

BEA 2025

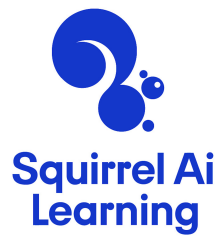
**The 20th Workshop on Innovative Use of NLP for Building
Educational Applications**

Proceedings of the Workshop

July 31 - August 1, 2025

The BEA organizers gratefully acknowledge the support from the following sponsors.

Gold Level



©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-270-1

Introduction

This year marks the 20th edition of the *Workshop on Innovative Use of NLP for Building Educational Applications*. As in previous years, we are happy to welcome a plethora of work on various aspects and types of educational applications – from traditionally popular tasks around language learning to novel applications related to teaching math and programming languages. This year, we have also extended BEA to a 2-day event, which allowed us to accept more valuable work from our authors: in total, we received a record number of 169 submissions, and from these, we have accepted 12 papers as talks and 63 as poster and demo presentations, for an overall acceptance rate of 44 percent. As in previous years, we have put the main emphasis on the high quality of research when selecting the papers to be accepted, but we also hope that we have managed to bring together a diverse program. One aspect in which BEA continues to excel is the range of languages that are covered by the work submitted and presented at our workshop: this year, accepted papers feature work on educational applications developed for Arabic, English, Estonian, Finnish, Germanic languages, Indian languages, Italian, Romanian, Russian, and Spanish.

In addition to the diverse oral, poster and demo presentations, this year, Kostiantyn Omelianchuk from Grammarly will give a keynote on *How LLMs Are Reshaping GEC: Training, Evaluation, and Task Framing*. BEA 2025 will also include, for the first time, a half-day tutorial on *LLMs for Education: Understanding the Needs of Stakeholders, Current Capabilities and the Path Forward*. Finally, BEA 2025 has hosted a shared task on *Pedagogical Ability Assessment of AI-powered Tutors*, which attracted a large number of participants, and the program includes an oral presentation on the shared task from the organizers as well as extended poster sessions for shared tasks participants presenting their systems.

Last but not least, we would like to thank everyone who has been involved in organizing the BEA workshop this year. We are particularly grateful to our sponsors who keep providing their support to BEA: this year, our sponsors include Cambridge University Press & Assessment, Duolingo English Test, Grammarly, National Board of Medical Examiners, SigIQ.ai, and Squirrel Ai Learning.

BEA 2025 Organizing Committee

Organizing Committee

General Chair

Ekaterina Kochmar, MBZUAI

Program Chairs

Andrea Horbach, Hildesheim University

Ronja Laarmann-Quante, Ruhr University Bochum

Marie Bexte, FernUniversität in Hagen

Publication Chair

Anaïs Tack, KU Leuven, imec

Shared Tasks Chairs

Victoria Yaneva, National Board of Medical Examiners

Bashar Alhafni, MBZUAI

Sponsorship Chair

Zheng Yuan, University of Sheffield

Jill Burstein, Duolingo

Program Committee

Chairs

Bashar Alhafni, MBZUAI
Marie Bexte, FernUniversität in Hagen
Jill Burstein, Duolingo
Andrea Horbach, CAU Kiel
Ekaterina Kochmar, MBZUAI
Ronja Laarmann-Quante, Ruhr University Bochum
Anaïs Tack, KU Leuven; imec; UCLouvain
Victoria Yaneva, National Board of Medical Examiners
Zheng Yuan, University of Sheffield

Program Committee

Rania Abdelghani, Hector Institute of Educational Sciences and Psychology, University of Tübingen
Tazin Afrin, NBME
Syeda Sabrina Akter, George Mason University
Ali Al-Laith, University of Copenhagen
Giora Alexandron, Weizmann Institute of Science
David Alfter, Gothenburg University
Jatin Ambasana, Indian Institute of Technology Bombay
Jiyuan An, Beijing Language and Culture University
Antonios Anastasopoulos, George Mason University
Nico Andersen, DIPF | Leibniz Institute for Research and Information in Education
Aitor Arronte Alvarez, University of Hawaii at Manoa
Yuya Asano, University of Pittsburgh
Nischal Ashok Kumar, University of Massachusetts Amherst
Berk Atil, Pennsylvania State University
Shiva Baghel, Extramarks
Xiaoyu Bai, University of Potsdam
Jinhyun Bang, Samsung Research
Stefano Banno, University of Cambridge
Mohmaed Basem, MSA University
Michael Gringo Angelo Bayona, Trinity College Dublin
Lee Becker, Pearson
Beata Beigman Klebanov, Educational Testing Service
Milena Belosevic, Bielefeld University
Enrico Benedetti, Utrecht University
Luca Benedetto, University of Cambridge
Maryam Berijanian, Michigan State University
Kay Berkling, Cooperative State University, Karlsruhe
Ummugul Bezirhan, Boston College, TIMSS and PIRLS International Study Center
Krishnakant Bhatt, IIT Bombay
Souvik Bhattacharyya, Lowe's
Abhidip Bhattacharyya, University of Massachusetts, Amherst
Serge Bibauw, Université catholique de Louvain
Louise Bloch, University of Applied Sciences and Arts Dortmund
Allison Bradford, University of California, Berkeley

Ted Briscoe, MBZUAI
 Dominique Brunato, Institute of Computational Linguistics A. Zampolli" (ILC-CNR), Pisa
 Ana-Maria Bucur, Interdisciplinary School of Doctoral Studies
 Luciano Cabral, IFPE
 Andrew Caines, University of Cambridge
 Chris Callison-Burch, University of Pennsylvania
 Jie Cao, University of Oklahoma
 Dan Carpenter, North Carolina State University
 Dumitru-Clementin Cercel, University Politehnica of Bucharest
 Sophia Chan, Educational Testing Service Canada
 Ignatios Charalampidis, University of Tuebingen
 Andreas Chari, University of Glasgow
 Danqing Chen, Technical University of Munich
 Lei Chen, Jinan University
 Mei-Hua Chen, Department of Foreign Languages and Literature, Tunghai University
 Longfeng Chen, South China University of Technology
 Artem Chernodub, ZenDesk
 Mihail Chifligarov, Ruhr University Bochum
 Luis Chiruzzo, Universidad de la Republica
 Hyundong Cho, USC, Information Sciences Institute
 Jinho D. Choi, Emory University
 Evgeny Chukharev, Iowa State University
 Yan Cong, Purdue University
 Mark Core, University of Southern California
 Sofia Correa Busquets, Pontificia Universidad Católica de Chile, National Center for Artificial Intelligence Chile, Foundational Research on Data Millenium Institute
 Steven Coyne, Tohoku University / RIKEN
 Scott Crossley, Georgia State University
 Syaamantak Das, Indian Institute of Technology Bombay
 Mihai Dascalu, University Politehnica of Bucharest
 Tirthankar Dasgupta, Tata Consultancy Services Ltd.
 Orphee De Clercq, LT3, Ghent University
 Kordula De Kuthy, Universität Tübingen
 Michiel De Vrindt, KU Leuven
 Jasper Degraeuwe, Ghent University
 FATIMA DEKMAK, American University of Beirut
 Carrie Demmans Epp, University of Alberta
 Dorottya Demszky, Stanford University
 Aniket Deroy, IIT Kharagpur
 Chris Develder, Ghent University
 Srijita Dhar, Chittagong University of Engineering & Technology
 Yuning Ding, FernUniversität in Hagen
 Rahul Divekar, Educational Testing Service
 George Duenas, Universidad Pedagogica Nacional
 Marius Dumitran, University of Bucharest
 Yo Ehara, Tokyo Gakugei University
 Walid El Hefny, Leibniz-Institut für Wissensmedien (IWM)
 Mohamed Elaraby, University of Pittsburgh
 Ron Eliav, Bar-Ilan University
 Jordan Esiason, North Carolina State University
 Yao-Chung Fan, National Chung Hsing University

Effat Farhana, Auburn University
 Mariano Felice, British Council
 Nigel Fernandez, University of Massachusetts Amherst
 Michael Flor, Educational Testing Service
 Jennifer-Carmen Frey, EURAC Research
 Benjamin Gagl, University of Cologne
 Thomas Gaillat, Rennes 2 university
 Martina Galletti, Sony Computer Science Laboratories - Paris | Sapienza University of Rome
 Diana Galvan-Sosa, University of Cambridge
 Ashwinkumar Ganesan, Amazon Alexa AI
 Rujun Gao, Texas A&M University
 Lingyu Gao, Toyota Technological Institute at Chicago
 Ritik Garg, Extramarks Education Pvt. Ltd.
 Voula Giouli, Aristotle University of Thessaloniki / ILSP, ATHENA RC
 Sebastian Gombert, DIPF | Leibniz Institute for Research and Information in Education
 Kiel Gonzales, University of the Philippines Diliman
 Mark Edward Gonzales, De La Salle University
 Cyril Goutte, National Research Council Canada
 Pranav Gupta, Lowe's
 Abigail Gurin Schleifer, Weizmann Institute of Science
 Eleonora Guzzi, Universidade da Coruña
 Ching Nam Hang, Assistant Professor, Yam Pak Charitable Foundation School of Computing and Information Sciences, Saint Francis University, Hong Kong
 Ikhlusal Hanif, Universitas Indonesia
 Jiangang Hao, Educational Testing Service
 Ahatsham Hayat, University of Nebraska-Lincoln
 Ping He, Northeastern University
 Nicolas Hernandez, Nantes University
 Michael Holcomb, University of Texas Southwestern Medical Center
 Matias Hoyle, Stanford University
 Chieh-Yang Huang, MetaMetrics Inc
 Aiden Huang, Acton-Boxborough Regional High School
 Chung-Chi Huang, Frostburg State University
 Anna Huelsing, CAU
 Leo Huovinen, Metropolia University of Applied Sciences
 Catherine Ikae, Applied Machine Intelligence, Bern University of Applied Sciences, Switzerland
 Fareya Ikram, University of Massachusetts Amherst
 Joseph Marvin Imperial, University of Bath
 Radu Tudor Ionescu, University of Bucharest
 Raunak Jain, Intuit
 Suriya Prakash Jambunathan, New York University
 Qinjin Jia, North Carolina State University
 Helen Jin, University of Pennsylvania
 Abel John, Stanford University
 Douglas Jones, MIT Lincoln Laboratory
 Edmund Jones, Cambridge University Press & Assessment
 Léane Jourdan, Nantes University
 Samarth Kadaba, Stanford University
 Indika Kahanda, University of North Florida
 Tomoyuki Kajiwara, Ehime University
 Honeiah Karimi, Cambium Assessment

Anisia Katinskaia, University of Helsinki
 Fatemeh Kazemi Vanhari, McMaster University
 Elma Kerz, Exaia Technologies
 Fazel Keshtkar, St. John's University
 Samin Khan, Stanford University
 Darya Kharlamova, National Research University Higher School of Economics
 Harksoo Kim, Konkuk University
 Han Kyul Kim, University of Southern California
 Levi King, Indiana University
 Kasper Knudsen, ITU
 David Kogan, Google
 Mamoru Komachi, Hitotsubashi University
 Charles Koutcheme, Aalto University
 Joni Kruijsbergen, LT3, Ghent University
 Andrei Kucharavy, HES-SO Valais-Wallis
 Aayush Kucheria, Aalto University
 Roland Kuhn, National Research Council of Canada
 Gaurav Kumar, University of California San Diego
 Murathan Kurfali, RISE Research Institutes of Sweden
 Alexander Kwako, University of California, Los Angeles
 Kristopher Kyle, University of Oregon
 Yunshi Lan, East China Normal University
 Antonio Laverghetta Jr., Pennsylvania State University
 Jaewook Lee, UMass Amherst
 Celine Lee, Cornell University
 Seolhwa Lee, Technical University of Darmstadt
 Travis Lee, Tennessee Tech University
 Bernardo Leite, Faculty of Engineering - University of Porto
 Arun Balajee Lekshmi Narayanan, University of Pittsburgh
 Xu Li, Zhejiang University
 Hariz Liew, Singapore University of Social Sciences
 Chuan-Jie Lin, National Taiwan Ocean University
 Yudong Liu, Western Washington University
 Naiming Liu, Rice University
 Zhexiong Liu, University of Pittsburgh
 Julian Lohmann, Christian Albrechts Universität Kiel
 Benny Longwill, Educational Testing Service
 Anastassia Loukina, Grammarly Inc
 Crisron Rudolf Lucas, University College Dublin
 Zhihao Lyu, CU Boulder
 Sarah Löber, University of Tübingen
 Denise Löfflad, Leibniz-Institut für Wissensmedien Tübingen
 Wanjing (Anyu) Ma, Stanford University
 Jakub Macina, ETH Zurich
 Lieve Macken, Ghent University
 Nitin Madnani, Duolingo
 Hang Man, The University of Hong Kong
 Zhenjiang Mao, University of Florida
 Jacek Marciniak, Adam Mickiewicz University
 Arianna Masciolini, University of Gothenburg
 Sandeep Mathias, Presidency University

Kaushal Maurya, MBZUAI
 Hunter McNichols, University of Massachusetts Amherst
 Detmar Meurers, Leibniz-Institut für Wissensmedien (IWM)
 Noah-Manuel Michael, Kiel University
 Amit Mishra, AmityUniversityMadhyaPradesh
 Daniel Mora Melanchthon, Leibniz Institute for Science and Mathematics Education
 Sai Sathvik Motamarri, PES University
 Phoebe Mulcaire, Duolingo
 Laura Musto, Universidad de la Republica
 Karthika N J, Indian Institute of Technology Bombay
 Farah Nadeem, LUMS
 Numaan Naeem, MBZUAI
 Ryo Nagata, Konan University
 Sungjin Nam, ACT, Inc
 Diane Napolitano, The Washington Post
 Aneet Narendranath, Michigan Technological University
 Léo Nebel, LIP6 - Sorbonne Université
 Kamel Nebhi, Education First
 Seyed Parsa Neshaei, EPFL
 Huy Nguyen, Amazon
 Gebregziabihier Nigusie, Mizan-Tepi University
 S Jaya Nirmala, National Institute of Technology Tiruchirappalli
 Sergiu Nisioi, Human Language Technologies Research Center, University of Bucharest
 Adam Nohejl, Nara Institute of Science and Technology
 Eda Okur, Intel Labs
 Kostiantyn Omelianchuk, Grammarly
 Amin Omidvar, PhD student at the Department of Electrical Engineering and Computer Science, York University
 Joshua Otten, GeorgeMasonUniversity
 Daniel Oyeniran, University of Alabama
 Ulrike Pado, HFT Stuttgart
 Sankalan Pal Chowdhury, ETH Zurich
 Nisarg Parikh, University of Massachusetts, Amherst
 Jeiyoon Park, SOOP
 Manooshree Patel, University of California, Berkeley
 Kaushal Patil, University of Southern California
 Kseniia Petukhova, MBZUAI
 Henry Pit, University of Melbourne
 Long Qin, Alibaba
 Mengyang Qiu, Trent University
 Marti Quixal, University of Tuebingen
 Chatrine Qwaider, MBZUAI
 Md. Abdur Rahman, Southeast University
 Vatsal Raina, University of Cambridge
 Sparsh Rastogi, Thapar Institute of Engineering and Technology
 pranshu rastogi, Independent Researcher
 Manav Rathod, University of California, Berkeley
 Hanumant Redkar, Goa University, Goa
 Robert Reynolds, Brigham Young University
 Saed Rezayi, National Board of Medical Examiners
 Luisa Ribeiro-Flucht, University of Tuebingen

Frankie Robertson, University of Jyväskylä
 Shadman Rohan, Center for Computational & Data Sciences, IUB
 Donya Rooein, Bocconi University
 Aiala Rosá, Instituto de Computación, Facultad de Ingeniería, Universidad de la República
 Allen Roush, University of Oregon
 Alla Rozovskaya, Queens College, City University of New York
 Josef Ruppenhofer, Fernuniversität in Hagen
 Stefan Ruseti, University Politehnica of Bucharest
 Johannes Rückert, University of Applied Sciences and Arts Dortmund
 Mariam Saeed, Applied Innovation Center
 Trishita Saha, IIT Hyderabad
 Jonathan Sakunkoo, Stanford University
 Annabella Sakunkoo, Stanford University OHS
 Omer Salem, Cairo University
 Nicy Scaria, Indian Institute of Science
 Nils-Jonathan Schaller, Leibniz Institute for Science and Mathematics Education
 Veronica Schmalz, KU Leuven
 Stephanie Schoch, University of Virginia
 Matthew Shardlow, Manchester Metropolitan University
 Mayank Sharma, Graduate Student, Stanford University
 Kevin Shi, University of California, Berkeley
 Mariana Shimabukuro, Ontario Tech University
 Hyo Jeong Shin, Sogang University
 Gyu-Ho Shin, University of Illinois Chicago
 Astha Singh, Iowa State University
 Li Siyan, Columbia University
 Lucy Skidmore, British Council
 Anastasia Smirnova, San Francisco State University
 Mariia Soliar, Leibniz-Institut für Wissensmedien (IWM)
 Mayank Soni, ADAPT Centre, Trinity College Dublin
 Alexey Sorokin, Moscow State University
 Anna Sotnikova, EPFL
 KV Aditya Srivatsa, MBZUAI
 Maja Stahl, Leibniz University Hannover
 Felix Stahlberg, Google Research
 Katherine Stasaski, Salesforce Research
 Helmer Strik, Centre for Language and Speech Technology (CLST), Centre for Language Studies (CLS), Radboud University Nijmegen
 David Strohmaier, University of Cambridge
 Hakyung Sung, University of Oregon
 Abhijit Suresh, Graduate Student
 Andreas Säuberli, LMU Munich
 Chuangchuang Tan, Beijing Jiaotong University
 CheeWei Tan, Nanyang Technological University
 Wenjia Tan, University of Macau
 Nhat Tran, University of Pittsburgh
 Felipe Urrutia, Center for Advanced Research in Education
 Masaki Uto, The University of Electro-Communications
 Takehito Utsuro, University of Tsukuba
 Martin Vainikko, University of Tartu
 Sowmya Vajjala, National Research Council

Piper Vasicek, Brigham Young University
 Justin Vasselli, Nara Institute of Science and Technology
 Giulia Venturi, Institute of Computational Linguistics Antonio Zampolli"(ILC-CNR)
 Anthony Verardi, Duolingo
 Amit Verma, Guvi Geek Network
 Elena Volodina, University of Gothenburg
 Anh-Duc Vu, University of Helsinki
 Deliang Wang, The University of Hong Kong
 Nikhil Wani, University of Southern California
 Taro Watanabe, Nara Institute of Science and Technology
 Yuchen Wei, Pennsylvania State University
 Alistair Willis, The Open University
 Steven Wilson, University of Michigan-Flint
 Anna Winklerova, Faculty of Informatics Masaryk University
 Hanna Woloszyn, University of Cologne
 Simon Woodhead, Eedi
 Anna Wroblewska, Faculty of Mathematics and Information Science, Warsaw University of Technology
 Changrong Xiao, Tsinghua University
 Hiroaki Yamada, Institute of Science Tokyo
 Haiyin Yang, University of Florida
 Roman Yangarber, University of Helsinki
 Sahar Yarmohammadtoosky, NBME
 Hanling Yi, Intellifusion, Inc.
 Su-Youn Yoon, EduLab
 Marcos Zampieri, George Mason University
 Alessandra Zarcone, Technische Hochschule Augsburg
 Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education
 Kamyar Zeinalipour, University of Siena
 Torsten Zesch, Computational Linguistics, FernUniversität in Hagen
 Franklin Zhang, Bellevue College
 Mike Zhang, Aalborg University
 Jing Zhang, Emory University
 Yiling Zhao, Stanford University
 Yang Zhong, University of Pittsburgh
 Yiyun Zhou, NBME
 Ej Zhou, University of Cambridge
 Jessica Zipf, University of Konstanz
 Michael Zock, CNRS-LIS
 Leonidas Zotos, University of Groningen
 Bowei Zou, Institute for Infocomm Research
 Robert Östling, Department of Linguistics, Stockholm University

Keynote Talk

How LLMs Are Reshaping GEC: Training, Evaluation, and Task Framing

Kostiantyn Omelianchuk
Grammarly

Abstract: This keynote will explore the evolving role of Large Language Models (LLMs) in training and evaluating Grammatical Error Correction (GEC) systems, using Grammarly as a case study. It will cover the shift from primarily using human-annotated corpora to semi-synthetic data generation approaches, examining its impact on model training, evaluation practices, and overall task definition. Key topics include task definition challenges, trade-offs between data types, observed biases in models, and recent advances in LLM-based evaluation techniques. The talk will also explore scalable approaches for multilingual GEC and outline implications for future research.

Bio: Kostiantyn Omelianchuk is an Applied Research Scientist and Area Tech Lead at Grammarly, where he works on practical applications of NLP, with a primary interest in Grammatical Error Correction (GEC). He has over nine years of experience in the field and has co-authored several papers, including GECToR: Grammatical Error Correction – Tag, Not Rewrite, a widely used approach in the GEC community. His research explores edit-based modeling, the use of large language models for text correction and simplification, and the transition from human-annotated to synthetic data for training and evaluation. His recent work focuses on multilingual GEC, LLM-based evaluation methods, and synthetic data generation.

Table of Contents

<i>Large Language Models for Education: Understanding the Needs of Stakeholders, Current Capabilities and the Path Forward</i>	
Sankalan Pal Chowdhury, Nico Daheim, Ekaterina Kochmar, Jakub Macina, Donya Rooein, Mrinmaya Sachan and Shashank Sonkar	1
<i>Comparing human and LLM proofreading in L2 writing: Impact on lexical and syntactic features</i>	
Hakyung Sung, Karla Csuros and Min-Chang Sung	11
<i>MateInfoUB: A Real-World Benchmark for Testing LLMs in Competitive, Multilingual, and Multimodal Educational Tasks</i>	
Marius Dumitran, Mihnea Buca and Theodor Moroianu	24
<i>Unsupervised Automatic Short Answer Grading and Essay Scoring: A Weakly Supervised Explainable Approach</i>	
Felipe Urrutia, Cristian Buc, Roberto Araya and Valentin Barriere	38
<i>A Survey on Automated Distractor Evaluation in Multiple-Choice Tasks</i>	
Luca Benedetto, Shiva Taslimipoor and Paula Buttery	55
<i>Alignment Drift in CEFR-prompted LLMs for Interactive Spanish Tutoring</i>	
Mina Almasi and Ross Kristensen-McLachlan	70
<i>Leveraging Generative AI for Enhancing Automated Assessment in Programming Education Contests</i>	
Stefan Dascalescu, Marius Dumitran and Mihai Alexandru Vasiluta	89
<i>Can LLMs Effectively Simulate Human Learners? Teachers' Insights from Tutoring LLM Students</i>	
Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang and Mrinmaya Sachan	100
<i>Adapting LLMs for Minimal-edit Grammatical Error Correction</i>	
Ryszard Staruch, Filip Gralinski and Daniel Dzienisiewicz	118
<i>COGENT: A Curriculum-oriented Framework for Generating Grade-appropriate Educational Content</i>	
Zhengyuan Liu, Stella Xin Yin, Dion Hoe-Lian Goh and Nancy Chen	129
<i>Is Lunch Free Yet? Overcoming the Cold-Start Problem in Supervised Content Scoring using Zero-Shot LLM-Generated Training Data</i>	
Marie Bexte and Torsten Zesch	144
<i>Transformer Architectures for Vocabulary Test Item Difficulty Prediction</i>	
Lucy Skidmore, Mariano Felice and Karen Dunn	160
<i>Automatic concept extraction for learning domain modeling: A weakly supervised approach using contextualized word embeddings</i>	
Kordula De Kuthy, Leander Gierbach and Detmar Meurers	175
<i>Towards a Real-time Swedish Speech Analyzer for Language Learning Games: A Hybrid AI Approach to Language Assessment</i>	
Tianyi Geng and David Alfter	186
<i>Multilingual Grammatical Error Annotation: Combining Language-Agnostic Framework with Language-Specific Flexibility</i>	
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>	
Ayméric 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>	
Wanjing (Anyu) Ma, Michael Flor and Zuwei 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 Koutchme, Nicola Dainese and Arto Hellas	564
<i>Analyzing Interview Questions via Bloom's Taxonomy to Enhance the Design Thinking Process</i>	
Fatemeh 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 Wiecezarek 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 Góngora, Ignacio Sastre, Santiago Robaina, Ignacio Remersaro, Luis Chiruzzo and Aiala Rosá	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>	
Sofía Correa Busquets, Valentina Córdova Véliz 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 Ikhlasul 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

Program

Thursday, July 31, 2025

09:00 - 10:30 *Tutorial Session A*

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Tutorial Session B*

12:30 - 14:00 *Lunch Break / Birds of a Feather*

14:00 - 15:30 *Oral Session A*

A Bayesian Approach to Inferring Prerequisite Structures and Topic Difficulty in Language Learning

Anh-Duc Vu, Jue Hou, Anisia Katinskaia, Ching-Fan Sheu and Roman Yangarber

Enhancing Arabic Automated Essay Scoring with Synthetic Data and Error Injection

Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash and Ted Briscoe

Alignment Drift in CEFR-prompted LLMs for Interactive Spanish Tutoring

Mina Almasi and Ross Kristensen-McLachlan

You Shall Know a Word's Difficulty by the Family It Keeps: Word Family Features in Personalised Word Difficulty Classifiers for L2 Spanish

Jasper Degraeuwe

Assessing Critical Thinking Components in Romanian Secondary School Textbooks: A Data Mining Approach to the ROTEX Corpus

Madalina Chitez, Liviu Dinu, Marius Micluta-Campeanu, Ana-Maria Bucur and Roxana Rogobete

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

Felipe Urrutia, Cristian Buc, Roberto Araya and Valentin Barriere

15:30 - 16:00 *Coffee Break*

16:00 - 17:30 *Poster Session A*

Thursday, July 31, 2025 (continued)

A Survey on Automated Distractor Evaluation in Multiple-Choice Tasks

Luca Benedetto, Shiva Taslimipour and Paula Buttery

Increasing the Generalizability of Similarity-Based Essay Scoring Through Cross-Prompt Training

Marie Bexte, Yuning Ding and Andrea Horbach

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

Kordula De Kuthy, Leander Gierbach and Detmar Meurers

Automated Scoring of a German Written Elicited Imitation Test

Mihail Chifligarov, Jammila Laâguidi, Max Schellenberg, Alexander Dill, Anna Timukova, Anastasia Drackert and Ronja Laarmann-Quante

Challenges for AI in Multimodal STEM Assessments: a Human-AI Comparison

Aymeric de Chillaz, Anna Sotnikova, Patrick Jermann and Antoine Bosselut

Don't Score too Early! Evaluating Argument Mining Models on Incomplete Essays

Nils-Jonathan Schaller, Yuning Ding, Thorben Jansen and Andrea Horbach

LangEye: Toward 'Anytime' Learner-Driven Vocabulary Learning From Real-World Objects

Mariana Shimabukuro, Deval Panchal and Christopher Collins

Explaining Holistic Essay Scores in Comparative Judgment Assessments by Predicting Scores on Rubrics

Michiel De Vrindt, Renske Bouwer, Wim Van Den Noortgate, Marije Lesterhuis and Anaïs Tack

Name of Thrones: How Do LLMs Rank Student Names in Status Hierarchies Based on Race and Gender?

Annabella Sakunkoo and Jonathan Sakunkoo

Enhancing Security and Strengthening Defenses in Automated Short-Answer Grading Systems

Sahar Yarmohammadtoosky, Yiyun Zhou, Victoria Yaneva, Peter Baldwin, Saed Rezayi, Brian Clauser and Polina Harik

EduCSW: Building a Mandarin-English Code-Switched Generation Pipeline for Computer Science Learning

Ruishi Chen and Yiling Zhao

Thursday, July 31, 2025 (continued)

Paragraph-level Error Correction and Explanation Generation: Case Study for Estonian

Martin Vainikko, Taavi Kamarik, Karina Kert, Krista Liin, Silvia Maine, Kais Allkivi, Annekatrin Kaivapalu and Mark Fishel

Can LLMs Reliably Simulate Real Students' Abilities in Mathematics and Reading Comprehension?

KV Aditya Srivatsa, Kaushal Maurya and Ekaterina Kochmar

Transformer Architectures for Vocabulary Test Item Difficulty Prediction

Lucy Skidmore, Mariano Felice and Karen Dunn

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

Hakyung Sung, Karla Csuros and Min-Chang Sung

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

Marius Dumitran, Mihnea Buca and Theodor Moroianu

Investigating Methods for Mapping Learning Objectives to Bloom's Revised Taxonomy in Course Descriptions for Higher Education

Zahra Kolagar, Frank Zalkow and Alessandra Zarcone

Using NLI to Identify Potential Collocation Transfer in L2 English

Haiyin Yang, Zoey Liu and Stefanie Wulff

Improving In-context Learning Example Retrieval for Classroom Discussion Assessment with Re-ranking and Label Ratio Regulation

Nhat Tran, Diane Litman, Benjamin Pierce, Richard Correnti and Lindsay Clare Matsumura

Comparing Behavioral Patterns of LLM and Human Tutors: A Population-level Analysis with the CIMA Dataset

Aayush Kucheria, Nitin Sawhney and Arto Hellas

UPSC2M: Benchmarking Adaptive Learning from Two Million MCQ Attempts

Kevin Shi and Karttikeya Mangalam

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

Thursday, July 31, 2025 (continued)

Automatic Generation of Inference Making Questions for Reading Comprehension Assessments

Wanjing (Anyu) Ma, Michael Flor and Zuowei Wang

Lessons Learned in Assessing Student Reflections with LLMs

Mohamed Elaraby and Diane Litman

Automated L2 Proficiency Scoring: Weak Supervision, Large Language Models, and Statistical Guarantees

Aitor Arronte Alvarez and Naiyi Xie Fincham

Advances in Auto-Grading with Large Language Models: A Cross-Disciplinary Survey

Tania Amanda Nkoyo Frederick Eneye, Chukwuebuka Fortunate Ijezue, Ahmad Imam Amjad, Maaz Amjad, Sabur Butt and Gerardo Castañeda-Garza

Exploring LLMs for Predicting Tutor Strategy and Student Outcomes in Dialogues

Fareya Ikram, Alexander Scarlatos and Andrew Lan

Temporalizing Confidence: Evaluation of Chain-of-Thought Reasoning with Signal Temporal Logic

Zhenjiang Mao, Artem Bisliouk, Rohith Nama and Ivan Ruchkin

18:00 - 21:00

Workshop Dinner

Friday, August 1, 2025

09:00 - 09:45 *Keynote Talk by Kostia Omelianchuk*

09:45 - 10:30 *Oral Session B*

LLMs in alliance with Edit-based models: advancing In-Context Learning for Grammatical Error Correction by Specific Example Selection
Alexey Sorokin and Regina Nasyrova

Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors
Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack and Justin Vasselli

MSA at BEA 2025 Shared Task: Disagreement-Aware Instruction Tuning for Multi-Dimensional Evaluation of LLMs as Math Tutors
Baraa Hikal, Mohmaed Basem, Islam Oshallah and Ali Hamdi

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Poster Session B*

Leveraging Generative AI for Enhancing Automated Assessment in Programming Education Contests
Stefan Dascalescu, Marius Dumitran and Mihai Alexandru Vasiluta

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

Towards a Real-time Swedish Speech Analyzer for Language Learning Games: A Hybrid AI Approach to Language Assessment
Tianyi Geng and David Alfter

LEVOS: Leveraging Vocabulary Overlap with Sanskrit to Generate Technical Lexicons in Indian Languages
Karthika N J, Krishnakant Bhatt, Ganesh Ramakrishnan and Preethi Jyothi

The Need for Truly Graded Lexical Complexity Prediction
David Alfter

Friday, August 1, 2025 (continued)

Educators' Perceptions of Large Language Models as Tutors: Comparing Human and AI Tutors in a Blind Text-only Setting

Sankalan Pal Chowdhury, Terry Jingchen Zhang, Donya Rooein, Dirk Hovy, Tanja Käser and Mrinmaya Sachan

Costs and Benefits of AI-Enabled Topic Modeling in P-20 Research: The Case of School Improvement Plans

Syeda Sabrina Akter, Seth Hunter, David Woo and Antonios Anastasopoulos

Are Large Language Models for Education Reliable Across Languages?

Vansh Gupta, Sankalan Pal Chowdhury, Vilém Zouhar, Donya Rooein and Mrinmaya Sachan

Span Labeling with Large Language Models: Shell vs. Meat

Phoebe Mulcaire and Nitin Madnani

STAIR-AIG: Optimizing the Automated Item Generation Process through Human-AI Collaboration for Critical Thinking Assessment

Euigyum Kim, Seewoo Li, Salah Khalil and Hyo Jeong Shin

End-to-End Automated Item Generation and Scoring for Adaptive English Writing Assessment with Large Language Models

Kamel Nebhi, Amrita Panesar and Hans Bantilan

bea-jh at BEA 2025 Shared Task: Evaluating AI-powered Tutors through Pedagogically-Informed Reasoning

Jihyeon Roh and Jinhyun Bang

K-NLPers at BEA 2025 Shared Task: Evaluating the Quality of AI Tutor Responses with GPT-4.1

Geon Park, Jiwoo Song, Gihyeon Choi, Juoh Sun and Harksoo Kim

IALab UC at BEA 2025 Shared Task: LLM-Powered Expert Pedagogical Feature Extraction

Sofía Correa Busquets, Valentina Córdova Véliz and Jorge Baier

TBA at BEA 2025 Shared Task: Transfer-Learning from DARE-TIES Merged Models for the Pedagogical Ability Assessment of LLM-Powered Math Tutors

Sebastian Gombert, Fabian Zehner and Hendrik Drachsler

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

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

Friday, August 1, 2025 (continued)

Analyzing Interview Questions via Bloom's Taxonomy to Enhance the Design Thinking Process

Fatemeh Kazemi Vanhari, Christopher Anand and Charles Welch

Exploring LLM-Based Assessment of Italian Middle School Writing: A Pilot Study

Adriana Mirabella and Dominique Brunato

Beyond Linear Digital Reading: An LLM-Powered Concept Mapping Approach for Reducing Cognitive Load

Junzhi Han and Jinho D. Choi

BLCU-ICALL at BEA 2025 Shared Task: Multi-Strategy Evaluation of AI Tutors

Jiyuan An, Xiang Fu, Bo Liu, Xuquan Zong, Cunliang Kong, Shuliang Liu, Shuo Wang, Zhenghao Liu, Liner Yang, Hanghang Fan and Erhong Yang

Jinan Smart Education at BEA 2025 Shared Task: Dual Encoder Architecture for Tutor Identification via Semantic Understanding of Pedagogical Conversations

Lei Chen

CU at BEA 2025 Shared Task: A BERT-Based Cross-Attention Approach for Evaluating Pedagogical Responses in Dialogue

Zhihao Lyu

SYSUpporter Team at BEA 2025 Shared Task: Class Compensation and Assignment Optimization for LLM-generated Tutor Identification

Longfeng Chen, Zeyu Huang, Zheng Xiao, Yawen Zeng and Jin Xu

Emergent Wisdom at BEA 2025 Shared Task: From Lexical Understanding to Reflective Reasoning for Pedagogical Ability Assessment

Raunak Jain and Srinivasan Rengarajan

Henry at BEA 2025 Shared Task: Improving AI Tutor's Guidance Evaluation Through Context-Aware Distillation

Henry Pit

TutorMind at BEA 2025 Shared Task: Leveraging Fine-Tuned LLMs and Data Augmentation for Mistake Identification

FATIMA DEKMAK, Christian Khairallah and Wissam Antoun

BD at BEA 2025 Shared Task: MPNet Ensembles for Pedagogical Mistake Identification and Localization in AI Tutor Responses

Shadman Rohan, Ishita Sur Apan, Muhtasim Shochcho, Md Fahim, Mohammad Rahman, AKM Mahbubur Rahman and Amin Ali

Friday, August 1, 2025 (continued)

LLM-Assisted, Iterative Curriculum Writing: A Human-Centered AI Approach in Finnish Higher Education

Leo Huovinen and Mika Hämäläinen

12:30 - 14:00 *Lunch Break / Birds of a Feather*

14:00 - 15:30 *Poster Session C*

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

Adapting LLMs for Minimal-edit Grammatical Error Correction

Ryszard Staruch, Filip Gralinski and Daniel Dzienisiewicz

Do LLMs Give Psychometrically Plausible Responses in Educational Assessments?

Andreas Säuberli, Diego Frassinelli and Barbara Plank

Towards Automatic Formal Feedback on Scientific Documents

Louise Bloch, Johannes Rückert and Christoph Friedrich

Transformer-Based Real-Word Spelling Error Feedback with Configurable Confusion Sets

Torsten Zesch, Dominic Gardner and Marie Bexte

Unsupervised Sentence Readability Estimation Based on Parallel Corpora for Text Simplification

Rina Miyata, Toru Urakawa, Hideaki Tamori and Tomoyuki Kajiwara

Exploiting the English Vocabulary Profile for L2 word-level vocabulary assessment with LLMs

Stefano Banno, Kate Knill and Mark Gales

Improving AI assistants embedded in short e-learning courses with limited textual content

Jacek Marciniak, Marek Kubis, Michał Gulczyński, Adam Szpilkowski, Adam Wiczarek and Marcin Szczepański

GermDetect: Verb Placement Error Detection Datasets for Learners of Germanic Languages

Noah-Manuel Michael and Andrea Horbach

Friday, August 1, 2025 (continued)

Automated Scoring of Communication Skills in Physician-Patient Interaction: Balancing Performance and Scalability

Saed Rezayi, Le An Ha, Yiyun Zhou, Andrew Houriet, Angelo D’Addario, Peter Baldwin, Polina Harik, Ann King and Victoria Yaneva

Can GPTZero’s AI Vocabulary Distinguish Between LLM-Generated and Student-Written Essays?

Veronica Schmalz and Anaïs Tack

A Framework for Proficiency-Aligned Grammar Practice in LLM-Based Dialogue Systems

Luisa Ribeiro-Flucht, Xiaobin Chen and Detmar Meurers

RETUYT-INCO at BEA 2025 Shared Task: How Far Can Lightweight Models Go in AI-powered Tutor Evaluation?

Santiago Góngora, Ignacio Sastre, Santiago Robaina, Ignacio Remersaro, Luis Chiruzzo and Aiala Rosá

Archaeology at BEA 2025 Shared Task: Are Simple Baselines Good Enough?

Ana Roşu, Jany-Gabriel Ispas and Sergiu Nisioi

NLIP at BEA 2025 Shared Task: Evaluation of Pedagogical Ability of AI Tutors

Trishita Saha, Shrenik Ganguli and Maunendra Sankar Desarkar

LLM-based post-editing as reference-free GEC evaluation

Robert Östling, Murathan Kurfali and Andrew Caines

Estimation of Text Difficulty in the Context of Language Learning

Anisia Katinskaia, Anh-Duc Vu, Jue Hou, Ulla Vanhatalo, Yiheng Wu and Roman Yangarber

Exploring task formulation strategies to evaluate the coherence of classroom discussions with GPT-4o

Yuya Asano, Beata Beigman Klebanov and Jamie Mikeska

EyeLLM: Using Lookback Fixations to Enhance Human-LLM Alignment for Text Completion

Astha Singh, Mark Torrance and Evgeny Chukharev

Decoding Actionability: A Computational Analysis of Teacher Observation Feedback

Mayank Sharma and Jason Zhang

Friday, August 1, 2025 (continued)

Thapar Titan/s : Fine-Tuning Pretrained Language Models with Contextual Augmentation for Mistake Identification in Tutor–Student Dialogues

Harsh Dadwal, Sparsh Rastogi and Jatin Bedi

Wonderland_EDU@HKU at BEA 2025 Shared Task: Fine-tuning Large Language Models to Evaluate the Pedagogical Ability of AI-powered Tutors

Deliang Wang, Chao Yang and Gaowei Chen

BJTU at BEA 2025 Shared Task: Task-Aware Prompt Tuning and Data Augmentation for Evaluating AI Math Tutors

Yuming Fan, Chuangchuang Tan and Wenyu Song

SmolLab_SEU at BEA 2025 Shared Task: A Transformer-Based Framework for Multi-Track Pedagogical Evaluation of AI-Powered Tutors

Md. Abdur Rahman, MD AL AMIN, Sabik Aftahee, Muhammad Junayed and Md Ashiqur Rahman

LexiLogic at BEA 2025 Shared Task: Fine-tuning Transformer Language Models for the Pedagogical Skill Evaluation of LLM-based tutors

Souvik Bhattacharyya, Billodal Roy, Niranjana M and Pranav Gupta

DLSU at BEA 2025 Shared Task: Towards Establishing Baseline Models for Pedagogical Response Evaluation Tasks

Maria Monica Manlises, Mark Edward Gonzales and Lanz Lim

LookAlike: Consistent Distractor Generation in Math MCQs

Nisarg Parikh, Alexander Scarlatos, Nigel Fernandez, Simon Woodhead and Andrew Lan

From End-Users to Co-Designers: Lessons from Teachers

Martina Galletti and Valeria Cesaroni

15:30 - 16:00 *Coffee Break*

16:00 - 17:15 *Oral Session C*

Down the Cascades of Omethi: Hierarchical Automatic Scoring in Large-Scale Assessments

Fabian Zehner, Hyo Jeong Shin, Emily Kerzabi, Andrea Horbach, Sebastian Gombert, Frank Goldhammer, Torsten Zesch and Nico Andersen

Direct Repair Optimization: Training Small Language Models For Educational Program Repair Improves Feedback

Charles Koutchme, Nicola Dainese and Arto Hellas

Friday, August 1, 2025 (continued)

Advancing Question Generation with Joint Narrative and Difficulty Control

Bernardo Leite and Henrique Lopes Cardoso

Intent Matters: Enhancing AI Tutoring with Fine-Grained Pedagogical Intent Annotation

Kseniia Petukhova and Ekaterina Kochmar

LLMs Protégés: Tutoring LLMs with Knowledge Gaps Improves Student Learning Outcome

Andrei Kucharavy, Cyril Vallez and Dimitri Percia David

17:15 - 17:30 *Closing Remarks*