From Human to Machine Coding: Evaluating ChatGPT 4o's Ability to Match Human Expertise in Qualitative Analysis

Mathew Franjul, Genevieve Martinez, Chose Tran
Department of Psychology, College of Science, The University of Texas at Arlington

Introduction

In training artificial intelligence (AI) and developing advanced machine learning algorithms, it is imperative to integrate human intelligence (Wu et al., 2022). Interest in reinforcement learning is rising due to the challenge of designing intelligent systems that operate and adhere to real-world environments, (Barto & Sutton, 1997). This study examines whether AI can code qualitative responses as effectively as human coders by utilizing two reinforcement learning methods. Given that coding in research is a time-intensive process demanding significant human effort, AI's potential to match human expertise in rating responses presents an opportunity to enhance efficiency.

To investigate this, we utilized ChatGPT-4o and analyzed its rating accuracy by comparing it to human-coded responses:

- **Reinforcement Learning Method #1**: AI was provided with a rubric and received human judgment to assess its ability to identify and categorize ratings appropriately.
- **Reinforcement Learning Method #2**: AI was provided with the same rubric and grading materials, along with additional examples of human knowledge ratings, prior to receiving human judgements.

By implementing these techniques, we examined whether AI could improve its rating accuracy. These findings may have significant implications for research efficiency:

- AI can serve as a tool in coding and reducing the manual workload for researchers
- AI can allow scholars to allocate more time toward complex analytical tasks

Methods

Data Source
Data for this study was sourced from a pre-existing dataset where human coders had previously analyzed qualitative responses. These human-coded responses served as the baseline for comparison.

AI Training & Coding Process
A machine learning-based AI model was trained to code qualitative responses using a structured rubric. The AI was provided with:

1. Rubric Guidance: A predefined set of coding criteria to ensure consistency.
2. Real-Time Feedback: AI outputs were iteratively refined based on discrepancies identified against human-coded examples.
3. Pre-Training on Human Judgments (only for Hypothesis II): A subset of human-coded responses was used to familiarize the AI with expected patterns before it received real-time feedback.

Study Design & Analysis
A within-subjects experimental design was used to examine the effectiveness of AI rating similarly to humans when receiving direct feedback on its coding judgements. In the study two independent variables were manipulated: (1) the presence of ongoing direct feedback to the AI and (2) the provision of a rubric along with human judgement. The dependent variables were the AI's coding assessments of quality, originality, and elegance.

Procedure
The repeated-measures ANOVAs were conducted to evaluate the effects of the experimental manipulations on coding quality, originality, and elegance. A separate ANOVA was conducted for each dependent variable under both conditions. Effect sizes were reported using partial eta squared (η^2). Statistical significance was set at $p < .05$.

Hypothesis Testing
We tested whether:

1. AI's coding accuracy would align with human coders when guided by rubrics and feedback.
2. Exposure to human-coded examples prior to feedback would improve AI's performance.

Research Goals

- **AI can match human coding accuracy** when guided by structured rubrics and real-time feedback, demonstrating its potential as a reliable qualitative coding tool.
- **Exposure to human-coded examples prior to feedback** significantly enhances AI's rating accuracy, suggesting that pre-training on human judgments boosts performance.
- **AI learns faster and delivers more efficient results** when provided with both human-coded examples and feedback, highlighting its value in reducing researcher workload and improving coding efficiency.

Figures

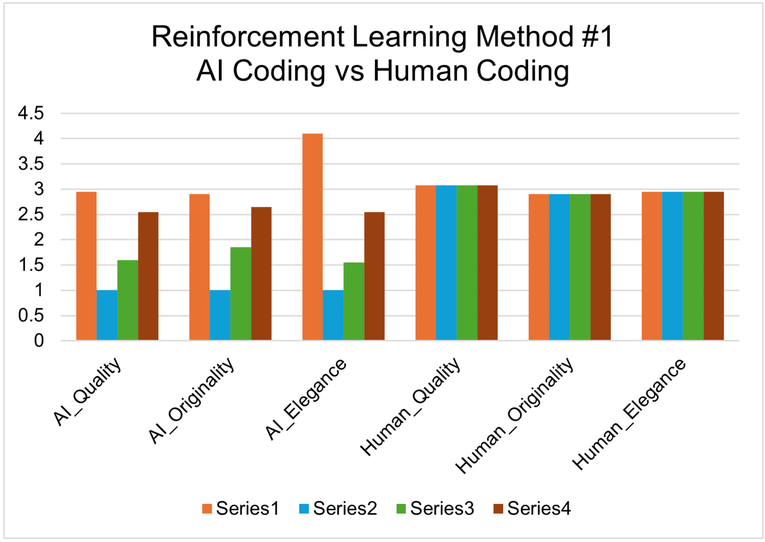

Figure 1. A within-subject ANOVA was conducted to show the progression of AI coding of each variable, quality, originality, and elegance, over four different times with receiving judgement between each time.

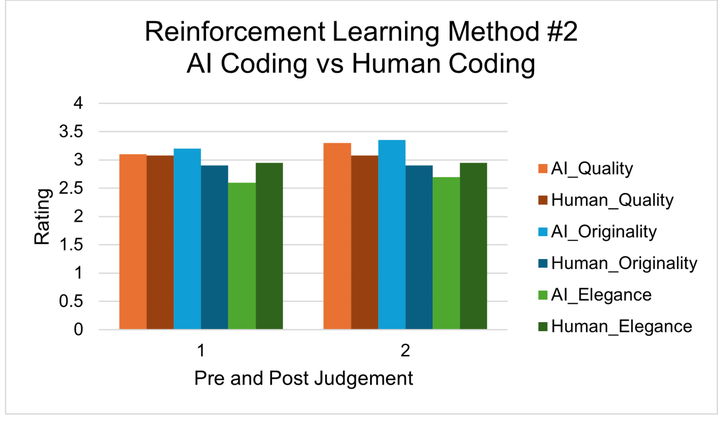

Figure 2. A within-subject ANOVA was conducted to show the change AI coding of each variable, quality, originality, and elegance, of before and after receiving feedback from human raters.

Results

Method 1 - Reinforcement Learning
Within-subjects ANOVA was conducted for each var comparing human and AI ratings across four trials.

- Quality
  - No significant difference in Trial 1
  - Significant difference in Trials 2, 3, and 4
- Originality
  - No significant difference in Trials 1 and 4
  - Significant difference in Trials 2 and 3
- Elegance
  - Significant difference in all four trials

Method 2 - Reinforcement Learning
Mauchly's Test of Sphericity was run for each variable; corrections were applied as needed. Repeated measures ANOVAs examined the effect of the rater.

- Quality
  - Sphericity not violated ($W = .983$, $\chi^2(2) = .316$, $p = .854$)
  - Sphericity-assumed ANOVA used
- Originality
  - Sphericity violated ($W = .592$, $\chi^2(2) = 9.44$, $p = .009$)
  - Greenhouse-Geisser correction applied ($\varepsilon = .710$)
  - Effect of rater approached significance, $F(1.42, 26.99) = 3.27$, $p = .068$, partial $\eta^2 = .147$
- Elegance
  - Sphericity violated ($W = .638$, $p = .018$)
  - Greenhouse-Geisser correction applied ($\varepsilon = .734$)
  - Effect of rater not significant, $F(1.47, 27.91) = 1.72$, $p = .201$, partial $\eta^2 = .083$

Discussion

Our research demonstrated that AI, particularly when exposed to human-coded examples, can effectively match human coding accuracy. Direct feedback alone showed some promise, but it required significantly more time to reach similar performance levels.

In **Reinforcement Learning Method #1,** limitations included:

- Slower Initial Accuracy: AI had to learn coding patterns from scratch, leading to more errors early on.
- Increased Training Time: More cycles of reinforcement were needed to reach acceptable accuracy levels.
- Constricted Coding: AI becomes overly rigid due to excessive feedback, limiting flexibility for improvement

In **Reinforcement Learning Method #2,** strengths included:

- Improved Initial Accuracy**:** AI performed better from the start, reducing the need for corrections
- Efficient Feedback Process**:** Pre-training allowed real-time feedback to be more effective.
- Better Alignment with Human Coders**:** AI internalized qualitative nuances, improving agreement with human-coded responses

This study highlights AI's potential as a qualitative coding tool when structured rubrics and feedback are applied. However, concerns about potential biases and AI's ability to capture deeper qualitative nuances remain. Future research should explore AI's adaptability across different datasets and coding frameworks.

References

1. Barto, A. G., & Sutton, R. S. (1997). *Reinforcement learning in artificial intelligence*. In J. W. Donahoe & V. P. Dorsel (Eds.), *Advances in psychology*, (Vol. 121, pp. 358–386). North-Holland. https://doi.org/10.1016/S0166-4115(97)80105-7
2. Wu, J., Huang, Z., Hu, Z., & Lv, C. (2023). Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving. *Engineering*, *21*(75-91). https://doi.org/10.1016/j.eng.2022.05.017