**Translational Abstract**

Independent human double screening of thousands of study titles and abstracts is a key part of high-quality systematic reviews to avoid an error-prone selection of relevant studies that can yield incorrect statistical results. However, human double-screening is resource-demanding, often slowing the review process significantly. Thus, to mitigate this issue, we evaluated whether it is possible to substitute one of the humans in the screening process with an artificial intelligence (AI) screener instead. In other words, we examined whether Large Language Models (LLMs) reached via OpenAI’s servers can conduct title and abstract screening similar to humans. To make this assessment, we developed a new benchmark scheme based on how well human screeners typically distinguish between relevant and irrelevant studies when they do title and abstract screening in high-quality reviews, and we conducted three classification experiments on psychological review data where we knew what titles and abstracts should be included. Across the experiments, we found that LLMs can conduct title and abstract screening on par with or in some cases even better than human screeners in terms of detecting relevant studies. We also found LLMs to be good at classifying irrelevant studies correctly.

of making a thorough assessment of AI tools is to compare their performance to human performance. We, therefore, investigated how well humans typically can distinguish relevant from irrelevant studies, when they do title and abstract screening in high-quality reviews. From this investigation

Our main aim with this paper that this method can be used in high-quality outlets such as *Psychological Bulletin* since it can save valuable human resources.