Empowering Youth with AI: An Evaluation of AI-Assisted Feedback in Juvenile Detention

Adam Fine, Ph.D., & Justin Richardson, M.S.

Youth Justice Lab
Arizona State University

August 13, 2025

Note: This white paper is based on a research article currently under peer review. The final, published version may differ.

Table of Contents

Introduction

Juvenile detention facilities are designed to provide stability, safety, and opportunities for rehabilitation for youth awaiting court hearings. However, research consistently shows that these environments often fail to produce positive long-term outcomes for youth, and may even contribute to worsened trajectories (Cauffman et al., 2021; Walker & Herting, 2020). One promising approach to improving youth experiences is to incorporate positive youth development principles—focusing on strengths, growth, and supportive relationships—into daily programming.

Journey.do, a social growth learning platform, provides youth with opportunities to learn and practice skills across life domains, share personal narratives, and receive personalized feedback from staff. Within Journey.do, providing high-quality, timely, and developmentally informed feedback is critical for maximizing the impact of this model. Yet, it is also a cognitively and emotionally demanding task for staff—especially in detention settings, where time and attention are scarce.

Advances in artificial intelligence (AI) present a potential opportunity: trained AI systems can generate personalized, strength-based first drafts of feedback that staff can then refine. The promise is not to replace staff, but to equip them with a tool that preserves the human relationship while reducing cognitive load and maintaining high feedback quality.

This evaluation examines whether AI-assisted feedback can match or even exceed the quality of human-generated feedback across key domains of review quality, and explores the implications for enhancing positive youth development in detention facilities.

As described in this white paper, we compared the quality of feedback between staff and AI. First, we trained a custom GPT model to give feedback using the same guidelines that staff use to review stories. Next, we selected 150 of the most recent youth submissions and their corresponding staff reviews from the three most popular modules on the Journey platform: What Got Me Here, Being Arrested as a Growth Opportunity, and Being Responsible for My Actions. Then, we generated an AI review for each of those 150 stories. Finally, human scorers evaluated these paired reviews on five major domains: 1) Empathetic and Active Listening; 2) Support and Affirmation; 3) Clarity and Tone; 4) Encouraging Growth and Progress; and 5) Developmentally informed Guidance. We found that the AI reviews consistently outperformed human reviews across each of the five major domains and also across the three modules. These results provide evidence that AI, when using a fine-tuned language model, can provide youth with high-quality feedback in a matter of seconds. Therefore, the present study supports the notion that AI could be a valuable tool in supporting youths' prosocial development, particularly if used to generate a first draft. In the following sections, we'll discuss the implications for positive youth development and the responsibilities of the juvenile justice system and its staff.

Methodology

Participants and Procedure

AI Model

Journey.do created a first generation of their AI model—a customized, private instance of ChatGPT-4o—that they trained with the same review guidelines used for human staff, including strength-based feedback, trauma-informed care principles, personalized acknowledgment of the youth's narrative content, and suggestions for reflection and next steps. To mimic real-world conditions, minor typographical errors were allowed, and responses were designed to sound natural and conversational at a developmentally appropriate 6th-grade reading level.

Story Database

In order to test the AI model, we created a database of 150 of the most recently accepted stories from youth participants in a large juvenile detention facility in the southwestern United States. These were actual youth stories that had been submitted and accepted on the platform between January and May 2024. To ensure we had variety in topic areas, we selected 50 stories from the three most popular modules:

1. What Got Me Here
2. Being Arrested as a Growth Opportunity
3. Being Responsible for My Actions

The 150 youth stories were then fed into Journey.do's trained AI. Thus, it generated 150 AI reviews of the same stories, which yielded 150 total pairs of human and AI-generated reviews of the same youth stories.

Rating Process

Five independent human scorers rated each review on 20 items across five domains of review quality (see Measures). They were not told which review in the pair was done by a human or the AI. We randomized the presentation order within each story pair to avoid bias.

Our methodological approach strengthens the validity and reliability of our findings. Using multiple scorers allowed us to capture a range of perspectives and reduce the influence of any single rater's subjective preferences or biases. Moreover, blinding scorers to whether a review was written by a human or AI helped prevent preconceived notions about AI or staff performance from influencing ratings. Further, randomizing the order of AI and human reviews within each pair further minimized potential carryover or contrast effects—for example, the risk that seeing one review first might raise or lower expectations for the second. Taken together, these procedures strengthened the rigor of the evaluation by ensuring that observed differences in review quality were attributable to the content itself rather than rater bias, order effects, or other confounding factors.

Measures

The quality of the feedback was assessed across five distinct domains developed for the study. Raters scored items on a 5-point Likert-type scale from 1 (Strongly Disagree) to 5 (Strongly Agree). The five domains were:

1. **Empathetic and Active Listening.** This domain assessed how well the review demonstrated understanding, validation, and attentiveness to what the youth expressed. Scorers rated four items (e.g., The review understood what the youth was trying to express; The review made the youth feel heard) that were mean-scored, with higher scores reflecting greater empathetic and active listening (α = .94).
2. **Support and Affirmation.** The domain assessed how well the review appreciated and affirmed the youth's ideas. Scorers rated four items (e.g., The review made the youth feel valued; The review made the youth feel understood) that were mean scored, such that higher scores indicated greater support, affirmation, and validation (α= .96).
3. **Clarity and Tone**. This domain assessed how understandable the review was and the extent to which it used a friendly tone. Scorers rated four items (e.g., The feedback in the review was clear; The review had an empathetic tone) that were mean-scored, with higher scores indicating greater clarity and friendliness (α = .85).
4. **Encouraging Growth and Progress.** This domain assessed how well the review encouraged youth to reflect, grow, and make progress. Scorers rated four items (e.g., The review pushed the youth to reflect and grow further; The review talked about what next steps the youth can take) that were mean-scored, with higher scores indicating more effective feedback for supporting reflection and growth (α = .87).
5. **Developmentally Informed Guidance.** This domain assessed how well the review incorporated trauma-informed care, positive youth development, and strength-based feedback. Scorers rated four items (e.g., The review was trauma-informed; The review was empowering) that were mean-scored, with higher scores indicating greater use of developmentally informed approaches (α = .91).

Analytic Approach

Paired-sample t-tests compared AI-generated and human-generated reviews across each domain. This method isolated quality differences while controlling for the specific story content, since each youth story had both a human and an AI review. Effect sizes were calculated using Cohen's d.

Results

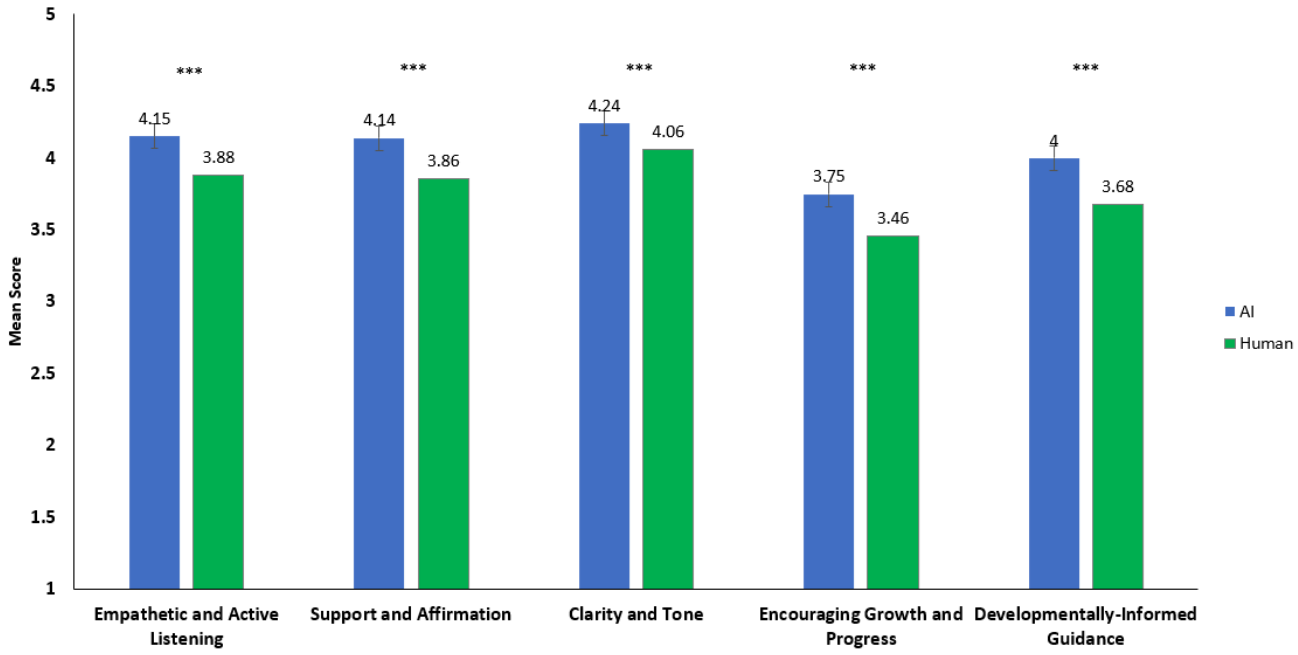
Review Quality: Do AI-Assisted Reviews Match or Exceed Human Feedback Quality?

As shown in Figure 1, the AI-generated reviews outperformed human-generated reviews across all five domains. The effect sizes indicated a consistent advantage for ChatGPT-4o over human reviewers. The observed Cohen's d values were .36 for empathetic and active listening, .36 for

support and affirmation, .34 for clarity and tone, .36 for encouraging growth and progress, and .43 for developmentally informed guidance. These effect sizes represent small-to-moderate differences, with the average Cohen's d across domains being .37 (range: .34 to .43), underscoring the model's consistent performance advantage while still leaving room for meaningful human contributions.

It is important to note that on a 1–5 scale, human reviewers' average scores across domains were in the mid-to-high 3s and low 4s, which reflects generally strong, high-quality feedback. This suggests that staff were already providing thoughtful, supportive, and understandable reviews. However, the AI reviews were consistently higher by about 0.18–0.32 points across domains, a difference large enough to be both statistically significant and practically meaningful given the narrow scale. In standardized terms, this translated into small-to-moderate effect sizes (d = .34–.43), with the largest gap in developmentally informed guidance. In other words, AI didn't just outperform in a technical sense — it improved already strong scores by a margin that was noticeable to the blinded reviewers and would likely be noticeable to a youth receiving the feedback.

Figure 1. Certificate Completion: t-test Results



Discussion

The Journey.do platform is designed to enable staff to support youth, but it can add to an already heavy staff workload (Sheppard et al., 2022). Providing individualized, developmentally informed feedback requires substantial time and emotional energy. Considering AI is becoming more common and more powerful (Chubb et al., 2022), this study examined whether it could serve as a supplementary tool for generating a first draft of high-quality, trauma-informed feedback for youth stories.

We trained a custom GPT model using the same review guidelines as staff, generated AI reviews for 150 recent youth submissions across three popular modules (What Got Me Here, Being Arrested as a Growth Opportunity, Being Responsible for My Actions), and compared them to staff-written reviews. Five blinded raters scored each review on five quality domains. AI reviews consistently outperformed human reviews across all domains, suggesting that a fine-tuned model can produce high-quality first-draft feedback within seconds.

Ultimately, AI-generated first-draft reviews consistently and significantly outperformed final-draft human-written reviews across all five metrics. This provides strong evidence that a fine-tuned AI can be a valuable tool, helping staff provide high-quality, developmentally-informed feedback in a fraction of the time.

Supporting Staff

Detention staff juggle safety, operations, education, and health responsibilities. Adding high-quality narrative feedback to their workload risks increasing burnout. AI tools can ease this burden by generating initial feedback drafts that are strength-based and trauma-informed, allowing staff to personalize and deepen the message. This approach saves cognitive bandwidth, reduces emotional fatigue, and preserves time for relationship-building.

Integrating AI into feedback workflows could help shift facility culture away from control and punishment toward positive change. Quicker, consistent feedback also reduces the risk that youth feel their progress is unnoticed, sustaining motivation. However, note that this evaluation did not test real-world workflow impacts, burnout reduction, or whether AI actually reduces review time, though the results are highly promising.

Important Considerations

This study does not advocate for AI replacing human reviews. The researchers strongly advocate for a model where AI acts as an assistant to, rather than a replacement for, human reviewers. While AI can generate high-quality, developmentally-informed feedback, it can also exhibit repetitive phrasing or a recognizable "patternicity" that can feel impersonal to youth. The proposed solution is for AI to be used as a tool that creates a strong first draft, which a staff member must then review closely and further individualize. This "human touch" is essential. Staff must build and leverage unique relationships with the youth in their care to customize the

feedback, ensuring the message feels genuine and appropriately tailored to that young person's specific situation and personality. This collaborative approach ensures the final feedback benefits from the consistency of AI while retaining the irreplaceable value of authentic human connection.

Thus, we propose a workflow where AI generates the initial, high-quality draft, which staff then review, personalize, and approve. This model could reduce the cognitive and emotional burden associated with writing feedback from scratch, freeing staff to focus on leveraging their unique relationships to make the messages more meaningful. This shift has the potential to change the facility's entire culture—moving from an environment focused on punishment and control to a positive change center focused on rehabilitation and support. By streamlining the delivery of strength-based feedback, AI helps ensure youths' efforts are noticed and validated in a timely manner, which is critical for sustaining their motivation.

Finally, note that we focused on independent scorers' perceptions rather than youth thoughts on reviews. Youth perceptions might vary from scorers and may disapprove of AI-assisted reviews as impersonal. Youth may value genuine human interaction, especially for sensitive issues. Despite safeguards, adolescents may worry about privacy when AI is involved, concerned about data collection, storage, and use. Youth feeling disconnected or distrustful of the feedback process would clearly be problematic, thus this warrants empirical inquiry.

Conclusion & Next Steps

These preliminary results suggest that when used as a first-draft tool, AI can help staff deliver high-quality, timely feedback without replacing the human relationship central to positive youth development. Properly integrated, AI-assisted reviews could expand staff capacity for meaningful engagement, align detention practices with rehabilitative goals, and ensure every interaction is an opportunity for positive growth.

Appendix

Feedback Quality Domains

The following items were used by raters to score the quality of each review on a 5-point scale (1-Strongly Disagree to 5-Strongly Agree).

1. Empathetic and Active Listening
    a. The review understood what the youth was trying to express.
    b. The review validated the youth's feelings and experiences.
    c. The review paid close attention to what the youth shared.
    d. The review made the youth feel heard.
2. Support and Affirmation
    a. The review made the youth feel valued.
    b. The review made the youth feel understood.
    c. The review made the youth feel appreciated.
    d. The review made the youth feel supported.
3. Clarity and Tone
    a. The feedback in the review was clear.
    b. The language was easy to understand.
    c. The review had an empathetic tone.
    d. The review had a conversational tone.
4. Encouraging Growth and Progress
    a. The review pushed the youth to reflect and grow further.
    b. The review highlighted what the youth did well.
    c. The review helps the youth make changes.
    d. The review talked about what next steps the youth can take.
5. Developmentally-Informed Guidance
    a. The review was trauma-informed (i.e., recognizes that past pain affects how you feel and act).
    b. The review aligned with positive youth development (i.e., grow and gain skills).
    c. The review was strength-based (i.e., focuses on promoting strengths and positive traits).
    d. The review was empowering.