IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/47AI/2026/4723026/
Vol. 2026 (12), Article ID 4723026

Personalized Learning via Generative AI: A Profile-Aware Adaptive RAG Framework for Neurodiverse Student Support*

Bassem Khaled ELGINDY, Abdelfatah HEGAZI and Nermeen MAGDY

Arab Academy for Science and Technology
& Maritime Transport College of Computing & Information Technology
Heliopolis, Cairo, Egypt

Correspondence should be addressed to: Bassem Khaled ELGINDY, bassem.elgindy@icloud.com

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

While large language models can produce coherent explanations of educational concepts, they are not necessarily consistent in maintaining learner-specific accessibility requirements when asked repeatedly. This can pose a problem for neurodiverse learners who might need literal language, spacing, structured steps, anchors for key words or a lessening of the writing load. Current educational AI systems and Retrieval-Augmented Generation systems typically aim at providing factual support or general tutoring, with less consideration of retrieving profile-based rules to guide the presentation of an explanation.

This paper presents a Retrieval-Augmented Generation framework for personalized learning support based on behavior. The system fetches the learner-profile constraints for Autism Spectrum Disorder, dyslexia, and motor-writing support for dysgraphia and conditions educational responses based on these constraints. The prototype features a Flask backend, a web interface, a JSON learner-profile knowledge base, optional vector retrieval, orchestration with OpenAI/Gemini/DeepSeek, auto_eval_v4 scoring, dashboard logging, and human feedback collection.
The prototype design was used in the evaluation by a combination of mixed methods. Adaptive RAG responses had a mean CompositeScore of 87.01, while baseline responses had a mean CompositeScore of 57.84, in the locked 578-row analysis snapshot. The mean RAG gain for the 262 matched model runs was 27.83 points. Human feedback had 23 rows with a mean rating of 4.78/5 and 100.0% helpful flags. Formative educational-suitability feedback was provided by four specialist reviews. Results show prototype level output alignment for accessibility, not clinical diagnosis or therapeutic validation.

Keywords: Retrieval-Augmented Generation; neurodiversity; dyslexia; autism.

Introduction

Generative artificial intelligence is being used more and more in education, as it can generate explanations, examples, summaries, quizzes, and learning scaffolds for a wide range of subjects. But, fluency in language generation does not necessarily mean accessibility of an educational response. Even if a response is factually correct, it can be challenging for a learner to use if it is visually dense, highly implicit, metaphorical, poorly structured, or has a high written output requirement. This is particularly relevant for neurodiverse learners such as learners with Autism Spectrum Disorder (ASD), dyslexia and motor-writing difficulties associated with dysgraphia.

Cite this Article as: Bassem Khaled ELGINDY, Abdelfatah HEGAZI and Nermeen MAGDY Vol. 2026 (12) "Personalized Learning via Generative AI: A Profile-Aware Adaptive RAG Framework for Neurodiverse Student Support " Communications of International Proceedings, Vol. 2026 (12), Article ID 4723026, https://doi.org/10.5171/2026.4723026

Neurodiverse learners may need adaptations to the explanation, but not to the academic level of the content. Literal language, predictable sequencing, direct explanations of hidden meaning, and less ambiguity, for example, may be helpful for ASD-related profiles. Short lines, spacing, keyword anchors, simplified layout and reduced visual crowding may be helpful for dyslexia-related profiles. Sentence starters, fill-in-the-blank structures, quick-response options, and less typing may benefit learners with motor-writing difficulties. These adaptations are not a reduction of the educational objective. Rather, they seek to convey the same educational message in a more learner-friendly way.

Reliable solution to the problem is not provided by generic interaction with large language models. A model can be manually prompted to use short lines, to not use idioms, or to simplify the explanation, but this requires repeated prompting by a learner or teacher. Manual prompting is unreliable, cumbersome and can be forgotten between sessions. The fine-tuning is also not optimal for this use case as educational accommodations should be transparent, editable and flexible to accommodate changing learner needs. A better way is to keep the rules for supporting the learner outside the model and access them during run-time.

This paper presents a Retrieval-Augmented Generation (RAG) approach for personalized educational support based on behavior. Traditional RAG is typically employed to fetch factual documents or evidence to enhance the content of an answer. The proposed framework, on the other hand, fetches learner-profile constraints that govern the structure and delivery of the answer. Information retrieved can include rules concerning literal wording, length of lines, visual spacing, step-by-step scaffolding, emphasis of key words, language preference, and burden of response. In this way, the system not only retrieves what to teach, it also retrieves how to teach.

The prototype developed and implemented includes a Flask backend, a web-based learner interface, a JSON learner-profile knowledge base, optional vector retrieval, adaptive prompt construction, multimodel execution with OpenAI, Gemini, and DeepSeek, automatic scoring, dashboard logging, and human feedback collection. The system takes in a learner identifier and task, fetches the corresponding profile rules, constructs an adaptive prompt, generates baseline and Adaptive RAG responses, scores the responses using auto_eval_v4, and stores the evidence for further analysis. This makes personalization more repeatable and auditable than a standalone chatbot interaction.

A mixed-method design is used to evaluate the prototype. The quantitative evidence is a snapshot of a locked dashboard_log.csv file with 578 rows of data, including baseline and Adaptive RAG outputs. Further evidence is provided through matched model comparison, human feedback, specialist review forms, screenshots and dashboard traces. The aim of the evaluation is to look at output alignment for accessibility, not clinical effectiveness or long-term learning outcomes.

This paper has three contributions. First, it introduces a profile-aware adaptive RAG architecture, which does not just retrieve factual content but also rules that support the learner. Second, it introduces multimodel orchestration, enabling the system to compare providers and choose the most powerful personalized response for a specific task, which involves using both OpenAI and Gemini. Third, it offers a clear prototype level assessment with automated scores, feedback and formative specialist review. The system is designed as an assistive educational prototype only and not as a diagnostic, therapeutic or clinical decision making tool.

Table 1. Summary of contributions.

Contribution	Description
Behavior-conditioned RAG	Retrieves learner-profile teaching constraints instead of only factual content.
Multi-model orchestration	Runs baseline and RAG outputs across OpenAI, Gemini, and DeepSeek and selects the best personalized response.
Transparent evaluation	Logs CompositeScore, RAG gain, latency, provider, mode, and feedback.
Mixed-method design	Combines system logs, survey means, human feedback, qualitative cases, and completed formative specialist review.
Ethical boundary	Frames the system as educational support, not diagnosis or treatment.

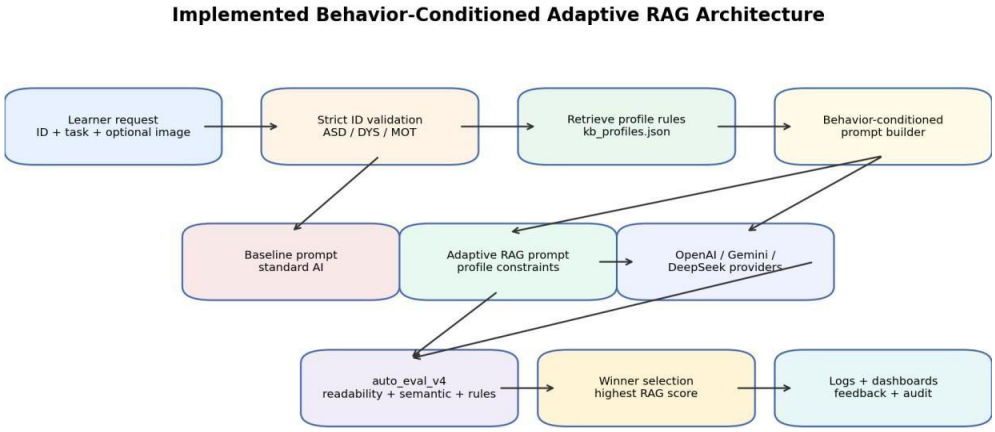

Figure 1. Implemented behavior-conditioned adaptive RAG architecture.

Related Work and Research Gap

The latest research on AI in inclusive education emphasizes the benefits of AI-assisted learning, including personalized learning, communication support, accessible content generation, feedback, and learning analytics. Scoping and systematic reviews also identify the same systems as having ethical, privacy, bias, transparency and human oversight concerns (Pagliara et al., 2024; Panjwani-Charania & Zhai, 2024). These issues are applicable for AI systems for neurodiverse learners as personalisation can include personal learner characteristics. AI research for dyslexia has mostly focused on screening, diagnosis or prediction. Eye-tracking, speech features, questionnaires, computer games, multimodal learning features, or deep-learning classifiers have been used in current research to detect dyslexia risk or reading difficulties (Alluhaidan et al., 2024; Santhiya & KanimozhiSelvi, 2023; Yap et al., 2025). These are significant as they encourage early detection and awareness. But detection does not necessarily lead to teaching accommodation. Research in the field of educational AI and autism spectrum disorder has focused on communication support, social inference, routine, predictability and structured learning environments. This research is relevant for LLM education because a general explanation might involve metaphors, implicit meaning, social inferences, or long background paragraphs that might not be appropriate for ASD-related learner profiles. Therefore, a behavior-conditioned generation system should be able to support literal language, structure, meaning clarification and ambiguity reduction. Recent studies on human-computer interaction suggest that LLMs can be beneficial for neurodivergent users, but also that they can have problems with overly neurotypical, inconsistent, wordy, metaphorical, and uncontrollable responses (Carik et al., 2025). This suggests the need for explicit control. The constraints should not be solely based on the learner repeatedly prompting the LLM to generate neurodivergent-accessible explanations. RAG provides an architectural way of connecting

Language models to external memory. In knowledge-intensive NLP, RAG is used to retrieve information to improve factuality and prevent over-reliance on the model's parametric memory (Lewis et al., 2020). The same architectural mechanism can be used in the context of personalised education from retrieving what to teach to retrieving how to teach.

The information to be retrieved is not just facts, but also constraints to support learners, such as visual spacing, literal wording, reduced ambiguity, step-by-step sequencing, and reduced writing load. Review of Literature. Table 2 provides a comparison of this work with other studies on inclusive AI, dyslexia support, neurodivergent use of LLMs, RAG, and model comparison. One thing that is common across all of these areas is that a multi-model RAG architecture is not implemented and auditable to retrieve learner-support behavior rules, but only facts.

Table 2. Summary of previous research and research gap positioning.

Literature stream	Main contribution	Unresolved gap addressed in this paper
AI in inclusive education	AI can support accessibility, assistive communication, adaptive educational content, and learner analytics.	Need implemented systems with transparent learner adaptation, privacy boundaries, and human oversight.
AI for learning disabilities	Reviews classify adaptive learning, tutoring systems, chatbots, and assistive tools.	Need runtime profile-conditioned generation rather than general classification of tools.
Dyslexia AI and detection	Uses ML, eye-tracking, speech, games, and multimodal features to detect risk.	Detection does not automatically generate accessible instructional explanations.
Dyslexia readability	Shows importance of spacing, line length, font choice, text density, and visual presentation.	Accessibility rules are rarely connected to LLM generation through learner profiles.
ASD educational communication	Highlights predictability, routine, literal language, explicit meaning, and reduced ambiguity.	Need runtime constraints that convert figurative or implied language into direct explanation.
Neurodivergent LLM use	Shows usefulness but also risks of verbosity, inconsistency, metaphor, and control difficulty.	Motivates behavior-conditioning instead of repeated manual prompting.
RAG and LLM grounding	Retrieves external information to improve factual grounding and memory update.	Traditional RAG retrieves knowledge content; this work retrieves learner-support behavior.
Universal Design for Learning	Provides a foundation for reducing barriers and offering multiple means of representation.	Needs executable AI pipeline for profile-conditioned generation.
Multi-model comparison	Model behavior varies across providers and tasks.	Supports per-request orchestration, benchmarking, and logging.
Expert-informed review	Professional judgment can assess clarity, accessibility, and educational suitability.	Needs careful boundary as formative educational feedback, not clinical validation.

Proposed Model

This section presents the proposed behavior-conditioned adaptive RAG model. The model is not a diagnostic or clinical decision making device. Instead, it is an educational assistive prototype that extracts the learner-support rules from a profile knowledge base and conditions the format, language and structure of educational explanations. Traditional RAG usually fetches factual statements to ground the content, but the proposed model fetches behavioral and pedagogical constraints, such as literal wording for ASD-related profiles, reduced visual density for dyslexia-related profiles, scaffolded response structures, and reduced written-output burden for motor-support profiles.

The prototype is a pipeline. The user enters a student ID, a task description and (optionally) an image. The backend verifies the identifier, fetches the profile, creates a prompt based on the profile, calls multiple LLM providers, rates the results, chooses the best result, sends the result back to the user interface, and logs the rows of all the models. The solution comprises of a user interface, backend, knowledge base, scoring engine, dashboards and feedback database. kb_profiles.json is the knowledge base of profiles. A condition, strategy_id, ASD support level, dyslexia focus, motor support level, age, grade band, language preference and clinician template identifier can be included in a profile. Rules are associated with the strategy_id. These rules are human-readable, and can be viewed and edited without retraining.

Optional vector search can be done via Chroma, but deterministic JSON search is the default. The validation of ID is deliberately stringent. ASD-, DYS-, and MOT- are allowed. Anonymous IDs are not allowed and are logged to a security log. This is in line with the research purpose of the system: adaptation should be via a profile rather than any free labels. If a new ID is not already in the knowledge base, then it is given a default ASD, DYS or MOT strategy using prefix fallback. The primary behaviour conditioning tool is the prompt builder. It starts with general restrictions, such as beginning with facts, restricting the number of items, and retaining important technical words in the task. It then appends prefix specific rules. DYS prompts are bullet points, double line breaks, bold technical terms, short definitions, example labels, and low crowding. ASD encourages the use of numbered literal steps, command verbs, clarification of meaning, sequence markers, idiom translation, and reasons. MOT encourages the use of fill-in-the-blanks, rapid-response buttons and mind map-like micro-lines. The language of choice is English or Arabic. This is very important for the Egyptian university system as accessible educational AI should support the use of local languages and R2L display. Clinician template IDs allow for professional verification, but not clinical validation. The

LLM layer is powered by OpenAI, Gemini, and DeepSeek. Each provider gives a response to the task and a RAG response to the behaviour conditioned prompt. RAG CompositeScore chooses personalised responses, making the provider decision an empirical runtime choice. Provider-specific routes are provided to support images. Vision-capable paths can be used to provide images to OpenAI and Gemini. The current version of DeepSeek is text-only, so the system utilizes a vision-capable bridge to describe the image and then passes on to DeepSeek. This enables the architecture to compare providers across modalities.

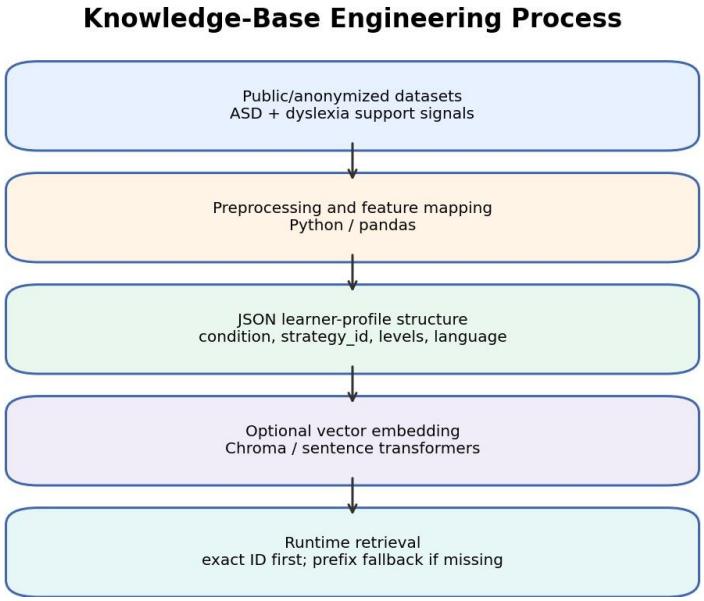

Figure 2. Knowledge-base engineering process.

Table 3. Implementation artifacts.

Component	File / artifact	Function
Backend API	app.py	Validation, retrieval, prompt construction, LLM calls, scoring, analytics routes.
Frontend	index.html	Student UI, task entry, model selection, image upload, voice input, feedback, and Dyslexia Guided Reading pilot.
Knowledge base	kb_profiles.json	Learner profiles, strategy rules, templates, prefix fallback support.
Scoring	scoring.py	auto_eval_v4 CompositeScore and submetrics.
Vector store	vector_db_manager.py	Optional Chroma profile ingestion and lookup.
Dashboards	analytics_dashboard.py, benchmarking_dashboard.py, dashboard.html	Score, latency, feedback, model comparison, and fairness/winner views.
Logs	dashboard_log.csv, human_feedback.csv	Quantitative and qualitative evaluation evidence.

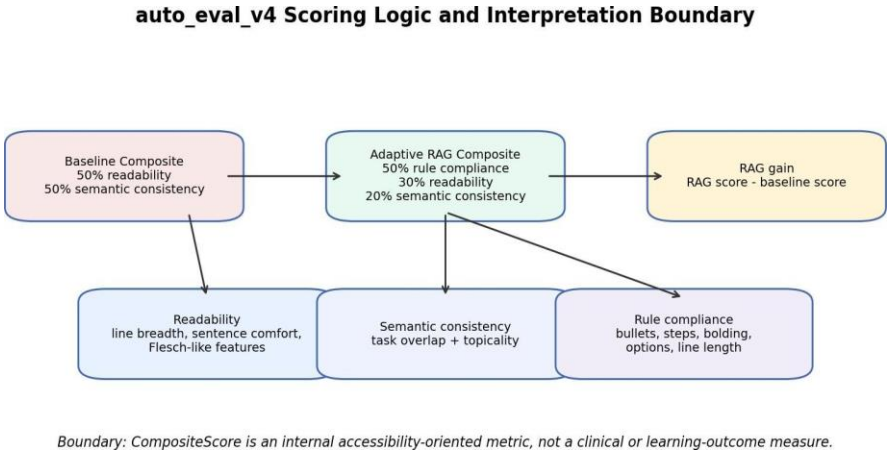

Figure 3. auto_eval_v4 scoring logic and interpretation boundary.

Expert-Informed Rule Refinement

Four completed specialist review forms indicated that the system was educationally promising, while also showing that the profile rules needed clearer differentiation between ASD and dyslexia. This feedback led to changes in both the backend rule-selectio, but also that the profile rules needed clearer differentiation between ASD and dyslexia. This feedback led to changes in both the backend rule-selection logic and the knowledge base. The review forms are treated as educational suitability feedback only; they do not constitute diagnosis, treatment, or clinical validation.

In app.py, dyslexia focus auto-resolution was expanded so that strategy_id values can resolve visual, RAN/fluency, attentional, phonological, surface, and double-deficit patterns. ASD scaffolding was strengthened by adding explicit figurative-to-literal conversion for Level 1, predictable sequence words for Level 2, and low-sensory wording for Level
3. The ASD prompt branch now includes an explicit rule for idiom and implied-language conversion.

In kb_profiles.json, the strategy library was expanded with ASD social-pragmatic, sensory-lowload, and executivefunction strategy families. Dyslexia strategy families were expanded with phonological, surface, RAN, attentional, and double-deficit support. The DYS prompt room was also increased from the common 60-word default to 95 words where needed, because expert feedback suggested that very short responses can be visually clean but pedagogically underexplained.

A frontend Dyslexia Guided Reading pilot was added. It splits Adaptive RAG text into three-word chunks, unlocks one chunk at a time, and uses browser speech recognition to let the learner read the current chunk before revealing the next one. This is an assistive UX feature; it does not change scoring logic or diagnose reading ability.

Table 4. Differentiated ASD, DYS, and MOT rule families.

Profile family	Core rules after expert-informed refinement	Why it differs
ASD	Numbered steps; literal command verbs; explicit idiom/social meaning conversion; predictable sequence markers; low-sensory wording for higher support levels.	Focuses on meaning clarity, ambiguity reduction, predictability, and social-pragmatic interpretation.
DYS	Bulleted lines; blank line between bullets; bold keyword anchors; short definitions; two labeled examples; low visual crowding; guided reading chunks.	Focuses on visual stability, decoding load, spacing, line length, fluency, and reading confidence.
MOT	Fill-in-the-blank summaries; quick-response options; voice input; micro-actions; mind-maplike structure.	Focuses on reducing written-output burden and helping the learner start the task.

Evaluation Methodology

The assessment is mixed method since adaptive educational quality is multidimensional. A response may be readable but semantically incorrect, semantically correct but not readable, or technically correct but pedagogically inappropriate. The quantitative strand is repeated measurements of internal properties, from row to row. The qualitative strand identifies if the responses are pedagogically appropriate, and involves user feedback, case-by-case interpretation and formative specialist review. The main quantitative measure is CompositeScore. The baseline is based on 50% readability and 50% semantic consistency. RAG has 50% rule compliance, 30% readability and 20% semantic consistency. RAG is rule compliant as the aim of personalised generation is to adhere to the rules of the profile. These rules are not applied to the baseline. Readability is a combination of Flesch-like measures, sentence comfort and line length. Lexical overlap with the task and topicality are aspects of semantic consistency. Bullets, bold, steps, line length, spacing, options, sentence starters and motor scaffolds are all part of compliance. The quantitative evidence is dashboard_log.csv, which includes time, student ID, model, mode, CompositeScore, score method, RAG gain, and latency. The official analysis snapshot in this paper is 578 rows. A newer raw log contains additional exploratory rows, but the paper still has a locked snapshot of the submission with 578 rows in order to avoid mixing development data with the reporting data. The paper uses three levels of output behavior for baselines and comparisons: a baseline response without profile-conditioned rules, an Adaptive RAG response with retrieved learnerprofile constraints, and a multi-provider comparison (OpenAI, Gemini, DeepSeek). This version does not yet have a baseline for formatting. This is a negative because it would give us the ability to differentiate between general formatting and profile-conditioned behavior. The expert-review strand is deemed to be formative and interpretive. Four specialist review forms were completed. These are the suitability ratings of profile rules, scoring of baseline and Adaptive RAG outputs, reviewer details, written notes and declarat declarations. These are considered as formative expert feedback on educational suitability, rather than clinical validation or evidence of learning outcomes.

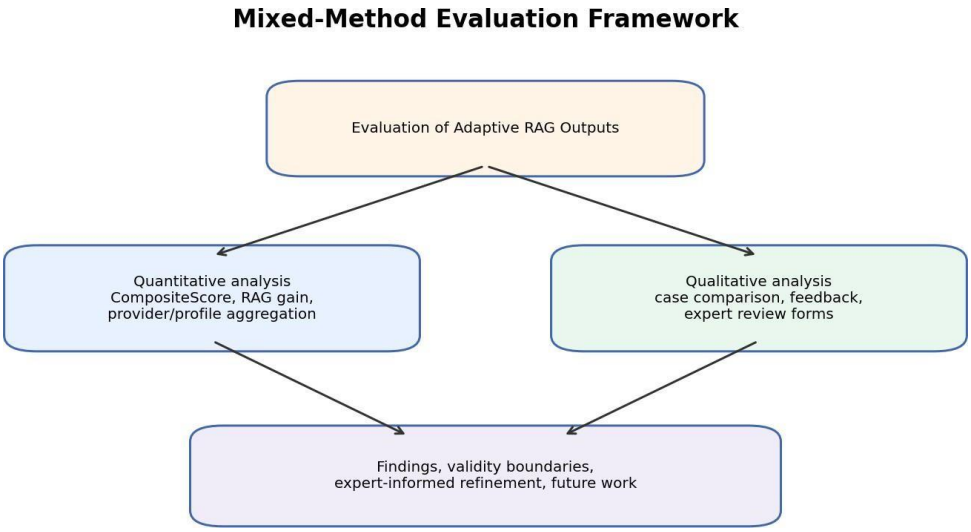

Figure 4. Mixed-method evaluation framework.

Table 5. Evaluation evidence sources and boundaries.

Evidence source	Size / status	Use	Boundary
ASD dataset	1,985 records	Profile engineering and EDA	Not diagnostic authority
Dyslexia dataset	210 task records from 70 participants	Reading-load rules	Not diagnostic authority
Knowledge base	2,063 profiles and expanded strategy families	Runtime retrieval	Educational scaffolds only
Dashboard log	578 rows locked for paper	Baseline vs RAG analysis	Internal metric evidence
Human feedback	23 rows	Ratings and comments	Small convenience sample
Student survey	N=40 summary	Pilot usability	Self-report only
Doctor/expert forms	4 completed,	Preliminary educational suitability review	Not clinical validation

Results

The main outcome is a steady increase in the internal Composite Score, based on the student profile. The 279 rows of baseline data had a mean score of 57.84. The average RAG score for the 295 RAG rows was 87.01. This is a descriptive improvement of 29.17 points. This should be read as an improvement in structure, readability and adherence to profile rules and is an automated accessibility measure, not a clinical learning measure.

outcomes. We also did a paired analysis where the baseline and RAG rows could be paired by time, student and model. This produced 262 pairs. The paired gain was 27.83 with a standard deviation of 20.31. The approximate 95% confidence interval was (25.36, 30.30). The effect size (Cohen dz) within pairs was 1.37. These numbers show a large positive effect in the internal metric system, but do not prove clinical effectiveness or long-term learning gains. Model differences are reflected in the provider results. DeepSeek had the best mean RAG Composite Score in the locked log, OpenAI had a good score and low latency. The mean score of Gemini was lower in this task mix, but it is useful for long and multimodal tasks. This validates the proposed approach: providers to be selected or assessed based on task and profile, not name. Human feedback gives a user's point of view. The feedback file has 23 rows, 4.78/5 rating, and 100.0% helpful. Users reported that scaffolds were clearer, less confusing, and helpful to get started on tasks.

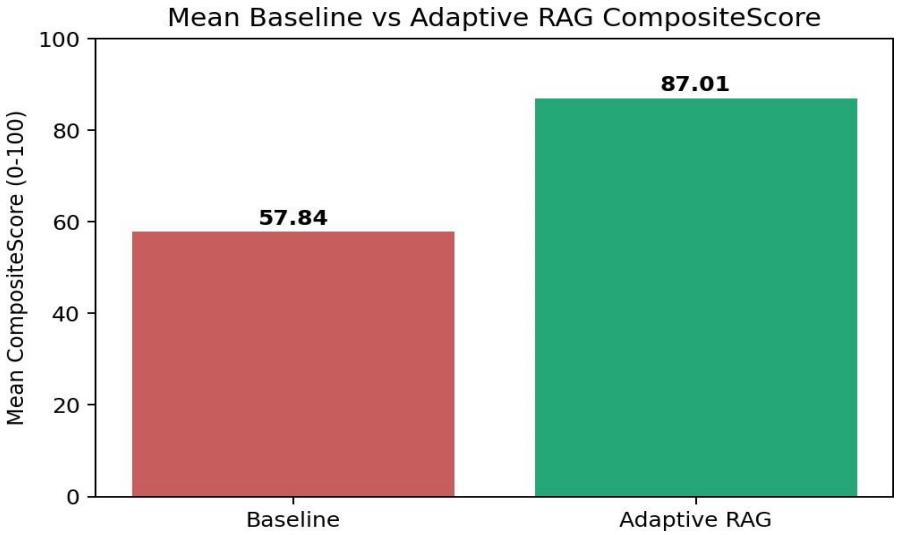

Figure 5. Mean baseline versus adaptive RAG CompositeScore.

This is a measure of usability, not a controlled measure of learning outcomes.

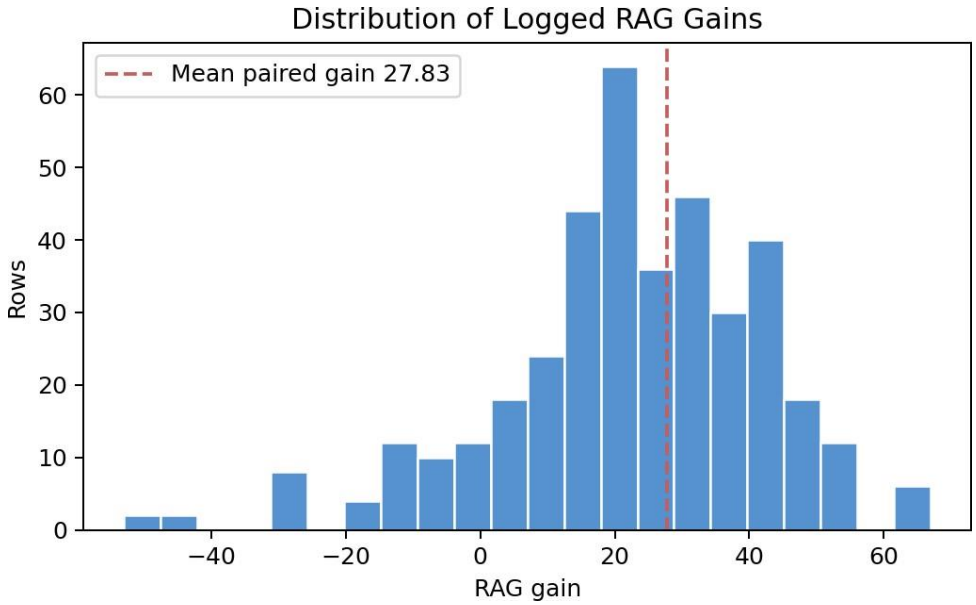

Figure 6. Distribution of logged RAG gains.

Table 6. Main quantitative results.

Metric	Value	Interpretation
Dashboard rows	578	Official paper snapshot of operational prototype log size
Baseline mean	57.84	Standard AI without profile rules
Adaptive RAG mean	87.01	Profile-conditioned generation
Descriptive difference	29.17	Mean RAG improvement
Matched pairs	262	Paired baseline/RAG comparisons
Mean paired gain	27.83	Gain within matched pairs
Approx. 95% CI	25.36 to 30.30	Internal metric uncertainty interval
Cohen dz	1.37	Large within-pair effect size
Mean feedback rating	4.78/5	User-level usability signal
Helpful flags	100.0%	Feedback helpfulness in the small feedback sample

Formative Specialist Review Results

Four specialist review forms were filled out during live demonstrations of the assistive educational prototype. The specialists reviewed a range of baseline and Adaptive RAG outputs and rated them according to pedagogical accuracy, linguistic match, structural accessibility, cognitive load reduction, profile match, safety and reduced ambiguity, semantic match to the task and educational usefulness. The forms provided evidence for the use of Adaptive RAG as an educational accessibility tool. The most strongly supported expert observations were clearer literal language for ASDrelated profiles, reduced visual crowding and stronger keyword anchoring for dyslexia-related profiles, and reduced written output for motor-support profiles. The expert feedback also informed implementation changes, such as the addition of ASD strategy families, the addition of dyslexia strategy families, better conversion of figurative language, and the Dyslexia Guided Reading pilot interface. These results are regarded as formative evidence of educational suitability. They do not support clinical validation, diagnosis, treatment or long-term learning outcomes.

Table 7. Summary of completed specialist review evidence

Evidence item	Final status	Use in paper	Boundary
Completed specialist forms	4 completed	Formative expert feedback on educational suitability	Not clinical validation
Profile-rule review	Completed	Supports refinement of ASD, DYS, and MOT rules	Rule suitability only
Baseline vs Adaptive RAG comparison	Completed for reviewed examples	Compares educational accessibility of outputs	Small expert sample
Written expert notes	Completed	Used to guide system refinements	Qualitative evidence only
Signed declarations	Completed where available	Supports traceability of expert review	Limited to reviewed cases

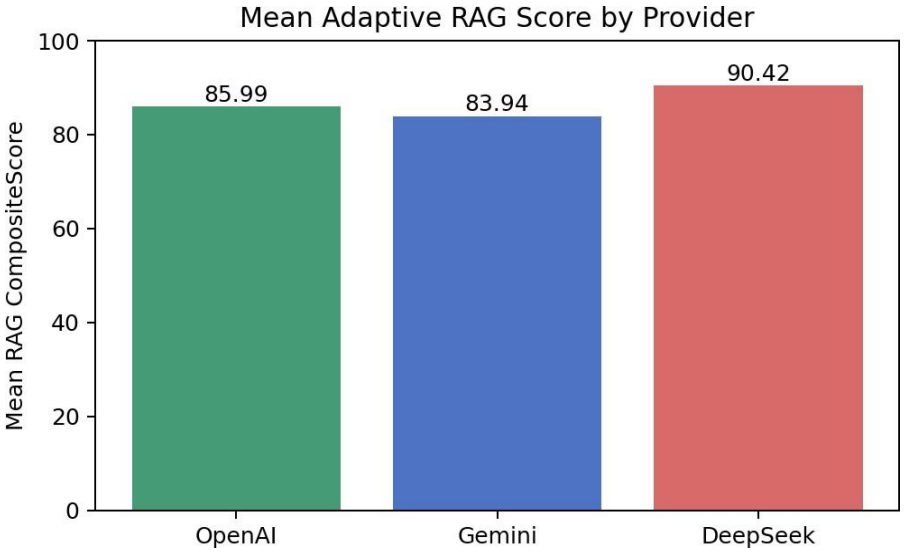

Figure 7. Mean Adaptive RAG CompositeScore by provider.

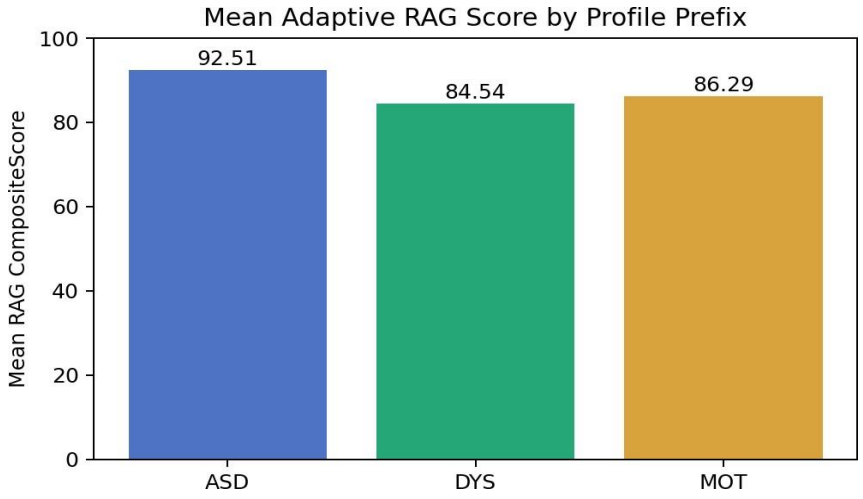

Figure 8. Mean Adaptive RAG CompositeScore by profile prefix.

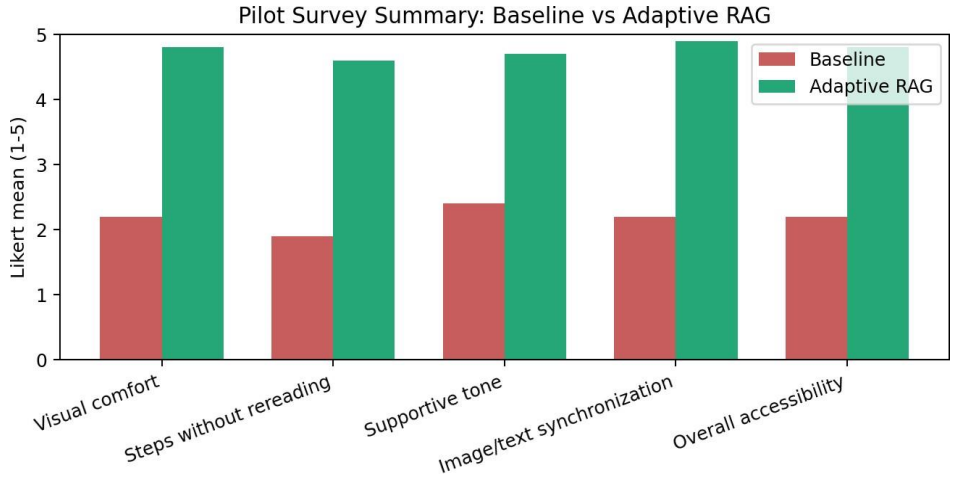

Figure 9. Pilot survey means for baseline and adaptive RAG.

Functional Testing Ledger

Table 7a and Table 7b present the functional-test ledger used to interpret the model gains. The table is intentionally split into two parts to avoid the layout problem that previously produced floating row fragments in the middle of the paper. The values are reported as prototype-level evidence from internal logs and demonstrations, not as clinical outcome measures.

Table 7a. Functional testing ledger, tests 1-10.

Test	Profile	Task	Winning model	Baseline	RAG	Gain	Prototype-level interpretation
1	DYS-1200	Physics (Science)	ChatGPT	45	90	+45	Line length and visual units reduced crowding.
2	ASD-50	Python coding	DeepSeek	41	94	+53	Literal zero-fluff logic reduced ambiguity.
3	DYS-1019	History timeline	Gemini	70	87	+17	Keyword anchors supported memory sequencing.
4	DYS-1996	Math word problem	Gemini/GPT	41	88	+47	Natural-language math improved accessibility.
5	ASD-50	Kitchen safety	DeepSeek	80	91	+11	IF/THEN procedure reduced ambiguity.
6	MOT-Omar	Event logistics	DeepSeek/GPT	61	92	+31	Mind-map and blanks reduced tracking load.
7	DYS-V	BMW tire change	Gemini	38	91	+53	Vision-to-text anchoring supported procedure.
8	ASD-Multi	iDrive analysis	Gemini	50	92	+42	Visual priming supported UI interpretation.
9	DYS-Multi	Electrical flow	Gemini	35	88	+53	Color anchors supported circuit tracking.
10	MOT-Multi	Part ID (N13 engine)	DeepSeek	58	91	+33	Step isolation reduced visual search.

Table 7b. Functional testing ledger, tests 11-20.

Test	Profile	Task	Winning model	Baseline	RAG	Gain	Prototype-level interpretation
11	PDA-100	SQL task refusal	Gemini	31	92	+61	Declarative tone preserved autonomy.
12	SPD-L2	iDrive schematics	DeepSeek	40	95	+55	Monochrome minimalism reduced sensory load.
13	MOT-Omar	Python refactoring	ChatGPT	60	91	+31	Active blanks reduced typing burden.
14	DYS-1200	DBMS normalization	ChatGPT	45	93	+48	Vertical anchors stabilized table reading.
15	ASD-50	API integration	DeepSeek	70	94	+24	Zero-fluff logic reduced unsupported assumptions.
16	DYS-V	Gearbox assembly	Gemini	38	91	+53	Vision-to-text sync supported assembly.
17	SPD-L2	Network topology	DeepSeek	42	90	+48	Reduced linguistic noise improved spatial focus.
18	PDA-100	Exam scheduling	Gemini	28	89	+61	Indirect language maintained autonomy.
19	DYS-1996	Calculus limits	Gemini	41	88	+47	Natural-language math reduced cognitive load.
20	ASD-Multi	Lab safety (N13)	DeepSeek	80	91	+11	Literal steps supported safety compliance.

Screenshot and Dashboard Evidence

The figures below are included as implementation evidence and qualitative examples. They show that the prototype is not only a conceptual pipeline but a working local system with a learner interface, baseline/adaptive output comparison, scoring display, dashboard analytics, API logging, and guided reading UX. The screenshots should be interpreted as artifact evidence, not as independent clinical evidence.

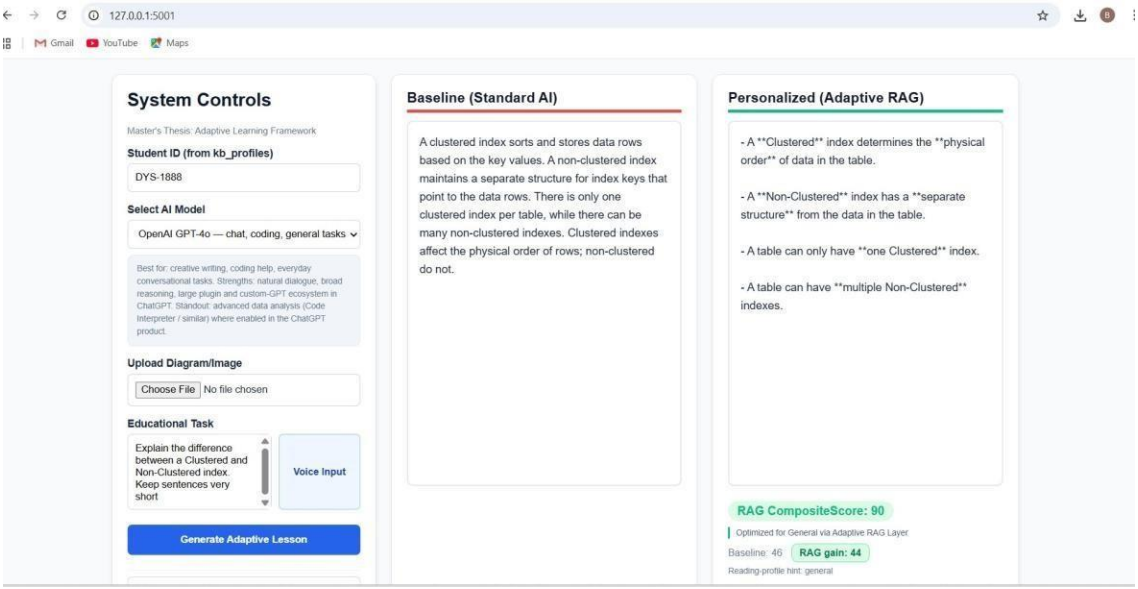

Figure 10. Dyslexia-oriented science example showing shorter lines, keyword emphasis, and a higher Adaptive RAG score.

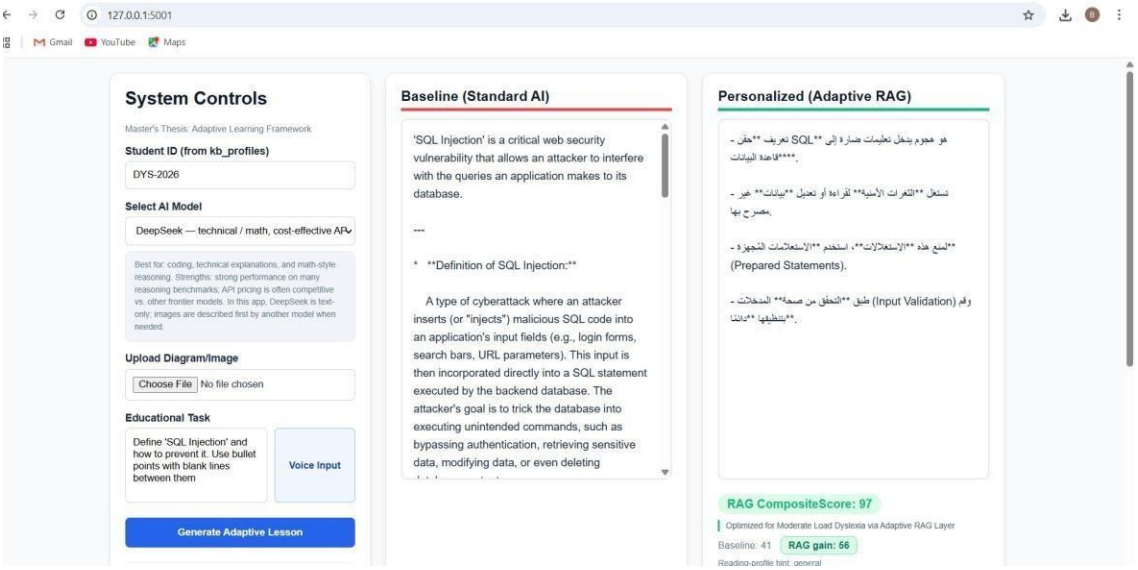

Figure 11. ASD-oriented coding example showing numbered literal steps and reduced conversational ambiguity.

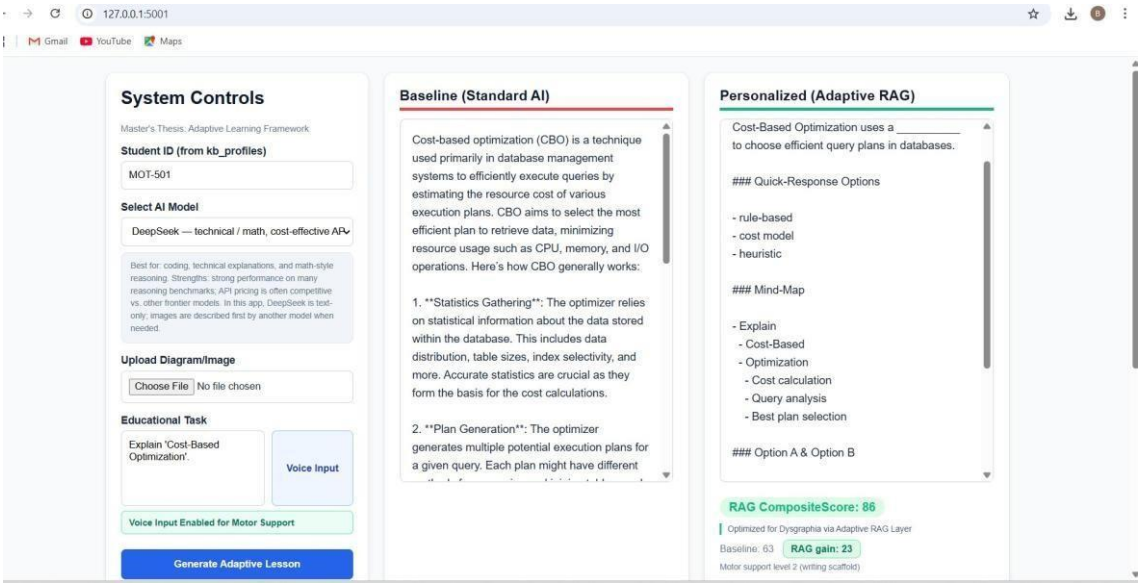

Figure 12. Multimodal physical-procedure example used for vision-to-text anchoring and visual scaffolding discussion.

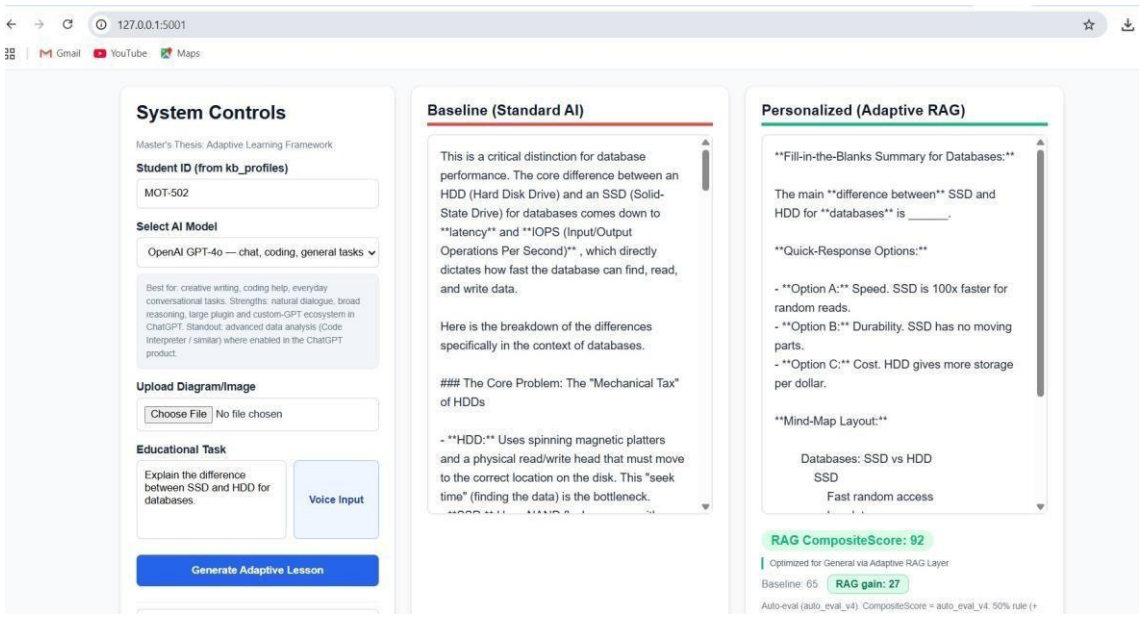

Figure 13. MOT support example showing fill-in-the-blank and quick-response scaffolding for reduced written-output burden.

Figure 14. Winner fairness dashboard example with provider counts and mean score quality.

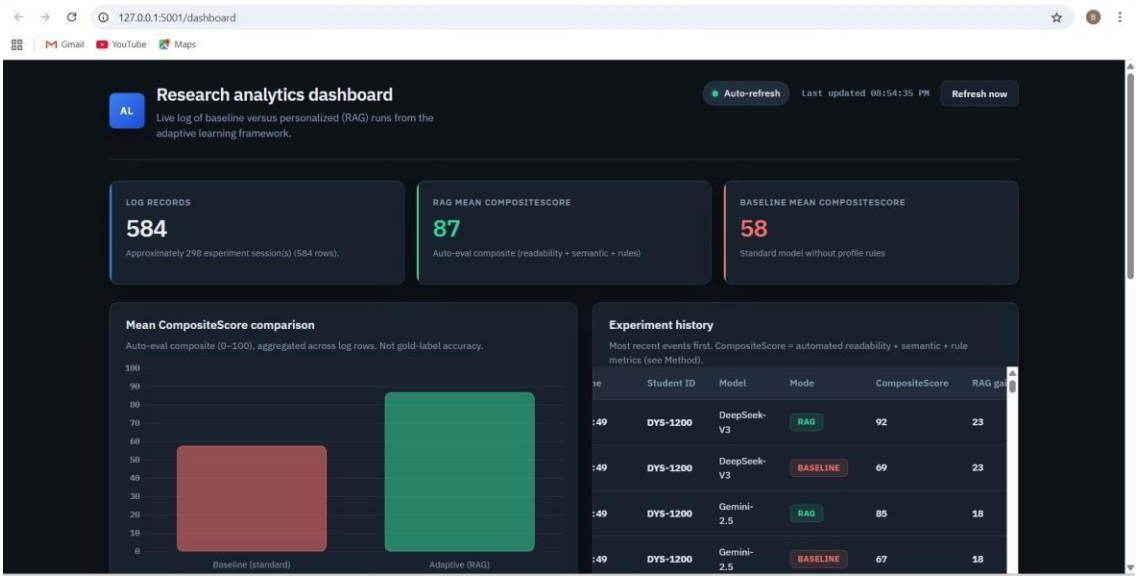

Figure 15. Research analytics dashboard showing logged records, baseline mean, and RAG mean in the local prototype.

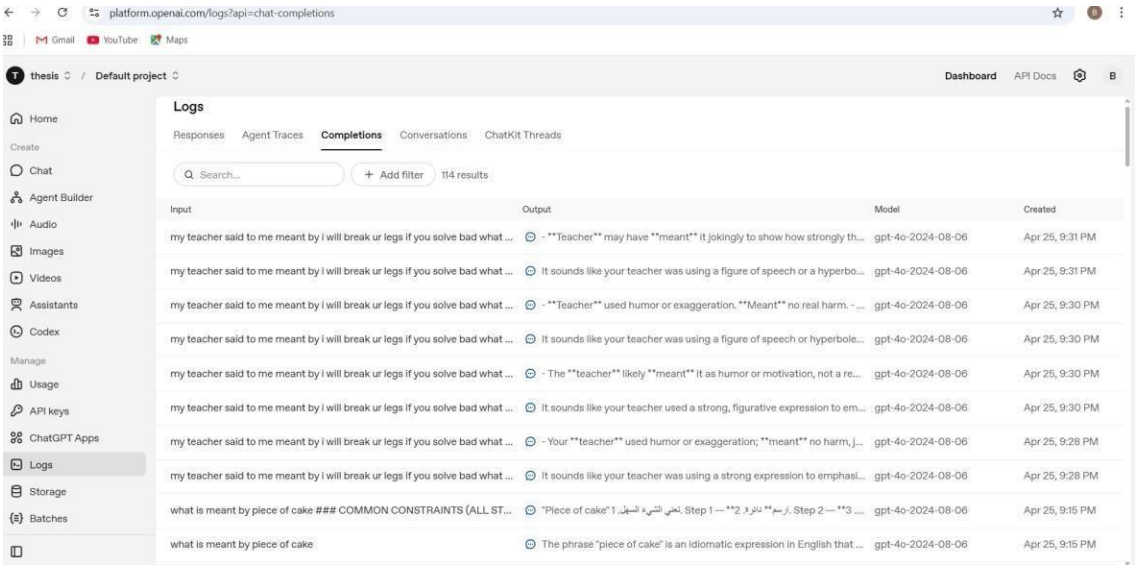

Figure 16. ChatGPT API log screenshot used as evidence that model version, prompt constraints, and outputs were traceable.

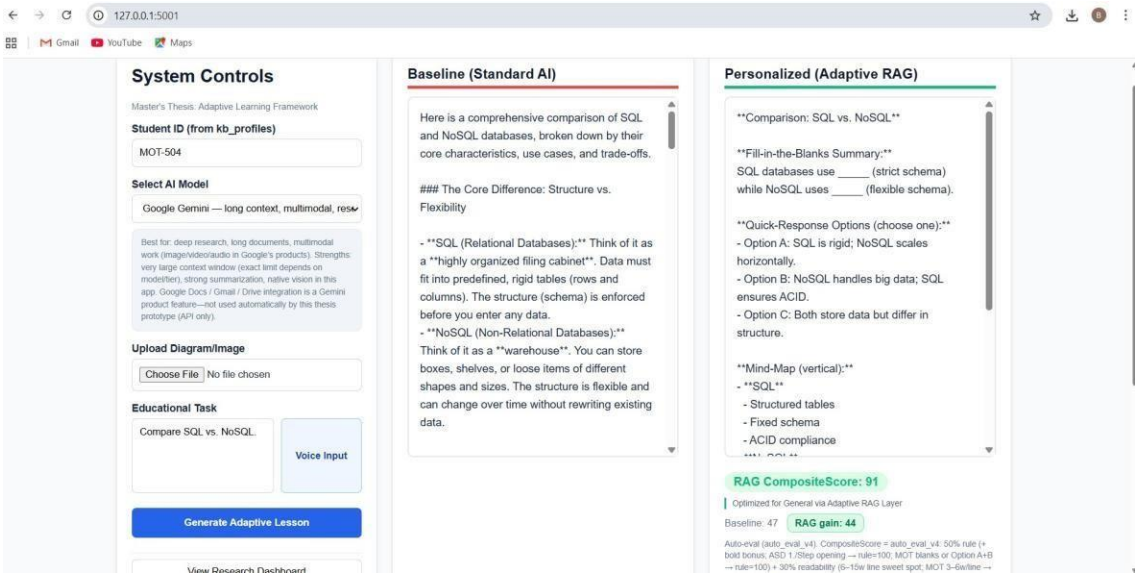

Figure 17. Dyslexia Guided Reading pilot interface where text is split into short chunks and future chunks remain locked.

Discussion

The results support the claim that personalization is a systems problem. A general model may be able to solve a task, but it may not always be able to know how to represent the task for a particular learner unless the constraints are given.

These constraints are explicitly provided in the RAG design. This removes the need for constant reminders and monitoring of rule application. The largest empirical finding is the difference between baseline and Adaptive RAG

CompositeScore. This is expected as the RAG prompt includes profile rules, but it is important as the process of injecting rules, running the model, scoring, ranking and logging is repeatable. An answer can be generated by a prompt,
and for many tasks, an information system can generate, score, rank, log, review and manage answers. Four completed specialist forms to support formative assessment process. They are not clinically valid, but offer

formative expert feedback on clarity, language fit, structure, cognitive load, safety, reduced ambiguity, and profile suitability. This is where expert feedback at prototype level comes in. The forms also helped with the system by fostering improved ASD literalization, increased dyslexia strategy families, and a guided reading interface. The main theory is that RAG can be used from knowledge grounding to behaviour grounding. The retrieved object is not only a factual statement in the context of education accessibility, but also a behavior contract: be literal, don't crowd, don't be ambiguous, provide options, or make writing easier. This could apply to other accessibility scenarios where the user requires presentation, language and interaction constraints. 7. The construct validity of CompositeScore is weak since it is a construct. It measures readability, semantic consistency and rule compliance, but not all of educational suitability. It can mean that
the output was consistent with the profile, but it does not mean that the student learned more. One particular construct-validity issue is circularity. The Adaptive RAG prompt includes profile rules and the CompositeScore includes rulecompliance features. So, some of the gain is due to successful compliance with the injected profile rules. This is desirable for development, but may be circular if used as a measure of learning. This paper introduces CompositeScore, an internal accessibility-focused metric to reduce overclaiming.

API variability and tasks logged are limitations in the internal validity. Models may change, latency may vary, and the tasks are not perfectly balanced. The paired analysis is helpful in comparing baseline and RAG results within the same row, but more controlled testing is needed. The prototype is based on public/anonymized data, local logs, and a convenience sample for feedback, which limits generalizability. To use the system in Egyptian schools or universities, institutional ethics approval or exemption, Arabic/English language testing, parental consent (if students are under 18), teacher training, profile-review governance, and school/university data-protection procedures would be required. The system should not be used as a diagnostic on an ethical basis. It is not a diagnosis of ASD, dyslexia, dysgraphia,

Any medical condition or ADHD. It adapts educational material once a profile has been selected or simulated for research purposes. It should not be a substitute for teachers, clinicians, parents or IEPs. Bias and fairness need ongoing monitoring as profile rules may over-constrain learning if too strong or not help if too weak.

Table 8. Validity and ethics safeguards.

Risk / limitation	Mitigation in current prototype	Future requirement
Automated metric overclaiming	CompositeScore is explicitly labeled internal and nonclinical.	Blinded specialist review and learning-outcome studies.
Circularity in score design	RAG gains are interpreted as rule-adherence and accessibility-oriented evidence.	Separate adherence metrics from independent pedagogical-quality rubrics.
Small feedback sample	Feedback is described as a usability signal only.	Larger recruitment under ethics approval.
Provider volatility	Logs record model names, latency, and scores.	Versioned model tracking and repeated evaluation.
Arabic and Egyptian context	Arabic examples and RTL support are included as prototype features.	Arabic datasets, Egyptian curriculum tasks, and local institutional deployment.
Privacy risk	Research IDs use ASD-/DYS-/MOT-prefixes, not names.	Authentication, encryption, consent, and data governance.
Rule misalignment	Rules are visible in kb_profiles.json and can be edited.	Clinician, educator, and learner review of profile rules.

Conclusion and Future Work

This paper presented a behavior-conditioned RAG system for neurodiverse education. The system fetches the teaching constraints from the learner profiles, enriches the prompts, executes multiple LLM providers, scores the responses, logs and visualises the results, collects feedback and visualises evidence. There is strong evidence of automated accessibility metrics in dashboard logs and positive evidence of usability signals in feedback/survey evidence. The innovation is the redefinition of RAG for education accessibility. It does not retrieve factual information, but teaching constraints. This makes personalization auditable, editable and repeatable. For ASD and dyslexia this is particularly relevant as response structure, literalness, line length and visual load can make or break an answer.

Future research should involve a larger number of specialists in the review panel, a larger number of learners, measure learning outcomes over time, improve Arabic and right to left formatting, develop Arabic datasets for dyslexia and ASD-related education, scale up the number of learners, measure learning outcomes over time, improve Arabic and right to left formatting, develop Arabic datasets for dyslexia and ASD-related education, add teacher control for learner profiles, benchmark local privacy preserving models, and compare RAG with fine-tuning, generic formatting-only prompting, and traditional tutoring systems. Future use in Egypt should be tested in schools and universities under institutional supervision, informed consent, teacher training and data management. Until this is done, the framework should be introduced as a robust prototype and research tool, not as a validated intervention. 9.

Commercial LLMs can change over time, so version strings of models, API dates, scoring version, dataset snapshots and code archives should be included in final submissions. This prevents reviewers from viewing the results as model independent and helps reruns. Ethics Statement. This study was a low-risk assessment of an assistive educational prototype. There were no required review and feedback activities. Data reported were anonymised or pseudonymised and no personal information, clinical diagnostic data or treatment data were kept in the learner-profile knowledge base.

The prototype is not designed to be a diagnostic, therapeutic or clinical decision support tool, but rather to adapt educational explanations. The educational suitability of the outputs reviewed was determined by expert reviews, not clinical efficacy. This study was performed as a low-risk assessment of an assistive educational prototype under academic supervision. Feedback and expert review was voluntary. No clinical diagnostic record, reported data were anonymized or pseudonymized. Availability of code and data. The prototype contains a Flask backend, HTML/JavaScript frontend, JSON learner-profile knowledge base, scoring, dashboard logs, feedback and analytics dashboards. Prior to submission, the source code and anonymized reproducibility artifacts should be archived with a commit hash.

The version of scoring.py, a snapshot of kb_profiles.json, a summary of dashboard_log.csv, an anonymized version of human_feedback.csv, a summary of survey, and an anonymized summary of expert-review should be included in this archive. Personally identifiable and clinical data should not be placed in the public archive.

Table 10. Reproducibility metadata to final submission before submission.

Artifact	Reported value
OpenAI model	gpt-4o-2024-08-06
Gemini model	gemini-1.5-pro-002 or final logged Gemini model string
DeepSeek model	DeepSeek-V3 or final deployed API model string
Scoring version	auto_eval_v4
Official log snapshot	dashboard_log.csv, 578 rows
Baseline rows	279
Adaptive RAG rows	295
Matched pairs	262
Knowledge base	kb_profiles.json snapshot used in final experiments
Feedback file	human_feedback.csv summary, 23 rows
Survey evidence	N=40 summary
Specialist review	4 completed specialist review forms
Code archive	Repository URL and commit hash to be inserted before submission

References

- Alluhaidan, A. S., Rashid, M., Aldehim, G., Basheer, S., & Sakri, S. (2024). Efficient deep learning model for detecting dyslexia in children from an online gamified test dataset. Journal of Disability Research, 3(8).
https://doi.org/10.57197/jdr-2024-0099
- American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.). American Psychiatric Association Publishing.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G. B., Lespiau, J.-B.,
- Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. Proceedings of the 39th International Conference on Machine Learning.
- Carik, B., Ping, K., Ding, X., & Rho, E. H. (2025). Exploring large language models through a neurodivergent lens:
- Use, challenges, community-driven workarounds, and concerns. Proceedings of the ACM on Human-Computer
- Interaction, 9(GROUP), 1–28. https://doi.org/10.1145/3701194
- CAST. (2024). Universal Design for Learning guidelines version 3.0. https://udlguidelines.cast.org/
- Dabaghi, K., D'Urso, S., & Sciarrone, F. (2024). Artificial intelligence and learning of students with dyslexia: A brief review. In Emerging Technologies for Education (pp. 155–169). Springer. https://doi.org/10.1007/978-981-97-42431_13
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrievalaugmented generation for large language models: A survey. arXiv. https://arxiv.org/abs/2312.10997
- International Dyslexia Association. (2002). Definition of dyslexia. https://dyslexiaida.org/definition-of-dyslexia/
- Iyer, L. S., Chakraborty, T., Reddy, K. N., Jyothish, K., & Krishnaswami, M. (2023). AI-assisted models for dyslexia and dysgraphia: Revolutionizing language learning for children. In AI-assisted special education for students with exceptional needs (pp. 186–207). IGI Global. https://doi.org/10.4018/979-8-3693-0378-8.ch008
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas for Navy enlisted personnel. Naval Technical Training Command.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.
- Mayer, R. E. (2020). Multimedia learning (3rd ed.). Cambridge University Press.
https://doi.org/10.1017/9781316941355
- Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO.
- Pagliara, S. M., Bonavolontà, G., Pia, M., Falchi, S., Zurru, A. L., Fenu, G., & Mura, A. (2024). The integration of artificial intelligence in inclusive education: A scoping review. Information, 15(12), Article 774.
https://doi.org/10.3390/info15120774
- Panjwani-Charania, S., & Zhai, X. (2024). AI for students with learning disabilities: A systematic review. In AI in learning: Designing the future (pp. 469–493). Oxford University Press.
https://doi.org/10.1093/oso/9780198882077.003.0021
- Rello, L., & Baeza-Yates, R. (2013). Good fonts for dyslexia. Proceedings of the 15th International ACM SIGACCESS
- Conference on Computers and Accessibility. https://doi.org/10.1145/2513383.2513447
- Rello, L., Pielot, M., & Marcos, M.-C. (2016). Make it big! The effect of font size and line spacing on online readability. Proceedings of the CHI Conference on Human Factors in Computing Systems.
- Santhiya, S., & KanimozhiSelvi, C. S. (2023). A study on dyslexia detection using machine learning techniques for checklist, questionnaire and online game based datasets. Applied and Computational Engineering, 5(1), 837–842.
https://doi.org/10.54254/2755-2721/5/20230722

- Shaywitz, S. E., & Shaywitz, B. A. (2005). Dyslexia-specific reading disability. Biological Psychiatry, 57(11), 1301– 1309.
- Shilaskar, S., Bhatlawande, S., Deshmukh, S., & Dhande, H. (2023). Prediction of autism and dyslexia using machine learning and clinical data balancing. 2023 International Conference on Advances in Intelligent Computing and
- Applications (AICAPS), 1–11. https://doi.org/10.1109/AICAPS57044.2023.10074161
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285.
- Taş, T., Bülbül, M. A., Haşimoğlu, A., Meral, Y., Çalışkan, Y., Budagova, G., & Kutlu, M. (2023). A machine learning approach for dyslexia detection using Turkish audio records. Turkish Journal of Electrical Engineering and Computer
- Sciences, 31(5), 892–907. https://doi.org/10.55730/1300-0632.4024
- UNESCO. (2023). Guidance for generative AI in education and research. UNESCO.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
- Yap, J. R., Aruthanan, T., & Chin, M. (2025). Rewriting the script: A scoping review of the role of artificial intelligence in dyslexia research and education. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3526189