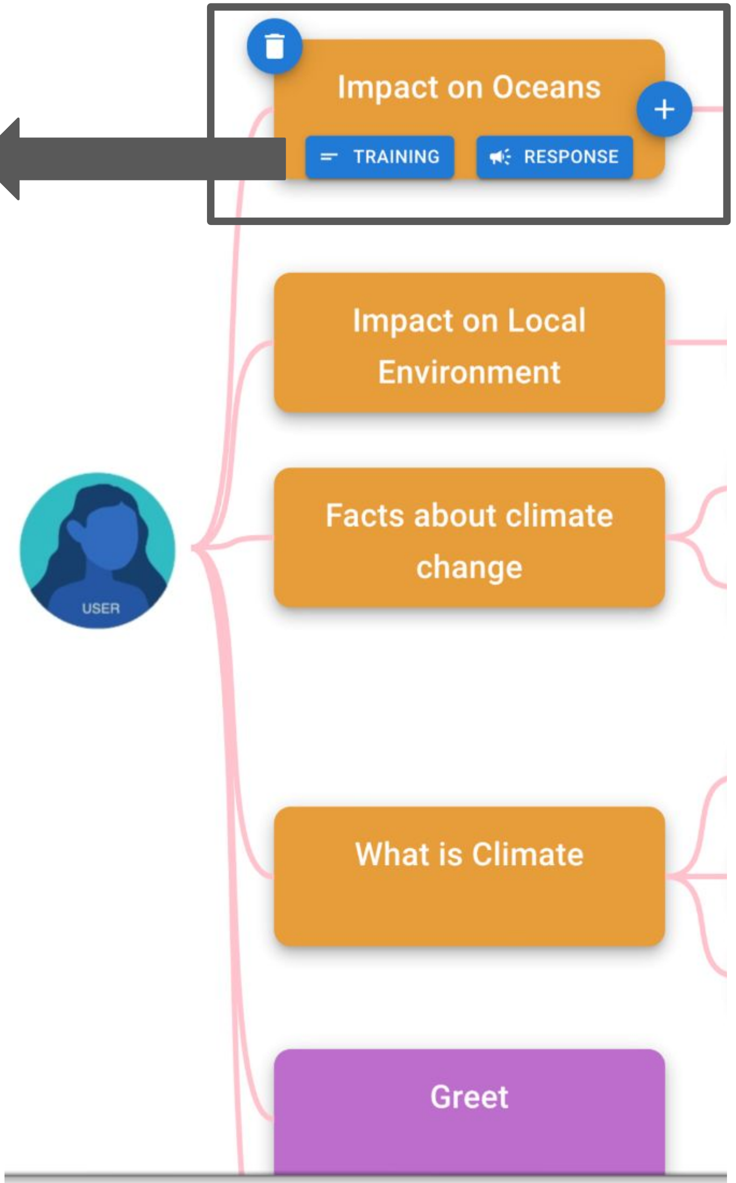
AMBY



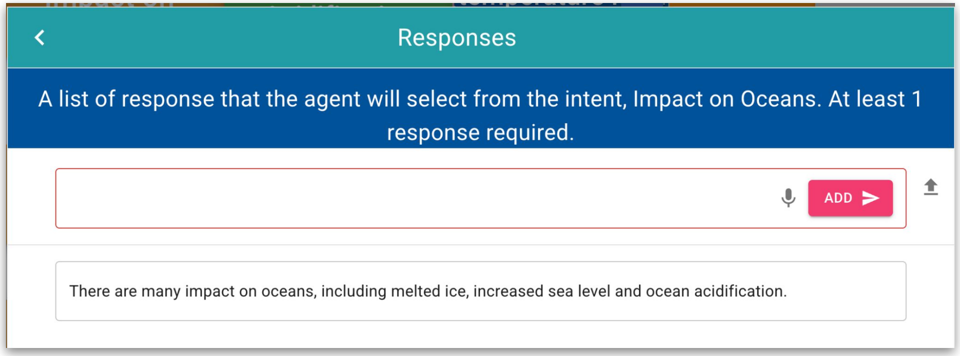
Training Phrases



Intents



Responses



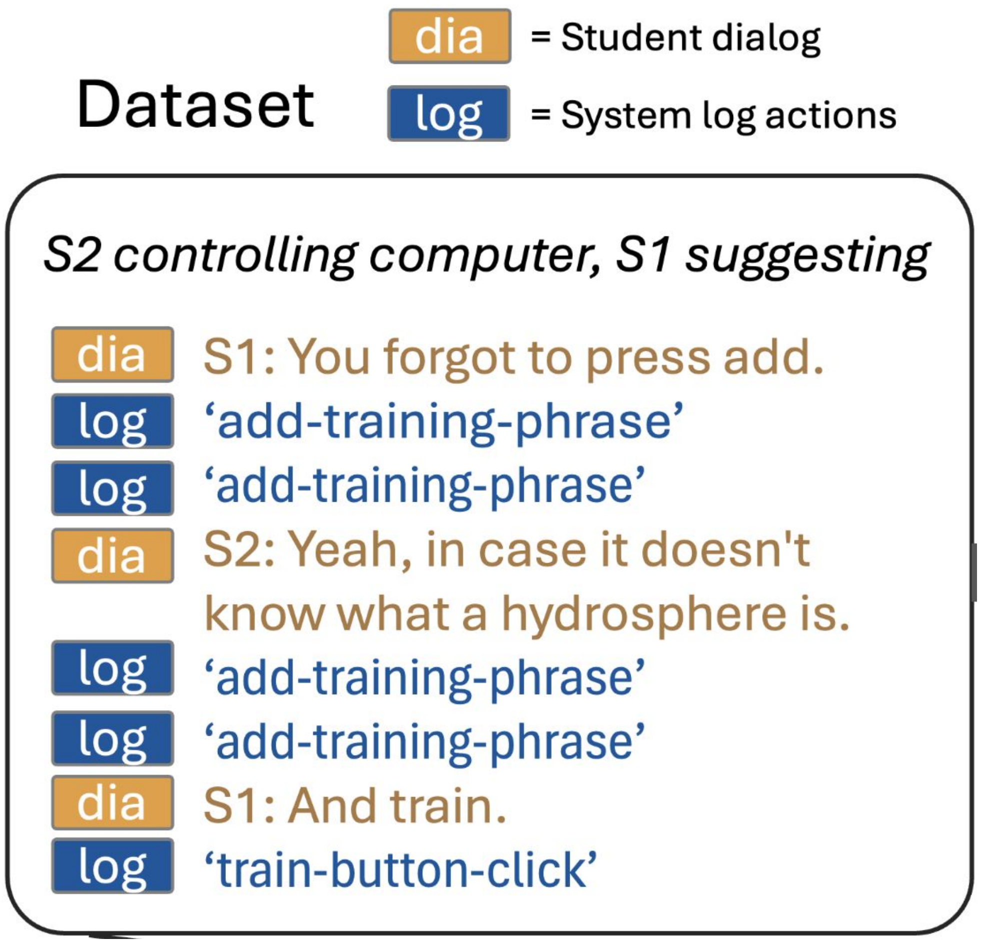
Dataset

Dialogue data:

- 121 30-minute collaboration sessions
- Human-transcribed
- Each session contains an average of 278 utterances (SD = 108.7)

Log data:

- 23 types of timestamped user interaction logs
- Average of 7 intent training requests per session



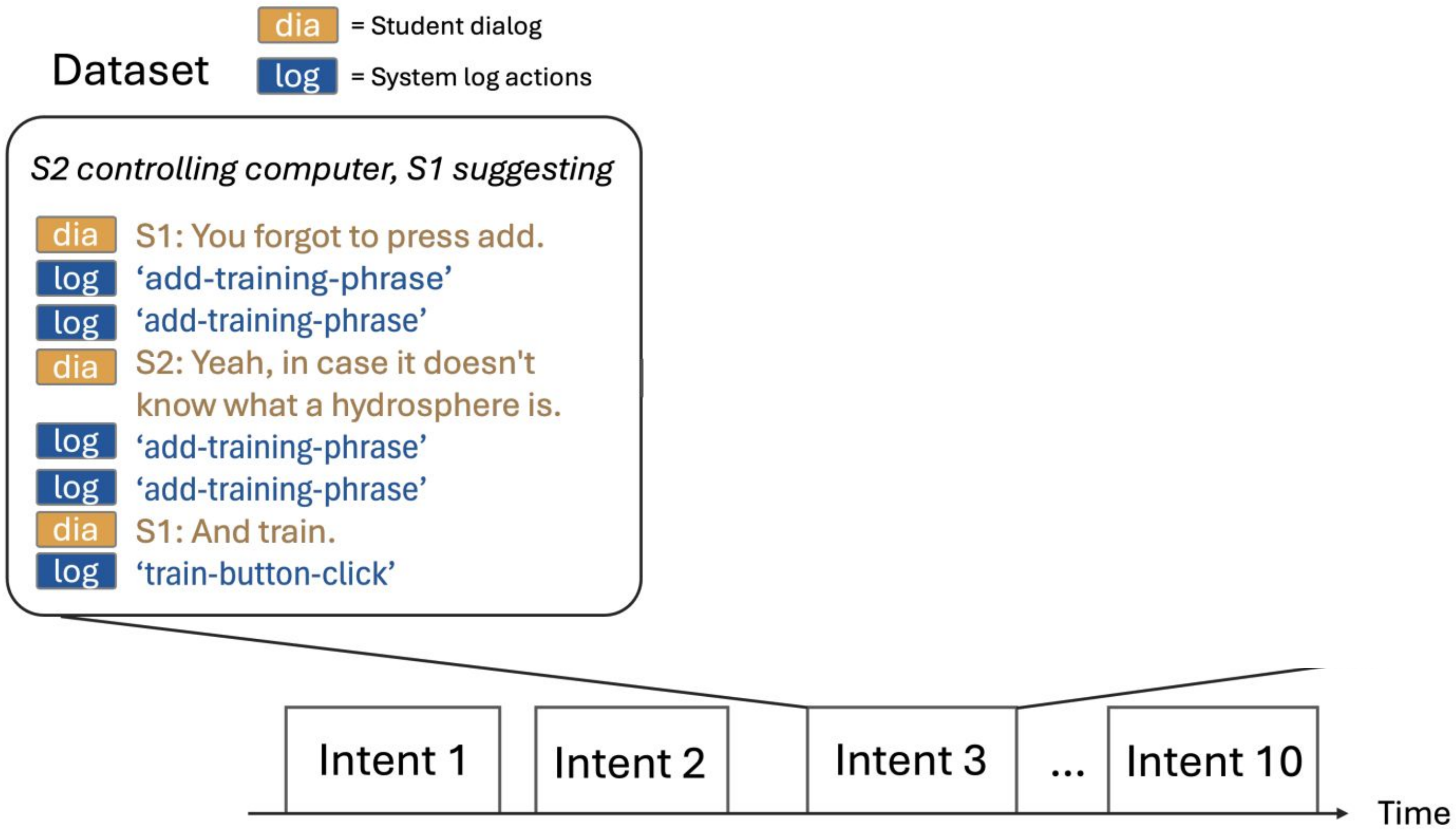
Outcome (Project Quality) Measures

- **Training Phrase Count (productivity)**: number of phrases input by students for training the chatbot
- **Lexical Density (content richness)**: the proportion of content words (nouns, adjectives, verbs, and adverbs) to total words
- **Lexical Variation**: the ratio of unique content words to total content word

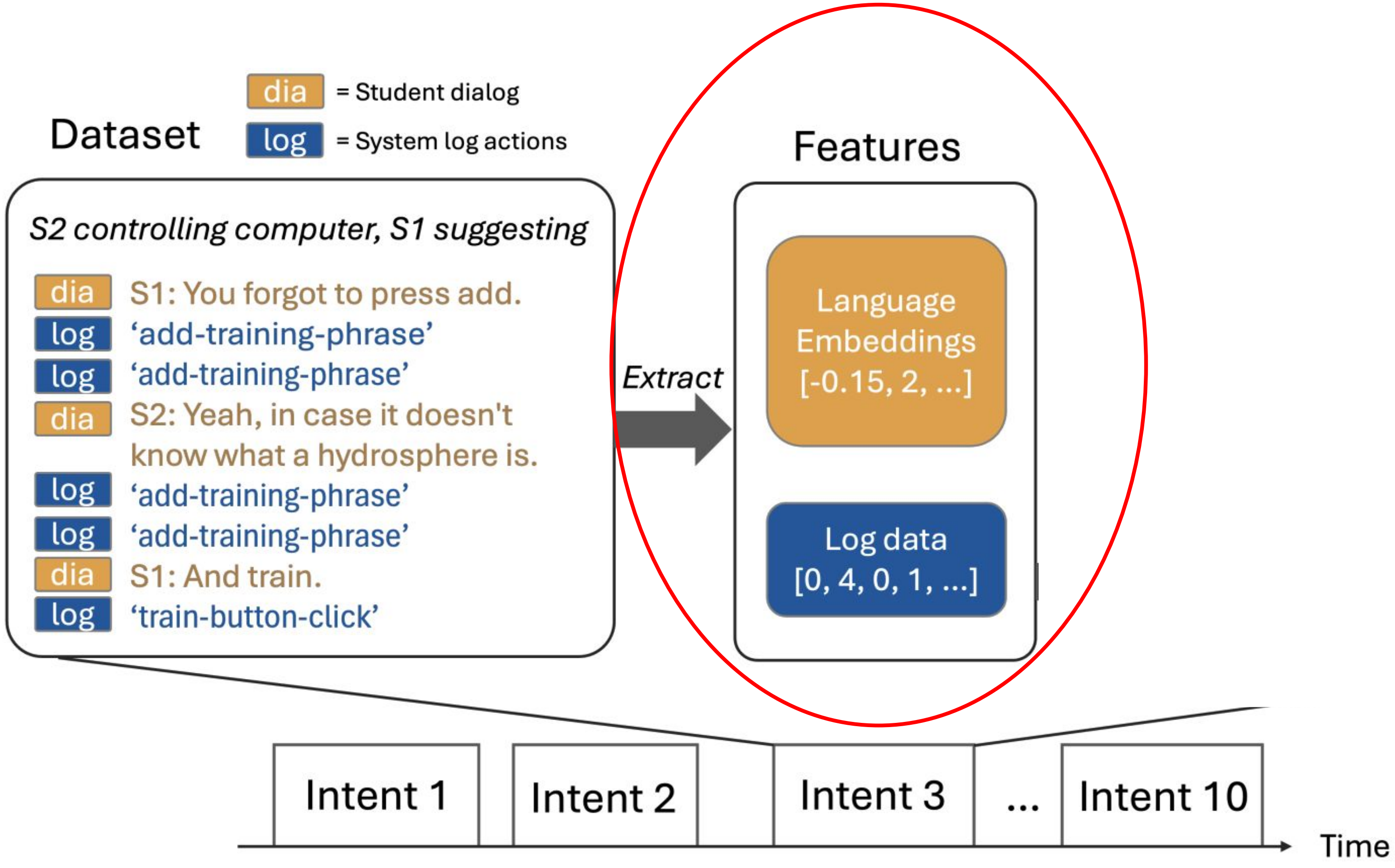
Justification of these measures:

- Alignment with key AI learning objectives
- Learning curve analyses
- Correlations with final project grades

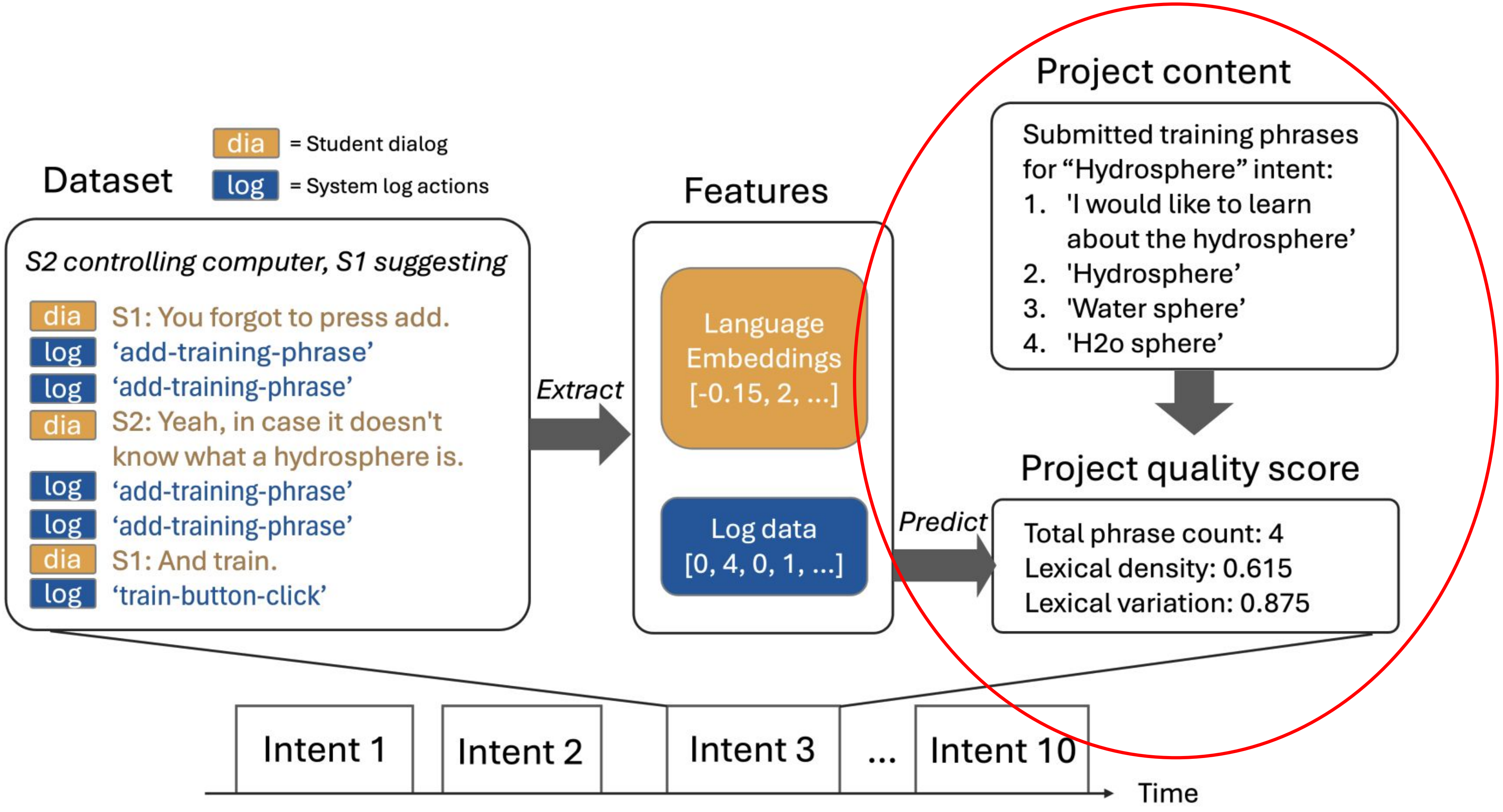
Data Wrangling and Segmentation



Data Wrangling and Segmentation



Data Wrangling and Segmentation



Machine Learning

Goal:
Predict project quality metrics (productivity, content richness, lexical variation) from **dialogue** and **log data** *together and in isolation*.

Machine Learning

Goal:
Predict project quality metrics (productivity, content richness, lexical variation) from **dialogue** and **log data** *together and in isolation*.

Model Architecture

- Feedforward neural network (2-4 hidden layers; CV-tuned)
  ReLU activation, dropout regularization (0-50%; CV-tuned)
- Optimized with **Adam** and early stopping (patience: 2 epochs)

Machine Learning

Goal:
Predict project quality metrics (productivity, content richness, lexical variation) from **dialogue** and **log data** *together and in isolation*.

Model Architecture

- Feedforward neural network (2-4 hidden layers; CV-tuned)
  ReLU activation, dropout regularization (0-50%; CV-tuned)
- Optimized with **Adam** and early stopping (patience: 2 epochs)

Evaluation Method

- 5-fold **student-level cross-validation**
- Tested on **33% held-out** set
- Performance metric: **AUC (median split)** with **95% bootstrapped confidence intervals**

Results

- RQ1: How well can student project quality be predicted from single modalities (dialogue, log data)?
- RQ2: To what extent does the multimodal fusion of these data sources enhance predictive accuracy?

Results: Unimodal Models

Outcome	Log Only AUC [95% CI]	Dialogue Only AUC [95% CI]
Training Phrase Count	**0.8053 [0.7470, 0.8604]***	0.5971 [0.5250, 0.6671]
Lexical Density	0.5112 [0.4556, 0.5655]	**0.6551 [0.5920, 0.7168]**
Lexical Variation	**0.6016 [0.5418, 0.6615]**	0.5260 [0.4579, 0.5933]

*0.6865 when excluding training-phrase setup transactions

Results: Unimodal Models

Outcome	Log Only AUC [95% CI]	Dialogue Only AUC [95% CI]
Training Phrase Count	**0.8053 [0.7470, 0.8604]***	0.5971 [0.5250, 0.6671]
Lexical Density	0.5112 [0.4556, 0.5655]	**0.6551 [0.5920, 0.7168]**
Lexical Variation	**0.6016 [0.5418, 0.6615]**	0.5260 [0.4579, 0.5933]

*0.6865 when excluding training-phrase setup transactions

Results

- RQ1: How well can student project quality be predicted from single modalities (dialogue, log data)?
- RQ2: To what extent does the multimodal fusion of these data sources enhance predictive accuracy?

Results: Multimodal Models

Outcome	Best Unimodal	Multimodal
Training Phrase Count	0.8053 [0.7470, 0.8604] (Log)	0.8301 [0.7732, 0.8822]
Lexical Density	0.6551 [0.5920, 0.7168] (Dialogue)	0.5700 [0.5042, 0.6352]
Lexical Variation	0.6016 [0.5418, 0.6615] (Log)	0.6089 [0.5438, 0.6727]

Discussion of Main Results

Log Data best predicts **productivity**

→ “Actions per minute” have shown similar insights into collaboration quality (Borchers et al., 2024)

→ Upside: Easy-to-generate proxies

→ Downside: Limited insight into what students do differently (there could be many confounds)

Discussion of Main Results

Differences between lexical variation (log data best) and lexical density (dialogue data best)

→ Both lexical variation and training phrase count might reflect distinct dimensions of productivity

→ *Surprising: Both measures are virtually uncorrelated (abs(r) < 0.03)*

Key Takeaway

- **Predictive value of modality *depends on the outcome being predicted***
- **Increasing evidence that the value of multimodal fusion in education depends on label, features, architecture, hyperparameter, and other modeling choices**
    - See, for instance, Wong et al., 2025; AIED 2025 best-paper nominated!

Looking Ahead and Applications in CS-EDU

Future Directions

- **Interpretability**: Apply SHAP or attention visualization to uncover **which features matter** most for each quality dimension.

- **Granularity**: Model individual student contributions and dialogue roles to better understand **collaborative dynamics**.

- **Real-time Adaptation**: Move toward **in-situ feedback**; flag low-quality input or disengagement during chatbot design sessions.
  a. *N.B.: Transcripts in this study were human-generated, though automated transcription might be feasible.*.

Looking Ahead and Applications in CS-EDU

Broader Applications

- **K-12 AI Literacy Tools**: Inform design of tools like AMBY to better scaffold **productive collaboration** and **linguistic diversity**.

- **Teacher Dashboards**: Provide educators with **process-level indicators** (e.g., engagement, content richness) for **formative assessment**.

- **Assessment Beyond Grades**: Promote **granular assessments** that value student thinking, not just final artifacts.
    a. *Potentially important in the LLM metacognitive laziness debate (see Fan et al., 2025; Weidlich et al., 2025).*

Conclusion

Contribution to CS Education

- Demonstrates the **feasibility of process-level prediction** in open-ended AI learning (with substantial room for improvement)

- Offers a pathway to a **scalable approach** for assessing project quality proxies in collaborative CS environments (e.g., for learning analytics and feedback)

- Echos recent research highlighting the **prediction task-dependent utility** of multimodal learning analytics.

Next Steps

- Improve **feature interpretability** and **real-time application**

- Broaden use to **other CS-EDU contexts** (e.g., block-based coding, data science) ***including through our open-source code***

Paper link: [***https://tinyurl.com/csedm-amby***](https://tinyurl.com/csedm-amby)

Combining Log Data and Collaborative Dialogue Features to Predict Project Quality in Middle School AI Education

Thank You!

Questions?

This work is supported by National Science Foundation DRL-2048480.

Code: ***https://github.com/conradborchers/collaboration-edm25***

Paper link: ***https://tinyurl.com/csedm-amby***

Let’s chat: ***cborcher@cs.cmu.edu*** | ***xtian9@ncsu.edu***