

FACET: Multi-Agent AI Supporting Teachers in Scaling Differentiated Learning for Diverse Students

Jana Gonnermann-Müller^{1,4}, Jennifer Haase^{3,4} and Nicolas Leins¹ and Moritz Igel¹ and Konstantin Fackeldey^{1,2} and Sebastian Pokutta^{1,2}

¹Zuse Institute Berlin

²Technische Universität Berlin

³Humboldt-Universität zu Berlin

⁴Weizenbaum Institute Berlin

{gonnermann-mueller, leins, igel, fackeldey, pokutta}@zib.de, jennifer.haase@hu-berlin.de

Abstract

Classrooms are increasingly heterogeneous, with learners varying in performance, motivation, language proficiencies, and learning differences such as dyslexia and ADHD. While teachers recognize the need for differentiated instruction, growing workloads make it an ideal that is often unrealized in practice. Current AI educational tools are predominantly student-facing and performance-centric, ignoring other aspects that shape learning outcomes. We introduce FACET, a teacher-facing multi-agent framework that supports differentiation across motivation, performance, and learning differences. Developed with educational stakeholders from the outset, the framework coordinates four specialized agents, including learner simulation, diagnostic assessment, material generation, and evaluation within a teacher-in-the-loop design. School principals ($N = 30$) shaped system requirements through participatory workshops, while in-service K–12 teachers ($N = 70$) evaluated material quality. Mixed-methods evaluation demonstrates strong perceived value for inclusive differentiation. Practitioners emphasized both the urgent need arising from classroom heterogeneity and the importance of maintaining pedagogical autonomy as a prerequisite for adoption. We discuss implications for future school deployment and outline partnerships for longitudinal classroom implementation.

1 Introduction

Today’s classrooms are increasingly diverse, with students varying widely in prior knowledge, motivation, language proficiency, and learning differences. This includes a significant number of learners with neurodiversity: an estimated 5% to 17% of schoolchildren have dyslexia [Habib and Giraud, 2013] and 11.4% of U.S. children aged 3–17 years received an attention-deficit/hyperactivity disorder (ADHD) diagnosis in 2022 [Danielson *et al.*, 2024]. Addressing such heterogeneity requires differentiated instruction, for example, adapting content and support to individual learners. Yet

differentiation significantly increases instructional complexity and teacher workload, creating a persistent gap between pedagogical ideals and what is feasible in everyday classroom practice. Artificial intelligence (AI) has emerged as a potential solution to this challenge. AI-driven educational platforms, such as Khan Academy’s Khanmigo in the United States [Khan Academy, 2025] or Squirrel AI in China [SquirrelAI, 2025], offer individualized learning paths. Yet these systems are predominantly student-facing, designed for direct learner support and interaction, such as [Zhang *et al.*, 2025], rather than for teacher support. Moreover, their adaptation is typically driven by performance metrics alone, leaving them largely insensitive to the motivational barriers, affective responses, and processing differences that characterize diverse learners. Educational research demonstrates that effective differentiation must extend beyond cognitive performance to include motivational factors and learning differences, such as reading and writing difficulties or ADHD [Bernacki *et al.*, 2021; Gaye *et al.*, 2024]. Evidence supports this approach: Stepwise cognitive strategies improve mathematics outcomes for students with ADHD [Iseman and Naglieri, 2011], whose working memory deficits require specially designed materials [Gaye *et al.*, 2024]. However, implementing such comprehensive differentiation is too time-intensive for most teachers, who increasingly turn to large language models (LLMs) (e.g., ChatGPT) for generating teaching materials [Schummel *et al.*, 2025]. However, using single LLMs presents several limitations: They exhibit instability in learner simulation and high prompt sensitivity [Mannekote *et al.*, 2025; Ilkou *et al.*, 2025], they often fail to align with specific curricula, learning goals, or classroom contexts, and they lack the scalability required for complex workflows, which becomes problematic when accommodating several learning groups and new requirements during the semester. Targeted multi-agent architectures offer a promising alternative to single LLM approaches [Wang *et al.*, 2025a; Haase and Pokutta, 2025]. They enable stable, reusable learner representations through specialized agents with distinct pedagogical roles. Teachers can use the same student personas throughout the school year, reducing prompting effort. Additionally, agents can be equipped with structured knowledge, such as curriculum-aligned tasks and learning goals, didactic

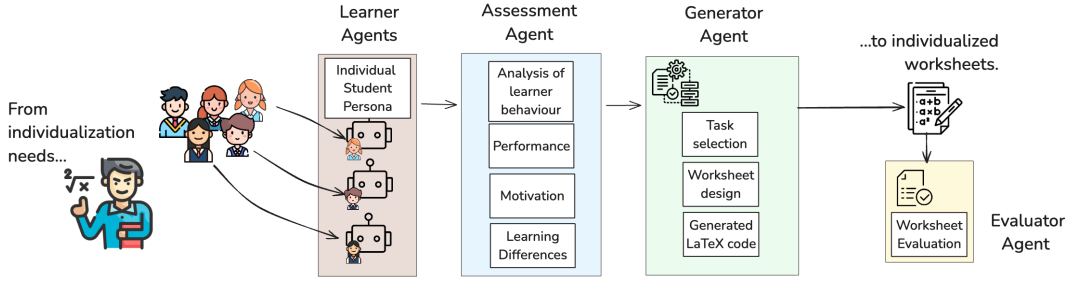


Figure 1: Core principles for individualized classroom material using a multi-agent framework

approaches, and guidelines for designing accessible teaching materials for students with dyslexia, ADHD, and other learning requirements.

Building on these advances, we present FACET, a teacher-facing multi-agent system designed to operationalize differentiation beyond performance while preserving teacher authority (see Figure 1). Teachers define learner profiles based on diverse characteristics, modify uploaded materials or generate new content, and review all outputs before classroom use. This approach keeps teachers in control, ensuring that AI-generated materials serve their pedagogical goals and fit their classroom realities. Therefore, FACET was developed through close collaboration with domain experts; from the beginning, mathematics teachers informed curriculum alignment and validated the didactic structure of the generated worksheets, while teachers with expertise in dyslexia reviewed accessibility adaptations. This close partnership with practitioners extends throughout our research, from workshops and teacher surveys presented in this paper to an ongoing longitudinal deployment across schools, ensuring that FACET remains grounded in authentic classroom needs.

Our key contributions are:

1. FACET, a modular four-agent architecture that separates learner simulation, diagnostic reasoning, material generation, and evaluation.
2. An operationalization of learner differentiation beyond performance through teacher-specifiable profiles, including cognitive and motivational aspects, as well as neurodiversity.
3. Mixed-methods evaluation with school principals ($N = 30$) and in-service teachers ($N = 70$) assessing workflow integration and material quality.

2 Related Work

Relevant work spans three areas: differentiated and inclusive instruction, multi-agent systems in education, and teacher-facing AI systems.

2.1 Differentiated and Inclusive Instruction

Research consistently shows that effective differentiation must address cognitive proficiency, motivation, and needs related to neurodiversity to improve learning outcomes and motivation [Bernacki *et al.*, 2021; Fariani *et al.*, 2023]. Self-determination theory highlights the importance of supporting autonomy, competence, and relatedness for sustained

engagement [Deci and Ryan, 2000], which are needs that performance-based metrics fail to capture. Neurodiversity refers to the natural variation in human brain function, often used in the context of autism spectrum disorder, ADHD, and learning disabilities. ADHD is characterized by developmentally inappropriate levels of inattention, hyperactivity, and impulsivity that persist over time [World Health Organisation (WHO), 2025]. Dyslexia is a specific learning disability characterized by difficulties with accurate and fluent word recognition, as well as poor spelling and decoding, typically stemming from a deficit in the phonological component of language, with potential secondary consequences for reading comprehension and broader academic development [Carroll *et al.*, 2025]. For learners with dyslexia or ADHD, learning and engagement depend heavily on instructional clarity, structure, and motivational support. Reading difficulties, language-processing challenges, and attention deficits can increase cognitive load, leading to reduced learning outcomes and disengagement. These learner needs translate into concrete requirements for differentiated instructional design, which requires deliberate design choices, including task sequencing, scaffolding, and instructional structure that go beyond content adaptation [Ilkou *et al.*, 2025]. For learners with dyslexia, research- and practice-oriented guidelines emphasize the importance of readable, well-structured worksheets that reduce extraneous cognitive load [Rello and Baeza-Yates, 2013]. Scaffolding strategies, such as worked examples, step-wise guidance, and partially completed representations, support learners' progression toward independent problem solving. Equally important are motivational cues that sustain engagement and reinforce self-efficacy. Bloom's revised taxonomy provides a structured framework for guiding learners from basic recall to higher-order thinking [Anderson and Krathwohl, 2001], enabling systematic task differentiation that balances challenge and support.

2.2 LLM and Multi-Agent Systems for Differentiated Education

Recent work further shows that LLMs can simulate stable personality traits and affective responses when equipped with detailed persona prompts specifying observable characteristics [Aher *et al.*, 2023; Wu *et al.*, 2025]. Crucially, without explicit constraints, LLM default to average-student personas; however, with structured profile specifications, LLM-based agents can produce cognitively diverse behavior, including neurodivergent profiles [Gonnermann-Müller *et al.*,

2026a; Gonnermann-Müller *et al.*, 2026b]. These capabilities enable the construction of reusable learner profiles [Wang *et al.*, 2025b; Huang *et al.*, 2026] that overcome the scalability limitations of earlier AI-based tutoring systems and single LLM approaches for teachers. Building on this foundation, emerging research explores LLM-based digital twins and multi-agent classroom simulations in which multiple agents represent heterogeneous learners, instructional roles, or pedagogical functions [Zhang *et al.*, 2025; Wu *et al.*, 2025]. Multi-agent architectures enable the separation of subtasks across specialized agents (e.g., learner simulation, pedagogical reasoning, assessment, adaptation), allowing intermediate reasoning steps to be inspected and corrected. Recent work demonstrates examples of student-facing multi-agent systems, such as the coordination of specialized agents for educational assessment; *SimClass*, a simulated classroom [Zhang *et al.*, 2025] and *GenMentor*, which coordinates multiple agents for goal mapping, skill gap identification, and adaptive content generation [Wang *et al.*, 2025a]. However, LLM-generated educational content carries inherent risks of factual hallucination and pedagogically misleading explanations [Hu *et al.*, 2025], underscoring the need for structured evaluation agents and teacher oversight in any generation pipeline.

2.3 Teacher-Facing AI Systems

Multi-agent teacher-facing AI systems have progressed more slowly than student-facing applications. Existing tools typically provide static instructional resources or limited customization, often requiring substantial manual effort. Recent work on LLM-based lesson planning suggests that AI can meaningfully reduce teachers’ planning workload [Hu *et al.*, 2025], yet these systems rarely incorporate explicit learner modeling and differentiation that includes cognitive and motivational aspects, as well as neurodiversity. Notable exceptions are emerging but remain domain-specific. *StoryLab* [Li *et al.*, 2025] demonstrates teacher-in-the-loop personalization for literacy, and *ARISE* [Gajewska *et al.*, 2025] employs multi-agent simulation for teacher training. Neither supports differentiated material generation across diverse learner profiles.

2.4 Synthesis

Existing work demonstrates that effective differentiation requires considering aspects beyond performance alone, including motivational factors and learning differences such as dyslexia and ADHD. Multi-agent architectures provide the foundational infrastructure to enable stable, reusable learner representations, the integration of knowledge bases ensuring curriculum alignment and didactic guidelines, and longitudinal learning as the system accumulates context from materials uploaded throughout the school year. However, teacher-facing systems have received limited attention. We translate the capabilities of agentic AI into a teacher-facing tool that produces curriculum-aligned, accessibility-compliant materials while preserving teacher authority. In close collaboration with schools, we developed FACET to address these gaps.

3 FACET: A Teacher-Facing Multi-Agent Framework

3.1 Design Rationale

FACET adopts a multi-agent framework that addresses the limitations of single LLM approaches through three foundational capabilities. (1) the system maintains stable learner representations through dedicated persona agents that consistently preserve learner characteristics across sessions, reducing the prompt sensitivity and output variability observed in single LLM approaches. (2) the architecture enables the integration of structured knowledge bases, ensuring curriculum alignment and adherence to didactic guidelines in generated worksheets. Therefore, the generator agent is equipped with foundational knowledge, such as the curriculum and evidence-based didactic theories that are embedded in the system and can be extended by teachers through uploading their own teaching materials. (3) the system supports longitudinal context accumulation: as teachers upload materials throughout the year, the system builds contextual knowledge, enabling differentiation.

3.2 Agent Specifications

The system architecture rests on a Python FastAPI backend coupled with a React frontend, ensuring a responsive teacher-in-the-loop experience. To guaranty data privacy and control over the generation process, we employ *GPT-OSS 120B*, an open-source LLM from OpenAI hosted on local infrastructure via Ollama.

FACET utilizes custom logic to orchestrate agent interactions, enforcing a strict pedagogical workflow, as illustrated in Figure 2. Communication between agents follows a pre-defined protocol to ensure stability: The learner agent generates natural language thought traces (thought protocol) of the student working on a task. These traces are passed to the assessment agent, which processes them into a strict JSON diagnostic format defined via a JSON schema in the system prompt. The generator agent enables the creation of the worksheet by receiving both the learner’s thought protocol and the structured diagnostic JSON. In contrast, the evaluator agent functions as an external reviewer; it receives only the generated worksheet and the original learner persona prompt to assess suitability, operating without access to the internal reasoning steps or intermediate diagnostics. Expert knowledge (curriculum standards, didactic guidelines) is integrated via context injection (system prompts), with plans to implement a Retrieval-Augmented Generation (RAG) pipeline to allow for scalable knowledge management in future iterations. To balance creativity with reliability, we apply a temperature of 0.9 to the learner, assessment, and generator agents, while restricting the evaluator agent to 0.1 for consistent diagnostic reasoning. The full generation cycle (simulation, diagnosis, and material creation) for one worksheet completes in approximately 30 seconds.

3.3 Learner Agents: Simulating Diverse Learner Profiles

Teachers create student descriptions that reflect the characteristics they commonly observe in their classrooms. Con-

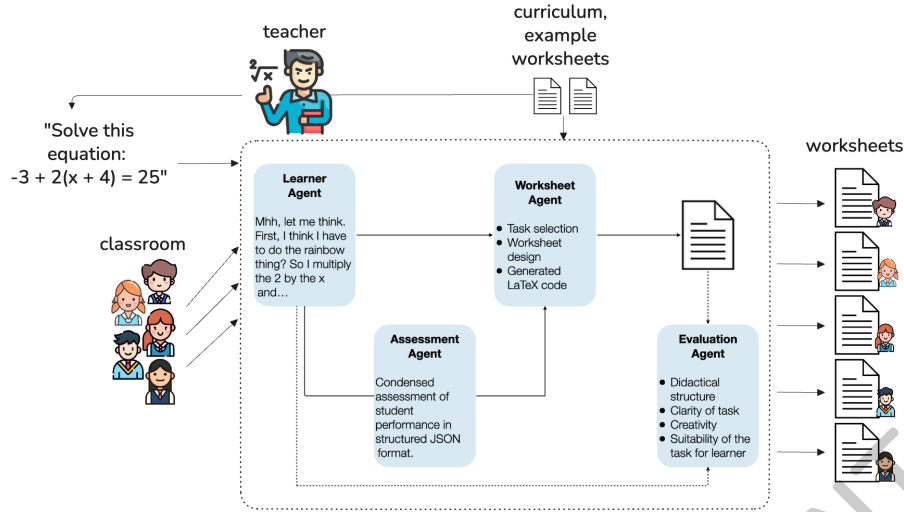


Figure 2: Core workflow: Teacher-defined learner profiles drive multi-agent generation of differentiated materials with human review

sistent with our framing of differentiation beyond performance, these profiles integrate diverse dimensions, such as prior knowledge levels, motivational aspects, reading and writing difficulties, and other cognitive and affective factors. The learner agents simulate student behavior based on teacher-defined characteristics. At the beginning of each process, the learner agent attempts to solve an assigned task while generating a structured transcript that captures reasoning steps, misconceptions, and emotional responses. Tasks can be uploaded by teachers or selected from curriculum-aligned repositories. By leveraging LLMs’ demonstrated capacity to simulate consistent personality traits and affective patterns [Aher *et al.*, 2023; Li *et al.*, 2024], FACET approximates realistic learner behavior as a basis for instructional adaptation. Crucially, a learner agent does not need to perfectly replicate a real student to be pedagogically useful; it must produce plausible differential engagement patterns that surface likely barriers, typical misconceptions, and cognitive load bottlenecks. Simulation fidelity depends critically on prompting specificity, which teacher-defined profiles in FACET are designed to ensure. Additional robustness is built into the architecture itself: rather than feeding the learner agent’s output directly to the material generator, the engagement trace is first passed to an independent diagnostic assessment agent, which structures and filters the simulation output before generation proceeds. The teacher’s final review provides a further safeguard, as all outputs are inspected and approved before classroom use. We emphasize that learner agents are not intended to serve as clinical or diagnostic models. Rather, they function as controllable simulations that test instructional materials against specified constraints (e.g., reduced text density, shorter instructions, and higher structure). To mitigate stereotyping risks, profiles are defined in terms of observable, learning-relevant needs, such as decoding difficulty, working-memory load, or sustained attention demands, rather than diagnostic labels alone.

3.4 Assessment Agent: Diagnostic Reasoning

The assessment agent analyzes simulated learner interactions with instructional materials, examining both cognitive processes and affective responses. Guided by a strict JSON schema provided in the system prompt to ensure structural consistency, it identifies indicators such as conceptual misunderstandings, low confidence, disengagement, or, conversely, fluent performance and high motivation. The resulting structured diagnostic summary provides an explicit foundation for subsequent generation and conditioning material adaptation based on identified needs.

3.5 Generator Agent: Differentiated Worksheets

The generator agent creates differentiated instructional materials grounded in the diagnostics produced by the preceding agents. It is equipped with structured knowledge bases that ensure both curriculum alignment and adherence to didactic principles. Teachers can specify high-level directives, such as instructional style, task selection strategies, and worksheet format, allowing FACET to align with individual pedagogical preferences while the agent contributes design decisions informed by established instructional frameworks embedded in the system or extended through teacher-uploaded materials. Critically, the generator incorporates accessibility guidelines identified in research as essential for neurodivergent learners, such as clear sans-serif fonts, sufficient line spacing, strong text-background contrast, single-column layouts, and explicit visual separation of tasks [Rello and Baeza-Yates, 2013]. Scaffolding strategies, worked examples, stepwise guidance, and partially completed representations are integrated based on Bloom’s revised taxonomy [Anderson and Krathwohl, 2001], a framework for progressing from foundational to complex cognitive tasks to systematically balance challenge and support. The decision to generate structured $\text{\LaTeX}$ code serves a dual purpose: it provides a robust, machine-readable format that ensures precise layout control for accessibility compliance (e.g., specific fonts,

spacing), while simultaneously enabling flexible export options. The system directly renders professional-grade PDFs or converts the source to .docx format, allowing teachers unfamiliar with $\text{\LaTeX}$ to easily edit materials in standard word processors.

3.6 Evaluator Agent: Pedagogical Quality Assurance

The evaluator agent provides structured feedback on generated materials along four pedagogically relevant dimensions aligned with our evaluation framework: Didactic structure, clarity, creativity, and format and design. This evaluation informs iterative system refinement while providing transparent feedback to teachers.

3.7 Teacher-In-The-Loop Interface

Consistent with the teacher-in-the-loop principle, educators retain full control throughout the workflow. Teachers interact with FACET at three points: (1) student profile specification via prompt templates or self-written prompts, (2) upload of materials or task selection from curriculum-aligned repositories, and (3) output review with accept/modify/reject controls. Teachers can decide to export materials as PDF, Word, or $\text{\LaTeX}$. This design reflects our core position: FACET augments teachers by scaling the production of differentiated materials, freeing cognitive and temporal resources for responsive instruction and meaningful student interaction, rather than replacing the contextual knowledge and professional judgment that only educators possess. Figure 2 depicts the workflow.

4 Evaluation Methods

We evaluated FACET through two complementary stakeholder engagements. (1) we conducted exploratory workshops with German school principals who will prospectively deploy the system in their schools, ensuring alignment with authentic institutional needs and workflow constraints. (2) to assess generalizability and broaden quality assessment beyond a single educational context, we surveyed practicing teachers from countries with high-performing education systems: the United Kingdom (UK), the United States (US), Canada, and Australia. While curricular standards differ across these countries, the evaluation dimensions, didactic structure, clarity, format, and creativity, assess general pedagogical quality rather than curriculum-specific content alignment, making cross-national teacher ratings appropriate for this purpose. This dual approach addresses two research questions:

RQ1: To what extent does FACET address teachers’ practical needs and integrate with existing workflows?

RQ2: To what extent can FACET generate pedagogically suitable materials for diverse learner profiles?

4.1 Workshop Study (RQ1)

Through two one-hour workshops with school principals ($N = 30$, $N = 15$ per session) from German schools, we examined practical relevance, usability, and classroom integration. After a tool demonstration, participants engaged in

structured discussions on system structure and usage, interaction possibilities and degrees of freedom, and intended output. Each session was led by one facilitator, with two researchers independently documenting participant responses. Themes were identified through inductive coding by two researchers; discrepancies were resolved through discussion until consensus was reached. Representative quotes in Table 1 are synthesized from these notes.

4.2 Survey Study (RQ2)

Building on the workshop findings, we conducted a quantitative online survey with K–12 in-service teachers recruited via Prolific. We derived tasks from the grade seven mathematics curriculum (area calculation) to ensure ecological validity. Teachers were presented with five example worksheets generated by FACET, each corresponding to distinct learner profiles: Profiles that represent varying combinations of mathematical proficiency and motivational orientation, or profiles that represent learners with dyslexia-related processing needs, specified in terms of observable, learning-relevant characteristics. The selection of student characteristics was informed by discussions with participating teachers during the workshops and reflects learner characteristics commonly observed in heterogeneous classrooms. All profiles used gender-neutral language to reduce potential confounding effects.

Participants were asked to review all five worksheets and evaluate them on a six-point Likert scale (1 = very low quality, 6 = very high quality). Each worksheet was rated along (1) didactic structure: How well is the worksheet structured from an educational perspective (e.g., support, sequence of tasks), (2) clarity of task formulation: How comprehensible are the tasks and the language?, (3) format and design: How appealing, clear, and reader-friendly is the worksheet? Does the layout support motivation and working with the material?, and (4) creativity: How original, motivating, and varied is the design of the tasks compared to typical teaching materials. For worksheet five, additional items addressed font readability, language simplification, and spacing/layout. Survey data was analyzed descriptively by computing means and standard deviations for each evaluation dimension across worksheets and learner profiles.

5 Results

This section presents findings from our mixed-methods evaluation, addressing the two RQs.

5.1 Workshop Findings (RQ1)

The thematic analysis of workshop sessions ($N = 30$ school principals) identified three recurring themes: practical relevance, usability and teacher-oriented design, and classroom integration and use cases. Table 1 summarizes the findings. Participants emphasize the strong need for AI-based differentiation support, noting that inclusive classrooms increasingly require addressing diverse needs (ADHD, learning difficulties, social-emotional challenges) that exceed individual teacher capacity. School principals highlighted the pedagogical value of supporting both downward and upward differenti-

<i>Practical Relevance</i>	
Differentiation	We need both scaffolding for struggling learners and enrichment for advanced students. But realistically, when time runs out, we end up just giving extra worksheets or moving students to the next topic, what we know that's not ideal, but what else can we do?
Teacher Workload	I have 34 students with completely different needs—ADHD, autism, learning difficulties, emotional challenges. With the teacher shortage, we're stretched thin. Differentiation isn't optional anymore, but it's impossible to sustain without help.
Alignment	If the AI generates tasks that don't match our curriculum or miss the right difficulty level, I won't use it. The materials have to be something I could actually hand out in class tomorrow.
Accessibility	For my students with dyslexia, formatting matters enormously. Clear layouts, less text on the page, explicit structure. These aren't nice-to-haves, they're essential for those kids to even access the content.
<i>Usability and Teacher-Oriented Design</i>	
Teacher-in-the-Loop	I need to stay in control. If the system generates something and I can't review it, change it, or throw it out—I won't trust it. I'm the one who knows my students.
Customization	My worksheets have a certain look; e.g. my symbols, my structure. Students know what to expect. Any new tool has to fit into that, not force me to start from scratch.
Configurable Learners	I don't want to write prompts from scratch every time. Give me templates I can adjust—pre-set profiles for different performance levels that I can tweak based on what I know about my class.
<i>Classroom Integration and Use Cases</i>	
Instructions	I need materials I can use flexibly, for the whole class, for group work, for individual practice. And please, automatic answer keys. I don't have time to create those separately.
Transparency of AI Reasoning	I like seeing the different learner personas. This makes me think about why a task is designed a certain way. But we have to be careful not to reduce students to stereotypes.
Generalization	This would be incredibly useful in other subjects too, e.g., languages, history. And we need profiles for ADHD students, not just for dyslexia. The more flexible, the better.

Table 1: Key themes from exploratory workshops with school principals ($N = 30$). Representative quotes synthesized from two exploratory workshops with German school principals.

ation, the importance of curriculum alignment and accessibility, and the necessity of reducing teacher workload amid conditions of teacher shortages and diverse learner needs. Accessible formatting for dyslexia, such as using clear layouts and reduced text density, was identified as critical. The teacher-in-the-loop design was viewed positively, as it preserves professional judgment, a prerequisite for trust and adoption.

5.2 Survey: Material Quality Assessment (RQ2)

Seventy K–12 teachers ($M_{\text{age}} = 40.5$, $SD = 9.5$; 73% female) evaluated five example worksheets generated by FACET. The majority of participants resided in the United Kingdom (57.14%) or the United States (34.29%), with smaller proportions from Canada (7.14%) and Australia (1.43%). Regarding school type, approximately half of the participants taught at primary schools (48.57%), while 42.86% taught at secondary schools, and 77.14% had experience teaching students aged 11–13 either currently or within the past two years, which reflects expertise in the age group for which the example worksheets were generated. Notably, participants demonstrated high openness toward AI integration, with 91.43% expressing a willingness to use AI for lesson preparation and 70.0% already incorporating AI into their lesson preparation practices. Table 2 presents the descriptive statistics for the worksheet evaluation.

Core quality dimensions. Didactic structure and clarity received consistently favorable ratings, exceeding the scale midpoint across worksheets for profiles 1–4 (range: 3.57 – 4.50). Worksheets targeting low-motivation learners (W1,

W2) achieved the highest ratings for didactic structure ($M = 4.41 - 4.50$) and clarity ($M = 4.23 - 4.46$), indicating effective scaffolding and motivational support.

Enrichment-oriented materials. Worksheet 3 (high-performing, highly motivated) received lower ratings across all dimensions (structure: $M = 3.87$; clarity: $M = 3.57$; format: $M = 3.10$; creativity: $M = 3.30$). This pattern suggests that the framework currently generates scaffolded content more effectively than enrichment-oriented materials for advanced learners.

Accessibility adaptation. The dyslexia-adapted worksheet (W5) yielded mixed results. Spacing/layout ($M = 4.07$) and language simplification ($M = 3.94$) exceeded the midpoint, indicating alignment with accessibility guidelines. However, font readability ($M = 3.30$) and overall suitability for reading difficulties ($M = 3.46$) fell near the midpoint, revealing that typographical requirements for dyslexia can be improved. This finding identifies a specific improvement target for the generation pipeline: while structural and linguistic adaptations are partially effective, typographical rendering requires targeted refinement.

Format and creativity. Format/design (range: 3.10 – 3.79) and creativity (range: 3.30 – 3.66) received moderate ratings near the midpoint. As noted in the survey instructions, teachers were asked to focus on pedagogical content rather than visual appearance, since the system allows teachers to upload their own example materials to customize the format and style of generated outputs. Nevertheless, these

Criterion	<i>M</i>	<i>SD</i>	<i>Mdn</i>
<i>WS 1: High Performance, low Motivation</i>			
Didactic Structure	4.50	1.02	5
Clarity of Task Formulation	4.46	1.05	5
Format and Design	3.79	1.36	4
Creativity	3.66	1.27	4
<i>WS 2: Low Performance, low Motivation</i>			
Didactic Structure	4.41	1.06	4
Clarity of Task Formulation	4.23	1.16	4
Format and Design	3.63	1.35	4
Creativity	3.50	1.41	4
<i>WS 3: High Performance, high Motivation</i>			
Didactic Structure	3.87	1.17	4
Clarity of Task Formulation	3.57	1.20	4
Format and Design	3.10	1.37	3
Creativity	3.30	1.34	3
<i>WS 4: Low Performance, high Motivation</i>			
Didactic Structure	4.03	1.20	4
Clarity of Task Formulation	3.80	1.19	4
Format and Design	3.29	1.46	3.5
Creativity	3.40	1.36	3
<i>WS 5: Dyslexia Profile</i>			
Suitability for Reading Difficulties	3.46	1.47	3.5
Font Readability	3.30	1.61	3
Language Simplification	3.94	1.40	4
Spacing and Layout	4.07	1.39	4

Table 2: Descriptive Statistics for Worksheet Evaluation $N = 70$. Items rated on a 6-point scale (1 = *very low*, 6 = *very high*).

dimensions represent opportunities for enhancement.

Overall, worksheets received favorable ratings, with most criteria scoring around $M = 4$ out of 6 and some reaching $M = 5$ out of 6, particularly for didactic structure and clarity of task formulation. However, results revealed specific areas for improvement: worksheets for high-motivation profiles (both low and high performance) received lower ratings for creativity and format/design, suggesting a need for more engaging and varied materials for motivated learners. The dyslexia-adapted worksheet showed room for improvement in font readability and suitability for reading difficulties, indicating that typographical adaptations require further refinement to meet accessibility guidelines. We are currently transitioning the dyslexia rendering path from standard L^AT_EX to XeLaTeX with OpenDyslexic and other validated dyslexia-friendly typefaces, combined with explicit character-spacing and line-height parameters, to close this gap before classroom deployment.

6 Limitations and Future Work

Our evaluation focused on grade-7 mathematics, however, the core four-agent pipeline is subject-agnostic, operating on teacher-defined profiles and uploaded materials rather than hard-coded subject rules. Learner profile dimensions, performance level, motivation, and neurodivergent characteristics apply equally to other subjects. Extending FACET to new subjects requires populating the knowledge base with domain-specific curriculum standards and task templates via the planned RAG pipeline, rather than architectural changes.

Our longitudinal deployment spans multiple subjects and grade levels in order to provide ecologically valid data on generalization. Furthermore, agents simulate behavior based on LLM simulation fidelity, which has not yet been validated against authentic student responses in our specific deployment context. To address this, future design iterations will incorporate nudges within the teacher-in-the-loop interface that actively prompt educators to cross-check simulated learner traces against their direct knowledge of students, for example, flagging implausible reasoning patterns or inviting teachers to annotate discrepancies between simulated and observed behavior. On the evaluation side, we are planning systematic comparisons of simulated learner outputs against existing student response datasets and authentic classroom data collected during longitudinal deployment:

School partnerships. We have established collaborations with 10 German primary, secondary, and inclusive schools, where five to ten teachers per school will integrate FACET into regular instruction over one semester, enabling evaluation of workflow integration, teacher trust, and student feedback.

Embedded validation mechanisms. Within deployment, teachers will systematically compare simulated learner behavior against actual student responses, flagging implausible patterns and documenting divergences. This practitioner-driven validation grounds simulation refinement in classroom reality.

Student feedback. The diverse school sample spans socioeconomic contexts, enabling an analysis of how different students perceive material quality.

7 Conclusion

We present FACET, a teacher-facing multi-agent system that addresses a pressing societal challenge: enabling differentiated instruction for heterogeneous classrooms where teachers face high workload demands. A four-agent architecture, consisting of learner simulation, diagnostic reasoning, material generation, and evaluation, operationalizes differentiation that incorporates aspects beyond performance while preserving educator authority. Our mixed-methods evaluation confirms that FACET meets genuine classroom needs. School principals validated the framework against real constraints, emphasizing teacher-in-the-loop control, configurable profiles, and curriculum alignment. Teachers rated scaffolding-oriented materials as classroom-ready, with a good didactic structure and clarity, though enrichment materials for advanced learners and dyslexia-specific formatting require further refinement. Generalizability across subjects and grade levels and the systematic validation of simulated learner behavior against real student responses remain important directions that our ongoing longitudinal deployment across partner schools will address. Developed through sustained collaboration with educational stakeholders, FACET demonstrates that AI can meaningfully support differentiated instruction when designed around teachers’ needs.

Ethical Statement

FACET use requires clear human oversight at all stages. All study data is anonymized/pseudonymized, and participation requires informed consent; participation is voluntary and appropriately compensated. We acknowledge that persona mimicking could be misunderstood. To address this, personas are treated as analytic abstractions rather than representations of real individuals. All outputs are labeled as AI-generated, and teachers retain full control to edit, reject, or withhold materials. Nevertheless, our commitment to data privacy is reinforced by a fully local infrastructure, while the framework is accessible via the web to international users. To maintain full data sovereignty, the system does not utilize third-party cloud providers or proprietary LLM APIs; all data is stored and processed exclusively on our self-hosted institutional servers and remains within our local network. It is not yet empirically established how students will perceive AI-adapted worksheets in real classrooms, including whether differentiation could feel stigmatizing, isolating, or like “special treatment.” Therefore, student-facing use is conditional on rigorous evaluation that includes psychosocial outcomes such as perceived fairness, autonomy, belonging, and self-esteem, not only performance.

Acknowledgements

Research reported in this paper was supported through the Research Campus Modal funded by the German Federal Ministry of Research, Technology and Space (BMFTR) (fund numbers 05M14ZAM, 05M20ZBM), by the German Federal Ministry of Research, Technology and Space (BMFTR), grant number 16DII133 (Weizenbaum Institute) and by the Zuse Institute Berlin via RISE@ZIB services for hosting the LLM models.

References

- [Aher *et al.*, 2023] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *ICML’23: Proceedings of the 40th International Conference on Machine Learning*, volume 17 of *ICML’23*, pages 337–371, Honolulu, Hawaii, USA, 2023. JMLR.org.
- [Anderson and Krathwohl, 2001] Lorin W. Anderson and David R. Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.
- [Bernacki *et al.*, 2021] Matthew L. Bernacki, Meghan J. Greene, and Nikki G. Lobczowski. A Systematic Review of Research on Personalized Learning: Personalized by Whom, to What, How, and for What Purpose(s)? *Educational Psychology Review*, 33:1675–1715, 2021.
- [Carroll *et al.*, 2025] Julia M. Carroll, Caroline Holden, Philip Kirby, Paul A. Thompson, Margaret J. Snowling, and the Dyslexia Delphi Panel. Toward a consensus on dyslexia: findings from a Delphi study. *Journal of Child Psychology and Psychiatry*, 66(7):1065–1076, July 2025.
- [Danielson *et al.*, 2024] Melissa L. Danielson, Angelika H. Claussen, Rebecca H. Bitsko, Samuel M. Katz, Kimberly Newsome, Stephen J. Blumberg, Michael D. Kogan, and Reem Ghandour. ADHD Prevalence Among U.S. Children and Adolescents in 2022: Diagnosis, Severity, Co-Occurring Disorders, and Treatment. *Journal of Clinical Child & Adolescent Psychology*, 53(3):343–360, 2024.
- [Deci and Ryan, 2000] Edward L Deci and Richard M Ryan. The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychol. Inq.*, 11(4):227–268, October 2000.
- [Fariani *et al.*, 2023] Rida Indah Fariani, Kasiyah Junus, and Harry Budi Santoso. A Systematic Literature Review on Personalised Learning in the Higher Education Context. *Technology, Knowledge and Learning*, 28(2):449–476, 2023.
- [Gajewska *et al.*, 2025] Ewelina Gajewska, Michal Wawer, Katarzyna Budzynska, and Jaroslaw A. Chudziak. Leveraging a Multi-agent LLM-Based System to Educate Teachers in Hate Incidents Management. In Alexandra I. Cristea, Erin Walker, Yu Lu, Olga C. Santos, and Seiji Isotani, editors, *Lecture Notes in Computer Science()*, volume 15881, pages 332–339, Cham, 2025. Springer Nature Switzerland.
- [Gaye *et al.*, 2024] Fatou Gaye, Nicole B. Groves, Elizabeth S. M. Chan, Alissa M. Cole, Emma M. Jaisle, Elia F. Soto, and Michael J. Kofler. Working memory and math skills in children with and without ADHD. *Neuropsychology*, 38(1):1–16, January 2024.
- [Gonnermann-Müller *et al.*, 2026a] Jana Gonnermann-Müller, Jennifer Haase, Nicolas Leins, Thomas Kosch, and Sebastian Pokutta. LLM-Based Educational Simulation: Evaluating Temporal Student Persona Stability Across ADHD Profiles, 2026.
- [Gonnermann-Müller *et al.*, 2026b] Jana Gonnermann-Müller, Jennifer Haase, Nicolas Leins, Thomas Kosch, and Sebastian Pokutta. Maintaining Stable Personas? Examining Temporal Stability in LLM-Based Human Simulation. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–6, Barcelona, Spain, April 2026. ACM.
- [Haase and Pokutta, 2025] Jennifer Haase and Sebastian Pokutta. Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research, October 2025. arXiv:2506.01839 [cs].
- [Habib and Giraud, 2013] Michel Habib and Kimberly Giraud. Chapter 23 - Dyslexia. In Olivier Dulac, Maryse Lassonde, and Harvey B. Sarnat, editors, *Pediatric Neurology Part I*, volume 111 of *Handbook of Clinical Neurology*, pages 229–235. Elsevier, 2013.
- [Hu *et al.*, 2025] Bihao Hu, Jiayi Zhu, Yiyang Pei, and Xiaoping Gu. Exploring the potential of LLM to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7, February 2025.

- [Huang *et al.*, 2026] Muhua Huang, Xijuan Zhang, Christopher Soto, and James Evans. Designing AI-Agents With Personalities: A Psychometric Approach. *Personality Science*, 7, 2026.
- [Ilkou *et al.*, 2025] Eleni Ilkou, Thomai Alexiou, Grigoris Antoniou, and Olga Viberg. Dyslexia and AI: Do Language Models Align with Dyslexic Style Guide Criteria? In Alexandra I. Cristea, Erin Walker, Yu Lu, Olga C. Santos, and Seiji Isotani, editors, *Artificial Intelligence in Education*, pages 32–47, Cham, 2025. Springer Nature Switzerland.
- [Iseman and Naglieri, 2011] Jackie S. Iseman and Jack A. Naglieri. A Cognitive Strategy Instruction to Improve Math Calculation for Children With ADHD and LD: A Randomized Controlled Study. *Journal of Learning Disabilities*, 44(2):184–195, 2011.
- [Khan Academy, 2025] Khan Academy. Khan Academy, 2025.
- [Li *et al.*, 2024] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. Quantifying AI Psychology: A Psychometrics Benchmark for Large Language Models, 2024.
- [Li *et al.*, 2025] Zhaohui Li, Feiwen Xiao, Jiaju Lin, Xiaohan Zou, Qingxiao Zheng, and Jinjun Xiong. StoryLab: Empowering Personalized Learning for Children Through Teacher-Guided Multimodal Story Generation. In Alexandra I. Cristea, Erin Walker, Yu Lu, Olga C. Santos, and Seiji Isotani, editors, *Lecture Notes in Computer Science()*, volume 15881, pages 285–292, Cham, 2025. Springer Nature Switzerland.
- [Mannekote *et al.*, 2025] Amogh Mannekote, Adam Davies, Jina Kang, and Kristy Elizabeth Boyer. Can LLMs Reliably Simulate Human Learner Actions? A Simulation Authoring Framework for Open-Ended Learning Environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29044–29052, 2025.
- [Rello and Baeza-Yates, 2013] Luz Rello and Ricardo Baeza-Yates. Good fonts for dyslexia. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–8, New York, NY, USA, 2013. Association for Computing Machinery.
- [Schummel *et al.*, 2025] Philip Schummel, Malte Teichmann, and Jana Gonnermann-Müller. Specifying ten roles of using ChatGPT in secondary education: a teacher’s perspective. In *Thirty-Third European Conference on Information Systems (ECIS 2025)*, Amman, Jordan, 2025.
- [SquirrelAI, 2025] SquirrelAI. Squirrel AI, 2025.
- [Wang *et al.*, 2025a] Tianfu Wang, Yi Zhan, Jianxun Lian, Zhengyu Hu, Nicholas Jing Yuan, Qi Zhang, Xing Xie, and Hui Xiong. LLM-powered Multi-agent Framework for Goal-oriented Learning in Intelligent Tutoring System. In *WWW ’25: Companion Proceedings of the ACM on Web Conference 2025*, pages 510–519, New York, NY, USA, 2025. Association for Computing Machinery. arXiv:2501.15749 [cs].
- [Wang *et al.*, 2025b] Yilei Wang, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(1):519, 2025.
- [World Health Organisation (WHO), 2025] World Health Organisation (WHO). International Classification of Diseases (ICD), 2025.
- [Wu *et al.*, 2025] Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. Embracing Imperfection: Simulating Students with Diverse Cognitive Levels Using LLM-based Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9887–9908, Vienna, Austria, 2025. Association for Computational Linguistics.
- [Zhang *et al.*, 2025] 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. Simulating Classroom Education with LLM-Empowered Agents. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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, 2025. Association for Computational Linguistics.