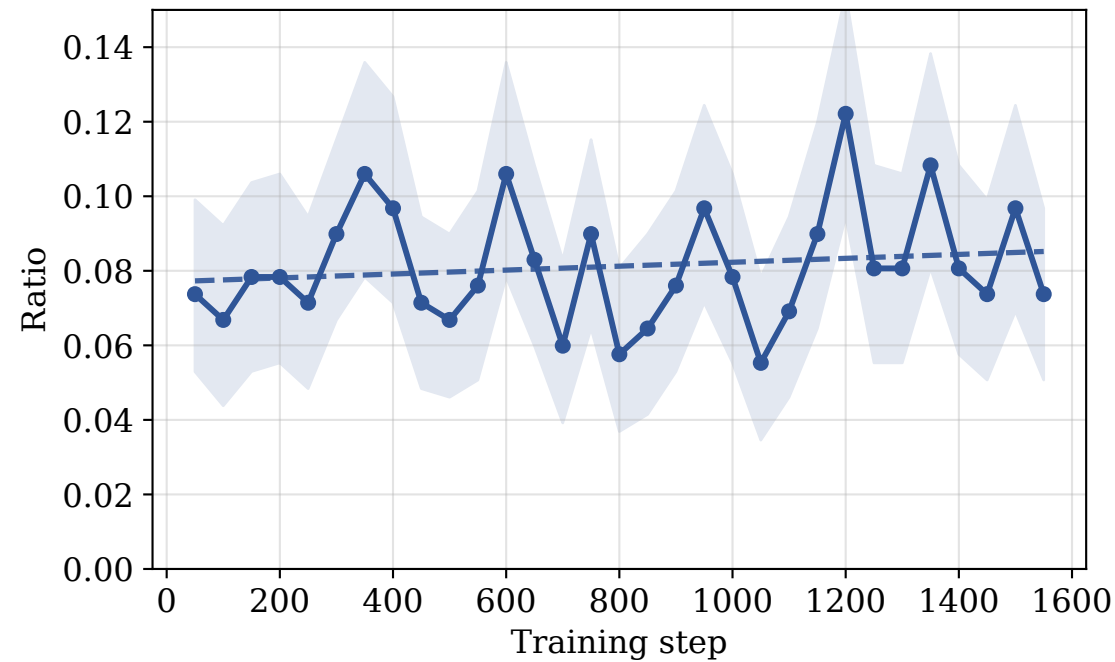
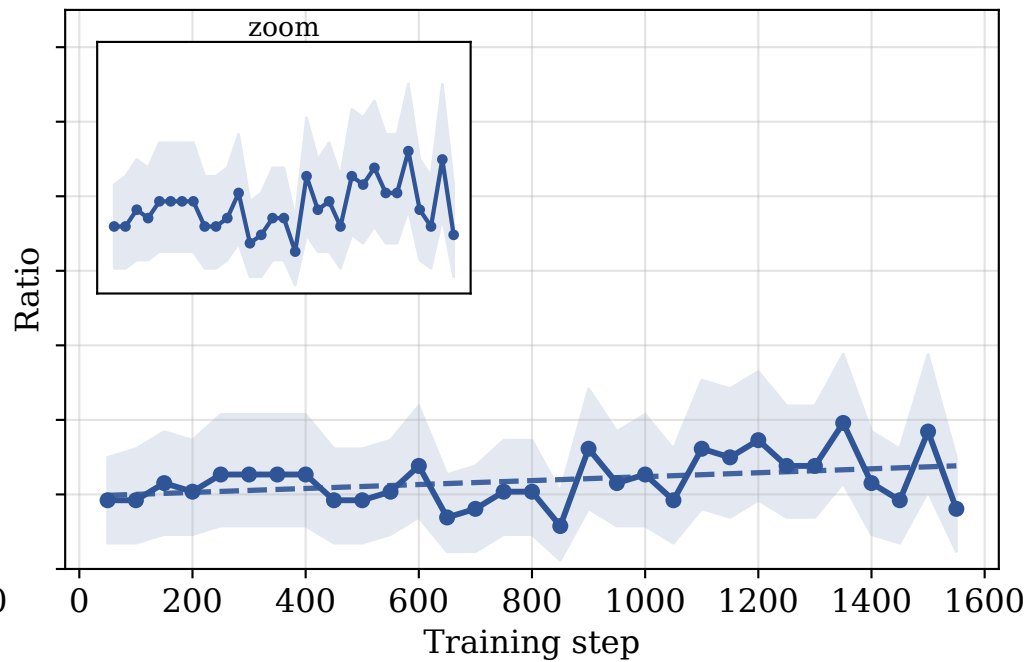


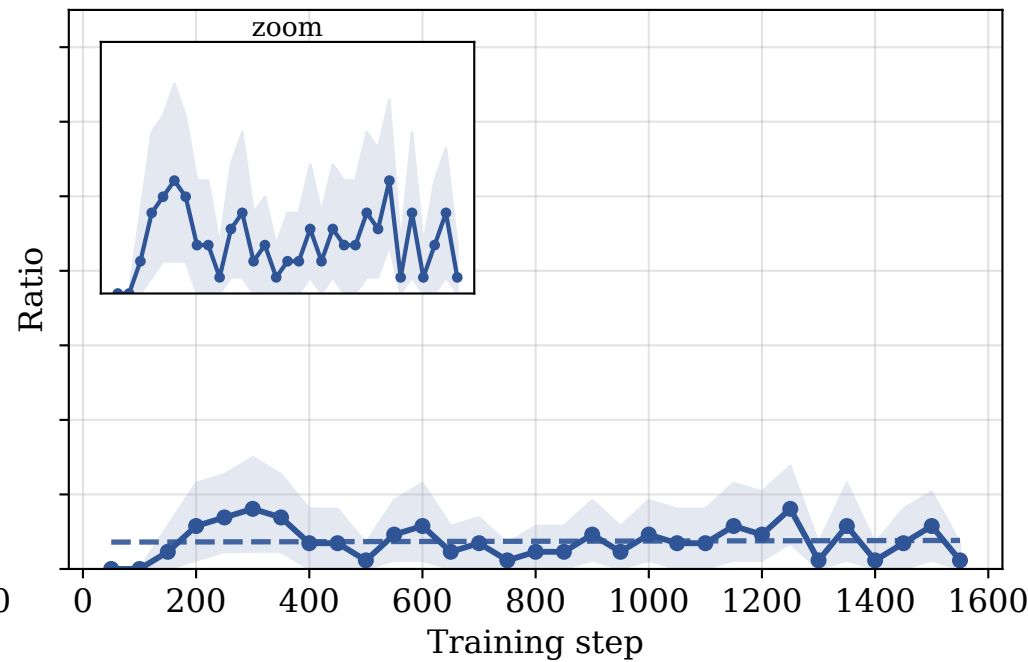
Words of "Aha!" String Detection  
(problem ratio per step)  
MATH-500, Qwen2.5-1.5B



LLM-Detected "Aha!"  
(problem ratio per step)  
MATH-500, Qwen2.5-1.5B



Formal "Aha!"  
(problem ratio per step)  
MATH-500, Qwen2.5-1.5B



Ratio 95% CI (bootstrap) Trend (WLS)