Opinions Divided on ChatGPT Bar Exam Performance

Researchers at the Massachusetts Institute of Technology (MIT) have raised concerns about the bar exam performance claims of ChatGPT, the latest version of OpenAI's AI language model. The original research suggested that ChatGPT outperformed 90% of human test-takers, but a Ph.D. candidate, Eric Martinez, argues that the actual performance may be lower, placing the AI model in the 68th percentile.

The discrepancy in percentile calculations revolves around the methodology used by the original researchers. Martinez's paper, titled "Re-Evaluating GPT-4's Bar Exam Performance," questions the percentile assigned to ChatGPT based on its Uniform Bar Exam (UBE) score of 297. While the original research used the February 2019 Illinois bar exam as a benchmark, Martinez suggests that the July exam provides a more accurate comparison. The July exam typically includes a higher percentage of retakers who had previously failed, resulting in lower scores overall. Based on this perspective, ChatGPT's percentile would be closer to the 68th percentile.

Martinez emphasizes that the widespread publicity surrounding ChatGPT's reported "90th percentile" performance raises concerns. If the AI model cannot adequately perform complex legal tasks, there is a risk that both lawyers and non-lawyers may rely on it for such tasks. This could have significant implications for legal accuracy and decision-making.

In response to the critique, Daniel Martin Katz, a law professor at Chicago-Kent, and Michael James Bommarito, a law professor at Michigan State, who conducted the original research in collaboration with two members from legal AI company Casetext, defend their conclusions and stand by the 90th percentile finding. However, they acknowledge the need to address points of confusion and misunderstanding in public discourse regarding their research.

It is worth noting that converting ChatGPT's UBE score into a percentile is challenging due to the lack of publicly available score distributions from the National Conference of Bar Examiners, which designs the exam. Additionally, states do not regularly or consistently release score distributions, further complicating the assessment.

Katz and Bommarito assert that their 90th percentile conclusion is conservative, as they excluded ChatGPT's high essay scores and relied on pre-COVID-19 pandemic results for comparison. They also suggest that anecdotal evidence indicates a decline in law student learning during the pandemic, which may have impacted exam performance.

OpenAI, the organization behind ChatGPT, has not yet responded to requests for comment regarding the recent debate surrounding the AI model's bar exam performance.

Differences in pass rates between the February and July bar exams can be significant. For instance, the pass rate for Illinois' most recent July exam was 68%, while the February exam had a pass rate of 43%.

The accuracy and implications of ChatGPT's bar exam performance remain a topic of discussion among researchers. While the original research claims a 90th percentile performance, a Ph.D. candidate has raised concerns about the methodology and suggests a lower percentile placement. The final version of the research paper is expected to address these concerns and clarify any misunderstandings.