Red Team Feedback Report: Team F

# Executive Summary

Team 6 (F) has proposed a comprehensive "AI Tutor" aimed at addressing the "black box" problem in EdTech through transparent scoring logic and a supporting governance framework.

The system utilizes a multi-agent LLM architecture (Evaluator, Tutor, SME) to provide feedback on student work.

**Strengths:**

* **Governance-First Approach:** The inclusion of a dedicated Governance Framework alongside the technical prototype is a significant strength, addressing critical adoption barriers like FERPA/COPPA compliance and stakeholder trust.
* **Clear Persona Definition:** The distinct dashboards for Admins (Anna), Teachers (Tom), Parents (Paula), and Students (Steve) effectively target specific stakeholder needs.
* **Innovative Scoring Logic:** The "Learning Signal Score" which separates Correctness from Explanation Quality is a valuable pedagogical innovation that discourages rote memorization.

**Weakness:**

* **Security Maturity:** The system currently lacks enterprise-grade security controls and specific guardrails required for deployment in a sensitive K-12 environment.
* **Pedagogical Risks:** The "always-on" live scoring mechanism may inadvertently induce stress or "reward hacking" behaviors rather than genuine learning.

# Key Findings and Recommendations

## Data Security and Privacy

1. Review data handling practices
   * **Finding:** While the presentation claims alignment with FERPA/COPPA, the technical architecture does not explicitly demonstrate enterprise-grade privacy enforcing features such as Single Sign-On (SSO), Data Boundary Controls, or Data Encryption in transit/at rest.
   * **Finding:** There is no clearly defined mechanism for the secure storage of vector database credentials used in the RAG (Retrieval-Augmented Generation) system.

**Recommendation:** Implement and document specific enterprise security features (SSO, Context-Aware Access). Explicitly define how RAG vector DB credentials are secured to prevent unauthorized alteration of lesson plans.

1. Test for potential data leakage
   * **Finding:** The current architecture relies on standard LLM agent orchestration without explicit "PII Guardrails." There is a risk that sensitive student data entered into the chat could be processed without redaction.

**Recommendation:** Integrate PII detection layers that redact or flag sensitive information (names, addresses, IDs) before data is sent to the LLM inference layer.

1. Evaluate input validation and sanitization
   * **Finding:** The system lacks visible "Jailbreak Guardrails" to identify attempts to circumvent AI safety protocols (e.g., prompt injection, role-playing attacks).
   * **Finding:** Moderation guardrails for content categories such as harassment, hate speech, and self-harm are not explicitly detailed in the agent architecture.

**Recommendation:** Implement a dedicated moderation endpoint (e.g., similar to OpenAI's moderation API) to sanitize both input prompts and model outputs.

## Technical Findings

1. **Hallucinations & Citations:** The SME (Subject Matter Expert) agent explains concepts (e.g., "AI Hallucinations") without providing supporting citations. This undermines the "Explainable" goal of the project. Standard RAG implementations should enforce citation grounding.
2. **Learning Style Limitations:** The prototype is heavily text-based (Reading/Writing focus). It currently lacks support for Visual, Auditory, or Kinesthetic learning styles (e.g., audio overviews, interactive quizzes), placing it behind competitors like Google NotebookLM.
3. **Feedback Loops:** There is no mechanism for students to provide feedback on AI responses (e.g., thumbs up/down), limiting the team's ability to detect edge cases or failure modes.

## Potential high-risk areas requiring deeper investigation

1. **Reward Hacking & Student Anxiety:** The "Live Scoring" feature is a high-risk design choice. Constant evaluation can trigger a threat response (cortisol spike) in students, inhibiting learning. Students may also learn to "game" the AI to maximize the score rather than understanding the material.
2. **RAG Integrity:** If the "Knowledge Base" (lesson files) can be compromised or if the retrieval mechanism fails, the AI Tutor may confidently provide incorrect information to students, which is a critical failure in an educational context.
3. **Algorithmic Bias:** While "fairness" is a goal, there is a need to investigate if the "Explanation Quality" scoring biases against students with lower English proficiency or different linguistic backgrounds.

|  |  |  |  |
| --- | --- | --- | --- |
| Risk Category | Risk Description | Likelihood | Impact |
| Technical | **AI Hallucination:** The Tutor provides incorrect facts without citations, misleading students. | High | High |
| Technical | **Lack of Accessibility:** The tool fails to support non-text learners or non-English speakers effectively. | High | Medium |
| Ethical | **Psychological Harm:** "Always-on" scoring creates a high-stakes environment, leading to student anxiety or disengagement. | Medium | High |
| Ethical | **Reward Hacking:** Students exploit the AI to get high scores without actual learning (e.g., pasting answers from other LLMs). | High | Critical |
| Security | **Jailbreaking/Prompt Injection:** Students bypass safety filters to generate inappropriate content. | Medium | High |
| Security | **PII Leakage:** Student personal data is inadvertently processed or stored insecurely. | Medium | Critical |

# Readiness for Deployment

**Current Status: Not Ready for Deployment**

**Justification:** While the *concept* and *governance framework* are mature, the **technical prototype requires significant hardening** before it is safe for a K-12 environment. The absence of explicit content moderation guardrails, PII protection mechanisms, and citation grounding for the SME agent presents unacceptable liability risks.

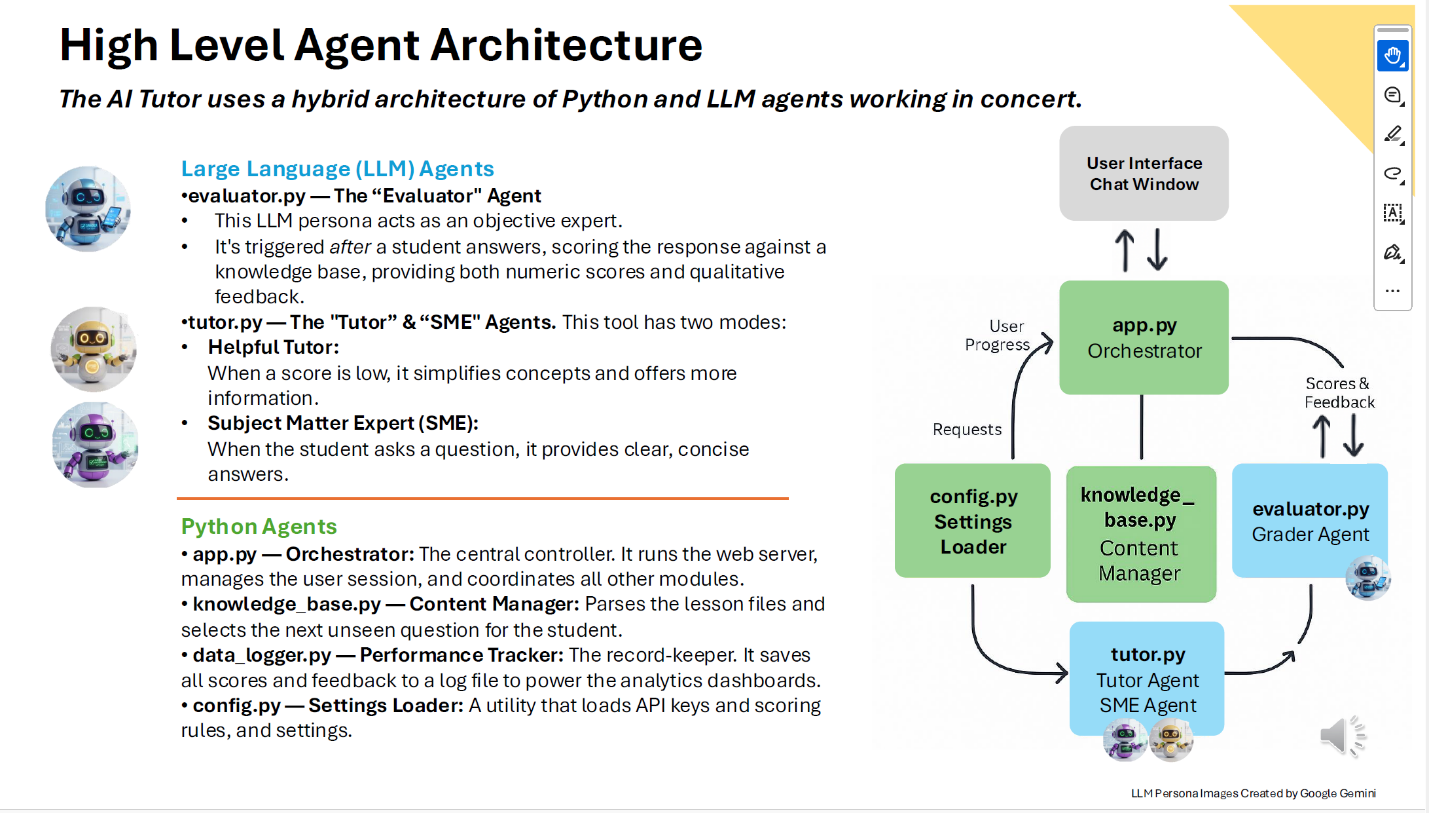
**Required Actions for Deployment Readiness:**

1. **Implement Safety Guardrails:** Deploy moderation endpoints and jailbreak detection.
2. **Grounding:** Enforce citations for all factual assertions made by the SME agent.
3. **Pedagogical Pivot:** Re-evaluate "Live Scoring" to reduce student anxiety; consider "On-Demand" analytics instead.
4. **Security Audit:** Define encryption standards and credential management for the RAG database.
5. **Ethical Audit:**- Audit with student health advisor for impact on mental wellbeing of students.

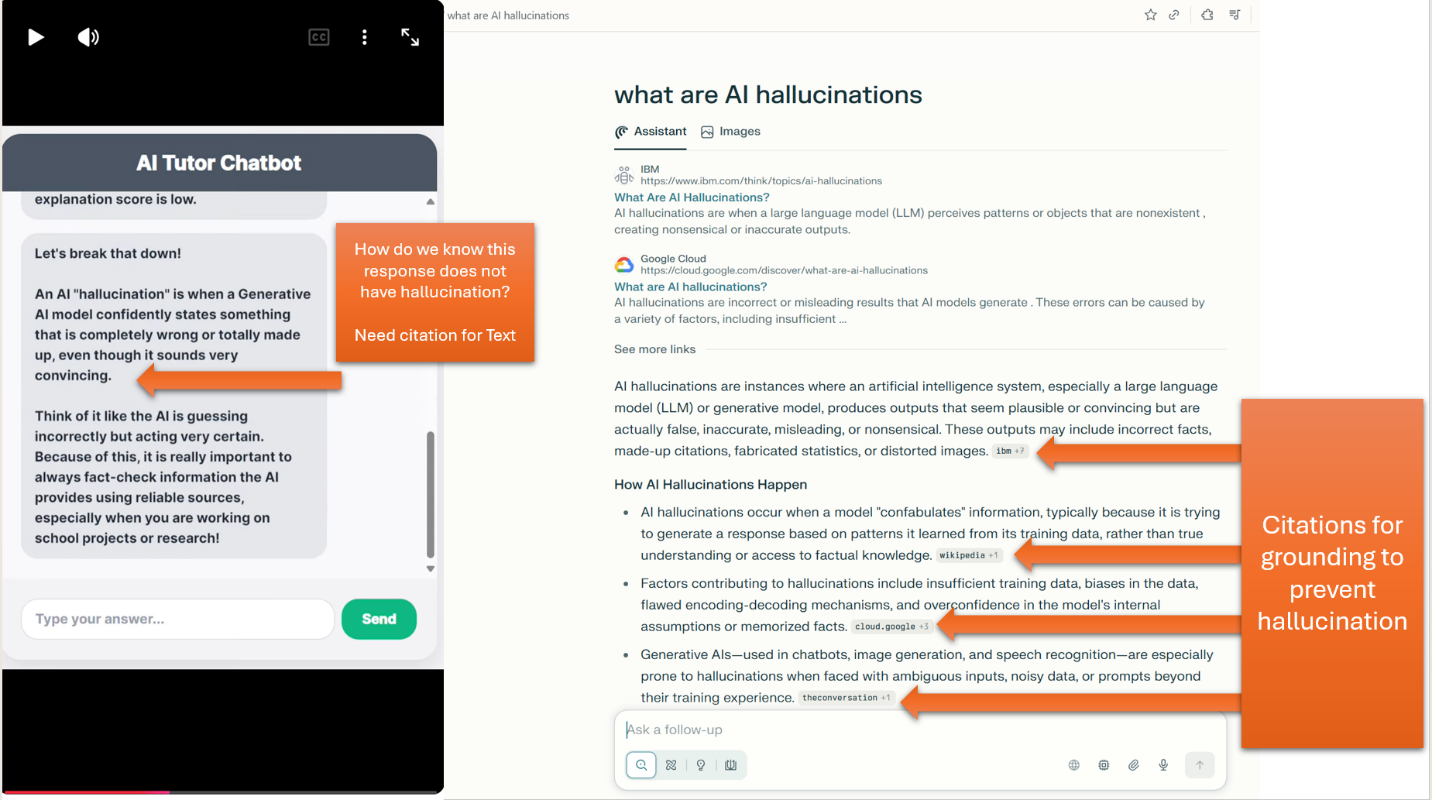
# Technical Findings

1. Verify core features work as documented
   1. *The core features, including student questioning and scoring responses with detailed grading explanations, are functioning as demonstrated in the YouTube videos.*
2. Test system with various input types (test with edge cases)
   1. *With limited access to the prototype, testing of the system with various input types and edge cases remains a future endeavor.*
3. Validate output quality and consistency
   1. *The YouTube videos demonstrate that the system's output quality and consistency are suitable for use as an AI Tutor capable of grading student responses and delivering constructive feedback.*
4. Evaluate the system behavior with unexpected data
   1. *With limited access to the prototype, testing of the system with various input types and edge cases remains a future endeavor.*
5. Document system limitations and failure modes

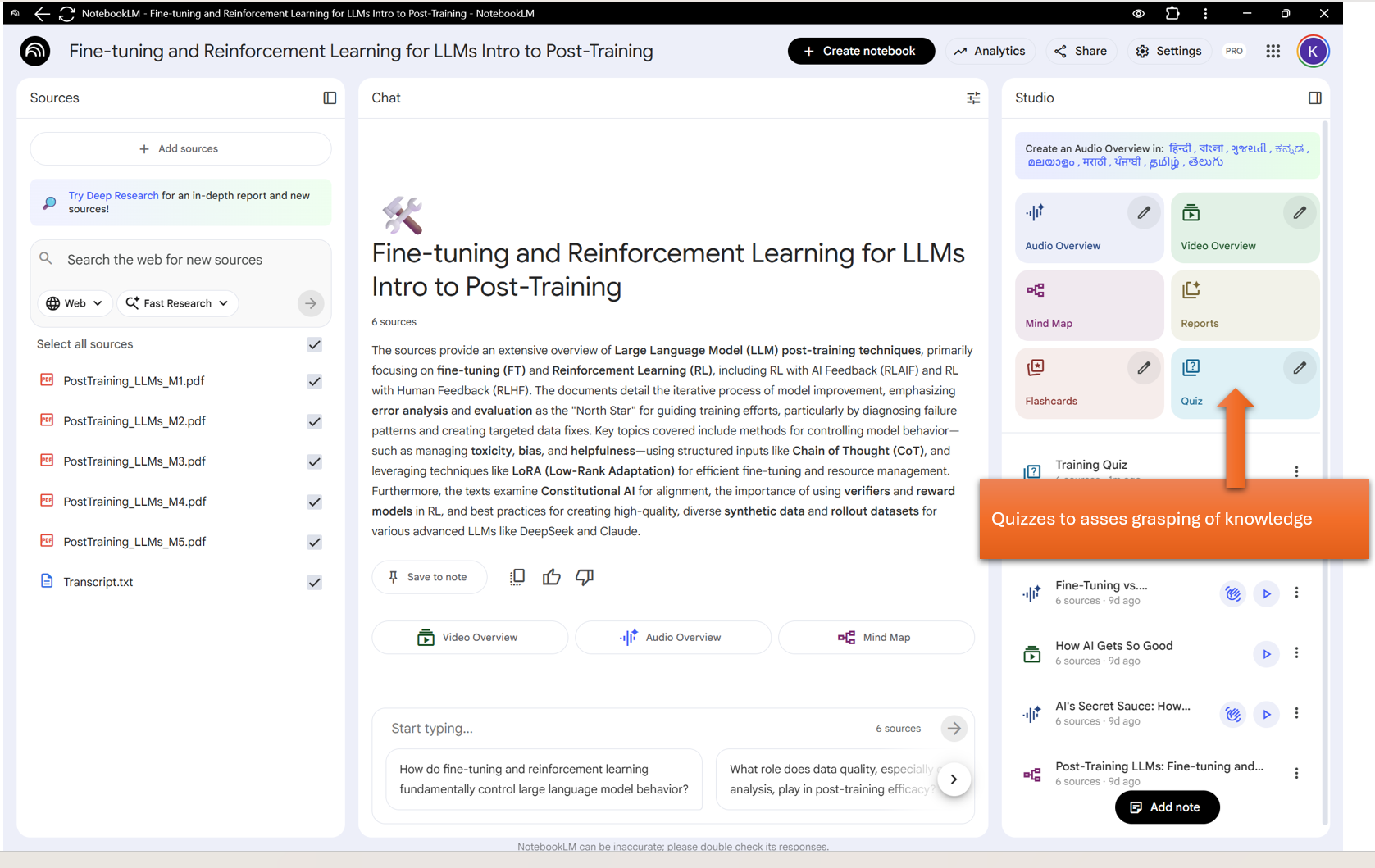
The high-level Agent architecture provides a clear depiction of the sub-agents; however, it does not sufficiently emphasize that the SME sub-agent (tutor.py) utilizes a RAG-based approach, refer image below.



The current architecture lacks mechanisms to reinforce the safeguards intended to prevent hallucination by the SME agent within tutor.py. For instance, as illustrated in the screenshot below, the AI Tutor provides an explanation of AI hallucination. However, without supporting citations, it is difficult to ascertain the accuracy of this explanation because the response is generated by a large language model. The industry standard for addressing hallucinations includes the provision of citations, as demonstrated in the screenshot on the right.



Although AI Tutor effectively assesses whether students initially grasp the material, it does not evaluate whether students have understood the responses provided by AI Tutor through follow-up quizzes. For context, Google's NotebookLM offers quizzes to help students assess their comprehension of the topic, as illustrated in the screenshot below.



## Moderation Guardrails

The present architecture does not include mechanisms to strengthen moderation guardrails across the following categories as detailed [here:](https://platform.openai.com/docs/guides/moderation#content-classifications)

|  |  |  |
| --- | --- | --- |
| Category | Description | Inputs |
| harassment | Content that expresses, incites, or promotes harassing language towards any target. | Text only |
| harassment/threatening | Harassment content that also includes violence or serious harm towards any target. | Text only |
| hate | Content that expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste. Hateful content aimed at non-protected groups (e.g., chess players) is harassment. | Text only |
| hate/threatening | Hateful content that also includes violence or serious harm towards the targeted group based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste. | Text only |
| illicit | Content that gives advice or instruction on how to commit illicit acts. A phrase like "how to shoplift" would fit this category. | Text only |
| illicit/violent | The same types of content flagged by the illicit category but also includes references to violence or procuring a weapon. | Text only |
| self-harm | Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders. | Text and images |
| self-harm/intent | Content where the speaker expresses that they are engaging or intend to engage in acts of self-harm, such as suicide, cutting, and eating disorders. | Text and images |
| self-harm/instructions | Content that encourages performing acts of self-harm, such as suicide, cutting, and eating disorders, or that gives instructions or advice on how to commit such acts. | Text and images |
| sexual | Content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness). | Text and images |
| sexual/minors | Sexual content that includes an individual who is under 18 years old. | Text only |
| violence | Content that depicts death, violence, or physical injury. | Text and images |
| violence/graphic | Content that depicts death, violence, or physical injury in graphic detail. | Text and images |

1. Jailbreak Guardrails: The current architecture does not clearly highlight features designed to reinforce jailbreak guardrails for the identification of attempts to circumvent AI safety protocols, including prompt injection, role-playing requests, or social engineering techniques.
2. PII Guardrails: The current architecture does not clearly highlight features designed to reinforce PII guardrails for the identification of PII shared by students.
3. Enterprise grade Privacy enforcing features: The current architecture does not clearly highlight following enterprise grade privacy features:
   1. SSO (Single sign-on): users log in with their existing organizational (work or school) accounts.
   2. Data Encryption - employs encryption for data both in transit and at rest, protecting against unauthorized access.
   3. Data Boundary Controls: enforce data residency and ensure data remains within the organization's data center or cloud.
   4. Secure Storage of Vector Database Credentials for RAG: The existing architecture lacks a clearly defined mechanism for the secure storage of credentials associated with the vector database in RAG. Insufficient access controls may expose lesson plans and content to unauthorized alteration, which could compromise the integrity of training materials and negatively impact students' academic progress.
   5. Context-Aware Access (CAA): control access based on context (e.g., user identity, device security status, network location)
   6. No AI Model Training on User Data: uploaded files, chats, and model outputs are not used to train the generative AI models.
   7. No Human Review: Data from accounts is not reviewed by human reviewers, ensuring confidentiality for sensitive information like student records (FERPA compliant).
   8. Admin Controls: Workspace administrators have control over turning AI Tutor on or off for their users and managing sharing settings (e.g., sharing within the organization only).

Reference links for more info:

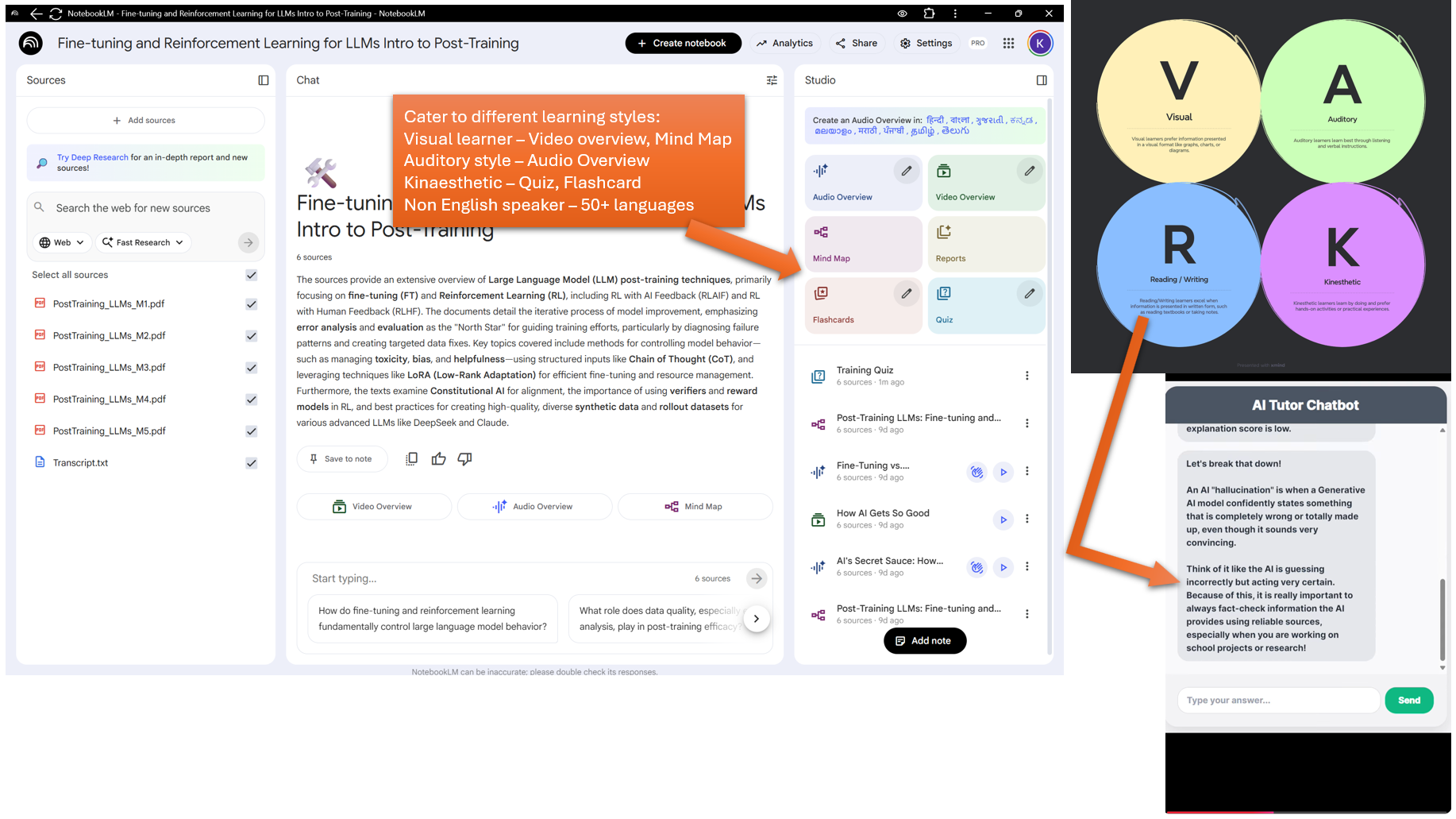
* <https://workspaceupdates.googleblog.com/2025/04/notebookLM-and-gemini-app-core-services-for-education-customers.html>
* <https://medium.com/google-cloud/notebooklm-enterprise-security-d49f70784621>
* <https://support.google.com/a/answer/15239506>

# Recommendations

While AI Tutor currently supports the reading/writing learning style through textual responses, its existing architecture does not sufficiently address other learning styles such as visual, auditory, or kinesthetic, nor does it provide adequate support for non-English speakers. For understanding with analogy, NotebookLM offers features tailored to various learning preferences, including audio overviews for auditory learners, flash cards and quizzes for kinesthetic learners, and video overviews and mind maps for visual learners. Furthermore, audio overviews are available in fifty different languages for non-English speakers: <https://blog.google/technology/google-labs/notebooklm-audio-overviews-50-languages/>



Source: <https://xmind.com/blog/learning-styles-definition-and-example>



Student Feedback Collection: The current system architecture lacks a mechanism to gather feedback from students, such as thumbs up or down responses. As a result, situations where students find the system unhelpful—especially in edge cases or unforeseen failure modes—are not captured, limiting opportunities to address user pain points. For context, platforms like NotebookLM incorporate a thumbs up/down feature, as shown in the image below.



Provide support rather than judgment for students at all times: The prototype or system architecture does not clarify whether scoring of responses is optional or if every response is consistently assessed, with scores made visible to parents and teachers. For students, the perception of being judged can significantly impact their mindset, as opposed to feeling genuinely supported. Research in biology indicates that environments perceived as judgmental can elevate cortisol and adrenaline levels, triggering a threat response and hindering learning. In contrast, supportive settings help regulate cortisol and raise oxytocin levels, which increase motivation and openness to new experiences and learning opportunities.

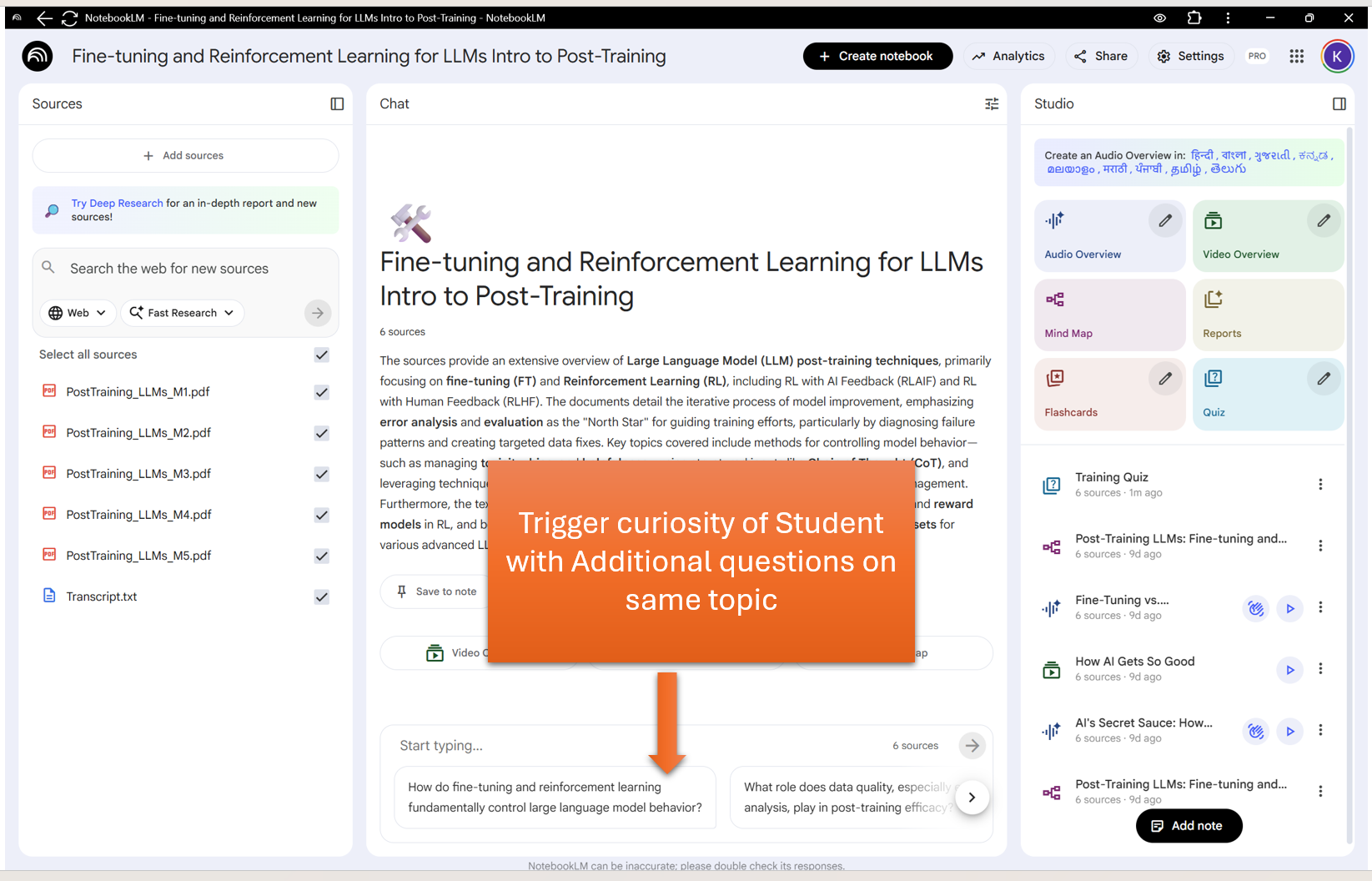
|  |  |  |  |
| --- | --- | --- | --- |
| Situation | Typical hormones / systems | Brain pattern | How it tends to feel |
| Feeling judged / criticized | ↑ Cortisol, ↑ adrenaline; strong HPA-axis & sympathetic activation | dACC + anterior insula (“social pain”), threat circuits more active | Tight chest, racing thoughts, self-attack, urge to escape or fight |
| Feeling supported / helped | Lower or more regulated cortisol; ↑ oxytocin; more parasympathetic activity | Amygdala threat response damped; better frontal regulation, bonding circuits engaged | Calmer body, warmer feelings, clearer thinking, more willingness to try things |

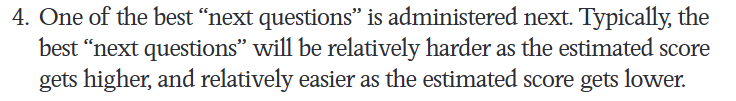
Therefore, implementing continuous scoring for every question may serve as a potential failure mode, ultimately hindering learning rather than facilitating it.

References:

* <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00014/full?utm_source=chatgpt.com>
* <https://www.sciencedirect.com/science/article/abs/pii/S0006322303004657?utm_source=chatgpt.com>

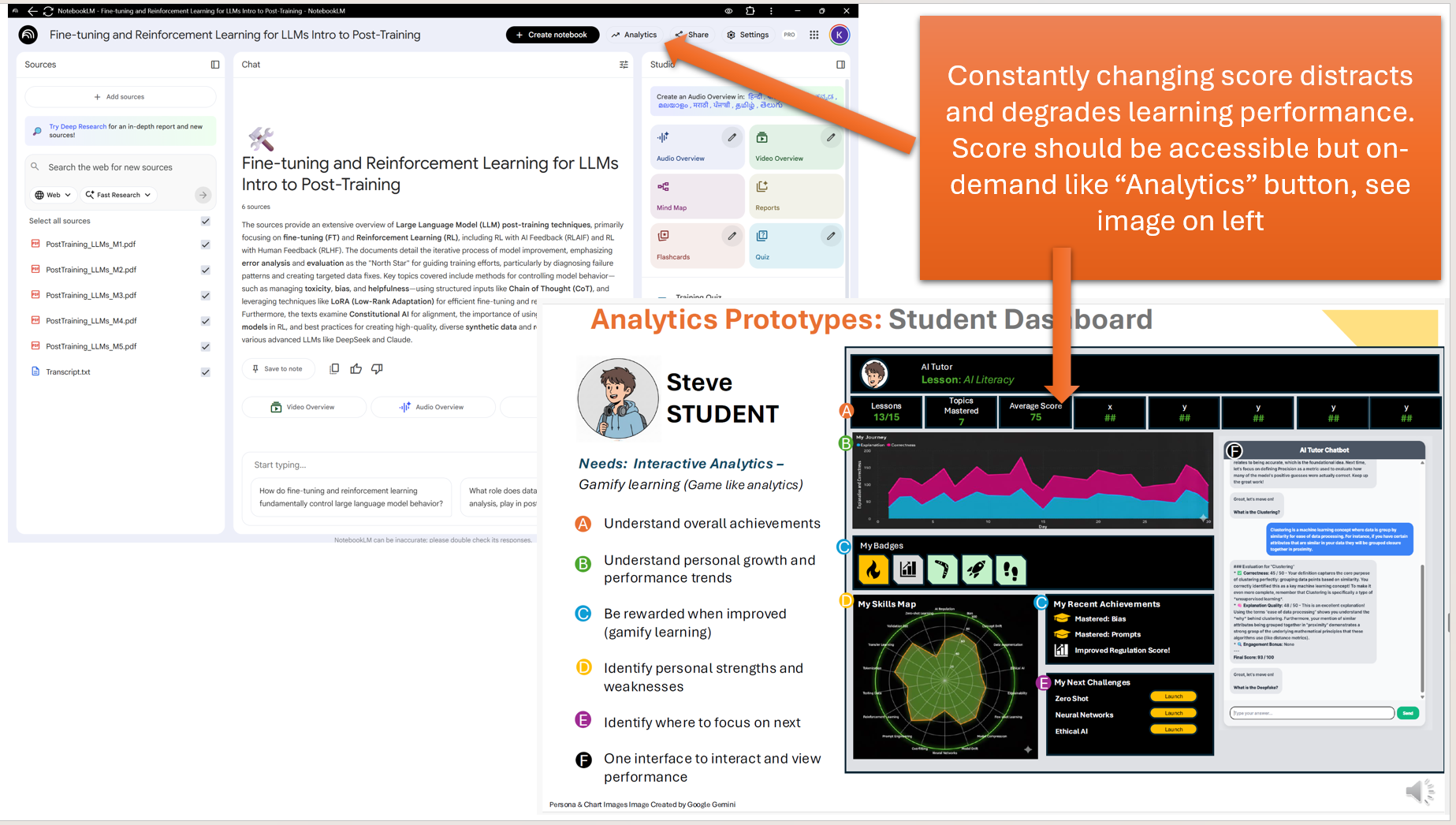
1. Tool could promote reward hacking: - The live scoring feature of AI Tutor tool may inadvertently encourage superficial learning rather than genuine understanding. As students become aware that their responses are monitored and evaluated by parents and teachers, they may be incentivized to answer every question correctly simply to demonstrate satisfactory progress and avoid negative consequences or receive rewards. For instance, students could potentially use external resources such as chatbots on personal devices to obtain the correct answers, subsequently entering them into the AI Tutor platform to achieve high scores and create an appearance of academic improvement. Furthermore, it is impractical for teachers to maintain exam conditions around the clock for AI Tutor usage. High performance on the AI Tutor could therefore foster a misleading impression of learning, providing students with opportunities to exploit the reward system rather than focusing on authentic comprehension and preparation for actual examinations. Consequently, there is a significant risk that features like live scoring may facilitate reward-driven behavior, undermining the intended educational objectives for all personas Ann Admin, Tom teacher, Paula parent and Steve student. A crucial consideration that AI tutors may overlook is that learning is facilitated by making mistakes; therefore, penalizing every mistake can trigger a survival response among students rooted in fear of punishment (from teachers and parents), rather than fostering a mindset focused on learning and growth.
2. Stimulate Student Curiosity: As an AI Tutor designed to facilitate learning, the system can offer a selection of follow-up questions that vary in difficulty or depth, allowing students to choose rather than following a predetermined sequence. For sessions with time constraints, scoring should be adaptive to reflect question difficulty, as achieving 100% on easy questions does not equate to attaining 100% on more challenging questions in terms of cognitive effort or knowledge demonstration. For context, the screenshot below illustrates the logic used for selecting subsequent questions in the GMAT exam’s computer adaptive testing framework.





Source:<https://www.gmac.com/-/media/files/gmac/research/validity-and-testing/demystifyingthegmat_computeradaptivetesting.pdf>

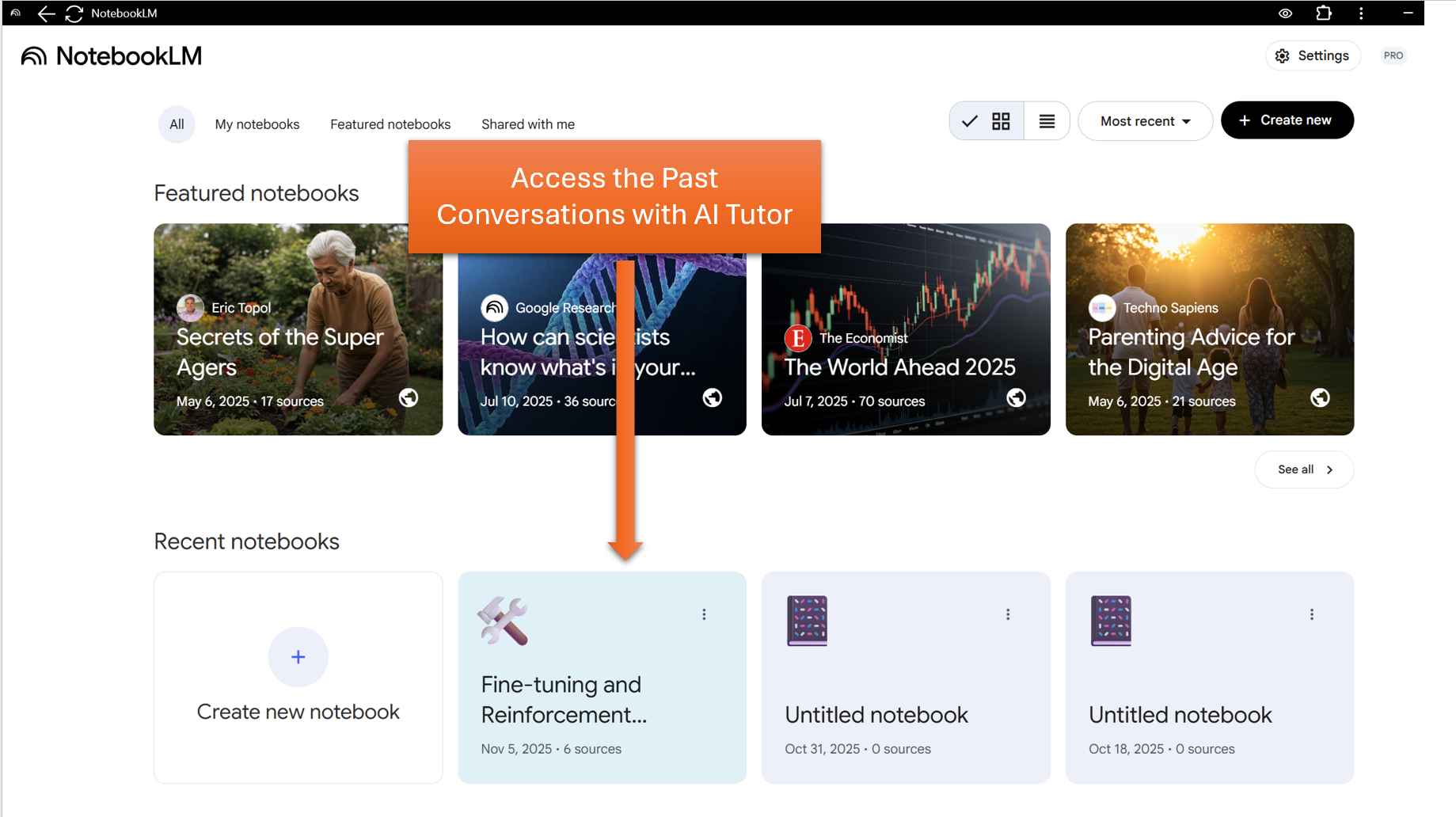
1. Frequent updates to the displayed score during learning sessions can be distracting, drawing attention away from the instructional content and interrupting the learner's ability to achieve a state of flow, which may negatively affect learning performance. While having access to up-to-date scores is beneficial for identifying areas for improvement, such information should be accessible on demand—similar to the “Analytics” button featured in Notebooklm, as illustrated in the image below—to help minimize distractions.



<https://www.youtube.com/watch?v=0rIjFCNay2Q>

<https://medium.com/a-teachers-hat/the-state-of-flow-while-learning-d1d15f332fa0>

1. Accessing Previous Conversations with AI Tutor: The current system architecture does not specify a method for retrieving a student’s past interactions with the AI Tutor. Enabling this capability could be beneficial for reviewing previous scores and educational progress. For reference, the screenshot of Notebooklm below shows past sessions with the AI Tutor.



1. Dark Mode/ Light Mode: Provide ability to personalize Luminance contrast and Text size to improve reading performance. For reference, the screenshot of Notebooklm below shows options to change mode.

Source: https://www.sciencedirect.com/science/article/pii/S0042698919302111

A screenshot of a computer

AI-generated content may be incorrect.

1. Gather Student Preferences: Enhance learning outcomes by tailoring response styles to each student's individual learning approach through Reinforcement Learning with Human Feedback (RLHF), utilizing preference collection as illustrated in the image below:

A screenshot of a computer

AI-generated content may be incorrect.

1. Discord support: Provide support on discord until dedicated enterprise support becomes available to students, teachers, parents and admin, refer image below:

A screenshot of a computer

AI-generated content may be incorrect.

1. Integrate with School Learning Management System (LMS): In order to ensure seamless transfer of grades, scores and learning history without manual effort, integrate the AI Tutor with the LMS such as Brightspace.
2. Parental Controls: Implement parental controls to restrict student access during late-night hours, helping prevent sleep deprivation and excessive usage. This measure prioritizes student wellbeing and reduces the risk of addiction, refer link below:

Source: https://help.openai.com/en/articles/12315553-parental-controls-on-chatgpt-faq