

BEA 2026

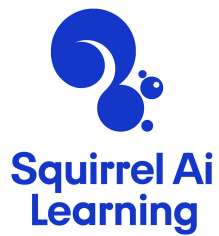
**The 21st Workshop on Innovative Use of NLP for Building
Educational Applications**

Proceedings of the Workshop

July 3-4, 2026

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

Gold Level



©2026 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 979-8-89176-409-5

Introduction

This year marks the 21st edition of the *Workshop on Innovative Use of NLP for Building Educational Applications*. As usual, we are happy to welcome papers 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, BEA continues being a 2-day workshop, with one significant change: we have one hybrid day and one fully virtual day. This change did not only allow us to accept more valuable work from our community, but also ensured better visibility for authors presenting online.

This year, we have received a total of 132 submissions, and from these, we have accepted 19 papers as talks and 44 as poster and demo presentations, for an overall acceptance rate of 48 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, with various languages, subject domains, and educational applications covered.

In addition to the diverse oral, poster and demo presentations, this year, we are hosting a panel on *Transitioning from Academia to the EdTech Industry*, with four distinguished speakers involved: Christine Bagarino (Social AI), Kai North (Cambium Assessment), Keelan Evanini (NBME), and Mariano Felice (British Council). BEA 2026 will also include, for the second time, a half-day tutorial on *Theory of Mind and Application in Educational Context*, organized by Effat Farhana (Auburn University), Maha Zainab (Auburn University), Qiaosi Wang (Carnegie Mellon University), Niloofar Mireshghallah (Carnegie Mellon University), Ramira van der Meulen (Leiden University), Max van Duijn (Leiden University). Last but not least, this year we have also hosted two shared tasks – on *Vocabulary Difficulty Prediction for English Learners* by Mariano Felice (British Council) and Lucy Skidmore (British Council), and on *Rubric-based Short Answer Scoring for German* by Sebastian Gombert (DIPF), Zhifan Sun (DIPF), Fabian Zehner (DIPF), Jannik Lossjew (IPN), Tobias Wyrwich (IPN), Berrit Katharina Czinczel (IPN), David Bednorz (IPN), Sascha Bernholt (IPN), Knut Neumann (IPN), Ute Harms (IPN), Aiso Heinze (IPN), and Hendrik Drachler (DIPF). Both tasks attracted a large number of participants, and the program includes oral presentations on the shared tasks from the organizers as well as talks from the tasks' winners and extended poster sessions for shared tasks participants presenting their systems.

Finally, 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, Squirrel Ai Learning, CATALPA, and British Council.

BEA 2026 Organizing Committee

Organizing Committee

General Chair

Ekaterina Kochmar, MBZUAI / University of Victoria

Program Chairs

Andrea Horbach, CAU Kiel / Leibniz Institute for Science and Mathematics Education

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 Chairs

Zheng Yuan, University of Sheffield

Jill Burstein, Duolingo

Virtual Chair

Stefano Bannò, Cambridge University

Program Committee

Chairs

Bashar Alhafni, MBZUAI
Stefano Bannò, University of Cambridge
Marie Bexte, FernUniversität in Hagen
Jill Burstein, Duolingo
Andrea Horbach, CAU Kiel / Leibniz Institute for Science and Mathematics Education
Ekaterina Kochmar, MBZUAI / University of Victoria
Ronja Laarmann-Quante, Ruhr University Bochum
Anais Tack, KU Leuven; imec
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
Soroosh Akef, Leibniz-Institut für Wissensmedien
Syeda Sabrina Akter, George Mason University
Vennela Akula, Auburn University
Erfan Al-Hossami, University of North Carolina at Charlotte
Ali Al-Laith, University of Copenhagen
Giora Alexandron, Weizmann Institute of Science
David Alfter, Gothenburg University
Norah Almousa, University of Pittsburgh
Aitor Arronte Alvarez, University of Hawaii at Manoa
Nischal Ashok Kumar, University of Massachusetts Amherst
Sina Bagheri Nezhad, Independent Researcher
Xiaoyu Bai, University of Potsdam
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, Telecom SudParis, IP Paris
Maryam Berijanlian, Michigan State University
Ummugul Bezirhan, Boston College, TIMSS and PIRLS International Study Center
Abhidip Bhattacharyya, University of Massachusetts, Amherst
Ted Briscoe, MBZUAI
Dominique Brunato, Institute of Computational Linguistics A. Zampolli (ILC-CNR), Pisa
Andrew Caines, University of Cambridge
Jie Cao, University of Oklahoma
Dumitru-Clementin Cercel, University Politehnica of Bucharest
Bharathi Raja Chakravarthi, University of Galway, Ireland
Imran Chamieh, Hochschule Ruhr West
Ignatios Charalampidis, University of Tuebingen
Andreas Chari, University of Glasgow

Mei-Hua Chen, Department of Foreign Languages and Literature, Tunghai University
Artem Chernodub, Zendesk
Luis Chiruzzo, Universidad de la Republica
Hyundong Cho, USC, Information Sciences Institute
Declan Clowry, Social AI
Longwei Cong, DIPF | Leibniz Institute for Research and Information in Education
Yan Cong, Purdue University
Mark Core, University of Southern California
Steven Coyne, Tohoku University / RIKEN
Orphee De Clercq, LT3, Ghent University
Kordula De Kuthy, IWM Tübingen
Jasper Degraeuwe, Ghent University
Carrie Demmans Epp, University of Alberta
Yuning Ding, Leibniz Institute for Science and Mathematics Education
Rahul Divekar, Bentley University
George Duenas, Universidad Pedagogica Nacional
Marius Dumitran, University of Bucharest
Andreea Dutulescu, National University of Science and Technology Politehnica Bucharest
Yo Ehara, Tokyo Gakugei University
Walid El Hefny, Leibniz-Institut für Wissensmedien (IWM)
Effat Farhana, Auburn University
Mariano Felice, British Council
Michael Flor, Educational Testing Service
Jennifer-Carmen Frey, EURAC Research
Momoka Furuhashi, Tohoku University / National Institute of Informatics
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
Lingyu Gao, Duolingo
Jie Gao, McGill University
Rujun Gao, Texas A&M University
Ritik Garg, Extramarks Education Pvt. Ltd.
Voula Giouli, Aristotle University of Thessaloniki / ILSP, ATHENA RC
Aron Gohr, Independent
Sebastian Gombert, DIPF | Leibniz Institute for Research and Information in Education
Kiel Gonzales, University of the Philippines Diliman
Mark Edward Gonzales, Agency for Science, Technology and Research (A*STAR), Singapore
Cyril Goutte, National Research Council Canada
Stuart Grey, University of Glasgow
Viktor Gulbrandsen, University of Oslo
Pranav Gupta, Cisco
Abigail Gurin Schleifer, Weizmann Institute of Science
Ibrahim Hallac, Østfold University of Applied Sciences
Ching Nam Hang, Assistant Professor, Yam Pak Charitable Foundation School of Computing and Information Sciences, Saint Francis University, Hong Kong
Jiangang Hao, Educational Testing Service
Hasnain Heickal, University of Massachusetts Amherst
Nicolas Hernandez, Nantes University
Jingying Hu, Purdue University
Chung-Chi Huang, Frostburg State University

Aiden Huang, Acton-Boxborough Regional High School
Phoebe Huang, University of Florida
Anna Huelsing, CAU
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
Md. Rakibul Islam, Ahsanullah University of Science and Technology
Elsayed Issa, Purdue University
Raunak Jain, Intuit
Abhishek Jariwala, Auburn University
Helen Jin, University of Pennsylvania
Léane Jourdan, Nantes University
Indika Kahanda, University of North Florida
Tomoyuki Kajiwara, Ehime University / The University of Osaka
Anisia Katinskaia, Swiss Data Science Center
Fateme Kazemi Vanhari, McMaster University
Elma Kerz, Exaia Technologies
Samin Khan, Stanford University
Sachiko Kimura, Rikkyo University, Dokkyo University
Mamoru Komachi, Hitotsubashi University
Maxim Konca, Goethe University Frankfurt
Katsunori Kotani, Kansai Gaidai 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
Ravi Kumar, Florida International University (FIU)
Alexander Kwako, University of California, Los Angeles
Antonio Laverghetta Jr., Pennsylvania State University
Jaewook Lee, UMass Amherst
Seolhwa Lee, Technical University of Darmstadt
Bernardo Leite, Faculty of Engineering - University of Porto
Arun Balajiee Lekshmi Narayanan, University of Pittsburgh
Hannah Levin, Stanford University
Jiazheng Li, King's College London
Xu Li, Southwest Petroleum University
Junhong Liang, MBZUAI
Wen Liang, Columbia University
Chuan-Jie Lin, National Taiwan Ocean University
Yudong Liu, Western Washington University
Zhexiong Liu, University of Pittsburgh
Denise Loefflad, Leibniz-Institut fuer Wissensmedien Tuebingen
Julian Lohmann, Christian Albrechts Universit t Kiel
Benny Longwill, Educational Testing Service
Anastassia Loukina, Grammarly Inc
Wanning Lu, Stanford University
Li Lucy, University of Washington
Sarah L ber, University of T bingen
Jakub Macina, ETH Zurich

Nitin Madnani, Duolingo
S.M Afridi Mahmud, Bangladesh University of Engineering and Technology
Maria Monica Manlises, De La Salle University
Zhenjiang Mao, University of Florida
Jacek Marciniak, Adam Mickiewicz University
Arianna Masciolini, Språkbanken Text; Department of Swedish, Multilingualism, Language Technology; University of Gothenburg
Sandeep Mathias, Presidency University
Hunter McNichols, University of Massachusetts Amherst
Detmar Meurers, Leibniz-Institut für Wissensmedien (IWM)
Daniel Mora Melanchthon, CAU Kiel / Leibniz Institute for Science and Mathematics Education
Phoebe Mulcaire, Duolingo
Laura Musto, Universidad de la Republica
Farah Nadeem, LUMS
Sungjin Nam, ACT, Inc
Diane Napolitano, The Washington Post
Léo Nebel, LIP6 - Sorbonne Université
Seyed Parsa Neshaei, EPFL
Duy Anh Nguyen, Independent Researcher
Huy Nguyen, Oracle
S Jaya Nirmala, National Institute of Technology Tiruchirappalli
Sergiu Nisioi, Human Language Technologies Research Center, University of Bucharest
Adam Nohejl, RIKEN
Kai North, Cambium Assessment Inc.
Similoluwa Okunowo, AIMS South Africa
Kostiantyn Omelianchuk, Grammarly
Amin Omidvar, PhD student at the Department of Electrical Engineering and Computer Science, York University
Daniel Oyeniran, University of Alabama
Alonso Palomino, DFKI (German Center for Artificial Intelligence Research)
Nisarg Parikh, University of Massachusetts, Amherst
Jeiyeon Park, Smilegate
Kseniia Petukhova, MBZUAI
Tomasz Pilka, Adam Mickiewicz University in Poznan
Pagnarith Pit, University of Melbourne
Lucie Polakova, Charles University
Long Qin, Alibaba
Mengyang Qiu, Saint Elizabeth University
Chatrine Qwaider, MBZUAI
Md. Abdur Rahman, Islamic University of Technology (IUT), OIC
Pranil Raichura, Independent Researcher
Anmol Rao, Arizona State University
pranshu rastogi, Independent Researcher
Sparsh Rastogi, Thapar Institute of Engineering and Technology
Manav Rathod, University of California, Berkeley
Hanumant Redkar, Goa University, Goa
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
Aiala Rosá, Instituto de Computación, Facultad de Ingeniería, Universidad de la República

Mahule Roy, University of Oxford
Alla Rozovskaya, Queens College, City University of New York
Josef Ruppenhofer, Fernuniversität in Hagen
Mariam Saeed, Applied Innovation Center
Pravish Sainath, Vanier College, Université de Montréal
Jonathan Sakunkoo, University of Oxford
Alexander Scarlatos, University of Massachusetts Amherst
Nils-Jonathan Schaller, Leibniz Institute for Science and Mathematics Education
Stephanie Schoch, University of Virginia
Dhruv Shah, Sardar Patel Institute of Technology
Natarajan Balaji Shankar, University of California Los Angeles
Matthew Shardlow, Manchester Metropolitan University
Qian Shen, University of Florida
Gyu-Ho Shin, University of Illinois Chicago
Hyo Jeong Shin, Sogang University
Ziqi Shu, Stanford University
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)
Alexey Sorokin, Moscow State University
Shankhalika Srikanth, Independent Researcher
Felix Stahlberg, Google Research
Katherine Stasaski, Google DeepMind
Helmer Strik, Centre for Language and Speech Technology (CLST), Centre for Language Studies (CLS), Radboud University Nijmegen
Parisa Suchdev, University of Vermont
Hakyung Sung, Rochester Institute of Technology
Andreas Säuberli, LMU Munich
Kyosuke Takami, Osaka Kyoiku University
Wenjia Tan, University of Macau
CheeWei Tan, Nanyang Technological University
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
Hariram Veeramani, UCLA
Giulia Venturi, Institute of Computational Linguistics Antonio Zampolli (ILC-CNR)
Anthony Verardi, Duolingo
Elena Volodina, University of Gothenburg
Anh-Duc Vu, University of Helsinki
Deliang Wang, The University of Hong Kong
Nikhil Wani, OpenThreads AI / University of Southern California
Taro Watanabe, Nara Institute of Science and Technology
Alistair Willis, The Open University
Yiheng Wu, University of Helsinki
Bushi Xiao, University of Florida
Hiroaki Yamada, Tokyo Metropolitan University
Bi-Cheng Yan, National Taiwan Normal University

Haiyin Yang, University of Florida
Roman Yangarber, University of Helsinki
Sahar Yarmohammadtoosky, NBME
Tahreem Yasir, North Carolina State University
Su-Youn Yoon, EduLab
Marcos Zampieri, George Mason University
Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education
Torsten Zesch, Computational Linguistics, FernUniversität in Hagen
Jing Zhang, Emory University
Yiling Zhao, Stanford University
Yang Zhong, University of Pittsburgh
Ej Zhou, University of Cambridge
Yiyun Zhou, NBME
Jessica Zipf, University of Konstanz
Michael Zock, CNRS-LIS
Robert Östling, Department of Linguistics, Stockholm University

Keynote Talk

Transitioning from Academia to the EdTech Industry

Christine Bagarino
Social AI

Bio: Christine Bagarino is the founder of Social AI, an AI native startup based in Tokyo, Japan, that builds AI technology for positive social impact. A former educator, curriculum developer, and Silicon Valley product manager, Christine developed Japan's first LLM-powered English learning chatbot approved for K12 public education. In 2023, she traveled to rural Vietnam for an EdTech pilot funded by the Asian Development Bank, which inspired her to make fun and quality English learning accessible to children all over the world.

Keynote Talk

Transitioning from Academia to the EdTech Industry

Kai North
Cambium Assessment

Bio: Dr. Kai North is a Senior Data Scientist at Cambium Assessment Inc. Kai earned his PhD in Information Technology at George Mason University, Virginia. He specializes in machine learning and natural language processing applied to educational and assessment technologies and has multiple top-tier publications at venues such as ACL, COLING, and ACM computing surveys.

Keynote Talk

Transitioning from Academia to the EdTech Industry

Keelan Evanini
NBME

Bio: Keelan Evanini has been a researcher in the field of natural language processing (NLP) for educational and conversational applications for more than 15 years. He is currently a Lead NLP Scientist at the National Board of Medical Examiners, where he conducts research into using conversational AI for interactive, scenario-based learning applications for medical students. Prior to joining NBME in 2025, he served as SVP of engineering and AI at Kasisto, a New York-based startup company that provides conversational AI solutions to financial institutions. From 2009-2020, he worked at Educational Testing Service as a research scientist and research director, where he conducted research into automated scoring of non-native spoken English for language proficiency assessment and spoken dialog systems for language learning applications. Keelan received a PhD in linguistics from the University of Pennsylvania in 2009 and is a Senior Member of the IEEE. He has published over 90 peer-reviewed papers and has been awarded 11 US patents.

Keynote Talk

Transitioning from Academia to the EdTech Industry

Mariano Felice
British Council

Bio: Mariano leads the artificial intelligence (AI) strategy for language learning and assessment at the British Council. His role involves researching the application of natural language processing (NLP) to language assessment, providing strategic guidance for the development and adoption of AI solutions, and promoting AI literacy and responsible use of new technologies.

With over a decade of experience, Mariano has worked on a wide range of topics, including grammatical error correction, automatic error typing, system evaluation, automated cloze test generation and item difficulty prediction. He has published numerous scientific papers in top-tier NLP conferences and is a regular speaker at international conferences as well as a reviewer for workshops, journals and conferences in his field. Mariano is also a visiting scholar at the University of Reading and the University of Cambridge.

Table of Contents

<i>Theory of Mind and Application in Educational Context</i>	
Effat Farhana, Maha Zainab, Qiaosi Wang, Niloofar Mireshghallah, Ramira van der Meulen and Max van Duijn	1
<i>Inferring Student Engagement via Real-Time Thermal–Visual Voice Activity Detection</i>	
Bradley Goodman	7
<i>Investigating Context-aware CTC for Pronunciation Assessment: Mitigating Peaky Behavior and Context Independency Assumption</i>	
Jiun-Ting Li, Tien-Hong Lo, Bi-Cheng Yan, Shih-Hsuan Chiu, Fu-An Chao and Berlin Chen	21
<i>A Survey of Automated Presentation Coaching: Systems, Methods, and Open Challenges</i>	
Wen Liang, Li Siyan, Zackary Rackauckas and Julia Hirschberg	34
<i>The Aftermath of DrawEduMath: Vision Language Models Underperform with Struggling Students and Misdiagnose Errors</i>	
Li Lucy, Albert Zhang, Nathan Anderson, Ryan Knight and Kyle Lo	48
<i>Criterial Features in German: Towards Interpretable NLP in Readability Assessment</i>	
Denise Loefflad, Sofia Kathmann, Heiko Holz and Detmar Meurers	63
<i>Letting Tutor Personas Speak Up for LLMs: Learning Steering Vectors from Dialogue via Preference Optimization</i>	
Jaewook Lee, Alexander Scarlatos, Simon Woodhead and Andrew Lan	78
<i>Towards Just-in-Time Adaptive Feedback: Enhancing Student Learning via Knowledge-Grounded LLM</i>	
Younghun Lee, Amir Bralin, Nobel Sanjay Rebello and Dan Goldwasser	93
<i>RABIT: Rationale-Based Distillation Towards Interpretable Automatic Speaking Assessment via a Small Language Model</i>	
Bi-Cheng Yan, Hong-Yun Lin, Fu-An Chao, Jiun-Ting Li and Berlin Chen	108
<i>Towards Pedagogically Aligned LLM Tutors for Math Mistake Remediation</i>	
Kseniia Petukhova, Tien Dat Nguyen and Ekaterina Kochmar	118
<i>Challenges in Machine Translation of Interactive Multimodal Exercises</i>	
Lucie Polakova, Miroslav Hrabal, Věra Kloudová, Michal Novák, Mariia Anisimova and Martin Popel	141
<i>Evaluating LLM Workflows for Generating Clinical Communication Assessment Items: A Comparative Study with Subject-Matter Experts</i>	
Christopher Runyon, Peter Baldwin, Ian Micir, Kevin Frome, Stephanie Mann, Saed Rezayi, Keelan Evanini and Victoria Yaneva	153
<i>Towards Self-Referential Analytic Assessment: A Profile-Based Approach to L2 Writing Evaluation with LLMs</i>	
Stefano Banno, Kate Knill and Mark Gales	174
<i>Using Interaction Log Data to Evaluate and Improve Feedback Accuracy in an Intelligent Language Tutoring System</i>	
Mariia Soliar, Leona Colling, Stephen Bodnar and Detmar Meurers	190
<i>A Bigger Catch: Fine-Grained Curriculum Standards Alignment on the MathFish Benchmark</i>	
Xinman Liu, Mayank Sharma and Xinyu Shi	208

<i>Through the Sentence Lens: Explainable Essay Scoring through Fine-Grained Predictions</i> Daniel Mora Melanchthon, Stefan Keller and Andrea Horbach	221
<i>Instruction-Following LLMs for Grammatical Error Correction: Analyzing Neutral-Anchored Instructional Sensitivity Across Editing Modes</i> Tolgahan Türker and Gülşen Eryiğit	234
<i>Assessing the Quality and Consistency of Automated Knowledge Component Generation using Instructor-generated Questions and LLMs</i> Jordan Esiason, Priyanka Khare, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng and Yeil Jeong	248
<i>Estimating LLM Grading Ability and Response Difficulty in Automatic Short Answer Grading via Item Response Theory</i> Longwei Cong, Sonja Hahn, Sebastian Gombert, Leon Camus, Hendrik Drachslers and Ulf Kroehne	259
<i>Using k-Shot Prompting with Large k for the Automated Scoring of a German Written Elicited Imitation Test</i> Malte Sternik, Ronja Laarmann-Quante and Anastasia Drackert	272
<i>Kelvi: A Morphological Parser to Support Tamil Literacy</i> Shankhalika Srikanth, Sabrina Yu, Sophia Chan and Madeline Solis de Ovando	281
<i>From Questions to Assessment Tuples: A Multi-Agent Framework with Bloom-Specialized Agents and Automated Verification</i> Gee-Lyle Wong, Runcong Zhao, Yulan He and Jiazheng Li	292
<i>Intent vs. Surface: Recovering Acoustic Realization from Modern ASR for Pronunciation Training</i> Seongjin Park	336
<i>KEYSCORE — Keystroke-enhanced Automated Essay Scoring</i> Nils-Jonathan Schaller, Daniel Mora Melanchthon, Thorben Jansen, Olaf Köller and Andrea Horbach	347
<i>EduMUSE: A Multimodal Educational Dataset with Automatically Extracted Instructional Context</i> Andreea Dutulescu, Stefan Ruseti, Mihai Dascalu and Danielle McNamara	358
<i>Opportunities and Challenges of LLMs in Education: An NLP Perspective</i> Sowmya Vajjala, Bashar Alhafni, Stefano Banno, Kaushal Maurya and Ekaterina Kochmar ..	368
<i>Fine-Grained Content Zone Prediction in German Argumentative Essays Using LLMs</i> Xiaoyu Bai and Manfred Stede	405
<i>Multi-step Large Language Model for Fine-Grained Feedback in Stepwise Linear Equation Solutions</i> Imran Chamieh, Torsten Zesch and Klaus Giebertmann	419
<i>Quality-Conditioned Agreement in Automated Short Answer Scoring: Mid-Range Degradation and the Impact of Task-Specific Adaptation</i> Abigail Gurin Schleifer, Moriah Ariely, Beata Beigman Klebanov, Asaf Salman and Giora Alexandron	432
<i>Using LLMs for item creation: Validating the potential of automatically generated sentence repetition test items for language assessment</i> Sarah Löber, Björn Rudzewitz, Yuan Chu, Mengyuan He, Shiqin Liu, Yushan Ye and Xiaobin Chen	443

<i>Comparative Evaluation of AI-Generated vs. Expert-written Answer Explanations for a Medical Education Self-Assessment</i>	
Yiyun Zhou, Francis O’Donnell and Victoria Yaneva	455
<i>What Aggregate Scores Hide: Per-Rule Evaluation of Russian Grammatical Error Correction</i>	
Anna Smirnova, Artyom Kopan, Vladislav Makeev and George Chernishev	463
<i>FinnGEC: Benchmarking Grammatical Error Correction for Finnish</i>	
Anh-Duc Vu, Mikhail Zolotilin, Jue Hou, Anisia Katinskaia, Yiheng Wu and Roman Yangarber	479
<i>From Metrics to Meaning: Rule-Grounded LLM Explanations for Data Literacy in the Case of Youth Football</i>	
Tomasz Piłka, Tomasz Kuczyński and Mateusz Czajka	492
<i>Sharing is Caring: Advantages of Sharing a Language Background with Learners as an Annotator of Learner Data in UD</i>	
Caroline Grand-Clement and Arianna Masciolini	503
<i>Classification of Student Struggle in Mathematics</i>	
Hannah Levin, Madhura Padwal and Nchimunya Mwiinga	513
<i>PERSA: Reinforcement Learning for Professor-Style Personalized Feedback with LLMs</i>	
Ravi Kumar, Utkarsh Grover, Xiaomin Lin and Agoritsa Polyzou	529
<i>Data-lean fine-tuning of models for evaluating teacher performance in a GenAI-led elicitation simulation</i>	
Beata Beigman Klebanov, Andrew Hoang, Jamie Mikeska, Benny Longwill, Sanjna Kashyap, Shreyashi Halder and Aakanksha Bhatia	546
<i>Multi-component student writing profiles for expert-aligned automated evaluation of English learner essays.</i>	
Russell Moore, Andrew Caines and Paula Buttery	562
<i>Policy-Sensitive Fairness Evaluation in Automated Scoring of Clinical Communication</i>	
Saed Rezayi, Le An Ha, Victoria Yaneva, Polina Harik, Janet Mee and Jason Snyder	574
<i>Noise Steering for Controlled Text Generation: Improving Diversity and Reading-Level Fidelity in Arabic Educational Story Generation</i>	
Haziq Khalid, Salsabeel Shapsough and Imran Zualkernan	581
<i>The Effects of Structured LLM-Generated Feedback on Programming Assignment Performance</i>	
Tsvetomila Mihaylova, Evanfiya Logacheva, Arto Hellas, Jing Fan, Francisco Castro, Bitá Akram, Narges Norouzi, Peter Brusilovsky and Juho Leinonen	598
<i>Interpretable Difficulty-Aware Knowledge Tracing in Tutor-Student Dialogues</i>	
Shuyan Huang, Alexander Scarlatos, Jaewook Lee and Andrew Lan	612
<i>Rubrics as Semantic Subspaces: A Unified Approach to Rubric-based Constructed Response Scoring across Short Answers and Essays</i>	
Sebastian Gombert, Sonja Hahn, Nico Andersen, Leon Camus, Zhifan Sun, Ngoc Nhu Hao Nguyen, Fabian Zehner, Longwei Cong, Alexander Mehler and Hendrik Drachslér	624
<i>Domain-Adaptive Pre-training for Automated Short Answer Grading in Conceptual Physics: Reliability, Question-Level Analysis, and Error Reduction</i>	
Shirin Lade, Alistair Willis, Jonathan Nylk and Oli Howson	635

<i>Measuring Optimal Challenge: Trajectory-Based Difficulty Alignment in Open-Ended Language Tutoring</i>	Ziqi Shu, Shuman Wang and Michael Hardy	651
<i>PeerMathDial: A Middle School Dialogue Dataset for Student Collaborative Math Problem Solving</i>	Murong Yue, Desmond Mcglone, Emily Slutz, Wenhan Lyu, Yixuan Zhang, Jennifer Suh and Ziyu Yao	668
<i>Effects of Varying LLM Access on Essay Writing Behavior</i>	Julia Christenson, Karin de Langis, Shirley Anugrah Hayati and Dongyeop Kang	685
<i>Assessment of L2 speech global dimensions using large audio language models</i>	Elsayed Issa and Mahmoud Ali	702
<i>Incentives Of EdTech: A Systematic Review Of EduNLP Research</i>	Gabrielle Gaudeau, Aoife O’Driscoll, Jasper Degraeuwe, Andrew Caines, Donya Rooein and Zeerak Talat	715
<i>Children’s English Reading Story Generation via Supervised Fine-Tuning of Compact LLMs with Controllable Difficulty and Safety</i>	Qian Shen, Fanghua Cao, Min Yao, Shlok Gilda, Bonnie Dorr and Walter Leite	751
<i>Transformer-based readability classifiers are worse than you think: Evidence from cross-domain Arabic readability assessment</i>	SARH ALZU’BI and Robert Reynolds	766
<i>Predicting Item Difficulty and Generating Reading Comprehension Items via an Annotated Repository</i>	Radhika Kapoor, Mayank Sharma, Sang Truong, Nick Haber, Ben Domingue and Maria Ruiz-Primo	777
<i>Generative-Evaluative Agreement: A Necessary Validity Criterion for LLM-Enabled Adaptive Assessment</i>	Grandee Lee, Yue Wang, Che Yee Lye and Luke Peh	798
<i>Evaluating LLM-Generated Formative Feedback for Undergraduate Mathematics Through the Lens of Feedback Theory</i>	Aron Gohr, Marie-Amelie Lawn, Kevin Gao, Inigo Serjeant and Stephen Heslip	813
<i>Confirming Correct, Missing the Rest: LLM Tutoring Agents Struggle Where Feedback Matters Most</i>	Tahreem Yasir, Wenbo Li, Sam Gilson, Sutapa Tithi, Xiaoyi Tian and Tiffany Barnes	819
<i>Retrieval-Augmented Tutoring for Algorithm Tracing and Problem-Solving in AI Education</i>	Mragisha Jain, Tirth Bhatt, Griffin Pitts, Aum Pandya, Peter Brusilovsky, Narges Norouzi, Arto Hellas, Juho Leinonen and Bitu Akram	841
<i>LLM-Powered but Rule-Grounded: Pedagogically Relevant Grammatical Error Characterization for Learner Model Construction</i>	Soroosh Akef, Amália Mendes, P Rebuschat and Detmar Meurers	851
<i>Teaching Through Analogies: A Modular Pipeline for Educational Analogy Generation</i>	Mariam Barakat and Ekaterina Kochmar	863
<i>Edit-level Majority Voting Mitigates Over-Correction in LLM-based Grammatical Error Correction</i>	Takumi Goto, Yusuke Sakai and Taro Watanabe	899
<i>Zero Shot Phonics: Evaluating Constraint-Adherent Phonics Story Generation in Large Language Models</i>	Maria Monica Manlises and Ethel Ong	914

<i>From Dialogue to Learner Modeling: Identifying Candidate Signals of Productive Use in LLM-Based Grammar Practice</i>	
Luisa Ribeiro-Flucht, Lanhua Huang and Xiaobin Chen	933
<i>Evaluating Adaptive Personalization of Educational Readings with Simulated Learners</i>	
Ryan Woo, Anmol Rao, Aryan Keluskar and Yinong Chen	941
<i>Toward Cross-Domain Automated Feedback: A Comparative Evaluation of Open-Source Models across Diverse Student Assessment Types</i>	
Muhammad Haseeb, Min Paing Hmue, Ahmad Imam Amjad, Maaz Amjad and Victor Sheng	951
<i>Findings of the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners</i>	
Mariano Felice and Lucy Skidmore	964
<i>SATLab at BEA 2026 Shared Task 1: Predicting the Difficulty of English Words for Three L1 Learners Using Primarily Psycholinguistic Features</i>	
Yves Bestgen	985
<i>UGA Threshold at BEA 2026 Shared Task 1: Predicting Vocabulary Acquisition Difficulty with Hand-Crafted SLA-Based Features</i>	
Emma Dalbo	992
<i>TeamXBC at BEA 2026 Shared Task 1: How AI (and I) won the shared task: Vibe and agentic coding solutions for practical machine learning problems</i>	
Xiaobin Chen	997
<i>SAAKTH at BEA 2026 Shared Task 1: L1-Aware English Vocabulary Difficulty Prediction with Hybrid Transformer and Psycholinguistic Features</i>	
Karthik Mattu, Adit Dhall, Arshad Naguru, Shubh Sehgal, Thejas Gowda and Hakyung Sung	1010
<i>SurreyCTS at BEA 2026 Shared Task 1: Semantic Funnelling and Entropy-based Multilingual Lexical Difficulty Prediction</i>	
Georgina Willoughby, Jordan Painter, Diptesh Kanojia, Emily Wells and Constantin Orasan	1016
<i>EduNLP at BEA 2026 Shared Task 1: Multi-Model Ensemble with Feature-Augmented Transformers for Vocabulary Difficulty Prediction</i>	
Avinash Kumar Sharma	1024
<i>AIDA at BEA 2026 Shared Task 1: A Two-Stage Framework for L1-Aware Vocabulary Difficulty Prediction with Representation Diversity and Residual Calibration</i>	
Seok Hyeon Cho, JunHyeok Choi, Sangeun Ji and Sung Won Han	1029
<i>Failure at BEA 2026 Shared Task 1: One Pipeline, Three L1s: A Unified Language-Agnostic System for Vocabulary Difficulty Prediction</i>	
Abid Hossain and Kamruzzaman Khan Alve	1041
<i>BoostedCats at BEA 2026 Shared Task 1: What Makes a Word Hard to Learn? Modeling L1 Influence on English Vocabulary Difficulty</i>	
Jonas Mayer Martins, Zhuojing Huang, Aaricia Herygers and Lisa Beinborn	1047
<i>uogal at BEA 2026 Shared Task 1: Ensemble of Multilingual Encoders with NMT Augmentation for L1-Aware Vocabulary Difficulty Prediction</i>	
bernardo stearns, John P. McCrae, Thomas Gaillat and Jefkine Kafunah	1065
<i>Jinnie's Lab at BEA 2026 Shared Task 1: Precalibration of Vocabulary Item Difficulty with Multilingual Transformers and Multi-Task Learning</i>	
Zhe Li, Pauline Aguinalde and Jinnie Shin	1077

<i>Glite at BEA 2026 Shared Task 1: Holistic Difficulty Models Dominate, Feature Engineering Closes the Gap in L1-Aware Vocabulary Difficulty Prediction</i>	
Vassili Philippov, Dmitrii Andreev, Pavel Katunin and Anton Nikolaev	1091
<i>NLP-Explorers at BEA 2026 Shared Task 1: DeBERTa–CatBoost Weighted Ensemble Approach for L1-Specific Vocabulary Difficulty Prediction</i>	
Tayyab Latif, Asifa Bibi, Sabur Butt, Grigori Sidorov and Alexander Gelbukh.	1106
<i>RETUYT-INCO at BEA 2026 Shared Task 1: Feature-Enriched mDeBERTa for Word Difficulty Prediction</i>	
Santiago Robaina, Aiala Rosá and Luis Chiruzzo	1113
<i>Token Titans at BEA 2026 Shared Task 1: Multilingual Lexical Complexity Prediction via Fine-Tuned XLM-RoBERTa with Ensemble Decoding</i>	
Anubhab Parashar and Sandeep Mathias	1119
<i>TOEBM at BEA 2026 Shared Task 1: Improving Lexical Difficulty Prediction with Context-Aligned Contrastive Learning and Ridge Ensembling</i>	
wicaksono M., Joanito Lopo, Tsamarah Nugraha, Ahmad Adi and Muhamad Nurfajri	1124
<i>Data Asgardians at BEA 2026 Shared Task 1: A Hybrid Transformer–Feature Ensemble for L1-Aware English Vocabulary Difficulty Prediction</i>	
Adrian Pineda, Sabur Butt and Héctor Ceballos Cancino	1137
<i>UOL@IDEM at BEA 2026 Shared Task 1: Neural Fusion and Feature-Rich Modeling for L1-Aware Vocabulary Difficulty Prediction</i>	
Nouran Khallaf and Serge Sharoff	1146
<i>Sakura at BEA 2026 Shared Task 1: What Makes Vocabulary Difficult?</i>	
Adam Nohejl, Xuanxin Wu, Yusuke Ide, Maria Riera Machin and Yi-Ning Chang	1161
<i>Report on the BEA 2026 Shared Task on Rubric-based Short Answer Scoring for German</i>	
Sebastian Gombert, Zhifan Sun, Fabian Zehner, Jannik Lossjew, Tobias Wyrwich, Berrit Czinczel, David Bednorz, Sascha Bernholt, Knut Neumann, Ute Harms, Aiso Heinze and Hendrik Drachslers	1179
<i>HFT at BEA 2026 Shared Task 2: Blunt-Edge Models for Hybrid Grading</i>	
Ulrike Pado	1193
<i>ASLAN at BEA 2026 Shared Task 2: Voting Across Scoring Paradigms</i>	
Marie Bexte, Yuning Ding, Josef Ruppenhofer, Nils-Jonathan Schaller, Daniel Mora Melanchthon, Torsten Zesch and Andrea Horbach	1201
<i>WSE Research at BEA 2026 Shared Task 2: Multi-Strategy Rubric-Based Short Answer Scoring for German</i>	
Jonas Gwozdz and Andreas Both	1210
<i>AMATI at BEA 2026 Shared Task 2: Automatic Short Answer Grading with Inductive Logic Programming and a Large Language Model</i>	
Alistair Willis and Aisling Third.	1217
<i>IWM-DKM at BEA 2026 Shared Task 2: Supplementing Supervised Fine-Tuning for Rubric-Based Short Answer Scoring</i>	
Kate Belcher, Marius De Kuthy Meurers, Kordula De Kuthy and Detmar Meurers	1224
<i>RETUYT-INCO at BEA 2026 Shared Task 2: Meta-prompting in Rubric-based Scoring for German</i>	
Ignacio Sastre, Ignacio Remersaro, Facundo Díaz, Nicolás De Horta, Luis Chiruzzo, Aiala Rosá and Santiago Góngora	1235

SDPA at BEA 2026 Shared Task 2: Efficient LLM Fine-Tuning for Rubric-based Short Answer Scoring
Zhexiong Liu and Jing Zhang 1244

Program

Friday, July 3, 2026

06:00 - 07:30 *Early Poster Session*

Using k-Shot Prompting with Large k for the Automated Scoring of a German Written Elicited Imitation Test

Malte Sternik, Ronja Laarmann-Quante and Anastasia Drackert

Fine-Grained Content Zone Prediction in German Argumentative Essays Using LLMs

Xiaoyu Bai and Manfred Stede

Sharing is Caring: Advantages of Sharing a Language Background with Learners as an Annotator of Learner Data in UD

Caroline Grand-Clement and Arianna Masciolini

The Effects of Structured LLM-Generated Feedback on Programming Assignment Performance

Tsvetomila Mihaylova, Evanfiya Logacheva, Arto Hellas, Jing Fan, Francisco Castro, Bitu Akram, Narges Norouzi, Peter Brusilovsky and Juho Leinonen

HFT at BEA 2026 Shared Task 2: Blunt-Edge Models for Hybrid Grading

Ulrike Pado

WSE Research at BEA 2026 Shared Task 2: Multi-Strategy Rubric-Based Short Answer Scoring for German

Jonas Gwozdz and Andreas Both

07:30 - 08:45 *Early Oral Session*

Domain-Adaptive Pre-training for Automated Short Answer Grading in Conceptual Physics: Reliability, Question-Level Analysis, and Error Reduction

Shirin Lade, Alistair Willis, Jonathan Nylk and Oli Howson

Using Interaction Log Data to Evaluate and Improve Feedback Accuracy in an Intelligent Language Tutoring System

Mariia Soliar, Leona Colling, Stephen Bodnar and Detmar Meurers

Towards Pedagogically Aligned LLM Tutors for Math Mistake Remediation

Kseniia Petukhova, Tien Dat Nguyen and Ekaterina Kochmar

What Aggregate Scores Hide: Per-Rule Evaluation of Russian Grammatical Error Correction

Anna Smirnova, Artyom Kopan, Vladislav Makeev and George Chernishev

Friday, July 3, 2026 (continued)

IWM-DKM at BEA 2026 Shared Task 2: Supplementing Supervised Fine-Tuning for Rubric-Based Short Answer Scoring

Kate Belcher, Marius De Kuthy Meurers, Kordula De Kuthy and Detmar Meurers

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

Theory of Mind and Application in Educational Context

Effat Farhana, Maha Zainab, Qiaosi Wang, Niloofar Miresghallah, Ramira van der Meulen and Max van Duijn

10:30 - 11:00 *Coffee Break*

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

Theory of Mind and Application in Educational Context

Effat Farhana, Maha Zainab, Qiaosi Wang, Niloofar Miresghallah, Ramira van der Meulen and Max van Duijn

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

14:00 - 15:45 *Oral Session A*

Using Interaction Log Data to Evaluate and Improve Feedback Accuracy in an Intelligent Language Tutoring System

Mariia Soliar, Leona Colling, Stephen Bodnar and Detmar Meurers

Estimating LLM Grading Ability and Response Difficulty in Automatic Short Answer Grading via Item Response Theory

Longwei Cong, Sonja Hahn, Sebastian Gombert, Leon Camus, Hendrik Drachsler and Ulf Kroehne

LLM-Powered but Rule-Grounded: Pedagogically Relevant Grammatical Error Characterization for Learner Model Construction

Soroosh Akef, Amália Mendes, P Rebuschat and Detmar Meurers

KEYSCORE — Keystroke-enhanced Automated Essay Scoring

Nils-Jonathan Schaller, Daniel Mora Melanchthon, Thorben Jansen, Olaf Köller and Andrea Horbach

15:45 - 16:00 *Coffee Break*

Friday, July 3, 2026 (continued)

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

Inferring Student Engagement via Real-Time Thermal–Visual Voice Activity Detection

Bradley Goodman

Letting Tutor Personas Speak Up for LLMs: Learning Steering Vectors from Dialogue via Preference Optimization

Jaewook Lee, Alexander Scarlatos, Simon Woodhead and Andrew Lan

A Bigger Catch: Fine-Grained Curriculum Standards Alignment on the MathFish Benchmark

Xinman Liu, Mayank Sharma and Xinyu Shi

Through the Sentence Lens: Explainable Essay Scoring through Fine-Grained Predictions

Daniel Mora Melanchthon, Stefan Keller and Andrea Horbach

Kelvi: A Morphological Parser to Support Tamil Literacy

Shankhalika Srikanth, Sabrina Yu, Sophia Chan and Madeline Solis de Ovando

From Questions to Assessment Tuples: A Multi-Agent Framework with Bloom-Specialized Agents and Automated Verification

Gee-Lyle Wong, Runcong Zhao, Yulan He and Jiazheng Li

Multi-step Large Language Model for Fine-Grained Feedback in Stepwise Linear Equation Solutions

Imran Chamieh, Torsten Zesch and Klaus Giebermann

Comparative Evaluation of AI-Generated vs. Expert-written Answer Explanations for a Medical Education Self-Assessment

Yiyun Zhou, Francis O'Donnell and Victoria Yaneva

Multi-component student writing profiles for expert-aligned automated evaluation of English learner essays.

Russell Moore, Andrew Caines and Paula Buttery

Policy-Sensitive Fairness Evaluation in Automated Scoring of Clinical Communication

Saed Rezayi, Le An Ha, Victoria Yaneva, Polina Harik, Janet Mee and Jason Snyder

Friday, July 3, 2026 (continued)

Rubrics as Semantic Subspaces: A Unified Approach to Rubric-based Constructed Response Scoring across Short Answers and Essays

Sebastian Gombert, Sonja Hahn, Nico Andersen, Leon Camus, Zhifan Sun, Ngoc Nhu Hao Nguyen, Fabian Zehner, Longwei Cong, Alexander Mehler and Hendrik Drachslar

Assessment of L2 speech global dimensions using large audio language models

Elsayed Issa and Mahmoud Ali

Children’s English Reading Story Generation via Supervised Fine-Tuning of Compact LLMs with Controllable Difficulty and Safety

Qian Shen, Fanghua Cao, Min Yao, Shlok Gilda, Bonnie Dorr and Walter Leite

Transformer-based readability classifiers are worse than you think: Evidence from cross-domain Arabic readability assessment

SARH ALZU’BI and Robert Reynolds

Predicting Item Difficulty and Generating Reading Comprehension Items via an Annotated Repository

Radhika Kapoor, Mayank Sharma, Sang Truong, Nick Haber, Ben Domingue and Maria Ruiz-Primo

Generative-Evaluative Agreement: A Necessary Validity Criterion for LLM-Enabled Adaptive Assessment

Grandee Lee, Yue Wang, Che Yee Lye and Luke Peh

Teaching Through Analogies: A Modular Pipeline for Educational Analogy Generation

Mariam Barakat and Ekaterina Kochmar

From Dialogue to Learner Modeling: Identifying Candidate Signals of Productive Use in LLM-Based Grammar Practice

Luisa Ribeiro-Flucht, Lanhua Huang and Xiaobin Chen

Toward Cross-Domain Automated Feedback: A Comparative Evaluation of Open-Source Models across Diverse Student Assessment Types

Muhammad Haseeb, Min Paing Hmue, Ahmad Imam Amjad, Maaz Amjad and Victor Sheng

SAAKTH at BEA 2026 Shared Task 1: L1-Aware English Vocabulary Difficulty Prediction with Hybrid Transformer and Psycholinguistic Features

Karthik Mattu, Adit Dhall, Arshad Naguru, Shubh Sehgal, Thejas Gowda and Hakyung Sung

SDPA at BEA 2026 Shared Task 2: Efficient LLM Fine-Tuning for Rubric-based Short Answer Scoring

Zhexiong Liu and Jing Zhang

Friday, July 3, 2026 (continued)

TOEBM at BEA 2026 Shared Task 1: Improving Lexical Difficulty Prediction with Context-Aligned Contrastive Learning and Ridge Ensembling
wicaksono M., Joanito Lopo, Tsamarah Nugraha, Ahmad Adi and Muhamad Nur-fajri

Saturday, July 4, 2026

09:00 - 10:30 *Oral Session B*

The Aftermath of DrawEduMath: Vision Language Models Underperform with Struggling Students and Misdiagnose Errors

Li Lucy, Albert Zhang, Nathan Anderson, Ryan Knight and Kyle Lo

Interpretable Difficulty-Aware Knowledge Tracing in Tutor-Student Dialogues

Shuyan Huang, Alexander Scarlatos, Jaewook Lee and Andrew Lan

Measuring Optimal Challenge: Trajectory-Based Difficulty Alignment in Open-Ended Language Tutoring

Ziqi Shu, Shuman Wang and Michael Hardy

Findings of the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners

Mariano Felice and Lucy Skidmore

Sakura at BEA 2026 Shared Task 1: What Makes Vocabulary Difficult?

Adam Nohejl, Xuanxin Wu, Yusuke Ide, Maria Riera Machin and Yi-Ning Chang

Report on the BEA 2026 Shared Task on Rubric-based Short Answer Scoring for German

Sebastian Gombert, Zhifan Sun, Fabian Zehner, Jannik Lossjew, Tobias Wyrwich, Berrit Czinczel, David Bednorz, Sascha Bernholt, Knut Neumann, Ute Harms, Aiso Heinze and Hendrik Drachsler

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Oral Session C*

EduMUSE: A Multimodal Educational Dataset with Automatically Extracted Instructional Context

Andreea Dutulescu, Stefan Ruseti, Mihai Dascalu and Danielle McNamara

Confirming Correct, Missing the Rest: LLM Tutoring Agents Struggle Where Feedback Matters Most

Tahreem Yasir, Wenbo Li, Sam Gilson, Sutapa Tithi, Xiaoyi Tian and Tiffany Barnes

Towards Just-in-Time Adaptive Feedback: Enhancing Student Learning via Knowledge-Grounded LLM

Younghun Lee, Amir Bralin, Nobel Sanjay Rebello and Dan Goldwasser

Saturday, July 4, 2026 (continued)

Evaluating LLM Workflows for Generating Clinical Communication Assessment Items: A Comparative Study with Subject-Matter Experts

Christopher Runyon, Peter Baldwin, Ian Micir, Kevin Frome, Stephanie Mann, Saed Rezayi, Keelan Evanini and Victoria Yaneva

Zero Shot Phonics: Evaluating Constraint-Adherent Phonics Story Generation in Large Language Models

Maria Monica Manlises and Ethel Ong

Evaluating Adaptive Personalization of Educational Readings with Simulated Learners

Ryan Woo, Anmol Rao, Aryan Keluskar and Yinong Chen

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

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

Investigating Context-aware CTC for Pronunciation Assessment: Mitigating Peaky Behavior and Context Independency Assumption

Jiun-Ting Li, Tien-Hong Lo, Bi-Cheng Yan, Shih-Hsuan Chiu, Fu-An Chao and Berlin Chen

A Survey of Automated Presentation Coaching: Systems, Methods, and Open Challenges

Wen Liang, Li Siyan, Zackary Rackauckas and Julia Hirschberg

Criteria Features in German: Towards Interpretable NLP in Readability Assessment

Denise Loefflad, Sofia Kathmann, Heiko Holz and Detmar Meurers

RABIT: Rationale-Based Distillation Towards Interpretable Automatic Speaking Assessment via a Small Language Model

Bi-Cheng Yan, Hong-Yun Lin, Fu-An Chao, Jiun-Ting Li and Berlin Chen

Challenges in Machine Translation of Interactive Multimodal Exercises

Lucie Polakova, Miroslav Hrabal, Věra Kloudová, Michal Novák, Mariia Anisimova and Martin Popel

Towards Self-Referential Analytic Assessment: A Profile-Based Approach to L2 Writing Evaluation with LLMs

Stefano Banno, Kate Knill and Mark Gales

Assessing the Quality and Consistency of Automated Knowledge Component Generation using Instructor-generated Questions and LLMs

Jordan Esiason, Priyanka Khare, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng and Yeil Jeong

Saturday, July 4, 2026 (continued)

Intent vs. Surface: Recovering Acoustic Realization from Modern ASR for Pronunciation Training

Seongjin Park

Opportunities and Challenges of LLMs in Education: An NLP Perspective

Sowmya Vajjala, Bashar Alhafni, Stefano Banno, Kaushal Maurya and Ekaterina Kochmar

Quality-Conditioned Agreement in Automated Short Answer Scoring: Mid-Range Degradation and the Impact of Task-Specific Adaptation

Abigail Gurin Schleifer, Moriah Ariely, Beata Beigman Klebanov, Asaf Salman and Giora Alexandron

Using LLMs for item creation: Validating the potential of automatically generated sentence repetition test items for language assessment

Sarah Löber, Björn Rudzewitz, Yuan Chu, Mengyuan He, Shiqin Liu, Yushan Ye and Xiaobin Chen

FinnGEC: Benchmarking Grammatical Error Correction for Finnish

Anh-Duc Vu, Mikhail Zolotilin, Jue Hou, Anisia Katinskaia, Yiheng Wu and Roman Yangarber

From Metrics to Meaning: Rule-Grounded LLM Explanations for Data Literacy in the Case of Youth Football

Tomasz Piłka, Tomasz Kuczyński and Mateusz Czajka

Classification of Student Struggle in Mathematics

Hannah Levin, Madhura Padwal and Nchimunya Mwiinga

PERSA: Reinforcement Learning for Professor-Style Personalized Feedback with LLMs

Ravi Kumar, Utkarsh Grover, Xiaomin Lin and Agoritsa Polyzou

Data-lean fine-tuning of models for evaluating teacher performance in a GenAI-led elicitation simulation

Beata Beigman Klebanov, Andrew Hoang, Jamie Mikeska, Benny Longwill, Sanjna Kashyap, Shreyashi Halder and Aakanksha Bhatia

Noise Steering for Controlled Text Generation: Improving Diversity and Reading-Level Fidelity in Arabic Educational Story Generation

Haziq Khalid, Salsabeel Shapsough and Imran Zualkernan

PeerMathDial: A Middle School Dialogue Dataset for Student Collaborative Math Problem Solving

Murong Yue, Desmond Mcglone, Emily Slutz, Wenhan Lyu, Yixuan Zhang, Jennifer Suh and Ziyu Yao

Saturday, July 4, 2026 (continued)

Evaluating LLM-Generated Formative Feedback for Undergraduate Mathematics Through the Lens of Feedback Theory

Aron Gohr, Marie-Amelie Lawn, Kevin Gao, Inigo Serjeant and Stephen Heslip

Retrieval-Augmented Tutoring for Algorithm Tracing and Problem-Solving in AI Education

Mragisha Jain, Tirth Bhatt, Griffin Pitts, Aum Pandya, Peter Brusilovsky, Narges Norouzi, Arto Hellas, Juho Leinonen and Bitu Akram

Edit-level Majority Voting Mitigates Over-Correction in LLM-based Grammatical Error Correction

Takumi Goto, Yusuke Sakai and Taro Watanabe

AIDA at BEA 2026 Shared Task 1: A Two-Stage Framework for L1-Aware Vocabulary Difficulty Prediction with Representation Diversity and Residual Calibration

Seok Hyeon Cho, JunHyeok Choi, Sangeun Ji and Sung Won Han

BoostedCats at BEA 2026 Shared Task 1: What Makes a Word Hard to Learn? Modeling L1 Influence on English Vocabulary Difficulty

Jonas Mayer Martins, Zhuojing Huang, Aaricia Herygers and Lisa Beinborn

uogal at BEA 2026 Shared Task 1: Ensemble of Multilingual Encoders with NMT Augmentation for L1-Aware Vocabulary Difficulty Prediction

bernardo stearns, John P. McCrae, Thomas Gaillat and Jefkine Kafunah

Jinnie's Lab at BEA 2026 Shared Task 1: Precalibration of Vocabulary Item Difficulty with Multilingual Transformers and Multi-Task Learning

Zhe Li, Pauline Aguinalde and Jinnie Shin

IWM-DKM at BEA 2026 Shared Task 2: Supplementing Supervised Fine-Tuning for Rubric-Based Short Answer Scoring

Kate Belcher, Marius De Kuthy Meurers, Kordula De Kuthy and Detmar Meurers

RETUYT-INCO at BEA 2026 Shared Task 2: Meta-prompting in Rubric-based Scoring for German

Ignacio Sastre, Ignacio Remersaro, Facundo Díaz, Nicolás De Horta, Luis Chiruzzo, Aiala Rosá and Santiago Góngora

UOL@IDEM at BEA 2026 Shared Task 1: Neural Fusion and Feature-Rich Modeling for L1-Aware Vocabulary Difficulty Prediction

Nouran Khallaf and Serge Sharoff

15:30 - 16:00

Coffee Break

Saturday, July 4, 2026 (continued)

16:00 - 16:45 *Panel*

16:45 - 17:15 *Oral Session D*

Incentives Of EdTech: A Systematic Review Of EduNLP Research

Gabrielle Gaudeau, Aoife O’Driscoll, Jasper Degraeuwe, Andrew Caines, Donya Rooein and Zeerak Talat

Effects of Varying LLM Access on Essay Writing Behavior

Julia Christenson, Karin de Langis, Shirley Anugrah Hayati and Dongyeop Kang

17:15 - 17:30 *Closing Remarks*