Sevda Hasanli
University of Ottawa
Canada

Biography

Sevda Hasanli, completed her master's in Clinical Psychology and is currently pursuing her PhD in the Neuropsychology Lab at the University of Ottawa. Her research spans from cognitive remediation in schizophrenia to exploring large language models and automated approaches for neuropsychological test scoring. Alongside academia, she has worked in government departments, applying psychological principles to enhance human–computer interaction and user experience in digital systems.

Can A Large Language Model (Llm) Match Humans In Describing And Rating Complexity Of Video Stimuli?

Abstract:

Many fields are incorporating computational tools to improve research efficiency. Large Language Models (LLMs) can now process video clips quickly and consistently, helping researchers handle exponentially more data. We explored whether LLMs can accurately describe and rate the complexity of video stimuli and match traditional human ratings of complexity. We compared the performance of three LLMs—LLaMA 3-8B, LLaVA 34B, and GPT-4o—in automatically describing 62 videos from the Database of Emotional Videos from Ottawa (DEVO-2). First, we ran Python scripts with structured prompts that fed each DEVO-2 video into the three LLMs. These scripts extracted frames from the videos and asked the models to identify objects, actions, and any critical details. We then manually checked each LLM description against actual video content. After this informal analysis, we chose the strongest model to continue processing the additional video clips. To our eyes, GPT-4o outperformed LLaMA and LLaVA, consistently generating the most detailed and contextually relevant descriptions of videos. Next, we prompted the LLMs to assess the visual complexity of each video. To validate the LLMs' assessment, we selected a previously published dataset with ratings of video complexity by 24 undergraduate participants which gave us a benchmark to compare LLM-generated complexity ratings. First, however, we examined reliability/consistency of the LLM ratings. We ran GPT-4o twice on the same video and compared the results. GPT-4o consistently reported similar complexity ratings from both runs (Spearman r=0.96, p=2.1e-11). Next, we compared LLM ratings with human judgments. The results show that GPT-4's scoring method aligns well with how people perceive complexity. (Spearman r=0.74, p=0.0002). This study provides strong evidence that LLMs, particularly GPT-4o, can describe and rate the complexity of video stimuli. As LLMs continue to advance, their role in psychological research will expand, offering new opportunities for innovation in the study of human cognition.