“ Evaluating Large Language Models as Judicial Decision-Makers“

Prof. Oren Gazal (University of Haifa)

Wednesday, 20 May 2026 at 6.15 PM CEST

Large Language Models (LLMs) are increasingly shaping various domains, yet their ability to align with human judgment remains a critical challenge. This study explores the extent to which LLMs can serve as judicial decision-makers by comparing their sentencing decisions to those of 123 retired judges on two fictional cases involving rape and violence. We evaluate GPT, Gemini, and Claude using zero-shot, few-shot, and chain-of-thought prompts. LLMs showed greater consistency, producing significantly lower sentence disparity than judges. To assess accuracy, we treated the judges' average sentence as a conservative benchmark—acknowledging that the "correct" sentence is unknown. If models outperform even this minimal standard, they are closer to any plausible ground truth. Remarkably, all LLMs deviated less from the judges' mean than the judges themselves, suggesting that when properly prompted, LLMs can deliver more accurate sentencing decisions than human judges.