

20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)

Vienna, Austria
31 July - 1 August 2025

Volume 1 of 2

ISBN: 979-8-3313-2401-8

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2025) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Large Language Models for Education: Understanding the Needs of Stakeholders, Current Capabilities and the Path Forward

Sankalan Pal Chowdhury, Nico Daheim, Ekaterina Kochmar, Jakub Macina, Donya Rooein, Mrinmaya Sachan and Shashank Sonkar 1

Comparing human and LLM proofreading in L2 writing: Impact on lexical and syntactic features

Hakyung Sung, Karla Csuros and Min-Chang Sung 11

MateInfoUB: A Real-World Benchmark for Testing LLMs in Competitive, Multilingual, and Multimodal Educational Tasks

Marius Dumitran, Mihnea Buca and Theodor Moroianu 24

Unsupervised Automatic Short Answer Grading and Essay Scoring: A Weakly Supervised Explainable Approach

Felipe Urrutia, Cristian Buc, Roberto Araya and Valentin Barriere 38

A Survey on Automated Distractor Evaluation in Multiple-Choice Tasks

Luca Benedetto, Shiva Taslimipoor and Paula Buttery 55

Alignment Drift in CEFR-prompted LLMs for Interactive Spanish Tutoring

Mina Almasi and Ross Kristensen-McLachlan 70

Leveraging Generative AI for Enhancing Automated Assessment in Programming Education Contests

Stefan Dascalescu, Marius Dumitran and Mihai Alexandru Vasiluta 89

Can LLMs Effectively Simulate Human Learners? Teachers' Insights from Tutoring LLM Students

Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang and Mrinmaya Sachan 100

Adapting LLMs for Minimal-edit Grammatical Error Correction

Ryszard Staruch, Filip Gralinski and Daniel Dzienisiewicz 118

COGENT: A Curriculum-oriented Framework for Generating Grade-appropriate Educational Content

Zhengyuan Liu, Stella Xin Yin, Dion Hoe-Lian Goh and Nancy Chen 129

Is Lunch Free Yet? Overcoming the Cold-Start Problem in Supervised Content Scoring using Zero-Shot LLM-Generated Training Data

Marie Bexte and Torsten Zesch 144

Transformer Architectures for Vocabulary Test Item Difficulty Prediction

Lucy Skidmore, Mariano Felice and Karen Dunn 160

Automatic concept extraction for learning domain modeling: A weakly supervised approach using contextualized word embeddings

Kordula De Kuthy, Leander Gierbach and Detmar Meurers 175

Towards a Real-time Swedish Speech Analyzer for Language Learning Games: A Hybrid AI Approach to Language Assessment

Tianyi Geng and David Alfter 186

Multilingual Grammatical Error Annotation: Combining Language-Agnostic Framework with Language-Specific Flexibility

Mengyang Qiu, Tran Minh Nguyen, Zihao Huang, Zelong Li, Yang Gu, Qingyu Gao, SILIANG LIU and Jungyeul Park 202

<i>LLM-based post-editing as reference-free GEC evaluation</i>	
Robert Östling, Murathan Kurfali and Andrew Caines	213
<i>Increasing the Generalizability of Similarity-Based Essay Scoring Through Cross-Prompt Training</i>	
Marie Bexte, Yuning Ding and Andrea Horbach	225
<i>Automated Scoring of a German Written Elicited Imitation Test</i>	
Mihail Chifligarov, Jammila Laâguidi, Max Schellenberg, Alexander Dill, Anna Timukova, Anastasia Drackert and Ronja Laarmann-Quante	237
<i>LLMs Protégés: Tutoring LLMs with Knowledge Gaps Improves Student Learning Outcome</i>	
Andrei Kucharavy, Cyril Vallez and Dimitri Percia David	248
<i>LEVOS: Leveraging Vocabulary Overlap with Sanskrit to Generate Technical Lexicons in Indian Languages</i>	
Karthika N J, Krishnakant Bhatt, Ganesh Ramakrishnan and Preethi Jyothi	258
<i>Do LLMs Give Psychometrically Plausible Responses in Educational Assessments?</i>	
Andreas Säuberli, Diego Frassinelli and Barbara Plank	266
<i>Challenges for AI in Multimodal STEM Assessments: a Human-AI Comparison</i>	
Aymeric de Chillaz, Anna Sotnikova, Patrick Jermann and Antoine Bosselut	279
<i>LookAlike: Consistent Distractor Generation in Math MCQs</i>	
Nisarg Parikh, Alexander Scarlatos, Nigel Fernandez, Simon Woodhead and Andrew Lan	294
<i>You Shall Know a Word's Difficulty by the Family It Keeps: Word Family Features in Personalised Word Difficulty Classifiers for L2 Spanish</i>	
Jasper Degraeuwe	312
<i>The Need for Truly Graded Lexical Complexity Prediction</i>	
David Alfter	326
<i>Towards Automatic Formal Feedback on Scientific Documents</i>	
Louise Bloch, Johannes Rückert and Christoph Friedrich	334
<i>Don't Score too Early! Evaluating Argument Mining Models on Incomplete Essays</i>	
Nils-Jonathan Schaller, Yuning Ding, Thorben Jansen and Andrea Horbach	345
<i>Educators' Perceptions of Large Language Models as Tutors: Comparing Human and AI Tutors in a Blind Text-only Setting</i>	
Sankalan Pal Chowdhury, Terry Jingchen Zhang, Donya Rooein, Dirk Hovy, Tanja Käser and Mrinmaya Sachan	356
<i>Transformer-Based Real-Word Spelling Error Feedback with Configurable Confusion Sets</i>	
Torsten Zesch, Dominic Gardner and Marie Bexte	375
<i>Automated L2 Proficiency Scoring: Weak Supervision, Large Language Models, and Statistical Guarantees</i>	
Aitor Arronte Alvarez and Naiyi Xie Fincham	384
<i>Automatic Generation of Inference Making Questions for Reading Comprehension Assessments</i>	
Wanqing (Anyu) Ma, Michael Flor and Zuowei Wang	398
<i>Investigating Methods for Mapping Learning Objectives to Bloom's Revised Taxonomy in Course Descriptions for Higher Education</i>	
Zahra Kolagar, Frank Zalkow and Alessandra Zarcone	415

<i>LangEye: Toward 'Anytime' Learner-Driven Vocabulary Learning From Real-World Objects</i> Mariana Shimabukuro, Deval Panchal and Christopher Collins	446
<i>Costs and Benefits of AI-Enabled Topic Modeling in P-20 Research: The Case of School Improvement Plans</i> Syeda Sabrina Akter, Seth Hunter, David Woo and Antonios Anastasopoulos	460
<i>Advances in Auto-Grading with Large Language Models: A Cross-Disciplinary Survey</i> Tania Amanda Nkoyo Frederick Eneye, Chukwuebuka Fortunate Ijezue, Ahmad Imam Amjad, Maaz Amjad, Sabur Butt and Gerardo Castañeda-Garza	477
<i>Unsupervised Sentence Readability Estimation Based on Parallel Corpora for Text Simplification</i> Rina Miyata, Toru Urakawa, Hideaki Tamori and Tomoyuki Kajiwara	499
<i>From End-Users to Co-Designers: Lessons from Teachers</i> Martina Galletti and Valeria Cesaroni	505
<i>LLMs in alliance with Edit-based models: advancing In-Context Learning for Grammatical Error Correction by Specific Example Selection</i> Alexey Sorokin and Regina Nasyrova	517
<i>Explaining Holistic Essay Scores in Comparative Judgment Assessments by Predicting Scores on Rubrics</i> Michiel De Vrindt, Renske Bouwer, Wim Van Den Noortgate, Marije Lesterhuis and Anaïs Tack	535
<i>Enhancing Arabic Automated Essay Scoring with Synthetic Data and Error Injection</i> Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash and Ted Briscoe	549
<i>Direct Repair Optimization: Training Small Language Models For Educational Program Repair Improves Feedback</i> Charles Koutcheme, Nicola Dainese and Arto Hellas	564
<i>Analyzing Interview Questions via Bloom's Taxonomy to Enhance the Design Thinking Process</i> Fateme Kazemi Vanhari, Christopher Anand and Charles Welch	582
<i>Estimation of Text Difficulty in the Context of Language Learning</i> Anisia Katinskaia, Anh-Duc Vu, Jue Hou, Ulla Vanhatalo, Yiheng Wu and Roman Yangarber	594
<i>Are Large Language Models for Education Reliable Across Languages?</i> Vansh Gupta, Sankalan Pal Chowdhury, Vilém Zouhar, Donya Rooein and Mrinmaya Sachan	612
<i>Exploiting the English Vocabulary Profile for L2 word-level vocabulary assessment with LLMs</i> Stefano Banno, Kate Knill and Mark Gales	632
<i>Advancing Question Generation with Joint Narrative and Difficulty Control</i> Bernardo Leite and Henrique Lopes Cardoso	647
<i>Down the Cascades of Omethi: Hierarchical Automatic Scoring in Large-Scale Assessments</i> Fabian Zehner, Hyo Jeong Shin, Emily Kerzabi, Andrea Horbach, Sebastian Gombert, Frank Goldhammer, Torsten Zesch and Nico Andersen	660
<i>Lessons Learned in Assessing Student Reflections with LLMs</i> Mohamed Elaraby and Diane Litman	672
<i>Using NLI to Identify Potential Collocation Transfer in L2 English</i> Haiyin Yang, Zoey Liu and Stefanie Wulff	687

<i>Name of Thrones: How Do LLMs Rank Student Names in Status Hierarchies Based on Race and Gender?</i>	
Annabella Sakunkoo and Jonathan Sakunkoo	697
<i>Exploring LLM-Based Assessment of Italian Middle School Writing: A Pilot Study</i>	
Adriana Mirabella and Dominique Brunato	708
<i>Exploring task formulation strategies to evaluate the coherence of classroom discussions with GPT-4o</i>	
Yuya Asano, Beata Beigman Klebanov and Jamie Mikeska	716
<i>A Bayesian Approach to Inferring Prerequisite Structures and Topic Difficulty in Language Learning</i>	
Anh-Duc Vu, Jue Hou, Anisia Katinskaia, Ching-Fan Sheu and Roman Yangarber	737
<i>Improving In-context Learning Example Retrieval for Classroom Discussion Assessment with Re-ranking and Label Ratio Regulation</i>	
Nhat Tran, Diane Litman, Benjamin Pierce, Richard Correnti and Lindsay Clare Matsumura .	752
<i>Exploring LLMs for Predicting Tutor Strategy and Student Outcomes in Dialogues</i>	
Fareya Ikram, Alexander Scarlatos and Andrew Lan	765
<i>Assessing Critical Thinking Components in Romanian Secondary School Textbooks: A Data Mining Approach to the ROTEX Corpus</i>	
Madalina Chitez, Liviu Dinu, Marius Micluta-Campeanu, Ana-Maria Bucur and Roxana Rogobete	780
<i>Improving AI assistants embedded in short e-learning courses with limited textual content</i>	
Jacek Marciniak, Marek Kubis, Michał Gulczyński, Adam Szpilkowski, Adam Wieczarek and Marcin Szczepański	794
<i>Beyond Linear Digital Reading: An LLM-Powered Concept Mapping Approach for Reducing Cognitive Load</i>	
Junzhi Han and Jinho D. Choi	805
<i>GermDetect: Verb Placement Error Detection Datasets for Learners of Germanic Languages</i>	
Noah-Manuel Michael and Andrea Horbach	818
<i>Enhancing Security and Strengthening Defenses in Automated Short-Answer Grading Systems</i>	
Sahar Yarmohammadtoosky, Yiyun Zhou, Victoria Yaneva, Peter Baldwin, Saed Rezayi, Brian Clauser and Polina Harik	830
<i>EyeLLM: Using Lookback Fixations to Enhance Human-LLM Alignment for Text Completion</i>	
Astha Singh, Mark Torrance and Evgeny Chukharev	841
<i>Span Labeling with Large Language Models: Shell vs. Meat</i>	
Phoebe Mulcaire and Nitin Madnani	850
<i>Intent Matters: Enhancing AI Tutoring with Fine-Grained Pedagogical Intent Annotation</i>	
Kseniia Petukhova and Ekaterina Kochmar	860
<i>Comparing Behavioral Patterns of LLM and Human Tutors: A Population-level Analysis with the CIMA Dataset</i>	
Aayush Kucheria, Nitin Sawhney and Arto Hellas	873
<i>Temporalizing Confidence: Evaluation of Chain-of-Thought Reasoning with Signal Temporal Logic</i>	
Zhenjiang Mao, Artem Bisliouk, Rohith Nama and Ivan Ruchkin	882

<i>Automated Scoring of Communication Skills in Physician-Patient Interaction: Balancing Performance and Scalability</i>	
Saed Rezayi, Le An Ha, Yiyun Zhou, Andrew Houriet, Angelo D’Addario, Peter Baldwin, Polina Harik, Ann King and Victoria Yaneva	891
<i>Decoding Actionability: A Computational Analysis of Teacher Observation Feedback</i>	
Mayank Sharma and Jason Zhang	898
<i>EduCSW: Building a Mandarin-English Code-Switched Generation Pipeline for Computer Science Learning</i>	
Ruishi Chen and Yiling Zhao	908
<i>STAIR-AIG: Optimizing the Automated Item Generation Process through Human-AI Collaboration for Critical Thinking Assessment</i>	
Euigyum Kim, Seewoo Li, Salah Khalil and Hyo Jeong Shin	920
<i>UPSC2M: Benchmarking Adaptive Learning from Two Million MCQ Attempts</i>	
Kevin Shi and Karttikeya Mangalam	931
<i>Can GPTZero’s AI Vocabulary Distinguish Between LLM-Generated and Student-Written Essays?</i>	
Veronica Schmalz and Anaïs Tack	937
<i>Paragraph-level Error Correction and Explanation Generation: Case Study for Estonian</i>	
Martin Vainikko, Taavi Kamarik, Karina Kert, Krista Liin, Silvia Maine, Kais Allkivi, Annekatrin Kaivapalu and Mark Fishel	953
<i>End-to-End Automated Item Generation and Scoring for Adaptive English Writing Assessment with Large Language Models</i>	
Kamel Nebhi, Amrita Panesar and Hans Bantilan	968
<i>A Framework for Proficiency-Aligned Grammar Practice in LLM-Based Dialogue Systems</i>	
Luisa Ribeiro-Flucht, Xiaobin Chen and Detmar Meurers	978
<i>Can LLMs Reliably Simulate Real Students’ Abilities in Mathematics and Reading Comprehension?</i>	
KV Aditya Srivatsa, Kaushal Maurya and Ekaterina Kochmar	988
<i>LLM-Assisted, Iterative Curriculum Writing: A Human-Centered AI Approach in Finnish Higher Education</i>	
Leo Huovinen and Mika Hämäläinen	1002
<i>Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors</i>	
Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack and Justin Vasselli	1011
<i>Jinan Smart Education at BEA 2025 Shared Task: Dual Encoder Architecture for Tutor Identification via Semantic Understanding of Pedagogical Conversations</i>	
Lei Chen	1034
<i>Wonderland_EDU@HKU at BEA 2025 Shared Task: Fine-tuning Large Language Models to Evaluate the Pedagogical Ability of AI-powered Tutors</i>	
Deliang Wang, Chao Yang and Gaowei Chen	1040
<i>bea-jh at BEA 2025 Shared Task: Evaluating AI-powered Tutors through Pedagogically-Informed Reasoning</i>	
Jihyeon Roh and Jinhyun Bang	1049

<i>CU at BEA 2025 Shared Task: A BERT-Based Cross-Attention Approach for Evaluating Pedagogical Responses in Dialogue</i>	
Zhihao Lyu	1060
<i>BJTU at BEA 2025 Shared Task: Task-Aware Prompt Tuning and Data Augmentation for Evaluating AI Math Tutors</i>	
Yuming Fan, Chuangchuang Tan and Wenyu Song	1073
<i>SYSUpporter Team at BEA 2025 Shared Task: Class Compensation and Assignment Optimization for LLM-generated Tutor Identification</i>	
Longfeng Chen, Zeyu Huang, Zheng Xiao, Yawen Zeng and Jin Xu	1078
<i>BLCU-ICALL at BEA 2025 Shared Task: Multi-Strategy Evaluation of AI Tutors</i>	
Jiyuan An, Xiang Fu, Bo Liu, Xuquan Zong, Cunliang Kong, Shuliang Liu, Shuo Wang, Zhenghao Liu, Liner Yang, Hanghang Fan and Erhong Yang	1084
<i>Phaedrus at BEA 2025 Shared Task: Assessment of Mathematical Tutoring Dialogues through Tutor Identity Classification and Actionability Evaluation</i>	
Rajneesh Tiwari and pranshu rastogi	1098
<i>Emergent Wisdom at BEA 2025 Shared Task: From Lexical Understanding to Reflective Reasoning for Pedagogical Ability Assessment</i>	
Raunak Jain and Srinivasan Rengarajan	1108
<i>Averroes at BEA 2025 Shared Task: Verifying Mistake Identification in Tutor, Student Dialogue</i>	
Mazen Yasser, Mariam Saeed, Hossam Elkordi and Ayman Khalafallah	1121
<i>SmolLab_SEU at BEA 2025 Shared Task: A Transformer-Based Framework for Multi-Track Pedagogical Evaluation of AI-Powered Tutors</i>	
Md. Abdur Rahman, MD AL AMIN, Sabik Aftahee, Muhammad Junayed and Md Ashiqur Rahman	1127
<i>RETUYT-INCO at BEA 2025 Shared Task: How Far Can Lightweight Models Go in AI-powered Tutor Evaluation?</i>	
Santiago G3n3gora, Ignacio Sastre, Santiago Robaina, Ignacio Remersaro, Luis Chiruzzo and Aiala Ros3	1135
<i>K-NLPers at BEA 2025 Shared Task: Evaluating the Quality of AI Tutor Responses with GPT-4.1</i>	
Geon Park, Jiwoo Song, Gihyeon Choi, Juoh Sun and Harksoo Kim	1145
<i>Henry at BEA 2025 Shared Task: Improving AI Tutor's Guidance Evaluation Through Context-Aware Distillation</i>	
Henry Pit	1164
<i>TBA at BEA 2025 Shared Task: Transfer-Learning from DARE-TIES Merged Models for the Pedagogical Ability Assessment of LLM-Powered Math Tutors</i>	
Sebastian Gombert, Fabian Zehner and Hendrik Drachsler	1173
<i>LexiLogic at BEA 2025 Shared Task: Fine-tuning Transformer Language Models for the Pedagogical Skill Evaluation of LLM-based tutors</i>	
Souvik Bhattacharyya, Billodal Roy, Niranjana M and Pranav Gupta	1180
<i>IALab UC at BEA 2025 Shared Task: LLM-Powered Expert Pedagogical Feature Extraction</i>	
Sof3a Correa Busquets, Valentina C3rdova V3liz and Jorge Baier	1187

<i>MSA at BEA 2025 Shared Task: Disagreement-Aware Instruction Tuning for Multi-Dimensional Evaluation of LLMs as Math Tutors</i>	
Baraa Hikal, Mohmaed Basem, Islam Oshallah and Ali Hamdi	1194
<i>TutorMind at BEA 2025 Shared Task: Leveraging Fine-Tuned LLMs and Data Augmentation for Mistake Identification</i>	
FATIMA DEKMAK, Christian Khairallah and Wissam Antoun	1203
<i>Two Outliers at BEA 2025 Shared Task: Tutor Identity Classification using DiReC, a Two-Stage Disentangled Contrastive Representation</i>	
Eduardus Tjitrahardja and Ikhlusal Hanif	1212
<i>Archaeology at BEA 2025 Shared Task: Are Simple Baselines Good Enough?</i>	
Ana Roşu, Jany-Gabriel Ispas and Sergiu Nisioi	1224
<i>NLIP at BEA 2025 Shared Task: Evaluation of Pedagogical Ability of AI Tutors</i>	
Trishita Saha, Shrenik Ganguli and Maunendra Sankar Desarkar	1242
<i>NeuralNexus at BEA 2025 Shared Task: Retrieval-Augmented Prompting for Mistake Identification in AI Tutors</i>	
Numaan Naeem, Sarfraz Ahmad, Momina Ahsan and Hasan Iqbal	1254
<i>DLSU at BEA 2025 Shared Task: Towards Establishing Baseline Models for Pedagogical Response Evaluation Tasks</i>	
Maria Monica Manlises, Mark Edward Gonzales and Lanz Lim	1260
<i>BD at BEA 2025 Shared Task: MPNet Ensembles for Pedagogical Mistake Identification and Localization in AI Tutor Responses</i>	
Shadman Rohan, Ishita Sur Apan, Muhtasim Shochcho, Md Fahim, Mohammad Rahman, AKM Mahbubur Rahman and Amin Ali	1266
<i>Thapar Titan/s : Fine-Tuning Pretrained Language Models with Contextual Augmentation for Mistake Identification in Tutor–Student Dialogues</i>	
Harsh Dadwal, Sparsh Rastogi and Jatin Bedi	1278