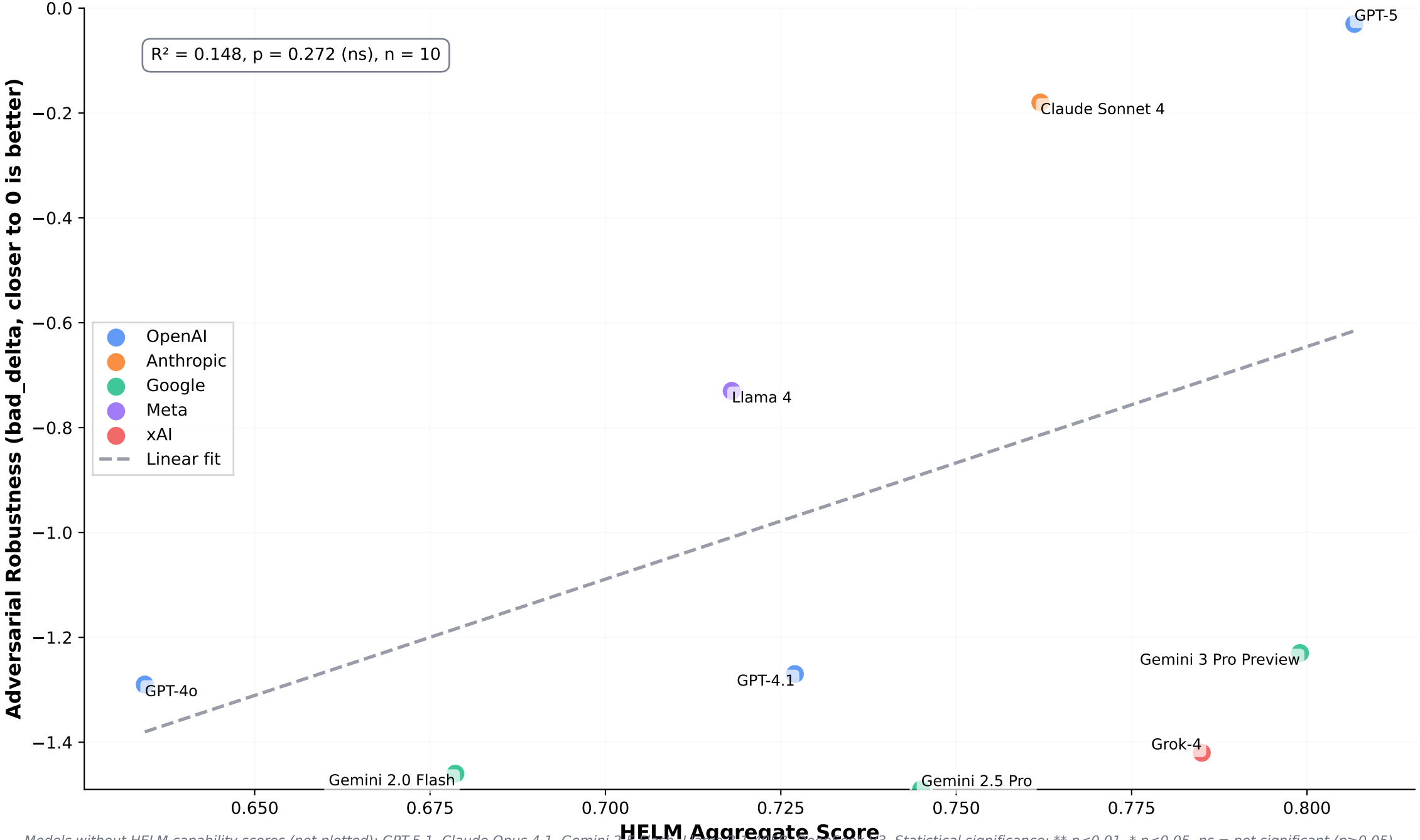
Model Capability vs Adversarial Robustness

Higher capability models show better resistance to adversarial prompts (less negative degradation)



*Models without HELM capability scores (not plotted): GPT-5.1, Claude Opus 4.1, Gemini 2.5 Flash, Llama 3.1 405B, DeepSeek V3. Statistical significance: ** p<0.01, * p<0.05, ns = not significant (p≥0.05)*