DATA AGGREGATION

Case reports (n = 183)

Clinical journals (n=79) & DSM-5-TR (n=104)

Fictitious cases (n = 61)

Clinician-authored

DATA CURATION

Curated case reports (n = 135)

Selected for high quality

Merge

Evaluation dataset (n = 196)

MODEL EVALUATION

Models: Claude Opus 4.5 (Anthropic), DeepSeek-V3.2 (DeepSeek), Gemini 3 Pro (Google), GPT-5.2 (OpenAI)

Task 1

Automated assessment of diagnostic accuracy (n = 169)

Task 2

Clinician-grounded quantitative-qualitative analysis of clinical reasoning (n = 30)

Task 3

Illustrative comparison with two psychiatry residents (n = 30)