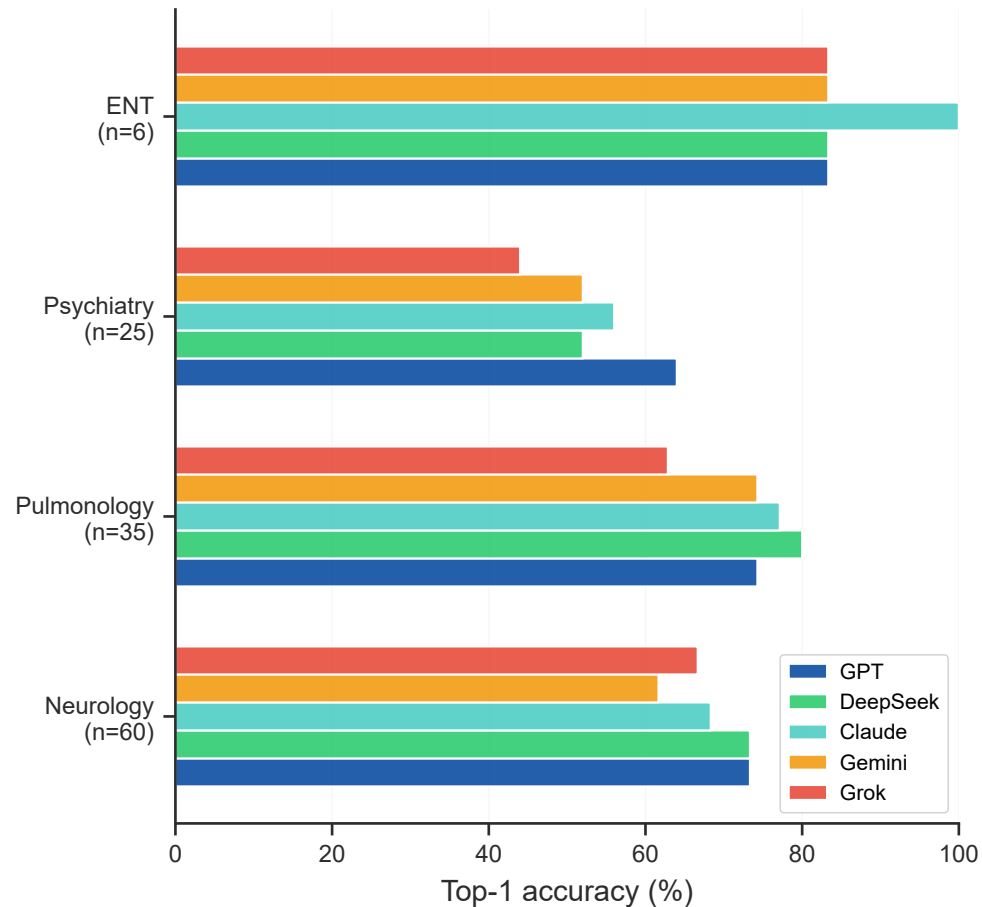**a Top-1 Diagnostic Accuracy**

b Diagnostic Failure Rate

Legend: GPT, DeepSeek, Claude, Gemini, Grok