LHI EXPO

UNIVERSITY OF
CAMBRIDGE

DEPARTMENT OF
COMPUTER SCIENCE
AND TECHNOLOGY

1 JULY 2025

TABLE OF CONTENTS

Welcome	2
Programme of the day	3
Poster session	4-17
List of presenters	18

WELCOME

Welcome to the second Learning & Human Intelligence Poster Exposition, the LHI Expo, at the Department of Computer Science & Technology. We have more than twenty posters and five talks detailing the work of Ada Computer Science, the Institute for Automated Language Teaching & Assessment (ALTA), the Raspberry Pi Computing Education Research Centre, and other independent researchers.

We hope you enjoy the event and welcome feedback at lhi-admins@cst.cam.ac.uk.

The Learning & Human Intelligence Group, Department of Computer Science & Technology, University of Cambridge.

www.cst.cam.ac.uk/research/lhi

PROGRAMME

11:00	**Start** - The Street, William Gates Building
11:15-11:30	**Talk** by *Dr Jessie Durk* RPCERC - "An Overview of Current Research in RPCERC" in Lecture Theatre 2
11:45-12:00	**Talk** by *Sol Dubock and Alex Lewin* Ada Computer Science - "Isaac Physics and Ada Computer Science in 2025" in Lecture Theatre 2
12:00-13:00	**Lunch** is served & Poster session
13:00 - 13:15	**Talk** by *Richard Diehl Martinez* CST - "Peering Inside LLMs with Pico: A Lightweight Framework for Learning-Dynamics Research" in Lecture Theatre 2
13:15-13:30	**Talk** by *Dr Luca Benedetto* ALTA - "Grammatical Gender Markers Bias LLM-based Educational Recommendations" in Lecture Theatre 2
13:45-14:00	**Talk** by *Dr Guy Emerson* ALTA - "Complexity-Theoretic Constraints on the Coherence of Cognitive Processes" in Lecture Theatre 2

POSTER SESSION

1. Investigating K-12 data ethics education

Authors: C. Johnson and D. Kirby

The field of computing ethics in higher education has been subject to research over a period of at least 40 years; in comparison, K-12 data ethics education is in its infancy. Growing demands for an AI curriculum in schools have highlighted the need to consider ethical as well as technical knowledge content. Given that there is no mandated AI curriculum in the UK, are young people being taught what they need to know about the ethical issues relating to the design and use of AI mediated/data-driven tools?

We are investigating how data ethics topics are delivered in the curriculum in three different contexts: (1) within an emerging AI curriculum; (2) within the current and legacy computing/CS curriculum; (3) within the broader ethics curriculum.

The project involves two phases:
1. a scoping literature review of the research published between 2023 and 2025 in the field of K-12 data ethics education (in progress)
2. an intervention study to investigate and develop UK K-12 students' knowledge, skills and understanding of data ethics (forthcoming)

2. Secondary teachers' experiences of teaching programming with LLMs in the classroom (Work in progress)

Authors: V. Cucuiat, K. Childs, B. Sheppard and J. Waite

This is a work-in-progress update. We describe a study with 15 secondary teachers in schools in England. In November 2024, participating teachers attended an in-person, one-day workshop where the research team introduced them to 1) feedback literacy, which is the ability to understand, interpret, and use feedback effectively to improve learning or performance, and 2) an LLM-powered tool called CodeHelp. Teachers were asked to integrate the use of CodeHelp into a lesson of their choice. Initial results suggest considerable variability in the lesson content and pedagogical approaches of the teachers. However, teachers were keen to see explanations that develop students' understanding of concepts rather than telling them how to fix their code. Future work may focus on new pedagogical approaches for learning to program with LLMs in the classroom.

3. Towards Using LLMs for Content Analysis

Authors: S. M. Nicolajsen and L. Gale

LLMs have the ability to empower researchers and others to undertake research that they might not otherwise be able to conduct. In this poster, we present early thoughts on how the process of content analysis can be expanded upon using LLMs. We consider potential applications, share some early personal experiences, offer cautionary notes, and reflect on the future of this method.

4. Student Use and Teacher Practice: A Scoping Review of Generative AI in Computing Education Using PICRAT

Authors: CA. Philbin and S. Sentance

Technology has promised to transform teaching and learning for over 25 years, from interactive whiteboards to personalised learning platforms. Generative AI large language model (LLM) chatbots, such as ChatGPT, are the latest innovation poised to reshape computing education. However, much of the research has taken a technology-driven approach to their integration rather than a student-centred one. This scoping review examines studies published between October 2022 and October 2024, using the PICRAT framework to map student activity with LLMs as passive, interactive, or creative, and educator use as replacing, amplifying, or transforming traditional teaching methods. Analysing 32 studies across K-12 and higher education, sourced following PRISMA-ScR guidelines, we find that use of AI chatbots primarily replicates or amplifies conventional instructional patterns, with most student interactions being passive or interactive rather than creative. While these tools improve efficiency, their potential for transformative learning, particularly in K-12 settings, remains under explored. This review offers insights for computing educators seeking to integrate AI into their teaching and highlights the value of using PICRAT, not only for planning and designing AI-enhanced learning activities, but also for evaluating their pedagogical impact over time.

5. A positive and inclusive classroom experience: primary school teachers' perspectives of physical computing

Authors: J. Durk and S. Sentance

The UK-wide, five-year, longitudinal 'Exploring Physical Computing in Schools' (EPICS) project is underway and investigates how engagement with physical computing can impact primary school children's confidence, creativity, and agency. Reporting on a subset of data collected in the first year of the project, we investigated primary school teachers' views and experiences of delivering introductory physical computing lessons using the BBC micro:bit to 8 and 9 year olds in the UK. We interviewed 10 individual teachers, and explored themes of positive learning environments and inclusivity. The teachers felt that a positive learning environment was realised through the children's enjoyment, their own modelling of excitement, and the cultivation of resilience. Teachers felt that the micro:bit lessons were inclusive, allowing all learners to feel a sense of achievement, particularly for those with additional needs. Our findings have implications for children's technological self-efficacy, creativity, and digital capital.

6. The Quiet Harm of Vibecoding: How AI Tools Can Erode Foundational Skills

Author: M. Cheung

Generative AI enables "vibe coding" an improvisational programming style prioritising speed and AI-generated suggestions. While this approach enhances productivity, it carries significant, subtle risks to foundational learning. In this qualitative study, I explored the experiences of five university students using AI coding tools. The findings reveal a key tension: participants are aware that AI over-reliance leads to shallow engagement and skill erosion. Critically, they also express a desire for AI that introduces "epistemic friction", tools that ask questions and provide guided support rather than delivering instant solutions. The uncritical pursuit of seamlessness in AI design can quietly undermine learning. This poster concludes that educational AI tools must evolve beyond mere productivity aids and be redesigned as scaffolds for reasoning, incorporating intentional friction to foster deeper reflection and intellectual autonomy.

7. The Ignorant Co-Learner: An "Artificial Ignorance" Approach to AI Design

Author: M. Cheung

Generative AI is often framed as an authoritative "answer machine" delivering seamless responses, a model that fosters passive reliance and discourages critical reflection. This poster challenges that paradigm by introducing artificial ignorance: an AI designed not to instruct, but to provoke inquiry by acting as an "ignorant co-learner". This approach is directly inspired by Jacques Rancière's philosophy of the "ignorant schoolmaster", who disrupts hierarchical teaching by withholding expert explanation to affirm the learner's own capacity to think independently. Translating this ethic to AI, we propose systems that relinquish claims to authority and instead introduce "epistemic friction", deliberate moments of uncertainty or conflict that compel users to question assumptions and construct meaning for themselves. This work attempts to reframe educational AI, shifting the goal from the automation of answers to the cultivation of critical, collaborative sense-making between humans and machines by outlining practical design principles that position users as reflective co-creators rather than passive recipients.

8. Data-driven problem-solving in K-12 education: A systematic literature review

Author: B. Whyte, M. Cheung, and K. Childs

The growing importance of data in an AI-driven world has led to calls for a rethinking of computing education (Shapiro et al., 2019; Tedre et al., 2021). Alongside traditional rule-based programming skills, learners must also work within a new data-driven paradigm in creating and interacting with emerging technologies. To understand what data science is taking place in K–12 settings, we conducted a systematic literature review of data-driven problem-solving. We have conducted an initial data extraction on the nature and scope of data science across a set of 98 papers. In this poster, we share preliminary findings from our review on which forms of data-driven problem-solving are taking place in educational settings. We report on who is learning about data science, in what countries and contexts it is being taught, and what tools are being used. We also share some emergent insights into future directions for this work.

9. Using a LLM to Mark Free-Text Questions on Ada Computer Science

Authors: S. Dubock and J. Waite

An ongoing study to evaluate the use of Large Language Models (LLMs) for marking free-text questions - those requiring students to write one or more sentences in response. This poster discuss how these questions have been integrated into the website, as well as a review of the performance of questions with high levels of student disagreement and what approaches we can take to resolve these.

10. Automatic Evaluation of LLM Generated Texts

Author: S. Taslimipoor

With the increasing use of large language models (LLMs) to generate text for various purposes—including exams and learning materials for language learners—there is a growing need to evaluate the quality of the generated content. Traditional text generation evaluation methods typically compare generated text to reference texts, ranking outputs based on their overlap or similarity to the reference texts. However, these approaches often struggle with open-ended generation and may fail to capture nuances such as coherence or contextual appropriateness. In this project, we aim to design reference-free metrics that can evaluate the quality of generated text across multiple dimensions in an unsupervised manner. We employ a dataset of LLM generated texts for a single-item reading comprehension task, where the texts are annotated by human experts for their effectiveness to be used for the task. We gather a set of reference-free text evaluation metrics and investigate how they are correlated with human judgments of the generated texts.

11. Dictionary Look-up Prediction

Authors: D. Strohmaier, G. Tyen, H. Gu, D. Nicholls, Z. Yuan, and P. Buttery

Knowing which words language learners struggle with is crucial for developing personalised education technologies. This poster presents the "dictionary look-up prediction" task as a means for evaluating the complexity of words in reading tasks using Read & Improve data. Read & Improve is an online learning platform for learners of English to practise reading comprehension exercises.

12. A Survey on Automated Distractor Evaluation in Multiple-Choice Tasks

Authors: L. Benedetto, S. Taslimipoor, and P. Buttery

We perform a comprehensive study of the approaches which are used in the literature for distractor evaluation in the context of multiple-choice questions (MCQs). We propose a taxonomy to categorise them, and highlight the need for novel evaluation techniques, since previously proposed ones are often not aligned with how distractors perform in exam settings.

13. Towards CEFR-targeted Text Simplification for Question Adaptation

Authors: L. Benedetto and P. Buttery

In this work, we study the feasibility of using CEFR-targeted text simplification to perform question adaptation. Focusing on reading comprehension multiple choice questions, our results suggest that it is indeed possible to tune the difficulty of existing questions by simplifying the associated reading passage. However, the models under consideration are not capable of a precise alignment with the CEFR (Common European Framework of Reference for Languages)

14. Rubrik's Cube: Testing a New Rubric for Evaluating Explanations on the CUBE dataset

Authors: D. Galvan-Sosa, G. Gaudeau, P. Kavumba, Y. Li, H. Gu, Z. Yuan, K. Sakaguchi, and P. Buttery

As organisations increasingly deploy AI systems to enhance customer experiences and accelerate product development, Large-Language Models (LLMs) have emerged as a popular choice for explanation generation due to their performance and usability. Interpretability is particularly important for explanation generation, where GenAI systems must not only provide answers but justify their reasoning to users. However, despite their widespread adoption, LLM explanations have been found to be unreliable, making it difficult for users to distinguish good from bad explanations. To address this issue, we designed Rubrik's CUBE, an education-inspired rubric and a dataset of 26k explanations, written and quality-annotated using the rubric by both humans and six open- and closed-source LLMs. The CUBE dataset focuses on two reasoning and two language tasks, providing the necessary diversity for us to effectively test our proposed rubric. Using Rubrik, we find that explanations are influenced by both task and perceived difficulty. Low quality explanations stem primarily from a lack of conciseness in LLM-generated explanations, rather than cohesion and word choice. Our work contributes to the responsible integration of GenAI into critical decision-making processes, providing a foundation for future advancements in explanation quality assessment.

15. Multiword Expressions and Grammatical Error Correction in Write & Improve Essays

Authors: S. Taslimipoor, C. Bryant, D. Nicholls, A. Caines, and P. Buttery

We describe the annotation work undertaken to identify multiword expressions (e.g. idioms, phrasal verbs, noun compounds) in English essays from the Write & Improve learning platform. This has involved 1850 learner essays at beginner, intermediate and advanced proficiency levels and a new error type for MWE errors. This is the first step in detecting MWE errors automatically and providing feedback to learners on MWE usage.

16. LLM-based post-editing as reference-free evaluation of grammatical error correction

Authors: R. Östling, M. Kurfali, and A. Caines

We investigate the use of Large Language Models (LLMs) as post-editors of English and Swedish texts, and perform a meta-analysis of a range of different evaluation setups using a set of recent GEC systems. We find that for the two languages studied in our work, automatic evaluation based on post-editing agrees well with both human post-editing and direct human rating of GEC systems. Furthermore, we find that a simple n-gram overlap metric is sufficient to measure post-editing distance, and that including human references when prompting the LLMs generally does not improve agreement with human ratings. The resulting evaluation metric is reference-free and requires no language-specific training or additional resources beyond an LLM capable of handling the given language.

17. Computational Modeling of Language Production

Author: Y. Gao

This research develops statistical models that learn probabilistic grammars from language production data, with applications to modeling individual learner grammar acquisition. Using Synchronous Hyperedge Replacement Grammar, our approach learns mappings between semantic representations and syntactic structures while requiring significantly less data than neural methods and maintaining full interpretability. We validate the framework on child language data (CHILDES) and WSJ articles (DeepBank). The system generates individual grammar profiles. Current extensions include individual grammar profiling, and learning trajectory simulation to predict optimal instructional sequences and identify minimal exposure requirements for grammar acquisition. This work provides a principled computational approach to language acquisition research with direct applications to personalized assessment and adaptive learning systems.

18. Extended many-facet Rasch models: Accounting for rater effects in automated essay scoring systems

Author: M. Elliott

The representation of a rater as a global scalar measure involves an assumption of uniformity of severity across the range of rating scales and criteria within each scale. We introduce extended rater representations of vectors or matrices of local measures relating to individual rating scales and criteria. We conclude that extended representations more naturally and completely reflect the role of the rater within the assessment process and provide greater inferential power than the traditional global measure of severity. This poster provides an overview of the research outputs from my PhD. It shows the narrative built through my thesis, which has produced two peer-reviewed journal articles and one conference presentation, plus further applications of the work, a description of the software package I have produced and suggested avenues for further research resulting from my studies.

19. Bilingual Small Language Models as Cognitive Proxies for LLM Interaction and Calibration

Author: S. Salhan

This presents a framework for high-precision Bilingual Small Language Models that simulate a diversity of multilingual (second language and successive/simultaneous bilingual acquisition) scenarios. These SLMs are cognitive proxies that will be used to calibrate LLM outputs in multi-agent interaction. It will describe preliminary work on cognitively-inspired interaction between LLMs and SLMs, developed in collaboration with Dr Diana Galvan-Sosa, Hongyi Gu, Dr Donya Rooein, Gabrielle Gaudeau, Dr Zheng Yuan, Dr Andrew Caines, Prof Paula Buttery.

20. Understanding the Individual Writer: Towards Personalised Feedback for Writing Development

Authors: D. Galvan-Sosa, A. Caines, P. Buttery, and D. Nicholls

Effective writing instruction requires integrated assessment approaches that provide evidence of student progress while maintaining coherence throughout the learning journey. Writing, a fundamental and complex human skill, is crucial for academic, professional, and creative expression. Our long-term vision is to significantly enhance writing instruction by prioritizing the individual learner, recognizing their unique challenges and writing patterns. We aim to tailor feedback to specific needs and learning styles, fostering a more personalized and effective learning experience. Our immediate goal focuses on understanding how Large Language Models (LLMs) can be effectively integrated into writing support systems. We employ a data-driven approach, analyzing diverse student writing datasets with advanced Natural Language Processing (NLP) techniques. This allows us to gain a deeper understanding of the factors contributing to high-quality writing, particularly at the sentence and discourse levels, ultimately guiding the development of more effective and personalized AI-powered writing tools.

21. The English Grammar Profile

Author: Ø. Andersen

An English Grammar Profile tagger that enables automated identification of grammatical constructions (classified by part of speech and CEFR level) in both learner texts and learning/examination materials has a number of applications, notably in assessment, formative feedback and content writing. Ongoing work includes further development of the EGP definitions to make them more precise and amenable to automation as well as manual annotation of full learner texts to facilitate more realistic evaluation of the tool.

22. IPA CHILDES & G2P+: Resources for Cross-Lingual Phonology and Phonemic Language Modeling

Authors: Z. Goriely and P. Buttery

We introduce two resources: (i) G2P+, a tool for converting orthographic datasets to a consistent phonemic representation; and (ii) IPA CHILDES, a phonemic dataset of child-directed and child-produced speech across 31 languages. Prior tools for grapheme-to-phoneme conversion result in phonemic vocabularies that are inconsistent with established phonemic inventories, an issue which GP2+ addresses by leveraging the inventories in the Phoible database. Using this tool, we augment CHILDES with phonemic transcriptions to produce IPA CHILDES. This new resource fills several gaps in existing phonemic datasets, which often lack multilingual coverage, spontaneous speech, and a focus on child-directed language. We demonstrate the utility of this dataset for phonological research by training phoneme language models on 11 languages and probing them for distinctive features, finding that the distributional properties of phonemes are sufficient to learn major class and place features cross-lingually.

23. ByteSpan: Information-Driven Subword Tokenisation

Authors: Z. Goriely, S. Salhan, P. Lesci, J. Cheng, and P. Buttery

Recent dynamic tokenisation methods operate directly on bytes and pool their latent representations into patches. This bears similarities to computational models of word segmentation that determine lexical boundaries using spikes in an autoregressive model's prediction error. Inspired by this connection, we explore whether grouping predictable bytes – rather than pooling their representations – can yield a useful fixed subword vocabulary. We propose a new information-driven subword tokeniser, ByteSpan that uses an external byte-level LM during training to identify contiguous predictable byte sequences and group them into subwords. Experiments show that ByteSpan yields efficient vocabularies with higher morphological alignment scores than BPE for English. Multilingual experiments show similar compression and Rényi efficiency for 25 languages.

24. BabyLM's First Words: Word Segmentation as a Phonological Probing Task

Authors: Z. Goriely and P. Buttery

Language models provide a key framework for studying linguistic theories based on prediction, but phonological analysis using large language models (LLMs) is difficult; there are few phonological benchmarks beyond English and the standard input representation used in LLMs (subwords of graphemes) is not suitable for analyzing the representation of phonemes. In this work, we demonstrate how word segmentation can be used as a phonological probing task, allowing us to study the representations learned by phoneme-based language models trained on child-directed speech across 31 languages. Following computational models of word segmentation, we present unsupervised methods for extracting word boundaries from a trained model using the observation that prediction-error peaks at the start of words. We also use linear probes to identify that these models implicitly track word boundaries, even when they do not appear in training.
This cross-lingual work corroborates statistical learning theories of acquisition and empirically motivates new methods for training subword tokenizers.

25. Adaptivity for Optimised Learning

Author: R. Moore

Here we present a walkthrough of frameworks for adaptive learning, based on a central predictive component that is able to accurately estimate future outcomes for a human learner. This informs either short-term recommendation or longer-term planning agents, that perform the customisation of pedagogical experiences for the user.This cross-lingual work corroborates statistical learning theories of acquisition and empirically motivates new methods for training subword tokenizers.

26. Bias Dynamics in BabyLMs: In Search of a Compact Sandbox for Lifecycle-Wide Debiasing

Author: F. Trhlik

Language models suffer extensively from encoding biases. However, most current bias-mitigation techniques address those biases only after pre-training. This research investigates BabyLMs, models trained on small-scale, child-aligned corpora, to determine whether they can provide a promising, computationally efficient sandbox for examining the full lifecycle of bias in LMs. We probe BabyLM variants and standard LM baselines for shared bias patterns across their corpora and layers, benchmark both post-hoc and in-training mitigation algorithms, and investigate how BabyLMs could democratise pre-model debiasing. Together, these experiments map the bias–performance trade-offs in small-corpus models and aim to position BabyLMs as an accessible platform for cost-effective, transparent bias mitigation.This cross-lingual work corroborates statistical learning theories of acquisition and empirically motivates new methods for training subword tokenizers.

PRESENTERS

ADA COMPUTER SCIENCE

Sol Dubock

ALTA INSTITUTE

Øistein Andersen
Luca Benedetto
Andrew Caines
Mark Elliott
Diana Galván Sosa
Yuan Gao
Russell Moore
Suchir Salhan
David Strohmaier
Shiva Taslimipoor
Filip Trhlik

COMPUTER SCIENCE & TECHNOLOGY

Richard Diehl Martinez
Guy Emerson
Zeb Goriely

RPCERC

Manni Cheung
Katharine Childs
Jessie Durk
Laurie Gale
Claire Johnson
Diana Kirby
Carrie Anne Philbin
Bobby Whyte

Cambridge ALTA

Institute for Automated Language Teaching and Assessment

RASPBERRY PI
COMPUTING EDUCATION RESEARCH CENTRE

ada

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF
COMPUTER SCIENCE AND TECHNOLOGY

Edited by Kata Szabo