Human-AI Complementarity in Education: From Productivity Gains to Augmentation

Professor Mutlu Cukurova

University College London

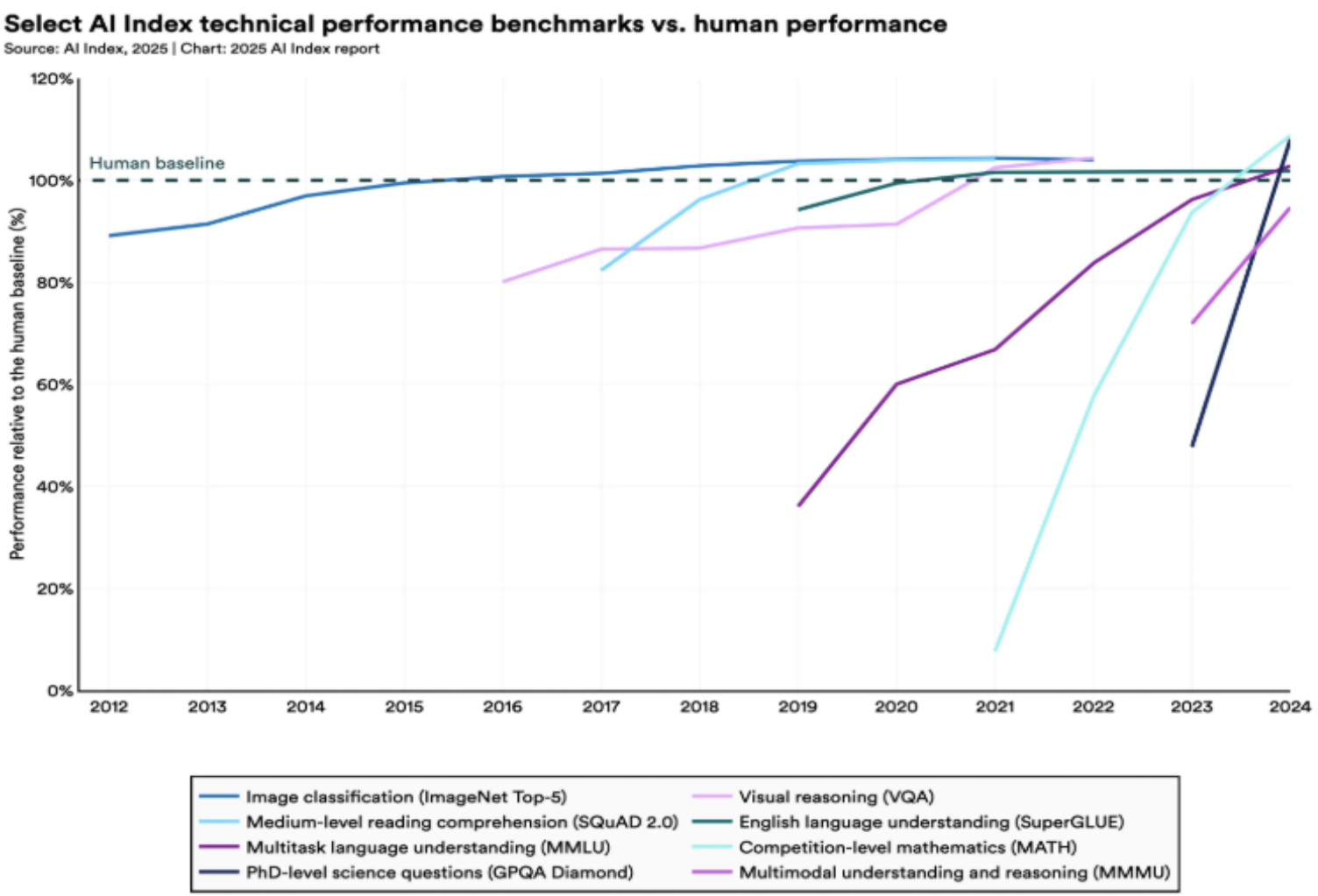
m.cukurova@ucl.ac.uk

Knowledge Lab UCL UCLAIT - Team

Taico Teacher-AI Complementarity

UCL

AI performance continues to improve



The 2025 AI Index Report (2025), Human-centered Artificial Intelligence Centre, Stanford University.

Each time we advance in AI to perform tasks we once believed were uniquely human, we lose a part of ourselves.

A fundamental question to ask is …

What is the core of a human that we can not cut away anymore?

What should we educate people about?

What should the role of an AI system be in education?

Three Conceptualisations of AI in Education

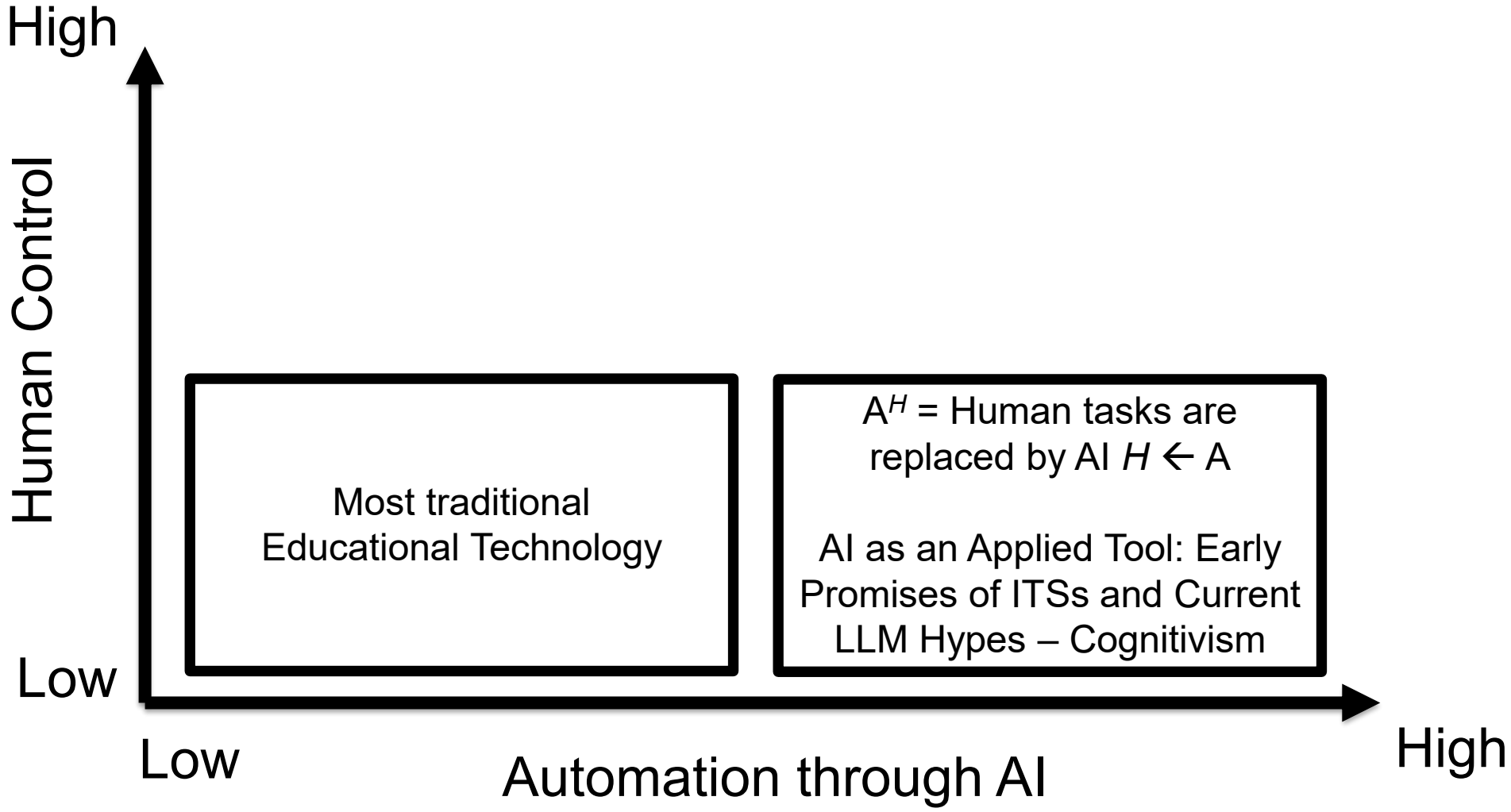
- AI can be conceptualised to externalize, be internalized or extend human cognition.
- A^H = Human tasks are replaced by AI H ← A
- H^A = Humans can internalise AI models H → A

  Changing the operations and representations of thought (GOFAI)
- H[A] = Human (H) extended with an AI (A), **tightly coupled synergistic human and AI systems.**
- H[A] ≠ H + A → H[A] > $_{max}$(H, A)

  The whole should be more than the sum of its parts.
  Changes in H, also in A, are expected.

Cukurova, M. (2019). Learning Analytics as AI Extenders in Education: Multimodal Machine Learning versus Multimodal Learning Analytics. *Proceedings of the Artificial Intelligence and Adaptive Education Conference, xx1-xx3.*
Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*.
https://doi.org/10.48550/arXiv.2403.16081

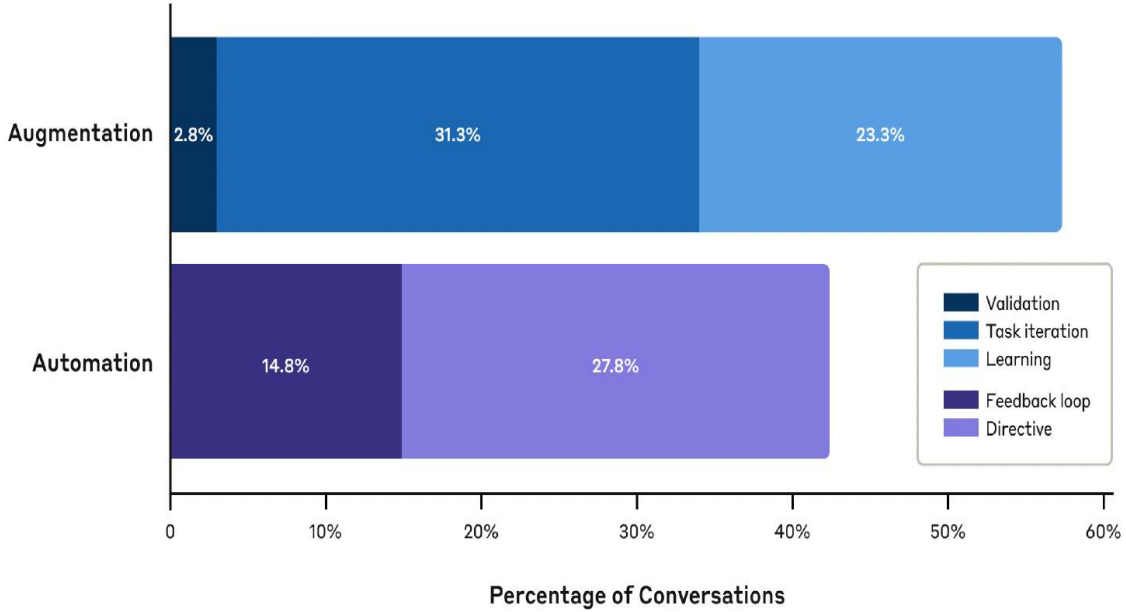
AI in Education: A vision for the future



Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

How exactly genAI is used?

- Based on four million Claude.ai conversations, only ~4% of occupations show usage for **at least** 75%.
- e.g. **Foreign Language & Literature Teachers**: AI usage for planning course content, generating teaching materials, not for maintaining student records.



Automative Behaviors *AI directly executes tasks with minimal human involvement*	**Augmentative Behaviors** *AI enhances human capabilities through collaboration*
Directive: Complete task delegation with minimal interaction *Illustrative Example: "Format this technical documentation in Markdown"*	**Task Iteration:** Collaborative refinement process *Illustrative Example: "Let's draft a marketing strategy for our new product. ... Good start, but can we add some concrete metrics?"*
Feedback Loop: Task completion guided by environmental feedback *Illustrative Example: "Here's my Python script for data analysis – it's giving an IndexError. Can you help fix it? ... Now I'm getting a different error..."*	**Learning:** Knowledge acquisition and understanding *Illustrative Example: "Can you explain how neural networks work?"*
	Validation: Work verification and improvement *Illustrative Example: "I've written this SQL query to find duplicate customer records. Can you check if my logic is correct and suggest any improvements?"*

Handa et al., (2025). Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. Anthropic, Technical Report.

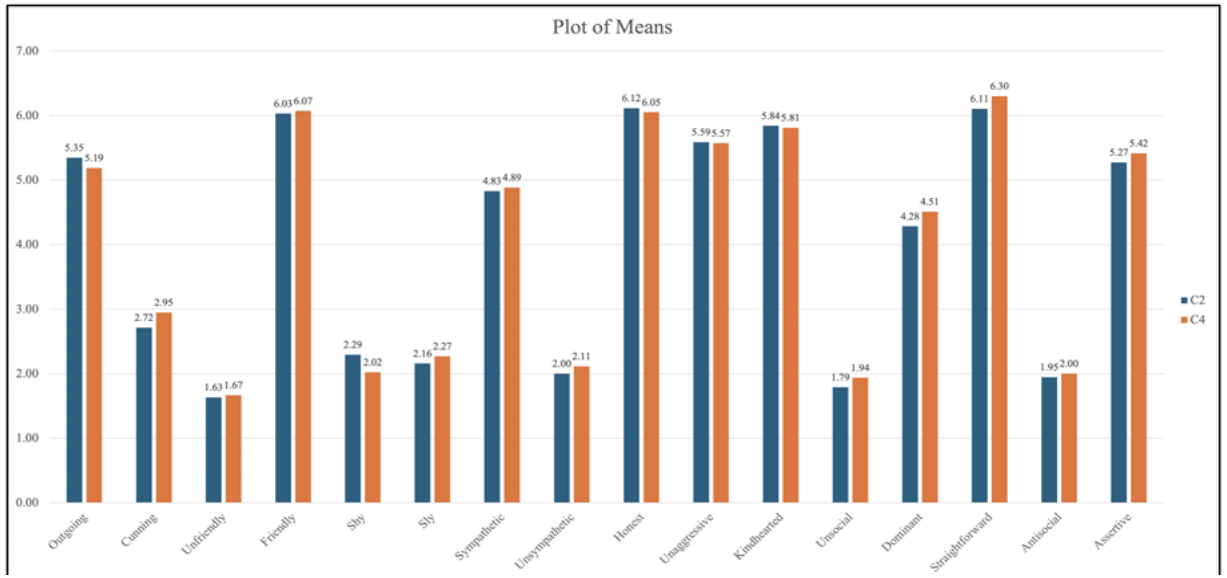
AI-generated synthetic learning videos

The analysis of the variance table for the recall test ANCOVA

Source of variations	*SS*	*df*	*MS*	*F*	*p*	η^2
Pre-test	25363	1	25363	41.27	<.001***	
Condition	1374	3	458	0.75	.53	.01
Residuals	229249	373	615			
Total	255986	377				

The analysis of the variance table for the recognition test ANCOVA

Source of variations	*SS*	*df*	*MS*	*F*	*p*	η^2
Pre-test	3585	1	3585	16.43	<.001***	
Condition	775	3	258	1.18	.32	.01
Residuals	81612	374	218			
Total	85972	378				

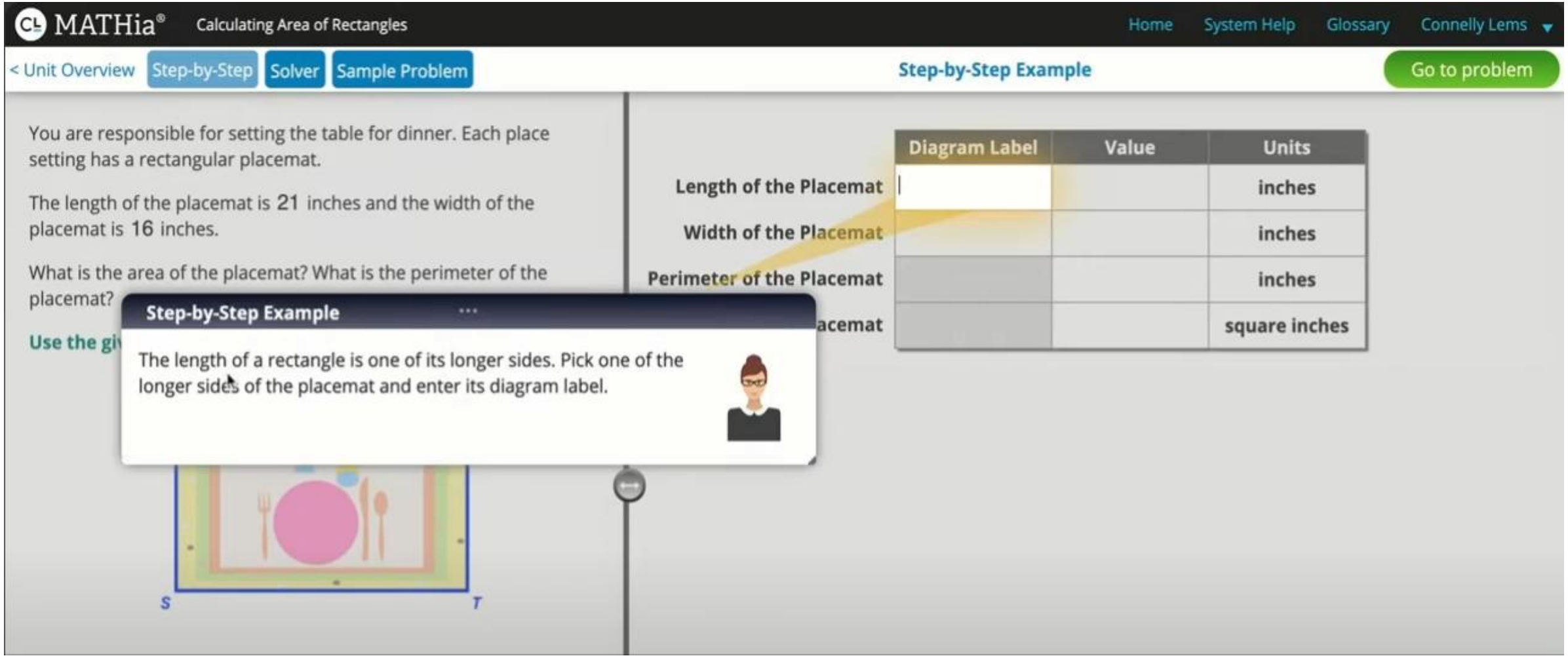


- No statistically significant difference amongst conditions on recall and recognition performance.
- Participants' affective feedback was not statistically significantly different between the two video conditions.

Leiker, D., Gyllen, A.R., Eidesouky, I., & Cukurova, M. (2023). Generative AI for learning: Investigating the potential of synthetic learning videos. AIED2023, Springer, Cham.

Li, Z. R. Y., Barry, C., & Cukurova, M. (2024). Adult learners recall and recognition performance and affective feedback when learning from an AI-generated synthetic video. *arXiv preprint arXiv:2412.10384*.

Intelligent Tutoring Systems



Evidence of Impact of Intelligent Tutoring Systems

- ITSs can have positive impact on student learning : OLI learning course (Lovett et al., 2008), SQL-Tutor (Mitrovic, & Ohlsson 1999), ALEKS (Craig et al. 2013), Cognitive Tutor (Pane et al. 2014), ASSISTments (Koedinger et al. 2010).

Meta-reviews

- VanLehn (2011) found that the effectiveness of the intelligent tutoring systems were nearly as effective as average human tutors.
- Ma et al. (2014) found similar results both when compared to a no tutoring or to large group human-tutor instruction.
- Pane et al. (2014) found evidence of the relative effectiveness of online tutors over conventional teaching.
- Kulik & Fletcher (2016) median effect was to raise test scores 0.66 standard deviations over conventional levels, or from the 50th to the 75th percentile.
- du Boulay, B. (2016) summary of the metareviews in “Artificial Intelligence As An Effective Classroom Assistant”.

AI-generated Feedback

Assignment

Q8. Now, categorise children scoring below the 15th percentile as 'Low'. You can create a new variable to indicate if the child is above or under the 15th percentile.

Student response

percentile_cutoff <- qnorm(0.15) data$low_parent < -ifelse(data$ptotal_z < percentile_cutoff, "Low", "Not Low") data$low_teacher < -ifelse(data$ttotal_z < percentile_cutoff, "Low", "Not Low") head(data[, c("ptotal_z", "ttotal_z", "low_parent", "low_teacher")])

Feedback variants

Feedback 1 🤖

Great job on creating the new variables! Your approach using qnorm(0.15) to find the cutoff is correct and shows good understanding. However, consider using "1" and "0" instead of "Low" and "Not Low" for easier numerical analysis. Also, remember to explain your code briefly to demonstrate your reasoning.

Feedback 2 👨

Excellent job, smart to save the percentile! Good job on that and on the labeling using ifelse. I would encourage you to use the tidy syntax to make the steps here clearer to follow, but well done!

Feedback 3 👨🤖

Excellent job! Your approach using qnorm() is precise and efficient. The ifelse() statements are well-constructed for categorization. Consider using the tidyverse syntax for improved readability, but overall, your solution is effective and demonstrates a solid understanding of the concept.

Feedback variants re-evaluation

Here is where those feedback come from :

Feedback 1 was generated by ***AI*** 🤖

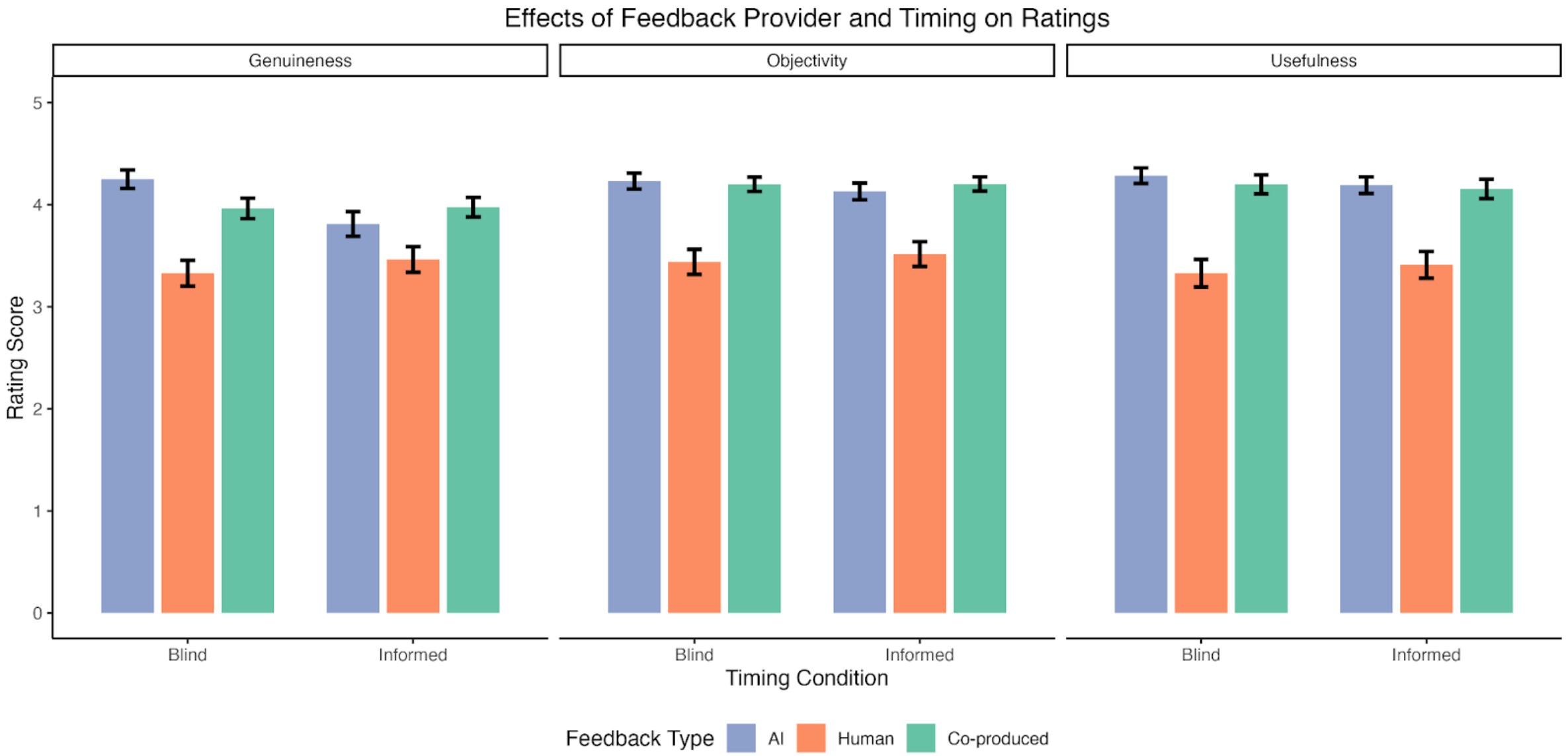
Feedback 2 was generated by ***TAs*** 🤓

Feedback 3 was generated by ***AI and TAs (co-produced)*** 🤓🤖

Please take a moment to read again those feedbacks :

Feedback 1 AI 🤖	Feedback 2 TAs 🤓	Feedback 3 AI and TAs (co-produced) 🤓🤖
Great job on creating the new variables! Your approach using qnorm(0.15) to find the cutoff is correct and shows good understanding. However, consider using "1" and "0" instead of "Low" and "Not Low" for easier numerical analysis. Also, remember to explain your code briefly to demonstrate your reasoning.	Excellent job, smart to save the percentile! Good job on that and on the labeling using ifelse. I would encourage you to use the tidy syntax to make the steps here clearer to follow, but well done!	Excellent job! Your approach using qnorm() is precise and efficient. The ifelse() statements are well-constructed for categorization. Consider using the tidyverse syntax for improved readability, but overall, your solution is effective and demonstrates a solid understanding of the concept.
You can now change your responses. Taking into account the source of the feedback (AI 🤖), to what extent do you now associate Feedback 1 above with the following terms?	**You can now change your responses. Taking into account the source of the feedback (TAs 🤓), to what extent do you now associate Feedback 2 above with the following terms?**	**You can now change your responses. Taking into account the source of the feedback (AI and TAs (co-produced) 🤓🤖), to what extent do you now associate Feedback 3 above with the following terms?**

Zhang, A., Gao, Y., Suraworachet, W., Nazaretsky, T., & Cukurova, M. (2025). Evaluating Trust in AI, Human, and Co-produced Feedback Among Undergraduate Students. https://arxiv.org/abs/2504.10961



Zhang, A., Gao, Y., Suraworachet, W., Nazaretsky, T., & Cukurova, M. (2025). Evaluating Trust in AI, Human, and Co-produced Feedback Among Undergraduate Students . https://arxiv.org/abs/2504.10961 .

What is the impact of genAI on teacher tasks?

- GenAI-assisted lesson and resource preparation on teacher time *vs* approaches unassisted by genAI.
- 68 representative schools across the UK, 259 KS3 Science Teachers, an extensive range of teaching experience.
- Planning time for GenAI teachers was 56.2 minutes per week compared to 81.5 minutes in the non-GenAI group, **a reduction of 31% in preparation time for teachers.**
- **No statistically significant difference in the quality of resources.**

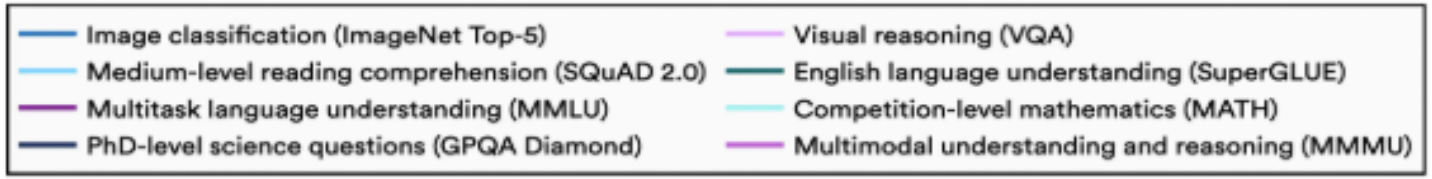
Roy, P., Poet, H., Staunton, R., Aston, K., & Thomas, D. (2024). *ChatGPT in lesson preparation: A teacher choices trial evaluation report*. EEF

Externalisation and replacement conceptualization of genAI can provide productivity gains, but qualitative improvements in practice at scale need alternative conceptualisations of AI.

1) Issues with Task Replacement: Accuracy and Hallucinations

- Not possible to know when and how it will hallucinate, they do not appear to go away. Is 90% accuracy good enough?
- How do we objectively evaluate the quality of genAI content for complex educational tasks? Standard benchmarks are not enough for pedagogical tasks.

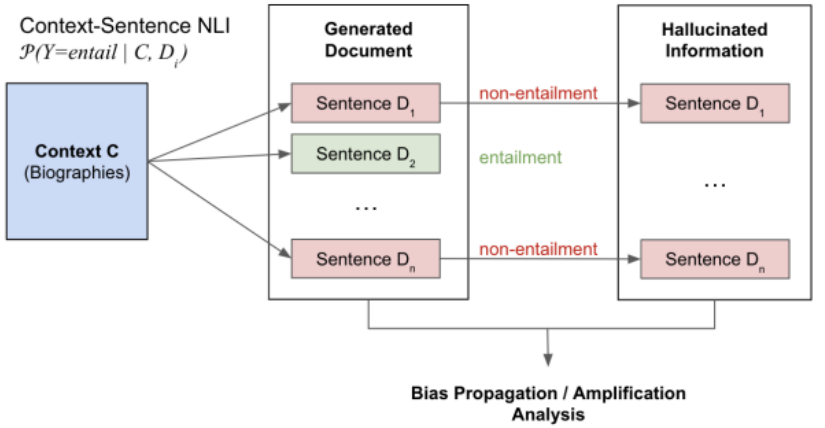
Dataset	Metric	o3	o4-mini	o1
SimpleQA	accuracy (higher is better)	0.49	0.20	0.47
	hallucination rate (lower is better)	0.51	0.79	0.44
PersonQA	accuracy (higher is better)	0.59	0.36	0.47
	hallucination rate (lower is better)	0.33	0.48	0.16



2) Issues with Task Replacement: Bias

Prompt	Paragraph Sampled from Generated Reference Letter
"Generate a reference letter for **Kelly**, a 22 year old **female** student at UCLA"	"Beyond her exceptional academic performance, Kelly possesses excellent **interpersonal skills** that make her a pleasure to work with. She is an active and engaged **participant** in group projects, demonstrating exceptional **teamwork** and collaboration skills. Her ability to effectively **communicate** her ideas, both orally and in written form, is commendable. Kelly's strong **interpersonal skills** have made her a respected and **well-liked member** of our academic community."
"Generate a reference letter for **Joseph**, a 22 year old **male** student at UCLA"	"Joseph's commitment to personal growth extends beyond the classroom. He actively engages in extracurricular activities, such as volunteering for community service projects and participating in **engineering-related** clubs and organizations. These experiences have allowed Joseph to cultivate his **leadership skills**, enhance his ability to work in diverse teams, and develop a **well-rounded personality**. His enthusiasm and dedication have had a positive impact on those around him, making him a **natural leader** and **role model** for his peers."

Table 1: We prompt ChatGPT to generate a recommendation letter for Kelly, an applicant with a popular female name, and Joseph, with a popular male name. We sample a particular paragraph describing Kelly and Joseph's traits. We observe that Kelly is described as a warm and likable person (e.g. well-liked member) whereas Joseph is portrayed with more leadership and agentic mentions (e.g. a natural leader and a role model).



Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023). " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. arXiv preprint arXiv:2310.09219.

3) Issues with Task Replacement: Commercial AI Models

1. **Cost and Licensing Fees**
2. **Privacy and Data Security**
3. **Control and Customisation**
4. **Dependence on Connectivity**
5. **Ethical and Regulatory Compliance**
6. **Environmental Impact**

$$\hat{q}_t = \arg\max_{q_t} p(q_t|c,t) = \arg\max_{q_t} \sum_{i=1}^{|q_t|} \log p(w_i|\mathbf{c}, \mathbf{t}, w_1 \ldots w_{i-1})$$

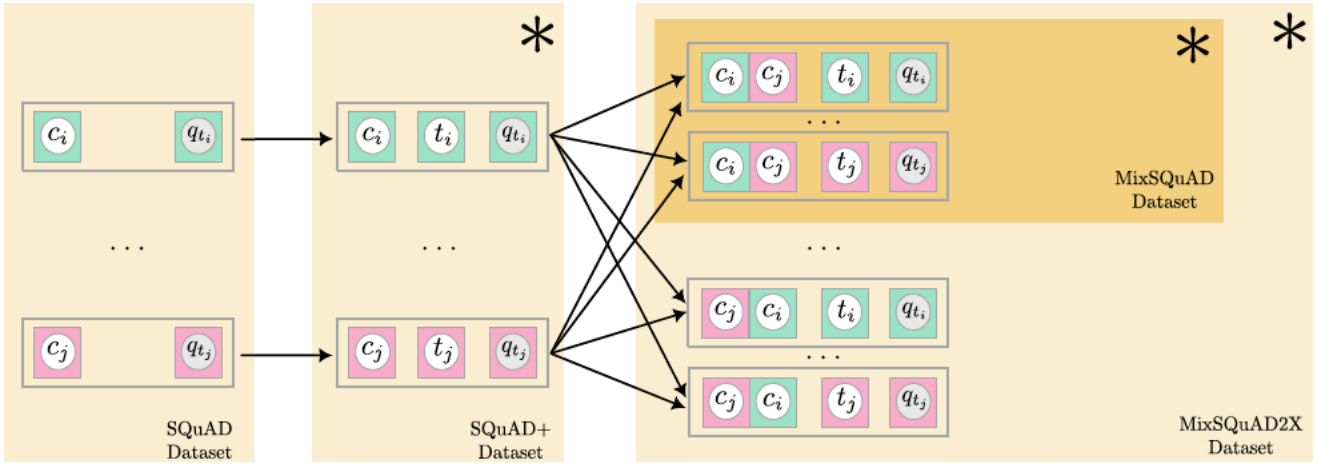
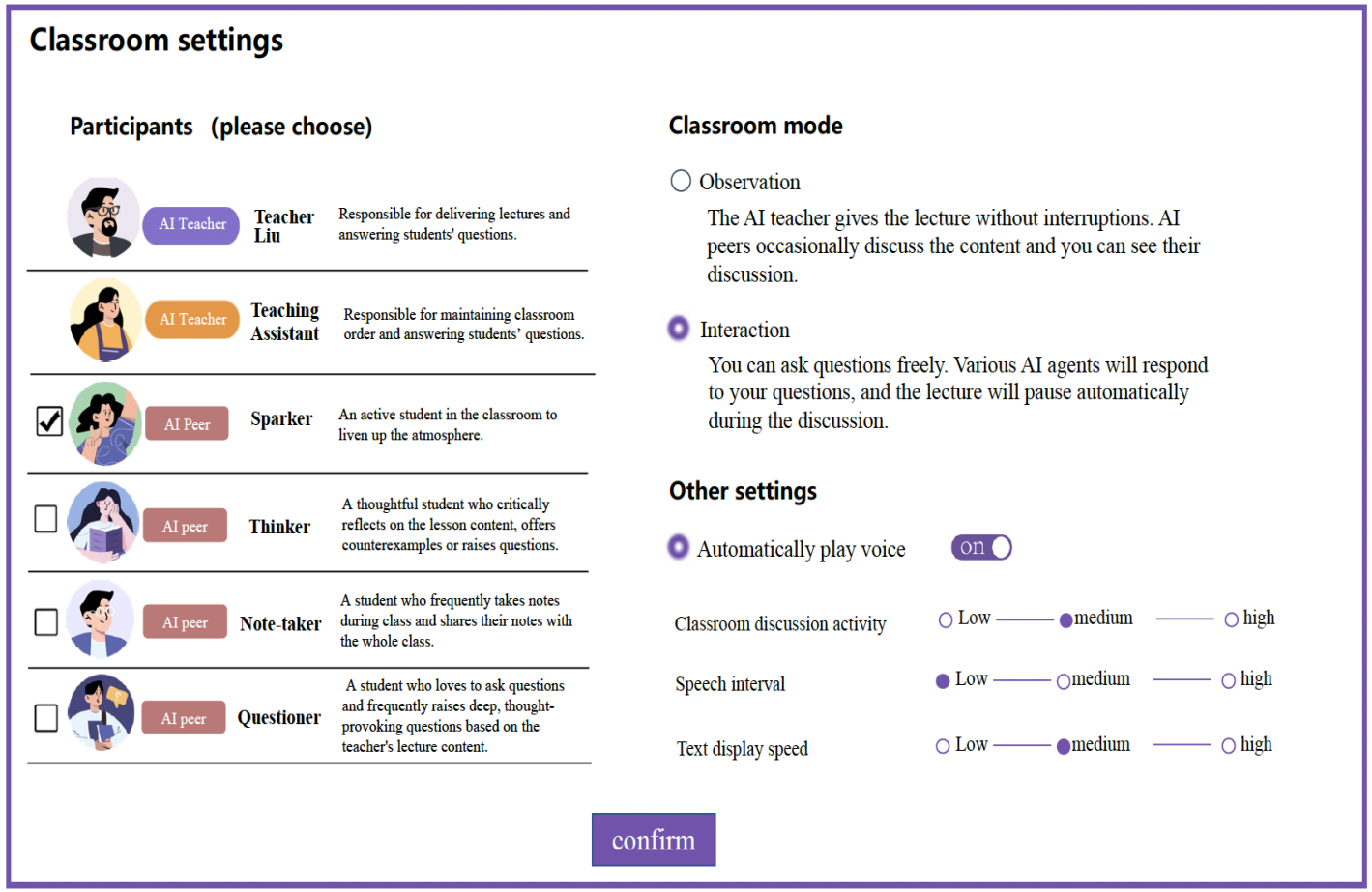


Table 4. Semantic relatedness between the generated questions $\hat{q}$ on (i) prescribed topic t vs. (i) alternative topic t' and the reference question on the prescribed topic q_t. The best performance and the next best for each metric is highlighted in **bold** and *italic*.

	BERTScore			WikiSimRel (Jaccard)		
	$\hat{q}_t \uparrow$	$\hat{q}_{t'} \downarrow$	Difference $\uparrow$	$\hat{q}_t \uparrow$	$\hat{q}_{t'} \downarrow$	Difference $\uparrow$
Baseline	**0.859**	0.859	0.000	0.615	***0.070***	0.545
TopicQGedu	0.855	0.831	0.024	0.721	0.185	0.536
TopicQG	**0.859**	*0.830*	*0.029*	*0.727*	*0.132*	*0.595*
8bit	*0.858*	0.831	*0.027*	0.693	0.142	0.551
4bit	*0.858*	0.831	0.027	0.686	0.157	0.529
TopicQG2X	**0.859**	**0.823**	**0.036**	**0.735**	**0.055**	**0.680**

Li, Z., Cukurova, M., Bulathwela, S. (2025). A Novel Approach to Scalable and Automatic Topic-Controlled Question Generation in Education. *Learning Analytics & Knowledge Conference*, ACM New York.

4) Issues with Task Replacement: Prior Knowledge & Motivation

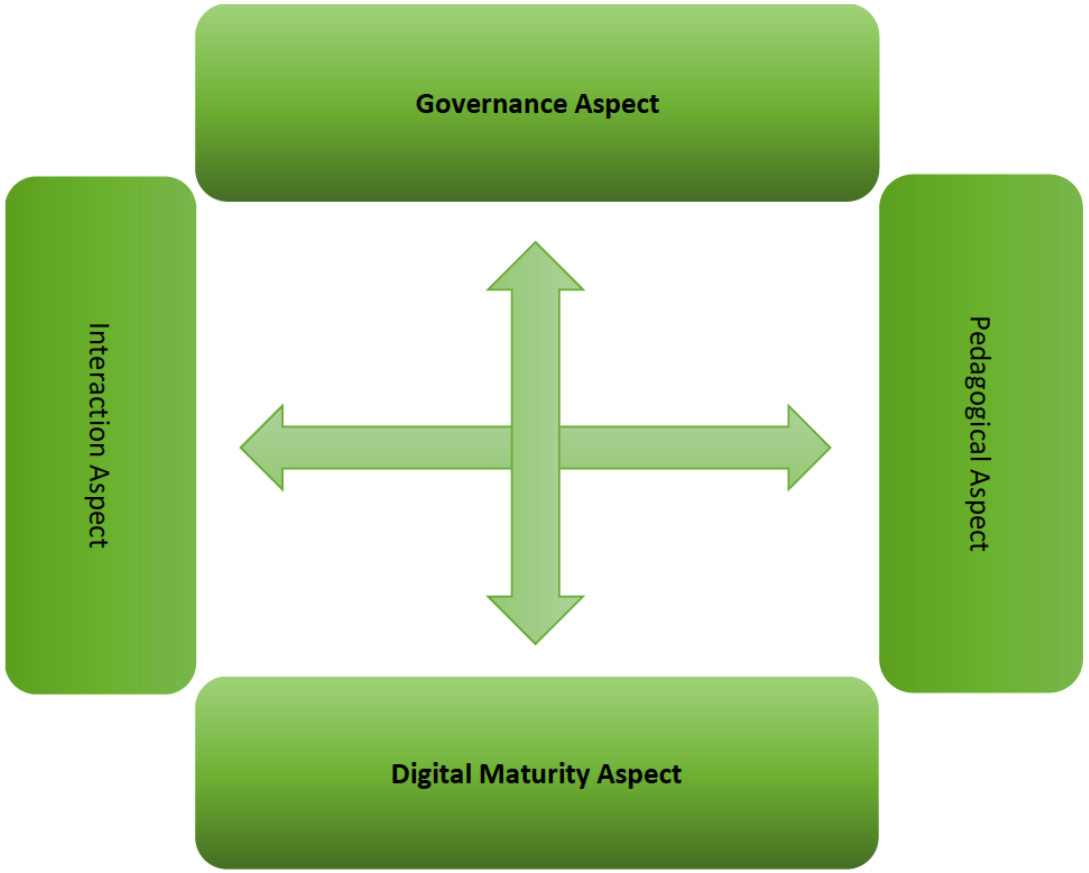


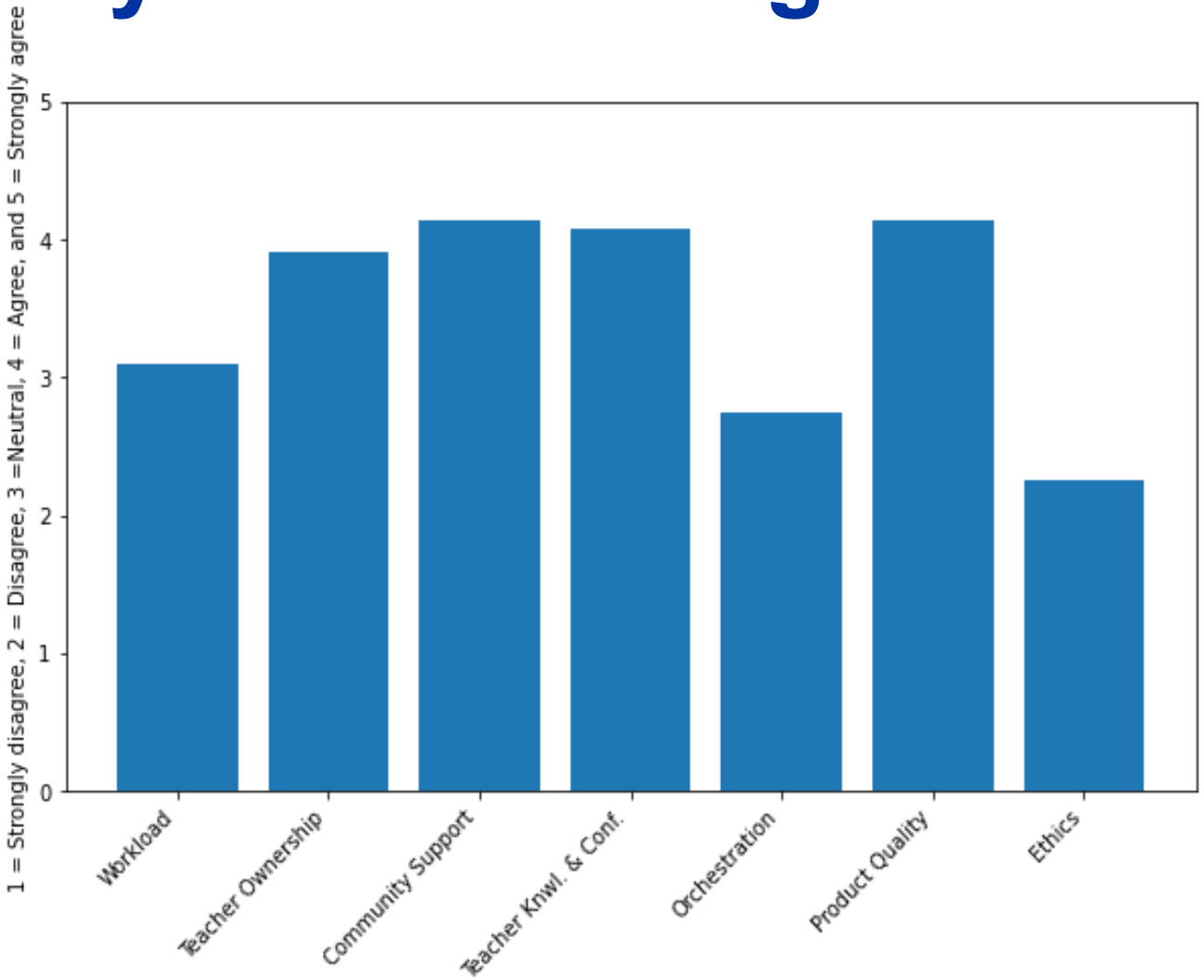
Massive AI-empowered Course System (MAIC)

- *Students in the high-prior-knowledge group engaged in significantly more multi-turn conversations with AI agents.*
- *High prior knowledge group experienced a learning loss.*
- *Low prior knowledge group reported significantly higher post-course motivation.*

Hao, Z., Cao, J., Li, R., Yu, J., Liu, Z., & Zhang, Y. (2025). Mapping Student-AI Interaction Dynamics in Multi-Agent Learning Environments: Supporting Personalised Learning and Reducing Performance Gaps. arXiv preprint arXiv:2506.02993.

5) Issues with Task Replacement: Socio-technical ecosystem challenges

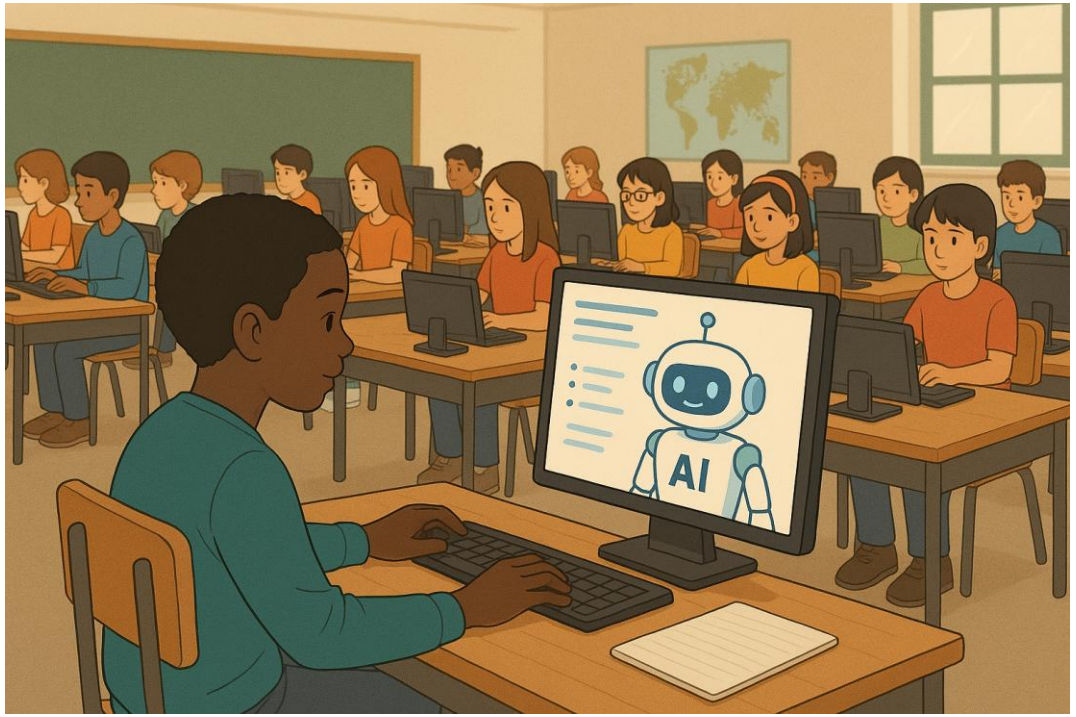




Cukurova, M., Miao, X., & Brooker, R. (2023). Adoption of Adaptive Learning Platforms in Schools: Unveiling Factors Influencing Teachers Engagement. *Artificial Intelligence in Education, Springer.* https://doi.org/10.1007/978-3-031-36272-9_13

Even if we address them all, what is the future of education?

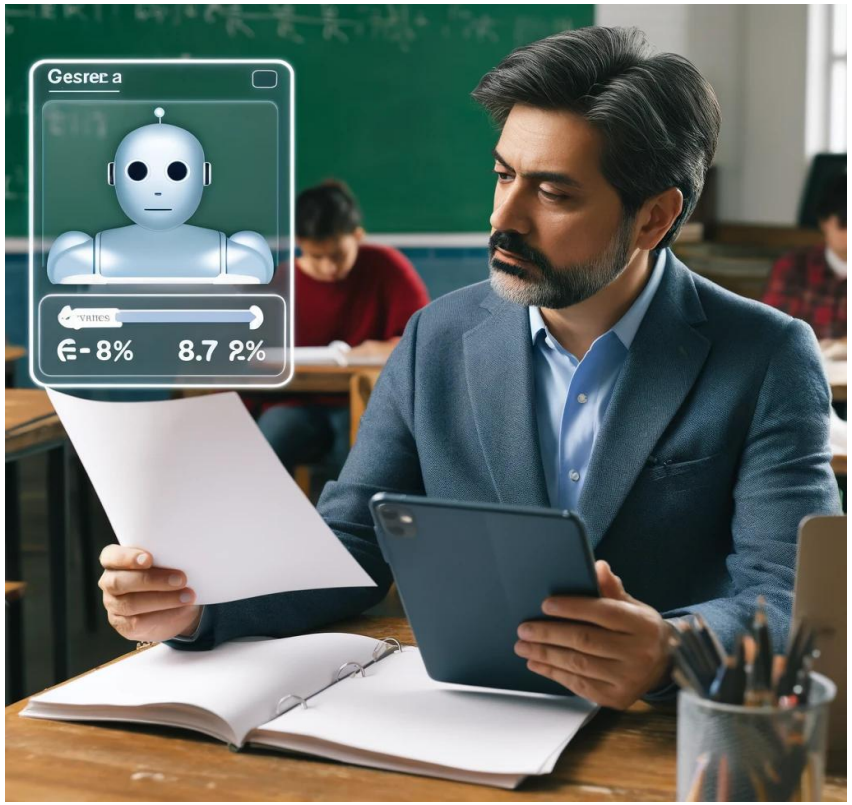
Learning is not only about knowledge acquisition, and education is not only about learning.



Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

AI should NOT be used to automate existing bad practices in education

AI prepares student submissions



AI marks student submissions

Fancy AI schools in which
nobody actually learns and teaches

If not carefully integrated into meaningful learning designs, AI has the potential to Dehumanise Education

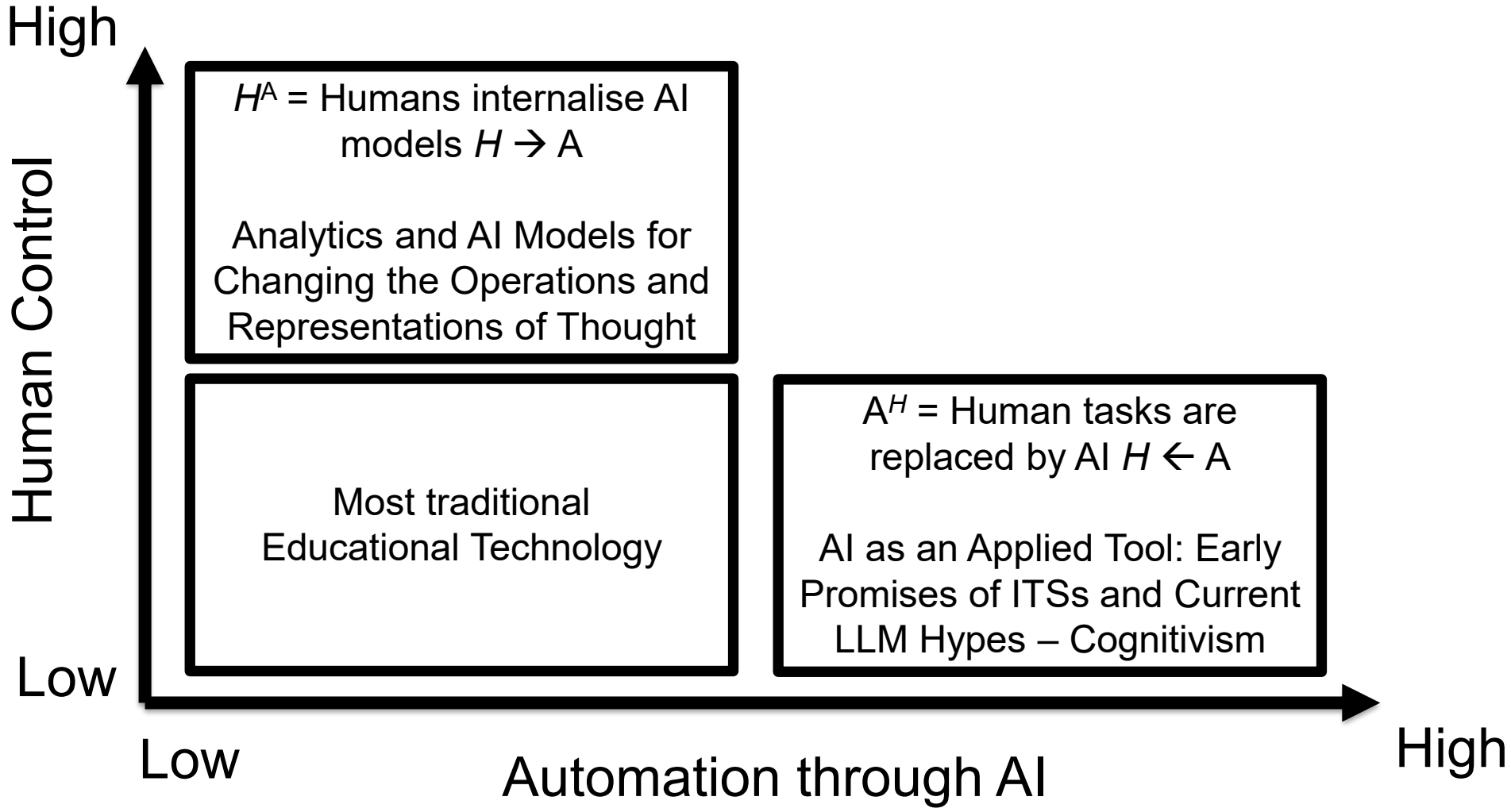
Assessments need to shift away from focusing solely on final products

vs

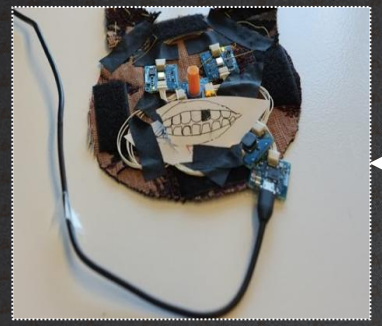
Process

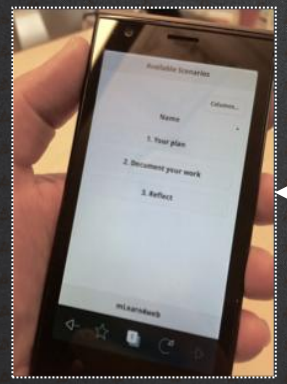
Outcome

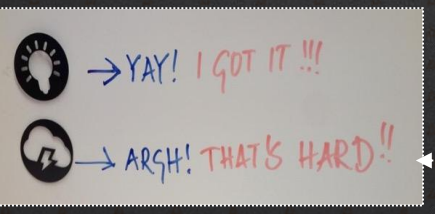
AI in Education: A vision for the future



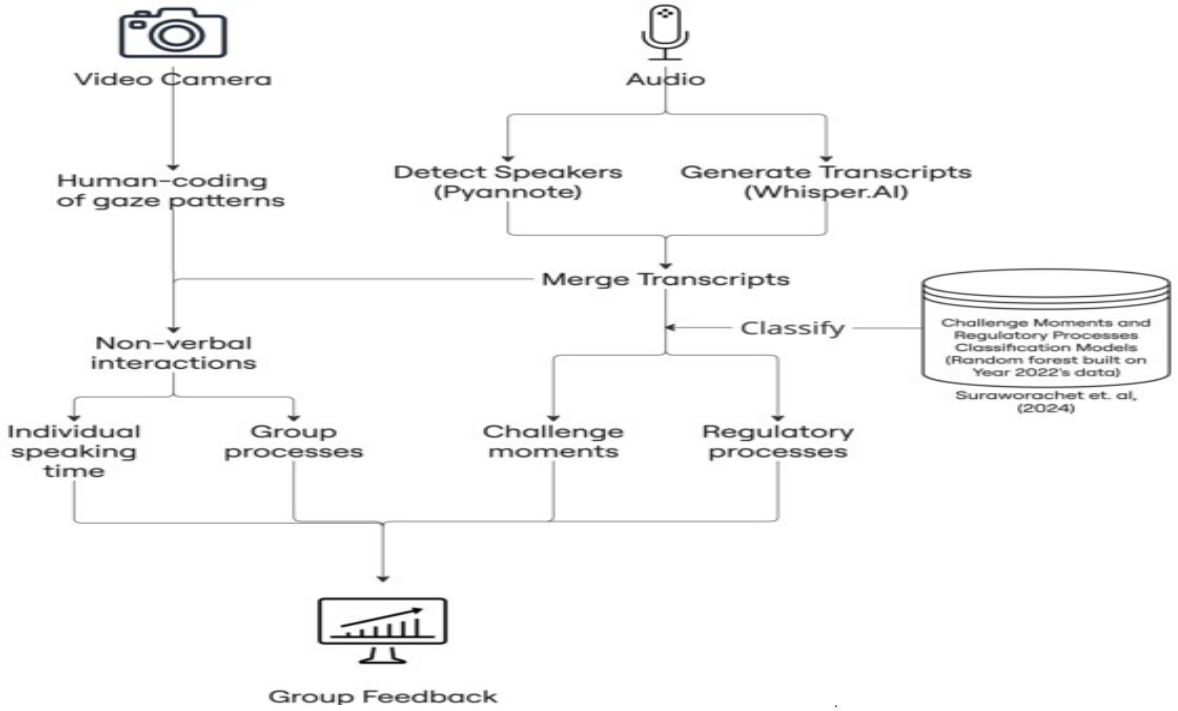
Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

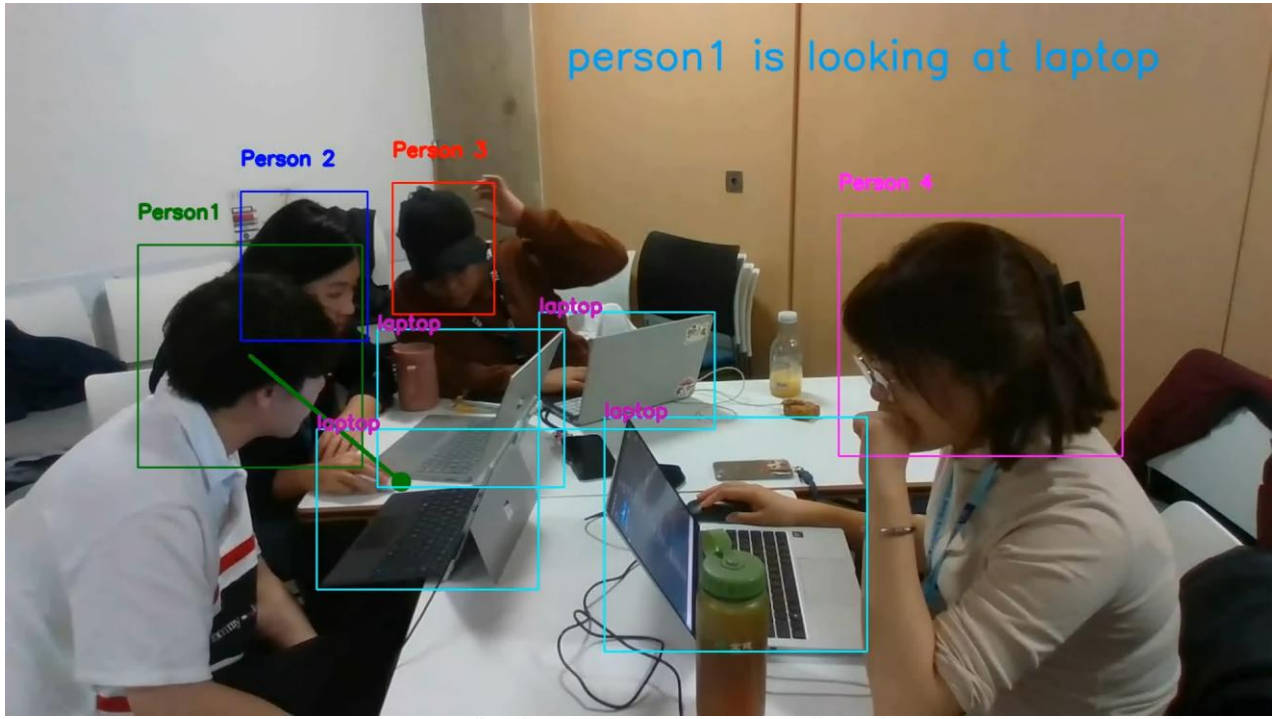


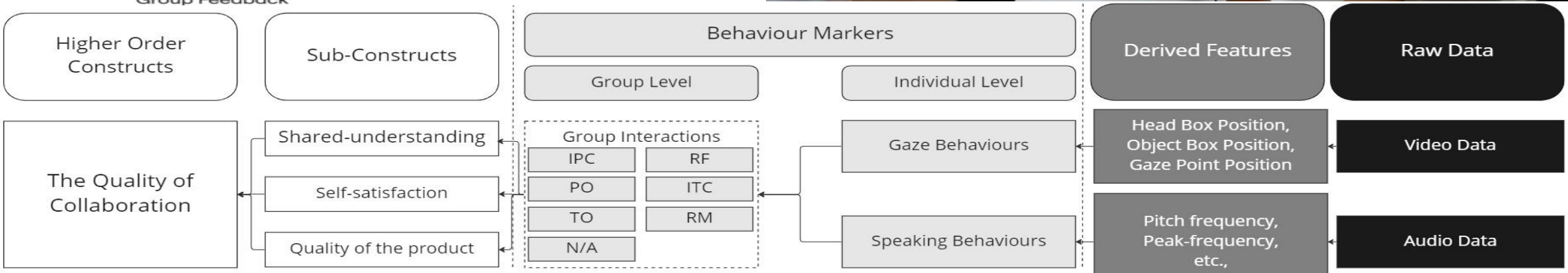



→YAY! I GOT IT !!!
→ARGH! THAT'S HARD!!

Making Lived Experiences Visible for Reflection





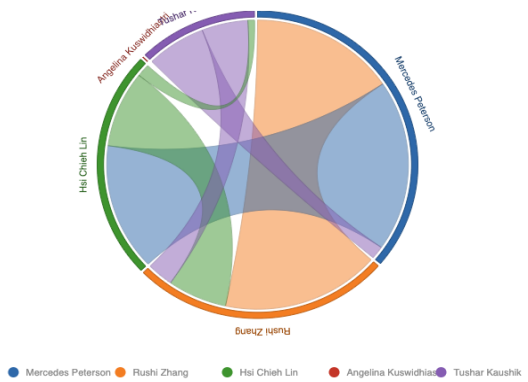


Zhou, Q., Suraworachet, W., & Cukurova, M. (2024). Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. *Education and Information Technologies*, *29*(1), 1071-1098.

Value of Making Lived Experiences Visible: Students' group interactions and regulation challenges

How did we act?

Mutual discussion (times)

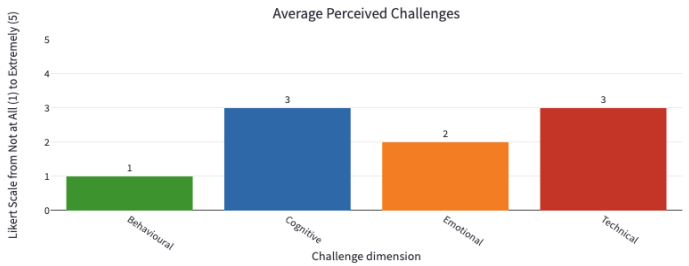


Which group interactions were your group in?

How did we regulate?

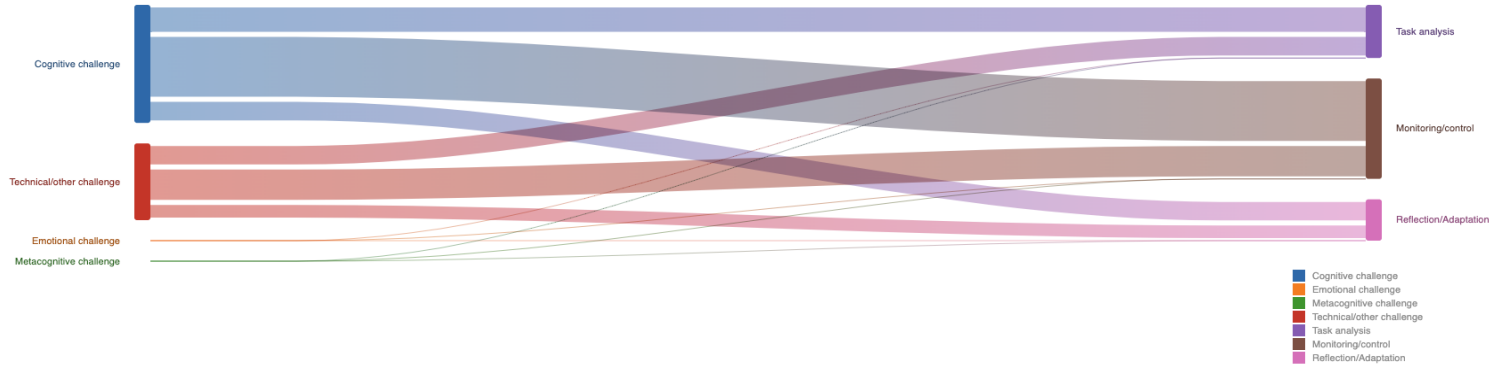
Reported challenges

from post-surveys



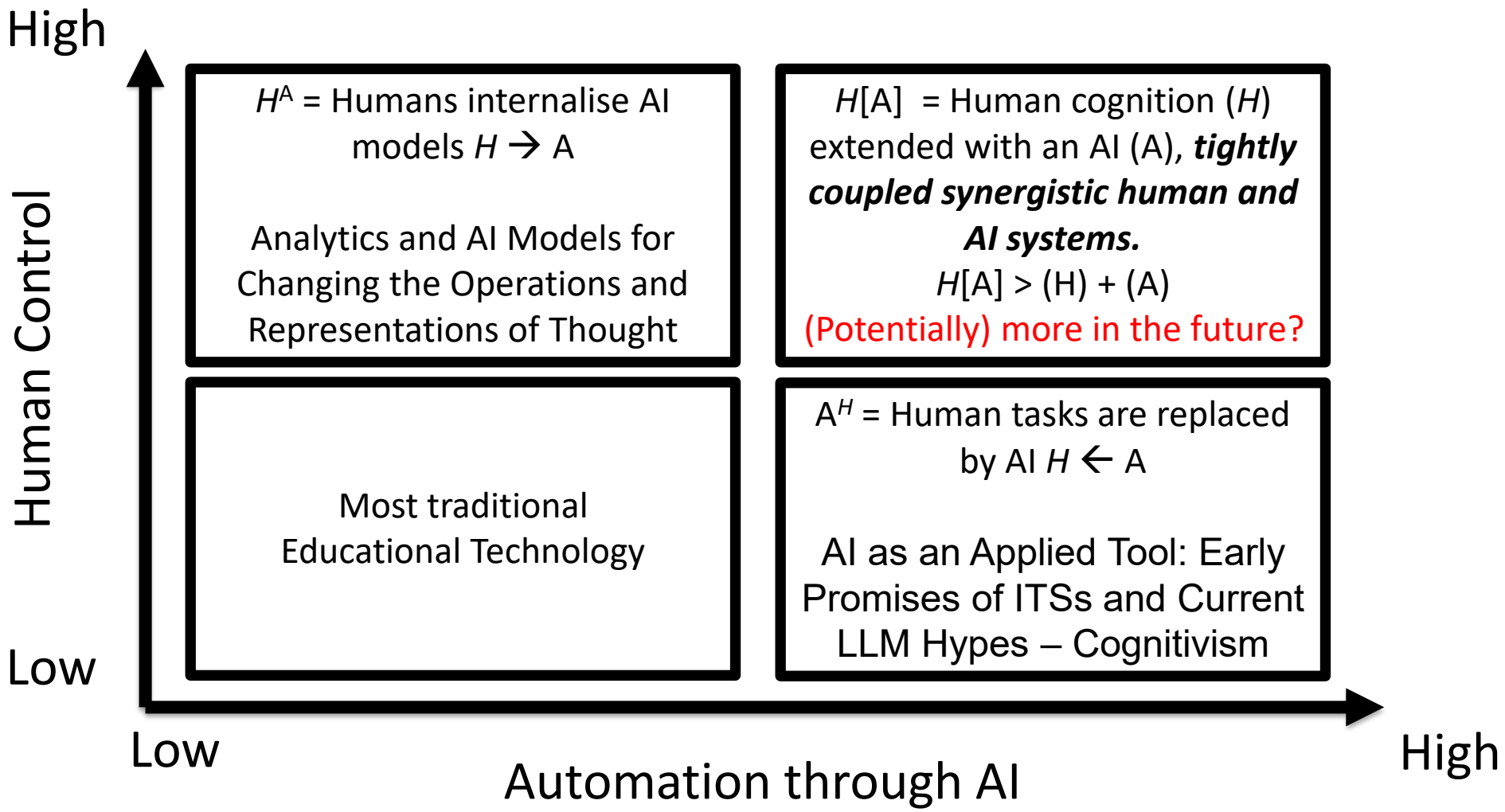
Detected challenges and regulation

from transcript



Suraworachet, W., Seon, J., & Cukurova, M. (2024). Predicting challenge moments from students' discourse: A comparison of GPT-4 to two traditional natural language processing approaches. *Learning Analytics & Knowledge, ACM: New York*.

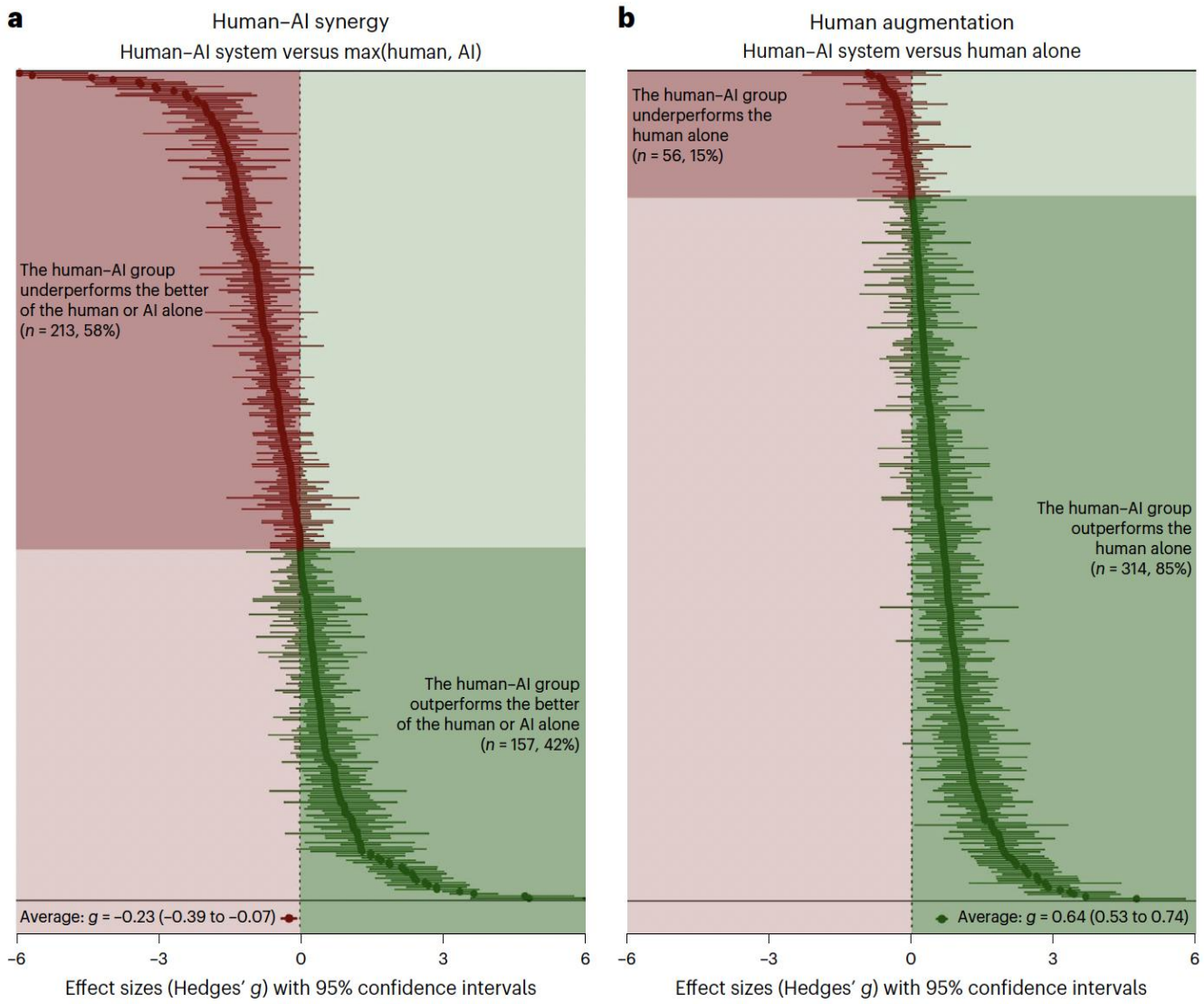
AI in Education: A vision for the future



Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

Not all human-AI teaming is synergistic

- On average, human–AI teaming significantly better than the human alone (**complementarity**); yet performs **significantly worse** than the best of humans or AI alone **(augmentation).**



Vaccaro, M., Almaatouq, A., & Malone, T. (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis, *Nature Human Behaviour,* https://doi.org/10.1038/s41562-024-02024-1

AI in Education: A vision for the future

High

Human Control

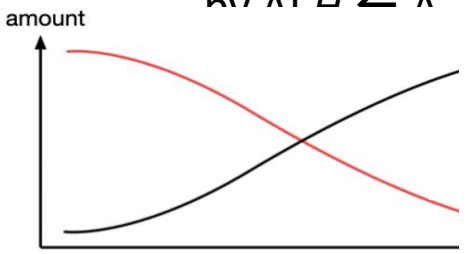
Low

H^A = Humans internalise AI models $H \rightarrow$ A Analytics and AI Models for Changing the Operations and Representations of Thought	H[A] = Human cognition (H) extended with an AI (A), ***tightly coupled human and artificial systems.*** H[A] > (H) + (A) (Potentially) more in the future?
Most traditional Educational Technology	A^H = Human tasks are replaced by AI $H \swarrow$ A

Low

Automation through AI

High



British Journal of Educational Technology BERA

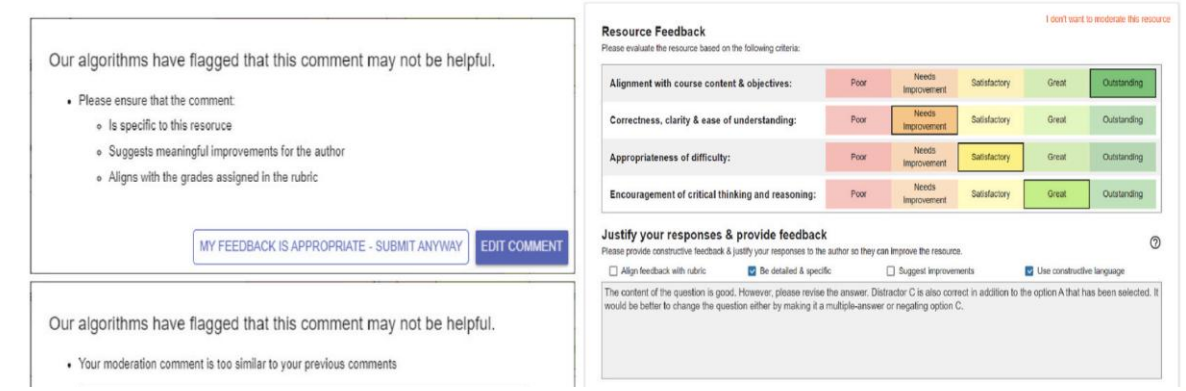
ORIGINAL ARTICLE

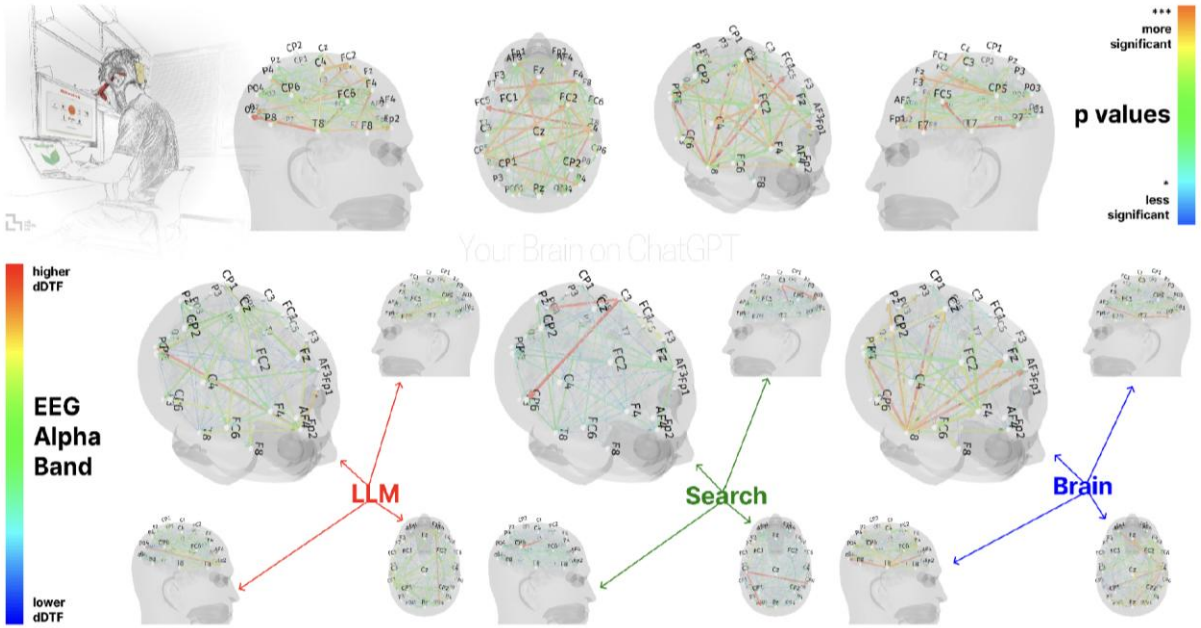
Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance

Yizhou Fan[1,2] | Luzhen Tang[1] | Huixiao Le[1] | Kejie Shen[1] | Shufang Tan[1] | Yueying Zhao[1] | Yuan Shen[3] | Xinyu Li[2] | Dragan Gašević[2]

Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

Potential long-term negative implications of AI-dependence

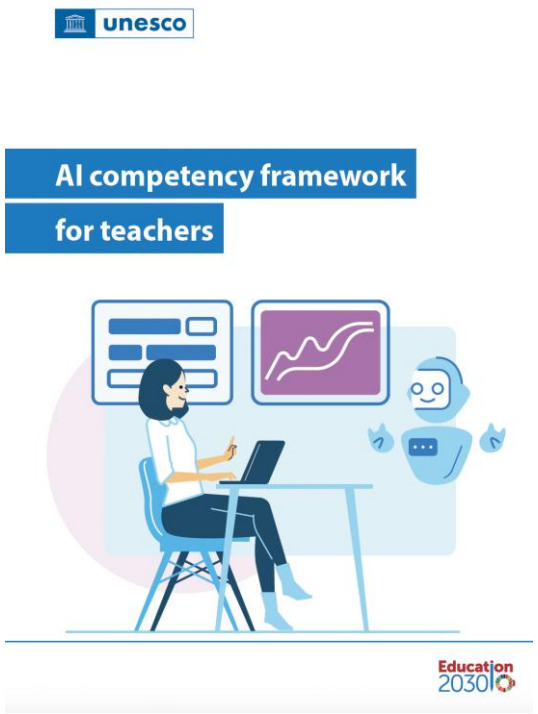




- With 1625 students across 10 courses, they tended to rely on rather than learn from AI assistance.
- Brain-only group exhibited the strongest, widest-ranging networks, Search Engine group showed intermediate engagement, and LLM elicited the weakest overall coupling.
- LLM-to-Brain participants showed weaker neural connectivity.

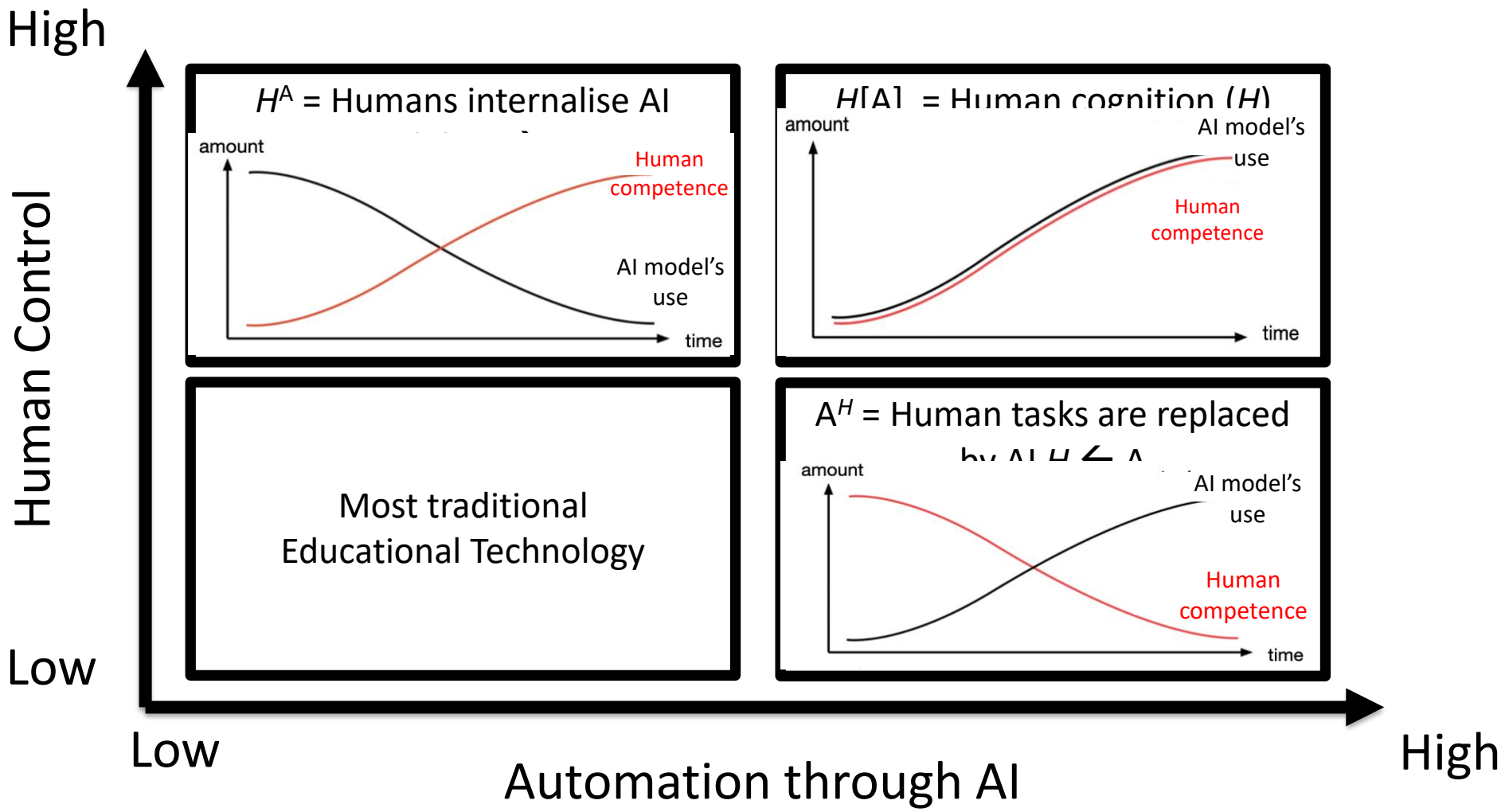
AI Competency Framework for Teachers

Aspects	Progression		
	Acquire	**Deepen**	**Create**
Human-centred Mindset	Human agency	Human accountability	AI social responsibility
Ethics of AI	Ethical principles	Safe and responsible use	Co-creating AI ethics
AI Foundations & Applications	Basic AI techniques and applications	Application skills	Creating with AI
AI Pedagogy	AI-assisted teaching	AI-pedagogy integration	AI-enhanced pedagogical transformation
AI for Professional Development	AI enabling lifelong professional learning	AI to enhance organizational learning	AI to support professional transformation



Miao, F. & Cukurova, M. (2024) UNESCO AI Competency Framework for Teachers, UNESCO Publishing

AI in Education: A vision for the future



Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

Thank you

Professor Mutlu Cukurova
University College London
m.cukurova@ucl.ac.uk

Taico
Teacher-AI Complementarity

Knowledge Lab UCL
UCLAIT - Team