Beyond Generation: Evaluating Large Language Models as Pedagogical Generators and Reviewers for Primary Lesson Plans in a Multilingual South Asian Context

Muhammad Raees Shujaan Azhar
Taleemabad, Islamabad, Pakistan
shujaan.azhar@taleemabad.com

***Abstract*—Quality lesson planning is a critical bottleneck in scaling effective primary education across Pakistan, where teachers serve diverse linguistic and curricular contexts spanning Urdu and English medium instruction. While large language models (LLMs) have shown promise in educational content generation, their capacity to *evaluate* lesson plans against domain-specific pedagogical standards remains largely unexplored, particularly in low-resource South Asian settings. In this paper, we make three contributions. First, we introduce a multi-dimensional evaluation rubric for primary lesson plans (Grades 1–5, subjects: English, Science, Mathematics, and Urdu) grounded in Bloom's Taxonomy, Assessment for Learning (AfL) strategies, and Pakistani national curriculum standards. Second, we present a benchmark study evaluating leading LLMs both as *generators* and as *reviewers* of lesson plans, measuring generator quality via our rubric and reviewer reliability via agreement with a strong reference reviewer that is itself validated against expert human judgments. Third, we describe a multi-agent lesson plan generation system that combines a generator agent and a rubric-guided reviewer agent in an iterative refinement loop, selecting agent models informed by our benchmark findings. Experimental results show that the strongest of seven open-weight models reaches 83% of the rubric maximum as a generator, and that all evaluated models score markedly higher on structural and factual criteria than on instructional-flow and deeper-learning criteria. A single reviewer-guided refinement pass lifts every model above 88% of the maximum. As *reviewers*, however, the same open models diverge from the frontier reference, scoring systematically more leniently, so automated generation reliably produces well-formed plans while dependable pedagogical review still favours a stronger model. Our rubric and benchmark dataset will be released upon publication to support reproducible research in AI-assisted curriculum development for underrepresented educational contexts.**

***Index Terms*—large language models, lesson plan generation, pedagogical evaluation, LLM benchmarking, Bloom's Taxonomy, Assessment for Learning, multi-agent systems, educational AI, Pakistan, primary education**

I. Introduction

Effective lesson planning is foundational to teaching quality, yet producing pedagogically sound lesson plans consistently and at scale remains a significant challenge in Pakistan's primary education system [1]. With tens of millions of primary school students across provinces including Punjab, Balochistan, and the Islamabad Capital Territory (ICT), and instruction delivered across Urdu and English mediums, the variability in lesson plan quality directly affects learning outcomes.

Recent advances in large language models (LLMs) have opened new possibilities for automating educational content creation [2]. However, generation alone is insufficient: a lesson plan must not only cover content but also target appropriate cognitive levels per Bloom's Taxonomy [3], embed formative assessment strategies consistent with Assessment for Learning (AfL) [4], and remain sensitive to the local curriculum framework and linguistic context.

Existing benchmarks for LLMs in educational contexts, such as those introduced by AI for Education [17] and lesson-planning assistants such as LessonPlanner [5], evaluate general pedagogical reasoning or assist drafting but do not address lesson plan generation or review in multilingual, curriculum-specific settings. No published work, to our knowledge, has systematically evaluated LLMs as *reviewers* of lesson plans against a grounded rubric, nor has any benchmark addressed the Pakistani primary curriculum context.

This paper addresses these gaps through three contributions:

1) **A Pedagogical Evaluation Rubric:** A validated, multi-dimensional rubric for assessing lesson plans for Grades 1–5 across English, Science, Mathematics, and Urdu, aligned to Bloom's Taxonomy, AfL strategies, and Pakistani curriculum standards.
2) **An LLM Benchmark:** A systematic evaluation of seven open-weight LLMs as both lesson plan generators and pedagogical reviewers, revealing key capability differences and cost-performance trade-offs relevant to deployment at scale.
3) **A Multi-Agent Generation System:** An agentic pipeline in which a generator agent produces lesson plans subsequently critiqued by a reviewer agent using the rubric, with iterative refinement until a quality threshold is reached. Model selection for each agent role is informed directly by benchmark findings.

The remainder of this paper is structured as follows. Section II reviews related work. Section III introduces our evaluation rubric. Section IV presents the LLM benchmark methodology and results. Section V describes the multi-agent

system. Section VI reports end-to-end system evaluation. Section VII discusses implications and limitations. Section VIII concludes.

II. Related Work

A. LLMs for Educational Content Generation

Commercial platforms such as MagicSchool AI, Khanmigo, Curipod, and Eduaide.ai have scaled AI-assisted lesson planning to millions of teachers [6], demonstrating clear demand for automated instructional content. However, none of these systems publishes a rubric-based evaluation of its lesson plan outputs; quality evidence is limited to teacher satisfaction surveys or perception studies. On the research side, Hu *et al.* [7] used GPT-4 to generate mathematics teaching plans for Chinese high schools and evaluated them with a Pedagogical Content Knowledge rubric, demonstrating strong output quality within that curriculum but providing no multilingual evaluation. Zheng *et al.* [8] introduced a three-stage pipeline combining retrieval-augmented generation with LLM self-critique and refinement for elementary mathematics, and Zheng *et al.* [9] extended this with a knowledge-base fine-tuning framework to improve structural coherence. The most contextually relevant deployed system is Shiksha Copilot [10], a Microsoft Research–Cornell collaboration serving 1,043 teachers in low-resource schools in Karnataka, India, which uses a human-in-the-loop RAG architecture and curator review process for lesson plan quality. While Shiksha Copilot shares the low-resource South Asian deployment context, it relies on human curator judgment rather than a formal rubric, and does not address the Pakistani curriculum, Urdu medium instruction, or systematic LLM benchmarking.

B. LLM-as-Judge and Automated Evaluation

Zheng *et al.* [12] established the LLM-as-judge paradigm with MT-Bench, showing GPT-4 achieves over 80% agreement with human raters on open-ended dialogue. Madaan *et al.* [13] demonstrated that a single LLM can improve its outputs iteratively through a generate-feedback-refine loop without weight updates, while Shinn *et al.* [14] extended this with episodic verbal memory. Hashemi *et al.* [15] addressed single-model evaluator bias through LLM-Rubric, training a calibration network over multiple LLM judges for multidimensional text evaluation and improving agreement with human assessors. The present paper advances these foundations into domain-specific pedagogical evaluation: our rubric operationalises Bloom's Taxonomy and AfL criteria for primary lesson plans, and we benchmark seven open-weight models as both generators and reviewers against a strong fixed reference reviewer that is itself validated against expert human verdicts, in the LLM-as-judge tradition [32], [33]. A committee-based reference such as a democratic Language Model Council [31] is a natural extension that we leave to future work.

C. Benchmarks for Pedagogical AI

Bai *et al.* [16] introduce EduBench with over 4,000 scenarios across nine educational contexts and twelve evaluation dimensions, and Lelièvre *et al.* [17] benchmark 97 models on professional teacher-examination questions covering declarative pedagogical knowledge. These works evaluate what models *know* about pedagogy; neither assesses whether models can *generate or score* a lesson plan against a structured rubric. Arif *et al.* [18] provide the most relevant South Asian benchmark, evaluating frontier LLMs across 14 Urdu NLP tasks and finding substantial performance gaps relative to English, a finding that motivates our Pakistan-specific, Urdu-inclusive evaluation framework. To our knowledge, no published benchmark addresses lesson-plan generation or review for Pakistani primary curricula in either English or Urdu medium.

D. Bloom's Taxonomy and AfL in Automated Systems

Scaria *et al.* [20] systematically evaluate five LLMs on generating questions at all six Bloom's levels, finding strong performance at lower cognitive levels (Remember, Understand) but high variance at Evaluate and Create, a pattern that motivates our rubric's Bloom's-alignment sub-criterion within C1. Li *et al.* [19] demonstrate that BERT-based classifiers can automatically assign learning objectives to Bloom's levels with F1 up to 0.95, establishing the NLP baseline for automated cognitive-alignment checks. Memarian and Doleck [21] review 35 studies on AI-assisted Assessment for Learning, concluding that while AI can generate formative feedback, no existing system embeds AfL principles *structurally* into a content generation workflow: the gap our rubric-grounded review stage fills. Du *et al.* [22] show that multi-agent debate reduces hallucinations and improves reasoning through iterative inter-agent critique, a mechanism that motivates the committee-based reference reviewer we outline as future work (Section IV-C).

E. Multi-Agent Systems in Education

The iterative generator–reviewer pattern has antecedents in general-purpose agent research [13], [14] and is increasingly applied to educational tasks. Zhang *et al.* [23] field a three-agent system (Evaluator, Optimiser, Question Analyst) for personalised instructional design, validated on Western mathematics benchmarks. Chen *et al.* [11] pair a Writer Agent with an Evaluator Agent in an iterative loop for lesson plan generation, demonstrating quality improvement over single-pass baselines. Both systems confirm the effectiveness of adversarial agent collaboration for instructional content, but neither incorporates a formally specified pedagogical rubric, curriculum grounding, or multilingual scope. The present work ties the reviewer agent to a validated, subject-specific rubric and selects agent models through a systematic benchmark: two design decisions absent from prior work.

III. Pedagogical Evaluation Rubric

A. Design Rationale

1) Motivation and Objectives: The rubric was developed with four purposes in mind. First, it closes the feedback loop between reviewer and author: lesson plan writers receive specific, actionable guidance tied to individual criteria rather than general comments. Second, it provides a common standard

for quality control, so that lesson plans are assessed against the same criteria regardless of subject, grade, or textbook. Third, it generates diagnostic feedback at the sub-criterion level, making it possible to distinguish a lesson plan that is structurally complete but pedagogically weak from one that fails on both counts. Fourth, by making the criteria explicit and public, the rubric supports professional development: authors can identify recurring weaknesses and build targeted competence over time.

2) Internal Gap Analysis: A systematic gap analysis of the prior Taleemabad review checklist (which evaluated lesson plans section-by-section against a binary checklist) identified five critical shortcomings that motivated a redesign.

(1) Fragmented evaluation approach. The checklist was compartmentalised by lesson plan section, preventing a unified assessment of overall instructional design and flow. This siloed view limited the reviewer's ability to identify cross-cutting strengths or weaknesses in pedagogical coherence.

(2) Absence of criterion prioritisation. All checklist items were treated as equally important. No criterion was dedicated to evaluating the instructional design of the lesson as a whole, which is foundational to pedagogical integrity.

(3) Ineffective scoring system. The binary present/absent scoring lacked diagnostic granularity. The absence of a dynamic rating scale reduced the tool's ability to distinguish minimally compliant lesson plans from high-quality ones, undermining the accuracy of feedback.

(4) Insufficient focus on practicality and time management. Practical feasibility (the degree to which a lesson plan can be delivered within available class time and with locally available resources) was underrepresented, despite being critical for real-world classroom delivery in low-resource Pakistani schools.

(5) Omission of key pedagogical frameworks. The checklist did not assess alignment to evidence-based pedagogical approaches including the Gradual Release of Responsibility (GRR) model [24], the Concrete–Pictorial–Abstract (CPA) framework for mathematics [25], inquiry-based learning for science, modelling and guided practice for languages, and a learner-centred approach emphasising student agency and active participation.

3) Classroom Observation Data: Classroom observation data collected across Taleemabad-partner schools further informed the rubric design. Observers found that teachers struggled most when lesson content was not broken into subject-appropriate steps and when practice activities ran over the available class time. They also noted a consistent gap in subject-specific technique: teachers needed concrete guidance on how to teach, not just what to teach. These observations pointed to distinct approaches per subject (the CPA progression for Mathematics, inquiry-based activities for Science, and modelling with guided practice for English and Urdu), which are reflected in Criteria C2 (Instructional Flow), C3 (Practicality), and C8 (Subject-Specific Review).

4) Contextual Constraints of Pakistani Primary Classrooms: Established Western lesson plan rubrics (such as the Danielson Framework for Teaching or Marzano's Instructional Model) assume conditions that do not hold in the majority of Pakistani primary schools: certified teachers with subject-specialist training, classrooms equipped with projectors and printed materials, and class sizes small enough to permit individualised instruction. Applying such instruments without modification would penalise lesson plans that are well-designed *for their actual delivery context* while rewarding features that are pedagogically irrelevant or practically undeliverable in Pakistan.

Field visits to Taleemabad-partner schools revealed conditions that directly shaped rubric design decisions. Teacher qualifications vary widely; many classroom practitioners hold only a basic teaching certificate and rely heavily on the lesson plan as the primary instructional scaffold. A rubric criterion that rewards vague, high-level guidance (acceptable for an experienced teacher) is counterproductive here: the lesson plan must be explicit enough for a low-skill practitioner to deliver without external support. This requirement is embedded in C4 (teacher-facing language) and the explicitness sub-criteria of C2.

Class sizes present a further constraint. Government schools across Punjab and the ICT regularly accommodate 60–80 students per classroom, and field visits recorded classes of close to 100 pupils spanning Grades 1–5 in a single room. At this scale, individualised instruction is not feasible, pair and group work requires explicit structural management, and any activity that depends on per-pupil materials is undeliverable. These realities are reflected in C3 (Practicality and Teacher Support), which penalises resource dependencies and over-scripting, and in C5 (Deeper Learning and Engagement), which requires structured peer interaction with explicit teacher instructions rather than open-ended collaboration.

The bilingual instructional context adds a further layer of complexity absent from Western frameworks. Pakistan's primary schools operate across Urdu and English medium curricula, and code-switching between languages within a single lesson is common practice [1]. A generic rubric treats language as a presentation variable; this rubric treats it as a pedagogical variable, motivating the subject- and language-specific sub-criteria in C8 and the separate reviewer model assignment for Urdu.

Finally, the fixed-textbook structure of Pakistani national curricula means that lesson quality cannot be evaluated independently of the prescribed content. Unlike curricula that give teachers flexibility to select source materials, Pakistan's ICT and Punjab curricula mandate specific textbook pages per lesson [27]. A lesson plan that ignores or misrepresents the textbook content (even if it scores well on generic pedagogical criteria) fails in its primary function. This is why textbook fidelity is enforced as a grounding constraint at the generation stage and assessed as a dedicated sub-criterion within C1.

5) External Frameworks: Three external frameworks provided additional grounding.

EdTech Hub. An independent evaluation of Taleemabad lesson plans conducted by EdTech Hub [26] employed eleven

criteria: Curriculum Alignment (Backward Design Framework); Instructional Flow (Gradual Release of Responsibility); Clarity; Conceptual Depth; Logical Sequencing; Factual Accuracy; Adaptability; Differentiation and Scaffolding; Teacher Usability; Assessment aligned to Bloom's Taxonomy [3]; and Inclusivity and Accessibility (encompassing rural versus urban settings and multilingual instructional needs). This framework confirmed the centrality of backward design and GRR to lesson quality evaluation and highlighted inclusivity as a criterion absent from Taleemabad's prior checklist.

Prevail Numeracy Review Tool. Prevail's review instrument, used to evaluate Taleemabad's 30-day numeracy programme, contributed two key design principles: mapping each lesson plan against specific Student Learning Outcomes (SLOs) to anchor evaluation to intended learning rather than surface form; and employing a four-point anchored scale applied holistically per criterion, which reduces rater bias while retaining the diagnostic resolution needed to distinguish partial from full criterion satisfaction.

AI for Education. The AI for Education initiative's work on pedagogical evaluation [17] informed the operationalisation of Assessment for Learning (AfL) strategies [4] (including learning objective specification, success criteria, formative check placement, and feedback loops) within automated review pipelines, directly shaping the Check-for-Understanding sub-criteria in C2 and the closing assessment sub-criterion in C5.

6) Iterative Development and Synthesis: The rubric was developed across three iterations. **Version 1** established five core criteria (Curriculum Alignment, Instructional Flow, Practicality, Accessibility, Subject-Specific Review) with a four-point rating scale and differential scoring to identify underperforming areas. **Version 2** introduced two additional criteria for Early Years lesson plans and Financial Literacy (FICO) content; added level-specific indicators and concrete examples for each rating level to reduce rater ambiguity; and formalised a percentage score threshold: plans scoring below 70% are flagged for mandatory revision and resubmission. **Version 3** (the present rubric) unified the lessons of all prior iterations into a nine-criterion instrument by: adding Structural Completeness as a gate check (C0); expanding Instructional Flow to twelve sub-criteria to operationalise GRR alongside prior-knowledge activation and active retrieval; disaggregating cultural relevance, factual accuracy, and deeper learning into dedicated criteria; mapping every sub-criterion to the Taleemabad Standard Framework codes it enforces; and automating scoring via an LLM reviewer agent, which required that each sub-criterion be specified with sufficient precision to support consistent machine scoring.

Four priorities ran through all three versions. Pedagogical approach was the first: GRR and learner-centred instruction needed to be embedded consistently, not left implicit. Instructional design was the second: clear SLOs, sequenced activities, and formative assessment checkpoints had to be present and assessable. Practicality was the third: a lesson plan is only useful if it can be delivered in a real Pakistani classroom, which means realistic time allocations and no dependency on equipment teachers do not have. Consistency was the fourth: the same criteria had to apply across all subjects, grades, and textbooks to allow meaningful comparison between lesson plans.

B. Rubric Dimensions

The rubric comprises eight common criteria (C0–C7) applied uniformly across all subjects, plus one subject-specific criterion (C8), totalling 44 common sub-criteria and a further 3–4 per subject. Each sub-criterion is scored 1–4 with anchored descriptors and is additionally tagged with the Taleemabad Standard Framework code(s) it enforces (e.g., Gradual Release of Responsibility, Assessment Rigour, Active Retrieval), enabling diagnostic roll-ups at the level of individual pedagogical standards. Table I summarises all criteria, sub-criterion counts, and maximum point allocations.

TABLE I
LESSON PLAN EVALUATION RUBRIC: CRITERIA AND MAXIMUM SCORES

ID	Criterion	# Sub	Max Pts
C0	Structural Completeness	6	24
C1	Curriculum Alignment	6	24
C2	Instructional Flow	12	48
C3	Practicality & Teacher Support	5	20
C4	Accessibility & Differentiation	4	16
C5	Deeper Learning & Engagement	4	16
C6	Cultural Relevance	4	16
C7	Factual Accuracy & Bias	3	12
	Common total (C0–C7)	44	**176**
C8	Subject-Specific (Eng / Other)	3/4	12/16
	Grand total (Eng)		**188**
	Grand total (Maths/Sci/Urdu)		**192**

1) C0: Structural Completeness: This gate criterion comprises six checks. Five verify that the required lesson plan sections are present and contain substantive, teacher-usable content: Student Learning Outcome (SLO), Hook, Explanation, Practice, and Conclusion; a section consisting only of a heading or placeholder line scores 1. The sixth check, *Single-Lesson Scope*, verifies that the plan delivers a single SLO within one 40-minute class. Unit- or chapter-plan red flags (a pacing table or weekly schedule, numbered "Lesson 1, Lesson 2, . . . " headers, multiple SLOs, or a stated duration of several hours) constitute a fundamental structural failure. This criterion is evaluated before quality scoring; a plan missing one or more sections receives automatic flags for the affected dimensions downstream.

2) C1: Curriculum Alignment: Six sub-criteria assess alignment to the target curriculum: (i) a clear, single-focus SLO with an observable action verb and an explicit Single National Curriculum (SNC) standard code; (ii) measurability of the SLO; (iii) scaffolded activity sequencing toward SLO mastery; (iv) fidelity to textbook content across all sections;

(v) alignment between SLO content and the assessment task (not merely the assessment type), together with phase coherence, so that every phase (Hook, I Do, We Do, You Do, exit ticket) targets the same SLO with no orphan activities; and (vi) consistency among the stated Bloom's level, the action verb, and the cognitive demand of the practice task. SLO inflation (bundling multiple skills into one objective) is penalised, capping the sub-criterion score at 2.

3) C2: Instructional Flow: The highest-weighted criterion (max 48 pts) covers twelve sub-criteria: relevance of the opening hook to the SLO; specificity of teacher instructions in the Explanation phase, including evidence-based scripted feedback for common student errors; teacher modelling with explicit think-aloud reasoning before student practice; smooth logical transitions between sections; provision of practice opportunities; placement and spacing of open-ended Check-for-Understanding (CFU) questions after each instructional phase; grade- and SLO-appropriate strategies; presence of higher-order thinking activities beyond recall; effectiveness of visual aids at the specific instructional moment of use; the Gradual Release of Responsibility (GRR) [24] structure (I Do $\rightarrow$ We Do $\rightarrow$ You Do) as an explicit standalone sub-criterion; prior-knowledge activation that surfaces and connects to what students already know; and active retrieval, a peer-mediated warm-up linked to a prior lesson's SLO. Treating GRR as a discrete sub-criterion rather than an implicit expectation allows the reviewer to distinguish lessons that include all three phases with appropriate scaffolding from those that collapse We Do and You Do or omit guided practice entirely.

4) C3: Practicality and Teacher Support: Evaluates practical classroom feasibility: realistic time allocation per section, totalling 35–45 minutes with enforced phase balance (I Do $\leq$25%, We Do and You Do $\geq$30% each); adequate time balance between Explanation and Practice; presence of teacher support materials (answer keys, misconception guidance, subject knowledge notes); use of low-cost, locally available resources with no dependency on projectors or internet connectivity; and sufficient professional discretion: scripts that are word-for-word throughout are penalised.

5) C4: Accessibility and Differentiation: Four sub-criteria assess inclusion: suggestions addressing *both* struggling and advanced learners (neither group alone scores above 2); named, concrete differentiation strategies (e.g., sentence starters, graphic organisers) rather than vague directives; age-appropriate student-facing language; and directive, empowering teacher-facing language accessible to low-skill practitioners.

6) C5: Deeper Learning and Student Engagement: Assesses active student participation (students produce output, not merely listen), structured peer interaction with explicit instructions, real-world application in an authentic Pakistani context, and a closing exit ticket that tests the SLO in a *new* context, has the teacher read success criteria aloud, and prompts students to self-predict their performance before submitting.

7) C6: Cultural Relevance and Representation: Four sub-criteria evaluate use of Pakistani names, settings, and scenarios; deliberate (not incidental) gender and background diversity; opportunities for students to connect content to their own cultural lives; and absence of stereotyping or culturally insensitive material.

8) C7: Factual Accuracy and Bias: Checks internal consistency and plausible textbook references (no hallucinated page numbers or fabricated exercise descriptions), academic content correctness (facts, formulas, grammar, scientific concepts), and absence of one-sided claims or harmful generalisations.

9) C8: Subject-Specific Review: A subject-specific criterion is appended to the common eight:

English (3 sub-criteria, max 12 pts): prioritised skill focus (reading, listening, speaking, or writing); use of named teaching strategies (e.g., phonics, shared reading); evidence of textbook material integration.

Mathematics (4 sub-criteria, max 16 pts): conceptual knowledge alongside procedural fluency; step-by-step computational guidance; real-world application with meaningful context; mathematical language and teacher–student dialogue.

Science (4 sub-criteria, max 16 pts): scientific concept clarity; hands-on or observation-based activities; real-life connections; scientific inquiry and CFU questions.

Urdu (4 sub-criteria, max 16 pts): The four Urdu sub-criteria are grounded in research on literacy acquisition in Pakistani primary schools and in the National Professional Standards for Teachers in Pakistan [27].

Prioritised Urdu skill focus. Effective Urdu lesson plans must target one primary literacy skill rather than addressing multiple skills superficially. The Pakistan Reading Project (PRP) [28], an empirically evaluated national literacy programme, identifies seven discrete literacy components: Print Concept, Phonemic Awareness, Phonics, Vocabulary, Fluency, Comprehension, and Writing, and demonstrates that lessons explicitly prioritising a single component (e.g., Phonics or Fluency) produce markedly stronger student literacy outcomes than lessons that treat these skills as undifferentiated background content. The PRP findings therefore motivate a sub-criterion that rewards deliberate, single-skill prioritisation as a modern, evidence-based instructional technique.

Named Urdu-specific teaching strategies. The inclusion of specific, named teaching strategies is essential for instructional clarity and quality assurance. Research by Hussain, Hussain, and Kousar [28] on the PRP demonstrates that schools utilising explicit, named strategies (such as Decoding and Blending and the Jalongo Method for phonetic spelling) achieve above-average student literacy outcomes compared to schools employing generalised methods. In addition, naming the strategy constitutes evidence of deliberate pedagogical alignment: it demonstrates that the teacher has intentionally matched the teaching method to the specific learning outcome [29]. Without such specification, the lesson plan fails to meet Pakistan's National Professional Standards for Teachers, which require demonstrable evidence of modern, active learning techniques.

Evidence of textbook material usage. Effective Urdu pedagogy requires what may be termed *textual fidelity*: bridging the gap between the prescribed textbook content (the intended curriculum) and the instructional materials and activities through which that content is made accessible. The PRP research model [28] emphasises that while the textbook defines the intended curriculum, a high-quality lesson plan must demonstrate the use of additional teaching materials (charts, flashcards, and word source activities) to render the text analytically and practically accessible to students. This aligns with Webb's Depth of Knowledge framework [30], which distinguishes between merely reading textbook material (recall) and analysing and applying it through structured classroom activities (deeper cognition).

Multiple strategy variations. A single instructional strategy is insufficient to serve the range of learner needs within a primary classroom. This sub-criterion rewards lesson plans that provide alternative approaches or variations, enabling teachers to adapt delivery without abandoning the pedagogical intent of the lesson.

C. Scoring Schema and Quality Thresholds

Each sub-criterion is rated on a four-point scale: 1 = absent; 2 = present but does not meet the criterion; 3 = present but only partially meets the criterion; 4 = fully present and meets the criterion. Criterion scores aggregate their sub-criteria linearly. The overall score is expressed as a percentage of the subject-specific maximum (188 pts for English; 192 pts for Mathematics, Science, and Urdu). Checks whose required context is unavailable (for example, the textbook pages for a content-alignment check, or the previous lesson's SLO for the retrieval check) are marked *not assessable* and excluded from the denominator rather than penalised, so the percentage reflects only the criteria that can be judged from the available evidence. Plans scoring below 70% are flagged for revision; plans at 70–84% are classified as *Acceptable*; plans at 85% and above are classified as *Strong*. These thresholds correspond to the team's operational quality bar in the authoring workflow; their empirical calibration against expert-scored plans is reported in the rubric validation below.

D. Rubric Validation

The rubric was validated in two ways. First, its criteria and sub-checks were authored and iterated with Taleemabad curriculum specialists, whose framework defines the pedagogical standards each check encodes (Section III-B). Second, to confirm that the automated reviewer applies the rubric as an expert would, curriculum specialists scored fully reviewed lesson plans on the same four-point scale, assigning one score per sub-check. Over eight such plans, spanning English (five), Mathematics (one), and Urdu (two), the specialist and reviewer criterion scores agree closely: a mean absolute difference of 1.7 percentage points and a Pearson correlation of 0.97 across the 72 criterion pairs, with the specialists scoring marginally stricter than the reviewer (mean signed difference -1.1 points). Agreement is tightest for English (mean absolute difference 1.1 points) and remains close for Urdu (2.6) and Mathematics (2.9). An earlier round using a coarser agree/disagree protocol over three further plans, including a Science plan, was consistent: specialists endorsed 135 of 139 automated sub-check judgments (97%).

These results support using the reviewer as a stand-in for expert judgment, which licenses its role as the fixed scorer in Sections IV-B and IV-C. They remain preliminary. The independently scored set is modest (eight plans, with a single Mathematics plan and no independently scored Science plan yet) and single-annotator; the plans are production outputs of the deployed generator; and scores were entered in an interface that also displays the automated review, so the observed agreement may be partly anchored to it. A larger, subject-balanced, and ideally blind set is accruing. Reporting full inter-rater reliability (Cohen's κ, Krippendorff's α) additionally requires at least two independent scorings per plan, which the current single-annotation tooling does not retain.

IV. Benchmarking LLMs for Lesson Plan Generation and Review

A. Benchmark Design

1) Test Set Construction: The benchmark test set comprises 18 lesson-plan generation tasks drawn from real production requests logged by the deployed system, so that each task reflects an authentic teacher-facing input rather than a synthetic prompt. Tasks span the study's four subjects on the ICT curriculum: five English, five Urdu, and five Mathematics tasks, one per grade from Grade 1 to Grade 5, together with three Science tasks. The English, Urdu, and Mathematics tasks each target a specific skill-based lesson type: reading, grammar, comprehension, and creative writing for the languages, and concrete, pictorial-and-abstract, and word-problem lessons for Mathematics. These cover the generator's full set of skill prompts. Science has no skill-specialised prompt in the pipeline and therefore uses the standard generation path; production Science coverage exists only for Grades 4–5, so the three Science tasks are drawn from that band. Every task is grounded in specific textbook pages retrieved from the OCR-processed curriculum database, and page ranges consisting of non-content front matter were excluded.

2) Ground Truth and Gold Standard: Expert human scoring does not scale to the full set of generated plans (seven models across every task). Generator quality is therefore assessed automatically by the rubric-guided reviewer described in Section IV-C, which assigns each plan a score on every sub-criterion of the v3 rubric. The reliability of this automated scorer is not assumed: its agreement with expert educators is measured separately against a set of human-annotated lesson plans (Section IV-C). All plans in the generator benchmark are scored by a single reference reviewer model held fixed across the study, so that differences in score reflect differences in generation quality rather than in the evaluator.

3) Models Evaluated: Seven open-weight models accessed through a unified model gateway are evaluated as generators: DeepSeek R1, DeepSeek V3.2, Qwen3 235B (Thinking),

TABLE II

GENERATOR EVALUATION: MEAN RUBRIC SCORE (% OF ASSESSABLE MAXIMUM) PER CRITERION, AVERAGED OVER 18 CANONICAL TASKS PER MODEL (KIMI K2: 17), EACH LESSON PLAN SCORED BY THE GEMINI 3.1 PRO REVIEWER AGAINST THE V3 RUBRIC. CRITERIA: C0 STRUCTURAL COMPLETENESS, C1 CURRICULUM ALIGNMENT, C2 INSTRUCTIONAL FLOW, C3 PRACTICALITY, C4 ACCESSIBILITY, C5 DEEPER LEARNING, C6 CULTURAL RELEVANCE, C7 FACTUAL ACCURACY, C8 SUBJECT-SPECIFIC. ROWS SORTED BY OVERALL SCORE.

Model	C0	C1	C2	C3	C4	C5	C6	C7	C8	Overall
DeepSeek V3.2	92.1	81.5	77.5	76.9	89.2	76.0	84.0	94.9	89.9	**83.3**
Kimi K2 Thinking	89.7	79.4	74.4	75.6	84.2	70.6	80.9	93.1	84.4	80.1
GLM 4.7	88.0	79.4	74.0	73.1	83.7	70.8	79.9	94.9	85.5	79.6
Qwen3 235B Thinking	89.6	76.2	72.1	72.8	84.0	71.2	82.3	88.4	83.9	78.6
DeepSeek R1	84.0	74.5	71.5	71.7	80.6	69.8	76.7	92.1	84.0	76.8
GLM 4.5	82.6	74.5	70.4	71.1	79.9	68.8	75.0	87.5	81.7	75.5
Qwen3 32B	77.8	70.8	64.8	67.8	73.6	66.0	72.9	80.1	80.0	71.2

Qwen3 32B, GLM 4.7, GLM 4.5, and Kimi K2 (Thinking). These constitute the deployed system's selectable model pool and are thus the practical candidates for cost-effective deployment at scale. A frontier model, Gemini 3.1 Pro, serves as the fixed reference reviewer rather than as a generator candidate. All generators are driven through the identical production pipeline: the same subject- and skill-specific system prompts, the same textbook-grounded input schema, and the same structured-output format, with only the model identifier varied per request. Reasoning ("thinking") modes are enabled for the models that support them, matching the production configuration.

B. Generator Evaluation

1) Task and Metric: Each generation task provides the model with a grade, subject, and textbook page range; the corresponding page content is retrieved from the curriculum database and supplied as the grounding context. The model returns a single structured lesson plan in one pass, with no subsequent review-and-refine iteration (the iterative loop is evaluated separately in Section VI). Each plan is then scored by the reference reviewer against the v3 rubric, yielding a 1–4 rating on every sub-criterion. Criterion scores are aggregated to a percentage of the *assessable* maximum: checks whose required context is unavailable for a given plan (for example the Active-Retrieval check, which requires the previous lesson's objective) are marked not-assessable and excluded from the denominator rather than penalised, keeping scores comparable across plans. We report, per model, the mean percentage on each criterion (C0–C8) and the overall mean.

2) Results: Table II reports the results. DeepSeek V3.2 scored highest overall (83.3%), followed by Kimi K2 (80.1%) and GLM 4.7 (79.6%); Qwen3 32B scored lowest (71.2%), a range of twelve percentage points across the pool. The top-ranked model is a non-reasoning model with among the lowest per-plan cost, and it scored above the larger reasoning models, so reasoning capacity did not predict generation quality in this evaluation.

The per-criterion profile was consistent across all seven models. Every model scored highest on Factual Accuracy (C7), Structural Completeness (C0), and the subject-specific criterion (C8), indicating that the models produced complete, internally consistent, and subject-appropriate plans. Scores were lowest on Instructional Flow (C2), Deeper Learning and Engagement (C5), and Practicality (C3). These criteria capture the harder pedagogical requirements: an explicit Gradual Release structure (I Do, We Do, You Do), checks for understanding placed after each phase, higher-order and participatory activities, and realistic time and resource budgeting. The models were stronger at the structure of a lesson plan than at its instructional detail. This is the main result for practitioners: automated generation meets the structural requirements but still needs review against the instructional-flow and engagement criteria, which is the function of the reviewer described in Section V.

C. Reviewer Evaluation

1) Task and Metric: Each candidate model is given a fixed set of lesson plans together with the full v3 rubric specification and asked to return structured JSON scores for every sub-criterion, exactly as the deployed Reviewer Agent does (Section V). We measure how well each open-weight model reproduces the judgments of the fixed reference reviewer, Gemini 3.1 Pro. Comparison against a single human annotator does not scale to seven models across every task and is susceptible to individual rater bias; a frontier reference reviewer can instead be applied uniformly to every plan, and its own reliability is established independently against expert human scores (Section III-D): across eight independently scored plans the reference reviewer's criterion scores match specialist scores to a mean absolute difference of 1.7 points (Pearson $r = 0.97$). Using a strong model as an evaluation reference is consistent with the LLM-as-judge paradigm [32], [33]. A committee-based reference such as a Language Model Council [31], which aggregates a panel of judges to further reduce single-model bias, is a natural extension left to future work on cost grounds.

The reference reviewer scores every plan once; each candidate then reviews the same plans under the identical rubric and prompt. Agreement is reported per candidate as the mean absolute error (MAE) and mean signed bias of its overall rubric percentage relative to the reference, together with the Pearson (r) and Spearman (ρ) correlations of the overall score across plans, which capture whether a candidate ranks plans as the reference does. Per-criterion MAE identifies which rubric dimensions are hardest to evaluate consistently. We also report the quadratic-weighted Cohen's κ between each candidate and the reference over the per-criterion scores, binned to the four-point rubric scale, as a chance-corrected agreement measure that credits near-misses more than far-misses.

2) Results: All seven open-weight reviewers score more leniently than the reference: every model has a positive mean bias, from $+3.5$ points (Kimi K2) to $+12.5$ points (GLM 4.7), so none is stricter than Gemini on average. Absolute agreement (MAE) ranges from 6.7 points for DeepSeek V3.2, the closest, to 13.5 points for GLM 4.7, the furthest; even the best

TABLE III

REVIEWER EVALUATION: AGREEMENT OF EACH OPEN-WEIGHT REVIEWER WITH THE FIXED GEMINI 3.1 PRO REFERENCE ACROSS THE SAMPLED LESSON PLANS. MAE AND BIAS ARE IN RUBRIC PERCENTAGE POINTS; r AND ρ ARE THE PEARSON AND SPEARMAN CORRELATIONS OF THE OVERALL SCORE; κ_w IS THE QUADRATIC-WEIGHTED COHEN'S κ OVER THE PER-CRITERION SCORES BINNED TO THE FOUR-POINT RUBRIC SCALE. LOWER MAE AND HIGHER CORRELATION OR κ_w INDICATE CLOSER AGREEMENT WITH THE REFERENCE.

Reviewer	**n**	**MAE**	**Bias**	r	ρ	κ_w
DeepSeek V3.2	17	6.7	+4.8	0.51	0.45	0.35
Kimi K2 Thinking	14	7.4	+3.5	0.47	0.36	0.46
Qwen3 32B	19	9.4	+8.2	0.40	0.56	0.28
DeepSeek R1	12	9.5	+9.3	0.60	0.59	0.43
GLM 4.5	19	11.6	+11.5	0.59	0.65	0.37
Qwen3 235B Thinking	18	12.6	+12.4	0.36	0.43	0.32
GLM 4.7	14	13.5	+12.5	0.20	0.30	0.22

open reviewer is off by roughly seven points per plan. Rank agreement is weak to moderate (r from 0.20 to 0.60), and MAE and correlation do not move together: DeepSeek R1 and GLM 4.5 preserve much of the reference ordering ($\rho \approx 0.6$) while still inflating the absolute level, whereas GLM 4.7 barely tracks it ($r = 0.20$). Across criteria, disagreement concentrates on Cultural Relevance (C6) and Factual Accuracy (C7), the dimensions that most require contextual and verification judgment, and is smallest on Practicality (C3). Chance-corrected agreement tells the same story: quadratic-weighted Cohen's κ ranges from 0.22 (GLM 4.7) to 0.46 (Kimi K2), all only fair to moderate and none approaching substantial agreement, and it ranks the models slightly differently from MAE because a reviewer can sit close to the reference on average yet still disagree on which criteria to reward. No open model is a close substitute for the frontier reviewer.

D. Cross-Analysis: Generator vs. Reviewer Capability

Generation ability does not predict review ability. DeepSeek V3.2 is both the strongest single-pass generator (Table II) and the reviewer closest to the reference (Table III), but the roles otherwise decouple: GLM 4.7 is among the stronger generators yet the least reliable reviewer ($r = 0.20$, MAE 13.5), while Qwen3-32B, the weakest generator, is a middling reviewer. Writing a good lesson plan and judging one are distinct capabilities, and the latter is the scarcer. The uniform positive bias compounds this: an open model placed at the quality gate would pass plans that the frontier reviewer rejects, the costly failure mode for an automated reviewer. Together these results motivate the deployment split used in production and evaluated here: open-weight models for generation, where Table II shows they reach frontier-competitive quality (especially after one refinement pass) at a fraction of the cost and at the highest request volume, paired with a single frontier model as the fixed reviewer and quality gate, where its judgment justifies the premium at lower volume.

V. MULTI-AGENT LESSON PLAN GENERATION SYSTEM

A. System Architecture

Figure 1 illustrates the overall pipeline. The system comprises two specialised agents (a Generator Agent and a Reviewer Agent) orchestrated in an iterative refinement loop. Both agents are backed by large language models accessed through a unified model gateway, enabling flexible model selection without changes to the surrounding pipeline. The system supports two curricula (ICT and Punjab) across Grades 1–5 and four subjects (English, Urdu, Mathematics, Science), and is deployed in production serving teachers across the Islamabad Capital Territory, Punjab, and Balochistan regions of Pakistan.

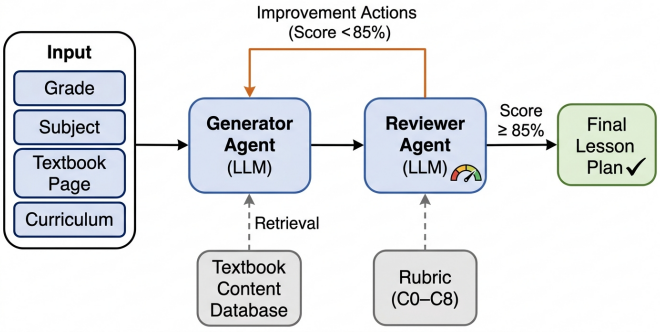

Fig. 1. Multi-agent lesson plan generation and refinement pipeline.

B. Textbook Grounding

The system grounds every lesson plan in the prescribed curriculum content rather than relying on the model's parametric knowledge. At generation time, the relevant textbook page or page range is retrieved from an OCR-processed database of primary school textbooks and passed to the Generator Agent as the primary content signal. Textbooks are indexed by curriculum, grade, and subject, covering both the ICT and Punjab curricula. This grounding addresses the Curriculum Alignment criterion (C1) of the rubric and reduces the risk of hallucinated textbook references.

C. Generator Agent

The Generator Agent receives a structured request specifying the grade level, subject, textbook page range, class size, and curriculum. The system prompt is assembled dynamically from three components: a cross-cutting policies block encoding constraints on cultural sensitivity, language register, and curriculum fidelity; a subject- and curriculum-specific pedagogical prompt selected from a curated prompt library; and a structured output format specification requiring the model to produce a lesson plan with five mandatory sections (Student Learning Outcome, Hook, Explanation, Practice, and Conclusion), each annotated with time allocations and teacher-facing instructions.

The prompt library is designed to cover the diversity of lesson types found in Pakistani primary schools. Separate

prompts exist for each subject–curriculum combination. For English and Urdu, skill-focused variants target specific literacy competencies: reading, comprehension, grammar, and creative writing. For Mathematics, prompts corresponding to the three CPA stages (Concrete, Pictorial, and Abstract [25]) and word problems are provided, ensuring the pedagogical approach is matched to the cognitive demand of the lesson.

Model selection for the Generator Agent is informed directly by the benchmark findings reported in Section IV. The same model gateway supports per-subject model assignment, allowing, for example, a different model to be used for Urdu generation than for English or Mathematics, reflecting the observed variation in multilingual capability across frontier models.

D. Reviewer Agent

The Reviewer Agent accepts a generated lesson plan and returns a structured pedagogical evaluation against the rubric described in Section III. The reviewer prompt is assembled by concatenating the eight common rubric criteria (C0–C7) with the subject-specific criterion (C8), together with an output specification requiring machine-readable structured feedback.

The reviewer evaluates the lesson plan against each sub-criterion and returns, for every criterion: a score, a rationale paragraph, a list of strengths, and a prioritised list of concrete improvement actions, each mapped to the specific sub-criterion that scored below the acceptable threshold. This format ensures that the refinement loop receives targeted, sub-criterion-level instructions rather than open-ended comments.

The reviewer also evaluates the pedagogical appropriateness of any visual aids present in the lesson plan. Rather than requiring a separate vision model call, image prompts are extracted from the lesson plan and passed as structured text context, allowing the reviewer to assess whether images would support instruction at the specific pedagogical moment where they appear.

The reviewer is invoked at a low temperature to maximise scoring consistency, a design choice validated by the reviewer reliability analysis in Section IV-C. Model selection for the Reviewer Agent follows the same benchmark-informed approach as the Generator Agent, with a separate model assignment available for Urdu.

E. Iterative Refinement Loop

After the Reviewer Agent scores a lesson plan, its structured improvement actions are injected into the Generator Agent as a targeted revision prompt. The generator is instructed to address each flagged deficiency while preserving all higher-scoring content unchanged, implementing a minimal-edit principle: only the weakest aspects of the lesson plan are rewritten, preserving content that already scores well and reducing per-iteration inference cost.

Convergence is reached when either the overall rubric score exceeds the "Strong" threshold (85%) or no sub-criterion scores below the acceptable level across all criteria. A maximum iteration cap bounds inference cost; lesson plans that have not converged within the cap are returned at their highest-scoring intermediate state, keeping per-LP cost practical at the scale of thousands of teachers.

F. Supporting Pipeline Components

Two additional components augment the core generation–review loop.

Visual content generation. When enabled, a secondary LLM call analyses the generated lesson plan and identifies locations where educational images would support instruction, generating descriptive prompts for each. A semantic similarity check against a cache of previously generated images avoids redundant generation: prompts with high similarity to an existing image reuse that image at no additional cost. Novel images are generated concurrently and embedded into the lesson plan. The generated images are evaluated under the Visual Aid Effectiveness sub-criterion (C2) of the rubric.

Bilingual translation. For English, Mathematics, and Science lesson plans, an optional translation stage produces a parallel bilingual version with key instructional content rendered in both English and Urdu. This supports teachers in mixed-medium classrooms and reflects the linguistic reality of Pakistani primary schools, where code-switching between English and Urdu is common practice [1].

G. Deployment Context

The system is deployed as a web service on cloud infrastructure, supporting both synchronous requests for interactive use and asynchronous processing with webhook callbacks for bulk generation workflows. The asynchronous mode is particularly relevant for Taleemabad's operational context, in which large batches of lesson plans are generated ahead of teacher training programmes. The pipeline currently serves two curricula across three administrative regions of Pakistan and is designed to support additional curricula and grade bands without architectural changes.

VI. System Evaluation

A. Experimental Setup

We evaluate the refinement loop of Section V on the same corpus as the generator benchmark (Section IV-B): the 124 lesson plans produced by the seven open-weight models across the 18 canonical tasks.[1] Each plan is refined by returning the model's own first-pass review to the editing agent (Section V-C): the per-criterion improvement actions emitted by the reviewer become the edit request, and the *original generator model* performs the revision, so that each model refines its own work rather than a stronger model's. The reviewer is held fixed at Gemini 3.1 Pro under the v3 rubric throughout, so that a score change reflects the generator's response to feedback and not reviewer variation. The single-pass results of Table II form

[1]Two of Kimi K2's cells are absent. One task was never generated, and one refinement call could not be completed by a reasoning-capable provider within the request timeout. Both follow from a single OpenRouter model identifier being served by several providers with differing support for reasoning mode, which leaves $n = 16$ for this model and $n = 18$ for the other six.

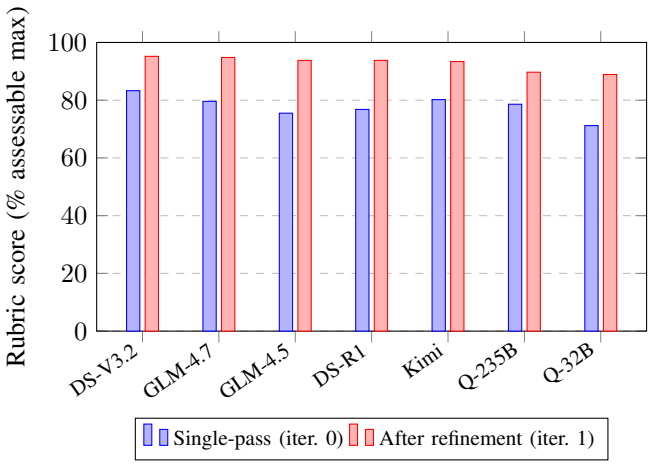

Fig. 2. Rubric score before (single-pass generation) and after one self-refinement iteration, per generator, with the reviewer (Gemini 3.1 Pro) held fixed. All seven models improve, and the weakest single-pass generators gain the most.

TABLE IV
SELF-REFINEMENT GAINS PER GENERATOR (MEAN RUBRIC %, ONE ITERATION, REVIEWER FIXED AT GEMINI 3.1 PRO).

Generator	Single-pass	Refined	Δ
DeepSeek V3.2	83.3	95.2	+11.9
GLM 4.7	79.6	94.8	+15.2
GLM 4.5	75.5	93.8	+18.3
DeepSeek R1	76.8	93.8	+16.9
Kimi K2	80.2	93.4	+13.2
Qwen3-235B	78.6	89.7	+11.0
Qwen3-32B	71.2	88.9	+17.7
Overall	**77.9**	**92.8**	**+14.9**

the iteration-0 baseline. Refinement is capped at two iterations and halts earlier when the first pass yields a large mean gain.

B. Rubric Score Improvement Across Iterations

A single refinement iteration raises the mean rubric score from 77.9% to 92.8%, a gain of 14.9 percentage points (Table IV, Fig. 2). The effect is uniform: all seven generators improve, by between 11.0 points (Qwen3-235B) and 18.3 points (GLM 4.5), and 121 of the 124 plans score higher after refinement. Because a single pass already produces a large and uniform improvement, the loop halts at one iteration and a second pass is not run.

C. Iteration Analysis

Refinement acts as a leveller. The gap between the strongest and weakest generator narrows from 12.1 points at iteration 0 (71.2% to 83.3%) to 6.3 points after refinement (88.9% to 95.2%), because the weakest single-pass generators gain the most: Qwen3-32B and GLM 4.5, the two lowest at iteration 0, rise by 17.7 and 18.3 points. Single-pass quality is thus a weak predictor of the refinement ceiling. Qwen3-235B, fourth at iteration 0, gains the least and finishes last, whereas GLM 4.7 climbs from mid-field to a near-tie for first.

Three of the 124 plans regress, the largest by 42 points. In each case a reasoning model corrupted the document structure during self-editing, dropping or garbling sections so that the revision scored below the original. These are the plans that fail to converge within the iteration cap, and they motivate adding a structural-integrity check to the editing step.

One limitation of this measurement is intrinsic to the design. Because the same model both critiques the plan and re-scores the revision, part of the observed gain reflects the generator acting on the specific feedback that its own judge will grade against. These figures are therefore best read as an upper bound on self-refinement under a fixed reviewer, rather than as an independent-judge measurement. The expert evaluation below addresses this directly.

D. Expert Teacher Evaluation

To distinguish genuine pedagogical improvement from reviewer-aligned gains, refined plans are additionally assessed by Taleemabad curriculum specialists against the rubric dimensions of Section III-B, blind to iteration and generator. This validation extends the human annotation set introduced in Section IV-C. At the present sample size the human results are preliminary, and a larger blind panel is in progress; we therefore report them as an agreement check on the fixed-reviewer scores above rather than as a standalone measure.

E. Qualitative Examples

The refinement actions are concrete and rubric-driven. In a Grade 4 Science plan on roots and stems (GLM 4.7), the first-pass reviewer scored Cultural Relevance and Representation at 7 of 16, noting the absence of any local grounding, and instructed the model to teach the concept through locally familiar flora, for example wheat or a rose bush, and to use diverse Pakistani names such as “Sana and Ali” in the practice items. The revised plan incorporated both and scored 16 of 16 on that criterion. In a Grade 2 Maths plan (Qwen3-235B), the single-pass objective conflated Roman-numeral recognition with unrelated addition and regrouping; the reviewer required a single-focus student learning outcome carrying its SNC code, and the revision narrowed the objective and added the code, raising Curriculum Alignment from 10 to 24 of 24. Both cases reflect the pattern behind the aggregate gains: the largest improvements arise in the instructional-design and contextualization criteria (Section III-B) that single-pass generation most often neglects.

VII. DISCUSSION

A. Key Benchmark Findings and Practitioner Implications

The generator benchmark (Table II) yields two findings with direct bearing on deployment. First, the quality ordering does not follow model scale or reasoning capability: the highest-scoring generator, DeepSeek V3.2, is a comparatively small non-reasoning model with among the lowest per-plan cost,

yet it exceeds the larger reasoning models on the rubric. For a system that must produce lesson plans at the scale of thousands of teachers, this indicates that a low-cost model is a defensible default and that reasoning-enabled models do not automatically justify their higher latency and cost. Second, the per-criterion profile is uniform across all seven models: they score highest on structural completeness, factual accuracy, and subject-specific coverage, and lowest on instructional flow, deeper learning, and practicality. The capabilities the models lack are the ones that separate a usable lesson plan from a merely well-formed one, namely an explicit gradual-release structure, checks for understanding, and realistic classroom pacing. A practitioner deploying these models should expect structurally sound output and should aim quality assurance at instructional design rather than at completeness.

B. The Reviewer Capability Gap

The clearest result of this study is an asymmetry between generating and reviewing. The open-weight models generate lesson plans at near-frontier quality (Table II), and a single self-refinement pass lifts every one of them above 88% of the rubric maximum. As reviewers, however, they diverge markedly from the frontier reference (Table III): all seven score more leniently, with a positive bias of up to 12.5 points, and their rank agreement with the reference ranges from negligible ($r = 0.20$) to only moderate ($r = 0.60$). Reviewing is the harder and less transferable capability, and a model's strength as a generator is a poor guide to its reliability as a judge. For a system that depends on an automated quality gate, this is the central practical finding: the open models are ready to generate but not to review, so the frontier reviewer earns its place precisely where the open models fall short.

C. Multilingual and Cultural Considerations

The benchmark evaluates Urdu generation on the same rubric as the other subjects, treating language as a pedagogical variable rather than a presentation variable. A detailed multilingual analysis, comparing per-subject quality and examining script handling and code-switching, requires a per-subject breakdown of the results and is left to future work; the present study establishes that Urdu lesson-plan generation can be scored on the same instrument as English, Mathematics, and Science.

D. Limitations

Several limitations bound these results. The rubric scores are produced by a single reference reviewer model rather than by human experts, so they inherit that model's biases; agreement with expert annotations is measured on a small set (Section IV-C), and a larger human study is needed to calibrate the automated scores. The generation benchmark covers one curriculum (ICT) and 18 tasks, with Science limited to the two grades for which production data exists, so per-subject and per-grade conclusions are indicative rather than definitive. Each plan is generated in a single pass; the effect of the iterative refinement loop is evaluated separately (Section VI) and is not reflected in Table II. Finally, models were accessed through a shared gateway that may route one model identifier across providers with differing capabilities and rate limits; occasional provider-side failures for reasoning models were handled by retrying, but this routing is a source of variance that a fully controlled study would remove.

VIII. Conclusion and Future Work

We presented three contributions toward AI-assisted lesson plan development for Pakistani primary education: a pedagogically grounded evaluation rubric, a benchmark study of LLMs as generators and reviewers of lesson plans, and a multi-agent generation-refinement system whose design is directly informed by benchmark findings. Our results demonstrate that current open-weight models generate structurally complete and factually accurate lesson plans but score lower on the instructional-design criteria that most affect classroom usefulness, which is the gap the rubric-guided reviewer is designed to close.

Future work will focus on: (i) fine-tuning a dedicated reviewer model on our rubric-labelled dataset to improve reviewer reliability at lower inference cost; (ii) expanding coverage to Grades 6–8 and additional subjects; (iii) integrating teacher-in-the-loop validation to capture practitioner feedback as a training signal; and (iv) open-sourcing the full benchmark dataset to encourage reproducible research in EdTech AI for South Asian and similarly underrepresented contexts.

References

[1] I. Ahmad, K. Rehman, A. Ali, I. Khan, and A. F. Khan, "Critical analysis of the problems of education in Pakistan: Possible solutions," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 3, no. 2, pp. 79–84, 2014. https://doi.org/10.11591/ijere.v3i2.1805

[2] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen, "Large language models for education: A survey and outlook," *arXiv preprint arXiv:2403.18105*, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2403.18105

[3] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain*. New York: David McKay, 1956.

[4] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7–74, 1998.

[5] H. Fan *et al.*, "LessonPlanner: Assisting novice teachers to prepare pedagogy-driven lesson plans with large language models," in *Proc. 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*, Pittsburgh, PA, USA, Oct. 2024, pp. 1–18. https://doi.org/10.1145/3654777.3676390

[6] S. Al-Raqqad, M. Al-Shammary, and R. Al-Anazi, "Technology review of Magic School AI: An intelligent way for education inclusivity and teacher workload reduction," *Education Sciences*, vol. 15, no. 8, p. 963, Jul. 2025. DOI: https://doi.org/10.3390/educsci15080963

[7] B. Hu, L. Zheng, J. Zhu, L. Ding, Y. Wang, and X. Gu, "Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design," *IEEE Trans. Learning Technologies*, vol. 17, pp. 1471–1485, 2024. DOI: https://doi.org/10.1109/TLT.2024.3384765

[8] Y. Zheng *et al.*, "Automatic lesson plan generation via large language models with self-critique prompting," in *Proc. 25th Int. Conf. Artificial Intelligence in Education (AIED)*, Recife, Brazil, Jul. 2024, pp. 134–147. DOI: https://doi.org/10.1007/978-3-031-64315-6_13

[9] Y. Zheng *et al.*, "Knowledge-enhanced large language models for automatic lesson plan generation," *Humanities and Social Sciences Communications*, vol. 12, no. 1784, 2025. DOI: https://doi.org/10.1057/s41599-025-06004-2

[10] D. V. Dennison *et al.*, "Shiksha Copilot: Teacher-AI collaboration for curating and customizing lesson plans in low-resource schools," *arXiv preprint arXiv:2507.00456*, Jul. 2025. [Online]. Available: https://arxiv.org/abs/2507.00456

[11] Y. Chen, Y. Yang, and G. Xu, "AgentLesson: LLM-based multi-agent system for educational lesson plan generation," in *Behavioural and Social Computing*. Singapore: Springer, 2025. DOI: https://doi.org/10.1007/978-981-95-7138-3_11

[12] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, New Orleans, LA, USA, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[13] A. Madaan *et al.*, "Self-Refine: Iterative refinement with self-feedback," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, New Orleans, LA, USA, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2303.17651

[14] N. Shinn *et al.*, "Reflexion: Language agents with verbal reinforcement learning," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, New Orleans, LA, USA, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2303.11366

[15] H. Hashemi, J. Eisner, C. Rosset, B. Van Durme, and C. Kedzie, "LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts," in *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, Bangkok, Thailand, Aug. 2024, pp. 13806–13834. [Online]. Available: https://aclanthology.org/2024.acl-long.745

[16] Y. Bai *et al.*, "EduBench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios," *arXiv preprint arXiv:2505.16160*, May 2025. [Online]. Available: https://arxiv.org/abs/2505.16160

[17] M. Lelièvre *et al.*, "Benchmarking the pedagogical knowledge of large language models," *arXiv preprint arXiv:2506.18710*, Jun. 2025. [Online]. Available: https://arxiv.org/abs/2506.18710

[18] S. Arif, A. H. Azeemi, A. A. Raza, and A. Athar, "Generalists vs. specialists: Evaluating large language models for Urdu," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, FL, USA, Nov. 2024. DOI: https://doi.org/10.48550/arXiv.2407.04459

[19] Y. Li, M. Rakovic, B. X. Poh, D. Gašević, and G. Chen, "Automatic classification of learning objectives based on Bloom's taxonomy," in *Proc. 15th Int. Conf. Educational Data Mining (EDM 2022)*, Durham, UK, Jul. 2022, pp. 530–537.

[20] N. Scaria, S. D. Chenna, and D. Subramani, "Automated educational question generation at different Bloom's skill levels using large language models: Strategies and evaluation," in *Proc. 25th Int. Conf. Artificial Intelligence in Education (AIED)*, Recife, Brazil, Jul. 2024. DOI: https://doi.org/10.1007/978-3-031-64299-9_12

[21] B. Memarian and T. Doleck, "A review of assessment for learning with artificial intelligence," *Computers in Human Behavior: Artificial Humans*, vol. 2, art. 100040, 2024. DOI: https://doi.org/10.1016/j.chbah.2023.100040

[22] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Proc. 41st Int. Conf. Machine Learning (ICML 2024)*, Vienna, Austria, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2305.14325

[23] X. Zhang, C. Zhang, J. Sun, J. Xiao, Y. Yang, and Y. Luo, "EduPlanner: LLM-based multi-agent systems for customized and intelligent instructional design," *arXiv preprint arXiv:2504.05370*, Apr. 2025. [Online]. Available: https://arxiv.org/abs/2504.05370

[24] P. D. Pearson and M. C. Gallagher, "The instruction of reading comprehension," *Contemporary Educational Psychology*, vol. 8, no. 3, pp. 317–344, Jul. 1983. DOI: https://doi.org/10.1016/0361-476X(83)90019-X

[25] J. S. Bruner, *Toward a Theory of Instruction*. Cambridge, MA: Harvard University Press, 1966.

[26] I. Baloch and A. Taddese, *EdTech in Pakistan: A Rapid Scan*. EdTech Hub Country Scan, Jun. 2020. DOI: https://doi.org/10.5281/zenodo.3911655. [Online]. Available: https://docs.edtechhub.org/lib/FIQDEKCI

[27] Policy and Planning Wing, Ministry of Education, Government of Pakistan, *National Professional Standards for Teachers in Pakistan*. Islamabad: Ministry of Federal Education and Professional Training, Feb. 2009. [Online]. Available: https://www.nacte.org.pk/assets/download/NationalProfessionalStandardsforTeachersinPakistan.pdf

[28] M. Hussain, N. Hussain, and M. Kousar, "Role of Pakistan Reading Project (PRP) in promoting Urdu reading skill at primary level in Azad Jammu and Kashmir," *International Research Journal of Education and Innovation*, vol. 3, no. 1, pp. 226–238, 2022. https://doi.org/10.53575/irjei.v3.01.21(22)226-238

[29] Office of Curriculum, Assessment and Teaching Transformation, University at Buffalo, "Teaching methods," University at Buffalo, 2024. [Online]. Available: https://www.buffalo.edu/catt/teach/develop/design/teaching-methods.html

[30] N. L. Webb, "Depth-of-knowledge levels for four content areas," unpublished paper, Wisconsin Center for Educational Research, University of Wisconsin–Madison, Mar. 28, 2002. [Online]. Available: http://ossucurr.pbworks.com/w/file/fetch/49691156/norm%20web%20dok%20by%20subject%20area.pdf

[31] J. Zhao, F. M. Plaza-del-Arco, B. Genchel, and A. C. Curry, "Language model council: Democratically benchmarking foundation models on highly subjective tasks," in *Proc. 2025 Conf. of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2025)*, Albuquerque, NM, USA, Apr. 2025, pp. 12395–12450. DOI: https://doi.org/10.18653/v1/2025.naacl-long.617

[32] J. Gu *et al.*, "A survey on LLM-as-a-judge," *arXiv preprint arXiv:2411.15594*, Nov. 2024. DOI: https://doi.org/10.48550/arXiv.2411.15594

[33] H. Li *et al.*, "LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods," *arXiv preprint arXiv:2412.05579*, Dec. 2024. DOI: https://doi.org/10.48550/arXiv.2412.05579