LearnLens: LLM-Enabled Personalised, Curriculum-Grounded Feedback with Educators in the Loop

Runcong Zhao[1], **Artem Bobrov**[1], **Jiazheng Li**[1], **Cesare Aloisi**[2], **Yulan He**[1]
[1]King's College London, [2]AQA
{runcong.zhao, artem.bobrov, jiazheng.li}@kcl.ac.uk
caloisi@aqa.org.uk, yulan.he@kcl.ac.uk

Abstract

Effective feedback is essential for student learning but is time-intensive for teachers. We present **LearnLens**, a modular, LLM-based system that generates personalised, curriculum-aligned feedback in science education. LearnLens comprises three components: (1) an *error-aware assessment* module that captures nuanced reasoning errors; (2) a *curriculum-grounded generation* module that uses a structured, topic-linked memory chain rather than traditional similarity-based retrieval, improving relevance and reducing noise; and (3) an *educator-in-the-loop* interface for customisation and oversight. LearnLens addresses key challenges in existing systems, offering scalable, high-quality feedback that empowers both teachers and students. A screencast video introducing the system[1] and the demo[2] are available online.

1 Introduction

Timely, high-quality feedback is a key driver of learning across all educational levels (Hattie and Timperley, 2007; Boud and Dawson, 2023). Effective feedback should be personalised, actionable, and dialogic: supporting student engagement and offering clear guidance for improvement (Carless, 2016). However, delivering such feedback at scale remains a major challenge, as it is time-intensive and adds to teachers' already substantial workload and growing levels of stress and fatigue (Jomuad et al., 2021). Recent advances in large language models (LLMs) offer a promising opportunity to automate feedback generation and ease teacher workload. Prior work (Mazzullo and Bulut, 2025; Liu et al., 2025) has explored their use for automatically generating scores and rationales for student answers, as well as for delivering evidence-based feedback in classroom settings. However, existing systems still fall short in several ways: (i) they focus narrowly on scoring, overlooking the learner's partial understanding and reasoning process (Mayfield and Black, 2020; Xie et al., 2022); (ii) they generate feedback directly from LLMs without sufficient curricular grounding, often leading to hallucinated, misaligned, or overly generic suggestions (Mazzullo and Bulut, 2025; Meyer et al., 2024); and (iii) they offer limited avenues for educator control, correction, or customisation during the feedback process (Swamy et al., 2025).

To address these limitations, we present **LearnLens**, a modular LLM-based system for personalised, curriculum-aligned feedback in science education. LearnLens consists of three core components, each directly tackling a corresponding shortcoming. First, an *error-aware assessment* module moves beyond binary correctness by aligning learner responses with structured mark schemes and identifying conceptual, factual, or linguistic errors, enabling more nuanced understanding of student reasoning. Second, a *curriculum-grounded generation* module produces feedback using a restructured, topic-linked memory chain aligned with the national curriculum. Unlike traditional similarity-based retrieval, which often introduces irrelevant or noisy content, our *Chain-of-Concept* framework integrates carefully curated knowledge and curriculum alignment into the retrieval process, ensuring that generated feedback is pedagogically coherent and tailored to specific learning objectives. Third, an *educator-in-the-loop* interface, paired with a separate student-facing interface, enables teachers and students to interact with the system in complementary ways. Teachers can monitor student performance, create customised questions, review and revise feedback using natural language, and optionally select from a suite of embedded verifiers that assess different aspects of feedback quality—such as factual accuracy, curricular relevance, and linguistic clarity.

[1] https://www.youtube.com/watch?v=mCUALVwDKNQ
[2] https://learnlens.co.uk/login

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 625–633
November 4-9, 2025

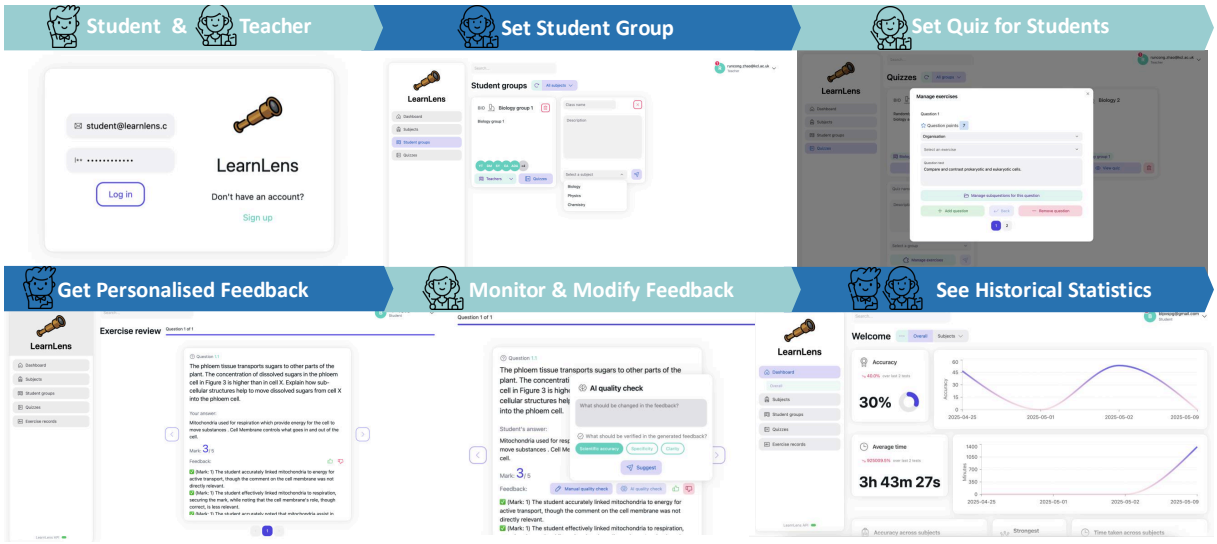

Figure 1: **LearnLens**: A modular LLM-based system delivering personalised, curriculum-aligned feedback through error analysis, topic memory, and educator oversight.

2 Architecture of LearnLens

We design **LearnLens** as a dual-interface intelligent feedback system that serves both teachers and students. Our core goal is twofold: (1) to reduce the workload for teachers while enhancing their ability to monitor student progress, and (2) to provide students with timely, personalised feedback that supports deeper learning. Powered by LLMs, **LearnLens** integrates structured knowledge management and human-in-the-loop collaboration into a scalable educational tool. Through dedicated teacher and student interfaces, the system delivers curriculum-aligned, high-quality support tailored to real classroom needs.

2.1 Teacher Interface: Less Grading, More Guiding

To reduce teacher workload while enhancing instructional oversight, **LearnLens** provides tools that support differentiated instruction, personalised assessment, and high-quality feedback. As shown in Figure 1, the teacher interface facilitates efficient classroom management through student grouping, quiz creation, and curriculum-aligned rubric generation. It also offers up-to-date analytic insights, including performance visualisation and verifier scores, to help teachers identify students in need of support and feedback requiring revision. This integrated, human-in-the-loop workflow enables teachers to focus less on repetitive grading and more on delivering targeted, meaningful guidance.

Student Grouping As illustrated by the *"Set Student Group"* module, teachers can organise students into groups based on subject, year level, and learning progress. Each group can then be assigned one or more teachers and a cohort of students, with appropriate permissions configured to control access to shared materials and feedback within that group. This structure enables differentiated instruction, collaborative planning, and role-based access control in a scalable and intuitive way.

Quiz Creation As illustrated by the *"Set Quiz for Students"* module, teachers can create quizzes and assign them to specific student groups. Each question is first associated with a topic from the national curriculum, ensuring alignment with prescribed learning objectives. For each selected topic, the system provides a curated bank of pre-authored questions that teachers can readily select. Alternatively, teachers may compose their own custom questions to address specific instructional goals. To further enhance flexibility and reduce authoring effort, teachers can also prompt the LLM to automatically generate questions based on the chosen topic and specified pedagogical requirements.

Curriculum-Aligned Rubrics To ensure consistency in model assessment, the system first generates a mark scheme for each question. Since the mark scheme impacts all subsequent assessments, we consider it a critical step where accuracy is prioritised over time efficiency. In line with inference-time scaling laws (Snell et al., 2024; Brown et al., 2024), the model is granted extended reasoning

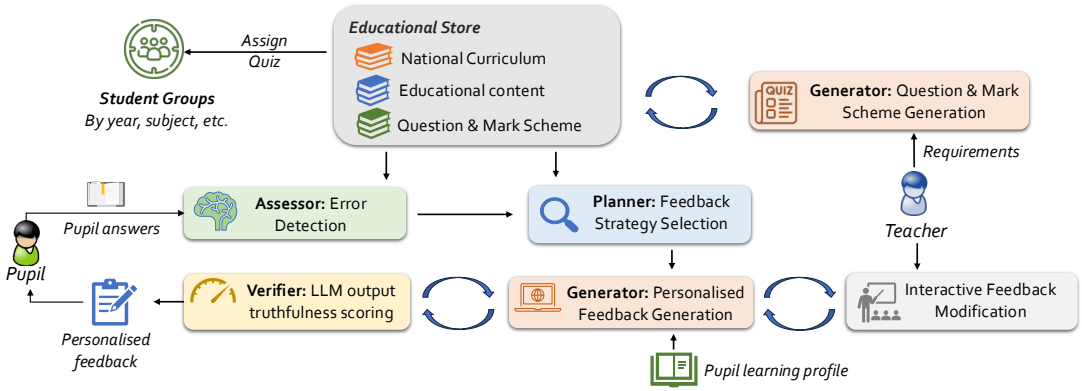

Figure 2: LearnLens: Overall framework.

time to explore diverse solution paths and consider multiple correct approaches. To further ensure pedagogical validity, the system retrieves relevant curriculum content for the target topic. This enables the model to distinguish between knowledge that students are expected to have mastered and content beyond their current level, thereby constraining the mark scheme within the appropriate learning scope.

Student Performance Visualisation To support effective teacher intervention, **LearnLens** provides a historical performance visualisation module that highlights students who may require additional support. Prior studies (Özer Özkan and OZKAN, 2025) and our interviews with teachers suggest that educators often struggle to monitor students' mastery levels in a timely manner. This module addresses that challenge by offering a clear, data-driven overview of student progress, enabling teachers to more efficiently identify learning gaps.

Verifier for Quality Assessment To track the quality of model-generated feedback and assist teachers in identifying responses that may require attention, each feedback instance is first self-evaluated against a set of educational criteria: *Scientific Accuracy* (whether misconceptions are correctly identified and explanations are valid), *Clarity* (whether the language is accessible to GCSE-level students[3]), and *Specificity* (whether the feedback clearly highlights what is correct, incorrect, or missing). For each criterion, the model assigns a self-verification score along with a justification.

As shown in Figure 2, **LearnLens** integrates a verification-and-revision loop into its assessment pipeline. To mitigate model-family bias, verification is performed by a separate foundation model from the generator. Feedback is iteratively refined based on verifier scores: if a response falls short, the model identifies weaknesses and revises it. The process continues until all criteria meet a threshold or the iteration limit is reached:

$$\text{stop} = \left[\min_i r_i \geq \tau\right] \vee \left[t \geq T_{\max}\right],$$

where r_i is the verifier score for the i-th criterion, τ is the acceptance threshold, and t is the current iteration. The highest-scoring feedback is returned. Examples with verification scores are shown in Appendix A.

Interactive Feedback Modification **LearnLens** integrates student performance visualisation and verifier scores to help teachers efficiently identify feedback that may require revision. Teachers can refine flagged feedback through a conversational interface, either by directly editing the content or by providing high-level suggestions. These modifications can be applied to a single question or propagated across the entire quiz. The AI agent then incorporates the teacher's input and generates an updated response accordingly. An illustrative example is provided in Appendix B.

Teacher intervention serves as a signal of dissatisfaction with the initial output. In such cases, the system dynamically allocates additional computational resources, allowing the model to reflect more deeply and produce higher-quality feedback. While faster refinement strategies were explored, they often compromised fidelity to teacher intent. Given this trade-off, we prioritise feedback quality to preserve user trust and learning effectiveness.

2.2 Student Interface: Faster Feedback, Better Understanding

From the student perspective, **LearnLens** delivers timely and tailored feedback that promotes contin-

[3]LearnLens can be tailored to students across all educational levels. In this demo, we focus on GCSE science.

uous learning. To ensure the feedback is both accurate and actionable for learning, the system combines reliable assessment, memory-driven strategy planning, and safe, context-aware generation.

Assessor To deliver *consistent, fine-grained* scoring, the Assessor maps each student answer $\hat{a}_i$ onto the curriculum-aligned mark scheme $\mathcal{C}_i = \{(c_k, w_k)\}_{k=1}^{K_i}$, where c_k is the k-th key concept and $w_k \in \mathbb{Z}_{>0}$ its weight. The raw score is the weighted sum of concept matches:

$$s_i = \sum_{k=1}^{K_i} w_k \, \textsc{Match}(c_k, \hat{a}_i),$$

with MATCH satisfied when the LLM detects that the concept appears in the answer, thus granting partial credit even if the final answer is wrong. For internal consistency we also obtain a *prompt score* s_i^{prompt} via a direct "grade-this-answer" prompt and trigger self-reflection whenever $s_i \neq s_i^{\text{prompt}}$ (Li et al., 2024). LearnLens is also robust to surface errors: grammatical or typographical mistakes do *not* affect the score s_i. Such issues are instead captured by a separate expression-quality flag $\delta_i \in \{0, 1\}$ (1 = major language issues), which is shown in the feedback but excluded from the numerical grade. This design (i) renders the weighting and partial credit scheme *transparent*: students and teachers can directly inspect the marks w_k allotted to each concept c_k and see how many were awarded, which removes the black box impression and eases disputes; and (ii) decouples conceptual understanding from writing quality.

Planner As shown in Figure 3, we organise past assessments by *curriculum topics* rather than raw embedding similarity used in earlier work (Gao et al., 2023). Each question is decomposed into sub-questions (nodes) v_i and labelled with one or more topics $\mathcal{L}(v_i) \subseteq \mathcal{T}$. We model these records as a topic graph

$$G = (V, E, \mathcal{T}), \quad E = \{(v_i, v_j) \mid \mathcal{L}(v_i) \cap \mathcal{L}(v_j) \neq \varnothing\}.$$

For a query q tagged with topics $C_q \subseteq \mathcal{T}$, retrieval is confined to the topic subgraph $G_q = G\big[\{\, v_j \mid \mathcal{L}(v_j) \cap C_q \neq \varnothing\}\big]$, and we return

$$\mathcal{R}(q) = \underset{v_j \in G_q}{\arg\,\text{top-k}}\, f(q, v_j),$$

where f is a FAISS-based ranker (Douze et al., 2024). Topic filtering removes cross-topic noise, yielding more coherent evidence; the module then analyses pupil errors and prior knowledge to choose the most effective feedback strategy.

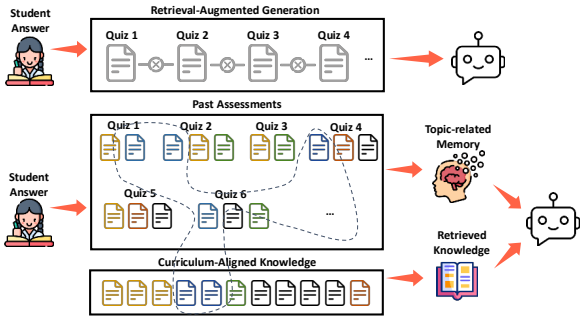

Figure 3: Comparison between traditional RAG and the proposed Chain-of-Concept approach. In RAG (top), past assessment records are stored sequentially without explicit logical structure, forcing the retriever to rely solely on surface-level similarity. This often introduces significant noise, especially as the database grows. In contrast, our approach (bottom) organises past assessments by topic-level relationships, enabling the model to retrieve more contextually relevant information and generate more personalised, coherent feedback.

Generator To generate high-quality feedback, **LearnLens** integrates the quiz content, the corresponding mark scheme, and the selected feedback strategy. We adopt an LLM-based self-reflection mechanism (Asai et al., 2024; Zhang et al., 2024), allowing the model to reason over the student's response. To ensure the safety and appropriateness of the generated content, a safety-aligned model is applied post-generation to filter out any harmful, biased, or otherwise inappropriate language.

3 Evaluation

3.1 Experimental Setup

We conducted a comparative evaluation of LLMs across multiple model families and sizes. Due to student data privacy concerns, all experiments were conducted via local deployment. Our evaluation includes Meta-Llama-3-8B-Instruct, Qwen2.5-32B-Instruct, and QwQ-32B.

3.2 Evaluation Results

3.2.1 User Feedback Results

To evaluate how well **LearnLens** aligns with real-world teaching needs, we collected responses from *30 participants*, all of whom had prior teaching experience in a STEM module at either high school or college level. Each participant interacted with the system and completed a structured questionnaire (Figure 4), which covered topics such as interface usability, system responsiveness, question quality, feedback usefulness, and overall user experience. The complete questionnaire is provided in Appendix C. Our questionnaire results, summarised in Figure 4 (full response distribution) and

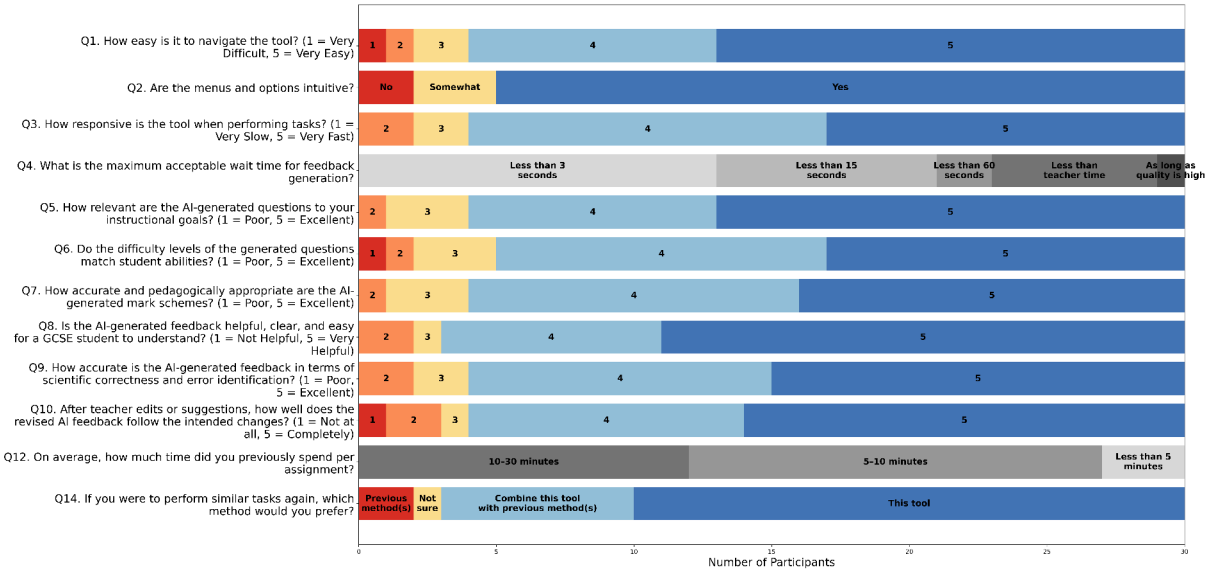

Figure 4: Full Questionnaire Response Distribution (N = 30 per Question). Shades of blue denote more favourable evaluations of the tool, while shades of red indicate dissatisfaction. Grey segments represent neutral factual responses that are not direct indicators of user sentiment toward the tool.

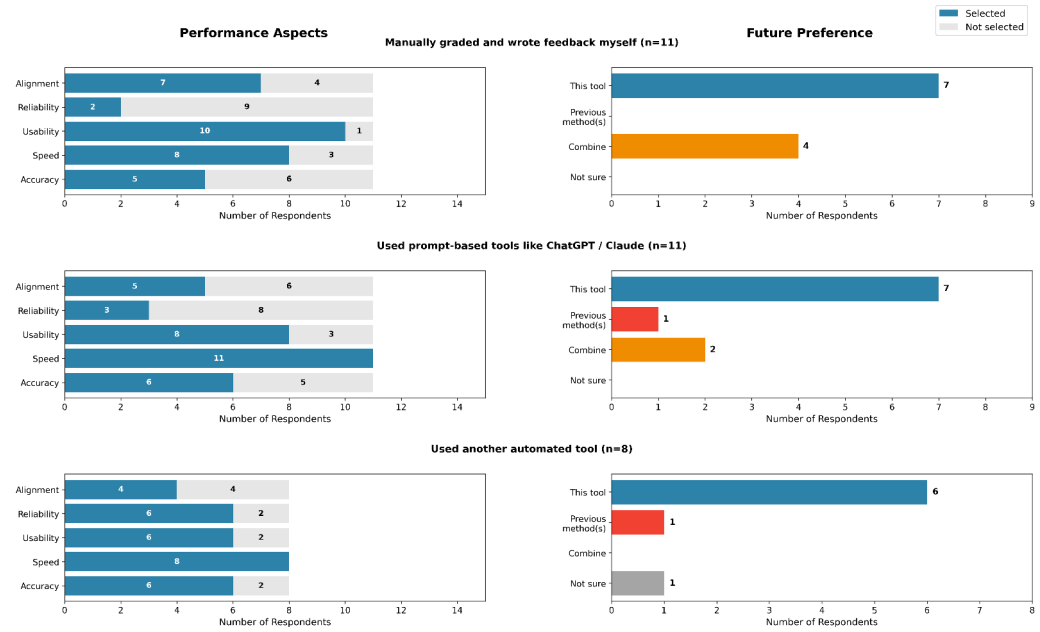

Figure 5: Grouped feedback by participants' prior methods (manual, prompt-based, or automated). Left: perceived improvements across five dimensions. Right: future preferences for task completion. Blue = positive toward our tool; red = preference for previous methods; orange = combine both; grey = neutral or other.

Figure 5 (grouped comparison), highlight the following themes:

Global Satisfaction and Usability Across the nine Likert-scale items (Q1, Q3, Q5–Q10) the mean rating never falls below 4.1/5, signalling a uniformly positive reception. Navigation ease and perceived responsiveness confirm that the dual-interface design achieves its principal UX goal: immediate, low-friction access to feedback tools. Likewise, high scores for relevance to instructional goals (4.4/5) and scientific accuracy (4.3/5) validate our curriculum-grounding strategy.

Categorical items echo this trend. About 80% of teachers describe the menu structure as "intuitive", and 75% expect results in under 15 seconds, a threshold LearnLens already meets in >90 % of cases during local deployment. These data confirm that the system's latency, a common pain-point for classroom AI, is acceptable for real-world use.

	Performance				Generation Latency & Cost			
Model	**MSE**	**Corr.**	**Acc.**	**±1 Acc.**	**Avg. (s)**	**Min. (s)**	**Max. (s)**	**Cost ($)**
LLaMA-3-8B-Ins	3.468	0.235	0.190	0.646	5.14	2.43	9.40	0.0040
Qwen2.5-32B-Ins	3.658	0.230	0.241	0.633	12.39	5.78	20.29	0.0096
QwQ-32B	3.506	0.323	0.329	0.709	45.49	32.27	57.36	0.0351
LearnLens	**3.190**	**0.388**	**0.354**	**0.747**	11.39	7.59	16.97	0.0099

Table 1: **LearnLens** performance and efficiency. MSE: mean squared error; Corr.: Pearson correlation; Acc.: exact match; ±1 Acc.: within–one–mark accuracy. Avg./Min./Max. are generation latencies; Cost is per request.

Efficiency Gains Q12 shows that half of teachers spent 10–30 minutes per assignment, whereas the median with **LearnLens** falls below 5 minutes. Q13 echoes this: 26/30 respondents rank *speed* as the top strength. By off-loading rubric matching and error spotting, **LearnLens** shifts teachers' time from grading to guidance, realigning workload rather than merely automating it.

Experience-Sensitive Priorities Figure 5 segments perceptions by prior practice. Manual graders value speed (91 %) and usability (82 %) most, unsurprising given their baseline workflow. Prompt-tool users distribute credit more evenly, highlighting LearnLens' balanced improvement across accuracy, reliability, and alignment. Their familiarity with LLM quirks likely sharpens expectations. Other-automation users rank accuracy (88 %) and reliability (75 %) highest, suggesting that our verifier-in-the-loop architecture successfully addresses limitations they encountered elsewhere.

A $\tilde{\chi}^2$ test ($\alpha = 0.05$) shows no significant difference in overall adoption intent across groups ($p = 0.63$); every cohort exceeds 65 % "continue using", with the automation-savvy group peaking at 85 %. Thus, while priorities diverge, and willingness to integrate LearnLens is stable.

Implications for Deployment To translate these findings into actionable next steps, we highlight three deployment priorities that will maximise user satisfaction and long-term adoption: (1) Latency ceiling. Meeting the sub-15-second expectation is critical; ongoing optimisation should therefore focus on inference batching and on-device caching. (2) Adaptive onboarding. Manual graders respond to time-saving narratives, whereas technical users are swayed by demonstrable accuracy safeguards; onboarding materials should branch accordingly. (3) Verifier transparency. High appreciation for accuracy among automation veterans highlights the value of surfacing verifier scores (§2.1); exposing these metrics to all users could further bolster trust.

In summary, the study confirms that LearnLens delivers broadly perceived pedagogical value, but supporting diverse existing workflows is crucial for its scalable adoption in classrooms.

3.3 Agent Performance and Efficiency

Table 1 reports results on *100 authentic student answers* to a five-mark short-essay science question. Baselines use the same foundation models as those employed in LearnLens, but perform both feedback generation and verification through direct prompts to the same model, without modular decomposition. For fairness, all methods share the same inputs (question, mark scheme, answer) and required outputs (score, feedback, verification).

LearnLens follows a *modular* pipeline: lightweight models handle routine subtasks, while larger models are invoked only when deeper reasoning is needed. Combined with efficient `vLLM` serving and speculative decoding, this design *achieves the strongest scoring quality without additional overhead*: MSE drops to 3.19 (8–13 % below any baseline), correlation rises to 0.388, and both exact and ±1 accuracies surpass the best baseline (QwQ-32B). At the same time, average generation latency is 11.4 s, comparable to Qwen2.5-32B and four times faster than QwQ-32B, while the per-request cost ($0.0099) matches Qwen2.5-32B and is 72 % cheaper than QwQ-32B.

4 Conclusion

We introduce **LearnLens**, a dual-view feedback system that supports both teachers and students through curriculum-aligned assessment and personalised feedback. By combining LLM capabilities with human-in-the-loop design, LearnLens reduces teacher workload and enhances student learning, offering a practical path toward scalable, classroom-ready AI assistance. While LearnLens shows promise in supporting teachers, we acknowledge the lack of student evaluation and plan to address this in future work.

Acknowledgments

This work was supported by the UK Department for Education through the AI Catalyst Fund - AI Tools for Education (grant no. 10144187) and the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2) and a Prosperity Partnership project with AQA (UKRI566). The authors also acknowledge the use of the King's Computational Research, Engineering, and Technology Environment (CREATE) at King's College London.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

David Boud and Phillip Dawson. 2023. What feedback literate teachers do: an empirically-derived competency framework. *Assessment & Evaluation in Higher Education*, 48(2):158–171.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *Preprint*, arXiv:2407.21787.

David Carless. 2016. Feedback as dialogue. In *Encyclopedia of Educational Theory and Philosophy*, pages 286–289. Springer.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *CoRR*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.

Perlito D Jomuad, Leah Mabelle M Antiquina, Eusmel U Cericos, Joicelyn A Bacus, Juby H Vallejo, Beverly B Dionio, Jame S Bazar, Joel V Cocolan, and Analyn S Clarin. 2021. Teachers' workload in relation to burnout and work performance. *International Journal of Educational Policy Research and Review*, 8(2):48–53.

Jiazheng Li, Hainiu Xu, Zhaoyue Sun, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2024. Calibrating llms with preference optimization on thought trees for generating rationale in science question scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5452–5479, Miami, Florida, USA. Association for Computational Linguistics.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162. Association for Computational Linguistics.

Elisabetta Mazzullo and Okan Bulut. 2025. Automated feedback generation for open-ended questions: Insights from fine-tuned llms. In *Proceedings of Large Foundation Models for Educational Assessment*, volume 264 of *Proceedings of Machine Learning Research*, pages 103–120. PMLR.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Vinitra Swamy, Davide Romano, Bhargav Srinivasa Desikan, Oana-Maria Camburu, and Tanja Käser. 2025. iLLuMinaTE: An LLM-XAI framework leveraging social science explanation theories towards actionable student performance feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28431–28439. AAAI Press.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 6467–6477. International Committee on Computational Linguistics.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3602–3622, Bangkok, Thailand. Association for Computational Linguistics.

Yasemin Özer Özkan and Murat OZKAN. 2025. Decoding teachers' dilemma: Unveiling the real obstacles

to implementing formative assessment in the classroom. *Journal of Qualitative Research in Education*, 41:95–118.

A Example: Personalised Feedback with Self-Verification Scores

To illustrate how **LearnLens** produces feedback with built-in quality assurance, we present an example of AI-generated personalised feedback accompanied by self-verification scores. The system evaluates student responses against a mark scheme, assigns partial credit where appropriate, and generates targeted feedback. Verification scores reflect the feedback's scientific accuracy, clarity, and specificity, while reasoning statements explain how each score was determined.

Sample Feedback with Verification

AI-Generated Feedback:

- **Mark: 0** – Include that cell X contains mitochondria, which are needed to provide energy for active transport.
- **Mark: 1** – Correctly identified active transport moves sugars against the concentration gradient from low (cell X) to high (phloem).
- **Mark: 0** – Mention that mitochondria enable aerobic respiration to release energy specifically needed for active transport.
- **Mark: 1** – Correctly explained energy is needed to move sugars against the concentration gradient via active transport.
- **Mark: 1** – Correctly stated active transport moves sugars from low to high concentration in the phloem.

Total Mark: 3/5

Verification Scores:

- *Scientific Accuracy:* **5**
- *Specificity:* **5**
- *Clarity:* **4**

Verification Reasoning:

- *Scientific Accuracy:* Full marks.
- *Specificity:* Feedback addresses specific content of the student's answer by highlighting what was right, wrong, or missing.
- *Clarity:* Language is generally clear and accessible to GCSE students, though some simplification could reduce potential confusion.

B Example: Interactive Feedback Modification

To illustrate how teacher input is incorporated into the feedback generation pipeline, we present an example of interactive feedback refinement. The original AI-generated feedback is revised based on a set of teacher-provided suggestions that emphasise clarity, brevity, and actionability. The differences introduced by the model in response to teacher suggestions are highlighted in red for clarity and ease of comparison.

Original AI Feedback

- **Mark: 0** – Add explicit mention of mitochondria in cell X. Specify that mitochondria in cell X facilitate aerobic respiration to provide energy needed for active transport.
- **Mark: 1** – Correctly identified active transport moving sugars against the concentration gradient from cell X to phloem.
- **Mark: 0** – Specify that energy released via respiration is required for active transport. Include "to release energy needed for active transport."
- **Mark: 1** – Correctly explained movement against the concentration gradient via active transport. Correct direction noted.
- **Mark: 1** – Correctly stated active transport from low (cell X) to high concentration (phloem).

Teacher Suggestions:

- Do not include general suggestions.
- Give short, specific instructions that students can implement.
- Keep the feedback concise.

Revised Feedback (with Highlighted Edits)

- **Mark: 0** – Specify that cell X contains many mitochondria, which are needed for aerobic respiration to produce energy for active transport. (concise)
- **Mark: 1** – Correctly identified active transport moving sugars against the concentration gradient from cell X to phloem. (unchanged)
- **Mark: 0** – Add that mitochondria carry out respiration to release energy specifically for active transport of sugars.
- **Mark: 1** – Correctly explained movement against the concentration gradient via active transport. Direction (low → high) is accurate.

- **Mark:** **1** – Correctly stated active transport from low (cell X) to high concentration (phloem). (unchanged)

C User Questionnaire

To better understand user needs and evaluate the usability and effectiveness of **LearnLens**, we designed a structured questionnaire for teachers. The survey covered five key dimensions: interface design, system performance, question creation and assessment, personalised feedback quality, and overall user experience. The full questionnaire is detailed below.

Section 1: Interface Design

Q1. How easy is it to navigate the tool? (1 = Very Difficult, 5 = Very Easy)

Q2. Are the menus and options intuitive?
- Yes
- No
- Somewhat

Section 2: Performance

Q3. How responsive is the tool when performing tasks? (1 = Very Slow, 5 = Very Fast)

Q4. What is the maximum acceptable wait time for feedback generation?
- Less than 3 seconds
- Less than 15 seconds
- Less than 60 seconds
- Less than the time a teacher would typically spend providing feedback
- As long as the feedback is high quality

Section 3: Question Creation

Q5. How relevant are the AI-generated questions to your instructional goals? (1 = Poor, 5 = Excellent)

Q6. Do the difficulty levels of the generated questions match student abilities? (1 = Poor, 5 = Excellent)

Q7. How accurate and pedagogically appropriate are the AI-generated mark schemes? (1 = Poor, 5 = Excellent)

Section 4: Personalised Feedback

Q8. Is the AI-generated feedback helpful, clear, and easy for a GCSE student to understand? (1 = Not Helpful, 5 = Very Helpful)

Q9. How accurate is the AI-generated feedback in terms of scientific correctness and error identification? (1 = Poor, 5 = Excellent)

Q10. After teacher edits or suggestions, how well does the revised AI feedback follow the intended changes? (1 = Not at all, 5 = Completely)

Section 5: Overall Experience & Suggestions

Q11. Before using our tool, how did (or would) you complete similar tasks? (Select all that apply)
- Manually graded and wrote feedback myself
- Used prompt-based tools like ChatGPT / Claude
- Used another automated tool
- Other

Q12. On average, how much time did you previously spend per assignment (assuming a format similar to standard past paper design)?
- Less than 5 minutes
- 5–10 minutes
- 10–30 minutes
- More than 30 minutes
- Other: _____

Q13. Compared to your previous method(s), in which aspects does this tool perform better? (Select all that apply)
- Accuracy
- Speed
- Usability
- Reliability
- Alignment with curriculum

Q14. If you were to perform similar tasks again, which method would you prefer?
- I would continue using this tool
- I would return to my previous method(s)
- I would combine this tool with previous method(s)
- I'm not sure yet