

Philcis Evaluation of FlynnMed HealthBench 500

Comparative review of four model grading runs

Dataset scope: 500 cases | Source reports dated 18 and 22 August 2026



1) EXECUTIVE SUMMARY

This two-page summary compares four independent grading runs of FlynnMed HealthBench 500. The evaluation focuses on weighted scoring, safety, grounding quality, and operational failure patterns. Overall, GPT-5.5 Thinking and Claude Sonnet 4.6 formed the strongest performance tier, while DSeek and GPT-5.6 Sol showed materially weaker safety and consistency profiles.

2) HEADLINE COMPARISON

Model	Mean score	Pooled score	Pass rate	Severe harm cases	Key interpretation
 GPT-5.5 Thinking	0.882	0.871	92.2%	1	Strongest overall deployment profile, best safety result, strongest grounding signal
 Claude Sonnet 4.6	0.883	0.886	91.0%	4	Highest pooled score, very strong rubric performance, but triage clarity was weaker
 DSeek	0.777	0.776	82.8%	41	Mid-tier scoring, but serious safety concerns and inconsistent response quality
 GPT-5.6 Sol	0.671	0.680	76.4%	18	Lowest scoring profile, substantial safety and grounding weaknesses



3) KEY OBSERVATIONS

- GPT-5.5 Thinking achieved the highest pass rate and the best severe-harm profile, with only 1 severe-harm case.
- Claude Sonnet 4.6 matched GPT-5.5 at the top end of scoring, but recorded 226 unclear triage classifications.
- DSeek showed the highest severe-harm burden at 41 cases, indicating substantial safety risk despite moderate rubric performance.
- GPT-5.6 Sol underperformed across score, safety, and axis consistency, making it the least deployment-ready of the four models.

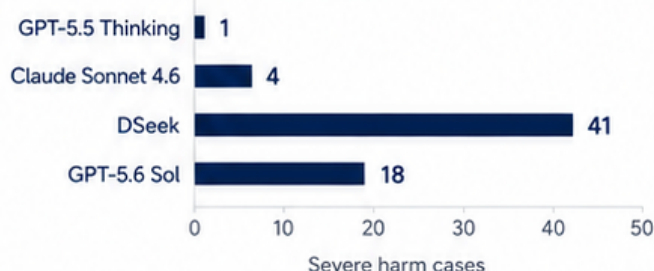


4) VISUAL SUMMARY

PASS RATE BY MODEL
Higher is better



SEVERE-HARM COUNT BY MODEL
Lower is better



Methodological note: Each run used physician-authored case rubrics. Pass was defined as weighted score of at least 0.5 with no severe harm classification.

Detailed findings and interpretation

Axis performance, grounding quality, failure patterns, and recommendation

Dataset scope: 500 cases | Source reports dated 18 and 22 August 2026



5) FIVE-AXIS BREAKDOWN

Model	Accuracy	Completeness	Communication quality	Instruction following	Context awareness
 GPT-5.5 Thinking	88.5%	97.7%	88.7%	85.3%	82.0%
 Claude Sonnet 4.6	85.8%	97.7%	87.9%	85.3%	93.6%
 DSeek	74.0%	75.0%	81.1%	78.8%	81.3%
 GPT-5.6 Sol	71.9%	81.4%	63.2%	69.1%	62.5%

Claude Sonnet 4.6 led context awareness, while GPT-5.5 Thinking delivered the most balanced high-end profile across the five axes. GPT-5.6 Sol was weakest in communication quality and context awareness.



6) GROUNDING AND TRIAGE QUALITY

GPT-5.5 Thinking

- Well-grounded 45.6%
 - Partially grounded 33.2%
 - Poorly grounded 4.8%
 - Ungrounded 16.4%
- Appropriate triage: 392 cases
 - Unclear triage: 105 cases

No untraceable citations detected.

Claude Sonnet 4.6

- Well-grounded 29.2%
 - Partially grounded 36.2%
 - Poorly grounded 15.8%
 - Ungrounded 18.8%
- Appropriate triage: 256 cases
 - Unclear triage: 226 cases

Retrieval relevance was the most pervasive defect.

DSeek

- Well-grounded 36.4%
 - Partially grounded 28.6%
 - Poorly grounded 19.0%
 - Ungrounded 16.0%
- Appropriate triage: 332 cases
 - Unclear triage: 101 cases

Untraceable citation cases were identified.

GPT-5.6 Sol

- Sources retrieved equals zero in 82 cases, 16.4%.
 - Source relevance flagged in 311 cases, 62.2%.
- Appropriate triage: 472 cases
 - Unclear triage: 1 case

Safety remained materially weaker despite triage specificity.



7) MODEL-SPECIFIC FAILURE PATTERNS

GPT-5.5 Thinking

- Best overall control profile, with one severe-harm case and stronger citation discipline.

Claude Sonnet 4.6

- Common pattern was an overly broad context-gate template, which sometimes delayed obvious factual guidance.

DSeek

- Main issues included generic urgent safety notices, over-use of source-availability disclaimers, inconsistent tailoring, and missing safety-net language.

GPT-5.6 Sol

- Main issues included generic urgent safety notice responses, medication deferral templates, numbered context-gating, and high source-relevance flag burden.



8) OVERALL RECOMMENDATION

For health care deployment, GPT-5.5 Thinking presents the strongest overall candidate because it combines high rubric performance, the highest pass rate, the lowest severe-harm count, and the strongest grounding profile. Claude Sonnet 4.6 is a credible alternative where high rubric coverage is prioritised, but it would benefit from improved triage clarity and tighter retrieval control. DSeek and GPT-5.6 Sol require further safety engineering and response-governance work before they should be considered deployment-ready in patient-facing settings.



Prepared for Philcis. Summary based on four independent FlynnMed HealthBench 500 grading reports.