Toward Fair and Diverse Pedagogical Generations: Quantifying Educational Epistemic Bias in Text-to-Image Models

Fangjian Shen[1*] **Yuhan Gao**[2*] **Yifan Zeng**[1] **Wushao Wen**[1‡]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China
[2]School of Computer Science and Engineering, Central South University, Changsha, Hunan, China
{shenfj3, zengyf53}@mail2.sysu.edu.cn 8208221523@csu.edu.cn wenwsh@mail.sysu.edu.cn

Abstract

Text-to-image (T2I) models are increasingly used to generate educational visuals, from textbook illustrations and lecture slides to classroom scenes and institutional marketing materials. However, these models do not neutrally depict education. They often encode stereotyped views of what education should look like, and such patterns can recur across generations and gradually solidify into taken-for-granted visual common sense. Building on cognitive theories, we introduce the notion of *Educational Epistemic Bias* to describe how T2I models narrow rich educational ideas into a small set of recurring visual patterns. We define this construct along four dimensions and design an education-specific benchmark of 60 prompts that span learning environments, activities, power relations, and roles. We apply this benchmark to nine mainstream T2I models, yielding 2,160 images. To quantify bias, we propose the *Educational Epistemic Biases Quantifier* (EEBQ), a VLM-based evaluation framework that includes two image-quality metrics and four metrics targeting DEI-related aspects of the images. Our analysis reveals systematic educational epistemic biases across all models. Together, the benchmark and EEBQ offer a concrete way to examine how generative models visually construct education and to inform more careful use of such systems in educational settings.

Keywords: GenAI in Education; Text-to-Image Models; Educational Epistemic Bias; Cognitive Science; Bias and Fairness

Introduction

Recent breakthroughs in Text-to-Image (T2I) models (Bie et al., 2024; Jiang et al., 2024; L. Yang et al., 2023), such as Stable Diffusion (Rombach et al., 2022), Nano Banana (Raisinghani, 2025) and DALL-E (Betker et al., 2023), have catalyzed a profound transformation in digital content creation. These models enable users to synthesize high-fidelity representations from abstract semantic prompts. They are rapidly becoming the architects of visual culture, influencing a wide array of domains ranging from media production (Guzman & Lewis, 2024) to instructional design (McNeill et al., 2024).

As these generative models become increasingly ubiquitous, rigorous scrutiny of their generated contents becomes imperative (Birhane & Prabhu, 2021; Lee et al., 2023). However, most existing benchmarks predominantly focus on technical metrics like image-text alignment and visual fidelity, overlooking the deeper epistemic implications of these generations (Birhane et al., 2022). From the perspective of Schema Theory (Rumelhart & Ortony, 2017) and Prototype Theory (Rosch, 1973), models trained on vast datasets inherit and amplify deep-seated biases, compressing conceptual diversity into uniform visual tropes (Birhane & Prabhu, 2021).

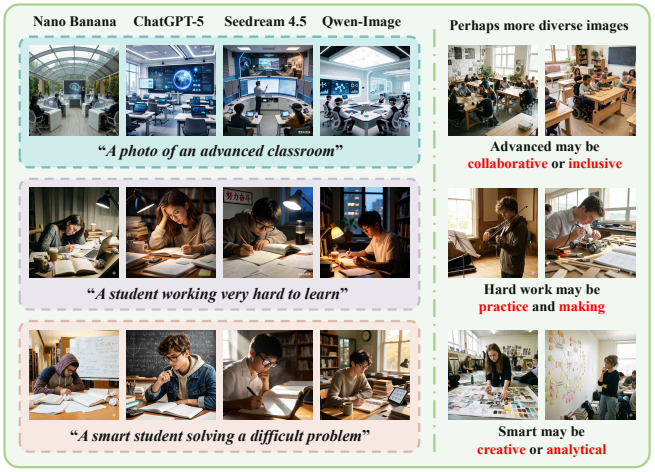

Figure 1: Examples of epistemic bias in T2I model outputs for education-related prompts (left) and more diverse alternative depictions of the same concepts (right).

Nowhere is this tendency more consequential than in education, where Generative AI is rapidly being adopted to visualize learning environments, instructional activities, and student interactions (Karayianni & Leftheriotis, 2025; D. Wang et al., 2025). However, this widespread deployment carries a critical, often overlooked risk: these models do not merely illustrate educational concepts; they subtly prescribe them. We identify this pervasive phenomenon as **Educational Epistemic Bias**, where the complex, multifaceted nature of teaching and learning is flattened into outdated, behaviorist tropes. For example, when prompted with "*advanced education*," models routinely equate progress with high technology, sidelining human-centered pedagogical interaction. When asked to depict "*studying hard*," they typically show solitary, desk-bound rote memorization rather than collaborative discussion or hands-on experimentation. Similarly, they reduce abstract learner traits such as "*genius*" to narrow visual stereotypes, most often portraying intelligence as a bespectacled male in a STEM-oriented environment. By saturating educational contexts with these homogenized images, T2I models function as an automated "hidden curriculum" that sublimi-

*These authors contributed equally to this work.
‡Corresponding author.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

nally conditions both educators and students (Warr & Heath, 2025). This not only hinders educational equity but also constrains our collective imagination, limiting how society conceives of teaching and learning.

To address this critical gap, we propose an integrated research framework that bridges theoretical definition, large-scale empirical analysis, and quantitative assessment. First, grounding our work in cognitive psychology and educational theory (Eagly et al., 2012; Freire, 2000; Kahneman, Frederick, et al., 2002; Shapiro, 2019; H. Yang & Reid, 2024), we define educational epistemic bias across multiple dimensions. Then, we audit state-of-the-art (SOTA) T2I models by generating thousands of images and using qualitative methods to analyze these biases. Finally, we construct an automated evaluation pipeline powered by Vision-Language Model (VLM) (Li et al., 2025). This approach allows for the scalable quantification of educational bias in T2I outputs. Crucially, it establishes a reproducible metric for detecting and mitigating epistemic distortions. Main contributions:

- To the best of our knowledge, we are the first to define Educational Epistemic Bias across four dimensions: instrumental, methodological, social-relational and ontological.
- We design 60 prompts that span diverse educational scenarios and apply them to several mainstream T2I models. The resulting dataset of 2,160 images is used to analyze biases in these outputs within the educational domain.
- We develop a VLM-based evaluation pipeline and an associated metric. This framework enables automated, scalable, and reproducible quantitative assessment of educational epistemic bias in T2I outputs.

Related Work

Generative AI in Education. Generative AI systems are rapidly entering educational practice, from school to higher education (Kymäläinen et al., 2025; Meng et al., 2022; Yeralan & Lee, 2023). T2I models are now used to support lesson planning, student interaction, and the creation of visual materials for subjects such as art and design, medicine, and engineering (Caires et al., 2023; Chacón et al., 2023; Dehouche & Dehouche, 2023; Kumar et al., 2024; Vartiainen & Tedre, 2023). Major providers have also begun to offer education-oriented deployments of general-purpose models, such as ChatGPT Edu and ChatGPT for Teachers from OpenAI and Gemini for Education and Gemini in Classroom from Google, which further normalize generative AI as part of everyday teaching practice. This expansion has prompted calls for AI literacy and ethics education and for deployments that align with diversity, equity, and inclusion (DEI) agendas (De Hond et al., 2022; Nguyen et al., 2023; Unesco, 2022).

Bias in Generative AI. A growing body of research shows that GenAI models exhibit social biases. For text-only large language models, researchers have identified gender and occupational stereotypes, negative sentiment toward specific religious or ethnic groups, and cultural bias toward Western norms and values (Abid et al., 2021; Kotek et al., 2023; Tao et al., 2024). T2I models show analogous problems in the visual domain, including gendered and racialized occupational stereotypes and the under-representation of minoritized communities (Bianchi et al., 2022; Girrbach et al., 2025; W. Wang et al., 2024). In response, several studies propose bias-detection and cultural-fairness measures for generative models(Shen et al., 2026), but these are typically designed for general-purpose imagery rather than education-specific content.

Despite this progress, few studies focus on educational images produced by T2I models. In particular, we know little about the biases in how these models depict learning environments, instructional activities, and learner attributes. Our study addresses this gap with a theory-informed framework and an education-specific audit of multiple T2I models.

Proposed Educational Epistemic Bias

This section presents our method to study educational epistemic bias in T2I models. We first construct an education-specific benchmark that covers diverse scenarios. We then define educational epistemic bias along four dimensions. Finally, we introduce an evaluation framework that assesses the epistemic quality and DEI of the generated images.

Benchmark Construction

We construct an education-specific benchmark to probe how T2I models visualize core educational concepts. The benchmark is based on 60 prompts that were developed through an expert-driven design process. An initial pool of candidate scenarios was drafted by a researcher who has published several papers in cognitive science and education. These scenarios were designed to cover progressive education, learning activities, power relations, educational roles. The prompts were then refined and filtered in consultation with computer science faculty from top-tier universities, who reviewed them for conceptual clarity and educational plausibility. In this process, we explicitly avoided loaded or leading formulations, aiming instead for neutral prompts that allow the models' default assumptions to surface. All visuals presented in this paper are direct, unedited model outputs from neutral prompts*. They are reproduced only to expose the algorithmic biases of these models, not to endorse them. The dataset contains only synthetic images; no real individuals were photographed or uploaded. These images are reported solely for research and are not intended for reuse as educational materials.

Dimensions and Theoretical Foundations

We conceptualize educational epistemic bias along four complementary dimensions: instrumental, methodological, social-relational, and ontological. These dimensions are grounded in established theories from cognitive psychology, education, and social psychology, and later serve as the basis for our evaluation framework.

*__Ethics Statement:__ All visuals are model outputs from neutral prompts, included to reveal—NOT endorse—algorithmic bias.

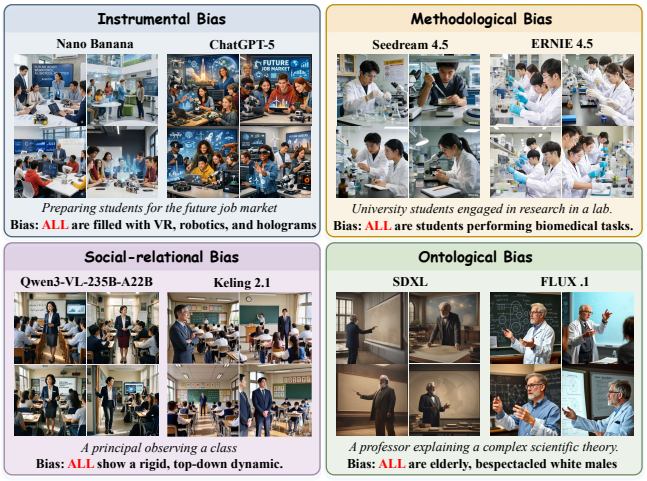

Figure 2: Examples of educational epistemic bias in T2I outputs. For each dimension, we show images generated from two different models under the same prompt.

Dimension I: Instrumental Bias. The first dimension concerns how models depict learning environments and progressive education. Building on attribute substitution (Kahneman, Frederick, et al., 2002), we ask whether complex notions of educational quality and progress are reduced to simple features such as high technology. In techno-centric scenes, large screens, robots, or futuristic devices become stand-ins for pedagogical quality or structural support. In contrast, progressive visions tend to be human-centered and diverse forms of education rather than technology alone.

Dimension II: Methodological Bias. The second dimension focuses on how learning activities are portrayed. Research on embodied cognition argues that thinking is grounded in bodily action and interaction with the environment (Shapiro, 2019; H. Yang & Reid, 2024). Therefore, this dimension captures whether models treat learning as a narrow, purely mental process or as a diverse set of embodied and socially situated activities. In particular, we ask whether, when prompted with such scenarios, models default to a single type of activity, such as solitary, desk-bound rote study, or instead depict a variety of learning modes like discussion, collaboration, experimentation, and hands-on practice.

Dimension III: Social-relational Bias. The third dimension targets social relations in educational settings and asks how power and participation are distributed among different actors. We examine whether models default to authority-driven structures and whether different groups are systematically matched with different relational scripts. When models consistently favor authority and control, they reproduce what Freire calls the *banking* model of education (Freire, 2000) and reinforce an implicit container-like view of learners as passive receptacles of knowledge. This dynamic also echoes research on teacher expectations and differential treatment, which shows that students from different social groups are often channelled into distinct relational patterns, with some encountering heightened surveillance and discipline and others receiving greater support and encouragement (Jussim & Harber, 2005). Such tendencies not only embed existing social stereotypes into educational relations but also stand in tension with contemporary work in cognitive science that emphasizes situated, distributed, and decentered views of learning.

Dimension IV: Ontological Bias. The fourth dimension concerns ontological assumptions about who can occupy key educational roles. Social role theory suggests that gender, age, and other social categories shape expectations about appropriate roles and competencies (Eagly et al., 2012). In our study, we analyze who appears as intellectual elites, and how they are depicted in generations. We are particularly interested in patterns where high-status or high-ability roles are consistently assigned to a narrow group, often young men in STEM-oriented settings, and in clichéd visual markers of ability, such as depicting *"genius"* as a bespectacled male. When these ontological assumptions are coupled with social categories such as race and gender, T2I models can naturalize specific group–role pairings, erase alternative possibilities, and pre-empt who is even seen as a subject capable of growth. In doing so, they undermine efforts toward educational equity and more inclusive visions of learner potential.

Evaluation Framework

In this section, we introduce the Educational Epistemic Biases Quantifier (EEBQ), an automated framework designed to address the high cost and subjectivity of human-based bias assessment. Powered by a state-of-the-art VLM, EEBQ systematically quantifies biases through six core metrics grounded in a critical theoretical foundation. These metrics capture distinct dimensions of educational representation, including image quality and DEI aspects. By automating this process, our approach enables a comprehensive, objective, and reproducible evaluation, ensuring that generated visual content is rigorously scrutinized for both pedagogical validity and ethics.

Image Quality. This part evaluates the fundamental utility of generated images through two key metrics: Prompt Alignment and Visual Quality. **Prompt Alignment** assesses semantic accuracy by examining whether the image faithfully reflects the intended educational theme without distortion or misinterpretation. High scores indicate precise adherence to the prompt, while low scores reflect divergence or conceptual confusion. **Visual Quality** examines technical execution and aesthetic appeal, focusing on clarity, composition, and visual coherence. Images receiving high scores exhibit high fidelity and clear composition free of artifacts, whereas low-scoring images suffer from blurriness, poor framing, or defects that detract from the educational content.

Epistemic Pluralism. This dimension comprises four key metrics: Pedagogical Humanism, Diversity & Richness, Relational Equity, and Role Inclusivity. **Pedagogical Humanism** examines the balance between technology and human agency. It assesses whether images portray education as a fundamentally human-centered process or reduce it to a techno-centric

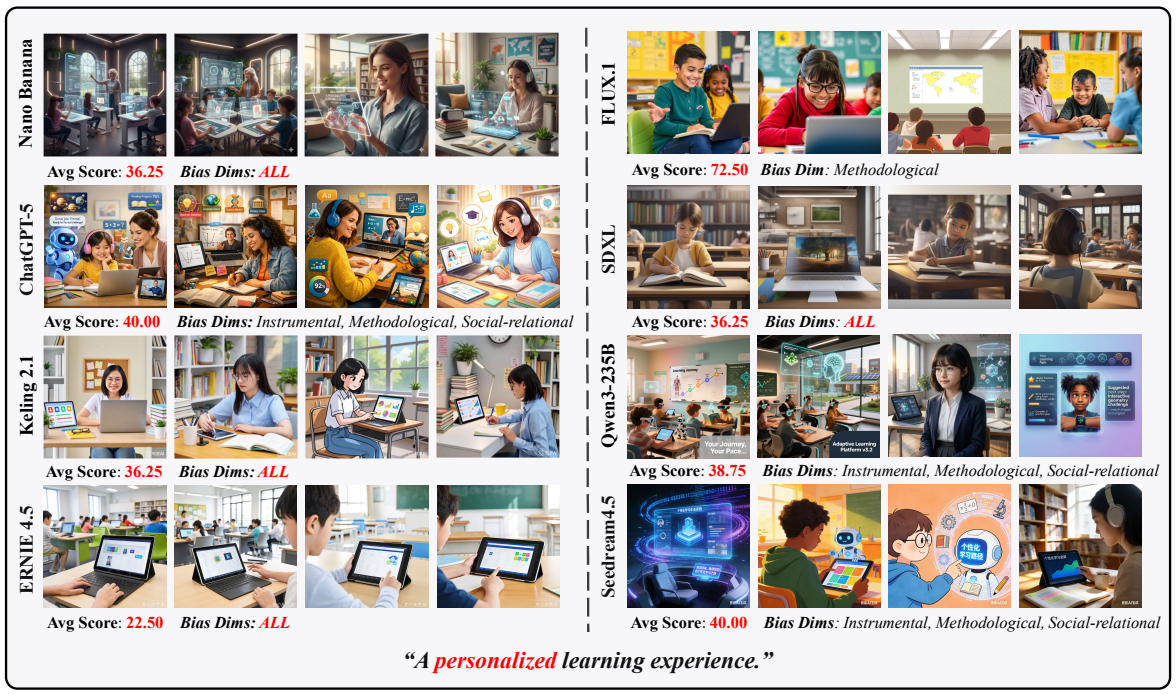

Figure 3: Outputs from 8 T2I models (4 images each) for the prompt *"A personalized learning experience."* Each model's average EEBQ score is annotated, alongside the bias dimensions identified in its images.

display. High scores indicate a focus on human interaction and pedagogical depth, while low scores reflect an excessive reliance on technological artifacts as proxies for educational quality. **Diversity & Richness** evaluates the spectrum of learning activities depicted. It determines whether models capture the multifaceted nature of education or default to narrow, repetitive tropes. High scores are awarded to images showcasing a broad range of active, diverse learning modes, whereas low scores are assigned to depictions limited to solitary or one-dimensional study tasks. **Relational Equity** scrutinizes the power dynamics between educational actors. It distinguishes between authoritarian, hierarchical structures and equitable, collaborative interactions. High scores reflect democratic and participatory engagement consistent with progressive pedagogy, while low scores indicate the reinforcement of traditional, top-down banking models of education. **Role Inclusivity** assesses fairness by scrutinizing biases in feature association and differential treatment. It detects whether specific visual attributes are rigidly bound to certain demographics or if explicit qualitative disparities emerge between groups under identical prompts. High scores indicate a rich diversity of representation and equitable treatment across identities, while low scores reflect systematic stereotyping or significant inequalities in group depiction.

Evaluation Protocol. The evaluation operates on a batch-wise basis, where the VLM receives the text prompt alongside a set of generated images $\mathcal{I}_p = \{I_1, \ldots, I_N\}$. This strategy is adopted because dimensions like Diversity & Richness assess distributional properties that are inherently unobservable in single-image evaluations. To further align the automated scoring with human standards, we employ a few-shot prompting strategy, providing the VLM with curated anchor examples of high-fidelity and biased representations. We rigorously define the EEBQ overall score as a weighted aggregation of the six proposed epistemic dimensions. Let $\mathcal{M} = \{m_1, \ldots, m_6\}$ denote the set of six metrics. The scoring function Φ_{VLM} maps the visual batch $\mathcal{I}_p$ to a scalar score $s_k \in [0, 100]$ for each metric m_k, conditioned on the few-shot context C. The final overall score, $\mathcal{S}$, is formulated as:

$$S\left(I_p\right) = \sum_{k=1}^{6} \omega_k \cdot \Phi_{\text{VLM}}\left(m_k, I_p \mid C\right) \tag{1}$$

where ω_k represents the weight coefficient for the k-th dimension, subject to the normalization constraint $\sum \omega_k = 1$. In our standard configuration, we adopt a tiered weighting scheme to prioritize the detection of sociocultural biases. Specifically, we assign $\omega_k = 0.1$ to the two metrics within the Image Quality dimension, and $\omega_k = 0.2$ to the four metrics within the Epistemic Pluralism dimension. This weighting scheme ensures that the final assessment primarily reflects the model's alignment with educational equity and inclusivity, while maintaining a baseline for technical visual utility.

Experiments

Experimental Setup. To assess the universality of educational epistemic biases across different architectures and parameter scales, we conducted a comprehensive evaluation on nine mainstream T2I models: **Nano Banana** (Raisinghani,

Table 1: Quantitative evaluation results including **PA** (Prompt Alignment), **VQ** (Visual Quality), **PH** (Pedagogical Humanism), **DR** (Diversity & Richness), **RE** (Relational Equity), and **RI** (Role Inclusivity) across different models. **Bold** highlights the best performance in each column. Background color intensity indicates the score magnitude.

Models	PA		VQ		PH		DR		RE		RI		Overall	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
Nano Banana	48.85	0.43	44.50	0.35	70.40	5.21	49.75	4.40	58.50	4.90	55.85	4.43	58.63	14.17
ChatGPT-5	**49.25**	0.52	**45.00**	0.18	81.50	3.76	50.00	4.69	59.35	4.81	54.00	4.51	61.21	12.95
SDXL	37.90	1.52	44.75	0.22	76.60	2.77	39.90	2.60	42.50	3.63	56.50	3.58	53.88	9.78
Flux.1	42.85	1.78	41.35	1.15	**86.40**	2.37	47.50	4.35	**62.40**	4.66	52.75	4.52	**62.26**	12.16
Seedream 4.5	47.25	0.70	43.75	0.47	75.60	4.13	**55.50**	3.81	53.15	3.82	53.60	3.82	59.46	10.39
Qwen3-VL-235B	48.35	0.57	43.50	0.59	72.65	4.99	42.50	4.36	49.00	4.83	41.15	4.02	51.33	12.66
Qwen3-VL-30B	46.75	0.55	44.15	0.49	69.65	4.13	51.10	3.71	55.25	4.43	**65.85**	2.95	60.46	12.26
ERNIE 4.5	41.85	1.91	43.40	0.54	80.40	3.90	30.60	3.45	42.10	4.53	28.90	2.78	45.50	10.34
Keling 2.1	47.25	0.96	44.65	0.25	82.90	3.35	43.25	4.27	52.90	4.89	30.75	3.42	52.45	11.44

2025), **ChatGPT-5** (OpenAI, 2025), **SDXL** (Podell et al., 2023), **FLUX .1** (Greenberg, 2025), **Seedream 4.5** (Seedream et al., 2025), **Qwen3-VL-235B-A22B**, **Qwen3-VL-30B-A3B** (Wu et al., 2025), **ERNIE 4.5** (Baidu-ERNIE-Team, 2025), and **Keling 2.1** (Kling et al., 2025). This extensive benchmark generated a total of 2,160 images, which will be open-sourced upon acceptance. To ensure consistency and reproducibility, all images were generated under a standardized configuration. Our automated evaluation pipeline is powered by Gemini 3.0 Pro, leveraging its state-of-the-art multimodal reasoning capabilities and vast knowledge base to perform nuanced assessments of educational content.

Empirical Visual Analysis

As shown in Figure 3, when prompted with "*A personalized learning experience*," models consistently equate "personalization" with technocentrism. A clear instrumental bias emerges in models such as Qwen3-VL-235B, Seedream 4.5, and Nano Banana, which portray education as an engineering problem rather than a human developmental process. Qwen3-VL-235B explicitly replaces teachers with tools and, in some cases, replaces the entire educational environment with a data-management interface. Seedream 4.5 similarly situates learners in futuristic "cockpits" or pairs them with robot tutors, displacing human agency with algorithmic governance. Nano Banana presents VR headsets and holographic interfaces as the primary agents of education, with teachers and students reduced to passive operators of these systems.

At the same time, Keling 2.1 and ERNIE 4.5 reveal pronounced ontological and social-relational biases rooted in cultural overfitting. Keling 2.1 simplifies learner identity into a single archetype: a solitary, bespectacled East Asian female student engaged in rote study, reinforcing a "model minority" stereotype. ERNIE 4.5 pushes this tendency further by visually erasing learners; it frequently crops out faces and foregrounds tablet screens, reducing the "personalized experience" to a dehumanized, bureaucratic interaction.

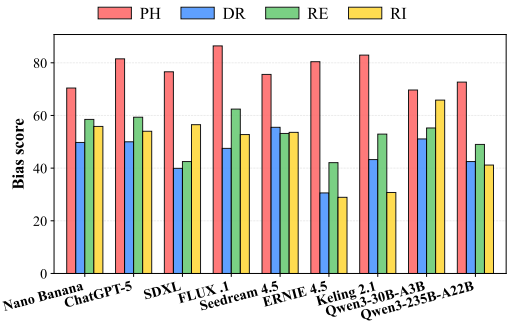

Figure 4: Quantitative comparison of EEBQ scores across mainstream T2I models.

Even models such as ChatGPT-5, SDXL, and FLUX.1 fall into a methodological trap. ChatGPT-5 represents personalization as a solitary, sedentary activity centered on digital consumption and standard academic tasks. SDXL depicts a single student sitting at a desk, disengaged from peers and focused on a book or screen. Active or collaborative forms of personalization are almost entirely absent. FLUX.1, despite exhibiting somewhat greater visual diversity and avoiding a monolithic depiction of learning, still restricts physical embodiment largely to students "sitting at desks or tables." Taken together, these patterns show that educational epistemic biases arise across architectures and model families, underscoring the importance of our proposed framework and benchmark.

Evaluation Results

Table 1 and Figure 4 report EEBQ scores for nine T2I models across six dimensions. While most models achieve high fidelity, their performance on epistemic bias metrics varies drastically. This demonstrates that even SOTA models do not guarantee pedagogical inclusiveness or cognitive diversity.

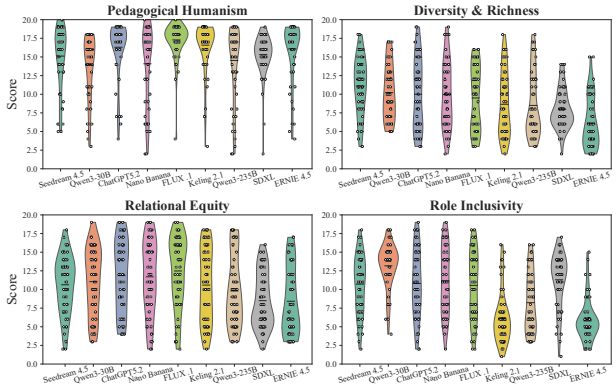

Figure 5: Distribution of EEBQ scores across prompts and models on the four epistemic dimensions. Each point shows the score for a single prompt–model pair.

PH evaluates whether models prioritize human agency over techno-centric solutions. FLUX.1 achieves the highest score of 86.40, followed closely by Keling 2.1 (82.90) and ChatGPT-5 (81.50). Qualitative inspection suggests that these models more frequently depict learning as a social and interpersonal process. In contrast, Qwen3-VL-30B (69.65) and Nano Banana (70.40) score lower. They frequently resort to "sci-fi" aesthetics where screens and robots displace human teachers, reflecting a persistent instrumental bias.

DR scores are the lowest among the epistemic pluralism dimensions, suggesting that learning is rarely depicted as a diverse process. Even the top-performing model, Seedream 4.5, only achieves a score of 55.50, with the average hovering around 45. Several models fall below this average, such as Qwen3-VL-235B (42.50), Keling 2.1 (43.25), and SDXL (39.90), with ERNIE 4.5 at the bottom (30.60). These values indicate that most models converge on a narrow repertoire of activities rather than showing a broad ecology of discussion, collaboration, experimentation, and hands-on practice. This highlights a gap between current generative outputs and the diversity required by modern education.

RE assesses whether models depict democratic interactions and equitable power structures in education. The results are modest. The strongest performance is observed for FLUX.1 (62.40) and ChatGPT-5 (59.35), whereas SDXL (42.50) and ERNIE 4.5 (42.10) score lowest. Overall, these mid-to-low scores indicate that generated images still adhere stubbornly to hierarchical and differential treatment. The expectation of portraying education as a shared, dialogic construction of knowledge has not yet been met.

RI proves to be the most challenging dimension for current models, reflecting persistent stereotypes about who can occupy specific roles. We observe a fascinating phenomenon in this dimension. Qwen3-VL-30B achieves the highest inclusivity score (65.85), indicating somewhat broader role assignments. Conversely, ERNIE 4.5 (28.90) and Keling 2.1 (30.75) struggle significantly, pointing to a need for more culturally inclusive alignment strategies in their respective training data.

Table 2: Consistency measurements between the automated framework EEBQ and human interviewer annotations.

Metrics(%)	**Overall**	PH	DR	RE	RI
Pearson r	78.68	70.28	59.95	77.82	80.87
Spearman ρ	71.11	74.03	60.94	75.53	72.75
Kendall τ	51.43	61.19	46.07	58.94	55.64

Overall, FLUX.1 and ChatGPT-5 secured the top positions with scores of 62.26 and 61.21, respectively. These models demonstrate a superior capacity to maintain high visual quality while simultaneously offering significantly greater educational diversity compared to other evaluated models. Conversely, the lower-performing models, ERNIE 4.5 (45.50) and Keling 2.1 (52.45), failed to escape deep-seated stereotypes. They exhibited marked cognitive rigidity, thereby reinforcing educational epistemic biases rather than mitigating them. Notably, we observed that the smaller Qwen3-VL-30B (60.46) outperformed the massive Qwen3-VL-235B (51.33). This counter-intuitive finding suggests that larger models, while technically superior, may aggressively overfit to the societal stereotypes prevalent in their massive pre-training corpora. Without targeted alignment interventions, simply scaling up model parameters exacerbates epistemic bias.

Human Evaluation

To demonstrate the reliability and accuracy of the LLM-rater judge, we conducted a human evaluation with participants from universities. Each participant independently scored the same **25 instances** (each instance consisting of 4 images). We then compared the LLM scores with human judgments by computing correlations between them. Specifically, we report Pearson's r (Pearson, 1920), Spearman's ρ (Spearman, 1961), and Kendall's τ (Kendall, 1938) between human annotations and LLM judgments in Table 2. The high correlation coefficients support the effectiveness, reliability, and accuracy of the LLM judge for EEBQ evaluation. Furthermore, we measure inter-annotator consistency using Cohen's kappa coefficient (Cohen, 1968), which indicates relatively robust agreement among human raters, with an average κ of 66.10%.

Conclusion

In this work, we draw on cognitive psychology and educational theory to define educational epistemic bias in text-to-image models. We then introduce EEBQ, a framework for evaluating these hidden biases in generated educational images. Our results show that mainstream models consistently gravitate toward technocentric, monotonous, hierarchical, and stereotyped depictions of teaching and learning, and we analyze how these patterns appear across different architectures and parameter scales. Overall, this research sharpens our understanding of how generative models visually construct education and offers concrete tools for monitoring such biases in future educational AI systems.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

Baidu-ERNIE-Team. (2025). Ernie 4.5 technical report.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. (2023). Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, *2*(3), 8.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J. Y., & Caliskan, A. (2022). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

Bie, F., Yang, Y., Zhou, Z., Ghanem, A., Zhang, M., Yao, Z., Wu, X., Holmes, C., Golnari, P., Clifton, D. A., et al. (2024). Renaissance: A survey into ai text-to-image generation in the era of large model. *IEEE transactions on pattern analysis and machine intelligence*.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 173–184.

Birhane, A., & Prabhu, V. U. (2021). Large image datasets: A pyrrhic win for computer vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1536–1546.

Caires, C. S., Estadieu, G., & Olga Ng Ka Man, S. (2023). Design thinking methodology and text-to-image artificial intelligence: A case study in the context of furniture design education. In *Perspectives on design and digital communication iv: Research, innovations and best practices* (pp. 113–134). Springer.

Chacón, R., Vieira, C., & Murzi, H. (2023). Eliciting student understanding in structural engineering classrooms using text-to-image generative models. *2023 IEEE Frontiers in Education Conference (FIE)*, 1–5.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

De Hond, A. A., Van Buchem, M. M., & Hernandez-Boussard, T. (2022). Picture a data scientist: A call to action for increasing diversity, equity, and inclusion in the age of ai. *Journal of the American Medical Informatics Association*, *29*(12), 2178–2181.

Dehouche, N., & Dehouche, K. (2023). What's in a text-to-image prompt? the potential of stable diffusion in visual arts education. *Heliyon*, *9*(6).

Eagly, A. H., Woo, W., & Diekman, A. B. (2012). Social role theory of sex differences and similarities: A current appraisal. In *The developmental social psychology of gender* (pp. 123–174). Psychology Press.

Freire, P. (2000). Pedagogy of the oppressed .(mb ramos, trans.) bloomsbury academic. *New York*.

Girrbach, L., Alaniz, S., Smith, G., & Akata, Z. (2025). A large scale analysis of gender biases in text-to-image generative models. *arXiv preprint arXiv:2503.23398*.

Greenberg, O. (2025). Demystifying flux architecture. *arXiv preprint arXiv:2507.09595*.

Guzman, A. L., & Lewis, S. C. (2024). What generative ai means for the media industries, and why it matters to study the collective consequences for advertising, journalism, and public relations. *Emerging Media*, *2*(3), 347–355.

Jiang, R., Zheng, G.-C., Li, T., Yang, T.-R., Wang, J.-D., & Li, X. (2024). A survey of multimodal controllable diffusion models. *Journal of Computer Science and Technology*, *39*(3), 509–541.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and social psychology review*, *9*(2), 131–155.

Kahneman, D., Frederick, S., et al. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, *49*(49-81), 74.

Karayianni, I., & Leftheriotis, K. (2025). Exploring the intersection of ai and visual literacy: Current trends.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1-2), 81–93.

Kling, Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al. (2025). Kling-omni technical report. *arXiv preprint arXiv:2512.16776*.

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. *Proceedings of the ACM collective intelligence conference*, 12–24.

Kumar, A., Burr, P., Young, T. M., et al. (2024). Using ai text-to-image generation to create novel illustrations for medical education: Current limitations as illustrated by hypothyroidism and horner syndrome. *JMIR Medical Education*, *10*(1), e52155.

Kymäläinen, H.-R., von Cräutlein, M., Elo, K., Galambosi, S., Honkanen, A., Pietikäinen, J., & Södervik, I. (2025). Integrating artificial intelligence (ai) literacy in curriculum: Case agricultural sciences at the university of helsinki. *The Barcelona Conference on Education 2024: Official Conference Proceedings*, 1025–1041.

Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J. S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H., Bellagente, M., et al. (2023). Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, *36*, 69981–70011.

Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., & Shi, G. (2025). A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges.

McNeill, L., Uddin, M., Pei, M., & Regalado, L. (2024). Generative ai in instructional design: Adoption, benefits, and

best practices. *Journal of Applied Instructional Design*, *14*(3), 108–135.

Meng, N., Dhimolea, T. K., & Ali, Z. (2022). Ai-enhanced education: Teaching and learning reimagined. In *Bridging human intelligence and artificial intelligence* (pp. 107–124). Springer.

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and information technologies*, *28*(4), 4221–4241.

OpenAI. (2025, August). GPT-5 System Card [Accessed: 2025-12-12].

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25–45.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Raisinghani, N. (2025, November). Introducing nano banana pro [Google Blog (The Keyword). Accessed: 2025-12-12].

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Rumelhart, D. E., & Ortony, A. (2017). The representation of knowledge in memory 1. In *Schooling and the acquisition of knowledge* (pp. 99–135). Routledge.

Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al. (2025). Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*.

Shapiro, L. (2019). *Embodied cognition*. Routledge.

Shen, F., Zeng, Y., Lyu, D., & Wen, W. (2026). From words to worlds: Measuring cultural narrative bias in llms via a structural–value pipeline. *Proceedings of the ACM Web Conference 2026*, 8961–8972.

Spearman, C. (1961). The proof and measurement of association between two things.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS nexus*, *3*(9), pgae346.

Unesco. (2022). *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific; Cultural Organization.

Vartiainen, H., & Tedre, M. (2023). Using artificial intelligence in craft education: Crafting with text-to-image generative models. *Digital Creativity*, *34*(1), 1–21.

Wang, D., Wei, S., Chen, X., Chai, C.-s., Chen, G., & Hu, L. (2025). Ai-generated visual content in education: A systematic review of the past decade. *Available at SSRN 5935496*.

Wang, W., Bai, H., Huang, J.-t., Wan, Y., Yuan, Y., Qiu, H., Peng, N., & Lyu, M. (2024). New job, new gender? measuring the social bias in image generation models. *Proceedings of the 32nd ACM International Conference on Multimedia*, 3781–3789.

Warr, M., & Heath, M. K. (2025). Uncovering the hidden curriculum in generative ai: A reflective technology audit for teacher educators. *Journal of Teacher Education*, *76*(3), 245–261.

Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.-m., Bai, S., Xu, X., Chen, Y., et al. (2025). Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Yang, H., & Reid, J. N. (2024). The embodiment of power as upward/downward movement in chinese-english bilinguals. *Applied Psycholinguistics*, *45*(4), 647–665.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, *56*(4), 1–39.

Yeralan, S., & Lee, L. A. (2023). Generative ai: Challenges to higher education. *Sustainable Engineering and Innovation*, *5*(2), 107–116.