

An LLM Judge Found a System Difference That Two Human Experts Did Not: Rubric-Based Evaluation of AI-Generated Montessori STEAM Mathematics Activities for Early Childhood

Zafer Savaş Kıvılcım

Altıeylül Science and Art Centre (BİLSEM)

Balıkesir, Türkiye

ORCID ID: 0000-0003-2429-7809

Şeyma Kıvılcım

Hatice Fahriye Eğinlioğlu Primary School

Balıkesir, Türkiye

ORCID ID: 0009-0006-5598-9444

Abstract

Large language models are increasingly used to score educational materials against rubrics, but are rarely validated against human experts. This study compared an LLM judge with two human raters on the same evaluation task. A structured prompt encoding six Montessori principles was given to four systems (ChatGPT, Gemini, Claude and Codex) in four independent sessions each, producing 48 patterning activities for children aged 3 to 6. A STEM education specialist and an in-service preschool teacher scored all 48 independently on a purpose-built 16-item rubric (MSDA), blind to the source system. An LLM scored the same activities under the same rubric. Agreement between the human raters was acceptable (ordinal Krippendorff's $\alpha = .47$; 70.6% exact, 94.9% within one scale point) and internal consistency high (Cronbach's $\alpha = .83$ and $.94$). The human raters found no difference between the systems ($H(3) = 3.50$, $p = .32$, $\epsilon^2 = .01$); no pairwise comparison approached significance. On the same data the LLM judge produced a large and statistically significant difference ($H(3) = 25.21$, $p < .001$, $\epsilon^2 = .51$). It scored every system lower than the humans, with the widest gaps on the STEAM criteria (B4 $\Delta = -1.61$), and agreed only modestly with the human mean ($\rho = .37$). On the human ratings the activities were broadly Montessori-consistent ($M = 4.59$); the weakest criteria were isolation of difficulty and prepared environment, while control of error, the expected difficulty, was handled well. Automated rubric scoring can therefore produce a ranking that human experts do not support, and needs validation against human raters before comparative conclusions are drawn.

Keywords: LLM-as-judge, automated scoring, rubric validity, Montessori education, STEAM, early childhood mathematics, generative artificial intelligence.