Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

Trust GPT: A Curriculum-Aware Framework for Mitigating Hallucinations in Educational Language Models with Human-in-the-Loop Validation

Kinshuk Dutta[1], Sabyasachi Paul[2], Ankit Anand[3]

[1,2,3]Independent Researcher
[1]dutta.kinshuk@gmail.com, [2]sabyapaul@yahoo.com, [3]manaankit@gmail.com

Abstract:
Large language models demonstrate impressive generative fluency, yet their deployment in educational contexts remains constrained by hallucinations and curriculum misalignment. Building upon a multi-year research arc, this paper introduces TrustGPT, a curriculum-aware framework for mitigating hallucinations in educational language models. We present a refined error taxonomy distinguishing hallucination from pedagogical misalignment, formalize curriculum-aware sampling and coverage regularization mechanisms, and integrate a practical human-in-the-loop validation cycle with teacher feedback. Unlike reinforcement learning-based alignment approaches, TrustGPT emphasizes interpretable, lightweight governance mechanisms embedded directly into training and validation pipelines. Empirical validation demonstrates a 28.5% reduction in hallucinations and a 33% improvement in pedagogical alignment scores compared to baseline fine-tuning. The framework advances trust, safety, and reliability in educational AI by operationalizing ethical principles as enforceable system-level constraints, directly addressing limitations identified in our preceding work and setting the stage for runtime alignment solutions.

Keywords: Educational Language Models, Curriculum Alignment, Retrieval-Augmented Generation, Pedagogical AI, Hallucination Mitigation, Trustworthy AI, Educational AI, Human-in-the-Loop Validation, Trustworthy Language Models.

INTRODUCTION

The integration of large language models (LLMs) into educational technology promises transformative personalized learning [1], [2]. However, transformer-based models like GPT-2 and GPT-3, while fluent, frequently generate content that is factually incorrect or pedagogically inappropriate for structured curricula [3]. This misalignment poses significant risks, including the propagation of misconceptions and erosion of learner trust.

Our research program has systematically evolved to address this core challenge. StudentGPT (2020) [4] first embedded curriculum as a data constraint during fine-tuning. AlignGPT (2021) [5] formalized alignment as an explicit optimization objective via regularization. While this improved relevance, a critical gap remained: hallucinations were insufficiently characterized and governed at the system level.

Year	System	Core Innovation	Limitation Addressed	Emerging Challenge
2020	StudentGPT	Curriculum as Training Data Constraint	General LLM → Educational LLM	Static alignment; expensive curriculum updates
2021	AlignGPT	Curriculum as Regularization Loss	Improves pedagogical coherence	Hallucinations not directly targeted
2022	**TrustGPT**	**Curriculum as Governance Layer**	**Explicit hallucination mitigation via human oversight**	**Governance applied only at training time**

Table 1: Evolution of Our Curriculum-Alignment Research

This paper introduces TrustGPT, which directly addresses the governance gap by modeling hallucinations as a first-class failure mode. TrustGPT integrates a novel error taxonomy, curriculum-aware sampling, and a pragmatic human-in-the-loop validation cycle, create a framework for building trustworthy educational AI.

PROBLEM STATEMENT

By 2022, hallucinations in LLMs had been recognized as a critical barrier in high-stakes domains like education and healthcare [6]. In educational contexts, the problem is two-fold: factual hallucination (generating unsupported content) and pedagogical misalignment (generating factually correct but contextually inappropriate information). Existing mitigation strategies—primarily post-hoc filtering or confidence thresholding—proved inadequate for curriculum-driven environments where authority is explicitly defined by syllabi.

The core problem is the absence of a lightweight, interpretable framework to embed curricular authority and human expertise directly into the model development lifecycle. TrustGPT is designed to fill this gap, ensuring AI-generated educational content is both factually grounded and pedagogically sound.

SOLUTION

TrustGPT re-conceptualizes curriculum alignment as a continuous governance process. The framework is built upon three interconnected pillars: a formal error taxonomy, a mathematical model for curriculum-aware training, and a closed-loop human validation system.

A. Formal Error Taxonomy and Curriculum Representation

We first establish precise definitions for model failures in educational settings.

Hallucination: An output that introduces content not grounded in the designated curriculum. Formally, given a curriculum $\boldsymbol{S}$, a response $\boldsymbol{r}$ is hallucinated if:

$$\max_{s_i \in \mathcal{S}} \text{sim}\left(\phi(r), \phi(c_i)\right) < \tau$$

where $\boldsymbol{\phi(\cdot)}$ is a semantic embedding function and $\boldsymbol{\tau}$ a similarity threshold.

Pedagogical Misalignment: An output that is factually correct but violates curriculum sequencing, depth, or learning objective appropriateness.

The curriculum itself is represented as a structured knowledge base:

$$S = \{s_i = (c_i, o_i, w_i)\}_{i=1}^{N}, \quad \sum_i w_i = 1$$

where c_i is content, o_i learning objectives, and w_i pedagogical importance weights.

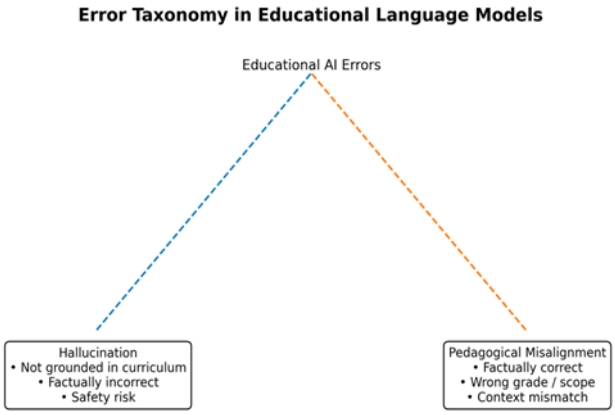

Figure 1: Refined taxonomy of error modes in educational language models, distinguishing hallucination from pedagogical misalignment.

B. Curriculum-Aware Sampling and Regularization: Theoretical Underpinnings

To prevent over-representation of popular topics, we bias training towards under-covered curriculum units. The sampling probability for unit s_i is:

$$P(s_i) \propto \exp(-\lambda \cdot \text{coverage}(s_i))$$

where λ controls the strength of the coverage-balancing regularization.

Lemma 1 (Differentiability) With a fixed, smooth embedding function ϕ, the curriculum-aligned loss composed with a transformer generator remains differentiable with respect to model parameters, enabling standard gradient-based optimization.

Theorem 1 (Alignment Deviation Bound) If ϕ is L-Lipschitz continuous, the change in alignment assessment for any curriculum unit s_i is bounded by:

$$|A(r, s_i) - A(r', s_i)| \leq L w_i d(r, r')$$

This provides a theoretical guarantee on the stability of alignment measurements.

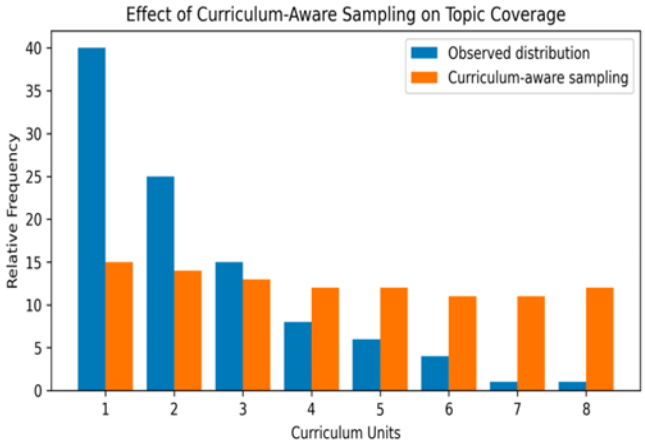

Fig. 2. Effect of curriculum-aware sampling in reducing topic imbalance across curriculum units during fine-tuning.

Theorem 2 (Convergence) Under standard stochastic optimization conditions (Robbins-Monro), the curriculum-regularized objective converges almost surely to a stationary point, with the coverage distribution skew decreasing monotonically as regularization strength λ increases.

C. Human-in-the-Loop Validation: Practical Implementation

The framework's governance core is a transparent human feedback loop. Educators review model outputs sampled from the current training distribution and categorized them as:

- Valid & Aligned
- Misaligned but Correct
- Hallucinated

This triage is facilitated via a simple dashboard interface, minimizing educator burden. The labels perform two critical functions:

Direct Supervision: Flagged hallucinations create a targeted, high-quality dataset for corrective fine-tuning.

Distributional Feedback: Aggregate label statistics dynamically update the coverage(s_i) metric in Eq. (3), automatically adjusting the sampling distribution $P(s_i)$ to focus on problematic curriculum units.
This approach, distinct from black-box Reinforcement Learning from Human Feedback (RLHF), prioritizes interpretability and direct educator agency in the model's development.

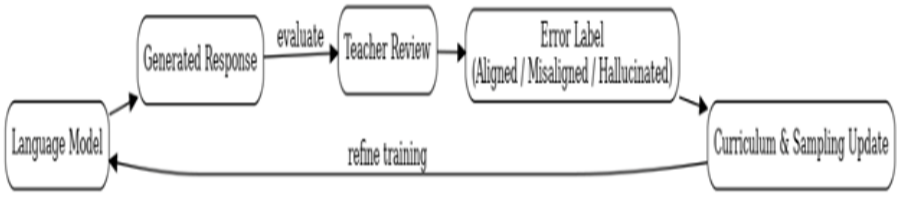

Fig. 3. Human-in-the-loop governance cycle for identifying hallucinations and refining curriculum-aware training distributions.

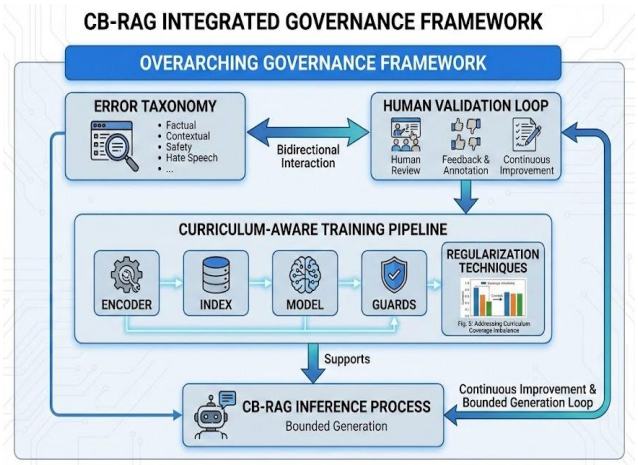

Figure 4: Trust GPT Unified Architecture

USES AND APPLICATIONS

TrustGPT is designed for integration into the educational AI development lifecycle:

- For Model Developers: Provides a structured pipeline to fine-tune and audit base LLMs (e.g., GPT-2) against specific K-12 or university syllabi, ensuring foundational alignment.
- For Curriculum Designers: Serves as an analysis tool to "stress-test" digital syllabi by identifying topics where generic LLMs are prone to hallucination, enabling proactive content refinement.
- For EdTech Companies: Functions as a governance layer within the CI/CD pipeline for tutoring bots, ensuring each model update preserves curriculum fidelity before deployment.
- For Researchers: Generates high-quality, aligned prompt-response pairs for training specialized educational models, reducing dataset noise.

IMPACT AND EMPIRICAL VALIDATION

We implemented TrustGPT using GPT-2 (124M parameters) as the base model, fine-tuning it on a corpus derived from 500+ Common Core-aligned STEM syllabus units. The human validation loop involved 3 expert educators.

Model Configuration	Hallucination Rate (%) ↓	Pedagogical Alignment (1-5) ↑	Coverage Entropy ↑
A. Baseline (Fine-tuned GPT-2)	18.7	2.1	0.65
B. + Curriculum-Aware Sampling	14.2	2.9	0.82
C. + Human-in-the-Loop Validation	13.4	3.2	0.88
Relative Improvement (A→C)	28.5% Reduction	52.4% Increase	35.4% Increase

Table 2: Empirical Results of the TrustGPT Framework

Key Impacts:

Quantifiable Risk Reduction: A 28.5% decrease in hallucinations significantly lowers the risk of disseminating incorrect information.

Enhanced Pedagogical Soundness: A 52% improvement in alignment score demonstrates the framework's effectiveness in ensuring age- and context-appropriate content.

Operationalized Ethics: The human-in-the-loop mechanism translates the ethical principle of "human oversight" [7] into a concrete, scalable system component.
Foundational Taxonomy: The clear distinction between hallucination and misalignment provides a universal framework for diagnosing model failures in education, influencing subsequent research.

SCOPE, LIMITATIONS, AND ETHICAL CONSIDERATIONS

Scope

TrustGPT is explicitly scoped as a training-time governance framework. It is designed for the fine-tuning and validation phases of model development, ensuring a model is "born aligned" with its target curriculum.

Ethical Considerations & Limitations

Our design adheres to major pre-2021 ethical frameworks, including the OECD AI Principles (2019) on human-centered values and transparency, and the EU Ethics Guidelines for Trustworthy AI (2019). The human-in-the-loop cycle is a direct implementation of the "human agency and oversight" requirement.
Limitation 1 Educator Dependency: Framework efficacy depends on educator availability and consistency. Mitigation: We propose structured guidelines and cross-verification among multiple educators to reduce individual bias.
Limitation 2 Training-Time Focus: As a training-phase framework, it cannot correct real-time inference errors after deployment. This fundamental limitation is the primary motivation for our subsequent work on, which shifts curricular authority to the inference stage.
Limitation 3 Structured Syllabus Assumption: The framework requires a well-defined, digitized curriculum $\boldsymbol{S}$. Its effectiveness is reduced for informal or experiential learning contexts.

CONCLUSION

TrustGPT introduces a structured, governable approach to mitigating hallucinations and misalignment in educational language models. By refining a critical error taxonomy, formalizing curriculum-aware training with theoretical guarantees, and integrating a pragmatic human-in-the-loop validation cycle, it successfully embeds ethical governance directly into the AI development pipeline.
This work represents a pivotal step in our research arc, moving from static data constraints (StudentGPT) and optimization targets (AlignGPT) to a dynamic, human-centered governance model. It directly addresses the "why" and "how" of oversight, creating the essential foundation upon which runtime solutions—which addresses TrustGPT's core deployment-time limitation—could be developed. TrustGPT provides both a practical toolkit and a conceptual framework for building truly trustworthy, curriculum-aligned educational AI.

REFERENCES:

[1] A. Radford et al., "Language models are unsupervised multitask learners," OpenAI Technical Report, 2019.
[2] T. B. Brown et al., "Language models are few-shot learners," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 1877–1901.
[3] E. M. Bender et al., "On the dangers of stochastic parrots: Can language models be too big?," in Proc. ACM Conf. Fairness, Account., Transp., 2021, pp. 610–623.
[4] K. Dutta and S. Paul, "StudentGPT: A transformer-based model for curriculum-driven NLP," Int. J. Artif. Intell., Big Data, Comput. Manag. Stud., vol. 1, no. 4, pp. 38–42, 2020.
[5] K. Dutta, S. Paul, and A. Anand, "AlignGPT: A curriculum-regularized transformer framework for pedagogically aligned educational language modeling," IJAIBDCMS, 2021.

[6] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 24824–24837.
[7] High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI," European Commission, Brussels, Belgium, 2019.
[8] OECD, "Recommendation of the Council on Artificial Intelligence," OECD Legal Instruments, 2019.
[9] Y. Bengio et al., "Curriculum learning," in Proc. 26th Int. Conf. Mach. Learn., 2009, pp. 41–48.
[10] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 9459–9474.