Counterfactual Evaluation Reveals Hidden Capability Profiles in Clinical LLMs and Agents

Matt Turk
Protege Data Lab
New York City, USA
matt.turk@withprotege.ai

Abstract

Two clinical AI systems can score nearly identically on coverage-based rubrics yet behave radically differently when their patient inputs change: one updates its recommendations to match the new clinical signal, the other produces the same output regardless. Standard evaluation cannot tell them apart. We introduce the **Causal Sensitivity Score (CSS)**, a pre-registered interventional metric that mutates oncology tumor-board cases along five clinically meaningful dimensions (biomarker flips, prior-treatment failures, biomarker strips, surgery-status changes, stage perturbations) and scores in $\{0, 0.5, 1.0\}$ whether each model's recommendations update in the pre-registered correct direction. Benchmarked against the published Consensus Match Score (CMS), a coverage-based weighted recall, six frontier models from three labs in single-shot inference on 224 cases rank in nearly opposite orders on the two metrics: all six change rank, the CMS-worst model becomes CSS-best, and one model that is upper-mid on CMS is dead last on CSS. We further surface a universal safety blind spot under our pre-registered scoring rule: every frontier model fails on surgery-status interventions ($\leq$ 17.2% CSS on Family D), a finding CMS does not expose. The metric transfers directly to tool-using agents: a ReAct-style experiment shows tool use lifts CSS for five of six models (+2.5 to +20.3pp), yet the lowest-CSS model retrieves the same chart sections as the others and still does not update its recommendations, suggestive of a structural-responsiveness deficit visible only under counterfactual evaluation. Cross-judge replication and three-rater medical-professional validation confirm the aggregate findings. Interventional pre-registered metrics like CSS complement coverage-based evaluation for clinical AI agents: they capture responsiveness signal coverage cannot, and offer a candidate dense reward for future agentic RL.

Keywords

Counterfactual evaluation, causal sensitivity, agent evaluation, RL reward signals, LLM-as-judge, clinical AI

1 Introduction

LLMs and LLM-powered agents are increasingly deployed in clinical AI (treatment recommendation, triage, tumor-board summarization), where evaluation determines whether they ship. The dominant paradigm scores outputs against reference behaviors via string similarity or LLM-as-judge rubrics [1, 7]. Both ask: *does the output look right?* Neither asks: *is the model updating its output for the right reasons?* An oncology AI that proposes FOLFIRINOX for a pancreatic case scores equally well on coverage-based metrics whether the patient is treatment-naive or whether it just always proposes FOLFIRINOX for pancreatic cases.

Why this matters for agent evaluation. Frontier deployments increasingly run as tool-using agents that fetch patient information themselves. Coverage-based metrics on agent outputs face a sharper look-right-vs.-be-right problem: an agent can make many tool calls, retrieve the right information, and still produce a recommendation that ignores what it found. Interventional metrics are the natural fit because they grade *behavioral responsiveness*: did the agent's output update appropriately when its tool returns were changed?

We introduce the **Causal Sensitivity Score (CSS)**, a pre-registered interventional metric. For each intervention (flipping HER2 status, injecting a prior failed therapy, removing biomarker mentions, toggling surgery status, etc.), CSS scores in $\{0, 0.5, 1.0\}$ whether recommendations update in the pre-registered correct direction. We evaluate six frontier models from three labs (OpenAI `gpt-5`, `gpt-5.4`, `gpt-5.4-mini`; Anthropic `claude-opus-4-7`, `claude-sonnet-4-6`; xAI `grok-4.20-0309-reasoning`) in two settings: single-shot LLM inference on all 224 expert-annotated tumor-board cases (§4), and a tool-using ReAct agent [6] on a 100-tuple Family D subset where interventions propagate through tool returns rather than the prompt (§4.6).

Our findings:

- **Rank reversal (single-shot).** CMS and CSS rank the six models in nearly opposite orders ($\rho = -0.49$; all six change rank); the CMS-worst model is CSS-best, and `gpt-5.4` is upper-mid on CMS (rank 4 of 6) but dead last on CSS. All six models fail Family D (surgery status) at $\leq$ 17.2% under the pre-registered scoring rule, a universal failure CMS does not expose.
- **Agent transfer.** CSS transfers to tool-using agents without modification. Tool use lifts CSS for five of six models on Family D (+2.5 to +20.3pp); `gpt-5.4` alone is essentially unchanged despite retrieving the same chart sections as the responsive five, suggestive of structural responsiveness rather than information access.
- **Validation.** Cross-judge replication (uniform Opus) preserves rank order ($\rho = +1.00$); three-rater medical-professional annotation on a 100-tuple subset confirms aggregate per-family failure rates (Family D: LLM mean 0.10 vs. human mean 0.09).

2 Method

2.1 Pre-registered Intervention Catalog

We curate 12 interventions across five clinically motivated families (Table 1). Each is specified in a YAML catalog with five fields committed *before* any model is evaluated: applicability filter, mutation rule (regex `replace`/`delete`/`insert`), pre-registered expected

Table 1: Pre-registered intervention families. "Eligible" counts the tuples that pass the catalog applicability filter; the per-family scored n in Table 3 excludes regex no-op mutations (App G), e.g., 73/153 eligible Family C tuples produced no-ops and are dropped.

Family	Mutation type	Eligible n
A: biomarker flip	replace HER2/ER/PD-L1 status	129
B: prior treatment	inject prior-line progression	269
C: biomarker strip	delete biomarker mentions	153
D: surgery status	toggle resection history	306
E: stage perturbation	change disease stage	5

output change, $\{0, 0.5, 1.0\}$ scoring rule, and family label A–E (full schema in App A).

Pre-registration rules out post-hoc family selection and the "metric designed to fit the result" critique. The catalog and scoring rules were authored by the author and have not yet undergone independent clinical vetting (App M).

2.2 Causal Sensitivity Score

For each (model m, intervention i, case c) where i applies, we generate baseline recommendations from the unmodified packet and intervened recommendations from the mutated packet. A judge LLM receives both, the pre-registered expected change, and the scoring rule, and emits $\{0.0, 0.5, 1.0\}$ for (no change / acknowledged but unchanged / updated correctly). Each tuple is processed through a two-stage pipeline (case summary → recommendations); the judge sees only the recommendations. CSS is the mean score across (c, i) tuples, in aggregate and per-family.

Self-judging avoidance. We use `gpt-5.4` as the default judge with `claude-opus-4-7` as the judge when `gpt-5.4` is the model under test, consistent with prior self-preference findings [7]. §4.5 reports a uniform-Opus replication.

2.3 Generalization to Tool-Using Agents

CSS requires only (a) inputs that admit pre-registered counterfactual mutations and (b) a pre-specifiable correct output-update direction. Both transfer directly to tool-using agents, where mutations can be applied to tool returns (e.g., flip a knowledge-base retrieval), to planning state, or to environment observations. The single-shot setting we report on is the cleanest experimental control; §4.6 runs the same protocol against ReAct agents and shows the metric and findings transfer.

3 Experimental Setup

Cohort. 224 oncology tumor-board cases, each with a chronological patient packet (median ~80k characters) and ground-truth treatment recommendations from a two-round expert oncologist consensus protocol. Treatments are labeled `strong` (clear consensus), `tacit` (tacit consensus), `mixed` (mixed evidence), or `refusal` (clear rejection).

Models. The six frontier models listed in §1, called via official APIs at default temperature.

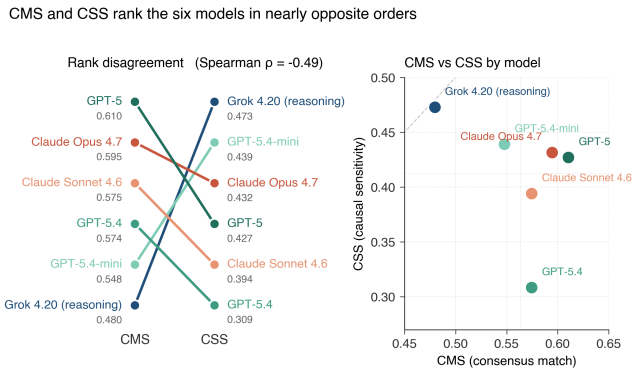

Figure 1: Rank disagreement between CMS and CSS for six frontier models from three labs. Spearman $\rho = -0.49$ (exact permutation $p = 0.36$ at $n = 6$, underpowered); all six models change rank between the two metrics. The CMS-worst model (`grok-4.20-reasoning`) is CSS-best; the CMS-best (`gpt-5`) is fourth on CSS.

Table 2: Headline: CMS vs. CSS for six frontier models. The two metrics rank the models in nearly opposite orders.

Model	CMS	rank	CSS	rank
`gpt-5`	**0.610**	1	0.427	4
`claude-opus-4-7`	0.595	2	0.432	3
`claude-sonnet-4-6`	0.575	3	0.394	5
`gpt-5.4`	0.575	4	**0.309**	**6**
`gpt-5.4-mini`	0.548	5	0.439	2
`grok-4.20-reasoning`	**0.480**	**6**	**0.473**	**1**

Comparison metric: Consensus Match Score (CMS). A published weighted recall against the oncologist-consensus treatment list:

$$\text{CMS} = 0.6R_{\text{strong}} + 0.2R_{\text{tacit}} + 0.15(1 - V_{\text{refusal}}) + 0.05P_{\text{extra}},$$

with R recalls of strong/tacit-consensus treatments, V_{refusal} the rate of recommending a rejected treatment, and P_{extra} judge-rated plausibility of off-list recommendations. CMS measures *output coverage* (does the recommendation overlap the consensus?); CSS, by contrast, measures *input responsiveness*.

4 Results

4.1 Rank Disagreement Between CMS and CSS

Table 2 and Figure 1 show the headline. The six models cluster within 13.1pp on CMS (0.480–0.610) but span 16.4pp on CSS (0.309–0.473); Spearman $\rho = -0.49$ (exact permutation $p = 0.36$ at $n = 6$, underpowered). *All six* models change rank: the most striking flip is `grok-4.20-reasoning` (CMS 6 → CSS 1) and `gpt-5` (CMS 1 → CSS 4). We treat the rank disagreement as a descriptive pattern; adding models will sharpen the inferential claim.

Table 3: Per-family CSS, six models, five intervention families. Bold = winner per row.

Family	opus-4-7	sonnet-4-6	gpt-5	gpt-5.4	gpt-5.4-mini	grok-4.20
A: biomarker flip (n=129)	**0.523**	0.415	0.504	0.326	0.477	0.496
B: prior treatment failure (n=269)	0.770	0.766	0.779	0.582	0.794	**0.868**
C: biomarker strip (n=80)	0.316	0.269	**0.406**	0.394	0.294	0.225
D: surgery status (n=306)	0.119	0.083	0.082	0.039	0.142	**0.172**
Aggregate	0.432	0.394	0.427	0.309	0.439	**0.473**

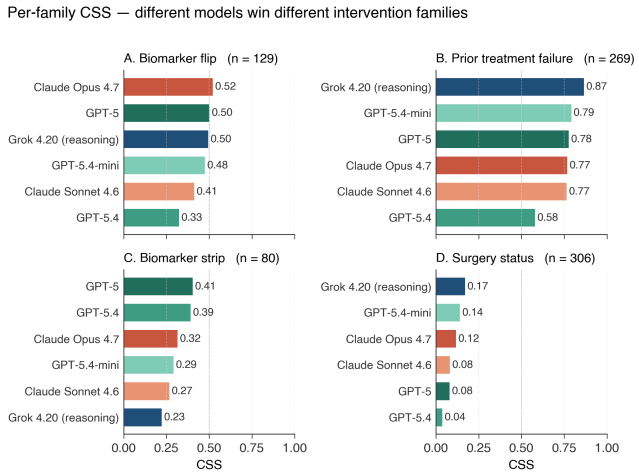

Figure 2: Per-family CSS for six frontier models, small-multiples view. Different models win different families: `claude-opus-4-7` on biomarker recognition (A), `gpt-5` on biomarker stripping (C), `grok-4.20` on prior-treatment (B) and surgery status (D). All six models fail catastrophically on Family D.

4.2 Per-Family Capability Profiles

Table 3 decomposes CSS by intervention family. Different models win different families: `claude-opus-4-7` on biomarker flips (A); `gpt-5` on biomarker stripping (C); `grok-4.20` on prior-treatment failure (B) and surgery status (D). `gpt-5.4` is dead last on A, B, and D; on C it is second (behind `gpt-5`). This per-family decomposition reveals capability profiles aggregate metrics destroy. (Family E, stage-perturbation, has only $n = 5$ eligible cases; we do not draw conclusions from it, App K.)

4.3 Universal Failure Mode: Family D

The strongest model on surgery-status interventions (`grok-4.20`) scores 17.2%; the weakest (`gpt-5.4`) scores 3.9%. *Every* frontier model from *every* lab fails to update treatment recommendations correctly when surgery status flips. This is a clinically meaningful safety finding (treatment timing depends entirely on whether the patient was resected) that CMS does not surface, because CMS only checks recommendation overlap with consensus, not behavioral change under a counterfactual.

Table 4: Per-model score distribution. "Wrong" = 0.0, "Partial" = 0.5, "Correct" = 1.0.

Model	Wrong	Partial	Correct
`grok-4.20-reasoning`	**40.8%**	23.8%	35.4%
`gpt-5`	50.6%	13.4%	**36.0%**
`gpt-5.4-mini`	45.2%	21.7%	33.1%
`claude-opus-4-7`	47.8%	18.0%	34.1%
`claude-sonnet-4-6`	54.0%	13.2%	32.8%
`gpt-5.4`	**59.6%**	19.1%	**21.3%**

4.4 Score-Distribution Diagnostics

The score distribution (Table 4) concretizes the two failure modes: `gpt-5.4` sits at 60% wrong-direction / 21% correct (the "looks fine on CMS, structurally less responsive" case); `grok-4.20-reasoning` is the inverse, with the lowest wrong-direction rate (40.8%) and second-highest correct rate (35.4%).

4.5 Judge Sensitivity (Cross-Judge Replication)

To rule out the asymmetric judge dispatch as a confound, we re-judge every tuple with `claude-opus-4-7` as a single judge for *all* models on the same 4,727 tuples. Rank order is identical under both configurations (Spearman $\rho = +1.00$); per-model inter-judge κ on $\{0, 0.5, 1.0\}$ is 0.61–0.69 for the five cross-judged models. Opus is a *stricter* judge than gpt-5.4: aggregate CSS drops 4–7pp under Opus for the five non-gpt-5.4 models. Since `gpt-5.4` was already Opus-judged by default, the original asymmetric dispatch was biased *against* `gpt-5.4` (held to a stricter standard than the other five, which were judged by the more lenient gpt-5.4), so its persistent last-place ranking is the harder of the two directions to obtain by chance. `gpt-5.4` still ranks last under uniform Opus judging; the deficit is not a judge-dispatch artifact (full table in App H). Construct validity holds across hard-case, latent, and cancer-category strata (App B).

4.6 Generalization Experiment: Tool-Using Agent

We re-run Family D in a tool-using agent setting: the agent has no chart in context and a single `read_chart_section(section)` tool over ten chart sections (demographics, diagnoses, biomarkers, medications, procedures, encounters, labs, vitals, allergies, overview). It investigates ReAct-style [6]; interventions mutate the underlying packet so the change propagates through retrieval. Judge and scoring are unchanged. We evaluate all six models on 100 tuples

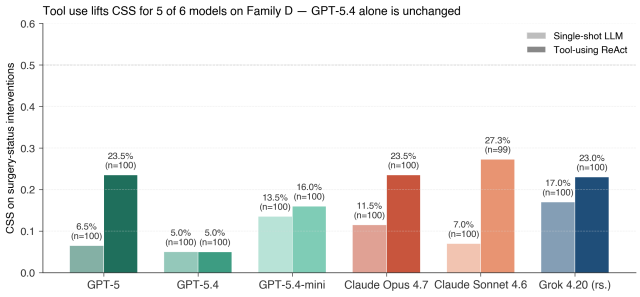

Figure 3: Tool use lifts CSS on Family D for 5 of 6 frontier models (gain +2.5 to +20.3pp). `gpt-5.4` is essentially unchanged (0.050 → 0.050), consistent with a structural responsiveness deficit rather than information-access failure. Up to 100 case-intervention tuples per model (one sonnet row dropped as missing data, $n = 99$).

each; all call the tool 7.0–8.3 times per case and query `procedures` at similar rates; they retrieve the same information.

Five of six models lift substantially under tool use, with single-shot→tool-using CSS gains of +17.0pp (`gpt-5`), +20.3pp (`claude-sonnet-4-6`, rank 4→1 on Family D), +12.0pp (`claude-opus-4-7`), +6.0pp (`grok-4.20`), +2.5pp (`gpt-5.4-mini`); `gpt-5.4` is essentially unchanged (0.050 → 0.050). Tool use helps 19–34 of up to 100 tuples for responders, only 9 for `gpt-5.4` (App C). The asymmetry is substantive: `gpt-5.4` retrieves the same sections as the responsive five and still does not update, suggestive of structural responsiveness rather than information access. Even the best tool-using model (0.273) sits well below 50%, so tool use mitigates but does not close the Family D blind spot.

4.7 Human Validation

Three medical-professional annotators independently scored 100 stratified (case, intervention, model) tuples on the same $\{0, 0.5, 1.0\}$ scale, blinded to model identity and LLM judge scores. Pairwise human-human Cohen's κ ranges 0.40–0.72; LLM-vs-majority κ = 0.46 (69/100 exact agreement, App L). Per-family LLM-vs-majority κ is highest on A (0.67) and lowest on C (0.07); D is 0.16. Crucially, the *aggregate* CSS rates agree closely: Family D LLM mean = 0.10 vs. human = 0.09; C is 0.30 vs. 0.33. Per-row case-level agreement is moderate, so headline claims should be read as population-level rather than per-case reliability.

Annotators flagged 37/100 rows as medically incoherent (29/49 in Family D); the universal Family D blind spot survives restriction to coherent-only rows (App J).

5 Limitations

Population-level claims. Per-row LLM-human κ for D and C is moderate-to-low (0.16, 0.07); headline claims should be read as population-level CSS properties rather than per-case reliability. **Regex-based counterfactuals.** Mutations admit three failure modes that score a correctly-refusing model 0.0 under the pre-registered rule: semantic no-ops (regex changes text but not meaning), incomplete propagation (one chart section changed while others still imply the original fact, which the model may treat as a data-entry error), and medical incoherence (29/49 D-rows). **Sample size.** $n = 6$ models renders the rank-correlation test underpowered ($p = 0.36$); the validation subset is $n = 100$ (15–49 per family), so per-family κ carries sampling noise. **Scope.** Findings are specific to oncology tumor-board cases; CSS methodology generalizes elsewhere given a domain-appropriate catalog. **Agent attribution.** The agent setting introduces variables besides tool use (sectioned retrieval, ReAct prompting, 8 KB cap); the gpt-5.4 persistence is consistent with structural responsiveness, not proof. Catalog provenance and camera-ready / future-work plans are in App M.

6 Related Work

LLM-as-judge reliability is established in Zheng et al. [7]; counterfactual probing of LLM reasoning in Saparov & He [4]; holistic and agent-eval aggregation in HELM [1] and AgentBench [3]; clinical LLM QA evaluation in Singhal et al. [5]; step-level process rewards in Lightman et al. [2]. CSS extends interventional probing to agentic clinical outputs, decomposes by intervention family, and operates as a candidate dense reward in the spirit of process supervision.

7 Conclusion

We introduce the Causal Sensitivity Score, a pre-registered counterfactual metric, and apply it to six frontier models on 224 oncology cases (single-shot) and a 100-tuple Family D subset (tool-using ReAct), benchmarked against the Consensus Match Score (CMS) coverage metric. Under the pre-registered scoring rule, CMS and CSS rank the models in nearly opposite orders: all six change rank, the CMS-worst model becomes CSS-best, and `gpt-5.4` sits at CMS rank 4 of 6 but is dead last on CSS. Every model fails Family D surgery-status interventions at $\leq 17.2\%$, a universal safety blind spot CMS does not expose. In the agent setting, tool use lifts CSS for five of six models (+2.5 to +20.3pp); `gpt-5.4` alone is essentially unchanged despite retrieving the same chart sections as the responsive five, a pattern consistent with structural responsiveness rather than information access. Cross-judge replication and three-rater medical-professional validation support the aggregate findings.

We see CSS as a complement to CMS rather than a replacement: it captures input-responsiveness signal coverage-based metrics cannot, transfers without modification from single-shot LLMs to tool-using agents, and offers a candidate dense reward for future agentic RL experiments. Together, these results indicate that the rank disagreement between CMS and CSS, the universal Family D failure, and the gpt-5.4 retrieval-without-update pattern are robust to judge choice and to human adjudication on the validation subset. Whether a clinical AI agent's recommendations update appropriately when the patient's facts change is an evaluation question worth measuring directly, and one that coverage-based metrics cannot answer.

References

[1] Percy Liang et al. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023).

[2] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. *International Conference on Learning Representations* (2024).

[3] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2024. AgentBench: Evaluating LLMs as Agents. In *International Conference on Learning Representations*.

[4] Abulhair Saparov and He He. 2023. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. *International Conference on Learning Representations* (2023).
[5] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv preprint arXiv:2305.09617* (2023).
[6] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*.
[7] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

Acknowledgments

We thank Engy Ziedan and Wes Hopkins of the Protege Data Lab team for their review and assistance in producing and validating these results, and the medical-professional annotators who graded the human-validation subset.

A Pre-Registered Intervention Catalog (Excerpt)

Each entry in the catalog has five fields, all committed before any model is evaluated. We include two representative entries here; the full catalog (12 entries across families A–E) is in the artifact release.

```
- id: A2_HER2_positive_to_negative
  family: A_biomarker_flip
  applicability:
    cancer_type: ["breast", "gastric"]
    requires_match_in_packet:
      "HER2[\\s-]?(positive|\\+)"
  mutation_type: replace
  mutation_pattern:
      "HER2[\\s-]?(positive|\\+)"
  mutation_replacement: "HER2-negative"
  expected_change:
    must_drop: ["trastuzumab", "T-DXd",
                "pertuzumab", "tucatinib"]
  scoring_rule:
    1.0: "all HER2-targeted recs dropped"
    0.5: "some dropped, or hedged in
          rationale only"
    0.0: "HER2-targeted recs unchanged"

- id: D1_remove_surgery_history
  family: D_surgery_status
  applicability:
    requires_match_in_packet:
      "\\[Procedure\\][^\\n]*?(resection|
       craniotomy|lumpectomy|mastectomy|
       gastrectomy|whipple|hepatectomy)"
  mutation_type: delete
  mutation_pattern:
      "\\[Procedure\\][^\\n]*?(resection|
       ...)[\\s\\S]*?(?=\\n---|\\n\\[)"
  expected_change:
    timing_must_shift:
      - adjuvant -> primary
      - post-operative -> definitive
    rationale_must_mention: "resection
        no longer documented"
  scoring_rule:
    1.0: "timing shifted as specified"
    0.5: "rationale acknowledges but
          recommendations unchanged"
    0.0: "no change"
```

B Sub-Population Stratification

CSS by latent-consensus flag and by cancer-category folder, six models. The gpt-5.4 deficit persists across all strata. Grok leads in 4 of 5 strata; in the mixed-evidence stratum, `gpt-5.4-mini` narrowly edges Grok (0.529 vs. 0.471).

Table 5: Sub-population CSS (all six models, all interventions).

Subset	opus	sonnet	gpt-5	gpt-5.4	mini	grok
Non-latent	0.483	0.446	0.458	0.321	0.492	0.500
Latent	0.422	0.385	0.422	0.306	0.430	0.468
Clear cons.	0.434	0.383	0.425	0.317	0.434	0.471
Tacit cons.	0.426	0.413	0.429	0.293	0.442	0.478
Mixed evid.	0.441	0.441	0.441	0.324	0.529	0.471

C Tool-Using Agent: Per-Case Effects

Per-model breakdown of how tool use changes the score on each of the 100 matched (case, intervention) tuples on Family D.

Table 6: Per-case effect of tool use on Family D. "Help" = tool-using CSS strictly higher; "Hurt" = tool-using CSS strictly lower; "Same" = unchanged.

Model	Help	Same	Hurt
`claude-sonnet-4-6`	34	64	1
`gpt-5`	33	64	3
`claude-opus-4-7`	28	66	6
`grok-4.20-reasoning`	20	73	7
`gpt-5.4-mini`	19	68	13
`gpt-5.4`	9	85	6

D Tool-Using Agent: Score Distribution Shift

Wrong / partial / correct breakdown for the agentic experiment, single-shot LLM (top half) vs. tool-using ReAct (bottom half), six models. The wrong-direction rate drops 6–30pp for the five responsive models; only 3pp for `gpt-5.4`.

E Tool-Use Pattern

F Compute and Cost

- **Single-shot baselines.** 6 models × 224 cases = 1,344 baseline inferences (gpt-5 and gpt-5.4-mini reused from prior delivery; remaining 4 models run fresh). Baseline CMS judge: 6 × 224 = 1,344 judge calls.

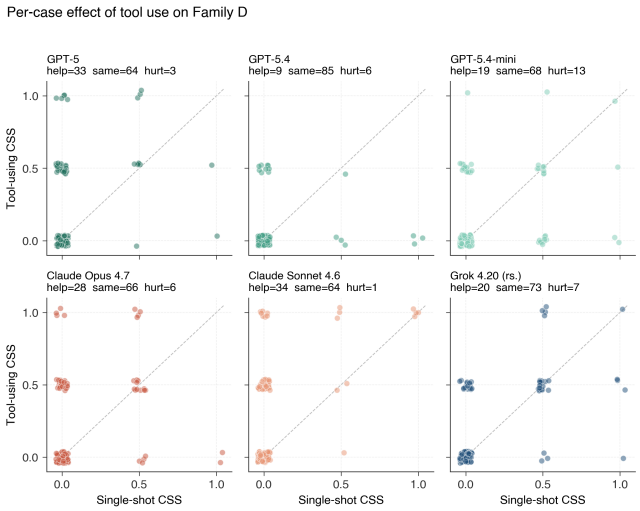

Figure 4: Per-case effect of tool use. Each point is one (case, intervention) tuple; x-axis is single-shot CSS, y-axis is tool-using CSS. Points above the diagonal are tuples where tool use helped. The mass of points above the diagonal for five models, and the near-diagonal cluster for `gpt-5.4`, makes the asymmetry visible at the per-case level.

Model	Wrong	Partial	Correct
Single-shot LLM (matched 100 tuples)			
`gpt-5`	89%	9%	2%
`gpt-5.4`	93%	4%	3%
`gpt-5.4-mini`	77%	19%	4%
`claude-opus-4-7`	79%	19%	2%
`claude-sonnet-4-6`	90%	6%	4%
`grok-4.20-reasoning`	71%	24%	5%
Tool-using ReAct (matched 100 tuples)			
`gpt-5`	61%	31%	8%
`gpt-5.4`	90%	10%	0%
`gpt-5.4-mini`	71%	26%	3%
`claude-opus-4-7`	61%	31%	8%
`claude-sonnet-4-6`	60%	26%	14%
`grok-4.20-reasoning`	59%	36%	5%

- **Single-shot interventions.** 6 models × 789 mutated tuples = 4,734 intervention-model inferences. CSS judge: 4,734 calls (with self-judge override for `gpt-5.4` routed to Opus).
- **Tool-using experiment.** 6 models × (50 baseline cases + 100 intervention tuples) = 900 trajectories, average 7.5 tool calls each, ≈ 7,600 tool-augmented LLM calls. CSS judge: 600 calls.

G Implementation Notes

Mutation no-ops. 73 of 862 eligible (case, intervention) tuples produced no-op mutations (the regex pattern was eligible at the catalog level but did not match the source packet); these are silently dropped from CSS scoring rather than scored as 0.

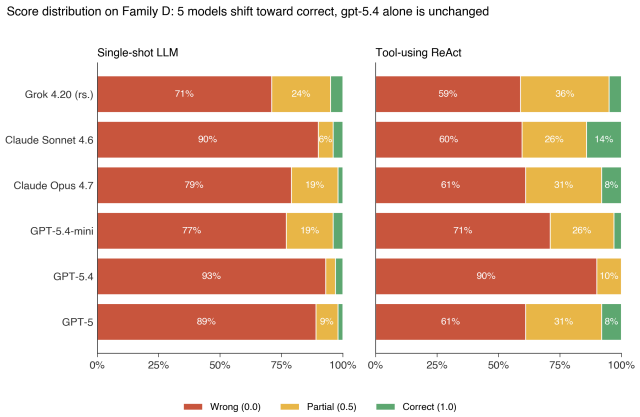

Figure 5: Score-distribution shift under tool use. Stacked bars show wrong/partial/correct breakdown for each model in single-shot (left) vs. tool-using (right) settings on Family D. The wrong-direction rate drops 6–30pp for the five responsive models; `gpt-5.4` barely moves.

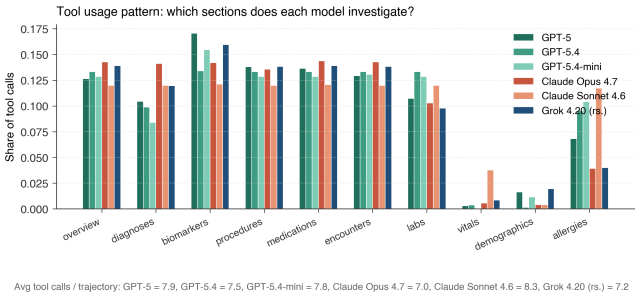

Figure 6: Tool-call distribution across the ten chart sections, averaged across 100 case-intervention tuples per model. All six models call the tool 7.0–8.3 times per case and query `procedures` (where surgery history lives) at similar rates; they retrieve roughly the same information. The CSS asymmetry across models is therefore not explained by retrieval differences.

Tool-using agent. ReAct loop with a 8-call cap (no model hit the cap; mean 7.0–8.3). Chart segmentation is regex-based on tagged entries (`[Medication]`, `[Procedure]`, `[Encounter]`, `[Lab Results]`, `[Vitals]`) and on header-delimited blocks (`PATIENT DEMOGRAPHICS`, `CONDITIONS`, `ALLERGIES`). Each section is capped at 8 KB.

Anthropic vs. OpenAI tool-call dispatch. The agent harness has provider-specific paths: OpenAI/xAI use the chat-completions `tool_calls` format; Anthropic uses the Messages API `tool_use` block format. Both wrap a single shared tool definition, so the agent sees identical tool semantics regardless of provider.

H Judge Sensitivity (Cross-Judge Replication) Table

Per-model CSS under the default judge dispatch (`gpt-5.4` for five models, Opus override for `gpt-5.4` as model) versus a uniform Opus-only judge on the same 4,727 valid tuples. Rank order is identical under both ($\rho_{\text{judges}} = +1.00$); inter-judge κ on $\{0, 0.5, 1.0\}$ labels is 0.61–0.69 (substantial). Note that `gpt-5.4`'s row is not a judge-swap: the default dispatch already routed it to Opus, so the +0.5pp shift reflects stochastic rerun variability between two Opus passes, not a different judge model.

Model	$\text{CSS}_{\text{default}}$	$\text{CSS}_{\text{Opus-only}}$	Δ	rank
`gpt-5`	0.427	0.359	−6.8pp	4
`claude-opus-4-7`	0.432	0.372	−6.0pp	3
`gpt-5.4`	0.309	0.314	+0.5pp	**6**
`claude-sonnet-4-6`	0.394	0.341	−5.3pp	5
`gpt-5.4-mini`	0.439	0.391	−4.8pp	2
`grok-4.20-reasoning`	**0.473**	**0.428**	−4.4pp	**1**

I Agentic Comparison: Single-Shot vs. Tool-Using ReAct

Family D matched 100 case-intervention tuples per model.

Model	Single-shot CSS	Tool-using CSS	Δ
`gpt-5`	0.065	0.235	+17.0pp
`gpt-5.4`	0.050	0.050	0.0pp
`gpt-5.4-mini`	0.135	0.160	+2.5pp
`claude-opus-4-7`	0.115	0.235	+12.0pp
`claude-sonnet-4-6`	0.070	**0.273**	**+20.3pp**
`grok-4.20-reasoning`	0.170	0.230	+6.0pp

J Human Validation: Family D Coherence Analysis

Of 49 Family D tuples in the human-annotated subset, 29 (59%) were flagged by at least one of three annotators as creating a medically incoherent scenario (e.g., curative-intent surgical resection inserted into a metastatic patient's chart). All values below are computed on the 49-tuple human-annotated subset (*not* the full 306-tuple Family D); n_{full} is the number of D-tuples for that model in the subset and n_{coh} is the subset after removing annotator-flagged rows. Per-model coherent-only CSS is the same or lower than the full-subset CSS for five of six models (`gpt-5.4-mini` is the exception):

Model	n_{full}	n_{coh}	LLM_{full}	LLM_{coh}	$\text{Human}_{\text{coh}}$
`gpt-5.4-mini`	8	3	0.313	0.333	0.000
`gpt-5`	8	4	0.125	0.125	0.000
`grok-4.20-reasoning`	8	5	0.125	0.100	0.400
`claude-opus-4-7`	9	4	0.056	0.000	0.000
`claude-sonnet-4-6`	8	1	0.000	0.000	0.000
`gpt-5.4`	8	3	0.000	0.000	0.000

The catalog issue is concentrated in insert-type mutations (D2 inserts a surgical resection; E1 inserts metastasis) on cases whose disease state is incompatible with the inserted event. For the camera-ready we will tighten D2 and E1 applicability filters with `¬metastatic` negative-match guards.

K Family E (Stage Perturbation, $n = 5$)

Reported for completeness; per-family numbers are noisy at $n = 5$ and we draw no conclusions. Per-model CSS: `opus-4-7` 0.800, `sonnet-4-6` 0.900, `gpt-5` 1.000, `gpt-5.4` 0.300, `gpt-5.4-mini` 0.900, `grok-4.20-reasoning` 1.000.

L LLM Judge vs. Human Majority Scatter

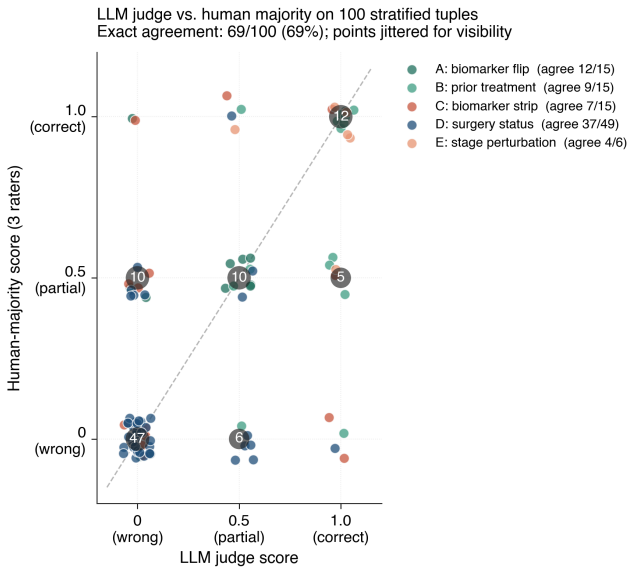

Figure 7: Case-level LLM judge vs. human-majority score on the 100-tuple validation subset. Points jittered for visibility on the $\{0, 0.5, 1.0\}$ grid; color by intervention family; counts shown in disagreement cells with ≥ 5 tuples. Exact agreement is 69/100. Most disagreement clusters at the partial-credit boundary (LLM scores 0.0 where humans score 0.5, and vice versa).

M Camera-Ready Refinements and Future Work

The workshop camera-ready window is three days. We scope camera-ready refinements to what is achievable in that window and label the rest as future work.

Catalog authorship. The intervention catalog and per-intervention scoring rules were authored by the author and have not yet undergone independent clinical vetting. The rules therefore reflect one researcher's interpretation of what each intervention should change about a model's output, which may differ from a practicing clinician's expectations and may miss clinically meaningful update patterns not enumerated in the catalog. An oncologist review of rule coverage and clinical alignment is a camera-ready item.

Camera-ready. (i) *Tightened insert-filter for D2 and E1.* We will add a ¬`metastatic` negative-match guard to D2 and E1 applicability filters so curative-intent resection and stage-perturbation insertions are not applied to already-metastatic charts, and re-run the Family D and E subsets on the filtered case set. (ii) *Refusal-credit branch in the scoring rule.* We will add a 0.5 credit when the model's rationale explicitly identifies the inserted scenario as medically incoherent and leaves recommendations unchanged, and re-judge the affected rows. (iii) *Bootstrap CIs over cases.* We will report case-resampled 95% intervals on per-model CSS and on the rank ordering to characterize stability of the headline reversals.

Future work (beyond camera-ready). (i) *Retrieval-controlled agent baseline* that feeds each model the exact retrieved snippets without the agent loop, to isolate tool-use structural-responsiveness effects from prompt-structure effects. (ii) *Semantic-no-op and incomplete-propagation audits* per mutated chart, requiring a separate validator and a sample-based calibration step. (iii) *Second human-annotation round* with expanded adjudication and refined per-intervention scoring rules to lift per-row κ in Families C and D. (iv) *Additional frontier models*, including Gemini and additional reasoning variants. (v) *Bootstrap CIs over judges* (additional uniform-judge passes), which are higher cost given per-pass compute.

Counterfactual validity (deep dive). CSS as computed treats every text-level mutation as a clinically meaningful counterfactual, but three failure modes can produce false low scores: (i) *semantic no-op* mutations where regex-level text changes do not change clinical meaning; (ii) *incomplete propagation* where one part of the chart is mutated while other parts continue to imply the original fact (e.g., deleting a Procedure block while encounter notes still refer to "post-op"); (iii) *medical incoherence* where the inserted scenario is medically impossible (e.g., curative resection on a metastatic patient), surfaced through human validation as 29/49 Family D rows. A model that correctly refuses to update on any of these is scored 0.0 under the pre-registered rule. The camera-ready filter tightening and refusal-credit branch above address (iii) directly; (i) and (ii) require the semantic-validity audits in future work.

Pre-registration tradeoff. Pre-registration commits us to a scoring rule before observing outputs, which protects against post-hoc cherry picking but cannot distinguish causal-sensitivity failure from correct refusal of incoherent or non-actionable inputs. We report results under the original rule and flag the medical-incoherence fraction transparently in §4.7; quantifying semantic no-ops and incomplete-propagation fractions requires the semantic-validity audits listed above.