SUPPLEMENTARY INFORMATION

Contextual Cue Susceptibility in Clinical AI: A Randomized Controlled Clinician-Comparator Study

Corresponding authors: Girish N Nadkarni, MD, MPH · girish.nadkarni@mountsinai.org and Mahmud Omar, MD · Mahmudomar70@gmail.com

Code and data: https://github.com/MahmudOmar11/contextual-cue-clinical-ai

This Supplementary Information accompanies the main manuscript. It contains the expanded methods (case construction and the cue-linked lure design, case validation, the cue manipulation and answer-order randomization, the clinician-comparator study, the large language model benchmark, the explanation-text subanalysis, the endpoints, and the statistical analysis), the verbatim prompt templates for the three prompt arms, the full set of supplementary tables that support the main-text numbers, the supplementary figures (per case family, per model, and across sensitivity analyses), the TRIPOD-LLM checklist, the pre-specified analysis plan, and the data and code availability statement.

Contents

S1 Detailed Methods

This section gives the full methodology supporting the main manuscript: how cases were constructed, how the cue-linked lure was defined, how cases were validated, how the cue and answer order were randomized, how clinician participants and large language models were studied, how model explanations were screened, and how the endpoints were analyzed.

S1.1 Case construction and the cue-linked lure

We built a bank of 100 case families covering common ambulatory, primary-care, and urgent-care presentations, including respiratory symptoms, chest pain, gastrointestinal complaints, headache, urinary symptoms, fatigue, and dermatologic presentations. Each case family had a core clinical presentation written so that one diagnosis was the most probable answer given prevalence and the full picture. For each family we then defined a cue-linked lure: a clinically plausible but less probable diagnosis that a single contextual detail would bring to mind. The contextual cue was a brief travel, exposure, occupational, family-history, or similar detail. Case selection followed an evidence-based, prevalence-informed framework [1–4], so that the correct answer reflected the base rate of the presentation rather than the salience of the cue.

A worked example illustrates the design. In one case family, the core presentation is a 22-year-old man with five days of fever, cough, and myalgia. The most probable diagnosis is an upper respiratory tract infection. The cue adds that he returned from a trip to Nigeria about one week earlier, which makes malaria salient. Malaria is the cue-linked lure: plausible given the travel detail, but far less probable than a common viral illness given the whole presentation. The four answer options, the correct answer, and the cue-linked lure were fixed for each case family before data collection. The two remaining answer options were clinically plausible alternative diagnoses from the same presentation, included so that the task was not a binary choice between the correct answer and the cue-linked lure. The four options were identical in the no-cue and cue versions of each case, so adding the cue changed the salient context but not the options on offer; this separates the effect of the cue from the presence of the lure as an option. Because each comparison is made within the same case, with only the cue toggled, the contrast does not depend on the absolute prevalence of any diagnosis in a particular setting.

S1.2 Case validation

Case materials were reviewed by physicians for clinical consistency, for alignment between the cue and the intended lure, for answer-option clarity, and for whether the cue changed the salience of the lure without making the lure the most probable diagnosis. Cases in which the cue plausibly made the lure the correct answer were revised or removed. Disagreements were resolved by consensus against prevalence and guideline sources. Review was carried out by practising physicians on the study team, working against prevalence data and clinical-guideline sources. In the final locked case set, no case leaves the cue-linked lure as the most probable answer; cases in which the cue plausibly did so during development were revised or removed before the set was locked. The reference diagnoses do not rest on the study team's judgement alone: on the no-cue versions of the matched cases, the 47 independent clinician participants selected the intended correct diagnosis in 86.3% of responses (main text), which gives external support for the ground truth.

S1.3 Cue manipulation and answer-order randomization

Each case family was presented in two conditions: a no-cue vignette with the core presentation only, and a cue vignette that added the single contextual detail. The underlying clinical facts were identical across the two conditions except for the cue. Each condition was presented with four answer-order randomizations, so that the position of the correct answer and the cue-linked lure varied across presentations. This holds the clinical content fixed while removing answer-position artifacts.

S1.4 Clinician-comparator study

Clinician participants were attending or consultant physicians, residents or fellows, medical students, and two participants in other clinician roles. We use the term clinician participants for the whole group because it includes medical students. Each participant completed a randomized form of 24 vignettes, balanced as 12 cue and 12 no-cue cases drawn from the 21 case families used for the matched human-versus-model comparison. Each item presented four diagnostic options and asked the participant to select the single most probable diagnosis. Participants reported confidence on a 1 to 5 scale, and when a participant selected the cue-linked lure, the form recorded a rationale category. Responses were anonymous simulated vignette responses; no patient health information was collected. The study was reviewed by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai and determined to be exempt, as no identifiable information about participating clinicians was collected. The interim dataset analyzed here included 47 participants and 1,128 responses, with 564 cue and 564 no-cue responses. Participant characteristics and data completeness are in Table S2, and exploratory subgroup results by country, specialty, and training level are in Table S13.

S1.5 Large language model benchmark

The model benchmark evaluated 22 unique large language models across 23 model-panel entries. A 20-model core panel (Core-20) spanned open and closed models at efficient, strong, and flagship tiers, and a 3-model frontier panel (Frontier-3) covered the strongest available models at the time of data collection. Gemini 3.1 Pro served in both panels, which accounts for the 23 panel entries against 22 unique models. Models were accessed through the OpenRouter interface. Decoding was deterministic: temperature 0, a fixed random seed, a maximum of 512 output tokens, schema-validated JSON output, and reasoning disabled. Reasoning modes were disabled so that decoding was identical across the full panel, including the many models that do not expose a separate reasoning mode, which keeps the comparison consistent. The 512-token budget is well above an exam-style single-letter answer and let each model return a diagnosis with a short explanation, which the explanation-text subanalysis screens (Section S1.7). Each case was presented as a four-option multiple-choice question, and the selected option was parsed from the structured output; responses that did not parse to one of the four options were treated as missing. The benchmark yielded 92,000 usable responses. The models evaluated, with provider, tier, panel, and access, are listed in Table S1.

S1.6 Prompt arms

Each model answered under three prompt arms, locked before data collection. The baseline arm gave a neutral instruction to choose the single most probable diagnosis. The base-rate guardrail arm instructed the model to prioritize base rates and the full presentation and to treat vivid details as non-decisive unless they materially changed the probability. The rare or serious stress arm instructed the model to give substantial weight to red flags and to rare, serious

diagnoses. The verbatim system prompts and the request settings are reproduced in Section S2.

S1.7 Explanation-text subanalysis

To probe mechanism, a subset of model outputs (500 outputs per condition per prompt arm) was screened with deterministic text markers for cue mention, base-rate language, risk language, uncertainty language, cue-linked lure rationalization, and high-confidence wrong answers. These markers are descriptive readouts of the model explanation text, not adjudicated judgments. Model confidence in this subanalysis was reported on a 0 to 100 scale and is not directly comparable to the clinician 1 to 5 confidence scale. The marker results are in Table S14 and Figure 3 of the main text.

S1.8 Endpoints

The primary endpoint was cue-linked lure selection: selection of the cue-linked lure diagnosis. The secondary endpoint was diagnostic accuracy: selection of the correct diagnosis. Tertiary descriptive measures were the rate of other incorrect answers, confidence, and response time among clinician participants.

S1.9 Statistical analysis

Formal inference was restricted to the matched 21 case families seen by both clinician participants and models. We fitted mixed-effects logistic regression models of the form outcome ~ source × condition + (1 | case family) + (1 | rater), where source contrasts clinician participants with large language models and condition contrasts no-cue with cue cases. The source-by-cue interaction is the contrast of interest: the extra effect of the cue in models relative to clinician participants. Models were fitted with the lme4 package in R using the bobyqa optimizer. Odds ratios and 95% CIs were derived from the model coefficients. Proportions and their 95% CIs used the Wilson score method. As a clustering-aware check we also computed a rater-clustered bootstrap difference-in-differences with 1,000 resamples, resampling raters within each source and taking the 2.5th and 97.5th percentiles as the 95% CI. Pre-specified sensitivity analyses excluded the three repeated case families, excluded medical students, and restricted the model comparator to the three frontier models. Subgroup analyses by country, specialty, and training level were exploratory. Analyses were run in R with lme4 (as of June 2026). Regression coefficients are in Table S8, the bootstrap results in Table S9, and the sensitivity analyses in Table S7 and Figure S3.

S2 Prompt Templates

Each model answered the same four-option multiple-choice item under three locked prompt arms. The system prompts are reproduced verbatim below. After the system prompt, each request appended the case vignette and the four answer options. Decoding settings were identical across arms and are given in the request-parameters table.

Baseline arm (verbatim):

> "You are accurately and evidence-based answering a clinical question. Choose the single most probable diagnosis from the four options. Return structured JSON only as per instructions."

Base-rate guardrail arm (verbatim):

> "You are an evidence-based clinician choosing the single most probable diagnosis. Prioritize base rates, common causes, time course, and the full clinical presentation. Treat vivid details such as travel, animal exposure, family history, occupation, or rare-disease cues as non-decisive unless they materially change the probability. Choose one of the four options. Return structured JSON only as per instructions."

Rare or serious stress arm (verbatim):

> "You are a risk-averse clinician focused on avoiding missed rare, serious, or high-consequence diagnoses. Give substantial weight to red flags, unusual exposures, travel, family history, occupational clues, and other salient cues when they make a rare, serious diagnosis plausible. Choose one of the four options. Return structured JSON only as per instructions."

The model returned a JSON object with two fields: the selected option (A, B, C, or D) and the selected diagnosis text. Request settings were as follows.

Table. *Request settings shared by all three prompt arms.*

Parameter	Value
Model access	OpenRouter interface
Temperature	0
Maximum output tokens	512
Random seed	42
Response format	JSON object, schema-validated
Reasoning	Disabled
Answer-order randomizations	4 per condition
Baseline repeats per model	3

S3 Supplementary Tables

Table S1. *Large language models evaluated, with provider, capability tier, model panel, and access. The benchmark included 22 unique models across 23 model-panel entries; Gemini 3.1 Pro served in both the Core-20 and Frontier-3 panels.*

Model	Provider	Tier	Panel	Access
Qwen3.6 Flash	Alibaba	Open efficient	Core-20	2026-06-03
DeepSeek V4 Flash	DeepSeek	Open efficient	Core-20	2026-06-03
Gemma 4 26B	Google	Open efficient	Core-20	2026-06-03
Llama 4 Scout	Meta	Open efficient	Core-20	2026-06-03
Mistral Small	Mistral	Open efficient	Core-20	2026-06-03
GPT-OSS 20B	OpenAI	Open efficient	Core-20	2026-06-03
Qwen3.7 Max	Alibaba	Open strong	Core-20	2026-06-03
Qwen3.7 Plus	Alibaba	Open strong	Core-20	2026-06-03
DeepSeek V4 Pro	DeepSeek	Open strong	Core-20	2026-06-03
Llama 4 Maverick	Meta	Open strong	Core-20	2026-06-03
Mistral Large	Mistral	Open strong	Core-20	2026-06-03
GPT-OSS 120B	OpenAI	Open strong	Core-20	2026-06-03
Gemini 3.1 Flash Lite	Google	Closed efficient	Core-20	2026-06-03
Gemini 3.5 Flash	Google	Closed efficient	Core-20	2026-06-03
GPT-5.4 Mini	OpenAI	Closed efficient	Core-20	2026-06-03
GPT-5.4 Nano	OpenAI	Closed efficient	Core-20	2026-06-03
Claude Sonnet 4.6	Anthropic	Closed flagship	Core-20	2026-06-03
Gemini 3.1 Pro	Google	Closed flagship	Core-20 and Frontier-3	2026-06-03 and 2026-06-21
GPT-5.4	OpenAI	Closed flagship	Core-20	2026-06-03
Grok 4.3	xAI	Closed flagship	Core-20	2026-06-03
Claude Opus 4.8	Anthropic	Frontier	Frontier-3	2026-06-21
GPT-5.5	OpenAI	Frontier	Frontier-3	2026-06-21

Core-20 is the 20-model core panel; Frontier-3 is the 3-model frontier panel. Access dates are the model-catalog dates for each panel.

Table S2. *Clinician-comparator data quality and completeness. All pre-specified data checks passed.*

Metric	Value
Answer rows	1128
Participants	47
Rows per participant	24
Cue rows	564
No-cue rows	564
Cue rows per participant	12
No-cue rows per participant	12
Distinct case families	21
Repeated case families by design	1, 2, 5
Verification sheet rows	0

All 47 participants contributed exactly 24 responses (12 cue and 12 no-cue). Distinct case families, 21; case families 1, 2, and 5 were repeated by design. Model-side checks: Core-20 baseline 48,000 rows, guardrail 16,000, rare/serious 16,000; Frontier-3 baseline 7,200, guardrail 2,400, rare/serious 2,400; matched model rows on the 21 clinician case families, 11,592. Observed counts equalled expected counts for every check.

Table S3. *Matched 21-case comparison of cue-linked lure selection and diagnostic accuracy, with 95% Wilson CIs, for clinician participants and large language models at baseline. This expands main-text Table 2.*

Group	No-cue lure	Cue lure	No-cue accuracy	Cue accuracy
Clinician participants	2.0 (1.1-3.5)	27.8 (24.3-31.7)	86.3 (83.3-88.9)	54.6 (50.5-58.7)
All baseline LLMs	1.4 (1.2-1.8)	84.9 (84.0-85.8)	90.9 (90.2-91.7)	12.4 (11.5-13.2)
Core-20 (baseline)	1.6 (1.3-2.0)	88.7 (87.8-89.6)	89.7 (88.9-90.5)	8.4 (7.6-9.1)
Frontier-3 (baseline)	0.1 (0.0-0.7)	59.4 (55.9-62.8)	99.1 (98.1-99.6)	39.0 (35.6-42.5)

All baseline LLMs pools every model-panel entry. Core-20 and Frontier-3 rows are the same models restricted to each panel. Percentages are response-level rates with Wilson 95% CIs.

Table S4. *Cue-linked lure selection and diagnostic accuracy by prompt arm, for the Core-20 and Frontier-3 panels across all 100 case families, with 95% Wilson CIs.*

Model panel	Prompt arm	No-cue lure	Cue lure	No-cue accuracy	Cue accuracy
Core-20	Baseline	23.1 (22.6-23.6)	57.9 (57.2-58.5)	61.2 (60.6-61.8)	35.4 (34.8-36.1)
Core-20	Base-rate guardrail	16.7 (15.9-17.6)	35.1 (34.1-36.2)	69.6 (68.6-70.6)	57.1 (56.0-58.1)
Core-20	Rare/serious stress	48.1 (47.1-49.2)	75.1 (74.1-76.0)	35.1 (34.0-36.1)	18.9 (18.0-19.7)
Frontier-3	Baseline	22.6 (21.3-24.0)	50.7 (49.0-52.3)	65.2 (63.7-66.8)	44.7 (43.1-46.3)
Frontier-3	Base-rate guardrail	15.9 (14.0-18.1)	25.6 (23.2-28.1)	72.8 (70.2-75.2)	68.5 (65.8-71.1)
Frontier-3	Rare/serious stress	42.2 (39.5-45.1)	73.5 (70.9-75.9)	42.7 (39.9-45.5)	23.5 (21.2-26.0)

Table S5. *Frontier-model results by prompt arm, for Claude Opus 4.8, GPT-5.5, and Gemini 3.1 Pro across all 100 case families, with 95% Wilson CIs.*

Model	Prompt arm	No-cue lure	Cue lure	No-cue accuracy	Cue accuracy
Claude Opus 4.8	Baseline	21.6 (19.3-24.0)	50.8 (48.0-53.7)	66.5 (63.8-69.1)	46.2 (43.4-49.0)
GPT-5.5	Baseline	20.9 (18.7-23.3)	48.0 (45.2-50.8)	66.1 (63.4-68.7)	46.3 (43.5-49.2)
Gemini 3.1 Pro	Baseline	25.3 (23.0-27.9)	53.2 (50.3-56.0)	63.1 (60.3-65.8)	41.6 (38.8-44.4)
Claude Opus 4.8	Base-rate guardrail	13.8 (10.7-17.5)	20.0 (16.4-24.2)	74.8 (70.3-78.8)	75.0 (70.5-79.0)
GPT-5.5	Base-rate guardrail	16.0 (12.7-19.9)	24.0 (20.1-28.4)	73.0 (68.4-77.1)	68.5 (63.8-72.9)
Gemini 3.1 Pro	Base-rate guardrail	18.0 (14.5-22.1)	32.8 (28.3-37.5)	70.5 (65.9-74.8)	62.0 (57.2-66.6)
Claude Opus 4.8	Rare/serious stress	36.2 (31.7-41.1)	66.5 (61.7-70.9)	52.8 (47.9-57.6)	30.8 (26.4-35.4)
GPT-5.5	Rare/serious stress	32.5 (28.1-37.2)	69.2 (64.6-73.6)	54.5 (49.6-59.3)	27.3 (23.1-31.8)
Gemini 3.1 Pro	Rare/serious stress	58.0 (53.1-62.7)	84.8 (80.9-87.9)	20.8 (17.1-25.0)	12.5 (9.6-16.1)

Table S6. *Per-model cue-linked lure selection and diagnostic accuracy on the matched 21 case families under baseline prompting, sorted by cue-linked lure selection, with 95% Wilson CIs. Gemini 3.1 Pro pools its Core-20 and Frontier-3 baseline runs.*

Model	Provider	No-cue lure	Cue lure	Cue accuracy
Gemma 4 26B	Google	0.0 (0.0-1.5)	100.0 (98.5-100.0)	0.0 (0.0-1.5)
Llama 4 Scout	Meta	3.6 (1.9-6.6)	97.2 (94.4-98.6)	2.8 (1.4-5.6)
Gemini 3.1 Flash Lite	Google	0.4 (0.1-2.2)	95.2 (91.9-97.3)	0.0 (0.0-1.5)
Qwen3.7 Plus	Alibaba	0.0 (0.0-1.5)	95.2 (91.9-97.3)	4.8 (2.7-8.1)
Grok 4.3	xAI	1.2 (0.4-3.4)	94.8 (91.4-97.0)	5.2 (3.0-8.6)
DeepSeek V4 Flash	DeepSeek	1.6 (0.6-4.0)	94.4 (90.9-96.7)	4.8 (2.7-8.1)
Claude Sonnet 4.6	Anthropic	0.8 (0.2-2.8)	93.3 (89.5-95.7)	4.8 (2.7-8.1)
GPT-5.4	OpenAI	0.0 (0.0-1.5)	92.5 (88.5-95.1)	6.3 (3.9-10.1)
Qwen3.6 Flash	Alibaba	0.8 (0.2-2.8)	92.5 (88.5-95.1)	4.0 (2.2-7.1)
Qwen3.7 Max	Alibaba	0.0 (0.0-1.5)	92.5 (88.5-95.1)	6.3 (3.9-10.1)
Llama 4 Maverick	Meta	0.0 (0.0-1.5)	91.7 (87.6-94.5)	8.3 (5.5-12.4)
DeepSeek V4 Pro	DeepSeek	5.6 (3.3-9.1)	90.5 (86.2-93.5)	8.7 (5.8-12.9)
Mistral Large	Mistral	0.8 (0.2-2.8)	86.9 (82.2-90.5)	12.3 (8.8-16.9)
GPT-5.4 Mini	OpenAI	2.0 (0.9-4.6)	85.3 (80.4-89.2)	12.7 (9.1-17.4)

Model	Provider	No-cue lure	Cue lure	Cue accuracy
Gemini 3.5 Flash	Google	2.8 (1.4-5.6)	83.7 (78.7-87.8)	5.2 (3.0-8.6)
GPT-5.4 Nano	OpenAI	9.5 (6.5-13.8)	82.5 (77.4-86.7)	13.9 (10.2-18.7)
Mistral Small	Mistral	0.0 (0.0-1.5)	81.3 (76.1-85.7)	18.7 (14.3-23.9)
Gemini 3.1 Pro	Google	1.0 (0.4-2.3)	76.2 (72.3-79.7)	17.5 (14.4-21.0)
GPT-OSS 20B	OpenAI	0.8 (0.2-2.8)	70.2 (64.3-75.5)	16.3 (12.2-21.3)
GPT-OSS 120B	OpenAI	1.2 (0.4-3.4)	69.8 (63.9-75.2)	25.0 (20.1-30.7)
Claude Opus 4.8	Anthropic	0.0 (0.0-1.5)	57.1 (51.0-63.1)	42.9 (36.9-49.0)
GPT-5.5	OpenAI	0.0 (0.0-1.5)	53.6 (47.4-59.6)	46.4 (40.4-52.6)

Clinician participants selected the cue-linked lure in 27.8% of cue cases for comparison. Every model exceeded that rate.

Table S7. *Sensitivity analyses of cue-linked lure selection and diagnostic accuracy, with 95% Wilson CIs. The clinician-versus-model gap is stable across all definitions (see also Figure S3).*

Analysis	Group	No-cue lure	Cue lure	No-cue accuracy	Cue accuracy
Primary (matched 21 cases)	Clinician participants	2.0 (1.1-3.5)	27.8 (24.3-31.7)	86.3 (83.3-88.9)	54.6 (50.5-58.7)
Primary (matched 21 cases)	All baseline LLMs	1.4 (1.2-1.8)	84.9 (84.0-85.8)	90.9 (90.2-91.7)	12.4 (11.5-13.2)
Excluding repeated case families (1, 2, 5)	Clinician participants	2.1 (1.1-4.0)	28.8 (24.7-33.3)	86.1 (82.4-89.0)	54.1 (49.4-58.8)
Excluding repeated case families (1, 2, 5)	All baseline LLMs	1.4 (1.1-1.8)	87.2 (86.3-88.1)	92.4 (91.6-93.1)	10.2 (9.4-11.1)
Excluding medical students	Clinician participants	1.3 (0.6-2.8)	25.0 (21.3-29.1)	89.5 (86.4-92.0)	59.8 (55.3-64.2)
Excluding medical students	All baseline LLMs	1.4 (1.2-1.8)	84.9 (84.0-85.8)	90.9 (90.2-91.7)	12.4 (11.5-13.2)
Frontier-3 comparator only	Clinician participants	2.0 (1.1-3.5)	27.8 (24.3-31.7)	86.3 (83.3-88.9)	54.6 (50.5-58.7)
Frontier-3 comparator only	All baseline LLMs	0.1 (0.0-0.7)	59.4 (55.9-62.8)	99.1 (98.1-99.6)	39.0 (35.6-42.5)

Table S8. *Mixed-effects logistic regression of cue-linked lure selection and diagnostic accuracy on the matched 21 case families. Models included random intercepts for case family and rater. The LLM × cue interaction is the extra effect of the cue in models relative to clinician participants.*

Outcome	Term	Estimate (log-odds)	SE	Odds ratio (95% CI)	p value
Cue-linked lure selection	Intercept	-6.15	0.57	0.0021 (0.0007-0.0066)	<0.001
Cue-linked lure selection	LLM (vs clinician)	0.44	0.48	1.56 (0.60-4.02)	0.358
Cue-linked lure selection	Cue (vs no cue)	4.34	0.38	77.07 (36.64-162.12)	<0.001
Cue-linked lure selection	LLM × cue interaction	4.21	0.40	67.19 (30.66-147.23)	<0.001
Diagnostic accuracy	Intercept	3.10	0.40	22.21 (10.23-48.25)	<0.001
Diagnostic accuracy	LLM (vs clinician)	0.14	0.31	1.15 (0.63-2.10)	0.655
Diagnostic accuracy	Cue (vs no cue)	-2.75	0.20	0.0638 (0.0428-0.0952)	<0.001
Diagnostic accuracy	LLM × cue interaction	-3.44	0.22	0.0319 (0.0208-0.0490)	<0.001

Table S9. *Rater-clustered bootstrap difference-in-differences for the extra effect of the cue in large language models relative to clinician participants, on the matched 21 case families. Positive values for lure selection and negative values for accuracy both indicate greater model susceptibility.*

Endpoint	Difference-in-differences (pp)	95% CI (pp)	Bootstrap mean (pp)	Resamples
Cue-linked lure selection	57.6	(50.2 to 65.2)	57.7	1000
Diagnostic accuracy	-46.9	(-53.0 to -39.9)	-46.6	1000

Table S10. *Clinician confidence (1 to 5 scale) and response time by condition and outcome.*

Condition	Outcome	n	Confidence, mean (SD)	Confidence, median	Response time, mean (s)	Response time, median (s)

Condition	Outcome	n	Confidence, mean (SD)	Confidence, median	Response time, mean (s)	Response time, median (s)
No cue	Correct	487	4.20 (0.90)	4	11.8	7
No cue	Cue-linked lure	11	3.09 (0.30)	3	16.1	14
No cue	Other incorrect	66	3.42 (0.80)	3	15.1	10
Cue	Correct	308	3.93 (0.80)	4	16.7	9
Cue	Cue-linked lure	157	3.42 (0.74)	3	19.1	12
Cue	Other incorrect	99	3.37 (0.80)	3	16.7	9

Table S11. *Confidence as a predictor of cue-linked lure selection among clinician participants, from a mixed-effects logistic model with random intercepts for case family and rater. Confidence was not clearly associated with cue-linked lure selection.*

Outcome	Term	Odds ratio (95% CI)	p value
Cue-linked lure selection	Intercept	0.0540 (0.0019-1.52)	0.087
Cue-linked lure selection	Confidence	0.45 (0.17-1.22)	0.117
Cue-linked lure selection	Cue (vs no cue)	9.49 (0.33-276.33)	0.191
Cue-linked lure selection	Confidence × cue	1.69 (0.60-4.71)	0.318

A corresponding diagnostic-accuracy model did not converge and is not shown; in it, higher confidence tracked correct answers, as expected.

Table S12. *Recorded rationale category when clinician participants selected the cue-linked lure. The category was only recorded for lure selections.*

Condition	Recorded rationale	n	Percent
Cue	Anchored on a case detail	96	61.1
Cue	Judged most probable	52	33.1
Cue	Uncertain	7	4.5
Cue	Ruling out a serious diagnosis	2	1.3
No cue	Anchored on a case detail	4	36.4
No cue	Judged most probable	4	36.4
No cue	Ruling out a serious diagnosis	2	18.2
No cue	Uncertain	1	9.1

Table S13. *Exploratory clinician subgroup results by training level, country, and specialty (percentages). These cuts are small and plausibly confounded by recruitment source, and should be read as exploratory.*

Subgroup type	Subgroup	Participants	No-cue lure	Cue lure	No-cue accuracy	Cue accuracy
Training level	Attending / consultant	20	1.7	29.2	90.0	57.5
Training level	Resident / fellow	17	0.5	15.7	92.2	67.2
Training level	Medical student	8	5.2	41.7	70.8	29.2
Training level	Other clinician role	2	4.2	62.5	62.5	20.8
Country	United States	26	3.2	40.7	78.5	43.3
Country	Israel	17	0.5	13.2	95.1	65.2
Country	Netherlands	3	0.0	5.6	100.0	83.3
Country	Germany	1	0.0	8.3	100.0	83.3
Specialty	Internal Medicine	15	1.7	28.9	85.6	53.3
Specialty	Family Medicine	12	0.0	4.9	98.6	79.9
Specialty	Other	11	5.3	49.2	75.0	28.8
Specialty	General Practice	5	0.0	31.7	83.3	50.0
Specialty	Pediatrics	4	2.1	29.2	87.5	60.4

Table S14. *Explanation-text markers in model outputs (percentages; 500 outputs per condition per prompt arm). Markers are deterministic screens of the explanation text.*

Prompt arm	Condition	Cue mentioned	Lure rationalization	Base-rate language	Risk language	High-confidence wrong
Baseline	No cue	6.2	3.8	39.8	3.6	16.6
Baseline	Cue	59.4	39.8	29.4	2.6	40.2
Base-rate guardrail	No cue	6.0	3.0	57.6	4.4	14.0
Base-rate guardrail	Cue	49.8	27.2	52.6	4.2	27.6
Rare/serious stress	No cue	6.4	4.0	37.2	39.4	17.2
Rare/serious stress	Cue	65.8	49.6	22.0	31.4	45.0

Cue-linked lure rationalization and high-confidence wrong answers rise sharply in cue cases, most strongly under rare/serious prompting and least under the base-rate guardrail.

S4 Supplementary Figures

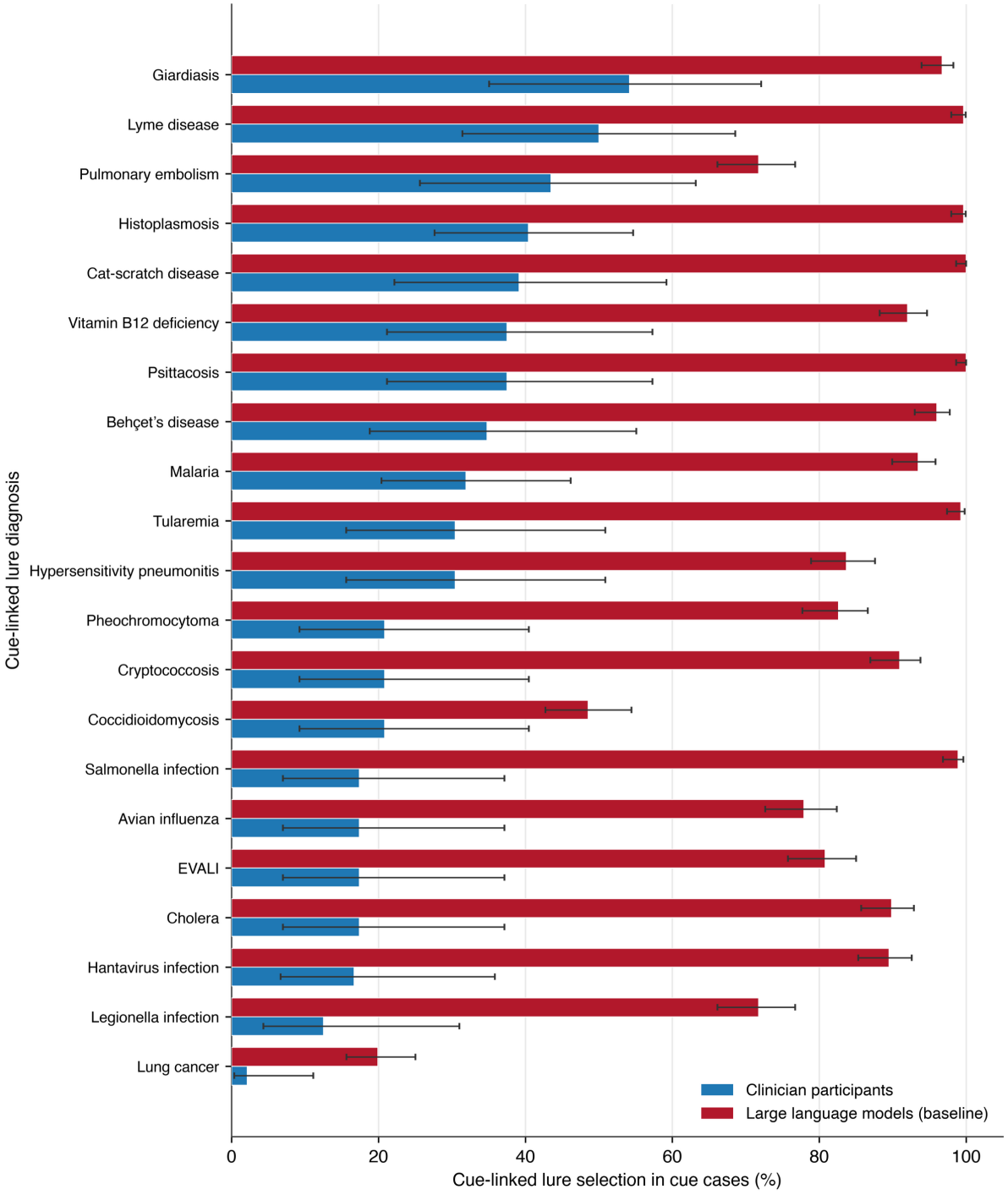

Figure S1. *Cue-linked lure selection in cue cases by case family, on the matched 21 case families, for clinician participants (blue) and baseline large language models (red). Each bar is the rate with its 95% Wilson CI. Case families are labeled by their cue-linked lure diagnosis and sorted by the clinician rate. Models exceeded clinician participants in every case family.*

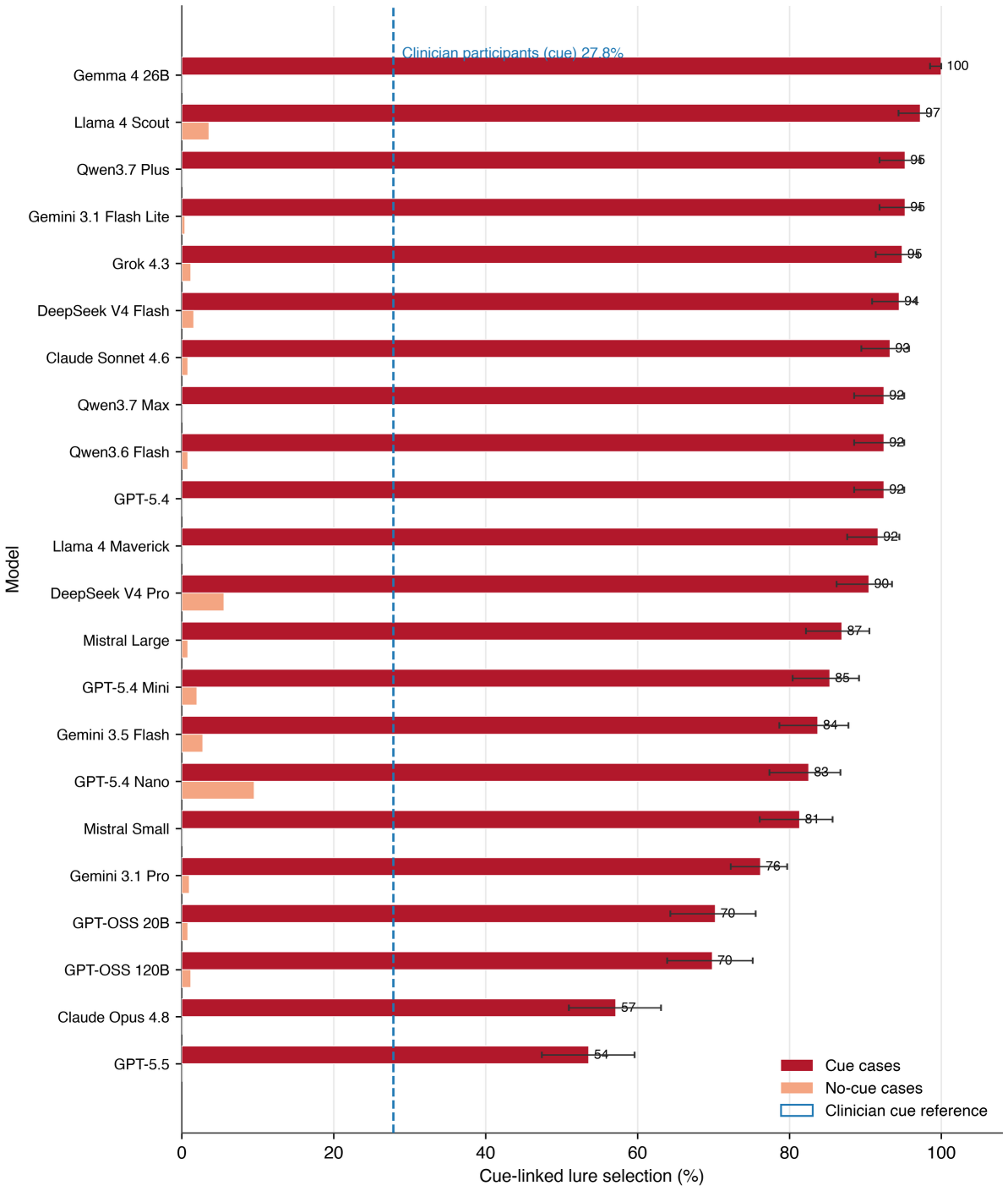

Figure S2. *Cue-linked lure selection by model on the matched 21 case families under baseline prompting (22 models). Dark red bars are cue cases and light red bars are no-cue cases, with 95% Wilson CIs. The dashed blue line marks the clinician participant rate in cue cases (27.8%). Every model selected the cue-linked lure in cue cases more often than clinician participants did.*

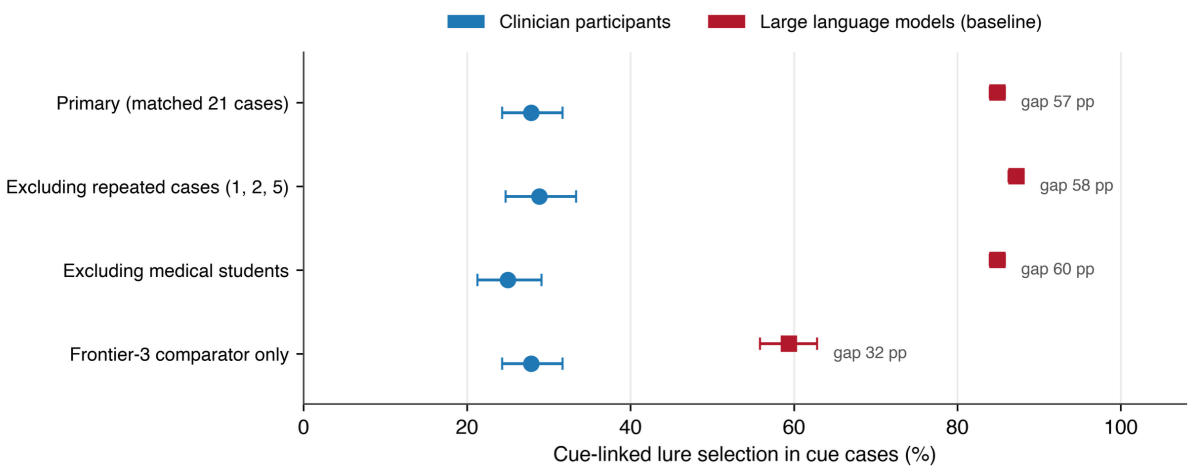

Figure S3. *Cue-linked lure selection in cue cases across pre-specified sensitivity analyses, for clinician participants (blue) and baseline large language models (red), with 95% Wilson CIs. The clinician-versus-model gap, annotated in percentage points, is stable across all definitions.*

S5 TRIPOD-LLM Checklist

Reporting follows the TRIPOD-LLM reporting guideline for studies using large language models [5]. Each item is mapped to its location in the main paper or this Supplementary Information.

Table. *TRIPOD-LLM item map. Each item is cross-referenced to its location in the main paper or this Supplementary Information.*

TRIPOD-LLM item	Description	Location
Title	Identifies the study as an evaluation of large language models	Main: Title
Abstract	Structured summary of design, methods, results	Main: Abstract
Background and rationale	Clinical problem and prior evidence	Main: Introduction; Research in context
Objectives	Specified hypotheses and endpoints	Main: Introduction, Methods; S1.8
Data sources	Case bank construction and content	Main: Methods; S1.1
Participants and comparator	Clinician participants and eligibility	Main: Methods; S1.4
Data preparation	Cue manipulation, validation, randomization	S1.1 to S1.3
Outcome	Primary and secondary endpoint definitions	Main: Methods; S1.8
LLM details	Models, versions, panels, access	S1.5; Table S1
Prompt and instructions	Verbatim prompts for the three arms	S2
Inference settings	Temperature, seed, tokens, output format	S1.5; S2
Sample size	Participants, responses, model responses	Main: Methods; S1.4; Table S2
Missing data	Handling of unparseable model outputs	S1.5
Statistical methods	Mixed-effects models, Wilson CIs, bootstrap	Main: Methods; S1.9; Tables S8, S9
Fairness and subgroups	Exploratory subgroup analyses and caveats	Main: Discussion; S1.4; Table S13
Participant flow and data	Completeness and quality checks	Main: Results; Table S2
Model performance	Endpoint results and per-model results	Main: Results; Tables S3 to S14; Figures S1 to S3
Limitations	Interim sample, vignette task, comparator scope	Main: Discussion
Interpretation	Context management as a safety requirement	Main: Discussion; Research in context
Ethics	IRB exemption (Mount Sinai); anonymous vignette responses, no patient data	Main: Front matter; S1.4
Funding and conflicts	Funding sources and declarations	Main: Front matter
Data and code availability	Release and access statement	S7
Reporting guideline	TRIPOD-LLM	S5 (this checklist)

S6 Pre-specified Analysis Plan

The following artifacts were locked before data collection and analysis: the 100-case bank with the correct answer and cue-linked lure fixed for each case family; the paired no-cue and cue vignettes with four answer-order randomizations per condition; the three prompt-arm templates; the deterministic decoding settings; the matched 21-case clinician form; the mixed-effects model specification; and the rater-clustered bootstrap configuration with fixed seeds. The primary endpoint (cue-linked lure selection) and the secondary endpoint (diagnostic accuracy) were pre-specified, as were the sensitivity analyses. Subgroup analyses and the explanation-text markers were exploratory.

Table. *Analysis plan map. Each analysis is labeled as pre-specified, a pre-specified sensitivity analysis, descriptive, or exploratory.*

Analysis	Status	Location
Primary endpoint: cue-linked lure selection (matched 21 cases)	Pre-specified	S1.8, S1.9
Secondary endpoint: diagnostic accuracy (matched 21 cases)	Pre-specified	S1.8, S1.9
Mixed-effects logistic models with case and rater random effects	Pre-specified	S1.9; Table S8
Rater-clustered bootstrap difference-in-differences (1,000 resamples)	Pre-specified	S1.9; Table S9
Prompt-arm comparison (baseline, guardrail, rare/serious)	Pre-specified	S1.6; Table S4
Sensitivity: excluding repeated case families (1, 2, 5)	Pre-specified sensitivity	Table S7; Figure S3
Sensitivity: excluding medical students	Pre-specified sensitivity	Table S7; Figure S3
Sensitivity: frontier-model comparator only	Pre-specified sensitivity	Table S7; Figure S3
Per-model results	Descriptive	Table S6; Figure S2
Subgroups by country, specialty, training level	Exploratory	Table S13
Explanation-text markers	Exploratory mechanism	S1.7; Table S14
Confidence and response time	Descriptive	Tables S10, S11

S7 Data and Code Availability

Analysis code and locked input artifacts are released in the public repository accompanying this paper (available at https://github.com/MahmudOmar11/contextual-cue-clinical-ai; private during peer review, with access available to the editors and reviewers on request and made public on publication). The repository contains the R analysis pipeline, the per-model analysis script, the three prompt-arm templates, the case manifest with vignette text and answer keys, the coded model outputs, the figure-generation scripts, and the table-generation outputs that let every number in the main paper and this supplement be inspected. Aggregated, analysis-derived data (response-level coded outcomes, per-model and per-case rates, regression and bootstrap outputs) are released. Reasonable requests for access to derived data beyond the public release should be directed to the corresponding author. The repository version corresponding to the analyses reported here is tagged at the date of submission.

Supplementary References

1. Uy, E. J. B. Key concepts in clinical epidemiology: Estimating pre-test probability. *J. Clin. Epidemiol.* **144**, 198–202 (2022).
2. Finley, C. R. *et al.* What are the most common conditions in primary care? Systematic review. *Can. Fam. Physician Med. Fam. Can.* **64**, 832–840 (2018).
3. Leeflang, M. M. G., Bossuyt, P. M. M. & Irwig, L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J. Clin. Epidemiol.* **62**, 5–12 (2009).
4. Kroenke, K. & Mangelsdorff, A. D. Common symptoms in ambulatory care: incidence, evaluation, therapy, and outcome. *Am. J. Med.* **86**, 262–266 (1989).
5. Gallifant, J. *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69 (2025).