Feedback Generation in Education using Large Language Models: A Survey of Recent Advances

Anonymous ACL submission

Abstract

Large language models (LLMs) have made it possible to generate formative educational feedback at scale, but naïve generation often fails to meet educational requirements. A rapidly growing line of work reframes feedback generation as a decision-making problem: systems think *pedagogically* before they speak, by making intermediate choices—what to target, what action to take, how much support to provide, and what evidence should justify the message—before realizing the feedback. This survey reviews recent 58 papers on LLM-based educational feedback generation that incorporates such deliberative structure. We organize existing systems by where decision-making lives: prompting and in-context planning, training-time alignment, inference-time candidate selection, and scripted pedagogical scaffolds. We highlight open challenges in signal and judgement reliability, construct validity, generalization, personalization and scaling. We conclude with recommendations for building more auditable, controllable, and pedagogically-grounded feedback systems.

1 Introduction

Feedback is widely regarded as one of the most powerful influences on learning and achievement. Done well, it can substantially boost student performance; done poorly, it can hinder it or even harm it (Hattie and Timperley, 2007; Mandouit and Hattie, 2023). *Formative feedback* is the information communicated to learners with the intention of modifying their thinking or behaviour to improve learning, and is most effective when it is timely, specific, credible, and supportive (Shute, 2008).

In parallel, rapid advances in the applications of large language models (LLMs) to education have opened new horizons for intelligent tutoring, automated assessment and content generation, and made it possible to automatically generate feedback at scale. Recent studies explore LLM feedback in educational settings, both when it succeeds and where it still fails. For example, LLMs have been widely studied for assessing student writing on criteria such as accuracy, specificity, and tone (Dai et al., 2024; Steiss et al., 2024; Escalante et al., 2023). Technical domains like computer programming have also benefited from LLM feedback identifying reasoning errors but it still sometimes struggle with subtle mistakes (Silva and Costa, 2025). Works using GPT models in assessment, formative (during instruction) and summative (evaluation), suggest the timely, personalised feedback is achievable, but with uneven quality and questionable reliability and impact on learning (Krumsvik, 2025; Jovic et al., 2025), and significant risks to academic integrity, the potential erosion of core skills, and the risk that AI feedback reinforces misconceptions or provides misleading information (Hasanein and Sobaih, 2023; García-López et al., 2025). Hallucination is another risk, underscoring the need for more robust mechanisms to verify and constrain model outputs in high-stakes domains including education (Huang et al., 2025; Sahoo et al., 2024).

Together, these strands point to a tension: educational feedback is an intervention that must be correct, pedagogically appropriate, and calibrated to the learner, which LLMs might not always get right through prompting, leading to generic, misaligned, or simply wrong feedback. This has led to a growing line of work that treats LLM-based feedback generation as a **deliberative** decision making process rather than a generation process. In this survey, we look at 58 papers (appeared in 2024/2025) that employs LLMs in feedback generation for education, addressing questions like what to optimize, where does the decision making live, what signals can guide the process, and where do experts fit in the loop.

1.1 Framing and Terminology

Feedback generation is an *instructional intervention* constrained by multiple objectives: it must be

content-correct, pedagogically aligned (e.g., hinting rather than revealing solutions), actionable, and appropriately calibrated in tone and confidence (e.g., not over- or under-praising). Feedback generation systems make at least one **intermediate decision** explicit before committing to the final output; typically selecting a feedback **target** (which error, misconception, or rubric dimension to address), an **action** (e.g., probe, hint, validate, request explanation, suggest revision, escalate), and **timing**. The feedback's *grounding signal* is an observable, external cue that anchors feedback planning, including unit test results, revised answer quality or instructor edits (Ma et al., 2025; Wang et al., 2025a; Mok et al., 2025).

Thus, a feedback policy

$$\pi_F : (I, T, C, S) \rightarrow (A_F, t_F)$$

maps a learner's input (I) for a task specification (T) under the learning context, curriculum and constraints (C) and the observable signals (S) into a timed feedback action (A_F, t_F). This framing allows comparing systems across domains by asking what learner artifact they utilize, what signals guide those decisions, and which constraints dominate.

1.2 Organisation

We organise this survey following a thematic flow: we discuss the domains and subject areas covered in the literature (§2), the architectures, mechanisms and workflows that are the backend of the surveyed systems (§ 3), then we discuss the gaps, open problems and opportunities that we observe from the pool of literature (§4). We also discuss evaluation strategies and rubrics in the Appendix (§C).

2 Domains and subject areas

2.1 Math

The works that handle math education in this survey (Table 1) concentrate in three settings: (i) *dialogue tutoring*, where planning is sequential and stateful, and feedback generation conditions on the full interaction to choose what is appropriate next with success predictors (Scarlatos et al., 2025b); (ii) *multi-choice questions setting*, where distractors provide explicit error handles, and feedback targets the misconceptions implied by, or explicitly associated with, the selected distractor (Scarlatos et al., 2024; McNichols et al., 2024a); and (iii) *open-ended response setting*, where free-form student solutions must be analysed for misconception targets, using reference/generated solution, and the intervention strength level should be chosen accordingly (McNichols et al., 2024b; Baral et al., 2024; Tonga et al., 2025; Pal Chowdhury et al., 2024).

2.2 Programming

Coding papers in the survey (Table 1) fall into three settings: (i) *outcome-grounded executable tasks*, where feedback should help pass code checks (e.g., unit tests), in single-turn (Chae et al., 2024) or multi-turn (Wang et al., 2025b) with appropriate type (explanation, hint, question) and strength (Lohr et al., 2025; Heickal and Lan, 2024); (ii) *classroom-wide workflows*, where feedback is generated at the aggregate level but get reviewed, edited, and dispatched by the instructor (Tang et al., 2025a,b), constrained by explicit course policies (Scholz et al., 2025); and (iii) *interactive learning interventions* where the system supports the learning process itself (e.g., proposing debugging hypotheses (Ma et al., 2024) or reflection on alternatives (Naik et al., 2024)), with explicit focus on when to intervene (Ghoochani et al., 2025).

2.3 Writing

Writing papers in our survey (Table 1) cluster into three settings: (i) *revision-grounded feedback*, where the quality of a draft and predicted revision improvement shape feedback (Nair et al., 2024), sometimes with hierarchical criteria (Yuan et al., 2024; Chamoun et al., 2024; Wang et al., 2025c) or consistency check across repeated submissions (Bhojan and Xin, 2025); (ii) *interactive cognitive support*, where feedback targets different cognitive stages of the creative process (e.g., ideation vs. evaluation), balancing being helpful with maintaining writer agency (Göldi et al., 2024); and (iii) *persona conditioning*, where feedback is a function of a writer-defined perspective that acts as an explicit lens with a stance that aligns with the writer's intent (Benharrak et al., 2024).

2.4 Other domains

We also identified literature that address spoken language assessment, multimodal language tutoring and STEM (e.g. science assignments), see Table 1.

3 Mechanisms

A *feedback mechanism* is the concrete approach in which a system performs decision making before committing to feedback. Mechanisms differ both in *where* decision making happens , *which*

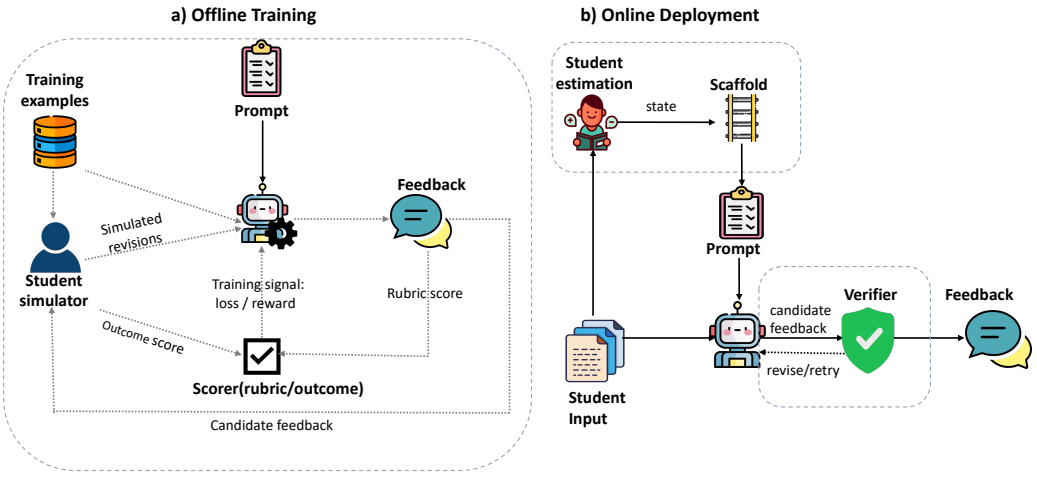

Figure 1: Overview of feedback generators: offline policy learning (Left), and scaffolded, verified feedback (Right).

intermediate decisions are available, *what* artifacts they introduce , and what supporting components they utilize. Note that these mechanisms are not mutually exclusive and can be synthesized.

3.1 Prompting and in-context learning

Here, the system uses an off-the-shelf LLM and describes the task and the constraints fully in the prompt. The feedback planning here happens within the model, selecting both (1) what to talk about (e.g., which error categories to address, or rubric) and (2) how (e.g., hint vs. explanation).

In writing, prompts explicitly guide the model to first elicit the relevant criteria and then generate critiques and suggestions that are traceable to them (Yuan et al., 2024), or to identify the most salient issues in drafts using a detailed rubric in zero- or few-shot settings (Rashkin et al., 2025), or using a rubric-based score (Han et al., 2024). In scientific writing, feedback pipelines similarly treat selecting what to focus on as a deliberate intermediate step, producing targeted revisions rather than broad, undirected commentary (Chamoun et al., 2024). Assessment workflow prompts produce multi-dimensional analytic judgments that can then be translated into actionable comments (Wang et al., 2025c). In math, in-context examples can help generate distractors and aligned explanations, constraining the feedback that follows (McNichols et al., 2024a). Finally, spoken-language assessment pipelines combine upstream transcription with rubric-aligned prompting to produce both scores and feedback (Shankar et al., 2025).

Prompting could also support personalization by injecting the right learner context (e.g., prior attempts, misconceptions, or prior feedback) in the prompt. In (Reddig et al., 2025) the intermediate planning step is to choose the right context/examples to include to make the feedback align with the student's needs.

3.2 Fine-tuning and preference learning

Here, an LLM is optimized at training time for feedback quality (*rubric-grounded*), or to the outcome after adopting the feedback (*outcome-grounded*).

Rubric-grounded preference optimization. Several surveyed systems defined good feedback by an explicit pedagogical rubric, involving correctness, non-revelation of answers, diagnostic value, actionability and tone. Fine-tuning or preference learning is used at training-time to optimize a feedback policy toward 'good' feedback.

This pattern is explicit in diagnostic math feedback, where contrastive pairs of rubric-labeled feedback candidates are used for offline preference learning with DPO (Scarlatos et al., 2024). A closely related approach appears in pedagogical-alignment training, where structured pedagogical rubric, designed to distinguish guidance from answer giving, is used to build a preference dataset (Sonkar et al., 2024). Training-time optimization also appears in multi-agent settings (Yang et al., 2025) where collaborative agents are optimized for group discussions with pedagogy-guided interaction modelling, where intervention timing and strategy are modelled explicitly.

Outcome-grounded optimization A stronger form of objective-based decision making targets task success after adopting the feedback instead

of rubrics, usually requiring student simulation or explicit tool usage for verification at training time. For tasks with executable outcome, the grounding signal for feedback is the execution results after the simulated learner acts on the proposed feedback. (Chae et al., 2024) trains a a feedback policy for the coding domain using Reinforcement Learning (RL) with an editor model (acting like a simulated student) that revises code given feedback, and unit test performance deciding rewards. In math tutoring, preference optimization is similarly used to push a tutor toward a hybrid objective that combines rubric-guided pedagogically-appropriate feedback with a prediction of learner success using knowledge tracing (Scarlatos et al., 2025b).

Beyond executable outcomes, (Dinucu-Jianu et al., 2025) frames the core tension of problem-solving vs. *teaching* problem-solving as a learnable policy, using RL to push tutoring toward pedagogically-desirable trajectories (e.g., reasoning). Synthetic data generation could also be utilized (Qian et al., 2025b), e.g. with simulated multi-agent teacher–student assignment–feedback loops (Zhang et al., 2025a), or multi-step synthetic Socratic dialogues between student, teacher and dean agents are used to fine-tune models for better pedagogical alignment (Liu et al., 2024a).

3.3 Runtime selection with learned verifiers

Another recurring mechanism is explicit inference-time feedback ranking of generated candidate feedback actions, evaluated by a learned scorer module. (Wang et al., 2025b) instantiates this mechanism in coding tutoring by sampling many candidate utterances for each turn, based on an estimated student state, then ranking them by a turn-by-turn scorer for the one most likely to make progress toward task completion. This is particularly useful when the action space is large (many plausible tutor moves).

LLM evaluators are also used to rank feedback candidates along pedagogical criteria in multi-stage agentic pipelines, e.g. in writing (Chamoun et al., 2024), or as a comprehensive automated evaluation layer that serves as a gatekeeper (Qian et al., 2025a). The reliability of these judges is a concern (Seo et al., 2025), but iterative evaluation and revision can stabilize feedback and reduce variance in reflective writing (Bhojan and Xin, 2025).

3.4 Scripted pedagogy scaffolds

Some of the surveyed works embed LLMs within external pedagogical scripts/ intervention policies that decide what kind of support to provide when. The scaffold provides explicit decision making for the learning stage and trigger conditions, intervention timing, type and stance, before relaying that to the LLM. It is argued that these scripts should be grounded in assessment evidence and decision modules (e.g., ZPD-informed adaptation) to determine their effective timeline (Cohn et al., 2025).

For academic writing, (Chen et al., 2025) introduce an 11-step scaffold to guide the learner from identifying key questions to readability and grammar checking. Programming feedback ladders impose ordered levels of support (identification to progressively more directive hints) (Heickal and Lan, 2024; Lohr et al., 2025), as well as in math (Tonga et al., 2025). Alternatively, finite state transducers with solution space decomposition guides the process (Pal Chowdhury et al., 2024).

In the collaborative SQL optimisation domain (Naik et al., 2024), types of reflection get triggered by pattern matching, prompting ChatGPT to produce timely, personalised interventions that help surface alternative solution paths. For peer review writing, inspired by the Cognitive Process Theory of Writing, assistance is explicitly conditioned on cognitive goals (ideation vs. evaluation) and constrains support through interface design (Göldi et al., 2024). Persona-conditioned feedback makes *stance* a planned variable, letting writers define audience lenses that shape feedback perspective/ specificity (Benharrak et al., 2024).

Alternative workflows that benefit from LLMs include *Learning-by-teaching* (Ma et al., 2024) where novices teach simulated students, with structured tasks that keep attention on constructing hypotheses and test suites. The *productive failure* approach (Puech et al., 2025) keeps students in a beneficial struggle regime before providing resolution. More generally, multimodal language tutoring can embed a script of "pedagogical instruction" to guide the tutoring and modality use (Liu et al., 2024b), and agentic tutors can hard-code Socratic guided questioning as a scaffold while grounding the interaction via retrieval (Knievel et al., 2025).

4 Open problems and opportunities

4.1 Signal reliability and judge validity

Educational feedback pipelines make intermediate choices using signals from rubrics, LLM-as-a-judge scores and mistake detectors. These signals can be noisy, biased, or brittle in ways that

materially affect downstream decisions, especially in subjective dimensions like tone, positivity, and perceived helpfulness. This could surface in rubric-grounded optimization (Scarlatos et al., 2024), as well as for automatic judges where claims about improved tutor behaviour are hindered by the evaluators diverging from human judgments (Scarlatos et al., 2025b), by bias and imperfect constraint checking (Ferraz et al., 2024), or sensitivity to surface forms such as verbosity (Seo et al., 2025).

A partial remedy is to shift to grounded verifiable signals when possible as judging can miss subtle issues (McNichols et al., 2024b; Baral et al., 2024). In programming, tool-grounded rewards (e.g. unit tests) reduce reliance on judges (Chae et al., 2024). Tool-interactive critique offers a primitive for grounding revisions in external evidence (Gou et al., 2024), while constraint decomposition turns rubrics into explicit constraints for selective refinement (Ferraz et al., 2024).

A more general requirement is to treat the evaluator/judge as a *first-class* component, not a convenience scorer, requiring **calibration and auditing**. Probing LLM evaluators shows that agreement and accuracy depend on evaluator choice and setup. Additional evaluators can act as *control signals* for quality gating and regeneration, e.g. the *Dean* framing which places a specialized feedback evaluator upstream of delivery to reject low-quality feedback and trigger regeneration (Qian et al., 2025a). Without human-grounded validation, adversarial/counterfactual stress tests, or uncertainty reporting for borderline decisions, pipelines risk *correlated failure modes*: generator and evaluator sharing blind spots, and high scores becoming self-referential rather than evidence of pedagogical quality, as could be seen in analytic writing assessment (Wang et al., 2025c).

Finally, reliability issues can also arise *before* planning starts, as in spoken-language assessment where errors and subgroup performance differences can propagate into scoring and feedback (Shankar et al., 2025), motivating **uncertainty-aware pipelines** that carry confidence from upstream perception into downstream decisions, and decision signals that are both **meaningful** and **auditable** in domains without crisp executability like writing.

4.2 Construct validity

Even when evaluation signals are reliable, they may measure imprecise or ill-posed educational constructs, e.g. unit tests capturing functional correctness in place of pedagogical gains in coding learning (Chae et al., 2024), superficial rubric-based revision improvements in essay writing that have a weak connection to learning outcome (Nair et al., 2024), or hallucinated or incomplete generated solution trees that could hinder math tutoring (Pal Chowdhury et al., 2024).

In the learning-science perspective, ensuring what systems optimize is meaningful calls for feedback design that is aligned with instructional goals and learner outcomes (Stamper et al., 2024). Process-centred interventions provide exemplars where the target is the learning behaviour itself, such as in hypothesis construction for debugging (Ma et al., 2024), or situated reflection that supports exploring alternatives (Naik et al., 2024). Similarly, a classroom randomized control trial found that GPT-4-driven interactive homework improves engagement and learning outcomes in grammar (Vanzo et al., 2025).

Still, what is lacking is a scalable way to connect intermediate decision quality (e.g. target selection, hint choice, escalation timing) to *real, validated outcomes* to inform feedback planning, without relying on expensive classroom trials. Explaining *why* gains occur and *for whom* require aligning what planners optimize with behaviours that plausibly drive learning (e.g., revision choices, practice quality, persistence), and reporting behavioural outcomes (what learners change vs. ignore), retention, and transfer. A concrete opportunity is to standardize reporting that connects feedback attributes to learner actions (edits made, strategy shifts).

4.3 Generalisation beyond simulators/proxies

Gains of some feedback generation research comes under simulated learners (e.g. dialogue knowledge tracing (Scarlatos et al., 2025a)), or revisers (e.g. simulated writing revisers (Nair et al., 2024)). Two risks are attached to that: (i) *fidelity mismatch* when the simulated behaviour diverges from the real learning, and (ii) *proxy overfitting* when a system fails to generalise beyond the simulation.

Furthermore, while the use of synthetic corpora (e.g. SCALEFeedback for CS assignments (Qian et al., 2025b)) and agentic simulations (e.g. SEFL self-feedback loops (Zhang et al., 2025a)) are expanding to counteract scarce classroom data, a transfer risk arises: improvements may reflect synthetic error patterns rather than authentic learning. Partial remedy for this includes acquiring **human evidence** that interventions survive reality, even

at small scale (e.g., debugging support (Ma et al., 2024)), and more systematic effort to **elicit simulation parameters from real trajectories** to calibrate simulators and stress-test parameter sensitivity in prompts and learner profiles.

4.4 Rubric-compliant generation

Rubric compliance at generation still lacks robustness, especially in long contexts or multi-turn settings. Even strong models often violate at least one constraint in real multi-constraint instructions (Ferraz et al., 2024), leading to phenomena like over-praise or over-inference (Guo et al., 2024) or drift in scaffolded-help behaviour (Sonkar et al., 2024).

A partial remedy is to enforce explicit criteria through structured critique or checkable tests. This makes **criterion selection and prioritization** a key problem: what to surface now under limited attention, as models can focus on low-hanging issues and miss the most salient problems otherwise (Rashkin et al., 2025). Another remedy is to ground critique with tools when possible: tool-interactive critique reduces hallucinations by grounding feedback in external verification (Gou et al., 2024), and outcome-grounded approaches make constraints become signals with suitable checkers (Chae et al., 2024). Yet, producing the intended *kind* of feedback remains unreliable (Lohr et al., 2025), motivating structured scaffolds such as feedback ladders (Heickal and Lan, 2024). These make strategy explicit (e.g., productive failure intent selection (Puech et al., 2025)).

Overall, the field still lacks a widely accepted **pedagogical constraint schema** that is simultaneously decomposable (Ferraz et al., 2024), checkable when possible (Chae et al., 2024), and aligned with learning-science constructs (hinting vs. telling, autonomy support, desirable difficulty), especially for long-form responses (Rashkin et al., 2025).

4.5 Overhead

Feedback mechanisms increase cost and latency through candidate sampling, verifier calls and iterative refinement. Long tool-interactive correction loops are one driver of overhead (Gou et al., 2024), but selective refinement can reduce unnecessary rewriting by focusing on detected violations (Ferraz et al., 2024). In classroom workflows, responsiveness also depends on human factors: instructor-facing systems mitigate workload via structured review and attention-aware batching (Tang et al., 2025b,a). Some approaches offload cost to offline training (Chae et al., 2024) or caching (Pal Chowdhury et al., 2024), but deployment feasibility still depends on online constraints. And with more planning, multi-agent loops, and evaluator gating, quality can improve but costs rise (Chamoun et al., 2024; Zhang et al., 2025a; Qian et al., 2025a).

A key opportunity is **adaptive budgeting**: put compute on verification/regeneration only when stakes or uncertainty are high. Spoken-language assessment makes this concrete as upstream uncertainty can be estimated to trigger conservative feedback, extra checks, or human review (Shankar et al., 2025). Also, characterizing the *cost-quality frontier* is needed: when additional candidates, refinement, verifier calls, or human review are worth it under strict classroom latency and constraints.

4.6 Safe scaling

Scaling feedback at the classroom level introduces systemic risk: errors can propagate widely if aggregation/reuse is faulty. Instructor-validation workflows mitigate this by structuring review and propagation, but they also expose the need for additional safeguards when issue detection is imperfect (Tang et al., 2025b,a). Tutor CoPilot supports live tutors with only suggestions (Wang et al., 2024b) but raises research questions about how AI reshape tutor behavior and student autonomy. Propagation safeguards must go beyond correctness to audit non-leakage, course policy compliance, subgroup fairness, and preservation of instructional intent.

Recent deployments highlight the stakes: automated project feedback can reach many students quickly (Ghoochani et al., 2025), and interactive homework tutoring can be deployed at classroom scale with measurable learning gains (Vanzo et al., 2025). The open problems are **monitoring and rollback**: detecting recurring misconceptions or policy-violations, preventing repeated errors across cases, and notifying affected learners. Safe scaling also requires governance of data and uncertainty. Privacy- and sensitivity-aware handling of learner data (audio, reflective writing) constrains what can be stored, retrieved, or reused (Shankar et al., 2025; Bhojan and Xin, 2025; Jovic et al., 2025), and uncertainty-aware control suggests that low-confidence cases may warrant different scaling policies (e.g., reduced propagation).

4.7 Adaptivity and Personalization

Many feedback generators remain generic, and when personalization is introduced (knowledge

tracing, writer-defined personas, attention-driven interfaces), it raises evaluation and safety questions. Learning-science framings argue personalization should be grounded in theory and evaluated beyond subjective preference (Stamper et al., 2024). A theory of adaptive scaffolding argues that pedagogical agents should modulate support based on learner signals while avoiding over- and under-scaffolding (Cohn et al., 2025), while in-context personalized tutor feedback relies on inferring accurate learner state and misconceptions (Reddig et al., 2025). Systems operationalize this via adaptive platforms that plan learning paths and per-step support (Chudziak and Kostka, 2025).

Personalization also extends beyond individuals: timely intervention in collaborative discussions requires modeling group progress and interaction dynamics (Yang et al., 2025). However, deployment evidence suggests motivational and context-aware tailoring remains uneven even when students prefer automated feedback (Ghoochani et al., 2025).

Two under-developed research directions are **calibration** (detecting when personalization is off: over-hinting, under-challenging, mismatched tone) and **equity** (auditing persona-conditioned policies for stereotyping or differential treatment).

4.8 Weak comparability

Comparability between feedback generation systems is hard because objectives and criteria are not made commensurate. And while the design space is converging, the field still needs standardized **objective bundles** that settles primary objective, constraints, acceptable proxies and required robustness checks, and enforcing explicit reporting of evaluator calibration and uncertainty when LLM judges are used. TutorGym and MathTutorBench provide building blocks by offering structured testbeds for tutoring behavior and pedagogical capability (Weitekamp et al., 2025; Macina et al., 2025), while efforts like SCALEFeedback (Qian et al., 2025b) support reproducible training and evaluation.

5 Discussion and Conclusion

This survey reviewed recent work (2024–2025) of LLM-based educational feedback generation. A consistent pattern emerges: making *pedagogically valid and reliable choices*—what to address, when to intervene, how much to reveal, what evidence to justify—under real deployment constraints. This survey contributes a cross-domain analysis that organizes mechanisms by *where* decision-making lives (prompting, training-time optimization, verification and selection, scaffolding).

A core view in the literature is moving from prompting to *feedback policies*: explicit mapping that connects observable grounding signals to feedback strategies, whether learned (e.g., preference optimization) or scripted (e.g., scaffolding rules), in a way that is auditable, stable, and safe at scale. These intermediate decisions make systems more comparable across domains. The surveyed work shows that even lightweight explicit structure—such as selecting criteria before drafting (Yuan et al., 2024; Chamoun et al., 2024), laddered levels of help (Heickal and Lan, 2024), or intent/state transitions that operationalize productive failure (Puech et al., 2025)—can convert otherwise underspecified "helpfulness" into controllable behavior.

Naïve feedback generation fails in high-stakes educational settings due to LLM propensity to violate some of the domain's multi-objective constraints (correctness, non-revelation, actionability, calibration, tone), especially in long contexts/multi-turn settings (Ferraz et al., 2024; Sonkar et al., 2024). Decision-making approaches make the constraints explicit, enabling mechanisms such as candidate selection with verifiers (Wang et al., 2025b), self-critique (Yuan et al., 2024), or rubric-grounded preference learning (Scarlatos et al., 2024).

Across mechanisms (§3), surveyed systems can be abstracted into a common pipeline: 1) **Grounding signals**: observable cues that anchor the intervention (unit tests, revisions, rubrics, instructor edits, interaction traces). 2) **Intermediate decisions**: discrete choices made *because* of those signals (what to target, what action to take, what strength, what evidence to cite, whether to escalate or verify). 3) **Feedback realization**: the surface language form, constrained by the action and policy (hint vs. explanation, Socratic prompt, tone).

This decomposition helps reconcile seemingly different approaches. Training-time approaches learn a policy that maps signals to decisions (§3.2); verifier-based approaches decide among candidates (§3.3); and scripted scaffolds implement explicit pedagogical control policies (§3.4). Instructor-in-the-loop systems further externalize decision-making into workflow primitives such as triage, verification, and reuse (Tang et al., 2025b), which are central to reliability at scale.

Crucially, this view is actionable as a *design and reporting template*. It forces clarity about what

signals are trusted, which intermediate decisions are exposed, and what realization constraints are enforced. This highlights when a system is implicitly relying on subjective judgments versus when it can anchor decisions in checkable outcomes.

This also highlights a practical implication: when reliable grounding signals exist (e.g., unit tests), systems can rely less on subjective evaluation and more on verifiable outcomes (Chae et al., 2024). When signals are inherently subjective or weakly grounded (e.g., writing quality), systems must invest in robust rubrics, evaluator calibration, and evidence-grounded critiques to avoid unfounded feedback (Chamoun et al., 2024; Seo et al., 2025). This leads to key takeaways:

- **Explicit intermediate decisions** (targets, actions, scaffold level, escalation) are a high-leverage route to auditability and controllability (Yuan et al., 2024; Heickal and Lan, 2024; Puech et al., 2025).
- **Signal validity is the bottleneck**: as systems become more policy-like, failure increasingly comes from mis-calibrated decision signals, especially judge signals (Seo et al., 2025; Qian et al., 2025a).
- **Evaluation must be decomposed** into action quality, feedback quality, and impact; otherwise gains are easily misattributed to surface fluency (Chae et al., 2024; Sonkar et al., 2024).
- **Cost-aware reliability** for real classrooms: verification/planning must be budgeted adaptively, not uniformly (Ferraz et al., 2024; Tang et al., 2025b).
- **Learning-science constructs** should increasingly become decision variables, not only prompt phrasing (Hattie and Timperley, 2007; Shute, 2008).

Decision-making and verification increase cost and latency, creating a tension between reliability and feasibility (§4.5). Many systems therefore implicitly assume either generous compute (e.g. multi-stage agentic pipelines) or settings where delays are acceptable. Classroom deployments, however, often require real-time responsiveness and robust safeguards against widespread mistake propagation (Tang et al., 2025b). This elevates two practical research directions: 1) **adaptive budgeting**: allocate verification and regeneration selectively based on uncertainty, stakes, and the availability of grounding signals, rather than uniformly (Ferraz et al., 2024). Concretely, triggers can be disagreement between candidates, missing or low-quality grounding signals, or high-stakes course policy cases (e.g., solution leakage); and 2) **monitoring and rollback**: as feedback is reused or broadcast at scale, systems need mechanisms to detect systematic failure patterns (e.g., recurring misconceptions, policy violations, subgroup disparities) and prevent repeated harm (Tang et al., 2025b).

Deployment realism also includes governance constraints beyond latency: privacy and sensitivity of learner data and equity risks that arise when evaluator bias or personalization policies behave differently across subgroups (Shankar et al., 2025; Nazaretsky et al., 2025). These affect safe scaling because they restrict what can be stored, reused, and propagated, and they shape which "signals" are acceptable as decision inputs.

Finally, the surveyed literature reinforces a learning-science point that is easy to lose in NLP-centric framing: educational feedback is an *instructional intervention*, not merely error correction (Hattie and Timperley, 2007; Shute, 2008). Several works explicitly model the tension between solving and teaching problem solving (Dinucu-Jianu et al., 2025), or structure support around productive struggle and scaffolding (Puech et al., 2025; Cohn et al., 2025). These directions suggest that the next wave of planning-based feedback research should more directly encode theoretically grounded constructs—e.g., autonomy support, desirable difficulty, and scaffold fading—as decision variables, and evaluate them with behavioral outcomes (what learners do next), not only perceived helpfulness.

Operationally, this means taking pedagogical principles as *choices* the system must make: e.g., selecting a Socratic prompt vs. a direct hint (autonomy support), delaying solution revelation until evidence of impasse (desirable difficulty), and reducing hint specificity as competence improves (scaffold fading). Framed this way, learning-science alignment becomes an implementable agenda rather than an after-the-fact description.

Conclusion. LLMs have made feedback generation widely accessible, but they emphasize the *policy* question. Across the recent work, we observe meaningful convergence on planning-based formulations that leverage grounding signals, intermediate decisions, and explicit workflows to improve reliability, controllability, and scalability. We identify open issues of trustworthy signals, construct validity, generalization, and safe scaling under classroom constraints. Addressing these requires closer integration between NLP mechanisms, educational theory and evaluation practice, alongside standardized reporting that makes objectives, signals, and intermediate decisions explicit.

6 Limitations

Temporal scope. This survey focuses on work published in 2024–2025, with the goal of capturing the most recent wave of LLM-based feedback systems that explicitly incorporate planning or decision-making. As a result, we do not attempt to comprehensively cover earlier foundational literature on intelligent tutoring systems, automated feedback, or pre-LLM approaches, and we may also miss very recent papers that appeared after our cutoff.

Search methodology and coverage. Our paper collection relied on a manual, keyword-driven search and screening process, rather than a fully exhaustive systematic review. We used queries centered on LLMs and feedback/planning, including terms such as LLM-generated feedback, formative feedback, automated feedback, LLM tutor/tutoring, hint generation, Socratic tutoring, rubric-based feedback, programming feedback, writing feedback, feedback policy, verification/verifier, and LLM-as-a-judge. Since terminology differs across NLP, learning sciences, and HCI/EdTech communities, and because some relevant work is framed as assessment, scaffolding, revision support, or instructional interventions rather than "feedback", our coverage is necessarily incomplete and may omit pertinent studies that use different descriptors or appear in adjacent venues.

Space constraints. ACL page limits require prioritizing the core synthesis and taxonomy in the main paper. We therefore place extended material such as expanded tables, evaluation details, dataset summaries, and additional examples or elaborations into the appendix. This improves readability of the main narrative but may reduce the detail available in the main text for some subsets of surveyed work.

References

Sami Baral, Eamon Worden, Wen-Chiang Lim, Zhuang Luo, Christopher Santorelli, and Ashish Gurung. 2024. Automated assessment in math education: A comparative analysis of llms for open-ended responses. *educational data mining*, pages 732–737.

Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined ai personas for on-demand feedback generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Anand Bhojan and Tan Li Xin. 2025. Reflexai: Optimizing LLMs for consistent and constructive feedback in reflective writing. In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 387–394. SciTePress.

Hyungjoo Chae, Taeyoon Kwon, Seungjun Moon, Yongho Song, Dongjin Kang, Kai Tzu-iunn Ong, Beong-woo Kwak, Seonghyeon Bae, Seung-won Hwang, and Jinyoung Yeo. 2024. Coffee-gym: An environment for evaluating and improving natural language feedback on erroneous code. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22503–22524, Miami, Florida, USA. Association for Computational Linguistics.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763.

Fumian Chen, Sotheara Veng, Joshua Wilson, Xiaoming Li, and Hui Fang. 2025. Coachgpt: A scaffolding-based academic writing assistant. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4051–4055.

Jarosław A Chudziak and Adam Kostka. 2025. Ai-powered math tutoring: Platform for personalized and adaptive education. In *International Conference on Artificial Intelligence in Education*, pages 462–469. Springer.

Clayton Cohn, Surya Rayala, Namrata Srivastava, Joyce Horn Fonteles, Shruti Jain, Xinying Luo, Divya Mereddy, Naveeduddin Mohammed, and Gautam Biswas. 2025. A theory of adaptive scaffolding for llm-based pedagogical agents. *arXiv preprint arXiv:2508.01503*.

Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tongguang Li, Dragan Gašević, and Guanliang Chen. 2024. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7:100299.

David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. 2025. From problem-solving to teaching problem-solving: Aligning LLMs with pedagogy using reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 272–292, Suzhou, China. Association for Computational Linguistics.

Juan Escalante, Austin Pack, and Alex Barrett. 2023. Ai-generated feedback on writing: insights into efficacy and enl student preference. *International Journal of Educational Technology in Higher Education*, 20(1):57.

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu,

Vivferraz2024llmek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. LLM self-correction with DeCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812, Miami, Florida, USA. Association for Computational Linguistics.

Iván Miguel García-López, Carina Soledad González González, María Soledad Ramírez-Montoya, and José Martín Molina-Espinosa. 2025. Challenges of implementing chatgpt on education: Systematic literature review. *International Journal of Educational Research Open*, 8:100401.

Fatemeh Ghoochani, Jonas Scharfenberger, Burkhardt Funk, Raoul Doublan, Mayur Jakharabhai Odedra, and Bennet Etsiwah. 2025. From feedback to formative guidance: Leveraging llms for personalized support in programming projects. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 398–403.

Andreas Göldi, Thiemo Wambsganss, Seyed Parsa Neshaei, and Roman Rietsche. 2024. Intelligent support engages writers through relevant cognitive processes. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Shuchen Guo, Ehsan Latif, Yifan Zhou, Xuan Huang, and Xiaoming Zhai. 2024. Using generative ai and multi-agents to provide automatic feedback. *arXiv preprint arXiv:2411.07407*.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293, Miami, Florida, USA. Association for Computational Linguistics.

Ahmed M. Hasanein and Abu Elnasr E. Sobaih. 2023. Drivers and consequences of chatgpt use in higher education: Key stakeholder perspectives. *European Journal of Investigation in Health, Psychology and Education*, 13(11):2599–2614.

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

Hasnain Heickal and Andrew Lan. 2024. Generating feedback-ladders for logical errors in programming using large language models. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 947–951.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Marina Jovic, Stavros Papakonstantinidis, and Robert Kirkpatrick. 2025. From red ink to algorithms: investigating the use of large language models in academic writing feedback. *Language Testing in Asia*, 15(1):59.

Christopher Knievel, Alexander Bernhardt, and Christian Bernhardt. 2025. Aitee–agentic tutor for electrical engineering. *arXiv preprint arXiv:2505.21582*.

Rune Johan Krumsvik. 2025. Gpt-4's capabilities for formative and summative assessments in norwegian medicine exams—an intrinsic case study in the early phase of intervention. *Frontiers in Medicine*, 12:1441747.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024a. Socraticlm: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721.

Zhengyuan Liu, Stella Xin Yin, Carolyn Lee, and Nancy F Chen. 2024b. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. In *2024 IEEE conference on artificial intelligence (CAI)*, pages 1258–1265. IEEE.

Dominic Lohr, Hieke Keuning, and Natalie Kiesler. 2025. You're (not) my type-can llms generate feedback of specific types for introductory programming tasks? *Journal of Computer Assisted Learning*, 41(1):e13107.

Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to teach programming in the ai era? using llms as a teachable agent for debugging. In *International Conference on Artificial Intelligence in Education*, pages 265–279. Springer.

SHengjie Ma, Chenlong Deng, Jiaxin Mao, Jiadeng Huang, Teng Wang, Junjie Wu, Changwang Zhang, and 1 others. 2025. Pou: Proof-of-use to counter tool-call hacking in deepresearch agents. *arXiv preprint arXiv:2510.10931*.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *arXiv preprint arXiv:2502.18940*.

Luke Mandouit and John Hattie. 2023. Revisiting "the power of feedback" from the perspective of the learner. *Learning and Instruction*, 84:101718.

Hunter McNichols, Wanyong Feng, Jaewook Lee, Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2024a. Automated distractor and feedback generation for math multiple-choice questions via in-context learning. *Preprint*, arXiv:2308.03234. ArXiv:2308.03234.

Hunter McNichols, Jaewook Lee, Stephen Fancsali, Steve Ritter, and Andrew Lan. 2024b. Can large language models replicate its feedback on open-ended math questions?

Jisoo Mok, Ik-hwan Kim, Sangkwon Park, and Sungroh Yoon. 2025. Exploring the potential of LLMs as personalized assistants: Dataset, evaluation, and analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10212–10239, Vienna, Austria. Association for Computational Linguistics.

Atharva Naik, Jessica Ruhan Yin, Anusha Kamath, Qianou Ma, Sherry Tongshuang Wu, Charles Murray, Christopher Bogart, Majd Sakr, and Carolyn P Rose. 2024. Generating situated reflection triggers about alternative solution paths: A case study of generative ai for computer-supported collaborative learning. In *International Conference on Artificial Intelligence in Education*, pages 46–59. Springer.

Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. 2024. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16636–16657.

Tanya Nazaretsky, Hagit Gabbay, and Tanja Käser. 2025. Can students judge like experts? a large-scale study on the pedagogical quality of ai and human personalized formative feedback. *Computers and Education: Artificial Intelligence*, page 100533.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.

Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26291–26311.

Keyang Qian, Yixin Cheng, Rui Guan, Wei Dai, Flora Jin, Kaixun Yang, Sadia Nawaz, Zachari Swiecki, Guanliang Chen, Lixiang Yan, and 1 others. 2025a. Dean of llm tutors: exploring comprehensive and automated evaluation of llm-generated educational feedback via llm feedback evaluators. *arXiv preprint arXiv:2508.05952*.

Keyang Qian, Kaixun Yang, Wei Dai, Flora Jin, Yixin Cheng, Rui Guan, Sadia Nawaz, Zachari Swiecki, Guanliang Chen, Lixiang Yan, and 1 others. 2025b. Scalefeedback: a large-scale dataset of synthetic computer science assignments for llm-generated educational feedback research. *arXiv preprint arXiv:2508.05953*.

Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2025. Help me write a story: Evaluating LLMs' ability to generate writing feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25827–25847, Vienna, Austria. Association for Computational Linguistics.

Jennifer M Reddig, Arav Arora, and Christopher J MacLellan. 2025. Generating in-context, personalized feedback for intelligent tutors with large language models. *International Journal of Artificial Intelligence in Education*, pages 1–42.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Scarlatos, Ryan S Baker, and Andrew Lan. 2025a. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 249–259.

Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025b. Training llm-based tutors to improve student learning outcomes in dialogues. In *International Conference on Artificial Intelligence in Education*, pages 251–266. Springer.

Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Improving the validity of automatically generated feedback via reinforcement learning. In *International Conference on Artificial Intelligence in Education*, pages 280–294. Springer.

Niklas Scholz, Manh Hung Nguyen, Adish Singla, and Tomohiro Nagashima. 2025. Partnering with ai: A pedagogical feedback system for llm integration into programming education. In *European Conference on Technology Enhanced Learning*, pages 243–248. Springer.

Hyein Seo, Taewook Hwang, Jeesu Jung, Hyeonseok Kang, Hyuk Namgoong, Yohan Lee, and Sangkeun Jung. 2025. Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences (2076-3417)*, 15(2).

Natarajan Balaji Shankar, Kaiyuan Zhang, Andre Mai, Mohan Shi, Alaria Long, Julie Washington, Robin Morris, and Abeer Alwan. 2025. Leveraging asr and

llms for automated scoring and feedback in children's spoken language assessments. In *Proc. SLaTE 2025*, pages 1–5.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Priscylla Silva and Evandro Costa. 2025. Assessing large language models for automated feedback generation in learning programming problem solving. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 116–124. PMLR.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650.

John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*, pages 32–43. Springer.

Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894.

Xiaohang Tang, Sam Wong, Zicheng He, Yalong Yang, and Yan Chen. 2025a. Reva: Supporting llm-generated programming feedback validation at scale through user attention-based adaptation. *arXiv preprint arXiv:2507.11470*.

Xiaohang Tang, Sam Wong, Marcus Huynh, Zicheng He, Yalong Yang, and Yan Chen. 2025b. Sphere: Supporting personalized feedback at scale in programming classrooms with structured review of generative ai outputs. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Junior Cedric Tonga, Benjamin Clement, and Pierre-Yves Oudeyer. 2025. Automatic generation of question hints for mathematics problems using large language models in educational technology. In *Proceedings of Large Foundation Models for Educational Assessment*, volume 264 of *Proceedings of Machine Learning Research*, pages 61–102. PMLR.

Alessandro Vanzo, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2025. Gpt-4 as a homework tutor can improve student engagement and learning outcomes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31119–31136.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. 2025a. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*.

Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. 2025b. Training turn-by-turn verifiers for dialogue tutoring agents: The curious case of LLMs as your coding tutors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12416–12436, Vienna, Austria. Association for Computational Linguistics.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024a. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. 2024b. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.

Zhengxiang Wang, Veronika Makarova, Zhi Li, Jordan Kodner, and Owen Rambow. 2025c. Llms can perform multi-dimensional analytic writing assessments: A case study of l2 graduate-level academic english writing. *arXiv preprint arXiv:2502.11368*.

Daniel Weitekamp, Momin N. Siddiqui, and Christopher J. MacLellan. 2025. Tutorgym: A testbed for evaluating ai agents as tutors and students. In *International Conference on Artificial Intelligence in Education*, pages 361–376. Springer.

Qihao Yang, Yu Yang, Sixu An, Tianyong Hao, and Guandong Xu. 2025. Llm-based collaborative agents with pedagogy-guided interaction modeling for timely instructive feedback generation in task-oriented group discussions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pages 9972–9980. International Joint Conferences on Artificial Intelligence.

Weizhe Yuan, Pengfei Liu, and Matthias Gallé. 2024. Llmcrit: Teaching large language models to use criteria. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7929–7960.

Mike Zhang, Amalie Pernille Dilling, Léon Gondelman, Niels Erik Ruan Lyngdorf, Euan D Lindsay, and Johannes Bjerva. 2025a. Sefl: Harnessing large language model agents to improve educational feedback systems. *arXiv e-prints*, pages arXiv–2502.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025b. Simulating classroom education with LLM-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*,

pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.

A Surveyed papers by domain

Table 1 presents an overview of all the surveyed papers in this work by their subject domain.

B Mechanisms Supporting Components

B.1 Student state estimation and simulation

Feedback can be guided by an estimate of what the learner *already knows*, leading to decision making over learner state at inference time, or what the learner is *likely to do next*, for feedback policies optimized against predicted responses to feedback.

The student state could be represented by the mastered (or lacking) knowledge bits, like in (Wang et al., 2025b) which uses knowledge tracing for state estimation from dialogue to guide tutor behaviour turn-by-turn. In math tutoring, predicted subsequent success is used as signals to improve tutor actions under sequential constraints (Scarlatos et al., 2025b). In writing, (Nair et al., 2024) uses a student simulator to produce revisions conditioned on candidate feedback and uses revision quality to induce preferences.

Beyond that, student simulators could also be used to facilitate tutoring data creation (Liu et al., 2024a) and system evaluation (Pal Chowdhury et al., 2024). On the other hand, classroom-scale simulation (SimClass) aims to capture multi-party instructional dynamics (Zhang et al., 2025b).

B.2 Scalable feedback aggregation

In classroom settings, qualified experts could be empowered to validate, edit, and reuse aggregated feedback to improve efficiency. The process of *triage, verify, and reuse* decides the common cases to address, the evidence to inspect, and the propagation strategy (Tang et al., 2025b,a), e.g., for large programming course cohorts (Ghoochani et al., 2025), and for live instruction (Wang et al., 2024b).

B.3 Self critique and refinement loops

Refinement loops are general purpose primitives that can be used with feedback systems to improve reliability at inference time under constraints.

Interactive critiquing uses tools, as in (Gou et al., 2024), to ground self-correction in external verification evidence. Agentic educational systems also leverage agents that critique and revise in stages as a plug-in refinement component within larger workflows (Zhang et al., 2025a).

On the other hand, multi-constraint requirements (e.g. pedagogical principles) can be decomposed by the model's inherent reasoning/in-context learning into explicit constraints to self-critique violations, and iteratively refine only the parts that fail (Ferraz et al., 2024; Yuan et al., 2024) in single- and multi-agent settings (Jiang et al., 2024; Guo et al., 2024). Iterative revise-and-review loops are also used to improve consistency and constructiveness in reflective writing feedback (Bhojan and Xin, 2025).

C Datasets and Evaluation

Evaluating feedback generation is more challenging than generic text because goodness depends on both **what intervention the system chooses** and **how that intervention is realized in language**. As a result, evaluation should distinguish decision action quality (whether the action is appropriate given the grounding signals) from feedback quality and impact (whether the final feedback is pedagogically valid and improves downstream outcomes).

Across the literature, authors operationalize "good feedback" via observable **evaluation signals**—unit test execution, rubric judgments, revision outcomes, learner-study measures, or instructor edits. When direct learning outcomes are expensive or unavailable, proxies are utilised such as predicted next-turn correctness, simulated revisions, or LLM-as-a-judge rubric satisfaction. Because decision making systems are frequently optimized against these signals, evaluation must clearly state: (i) what is optimized, (ii) what signal supports the claim, and (iii) how trustworthy the signal is (Stamper et al., 2024).

C.1 Datasets

Table 2 in the Appendix provides a summary of the datasets we surveyed, including artifact types, supervision signals, scale, and evaluation protocols.

C.2 Evaluation criteria

We observe four recurring objective families with unique evaluation requirements and threats to validity, but are not necessarily mutually exclusive:

(A) Feedback pedagogical validity: feedback that is a valid instructional intervention: content-correct, non-revealing, diagnostic, actionable, and appropriately calibrated in tone and confidence. This naturally pairs with rubric-based evaluations (Scarlatos et al., 2024; Sonkar et al., 2024;

Math	(Baral et al., 2024), (Chudziak and Kostka, 2025), (Dinucu-Jianu et al., 2025), (Liu et al., 2024b), (Liu et al., 2024a), (Jiang et al., 2024), (Puech et al., 2025), (Macina et al., 2025), (McNichols et al., 2024a), (McNichols et al., 2024b), (Pal Chowdhury et al., 2024), (Puech et al., 2025), (Reddig et al., 2025), (Scarlatos et al., 2024), (Scarlatos et al., 2025a), (Scarlatos et al., 2025b), (Tonga et al., 2025) (Wang et al., 2024a), (Wang et al., 2024b), (Weitekamp et al., 2025)
Programming	(Chae et al., 2024), (Ghoochani et al., 2025), (Qian et al., 2025b), (Qian et al., 2025a), (Heickal and Lan, 2024), (Lohr et al., 2025), (Ma et al., 2024), (Naik et al., 2024), (Scholz et al., 2025), (Tang et al., 2025b), (Tang et al., 2025a), (Wang et al., 2024a),
Writing	(Benharrak et al., 2024), (Chamoun et al., 2024), (Göldi et al., 2024), (Chen et al., 2025), (Bhojan and Xin, 2025), (Han et al., 2024), (Nair et al., 2024), (Jovic et al., 2025), (Naik et al., 2024), (Rashkin et al., 2025), (Wang et al., 2025c), (Yuan et al., 2024),
Other domains	(Guo et al., 2024), (Knievel et al., 2025), (Liu et al., 2024b), (Shankar et al., 2025), (Sonkar et al., 2024), (Zhang et al., 2025a), (Stamper et al., 2024), (Nazaretsky et al., 2025), (Seo et al., 2025), (Ferraz et al., 2024), (Yang et al., 2025), (Cohn et al., 2025), (Vanzo et al., 2025), (Zhang et al., 2025b)

Table 1: Surveyed papers by domain

Chen et al., 2025; Han et al., 2024; Liu et al., 2024a; Guo et al., 2024).

(B) Task-level and learning outcomes: feedback that leads to success on the task. In programming, this is often unit test execution success (Chae et al., 2024; Wang et al., 2025b), while in writing, revision quality improvement is usually used as a proxy (Nair et al., 2024), or instead through learner questionnaire (Han et al., 2024). In tutoring, it can be proxied by predicted next-step success under an explicit student model (Scarlatos et al., 2024). Some works measures learning outcome more directly using real participants with pre/post evaluations(Ma et al., 2024; Naik et al., 2024), providing stronger evidence of educational impact but are typically more expensive and harder to standardize across domains.

(C) Classroom efficiency: feedback insights that lead to improvements in throughput and reduce workload in classroom workflows (Tang et al., 2025b,a).

(D) Controllability and stability of pedagogical feedback: Whether a system can (i) follow a requested pedagogical policy (feedback type/criterion/scaffolding level) and (ii) do so consistently across runs. This is evaluated via type adherence in programming feedback (Lohr et al., 2025), coherent multi-step feedback ladders (Heickal and Lan, 2024), and repeated-run consistency for rubric-aligned reflective writing feedback (Bhojan and Xin, 2025). In writing assessment, multi-dimensional analytic scoring tests reliable dimension expression (Wang et al., 2025c); criterion-conditioned critique makes control explicit (Yuan et al., 2024); and agentic writing-feedback pipelines test grounded, targeted feedback under planning constraints (Chamoun et al., 2024). In tutoring, policy-level evaluation asks whether scaffolding adapts appropriately and preserves productive failure (Cohn et al., 2025; Puech et al., 2025; Dinucu-Jianu et al., 2025).

C.3 Evaluation signals

Decision making systems are constrained by what signals are observable in the setting. We group measurable evaluation signals into six classes.

Tool-grounded signals: Coding often uses unit tests as an outcome-grounded signal (Chae et al., 2024; Wang et al., 2025b). These are attractive because they are verifiable and scalable, but they primarily measure functional correctness, not pedagogical properties. Tool-grounded evaluation also extends beyond unit tests to tutor/testbed settings where correctness is defined over interaction trajectories: platforms report outcome-based measures (e.g., Success@N / task completion) (Chudziak and

Kostka, 2025), domain tutors check reasoning over structured artifacts (e.g., circuits) (Knievel et al., 2025), and testbeds/benchmarks score standardized tutor actions (e.g., next-step and bottom-out guidance) (Weitekamp et al., 2025; Macina et al., 2025).

Rubric-based signals: Often feedback is evaluated against rubrics that encode pedagogical constraints such as correctness, non-revelation, actionability, diagnostic value, and tone, like in MCQ diagnostic math feedback (Scarlatos et al., 2024) and Pedagogical Alignment (Sonkar et al., 2024). Rubric-based signals increasingly extend beyond "overall quality." They include feedback-type adherence in programming (Lohr et al., 2025), coherence of multi-step feedback ladders (Heickal and Lan, 2024), structured criteria for distractor/feedback alignment and pedagogical quality in math/hinting (McNichols et al., 2024a,b; Baral et al., 2024; Tonga et al., 2025), and multi-dimensional analytic writing rubrics that score distinct dimensions (Wang et al., 2025c). As LLM judges become common, studies evaluate judge consistency/accuracy (Seo et al., 2025) and develops validated feedback-evaluator suites (Qian et al., 2025a); preference-based optimization further requires transparent judge setup and calibration (Zhang et al., 2025a). Learner ratings may also diverge from expert judgments due to source credibility effects, motivating designs that disentangle perceived credibility from pedagogical quality (Nazaretsky et al., 2025). When rubrics utilise LLM judges, rater reliability becomes a key requirement. This should be calibrated through agreement with human labellers, consistency checks, or sensitivity analyses. Otherwise, scores could reflect judge bias or surface-form sensitivity rather than genuine improvements (Scarlatos et al., 2024; Tang et al., 2025b).

Revision improvement: Writing feedback is commonly evaluated by its downstream effect on revisions. Revision-based signals increasingly emphasize targeted feedback: scientific writing evaluates whether critiques are text-grounded and actionable for the intended revision (Chamoun et al., 2024); reflective writing measures rubric gains and cross-run consistency, treating iteration as part of the protocol (Bhojan and Xin, 2025); and higher-education deployments pair revision evidence with qualitative analysis of learner trust and adoption (Jovic et al., 2025). (Nair et al., 2024) uses a simulated reviser to generate and score revised drafts for preference training. This achievement-oriented proxy depends on how well the simulation reflects real learning.

Direct human evidence: Some works measure outcome with human studies. (Ma et al., 2024) reports pre/post-feedback outcomes for debugging skill in novices, while collaborative learning in (Naik et al., 2024) is measured by pre/post reports of interaction evidence. Writing-support and persona-conditioned systems often combine behavioural logs (e.g. feature use) with subjective outcomes (perceived usefulness) (Göldi et al., 2024; Benharrak et al., 2024). Other human studies include: large-course comparisons of peer vs. LLM feedback preferences/actionability (Ghoochani et al., 2025); school RCTs measuring learning/engagement gains from interactive homework tutoring (Vanzo et al., 2025); field studies of human–AI tutoring support in real-time instruction (Wang et al., 2024b); speech assessment pipelines combining ASR quality (e.g., WER) with agreement against human scoring (incl. subgroup analyses) (Shankar et al., 2025); and human judgments validating diagnosis/remediation quality in tutoring (Wang et al., 2024a; Reddig et al., 2025). These evaluations provide stronger evidence from the real world, but are typically small scale and harder to scale across settings.

Instructor-in-the-loop signals: Workflow systems must evaluate both the quality of finalized feedback and the validity of intermediate decision modules. e.g. issue detectors. (Tang et al., 2025b) reports component-level quality alongside human-coded feedback quality and workflow outcomes. (Tang et al., 2025a) evaluates instructor's feedback review efficiency and edit behaviour, in addition to aggregate measures like misconception coverage.

Benchmark and testbed signals: Standardized benchmarks and interactive testbeds evaluate tutor-like planning policies under shared protocols, e.g., MathTutorBench for open-ended pedagogical capabilities (Macina et al., 2025) and TutorGym for measuring action-level tutoring behaviors across domains (Weitekamp et al., 2025). Scalable evaluation can also rely on synthetic data and simulations, but this increases the need to report synthetic–real fidelity/coverage (Qian et al., 2025b) and to consider how simulation realism may affect conclusions (Zhang et al., 2025b).

C.4 Validity threats and failure modes

Evaluation conclusions depend on the reliability and validity of the underlying signals. We highlight four categories of validity threats.

Does the metric gauge the real educational construct? Learning science emphasizes evaluating instructional and learner outcomes directly when feasible, and otherwise being explicit about which proxy is optimized and why (Stamper et al., 2024), because passing unit tests might not imply conceptual understanding, rubric satisfaction does not guarantee learning outcome, and revision quality improvement may reflect surface-level edits. A recurring example is that liking or trusting feedback does not necessarily imply learning gains. Large-scale course deployments show that students may prefer LLM feedback and struggle to distinguish it from peer feedback (Ghoochani et al., 2025), while school-based evaluations highlight that stronger evidence comes from designs that include learning outcomes and engagement measures (e.g., pre/post tests and classroom RCTs), not only satisfaction (Vanzo et al., 2025). Similarly, human–AI tutoring support systems may change instructor practices and student experiences in ways that require measuring both pedagogical impact and workflow effects (Wang et al., 2024b).

Are judgments stable and calibrated? Rubric-based evaluation, especially with LLM judges, requires evidence of reliability: agreement with humans, consistency checks, and sensitivity analyses, to protect from judge idiosyncrasies (Scarlatos et al., 2024).LLM judging is not automatically trustworthy: studies evaluate judge stability/accuracy against humans (Seo et al., 2025) and develops validated feedback-evaluator suites trained on human-annotated data (Qian et al., 2025a). Preference "win rates" can be judge/prompt-sensitive, so reporting judge setup and sensitivity is essential for reproducibility (Zhang et al., 2025a); repeated sampling or aggregation can reduce variance when stability is part of the target construct (Bhojan and Xin, 2025).

Is the observed gain caused by better decision making? For systems that combine multiple components, ablations are needed to decide if gains come from better intermediate decisions or simply stronger generation. Component-level evaluation (e.g., verifier calibration, target-selection accuracy, propagation safety) helps isolate what got improved (Tang et al., 2025b,a). This is especially important for agentic pipelines with multiple interacting steps: targeted ablations can test whether planning improves groundedness/actionability beyond strong prompting (Chamoun et al., 2024), and pedagogy-steering/alignment methods need component-level checks to show gains reflect pedagogical behaviour rather than surface fluency (Puech et al., 2025; Dinucu-Jianu et al., 2025).

Does the gain generalize? The utilisation of simulated students and revisers supports scaling but may not match real classroom distributions. Proxies such as predicted next-turn correctness require validation and should not be interpreted as direct evidence of learning gains (Scarlatos et al., 2025b; Nair et al., 2024; Wang et al., 2025b). Generalization risk is amplified when evidence comes from synthetic or simulated settings: synthetic datasets require reporting fidelity/coverage against real student work (Qian et al., 2025b), and classroom simulations/testbeds depend on the realism of simulated students and tasks (Zhang et al., 2025b; Weitekamp et al., 2025; Macina et al., 2025). Domain-specific tutors add another generalization risk: performance can be brittle if evaluation tasks do not cover the range of authentic student misconceptions (Knievel et al., 2025).

C.5 What to report for comparability

Given diverse evaluation practices, we recommend that feedback generation systems report:

1. *Primary objective/constraints*: whether the goal is pedagogical validity, learning gains, workflow efficiency, or a combination thereof.

2. *Evaluation signals*: whether evaluation uses unit tests, rubrics/judges, revision outcomes, learner-study outcomes, or instructor edits.

3. *Reliability evidence*: report inter-rater agreement for human labellers, and for LLM judge report calibration against humans and stability checks (e.g., repeated-run agreement). Where possible, also report cross-model agreement to diagnose judge sensitivity

4. *Component-level validity*: For verifiers or propagation modules, report calibration/accuracy and safety checks.

5. *Intermediate decision evaluation*: When the system makes explicit intermediate decisions, report accuracy or calibration against reference.

6. *Decision-making design*: specify the intermediate decisions the system makes (e.g., feedback type, scaffold level, criterion/target selection), how these decisions are represented (labels, templates, plans), and how they are enforced or checked (constraints, verifiers, critique loops). If controllability is a goal, report adherence metrics (e.g., type-following rate, ladder coherence, criterion coverage) alongside general quality scores.

7. *Downstream outcome*: e.g. unit-test success, revision improvement in writing and throughput gains for workflows. When feasible, pair subjective ratings (e.g., perceived usefulness) with learning/engagement outcomes using stronger study designs (e.g., pre/post tests, randomized trials, and classroom field evaluations)

8. *Ablations*: to separate the generation effects from the contributions of the decision making mechanism.

9. *Limitations*: include validation evidence and scope claims when using simulated students/revisers or proxy models. If training/evaluation uses synthetic or simulated data, report fidelity/coverage analyses and discuss how conclusions might shift under real student distributions.

C.6 Discussion: Evaluation - separating action quality, feedback quality, and impact

A recurring lesson from the surveyed work is that feedback evaluation must distinguish *(i) decision/action quality* from *(ii) linguistic quality* and *(iii) downstream impact* (§C). Tool-grounded metrics (unit tests) can scale and are verifiable, but do not necessarily measure learning or pedagogical validity (Chae et al., 2024). Rubric-based judgments better capture pedagogical constraints (Scarlatos et al., 2024; Sonkar et al., 2024), but introduce rater variance and judge sensitivity, especially when LLMs are used as evaluators (Seo et al., 2025). Revision improvement provides an attractive proxy for writing, yet may conflate surface-level edits with genuine learning-relevant progress (Nair et al., 2024; Jovic et al., 2025). The strongest evidence still comes from human studies and deployments that measure learning and engagement outcomes (Ma et al., 2024; Vanzo et al., 2025), but these are expensive and difficult to standardize.

One implication is that evaluation should increasingly be reported as a *bundle* rather than a single number: a primary objective, its signal(s), reliability evidence for those signals, and ablations that isolate whether improvements are attributable to better intermediate decisions or simply stronger generation. Where "impact" is claimed, stronger designs (e.g., pre/post measures or randomized comparisons) are needed to separate learning gains from mere revision compliance (Ma et al., 2024; Vanzo et al., 2025). Finally, the rise of evaluator-centric pipelines (e.g., gatekeeping "Dean"-style evaluators) suggests that evaluator auditing and calibration should be treated as a first-class requirement, not an optional appendix (Qian et al., 2025a; Seo et al., 2025).

Paper	Task / Domain	Input	Output	Training signal	Size
(Scarlatos et al., 2025b)	Dialogue tutoring (math)	Tutoring dialogue context + student turn (MathDial)	Tutor response	DPO-style preferences favoring tutor turns that elicit correct next student responses; plus GPT-4o distillation to strengthen the base tutor	MathDial: 1,852 (28,125 tutor turns) Test: 361 (5,347 tutor turns) Distill: 6,250; DPO: 11,250 pairs
(Chae et al., 2024)	Code feedback	Problem description + wrong code (with tests in the environment)	NL feedback on how to fix code	Pairwise correct/incorrect feedback labels + unit-test-driven reward (COFFEEEVAL) used as supervision for improving feedback	COFFEE: 44,782; 742 problem sets COFFEE-TEST: 180 Avg 35.5 tests/problem
(Wang et al., 2025b)	Coding tutor + verification	Coding task context + dialogue history/turn	Tutor turn + turn-level verifier judgments	Supervised turn-by-turn verifiers used to assess tutoring turns during interactive coding tutoring	EvoCodeBench: 100 tasks
(Nair et al., 2024)	Writing feedback	Student draft (and prompt/context)	Feedback comments	"Closed-loop" learning signal based on LM-simulated revisions conditioned on feedback	363 datapoints (291/72) 873 feedback items
(Sonkar et al., 2024)	Pedagogically scaffolded tutoring	Conversation context (biology tutoring dialogues)	Structured tutor response (incl. pedagogical fields)	Preference learning (DPO/IPO/KTO) on synthetic preference pairs derived from structured tutoring signals (CLASS-style fields)	1,738 conversations total 600 SFT (4,942 QA) 600 LHP (4,921 QA) 450 test (3,701 QA)
(Rashkin et al., 2025)	Story writing feedback (eval dataset)	Story draft (often intentionally corrupted)	Feedback to improve story writing	Human ratings + automatic metrics (dataset is primarily for evaluation/analysis of feedback quality)	Test: 1,300 stories StoryFeedback: ∼84k pairs + human-rated subset
(Liu et al., 2024a)	Socratic math tutoring	Math problem + multi-round tutoring dialogue (simulated students)	Socratic tutor turns (questions/hints)	SFT on SocraTeach generated via "Dean-Teacher-Student" pipeline; plus augmentation for specific teaching abilities	SocraTeach: 35k multi-round (=208k single-round equiv.) + 22k extra single-round

Table 2: Summary of publicly available datasets