I think a particularly good use case for LLMs in science would be in studying the effects of various treatments for mental health.

Rather than limiting surveys to multiple choice responses or ratings on a numerical scale, a well-trained LLM could process free form responses to open-ended questions, allowing a more detailed/nuanced evaluation of a larger number of patients/participants.

Summarization and sentiment analysis could be used not only to track changes in individual patients over time, but also may be able to detect subtle changes and warning signs that would not be apparent in more restrictive answer formats. Over time, it may even be possible to combine the LLM with other predictive models and data to help identify subtle or hidden trends, or to identify at-risk individuals in need of some additional form of intervention in a way that would be difficult for a human expert to track across a large number of patients/participants.

As great as the potential benefits, applying LLMs and AI in this manner would need to be very closely monitored and carefully evaluated. A great place to start for this would be through the HELM (Holistic Evaluation of Language Models) framework. Specifically, the Massive Multitask Language Understanding (MMLU) scenario allows for the seven metrics (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency) to be measured using across specific knowledge-intensive domains. Quite a few of the subjects included in this benchmark would be highly relevant to this use case – such as Professional Medicine, Anatomy, Clinical Knowledge, and Professional Psychology.

Given the importance (and potential risks) of this application, manual evaluation by human experts will also be necessary. A framework combining escalation of specific “flagged” findings to a human expert as well as random spot-checks would be wise.