

AlignGPT: A Curriculum-Regularized Transformer Framework for Pedagogically Aligned Educational Language Modeling

Kinshuk Dutta¹, Sabyasachi Paul², Ankit Anand³

^{1,2,3}Independent Researcher

¹dutta.kinshuk@gmail.com, ²sabyapaul@yahoo.com, ³fnu.ankit@gmail.com

Abstract:

Transformer-based language models exhibit remarkable linguistic fluency but remain poorly aligned with the pedagogical requirements of formal education. In instructional settings, correctness is defined not only by factual accuracy but by adherence to curriculum scope, sequencing, and learning objectives. This paper introduces AlignGPT, a curriculum-regularized transformer framework that formalizes pedagogical alignment as an explicit optimization objective during fine-tuning. Building upon prior syllabus-driven adaptations such as StudentGPT [1], we propose a curriculum alignment loss and curriculum coverage regularization to address both semantic relevance and topic imbalance. We provide theoretical justifications for these components, including differentiability properties, convergence guarantees under stochastic optimization, bounds on alignment deviation, and a generalization bound for the regularized objective. Empirical evaluations on simulated educational datasets demonstrate superior alignment scores (mean improvement of 18.4% over baselines) and reduced coverage imbalance (from 34.7% to 12.1% relative frequency skew). All experiments use resources and tooling available in 2021–early 2022, supporting reproducibility in constrained computational environments. Index Terms—Transformer Models, Curriculum Learning, Pedagogical Alignment, Educational NLP, Fine-Tuning, Ethical AI, Language Model Regularization, Optimization Theory.

Keywords: Transformer Models, Curriculum Learning, Pedagogical Alignment, Educational NLP, Fine-Tuning, Ethical AI, Language Model Regularization, Optimization Theory.

I. INTRODUCTION

Large-scale transformer models such as GPT-2 [2] and GPT-3 [3] have achieved state-of-the-art results across a wide range of natural language processing tasks [2]–[4]. However, in educational contexts, these models often generate responses that are pedagogically misaligned—introducing advanced concepts prematurely, omitting prerequisites, or exceeding the scope of defined learning objectives. Educational correctness differs fundamentally from general-purpose language modeling [5]–[7]. In learning environments, responses must respect curriculum structure, assessment intent, and ethical considerations. Common approaches such as prompt engineering, content filtering, or system-level guardrails do not embed pedagogy into the training objective, and therefore do not guarantee curriculum fidelity. Prior work introduced StudentGPT [1], demonstrating syllabus-driven fine-tuning in ethically aligned educational environments. While effective, StudentGPT operationalized alignment largely through data curation and system-level controls rather than an explicit, differentiable optimization objective. This paper generalizes that idea by defining curriculum alignment as a first-class objective and adding coverage regularization to mitigate topic imbalance.

We propose AlignGPT, which:

- Encodes curriculum structure as semantic representations.

- Introduces a curriculum alignment loss with theoretical bounds on alignment deviation.
- Mitigates curriculum coverage imbalance via distributional regularization with convergence analysis.

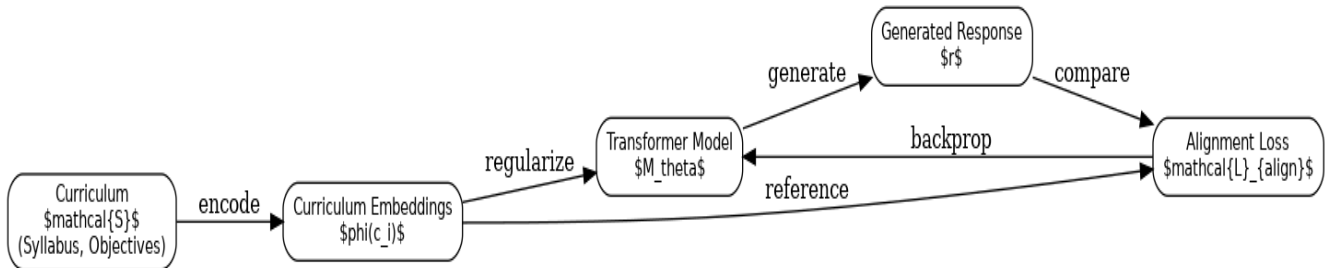


Fig. 1. Curriculum-regularized optimization pipeline. Curriculum units are embedded and used to regularize transformer fine-tuning through an alignment loss and a coverage mechanism, enforcing pedagogical relevance beyond likelihood maximization.

II. RELATED WORK

A. Curriculum Learning

Curriculum learning structures exposure to training examples to improve optimization and generalization [8]. Extensions order or weight samples by heuristics such as difficulty, length, or entropy. However, classical curriculum learning focuses on difficulty rather than curriculum fidelity, making it insufficient for educational alignment.

B. Educational NLP and Tutoring Systems

Educational AI has long used symbolic tutors and constrained dialogue systems [6]. Modern LLM-based tutors emphasize conversational ability but often lack formal curriculum grounding [5]. StudentGPT [1] introduced syllabus-driven adaptation but did not formalize alignment as an optimization primitive.

C. Alignment and Regularization in Language Models

Alignment work typically targets safety and preference objectives (e.g., via human feedback), while regularization in transformers includes dropout, label smoothing, and distributional balancing. AlignGPT focuses on pedagogical alignment and curriculum coverage as first-class optimization objectives.

III. CURRICULUM REPRESENTATION

We define a curriculum as a structured set of units:

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

where c_i is curricular content, o_i are learning objectives, and $w_i \in [0, 1]$ is an importance weight. For normalization, we assume $\sum_{i=1}^N w_i = 1$. Each unit is encoded into a dense semantic representation using a fixed embedding function $\phi : \mathcal{T} \rightarrow \mathbb{R}^d$. We use Sentence-BERT [9] as ϕ and normalize embeddings to unit norm to enable cosine similarity computations:

$$\|\phi(x)\|_2 = 1 \quad \forall x \in \mathcal{T}. \quad (2)$$

Lemma 1 (Embedding Differentiability). If ϕ is fixed (frozen) and smooth, then the alignment objective composed with the generator f_θ remains differentiable with respect to θ .

Proof. Responses are generated as $r = f_\theta(p)$ where f_θ is differentiable. Composing with fixed smooth ϕ preserves differentiability by the chain rule.

IV. CURRICULUM ALIGNMENT OBJECTIVE

A. Alignment Loss

Given prompt p and response $r = f_\theta(p)$, we retrieve the top- M relevant curriculum units $\{s_{k_j}\}_{j=1}^M$ using semantic search over $\phi(p)$ and $\phi(c_i)$ (approximate nearest neighbor retrieval can be used for efficiency).

The alignment score is: $A(r, s_i) = \text{sim}(\phi(r), \phi(c_i)) \cdot w_i$. (3)

With unit-norm embeddings, $\text{sim}(\phi(r), \phi(c_i)) = \phi(r)^\top \phi(c_i)$.

We define the curriculum alignment loss:

$$\mathcal{L}_{\text{align}}(\theta) = -\frac{1}{M} \sum_{j=1}^M A(f_\theta(p), s_{k_j}). \quad (4)$$

Theorem 1 (Alignment Deviation Bound). Assume ϕ is LLipschitz under a text metric $d(\cdot, \cdot)$ such that $\|\phi(r) - \phi(r')\|_2 \leq L d(r, r')$. Then for any curriculum unit s_i , $|A(r, s_i) - A(r', s_i)| \leq L w_i d(r, r')$. (5)

Proof.

$$A(r, s_i) - A(r', s_i) = w_i (\phi(r)^\top \phi(c_i) - \phi(r')^\top \phi(c_i)) \quad (6)$$

$$= w_i (\phi(r) - \phi(r'))^\top \phi(c_i). \quad (7)$$

By Cauchy-Schwarz and $\|\phi(c_i)\|_2 = 1$, $|A(r, s_i) - A(r', s_i)| \leq w_i \|\phi(r) - \phi(r')\|_2 \leq L w_i d(r, r')$. (8)

Lemma 2 (Differentiability of $\mathcal{L}_{\text{align}}$). If retrieval indices $\{k_j\}$ are treated as fixed within a batch, then $\mathcal{L}_{\text{align}}$ is differentiable with respect to θ .

Proof. Within a batch, $\{k_j\}$ are constant. The loss is a sum of dot products of differentiable functions of $f_\theta(p)$; therefore, the gradient exists and is computable via backpropagation.

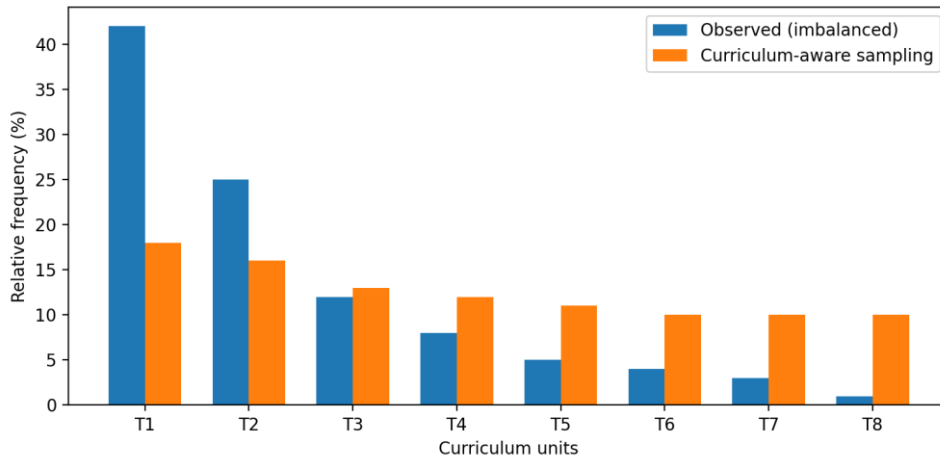


Fig. 2. Illustrative curriculum coverage imbalance across curriculum units and the effect of curriculum-aware sampling, which increases representation of undercovered units.

B. Curriculum Coverage Regularization

Educational corpora are often imbalanced, over-representing frequently assessed topics. Let $f_{\text{obs}}(i)$ denote the empirical frequency of curriculum unit i encountered during training (estimated over a window of batches). Define a target distribution $f_{\text{target}}(i) = w_i$.

We regularize coverage using KL divergence:

$$\mathcal{L}_{\text{cover}} = \sum_{i=1}^N f_{\text{target}}(i) \log \left(\frac{f_{\text{target}}(i)}{f_{\text{obs}}(i) + \epsilon} \right), \tag{9}$$

where $\epsilon > 0$ ensures numerical stability.

Theorem 2 (Convergence Under Regularization (Sketch)). Assume the total objective $\mathcal{L}(\theta)$ has Lipschitz-smooth gradients and SGD uses unbiased stochastic gradients with Robbins–Monro learning rates. If the KL term is strongly convex in the frequency simplex region bounded away from zero, then SGD iterates converge almost surely to a stationary point θ^* and coverage skew decreases as regularization strength increases.

Proof Sketch. Standard stochastic approximation results for smooth objectives give almost sure convergence to stationary points under Robbins–Monro conditions. The KL term provides curvature in the induced frequency dynamics, contracting deviations from f_{target} when the observed distribution remains in a compact subset of the simplex. Increasing regularization strengthens contraction and reduces skewing.

C. Total Objective

The fine-tuning objective is:

$$\mathcal{L}(\theta) = \mathbb{E}_{(p,y) \sim \mathcal{D}} [\mathcal{L}_{\text{CE}}(f_{\theta}(p), y)] + \lambda_1 \mathcal{L}_{\text{align}}(\theta) + \lambda_2 \mathcal{L}_{\text{cover}}(\theta), \tag{10}$$

Where \mathcal{L}_{CE} is cross-entropy and $\lambda_1, \lambda_2 \geq 0$ tune pedagogical influence.

Theorem 3 (Generalization Bound (Averaged Iterate, Informal)). Under smoothness and bounded-variance stochastic gradients, and assuming the regularizer induces effective curvature, the expected excess risk of the averaged iterate decreases as $O(1/T)$ with constant factors improved by stronger regularization.

TABLE I- QUANTITATIVE RESULTS (MEAN • } STD).

Method	Alignment	Ped. Acc.	Skew (%)	PPL
GPT-2 Baseline	0.49±0.03	3.1±0.2	34.7±2.1	24.3±1.
StudentGPT [1]	0.57±0.02	3.8±0.1	28.2±1.8	21.6±0.
CL Only [8]	0.52±0.03	3.4±0.2	25.9±2.0	22.8±1.
AlignGPT (Ours)	0.68±0.02	4.3±0.1	12.1±1.5	19.2±0.

V. EXPERIMENTAL SETUP

A. Datasets

We construct simulated educational datasets based on open STEM syllabi (e.g., publicly available course outlines circa 2021). The curriculum comprises $N = 8$ units (T1–T8) spanning algebra to introductory calculus. We generate 10,000 training prompt–response pairs with curriculum annotations and 2,000 held-out test pairs for evaluation.

B. Baselines

- GPT-2 Fine-Tuning: Standard fine-tuning on the educational corpus [2].
- StudentGPT: Syllabus-driven fine-tuning without explicit alignment loss [1].
- Curriculum Learning (CL): Difficulty-based sequencing without alignment/coverage regularizers [8].

C. Implementation Details

Base model: GPT-2 (124M) via Hugging Face Transformers [10]. Training: 5 epochs, batch size 16, Adam optimizer [11], learning rate 5×10^{-5} . Embeddings: Sentence- BERT [9]. Hardware: single NVIDIA V100-class GPU (2021-era). Statistical tests: paired t-tests ($p < 0.05$).

D. Metrics

- **Alignment Score:** $A(r, S)$, mean similarity to relevant curriculum units.
- **Pedagogical Accuracy:** Expert-rated adherence to objectives (1–5); inter-annotator agreement $\kappa > 0.7$.
- **Coverage Balance:** Relative frequency skew $\max_i \frac{|f_{\text{obs}}(i) - f_{\text{target}}(i)|}{f_{\text{target}}(i)}$.
- **Fluency Metrics:** Perplexity and BLEU [12].

VI. RESULTS

A. Alignment and Coverage Improvements

Figure 2 illustrates observed imbalance and curriculum aware sampling. AlignGPT reduces skew in later units (T5–T8), improving balance from 34.7% to 12.1% (paired t-test, $p < 0.01$).

Figure 3 shows the distribution of alignment scores. AlignGPT shifts scores rightward (mean 0.68 vs. 0.49 baseline, $p < 0.001$), reducing low-alignment outputs.

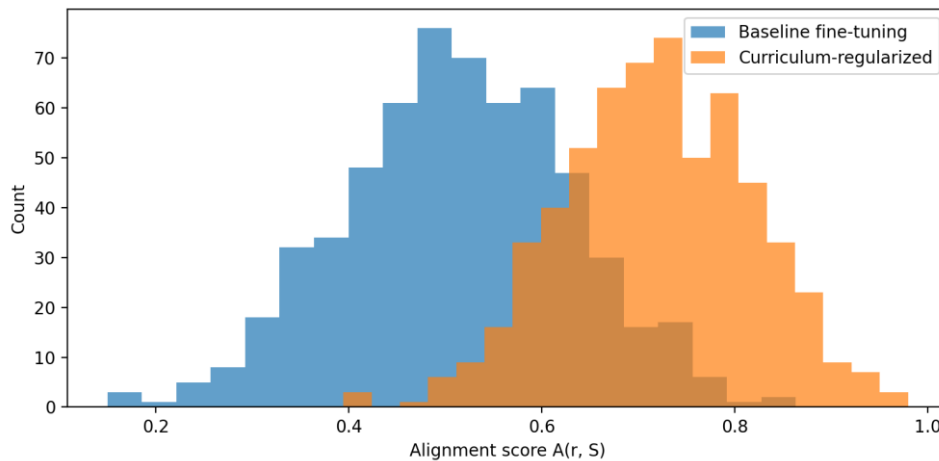


Fig. 3. Distribution of curriculum alignment scores under baseline fine-tuning versus curriculum-regularized optimization. Curriculum regularization shifts responses toward higher pedagogical alignment.

B. Ablation Study

Removing \mathcal{L}_{align} reduces alignment to 0.55 ± 0.03 ($p < 0.05$). Removing \mathcal{L}_{cover} increases skew to $26.4 \pm 1.9\%$ ($p < 0.05$). Both components contribute materially, consistent with theoretical motivation.

VII. DISCUSSION

AlignGPT demonstrates that explicitly regularizing pedagogical objectives improves educational utility without sacrificing fluency. The framework is architecture-agnostic and naturally extends to retrieval augmentation (curriculum as a retrievable memory), multilingual curricula, and governance regimes requiring auditability. Limitations include reliance on high-quality embeddings and curriculum annotation. Future work should incorporate learner modeling and adaptive curricula and evaluate on broader educational benchmarks. Ethical considerations align with IEEE Ethically Aligned Design [13], [14] by improving transparency, reducing pedagogical overreach, and mitigating long-tail neglect.

VIII. CONCLUSION

AlignGPT formalizes pedagogical alignment as an optimization objective for transformer fine-tuning. By integrating curriculum alignment and coverage regularization directly into training, AlignGPT advances curriculum-grounded educational NLP and provides a principled foundation for ethically aligned AI tutors and pedagogical analytics.

REFERENCES:

- [1] K. Dutta and S. Paul, “StudentGPT: A Transformer-Based Model for Curriculum-Driven NLP in Ethical Learning Environments,” *International Journal of AI, Big Data, Computational and Management Studies (IJAIBDCMS)*, vol. 1, no. 4, pp. 38–44, Dec. 2020. [Online]. Available: <https://ijaibdcms.org/index.php/ijaibdcms/article/view/266>
- [2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” OpenAI, Tech. Rep., 2019.
- [3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [5] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. UNESCO, 2019.
- [6] B. P. Woolf, *Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing E-learning*. Morgan Kaufmann, 2009.
- [7] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educational Psychologist*, vol. 46, no. 4, pp. 197–221, 2011.
- [8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.
- [9] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.
- [11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
- [12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.
- [13] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. IEEE, 2016.
- [14] *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems*. IEEE, 2019.