# InsPIRE Insight

## Learning Analytics with Scalable Bloom's Taxonomy Labeling of Socratic Chatbot Dialogues

By
Joel Weijia Lai
Centre for the Applications for Teaching & Learning Analytics for Students (ATLAS)
under the Institute for Pedagogical Innovation, Research & Excellence
Nanyang Technological University
Singapore
and
Co-authors: Kok Wai Lee & Yee Sin Ang
Singapore University of Technology and Design

## KEY IMPLICATIONS

- Scalable Bloom's Taxonomy labeling enables cognitive-level analytics of student–chatbot interactions, overcoming the limits of manual annotation.
- Classical ML models and fine-tuned LLMs provide complementary strengths, allowing institutions to balance accuracy, equity across labels, and real-time adaptability.
- Automated cognitive tagging can drive more pedagogically intelligent chatbots that scaffold students toward higher-order thinking.

## BACKGROUND

As higher education increasingly adopts educational chatbots to promote reasoning, reflection, and explanation, educators still lack scalable methods for assessing cognitive depth within these dialogues. Bloom's Taxonomy provides a structured lens for evaluating reasoning, yet manual annotation of conversational turns is resource-intensive and impractical at scale. The study addresses this challenge by developing a semi-supervised pipeline that automatically classifies student and chatbot utterances into Bloom cognitive levels, enabling richer learning analytics and more adaptive AI tutoring.

## FOCUS OF INITIATIVE

This initiative designs and evaluates a reproducible pipeline for automated Bloom's Taxonomy classification of 6,716 utterances collected from Socratic chatbot dialogues with 34 undergraduate statistics students. Only three students' conversations were expert-annotated, prompting the need for a scalable semi-supervised solution. The initiative introduces:

- A multi-label classification framework tailored to dialogic data, addressing overlap among Bloom levels.
- Two model tracks:
  1. Classical ML models (Logistic Regression, Linear SVM, XGBoost, MLP) using sentence embeddings and per-class calibrated thresholds.
  2. Fine-tuned lightweight LLM (GPT-4o-mini) trained on expert-labeled instances.
- A pseudo-labeling procedure to expand annotation coverage without reinforcing model bias.

Together, these enable Bloom-aware learning analytics for large-scale, authentic chatbot–student interactions.

## KEY OUTCOMES

- Performance:
  - Fine-tuned GPT-4o-mini achieved the highest micro-F1 (0.814), excelling at frequent lower-order cognitive classes.

  - The Linear SVM achieved the strongest macro-F1 (0.630), demonstrating balanced performance across rare higher-order skills.
- Cognitive Distribution Insights
  - Model-predicted profiles revealed strong dominance of Understanding across student utterances, with higher-order categories (Analyze, Evaluate, Create) occurring far less frequently.
  - Both LLM and classical models echoed this skew, although classical models better captured minority Bloom categories.
- Pedagogical Relevance: Automated detection of cognitive transitions provides a foundation for adaptive Socratic prompting, supporting deeper learning.

## SIGNIFICANCE OF OUTCOMES

*Implications for practice:* Educators can use automated Bloom-level analytics to identify whether chatbot interactions genuinely foster deeper reasoning, not merely factual recall. This enables refinement of chatbot prompts, targeted scaffolding, and design of learning activities that explicitly cultivate higher-order thinking.

*Implications for practice:* Educators can use automated Bloom-level analytics to identify whether chatbot interactions genuinely foster deeper reasoning, not merely factual recall. This enables refinement of chatbot prompts, targeted scaffolding, and design of learning activities that explicitly cultivate higher-order thinking.

*Proposed follow-up activities:*

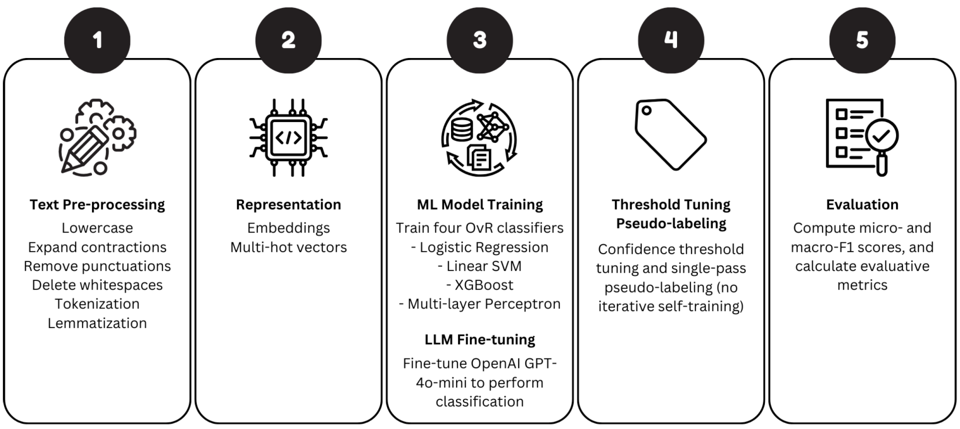
1. Expand expert annotation and model evaluation across diverse courses and institutions to enhance generalizability.
2. Integrate the Bloom classifier into live chatbot systems to enable real-time cognitive scaffolding.
3. Employ data augmentation or hybrid ensemble models to improve recognition of rare higher-order categories.

## PARTICIPANTS/SCOPE

34 undergraduate students in an NTU statistics course. 6,716 utterances analyzed across student–chatbot dialogues. Expert Bloom tagging performed for three students, representing 8.8% of the dataset; the remainder were pseudo-labeled.

## METHODOLOGY/APPROACH

The initiative employed a semi-supervised analytical pipeline integrating summarized in the following diagram:




Link to full article (if available): `https://doi.org/10.3390/computers14120555`
Contact Joel Weijia Lai at joellai@ntu.edu.sg for more information about the initiative.