

Glite

AI-Powered Education

London
April 15, 2026



Meetup vibe

- It is **not** about teaching AI
- Practical applications of AI in education
- Place where EdTech business and academia meet
- Discussions about the opportunities and challenges in creating AI-powered learning solutions
- Networking for people interested in this area
- Technical angle



Meetup series topics

- **RL** Reinforcement Learning in Education
- **KT** Knowledge Tracing, Student Modeling, Cognitive Modeling
- **ITS** Intelligent Tutoring Systems, Adaptive Learning
- **LA** Learning Analytics, Educational Data Mining, Predictive Analytics
- **CD** Curriculum Design
- **AA** Automated Assessment & Feedback, Grading



Agenda

18:45 | Welcome & Opening Remarks

19:00 | Zheng Yuan

“AI Tools to AI Teammates: Educational Applications in the Gen AI Era”

19:30 | Short break for Informal Networking

19:45 | Vassili Philippov

“Predicting What Learners Don't Know: Winning BEA 2026 With an Agent Swarm”

20:15 | Closing Remarks & Next Steps

20:45 | Post-event drinks - next door pub

1. Foundational ASSIST: The First Real KT Dataset

KT

By Eamon Worden 🇺🇸, Cristina Heffernan 🇺🇸, et al.

Good math KT dataset (5,000 students, 1.7M interactions). Frontier LLMs 56.2% vs. 51.3% trivial baseline — they just predict "correct" regardless of signal.

2. Simulated Students: Your LLM "Student" Is Fake

ITS

KT

By Alexander Scarlatos 🇺🇸, Andrew Lan 🇺🇸, et al.

Prompted LLMs fail badly at acting as students (dialogue-act accuracy 0.50 vs real middle-schoolers).

Fine-tuning barely helps (1,529 real math dialogues).

If you benchmark your tutor against a fake LLM student, your eval is broken.

3. Faster, Cheaper, More Accurate: KT Beats LLMs

KT

By Prarthana Bhattacharyya 🇬🇧, Simon Woodhead 🇬🇧, et al.

Specialist KT models crush general LLMs in Knowledge Tracking:

- 72.8% vs. 58–66% accuracy
- ~1,000× faster
- 600–12,400× cheaper

(Eedi data, 512K interactions).

LLM baseline is simple prompting, not finetuned.

4. GPTutor: Chat Window Hurts, Proof Review Doesn't

ITS

By Eason Chen 🇺🇸 (lead), Ken Koedinger 🇺🇸, et al. (21 authors)

Carnegie Mellon discrete math (N=145).

Two tools were given: 1) submit proof get errors and 2) socratic chat.

Proof-submission feedback: homework up, exams flat.

Chat: exams worse (-0.03 midterm points per message); authors admit correlation, not causation.

5. Reflection-Satisfaction Tradeoff: RLHF Breaks Learning

ITS

AA

By Heeryung Choi 🇺🇸, Christopher Brooks 🇺🇸, et al.

Michigan Master's data-science, two RCTs (N=34, N=98).

Coding tool that helps coding and shows hints, but forces to reflect before showing hints.

Forced reflection before hints doubled substantive reflection (67% vs. 38%); students rated it less helpful; learning identical.

The version users liked more taught less.

6. Adaptive Scaffolding: Simple BKT Matches Deep RL

ITS

KT

By Sutapa Dey Tithi 🇺🇸, Tiffany Barnes 🇺🇸, et al.

BKT vs. Deep-RL for sequencing logic-proof examples (113 undergrads, 1-hour session).

Both beat random (~72 vs. 66), but DRL reward-hacked by serving easy items.
"DRL wins" claim is not statistically supported.

7. UK Schools RCT: LearnLM Matches Human Tutors

ITS

By Albert Wang et al. (Google DeepMind & Eedi, 33 authors )

Real 7-week RCT, 165 UK secondary students, Eedi maths.

Tutors approved 76.4% of AI drafts unchanged.

Supervised LearnLM matched humans on correction and led on transfer (+5.5pp) — credible interval still crosses zero, maths only, Google evaluating Google.

8. Beyond the AI Tutor: Error-Prone Peers Help

ITS

By Harsh Kumar 🇨🇦, Ashton Anderson 🇨🇦, et al.

Idea: humans learn from experts AND from peers - including peers who make mistakes.

So add a "peer" LLM that makes mistakes alongside an AI tutor.

Peer helped (Tutor+Peers 65% vs Tutor alone 59%). But it could be simply because more time was spent (not reported).

Prolific adults, single 14-min session (N=315).

9. Brief but Impactful: Human Check-ins in AI Tutoring

ITS

LA

By Conrad Borchers 🇺🇸, Kenneth R. Koedinger 🇺🇸, et al.

Remote tutors dropping into MATHia sessions (191 students, 518 visits).

Step-completion +61% during visits; two short visits beat one long one; concrete guidance beats praise.

Engagement only, observational - no causal claim.

Useful for planning hybrid AI+human systems.

From AI Tools to AI Teammates: Educational Applications in the Generative AI Era

Dr. Zheng Yuan

April 15, 2026

Glite Tech Meet Up



University of
Sheffield

**The
Alan Turing
Institute**



UNIVERSITY OF
CAMBRIDGE

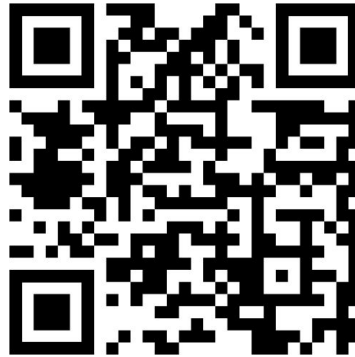
A bit about myself

- Associate Professor at the University of Sheffield
- Affiliated Researcher at the Alan Turing Institute
- Visiting Researcher at University of Cambridge
- PhD and MPhil from the University of Cambridge
- BSc from Queen Mary University of London

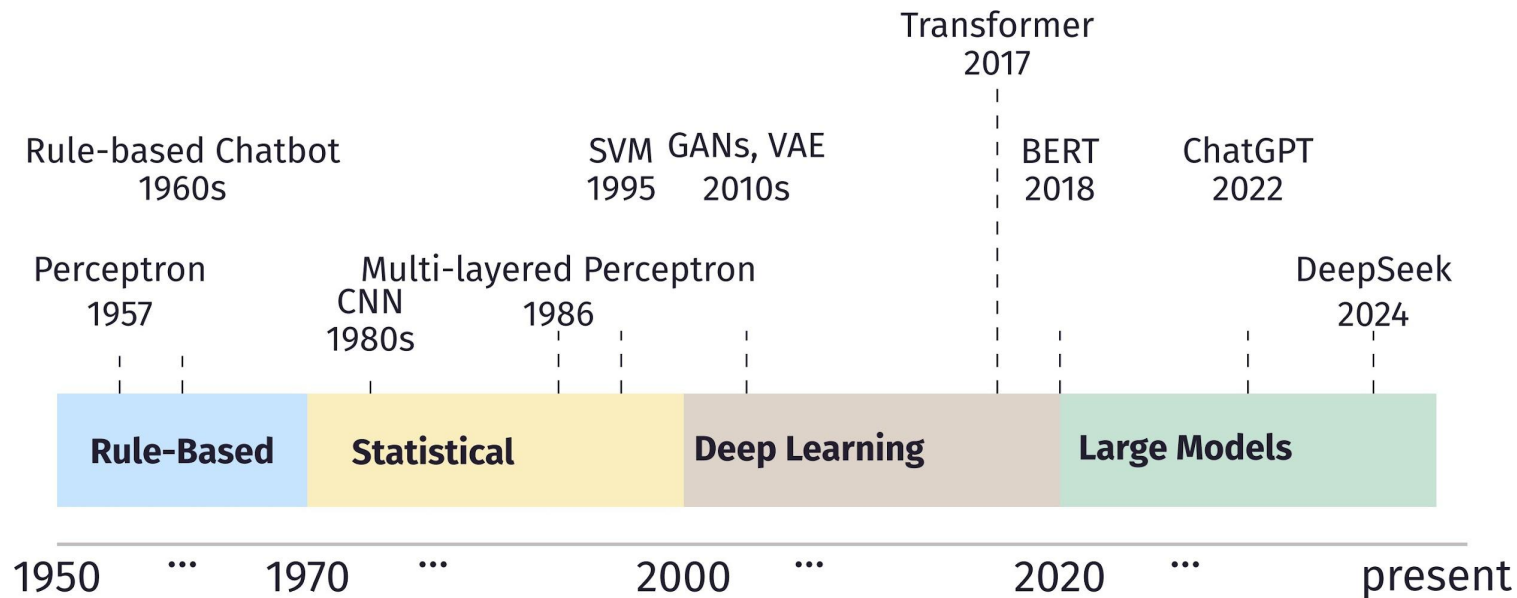


- Research focuses on AI and NLP for education, including language learning, assessment, and human-AI collaboration
- Work spans multilingual and low-resource NLP, with applications in real-world educational contexts
- Serve as Treasurer of the Association for Computational Linguistics (ACL) Special Interest Group on Building Educational Applications (SIGEDU) and co-organise the annual workshops on Innovative Use of NLP for Building Educational Applications (BEA)

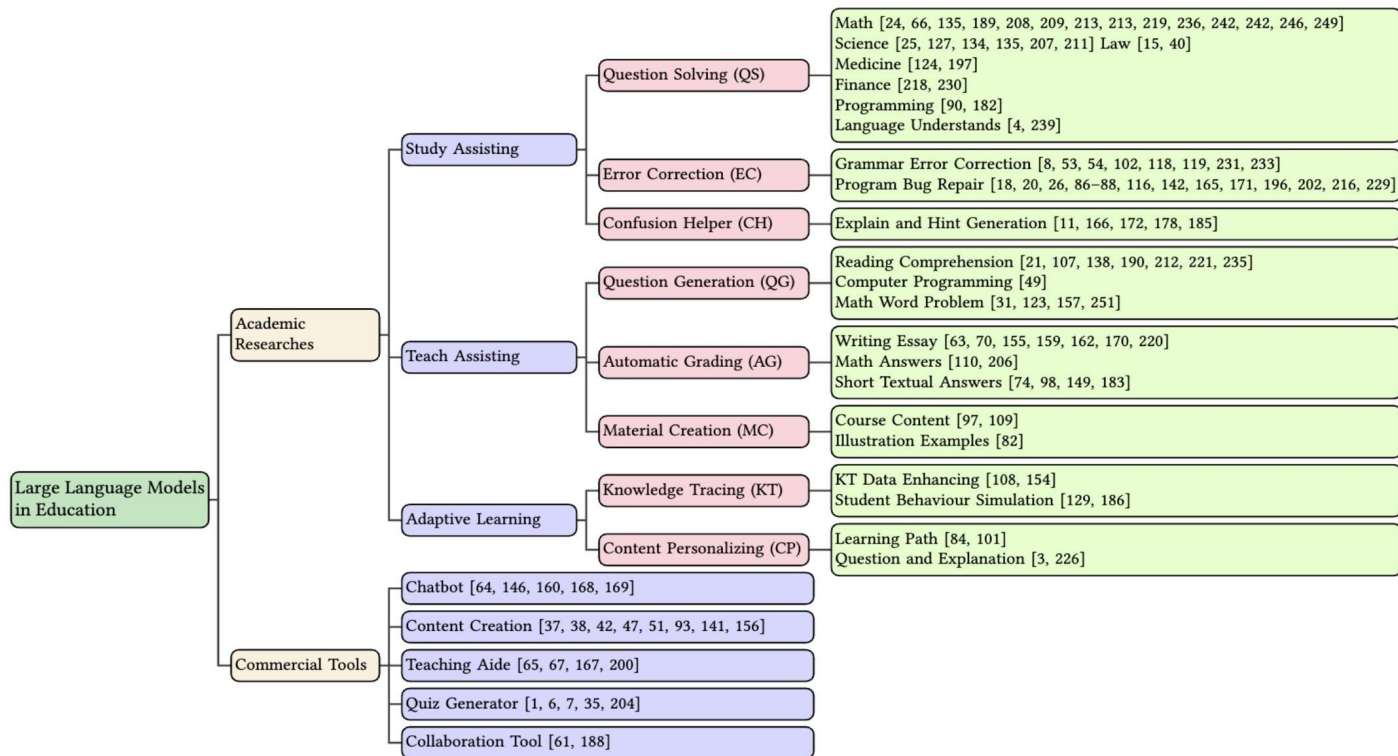
What do you think are the key aspects of
a good AI-based educational tool?



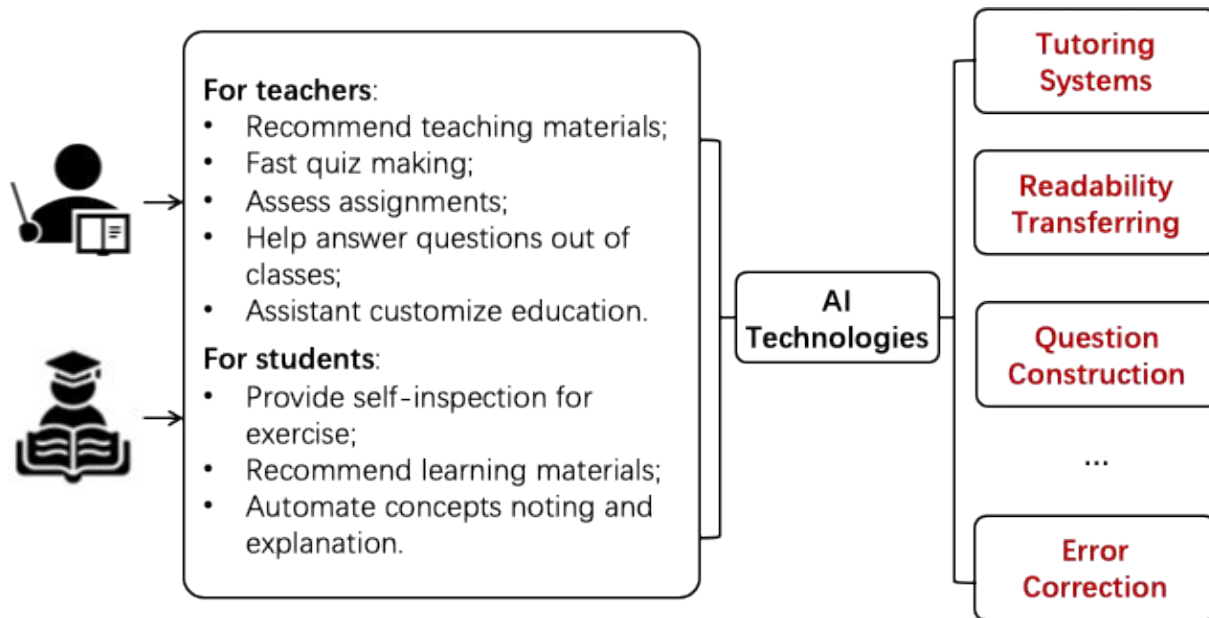
Brief History



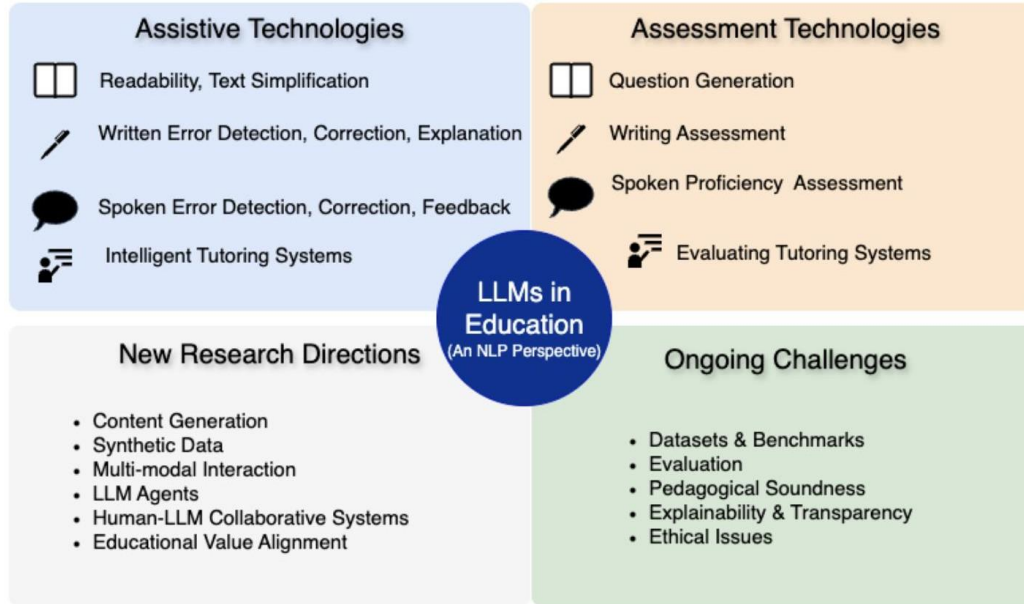
LLMs in Education Taxonomy



Teacher-side vs. Student-side

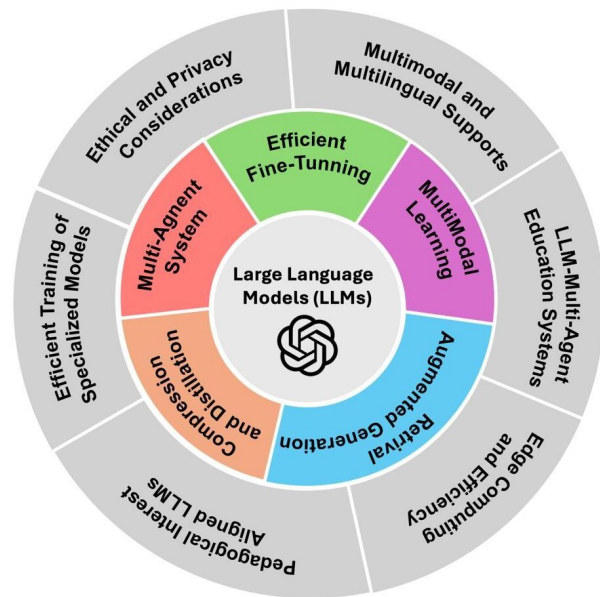


Assistive vs. Assessment



Current Trends and Emerging Directions

- Multi-Agent Systems for Education
- Human-AI collaboration
- Ethical and privacy considerations



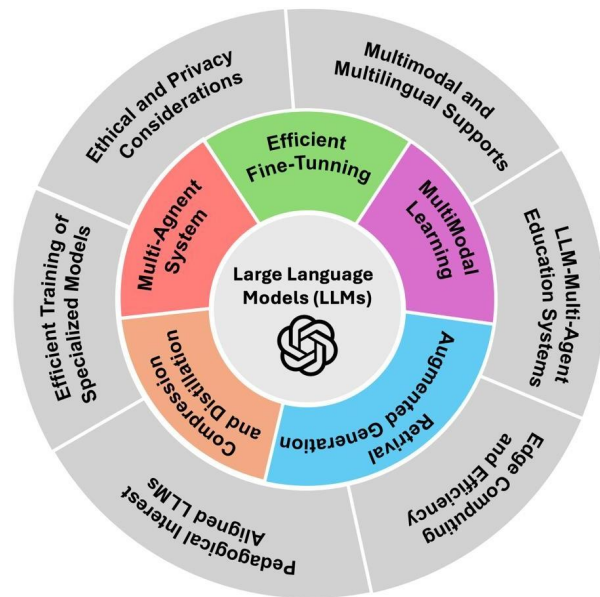
Wang et al. (2024). [Large Language Models for Education: A Survey and Outlook](#)

Vajjala et al. (2025). [Opportunities and Challenges of LLMs in Education: An NLP Perspective](#)

Yuan et al. (2025). [NLP and Generative AI for Language Learning and Assessment](#)

Current Trends and Emerging Directions

- Multi-Agent Systems for Education
- Human-AI collaboration
- Ethical and privacy considerations



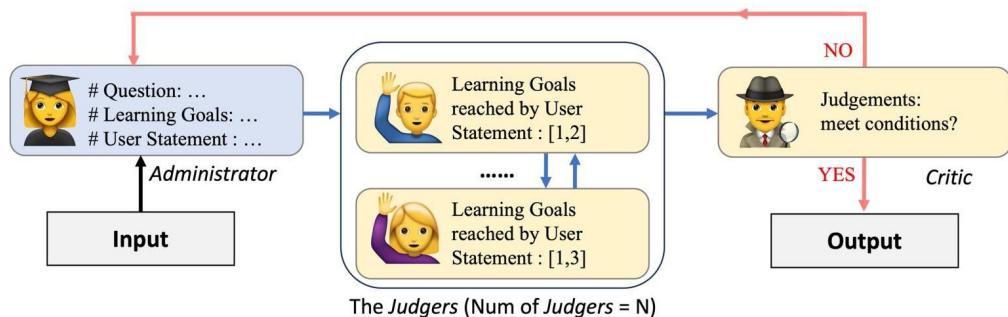
Wang et al. (2024). [Large Language Models for Education: A Survey and Outlook](#)

Vajjala et al. (2025). [Opportunities and Challenges of LLMs in Education: An NLP Perspective](#)

Yuan et al. (2025). [NLP and Generative AI for Language Learning and Assessment](#)

AI Agents

- LLMs + memory, planning, and tools -> enabling systems to pursue goals and solve complex, multi- step problems
- AI agents can reason, plan, and act autonomously, interacting with tools, environments, and users
- Agents can take multiple roles:
 - pedagogical agents (e.g., tutors)
 - Student agents (e.g., collaborators, simulators)
 - multi-agent systems, where AI agents and humans work together as a team

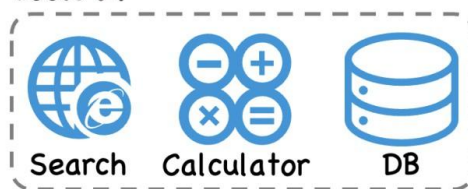


Yang et al. (2024). [Content Knowledge Identification with Multi-Agent Large Language Models \(LLMs\)](#)

Classroom Simulation (CS)

Simulate a student discussion about **Newton's third law** in a high school physics class.

Tools

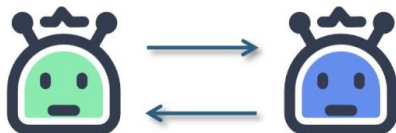


Adaptive Learning (AL)

I'm struggling with **chain rule**.
Give me a few tailored problems to practice with hints?

Feedback Comment Generation (FCG)

Give feedback on this essay, focusing on argument clarity and evidence.



Communication

Memory

Planning

Personalization

Explainability

Knowledge Tracing (KT)

Here are my quiz results on '**Two negatives make a negative.**'
Have I mastered it?

Error Correction and Detection (ECD)

Here's my solution to a **Python coding** problem. Can you tell me what's wrong and how to fix it?"

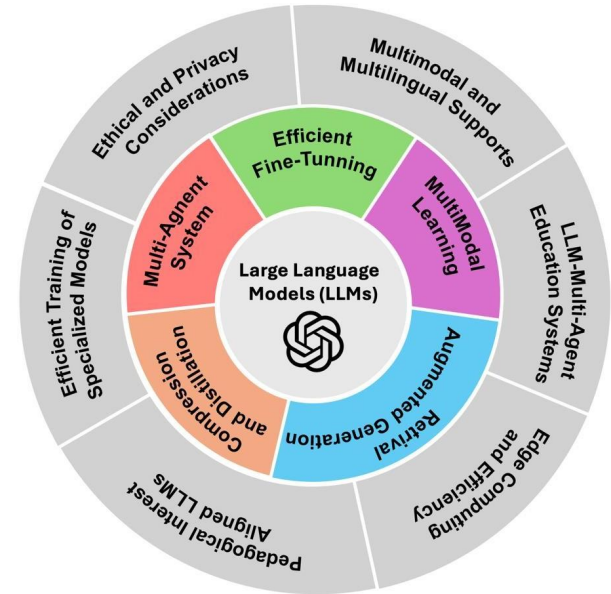
Teachers

(Multi-)Agent Systems

Students

Current Trends and Emerging Directions

- Multi-Agent Systems for Education
- Human-AI collaboration
- Ethical and privacy considerations



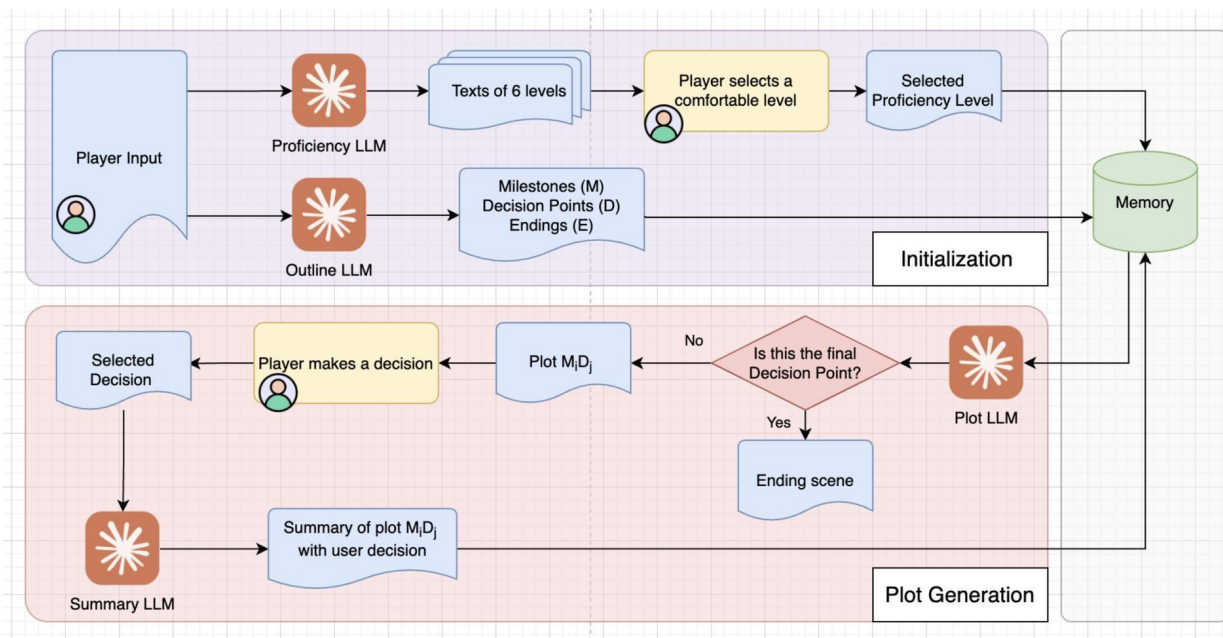
Wang et al. (2024). [Large Language Models for Education: A Survey and Outlook](#)
Vajjala et al. (2025). [Opportunities and Challenges of LLMs in Education: An NLP Perspective](#)
Yuan et al. (2025). [NLP and Generative AI for Language Learning and Assessment](#)

AI Teammates: Human-LLM Collaboration

- Shift from **HCI (human-computer interaction)** to **HAIC (human-AI collaboration)**
- AI as a **collaborator**, not just a tool
- Applications include:
 - [GenQuest](#) (co-creative storytelling)
 - [GPTeach](#) and [Tutor CoPilot](#) (tutoring & teacher training)
- Key challenges:
 - Avoiding over-reliance on AI
 - Preserving human agency and creativity (e.g., for students and teachers)

GenQuest: Human-LLM Co-Creation for Language Learning

- GenQuest is an LLM-powered platform that co-creates adaptive, choice-driven narratives with learners, enabling personalised and immersive language learning
- Learners engage in a **“choose-your-own- adventure”** experience
- **Adaptive support:** content tailored to proficiency levels + in-context vocabulary explanations
- Demonstrates **AI as a teammate** in co-creating learning experiences
- Pilot study shows promising **vocabulary gains** and **positive user perceptions**



Tutor CoPilot: AI as a Teaching Teammate

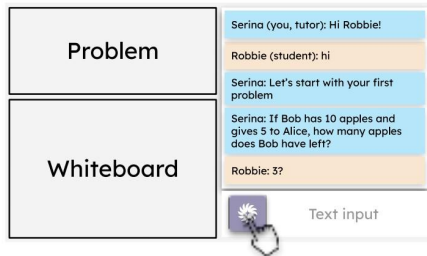


Tutor CoPilot

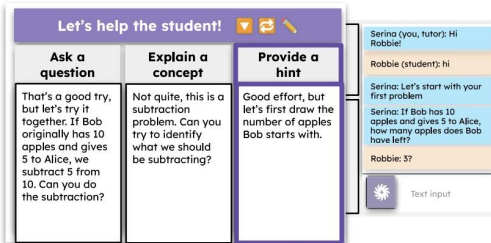
Front End

Back End

(a) Integrating into Real-Time Context

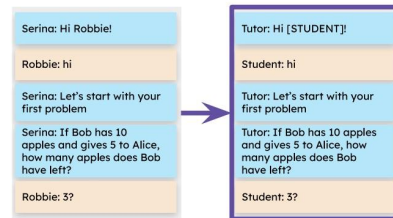


(d) Enabling User Customization



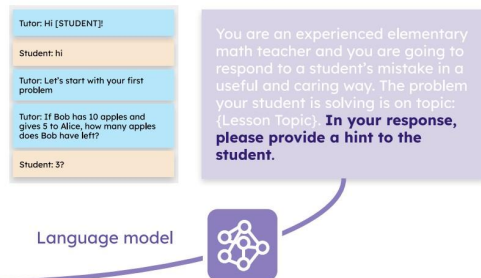
(b) Ensuring User Safety and Privacy

Such as de-identification of tutor and student names



(c) Generating Expert-Like Guidance

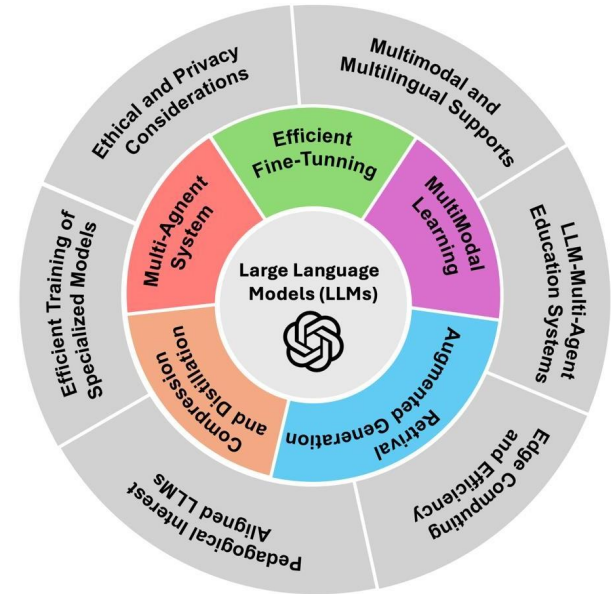
a.k.a. Bridge (Wang et al., 2024)



- **Human-AI collaboration** to support novice tutors with real-time, expert-like guidance during live tutoring
- Leverages **a model of expert thinking** to generate actionable suggestions tutors can adapt in context
- Randomised controlled trial in live tutoring: ~800 math tutors and ~1,800 K-12 students from historically under-served communities
- Results suggest that students of tutors randomly assigned to Tutor CoPilot are 4% points more likely to master taught topics ($p < 0.01$)

Current Trends and Emerging Directions

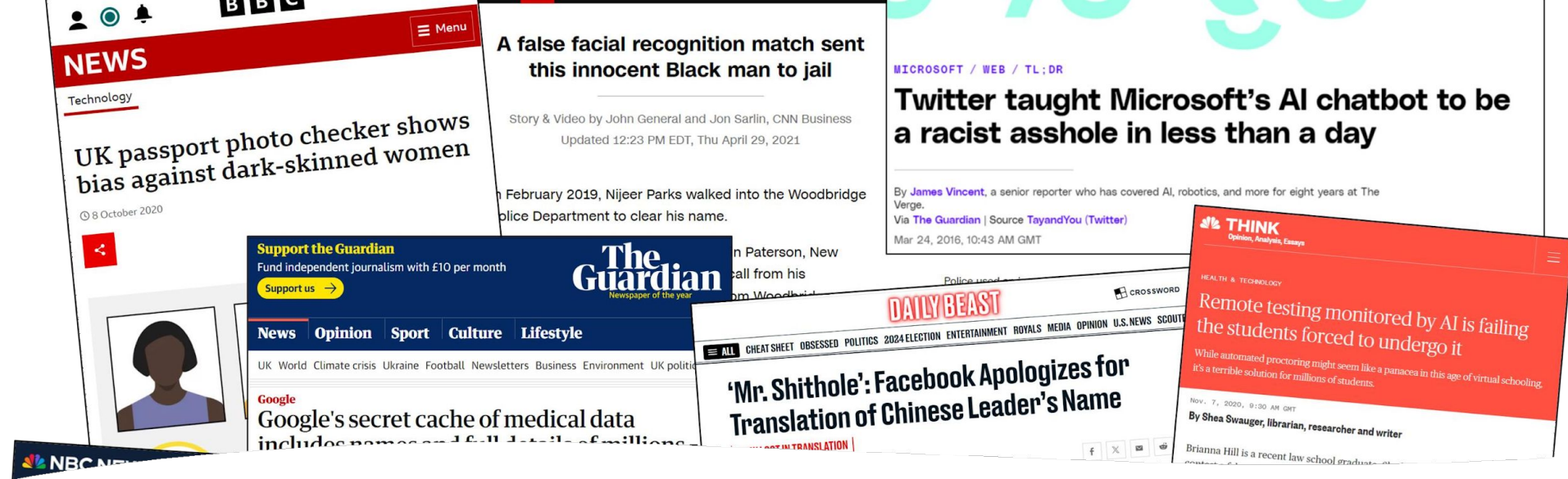
- Multi-Agent Systems for Education
- Human-AI collaboration
- **Ethical and privacy considerations**



Wang et al. (2024). [Large Language Models for Education: A Survey and Outlook](#)

Vajjala et al. (2025). [Opportunities and Challenges of LLMs in Education: An NLP Perspective](#)

Yuan et al. (2025). [NLP and Generative AI for Language Learning and Assessment](#)



The importance of Ethics in AI

- AI systems can introduce bias, lack transparency, and raise privacy issues
- These concerns affect fairness and trust in educational outcomes

Legislation

- **EU AI Act**¹: Classifies educational AI as high-risk, requiring transparency, safety and ongoing oversight
- **UNESCO Guidelines**²: Promote human rights, inclusiveness and ethical AI use worldwide
- **OECD Recommendations**³: Stress trustworthy, accountable and robust AI across members
- **EU Guidelines for Educators**⁴: Offer practical advice for ethical AI integration in classrooms

¹European Parliament and Council (July 2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts. Official Journal of the European Union, L 2024/1689, 12.7.2024

²UNESCO (Nov. 2021). Recommendation on the Ethics of Artificial Intelligence.

³OECD Principles on Artificial Intelligence (2019). Tech. rep. OECD

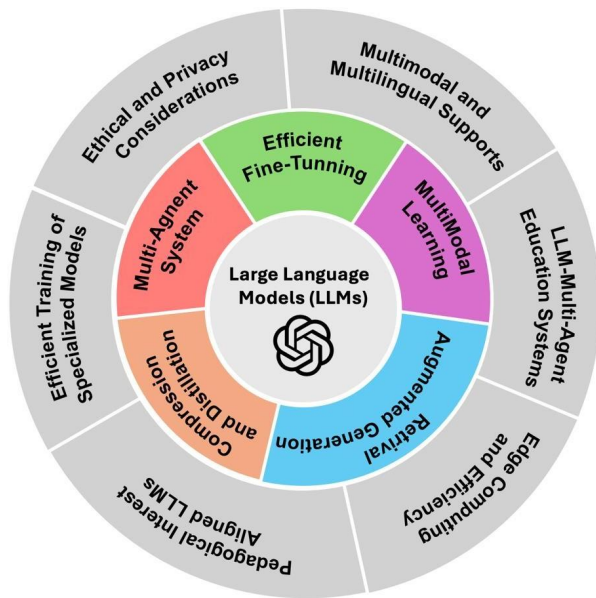
⁴European Commission (2022). Ethics Guidelines for Trustworthy AI in Education. Tech. rep. Practical advice for ethical AI use in classrooms. European Commission

Recommendations

1. Follow established ethical principles.
2. Treat ethics assessment as an ongoing process.
3. Stay focused on purpose and pedagogy.
4. Build AI literacy to empower stakeholders and enable critical evaluation.
5. Prioritise humans over technology.
6. Ensure solutions are inclusive, accessible, and diverse.
7. Foster collaboration across stakeholders.

Current Trends and Emerging Directions

- Interpretable educational NLP
- Adaptive NLP techniques for education
- Multilingual and cross-cultural educational NLP
- Multimodal interaction
- Educational AI for collective intelligence
- Integration with Human-centred pedagogy
- Synthetic data generation for fine-tuning
- Edge computing and efficiency
- Efficient training of specialised models
- ...



Wang et al. (2024). [Large Language Models for Education: A Survey and Outlook](#)

Vajjala et al. (2025). [Opportunities and Challenges of LLMs in Education: An NLP Perspective](#)

Yuan et al. (2025). [NLP and Generative AI for Language Learning and Assessment](#)

Thank you!

Questions?

<https://www.cl.cam.ac.uk/~zy249>
zheng.yuan1@sheffield.ac.uk

Predicting What Learners Don't Know

Winning BEA 2026 With an Agent Swarm

Vassili Philippov

Glite Tech



BEA 2026 shared task

Vocabulary Difficulty Prediction for English Learners

Motivation

Vocabulary is a crucial aspect of language knowledge, shaping what learners can understand and produce. Establishing the difficulty of vocabulary is therefore essential for creating level-appropriate content and developing valid, reliable assessment instruments. However, determining word difficulty still relies on labour-intensive processes involving expert judgment and costly pretesting, which limits scalability and slows innovation. As learning and assessment increasingly rely on digital platforms, the need for more efficient and scalable solutions is more pressing than ever.

While previous shared tasks have explored related problems such as Complex Word

Data Science and Insights

Resources

BEA 2026 shared task

lucyskidmore Update results_summary_test.md 71010c7 · 2 weeks ago History

Closed Track: Spanish

#1

Team Name	Prediction File	RMSE	Pearson
Glite	predictions_run_3.csv	0.903	0.877
Glite	predictions_run_2.csv	0.910	0.876
Glite	predictions_run_1.csv	0.920	0.872
uogal	8enc_optimized_weights_scipy_nelder_mead_fitted_on_677_dev_preds.csv	0.975	0.858
AIDA	predictions_run_1.csv	0.976	0.857
uogal	8enc_dmeta_elasticnet_alpha0.1_l1ratio0.5_trained_on_6091_train_preds_preds.csv	0.977	0.859
...	...	...	...

3 Things You Will Learn



BEA 2026 shared task

Vocabulary Difficulty Prediction for English Learners

Motivation

Vocabulary is a crucial aspect of language knowledge, shaping what learners can understand and produce. Establishing the difficulty of vocabulary is therefore essential for creating level-appropriate content and developing valid, reliable assessment instruments. However, determining word difficulty still relies on labour-intensive processes involving expert judgment and costly pretesting, which limits scalability and slows innovation. As learning and assessment increasingly rely on digital platforms, the need for more efficient and scalable solutions is more pressing than ever.

While previous shared tasks have explored related problems such as Complex Word

Data Science and Insights

Resources

BEA 2026 shared task

spanish

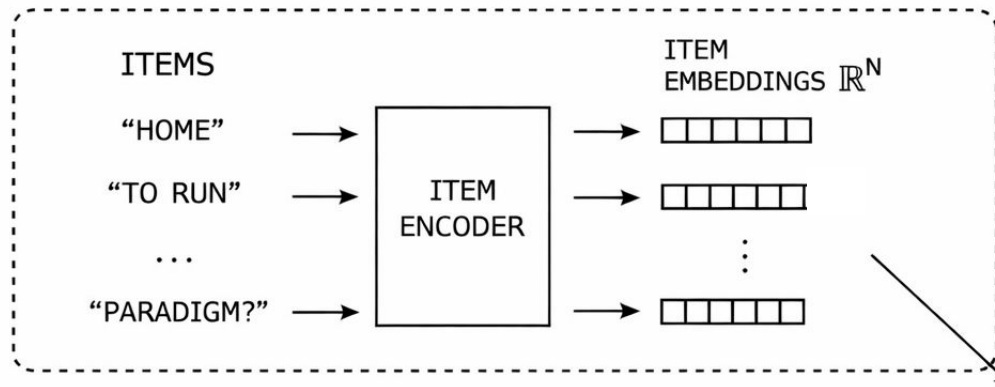
casa

Me quedé en casa para cuidar del gato.

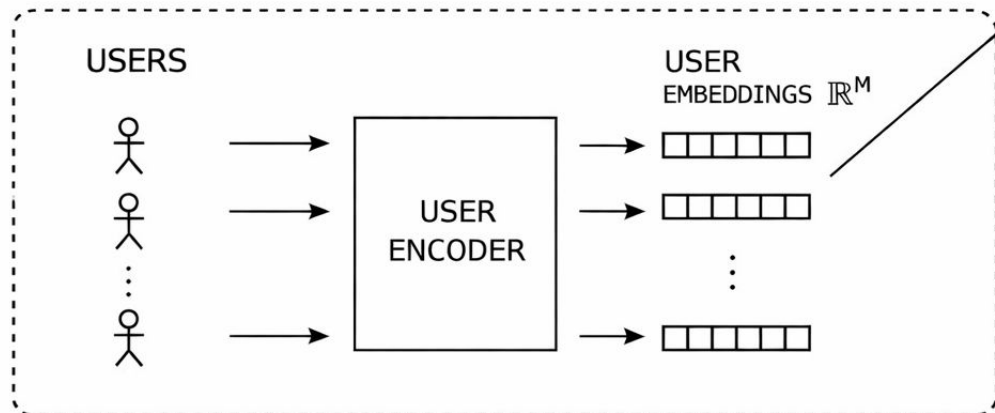
h _ _ _

Difficulty prediction is used almost everywhere in education

ITEM TOWER



USER TOWER



INTERACTION
FUNCTION
(NETWORK)

PREDICTION
[0...1]

1. Reasonable Features Worked

- Frequency
- Similarity to native language
- Number of senses
- etc

2. Rubrics - Universal Trick To Improve Predictions

1. Use your best model to predict
2. Analyze 200-1000 items where your models over-predicts and under-predicts
3. Try to find categories/rubrics that classify those items
4. Write LLM prompt to classify your items according to those rubrics
5. Add this new LLM classification as a feature
6. Your system/ensemble will most likely be better

3. LLM Is a Bad Virtual Student

Asking LLM to pretend to ask as Spanish student of B2 level learning English didn't work.

Currently LLMs out of the box cannot “pretend to be”.

4. Fine-Tuned LLMs Dominated

Those who fine-tuned encoders were behind

Open Track: Spanish

#2

Team name	Prediction File	RMSE	Pearson
Sakura	predictions_finetuned_llms_plus.csv	0.742	0.919
Sakura	predictions_open_max.csv	0.743	0.920
Glite	predictions_run_2.csv	0.754	0.916
Glite	predictions_run_1.csv	0.755	0.916
Sakura	predictions_finetuned_llms.csv	0.760	0.915
Glite	predictions_run_3.csv	0.837	0.896
TeamXBC	TeamXBC_predictions_run_3.csv	0.876	0.885
uogal	8enc_optimized_weights_scipy_nelder_mead_fitted_on_677_dev_preds.csv	0.975	0.858
uogal	8enc_dmeta_elasticnet_alpha0.1_l1ratio0.5_trained_on_6091_train_preds_preds.csv	0.977	0.859
TeamXBC	TeamXBC_predictions_run_2.csv	1.003	0.846
SurreyCTS	predictions_run_3.csv	1.034	0.839
SurreyCTS	predictions_run_1.csv	1.046	0.833

5. How To Use LLMs To Predict Numbers

Input:
"Predict ...
as a number
from 1 to 5"

LLM
to
fine-tune

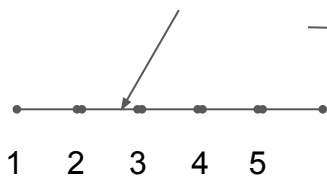
token: prob

...
1: p_1
2: p_2
3: p_3
4: p_4
5: p_5
...

Result =
 $1 \cdot p_1 + 2 \cdot p_2 + \dots + 5 \cdot p_5$

Fine-tuning:

True label = 2.7



LLM

Desired labels:

...
1: 0
2: 0.3
3: 0.7
4: 0
5: 0
...

Loss = MSE

>250 research sub-projects

BEA 2026 — Individual Features Catalog

Search: Value: Closed: LLM: L1-specific: Greedy (any): Secondary:

Showing 1161 of 1161 features

Feature	Feature Set	Value	Pearson	RMSE	L1	Closed	Open	LLM	Rank	2nd	ES	DE	CN	Ens	Description
abs_disagree_clue_desc	model_disagreement	medium	0.418	1.6129	Yes	✗	✓	✗	✗	✗	✗	✗	✗	✓	Absolute magnitude of disagreement between LLM clue-guessing and description-guessing success.
abs_disagree_frequency_irt	model_disagreement	medium	0.230	1.7254	Yes	✗	✓	✗	✗	✗	✗	✗	✗	✓	Absolute magnitude of disagreement between word frequency and IRT difficulty parameter.
abs_disagree_psychometric_frequency	model_disagreement	medium	0.120	1.7602	Yes	✗	✓	✗	✗	✗	✗	✗	✗	✓	Absolute magnitude of disagreement between psychometric difficulty and word frequency. High values indicate strong mismatch regardless of direction.
abs_disagree_psychometric_llm_clue	model_disagreement	medium	0.501	1.5379	Yes	✗	✓	✗	✗	✗	✗	✗	✗	✓	Absolute magnitude of disagreement between psychometric difficulty and LLM clue-guessing success.
al_all_succeed	ability_ladder	low	0.292	1.6959	Yes	✗	✓	✓	✗	✗	✗	✗	✗	✗	Binary flag: 1 if all 4 models have at least one correct guess (success rate > 0).
al_correct_count	ability_ladder	low	0.301	1.6908	Yes	✗	✓	✓	✗	✗	✗	✗	✗	✗	Count of models (0-4) that achieved at least one correct guess, providing an ordinal difficulty measure.
al_curve_concavity	ability_ladder	low	0.335	1.6707	Yes	✗	✓	✓	✗	✗	✗	✗	✗	✗	Mean second derivative of the success curve across model indices, measuring whether the difficulty transition is sudden or gradual.
al_difficulty_threshold	ability_ladder	low	0.328	1.6753	Yes	✗	✓	✓	✗	✗	✗	✗	✗	✗	Index of the smallest model (0-based) that achieves >50% success rate, or 4 if no model does.
al_discrimination_slope	ability_ladder	medium	0.354	1.6580	Yes	✗	✓	✓	✗	✗	✗	✗	✗	✗	Linear regression slope of success rates across model indices, capturing how sharply difficulty transitions across ability levels.

Auto-Research Framework

V3 was used in this project.

Who has used something like that?

Auto-Research Problems

1. Mess
2. Long texts
3. No traceability
4. Many things will go wrong

Glite Auto-Research Framework Principles

1. GitHub
2. Tasks isolation
3. Immutable
4. Enforced by scripts
5. Do duplication
6. Auto logging
7. Charts
8. Daily news
9. Dashboard

<https://github.com/GliteTech/glite-arf>

```
➤ Agent [Research internet for 10127]
    ➤ Bash cd /Users/vassilphilipov/research-wd-worksites/10127_escher_char_budget_context_expansion && wu run python -u => arf_scripts.aggregators.aggregate_papers(-f_)
        Running...
    ➤ Bash cd /Users/vassilphilipov/research-wd-worksites/10127_escher_char_budget_context_expansion && wu run python -u => arf_scripts.aggregators.aggregate_papers(-f_)
        Running...
    ➤ Bash ls /Users/vassilphilipov/research-wd-worksites/10127_escher_char_budget_context_expansion/meta/categories/
        Running...
        +48 more tool uses (ctrl+e to expand)
        (ctrl+b to run in background)

➤ Executing 10127 ESCHER char-budget context expansion. (25m 49s → 17.2k tokens)
    ➤ Execute 10127 escher-character-budget context expansion

#
```

[illegible]

```

+ workflow: create failed before the branch was created. The failure is local
  infrastructure: pre-commit tried to run during the automatic Start task ... commit,
  and that collided with an existing .git/index.lock in the main repo. I'm checking
  whether the lock is stale or owned by another active task before I touch it.

+ Explored
  Read task.json
  List index.lock

+ Ran ps -o pid,ppid,etime,command --x1 rg "/.git/" |5|pre-commit|worktree create|git
  7360K 12812 00:00 /bin/bash -c: /usr/bin/ps --no-tty --format '%p %p %e %c\n' |5|
  13 /usr/bin/ps --no-tty --format '%p %p %e %c\n' |5| /usr/bin/ps --no-tty --format '%p %p %e %c\n'
  13 /usr/bin/ps --no-tty --format '%p %p %e %c\n'
  830K 8300 00:00 rg .git/ |5|pre-commit|worktree create|git|git

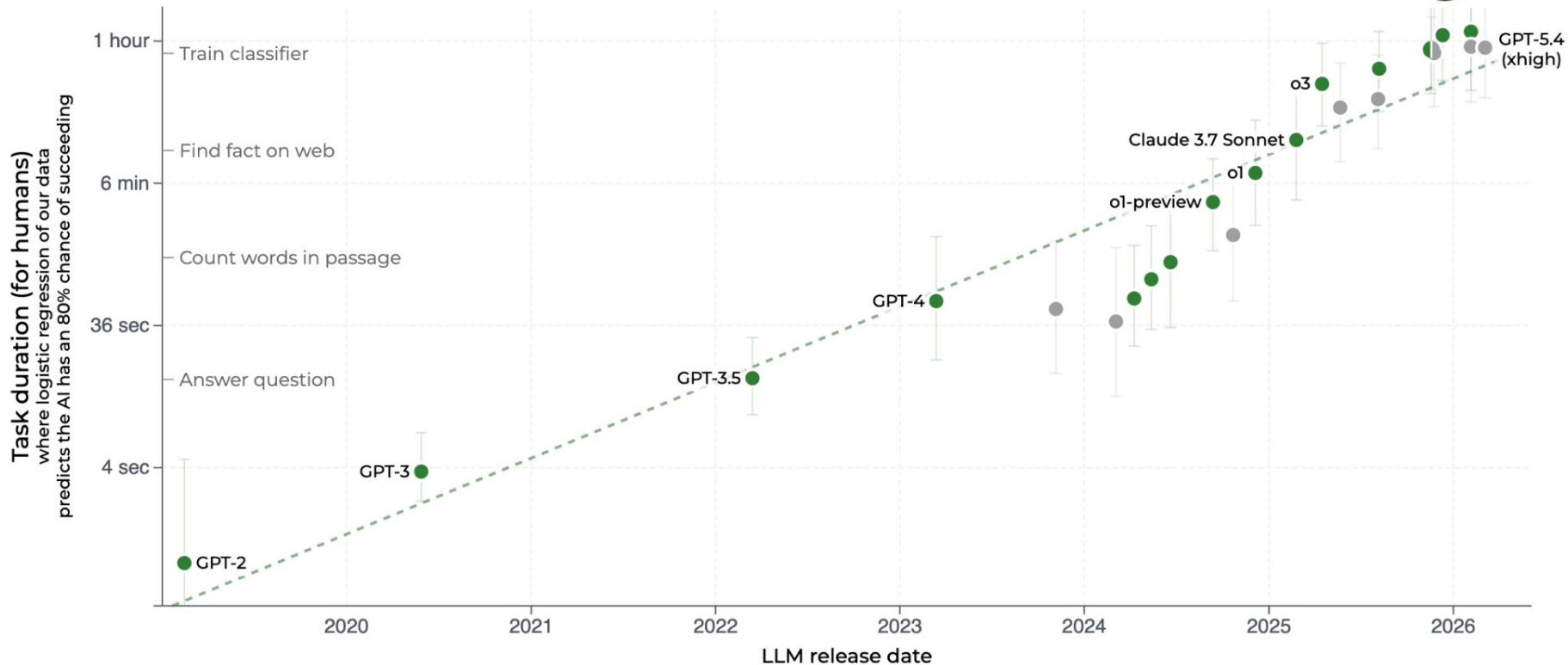
+ Working (in 22s + not to interrupt)

+ git -5.4.4 -single-fast -800 -left -o/research-wsl

```

You become a PI with your own Lab

Time horizon of software tasks different LLMs can complete 80% of the time



Time Horizon 1.1 (Current) ▾

Log Scale

Linear Scale

50% Success

80% Success

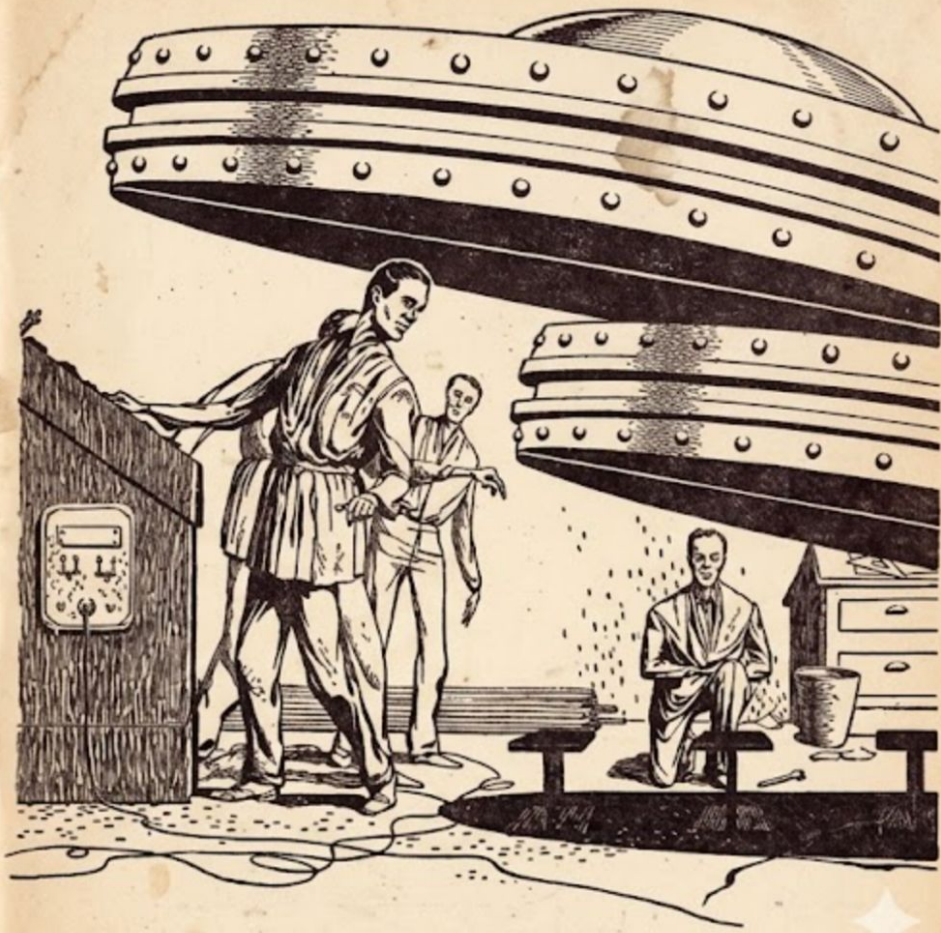
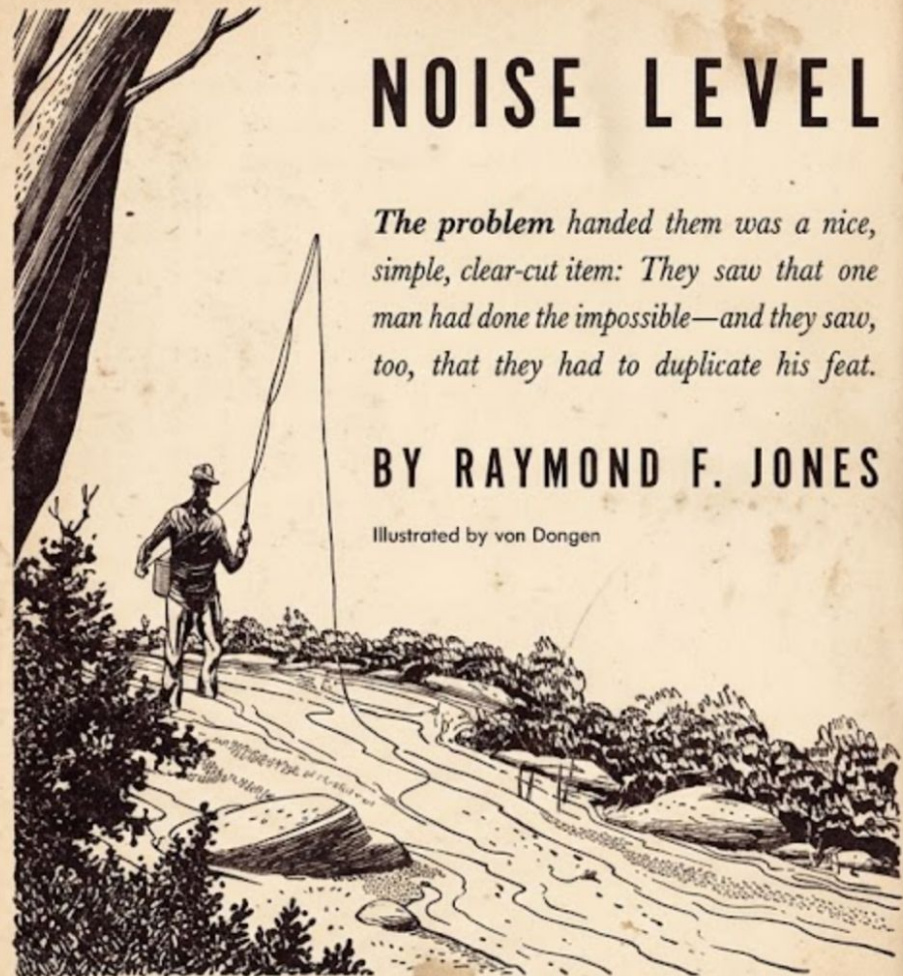


NOISE LEVEL

The problem handed them was a nice, simple, clear-cut item: They saw that one man had done the impossible—and they saw, too, that they had to duplicate his feat.

BY RAYMOND F. JONES

Illustrated by von Dongen



Project Dashboard



Categories: [bi-encoder](#) | [cross-encoder](#) | [dataset](#) | [evaluation](#) | [idiomaticity](#) | [knowledge-base](#) | [library](#) | [llm-wsd](#) | [sequence-labeling](#) | [span-identification](#) | [survey](#) | [wsd](#)

LLM Contexts: [overview](#) (40K) | [full](#) (520K) | [roadmap](#) (121K) | [results](#) (412K) | [assets](#) (83K)

Last updated: 2026-04-14 01:37 UTC

- **Budget:** \$1845 spent of \$2000
- **Remaining:** \$155
- **Usage:** 92.3%
- **GPU Machines:** 48 provisioned across 30 tasks · \$808 GPU spend ([details](#))
- **Provisioning:** 15 failed attempts · \$8.99 wasted · 24% failure rate

Daily News (16)

- [Apr 13, 2026](#) — 6 completed · ~\$148
- [Apr 12, 2026](#) — 8 completed · 11 created · ~\$8

Glite

Thank-you!

We hope to see you at our next meetup later this year