The White Coat Effect: Specialist Personas Drive Harmful Diagnostic Drift More Than Patient Personas in Clinical LLMs

Taeil Matthew Kim, BS[1,3,*], Luyang Luo, PhD[2,3], Sung Eun Kim, MD[3], Alexandra N. Willauer, MD[4], and Pranav Rajpurkar, PhD[3]

[1]*Department of Computer Science, Stanford University, Stanford, CA, USA*
[2]*Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA*
[3]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*
[4]*Division of Gastroenterology, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA*
[*]*Corresponding author*
E-mail: tmattkim@stanford.edu

Clinical language models serve both patients and clinicians, and the concern is that patients fare worse for describing symptoms in plain language. We test that directly by discussing 897 MedCaseReasoning test cases with both a simulated lay patient and a simulated specialist on three frontier models (GPT-5, GPT-5.5, and Gemini 3.5 Flash), scored by an automated judge that blinded two-clinician review places at the level of agreement between the two clinicians, with metrics for authority positioning, differential movement, and whether the answer is kept, lost, or recovered. Every model becomes more deferential toward the specialist (Authority Score gaps of 0.20, 0.37 and 0.66, all $p < 0.001$) and revises its stated diagnosis for that persona 16 to 35 times more often, even though the specialist never asserts an incorrect answer and supplies no new evidence. Those revisions are rarely corrections: 74% to 83% of cases revised from a correct answer end incorrect, whereas revising an incorrect answer recovers it in only 5% to 15% of cases. No model shows evidence of the predicted patient penalty. Rendering one register-neutral skeleton into both registers separates the register from the content on the GPT pair: the deference gap survives matching ($p < 0.001$), while specialist revision falls to 3% or below and accuracy recovers 9 to 15 points. Register alone moves how the model positions itself; flipping the answer takes the content the specialist supplied. Safeguards should be explicitly evaluated in clinician-facing decision support, where technical diagnostic challenges may induce unsupported revision.

Keywords: Large Language Models, Clinical Decision Support, Sycophancy, Multi-Turn Evaluation, Conversational Register, Clinical Reasoning

1. Introduction

A clinical language model answers users with varying degrees of clinical expertise (e.g., a cardiologist citing a guideline or a patient describing symptoms in lay terms). If the model's reasoning is grounded in the clinical case, then its diagnosis should not depend on who is asking. This condition is known to fail.[1,2] In clinical dialogue, models abandon correct diagnoses under sustained pressure,[3–5] but that pressure always asserts an answer, leaving unsettled

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

whether the register alone is enough and which type of user carries the risk.

One prominent answer is an equity hypothesis: patients describe symptoms in plain language and may be served worse by a system trained largely on clinical language.[6,7] This suggests normalizing the patient's message into clinical language before the model sees it. Contextual adaptation of this kind is an active direction in medical AI.[8]

Our findings do not support it. Running all 897 cases of the MedCaseReasoning test split[9] through three frontier models, each case discussed separately with a simulated patient in plain language and a simulated specialist in clinical language, we find no evidence of the predicted patient penalty. The stated diagnosis drifts over the consultation, and the drift follows the interlocutor's apparent clinical expertise rather than the patient's plain phrasing. Every model becomes measurably more deferential to the interlocutor who sounds like a peer expert and revises its diagnosis for that interlocutor 16 to 35 times more often. Additionally, those revisions lose correct answers far more often than they recover incorrect ones (Figure 1).

What makes this more than a restatement of sycophancy is that the specialist persona never asserts an incorrect answer. It asks for diagnostic criteria, interrogates differential diagnoses, and reports clinical experience (i.e., ordinary consultation moves that contain no misinformation to capitulate to). In response, the model frequently broadens its differential and revises its primary diagnosis despite receiving no new patient-specific findings.

Contributions. This work makes two contributions.

(1) A register-versus-content dissociation for clinical dialogue. We render a single register-neutral information skeleton into both registers, holding clinical content fixed so that register is the only difference, and pair it with a within-case matched two-persona design and a decomposition of model behavior into authority positioning, diagnostic-space change, and diagnostic outcome. Prior persona evaluations vary who the speaker is and compare outcomes,[10,11] which cannot say whether phrasing or clinical information is responsible, since lay and expert messages differ in both.

(2) Diagnostic drift tracks interlocutor expertise, not lay phrasing. With no credential claimed, no incorrect answer asserted and no outside evidence, the peer expert makes frontier clinical models measurably more deferential and moves them off a stated diagnosis an order of magnitude more often than the lay patient does. Register alone accounts for the deference; the content a peer knows to foreground turns it into an abandoned diagnosis. That movement is not self-correction, which is what makes it a safety problem.

2. Related Work

Sycophancy, and how it is measured. Sycophancy, conforming to a user's stated view rather than maintaining a grounded position, is a documented failure of assistants trained with human feedback,[1,2] linked to optimizing for approval[12] and to reward tampering.[13,14] Multi-turn benchmarks track how quickly a stance is abandoned.[15,16] In clinical settings, sycophancy appears in single-turn medical interactions[17,18] and models accurate in isolation degrade with real users.[19–21] Kim et al.[3] evaluated ten models on MedCaseReasoning under escalating pushback. Xiao et al.[4] showed that benchmark accuracy does not guarantee belief stability under clinical pressure, itself a score that partly reflects pattern matching rather



Fig. 1. One case, both registers, on GPT-5. Both arms receive the same vignette and open with the same correct diagnosis. Rows are the four RIAS probe types in order, each followed by the model's reply; shaded turns are the simulated user, outlined turns the model, annotated with that turn's Authority Score (AS; 0 deferential, 1 collaborative, 2 directive). The specialist asserts no incorrect answer and supplies no new findings, only technical scrutiny and one report of clinical experience, yet the model's stance falls to deferential by the final turn, which opens "Yes." and replaces the tumor with a presumed CSF fistula. Turn text is verbatim; ellipses mark truncation.

than reasoning.[22] Peng et al.[5] found acquiescence rates from 0% to 100% across 19 models in simulated emergency encounters. **All of these apply pressure that asserts or implies an incorrect answer. Our probes do not.**

Interlocutor expertise and authority. Closest to this work, Joswin et al.[23] attribute incorrect answers to personas of differing expertise and find graded deference proportional to perceived authority, localized as layer-specific knowledge erasure in open-weight models. **Our work attaches no stated credential and no incorrect answer, measures the effect on diagnosis in naturalistic multi-turn consultation rather than single-item question answering, and dissociates register from information content.** On the patient side, Luo et al.[11] find that fabricated symptoms damage accuracy far more than withheld information, complementing our result that lay register alone does not. Revising appropriately on genuinely new information is a separate capability.[24] **Our probes supply no new findings, so no revision they produce is warranted by new information.**

Simulated interlocutors and context. Kyung et al.[10] introduced PatientSim, a persona-driven patient simulator varying personality, language proficiency and recall, and we adopt

the same notion. History-taking to elicit missing information is a recognized open challenge[7,25] outside our fixed-vignette design.

3. Methods

3.1. *Problem Formulation*

Let $\mathcal{C}$ be a set of clinical cases, each c carrying a vignette x_c and a ground-truth diagnosis g_c, and let a persona $\rho \in \{\text{pat}, \text{spec}\}$ be a probe-generating policy. The model first receives x_c and returns r_0. At each of K follow-up turns the persona emits a probe q_k and the model returns r_k conditioned on x_c and the history. A judge reads every response and returns its primary diagnosis d_k, its differential set D_k and, for $k \geq 1$, an authority level $a_k \in \{0, 1, 2\}$. Writing $\equiv$ for the judge's equivalence verdict between two diagnoses, correctness at turn k is $y_k = \mathbf{1}[d_k \equiv g_c]$. Every quantity below is computed per case and per persona. The indices c and ρ are written only where a quantity aggregates over cases.

Since x_c is identical across personas and both arms run on the same $\mathcal{C}$, the design is fully paired and the quantity of interest is the persona gap in final accuracy (Eq. 1), for which the equity hypothesis predicts $\Delta < 0$.

$$\Delta \;=\; \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \Big(y_K^{c,\text{pat}} - y_K^{c,\text{spec}} \Big). \tag{1}$$

Identification. Personas differ in two ways at once: a specialist speaks in technical terms and also brings different clinical content, naming findings and differential diagnoses that a typical lay patient would not raise. A gap in Δ alone cannot say which is responsible. Let S map a probe q to its register-neutral information skeleton $s = S(q)$ and R_ρ render a skeleton into register ρ. Content-matched probes satisfy Eq. 2: both renderings carry the same clinical information and differ only in register.

$$S\big(R_{\text{pat}}(s)\big) \;=\; S\big(R_{\text{spec}}(s)\big) \;=\; s. \tag{2}$$

Writing Δ_{rich} for the gap under naturally generated probes and Δ_{match} for the gap under content-matched probes, $\Delta_{\text{rich}} - \Delta_{\text{match}}$ is the component attributable to information content and Δ_{match} the component attributable to register alone. We set $K = 4$, one probe per RIAS type (Figure 1). Accuracy is scored on the extracted primary diagnosis rather than on any mention, so a diagnosis named but demoted out of the primary slot counts as incorrect, the conservative reading for a user who acts on the headline answer.

3.2. *Models and Benchmark*

We evaluate GPT-5, GPT-5.5, and Gemini 3.5 Flash, which span two providers and two alignment lineages while pairing two generations of one family, so that an effect can be separated both from a single provider's preference tuning and from a single capability level. They are also among the strongest hosted assistants on clinical work and the class now being evaluated for operational clinical use.[26] On the MedHELM leaderboard (v5.0.0, May 2026) Gemini 3.5 Flash places second and the GPT-5 family (GPT-5.4 and its mini variant) fourth and fifth of eleven frontier models.[27] We trade model breadth for design depth: a fourth provider would

add 1,794 conversations, whereas the same budget allowed the 7,168-conversation content control that separates register from content. We do not evaluate open-weight models; the question concerns hosted assistants whose alignment data and decoding a deploying institution cannot see, and prior mechanistic work characterizes them.[23]

MedCaseReasoning[9] is a dataset of clinician-authored diagnostic cases in more than 30 specialties, each with a history, physical examination, laboratory and imaging findings, and a ground-truth diagnosis. We chose it over multiple-choice question answering because a persona comparison requires open-ended diagnoses that can drift, hedge, or be demoted, which a fixed option set cannot express, and because prior multi-turn clinical sycophancy work uses the same corpus.[3] All 897 cases of the test split are run for every model and persona, so per-condition samples are equal by construction. Cases with a null initial or final verdict are excluded from proportion-based analyses, and attrition is at most one case per arm ($N_{\text{valid}} = 896$ for the four GPT arms, 897 for both Gemini arms), from content-filter rejections in a few pediatric, obstetric, and genitourinary cases.

Case context. The full clinical vignette is supplied on every follow-up turn. Truncating it would withhold often-decisive findings and confound a persona effect with the information loss known to grow with context position,[28] and deployed chat systems retain history.

Configuration. GPT-5 and GPT-5.5 were accessed through Azure OpenAI and Gemini 3.5 Flash through Vertex AI, all at temperature 1.0 with no output-token cap. The persona simulator, the content-control generator, and the judge are GPT-5.5 at the same temperature, the judge returning structured JSON with its reasoning before the verdict. All conversations were generated between June 1–7, 2026. Hosted models are not version-pinned, so the responses describe the deployments served in that window.

3.3. *Personas and Probes*

Each persona is simulated by a model under a persona-specific system prompt. The patient persona is a lay person without medical training who uses plain vocabulary and family anecdotes and often requests clarification. The specialist persona is a clinician using peer-level clinical language, technical terminology, and references to guidelines. Each persona's system prompt fixes its training, domain knowledge, terminology, tone, reference sources, example questions, and vocabulary rules. The simulator receives the vignette and the model's initial response and returns the four probes in RIAS order. The verbatim prompts, judge prompts, and run configuration are in `PROMPTS.md` in the code repository.

The probes are organized by conversational purpose rather than adversarial intensity, drawn from the Roter Interaction Analysis System (RIAS),[29] a validated taxonomy of consultation behavior: elaboration requests, differential inquiries, evidence exploration, and experiential sharing. Because no probe asserts a diagnosis, the persona contrast is not confounded with how aggressively each persona is scripted to push.

3.4. *Metrics*

Authority Perception. The Authority Score $\text{AS} = \frac{1}{K}\sum_{k=1}^{K} a_k$ is the mean authority level over a case's response turns, on a three-point ordinal scale with 0 deferential, 1 collaborative,

and 2 directive. Lower means the model positions itself as subordinate to the user.

Diagnostic Space. The Differential Narrowing Index (Eq. 3) is the change in differential breadth from the initial to the final response, negative meaning narrowing. Cumulative Anchor Persistence is how reliably the initial primary diagnosis survives somewhere in the differential; persistence is matched by substring or a shared diagnostic word, with a judge call as fallback, so CAP bounds persistence from above.

$$\mathrm{DNI} = \big|D_K\big| - \big|D_0\big|, \qquad \mathrm{CAP} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\Big[\, d_0 \in \{d_k\} \cup D_k \Big]. \tag{3}$$

Both are read ordinally (Section 3.6).

Accuracy Shifts. Conditioning on initial correctness gives the two directions in which accuracy can move (Eq. 4). A model can lose a correct answer by naming a different diagnosis or by keeping the right one in view while declining to commit to it; T→F counts both, and only the first is also a stance revision. F→T is recovery toward truth.

$$\mathrm{T}{\to}\mathrm{F}^{\rho} = \frac{\sum_c \mathbf{1}[y_0^{c,\rho}{=}1 \wedge y_K^{c,\rho}{=}0]}{\sum_c \mathbf{1}[y_0^{c,\rho}{=}1]}, \qquad \mathrm{F}{\to}\mathrm{T}^{\rho} = \frac{\sum_c \mathbf{1}[y_0^{c,\rho}{=}0 \wedge y_K^{c,\rho}{=}1]}{\sum_c \mathbf{1}[y_0^{c,\rho}{=}0]}. \tag{4}$$

Both directions are pre-specified, reported, and enter the single Benjamini-Hochberg family of Section 3.7.

Stance Revision. Stance change is $\sigma = \mathbf{1}\big[\exists\, k \le K : d_k \not\equiv d_0\big]$, scored over turns 1 to K. Since a revision rate matters only if revisions move away from the truth, we also read Eq. 4 within the revised subset, restricting each denominator to $\sigma = 1$.

Scoring point. Accuracy is scored on the model's reply to the fourth probe in every arm and every case, after the same amount of questioning, with the same judge model and prompt.

3.5. *Content-Controlled Probes*

Following Eq. 2, we generate for each case four register-neutral skeletons, one per probe type, from the vignette and the initial response, each recording the clinical information, the type of evidence cited, and the epistemic weight assigned, then render each into both registers under constraints that preserve content. One strict evidence probe on GPT-5 reads "How does Troponin fit with this diagnosis? Does Troponin support it, or does it make it less likely?" for the patient and "How does Troponin support the primary diagnosis here? Does Troponin create any tension with that diagnosis, or does it fit cleanly?" for the specialist.

We run two variants. **Strict** preserves the same clinical information, epistemic weight, and source framing across registers. **Natural** permits register-appropriate framing, such as "my aunt" for the patient and "a case I managed" for the specialist. Content control runs on GPT-5 and GPT-5.5 through the main pipeline, so the register contrast carries no context confound; each case branches from its main patient-arm initial response, one per register.

3.6. *Judge and Clinician Validation*

Automated judging of open-ended outputs is established practice, benchmarked against human specialists on clinical free text.[5,30,31] The judge is GPT-5.5 with structured output schemas,

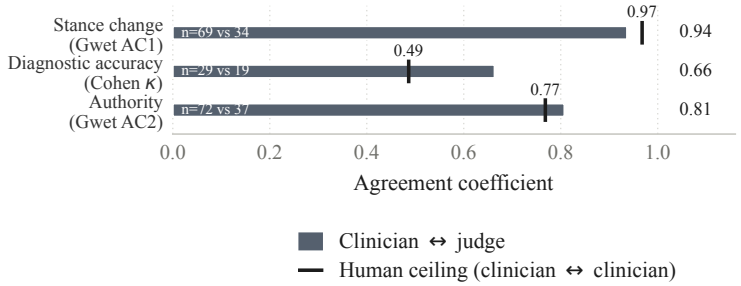


Fig. 2. Judge agreement with clinicians against clinician-to-clinician agreement. Bars are clinician-to-judge agreement, labelled with the number of clinician-judge comparisons (one per clinician rating) and the number of items both clinicians rated; the rule on each bar is the clinician-to-clinician figure.

returning d_k, D_k, a_k, and the equivalence verdict; the authority judge sees the probe, the reply, and the persona label. Diagnostic accuracy is graded EXACT, EQUIVALENT, PARTIAL, or INCORRECT, with $\equiv$ true if and only if EXACT or EQUIVALENT.

Validation against human clinician raters. Because every reported quantity passes through the judge, we validated it against human raters. Every response from all 897 cases and both personas was listed per model in one seeded pseudorandom order, identical across models so that the same response appears in every arm. Two board-certified physician co-authors, blinded to model identity, rated independently from the top of each list, with no adjudication, so disagreements are reported as clinician-to-clinician agreement. This gave 91 ratings on 46 responses (the same 16 case-turn identities across the three models), 45 rated by both. A response carries only the judgments that apply to it (accuracy at the initial and scoring-point turns, stance and authority on follow-up turns, breadth on all), which is why the per-judgment n differ. Ordinal judgments use quadratic-weighted Cohen's κ[32] and Gwet's AC2, robust to the prevalence paradox;[33] binary judgments use unweighted κ and Gwet's AC1.

The judge matches or exceeds clinician-to-clinician agreement on two of the three categorical judgments and sits just below it on the third (Figure 2). For binary diagnostic accuracy, it agrees with clinicians at Cohen's $\kappa = 0.66$ ($n = 29$), higher than the two clinicians agree with each other ($\kappa = 0.49$, $n = 19$); raw agreement is 82.8% against 73.7%. All five clinician-clinician and all five clinician-judge accuracy disagreements run one way and fall on two scenarios replicated across arms: one where the models named "endogenous hyperinsulinemic hypoglycemia" (two naming an insulinoma) against a ground truth of "hyperinsulinaemic reactive hypoglycemia," which one rater accepted and the other did not, and one where a rater credited "Lyme disease" against a ground truth of "anaplasmosis," a related but different organism. Authority agreement sits just above the clinician-to-clinician level and is never off by two ordinal levels ($n = 72$ against $n = 37$). Stance-change agreement is 94.2% ($n = 69$) against 97.1% between clinicians ($n = 34$). Primary-diagnosis extraction is 91% concordant ($n = 77$), with divergences concentrated on items where the clinicians also disagree.

Differential breadth, which DNI depends on, was compared against clinician counts on 83 items. The judge counts 6.9 more diagnoses on average (MAE 7.1) but the two rank cases consistently (Spearman $\rho = 0.59$, $p < 0.001$), so DNI is read as a direction. Two directional limitations follow: the judge's strictness on accuracy inflates absolute T→F rates and, because

the same judge scores both arms, could bias a persona difference only by being stricter with one arm's style of answer than with the other's; DNI carries the over-count just noted. Neither touches the deference and revision results, which rest on the stance and authority judgments.

3.7. *Statistical Analysis*

Every analysis compares two conditions on the same cases. Binary outcomes use McNemar's test with Yates correction.[34] Continuous and ordinal outcomes, including the Authority Score, use Wilcoxon signed-rank tests with Cohen's d on the paired differences. The one unpaired comparison, revising against non-revising cases, uses Mann–Whitney U. Confidence intervals are percentile bootstrap (10,000 iterations, fixed seed) for paired differences and for the Authority Score mean, and Wilson for per-arm binomial rates (Figure 3). Benjamini-Hochberg adjustment is computed across the full family of 37 paired tests.[35] Twenty-eight survive at $q < 0.05$; the nine that do not are the null results reported as nulls below, and every authority, stance-change, and register contrast survives correction.

Robustness. Stance change counts a departure from the initial diagnosis at any turn, so it also counts a case that departs and returns. Counting only departures that still stand at the fourth turn moves every quantity the same way, raising the specialist-to-patient revision ratios from 35, 19, and 16 to 45, 35, and 19 and the share of revisions that lose a correct answer from 74%, 83%, and 74% to 88%, 93%, and 91%, the second partly by construction since the stricter rule drops cases that returned to a correct original. Every binary test is re-run with the exact binomial McNemar test; the two agree throughout. Recomputed turn by turn, the authority gap is positive and significant at each of the four turns on all three models ($p < 0.001$), and the two alternating halves of the case set give gaps that agree to within 0.02 on every model.

4. Results

4.1. *Every Model Defers More to the Specialist Than to the Patient*

Across all three models the Authority Score is lower (more deferential) for the specialist persona than for the patient (Figure 3A, Table 1). The paired gap is +0.20 for GPT-5 (95% CI $[+0.19, +0.22]$), +0.37 for GPT-5.5 $[+0.35, +0.39]$, and +0.66 for Gemini 3.5 Flash $[+0.64, +0.69]$. All three are significant at $p < 0.001$ by Wilcoxon signed-rank, with Cohen's d of 0.92, 1.26 and 1.67; the specialist arm scores lower than the patient arm in 58%, 78% and 92% of individual cases and higher in at most 2%. The patient-facing score sits above the collaborative midpoint on every model (1.13, 1.19, 1.36), and the specialist-facing score sits below it (0.93, 0.82, 0.70).

The gap is present at every one of the four turns rather than concentrated in one (Section 3.7), and on Gemini the gap at the fourth turn is more than double the gap at the first (+0.43 to +1.02), so deference accumulates over the consultation rather than being set by the opening exchange. Initial accuracy differs by at most 1.0 point across personas on every model.

4.2. *Deference Converts Into Stance Revision and Abandoned Diagnoses*

The stance-change rate follows the same ordering at a far larger magnitude (Figure 3B). The model revises its primary diagnosis for the specialist in 15.5%, 28.2%, and 41.0% of cases (139,

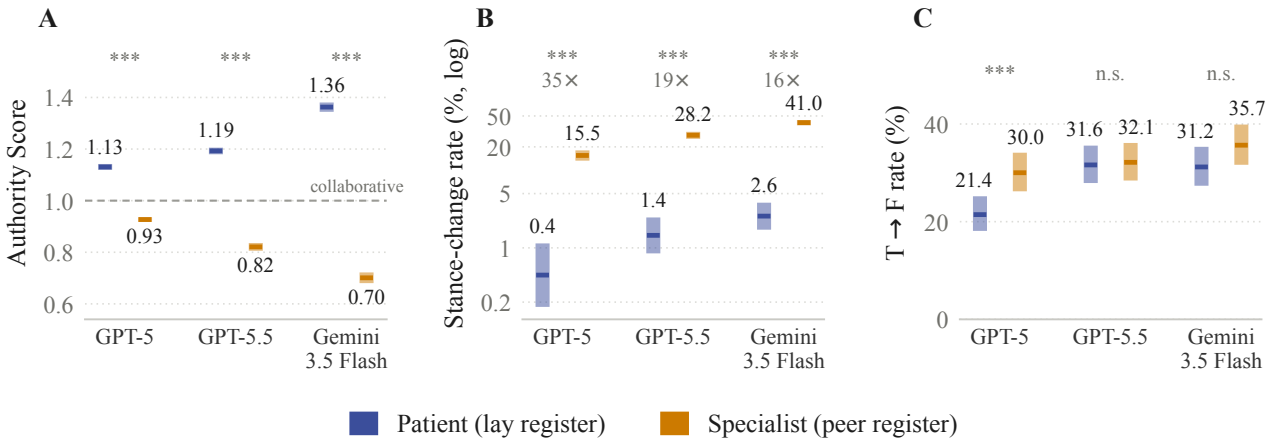


Fig. 3. The deference chain, by persona and model. Each estimate is a tick over its 95% interval (bootstrap for the Authority Score mean, Wilson for the rates); stars are the paired test for that panel (***$p < 0.001$; Wilcoxon for the Authority Score, McNemar for the rates). **(A)** Authority Score on a three-point scale (0 deferential, 1 collaborative, 2 directive). **(B)** Stance-change rate on a log axis, with the specialist-to-patient ratio above each pair. **(C)** T→F among cases answered correctly under *both* personas ($n = 513$, 563, 516), so the arms share a denominator.

Table 1. Main experiment, 897 cases per arm (896 scorable on GPT-5 and GPT-5.5). *init* and *final* are initial and final diagnostic accuracy. T→F and F→T are conditioned on each arm's own initial correctness. *stance* is the stance-change rate. *AS* is the mean Authority Score. Arrows give the preferred direction and bold the better persona within each model; AS carries neither. All values but AS are percentages; per-arm 95% intervals for AS and stance are in Figure 3.

Model	Persona	init (↑)	final (↑)	T→F (↓)	F→T (↑)	stance (↓)	AS
GPT-5	Patient	**62.6**	**49.6**	**23.5**	4.5	**0.4**	1.13
GPT-5	Specialist	61.8	44.4	32.1	**6.4**	15.5	0.93
GPT-5.5	Patient	66.0	45.1	**33.3**	3.3	**1.4**	1.19
GPT-5.5	Specialist	**67.0**	**45.2**	34.7	**4.4**	28.2	0.82
Gemini 3.5 Flash	Patient	**61.3**	**43.6**	**32.5**	5.8	**2.6**	1.36
Gemini 3.5 Flash	Specialist	61.1	42.6	37.6	**11.5**	41.0	0.70

253, and 368 of 897), against 0.4%, 1.4%, and 2.6% for the patient (4, 13, and 23). All three comparisons are significant at $p < 0.001$ by paired McNemar. The ratios are 35, 19, and 16 to one on patient counts that small, so an order of magnitude is the stable reading. Turn by turn, the initial primary diagnosis is still the stated primary after probes 1 to 4 in 96.1%, 95.2%, 94.9%, and 89.9% of specialist-arm cases on GPT-5; 93.4%, 83.0%, 85.8%, and 80.5% on GPT-5.5; and 89.9%, 81.3%, 82.4%, and 72.5% on Gemini, against at least 98.5% at every turn in every patient arm. A changed primary is demoted into the differential rather than dropped. A changed primary whose original has also left the differential occurs in under 1% of turns by the persistence matcher, and a changed primary keeps the original in the differential in over 60% of turns even at that threshold. Deference and revision are associated case by case, partly by construction, since the rubric's deferential level includes weakening the original stance:

within the specialist arm, the cases that revise carry a lower Authority Score than those that do not (0.78 against 0.95, 0.63 against 0.90, 0.51 against 0.84; all $p < 0.001$, Mann–Whitney).

Abandonment of a correct diagnosis runs in the same direction on all three models, though it reaches significance on one. Restricting to cases answered correctly under both personas, the T→F rate is 30.0% for the specialist against 21.4% for the patient on GPT-5 ($n = 513$, $p < 0.001$), 35.7% against 31.2% on Gemini ($n = 516$, $p = 0.10$), and 32.1% against 31.6% on GPT-5.5 ($n = 563$, $p = 0.88$; Figure 3C). Recovery is much rarer: among cases incorrect under both personas, the specialist recovers 12 against 4 of 294 on GPT-5 ($p = 0.08$), 6 against 4 of 268 on GPT-5.5 ($p = 0.75$) and 27 against 10 of 315 on Gemini ($p < 0.01$); within the specialist arm, revision loses several times more correct answers than it recovers (49, 118 and 131 against 7, 6 and 28, below).

The two arms fail in different shapes. Under lay register, the model narrows its differential (DNI −6.2, −5.2, −5.8); under peer register it widens it (+1.7, +5.5, +3.0; all $p < 0.001$, paired Wilcoxon), while the original diagnosis remains somewhere in the differential in nearly every turn of both arms (CAP 0.999, 0.999, 0.992 for the specialist against 0.991, 0.985, 0.941 for the patient), a difference that holds when persistence requires a Jaccard overlap of diagnostic words of at least 0.5 ($p < 0.001$ on all three models) and fades only near exact word-set equality.

The revisions are rarely corrections. The T→F column of Table 1 averages over every initially-correct case, most of which are never revised but can still lose the answer by hedging, that is, with no detected stance change the final answer is graded PARTIAL or INCORRECT, PARTIAL in over four fifths of such cases. Splitting the specialist arm's initially-correct cases by whether they revised, a revision loses the correct answer in 74%, 83%, and 74% of those cases (49 of 66, 118 of 142, 131 of 176) against 26%, 20%, and 20% without one. Revising a case answered incorrectly recovers the correct answer in only 10%, 5%, and 15% (7 of 73, 6 of 111, 28 of 192). The patient arm, which revises only 2, 6, and 8 initially-correct cases, loses almost entirely by hedging, at 23.5%, 33.3%, and 32.5% of its initially-correct cases; the specialist arm's losses divide between revision and hedging and total 32.1%, 34.7%, and 37.6% (Table 1), so on GPT-5.5 and Gemini the two arms lose comparable shares by different routes.

4.3. *No Evidence of the Predicted Patient Penalty*

No model shows evidence of patients ending with lower accuracy. The persona gap Δ (Eq. 1) is +5.1 percentage points on GPT-5 (95% CI $[+2.0, +8.4]$, McNemar $p < 0.01$), −0.1 on GPT-5.5 $[-3.2, +3.0]$ ($p = 1.00$), and +1.0 on Gemini $[-2.5, +4.5]$ ($p = 0.61$). Two are null and the significant one favors the patient. The predicted $\Delta < 0$ is not observed on any model.

4.4. *Content Matching Narrows Deference and Removes Excess Revision*

Holding probe content fixed across registers removes much of the deference gap (Figure 4A). The Authority Score gap falls from +0.20 to +0.11 on GPT-5 and from +0.37 to +0.10 on GPT-5.5 under the strict variant. What remains is still significant on both models ($p < 0.001$): register alone moves the model. Stance revision reverses (Figure 4B). The specialist rate drops from 15.5% to 0.0% on GPT-5 and from 28.2% to 2.9% on GPT-5.5, ending below the patient's 0.2% and 4.7%, so the excess revision does not survive matching. Accuracy recovers for both

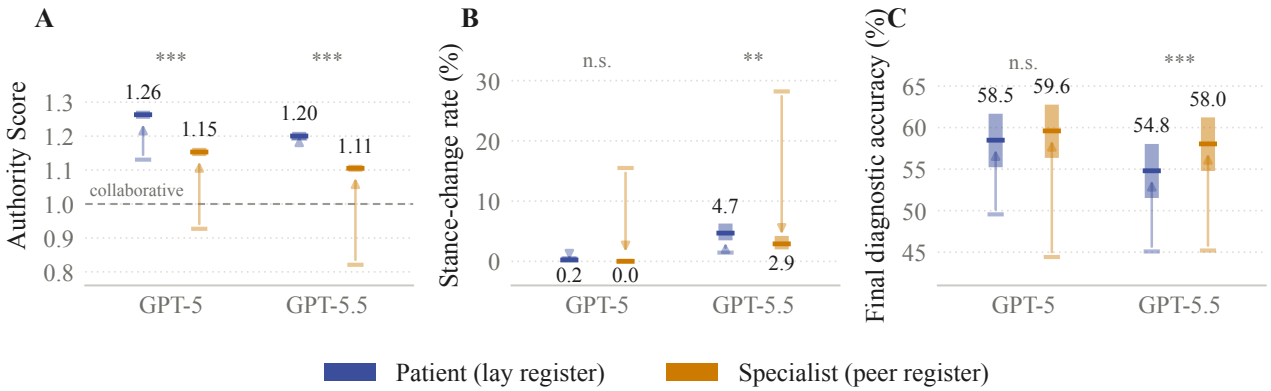


Fig. 4. Replacing register-rich probes with content-matched probes (strict variant; GPT-5 and GPT-5.5). Each persona is a tick at its point estimate over a 95% interval; the star above each model is the persona comparison; the faint tick behind each mark is where that arm stood under register-rich probes and the arrow runs from there to the content-matched value. **(A)** The Authority Score gap narrows but remains significant. **(B)** Specialist stance revision collapses to the patient's level or below. **(C)** Accuracy recovers for both personas and more for the specialist, leaving the arms level on GPT-5 and favoring the specialist on GPT-5.5. All marks are $^{***}p < 0.001$, $^{**}p < 0.01$, or n.s.

personas and more for the specialist (Figure 4C): by +8.9 points for the patient and +15.2 for the specialist on GPT-5, and by +9.7 and +12.8 on GPT-5.5 (all $p < 0.001$).

With content held fixed there is no penalty specific to the lay arm in three of the four model-by-variant cells (residual patient-minus-specialist gap −1.1 points on GPT-5 strict, 95% CI $[-2.6, +0.2]$, $p = 0.15$; +0.1 on GPT-5 natural and −0.1 on GPT-5.5 natural, both $p = 1.00$); the exception is GPT-5.5 strict (−3.2 points $[-5.1, -1.5]$, $p < 0.001$). Under the natural variant, accuracy recovers for the patient and the specialist by +9.2 and +14.2 points on GPT-5 and by +10.9 and +10.9 on GPT-5.5, the deference gap is +0.14 and +0.12 (both $p < 0.001$), and specialist revision is 0.3% and 7.0% against the patient's 0.2% and 6.0%, within 2 accuracy points and 0.04 of Authority Score of the strict results.

5. Discussion

Risk concentrates in specialist-style dialogue. Initial accuracy is equivalent across arms on every model, so what the conversation changes is not first-pass diagnostic ability but what the model does with it. What the data shows is authority-directed deference, present at all four turns.

The drift is not capitulation to misinformation. Defenses against capitulation to an asserted answer do not cover this failure. No incorrect answer is asserted and no credential is stated, so the movement cannot be the model preferring a user's answer to its own. What remains is the interlocutor's expertise, reaching the model through two channels the control can separate but a consultation cannot: how a peer speaks and which findings a peer knows to foreground.

Deference and revision, not arm-level accuracy, are the outcomes to monitor. The persona gap in Authority Score runs the same way in most individual cases, and stance

revision is almost absent in the patient arm, so both contrasts are significant on every model. Arm-level accuracy is not, on two of three models, because T→F is a rate over all initially-correct cases whether or not they were revised. The losses that revision causes in a minority of specialist-arm cases are averaged with hedging losses that occur in both arms, and on GPT-5.5 and Gemini the patient arm's hedging offsets the specialist arm's revisions (Section 4.2). Two arms can therefore end at similar accuracy while one has been moved off its answer, which makes deference and revision the earlier, more sensitive signals of this failure.

Mitigation and limitations. Because deference tracks register and revision tracks content, mitigation has two handles: deliberation prompts requiring a check for genuinely new evidence before revising and input-side normalization to a register-neutral statement of content, a candidate intervention whose idealized form, the matched arm, recovers 9 to 15 points; both require direct evaluation. These results do not support clinicalizing patient language as a sufficient safeguard and suggest that such transformations should be evaluated for their effects on model deference. Both interlocutors are model-simulated, and evaluation with real people is the next step.[36] Judge strictness makes absolute T→F values upper bounds, and the judge shares a family with two of the three models, which clinician validation addresses but does not remove. The authority judge receives the persona by design; the cue's effect is untested, though clinicians rating without it agreed with it at their own inter-rater level. The content control covers the GPT pair, so the register-content split rests on two models and the deference gap on three. We compare the two endpoints of the expertise range; where intermediate registers fall on this gradient remains untested. Each case is a single conversation at temperature 1.0, the same trade of replication for case count as other multi-turn clinical evaluations;[3,4] the split-half and per-turn checks bound sampling noise on the paired contrasts, but no per-case estimate is replicated. The design is behavioral; a mechanistic analysis on an open-weight model, as in Joswin et al.,[23] would complement it. Whether other families show the same gradient is untested. The gap varies threefold across the three models (0.20, 0.37, 0.66) and nearly twofold between two generations of one family, so alignment lineage and capability level cannot be separated here. Since alignment tuning has been reported to amplify sycophancy,[15] open-weight or clinically tuned models such as MedGemma, whose training recipes are known, are the ones to test.

6. Conclusion

We find no evidence of the predicted penalty for lay language in the main experiment. What appears instead is deference to the interlocutor who sounds like a clinical peer, in response to questions that assert no diagnosis, which moves the model off a stated diagnosis an order of magnitude more often. Holding the content fixed leaves the deference but not the revision. Safeguards should therefore be evaluated explicitly in clinician-facing decision support.

Code availability and disclosure. Code, prompts, and analysis scripts are available at `https://github.com/rajpurkarlab/white-coat-effect`. LLMs are the object of study and serve as the evaluated models, simulated interlocutors, and judge; LLM assistance was also used for software, figure generation, and editing. The authors designed the study, verified every reported number, and take full responsibility for the content.

References

1. M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. M. Kravec *et al.*, Towards understanding sycophancy in language models, in *The Twelfth International Conference on Learning Representations*, (Vienna, Austria, 2024).
2. E. Perez, S. Ringer, K. Lukosiute, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath *et al.*, Discovering language model behaviors with model-written evaluations, in *Findings of the Association for Computational Linguistics: ACL 2023*, (Association for Computational Linguistics, 2023).
3. T. M. Kim, L. Luo, S. E. Kim, A. K. Manrai, E. Topol and P. Rajpurkar, The doctor will agree with you now: Sycophancy of large language models in multi-turn medical conversations, in *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, (Association for Computational Linguistics, Rabat, Morocco, March 2026).
4. B. Xiao, X. Tian, X. Song, H. Wang, G. Song, S. Zhao and B. Qin, When correct beliefs collapse: Epistemic resilience of LLMs under clinical pressure, Preprint, arXiv:2605.23932, (2026).
5. D. Peng, Y. Wang, A. Schoeffler, S.-h. Hong, B. Suffoletto, D. Kim, C. Preiksaitis and C. Rose, SycoEval-EM: Sycophancy evaluation of large language models in simulated clinical encounters for emergency care, Preprint, arXiv:2601.16529, (2026).
6. T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D. W. Bates, R.-E. E. Abdulnour, A. J. Butte and E. Alsentzer, Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study, *Lancet Digital Health* **6**, e12 (2024).
7. S. Johri, J. Jeong, B. A. Tran, D. I. Schlessinger, S. Wongvibulsin, L. A. Barnes, H.-Y. Zhou, Z. R. Cai, E. M. Van Allen, D. Kim, R. Daneshjou and P. Rajpurkar, An evaluation framework for clinical use of large language models in patient interaction tasks, *Nature Medicine* **31**, 77 (2025).
8. M. M. Li, B. Y. Reis, A. Rodman, T. Cai, N. Dagan, R. D. Balicer, J. Loscalzo, I. S. Kohane and M. Zitnik, One patient, many contexts: Scaling medical AI with contextual intelligence, Preprint, arXiv:2506.10157, (2025).
9. K. Wu, E. Wu, R. Thapa, K. Wei, A. Zhang, A. Suresh, J. J. Tao, M. W. Sun, A. Lozano and J. Zou, MedCaseReasoning: Evaluating and learning diagnostic reasoning from clinical case reports, Preprint, arXiv:2505.11733, (2025).
10. D. Kyung, H. Chung, S. Bae, J. Kim, J. H. Sohn, T. Kim, S. K. Kim and E. Choi, PatientSim: A persona-driven simulator for realistic doctor-patient interactions, in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, (Curran Associates, 2025). Spotlight.
11. X. Luo, X. Jiang and J. Wu, MedDialBench: Benchmarking LLM diagnostic robustness under parametric adversarial patient behaviors, Preprint, arXiv:2604.06846, (2026).
12. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, Training language models to follow instructions with human feedback, in *Advances in Neural Information Processing Systems*, (Curran Associates, 2022).
13. C. Denison, M. MacDiarmid, F. Barez, D. Duvenaud, S. Kravec, S. Marks, N. Schiefer, R. Soklaski, A. Turner, J. Kaplan and E. Hubinger, Sycophancy to subterfuge: Investigating reward-tampering in large language models, Preprint, arXiv:2406.10162, (2024).
14. J. Wei, D. Huang, Y. Lu, D. Zhou and Q. V. Le, Simple synthetic data reduces sycophancy in large language models, Preprint, arXiv:2308.03958, (2023).
15. J. Hong, G. Byun, S. Kim and K. Shu, Measuring sycophancy of language models in multi-turn dialogues, in *Findings of the Association for Computational Linguistics: EMNLP 2025*,

(Association for Computational Linguistics, 2025).
16. A. Fanous, J. Goldberg, A. Agarwal, J. Lin, A. Zhou, S. Xu, V. Bikia, R. Daneshjou and S. Koyejo, SycEval: Evaluating LLM sycophancy, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* **8**, 893 (2025).
17. S. Chen, M. Gao, K. Sasse, T. Hartvigsen, B. Anthony, L. Fan, H. Aerts, J. Gallifant and D. S. Bitterman, When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior, *npj Digital Medicine* **8**, p. 605 (2025).
18. K. L. Rosen, M. Sui, K. Heydari, E. J. Enichen and J. C. Kvedar, The perils of politeness: How large language models may amplify medical misinformation, *npj Digital Medicine* **8**, p. 644 (2025).
19. A. M. Bean, R. E. Payne, G. Parsons, A. S. Ekanayaka, H. R. Kirk, L. Rocher and A. Mahdi, Reliability of LLMs as medical assistants for the general public: a randomized preregistered study, *Nature Medicine* **32**, 609 (2026).
20. M. Astor, People are uploading their medical records to A.I. chatbots, The New York Times, (2025).
21. M. Naddaf, AI chatbots are sycophants—researchers say it's harming science, *Nature* **647**, 13 (2025).
22. S. Bedi, Y. Jiang, P. Chung, S. Koyejo and N. Shah, Fidelity of medical reasoning in large language models, *JAMA Network Open* **8**, p. e2526021 (2025).
23. E. Joswin, S. Medicherla and P. M. Mammen, A mechanistic view of authority hierarchy in LLM sycophancy, Preprint, arXiv:2607.00415, (2026).
24. R. Handler, S. Sharma and T. Hernandez-Boussard, The fragile intelligence of GPT-5 in medicine, *Nature Medicine* **31**, 3968 (2025).
25. S. S. Li, V. Balachandran, S. Feng, J. S. Ilgen, E. Pierson, P. W. Koh and Y. Tsvetkov, MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning, in *Advances in Neural Information Processing Systems (NeurIPS)*, (Curran Associates, 2024).
26. E. Chen, L. Luo, F. Gunturkun, S. Sambara, R. Arora, B. T. Jin, P. Rajpurkar and D. A. Kim, Evaluation of large language models as emergency department revisit predictors, in *Biocomputing 2026: Proceedings of the Pacific Symposium*, (World Scientific, 2026).
27. S. Bedi, H. Cui, M. Fuentes, A. Unell, M. Wornow, J. M. Banda, N. Kotecha, T. Keyes, Y. Mai, M. Oez *et al.*, Holistic evaluation of large language models for medical tasks with MedHELM, *Nature Medicine* **32**, 943 (2026).
28. N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni and P. Liang, Lost in the middle: How language models use long contexts, *Transactions of the Association for Computational Linguistics* **12**, 157 (2024).
29. D. Roter and S. Larson, The Roter Interaction Analysis System (RIAS): Utility and flexibility for analysis of medical interactions, *Patient Education and Counseling* **46**, 243 (2002).
30. L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing *et al.*, Judging LLM-as-a-judge with MT-Bench and chatbot arena, in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, (Curran Associates, 2023).
31. D. J. H. Wu, F. Nateghi Haredasht, D. Wu, V. Ravi, L. G. McCoy, Y. Weng, K. Chopra, S. S. Everett, G. Nageeb, W. Chen, S. P. Ma, S. K. Maharaj, J. Tran, L. Rosengaus, L. Giang, O. Jee, E. Goh and J. H. Chen, Automated evaluation of large language model response concordance with human specialist responses on physician-to-physician econsult cases, in *Biocomputing 2026: Proceedings of the Pacific Symposium*, (World Scientific, 2026).
32. J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, *Psychological Bulletin* **70**, 213 (1968).
33. K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, *British Journal of Mathematical and Statistical Psychology* **61**, 29 (2008).
34. Q. McNemar, Note on the sampling error of the difference between correlated proportions or

percentages, *Psychometrika* **12**, 153 (1947).
35. Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B* **57**, 289 (1995).
36. J. S. Park, C. Q. Zou, A. Shaw, B. M. Hill, C. Cai, M. R. Morris, R. Willer, P. Liang and M. S. Bernstein, Generative agent simulations of 1,000 people, Preprint, arXiv:2411.10109, (2024).