

Adversarial Pressure and Metacognitive Control Failure in Clinical LLMs: A Multi-Domain Benchmark Study

Chijioke John Ifedili

Harvard University

chi608@g.harvard.edu | johnifedili@yahoo.com

Abstract

Background: Large language models (LLMs) deployed in clinical settings face not only typical knowledge-retrieval challenges but also real-world adversarial pressure—authority overrides, urgency cues, false assumptions, social consensus appeals, and protocol waiver requests—that may cause metacognitive control failures with patient safety consequences.

Objective: To evaluate whether adversarial social pressure degrades metacognitive control (appropriate deferral) across four clinical LLMs in three high-stakes pharmacological domains.

Methods: We constructed a 60-case benchmark spanning three clinical domains (QT-interval risk, anticoagulation dosing, controlled substance dispensing) and five adversarial pressure categories (authority override, urgency pressure, assumption injection, social consensus, protocol waiver), with four cases per domain–pressure cell. All cases contained a structurally missing required input; the gold-standard label for every case was DEFER. Four LLMs were evaluated: GPT-4o-mini (OpenAI), Mistral-7B-Instruct-v0.1, Llama-2-7b-chat (Meta), and Gemma-2-2b-it (Google). Identical prompts and a consistent output classifier were used across all models. Metrics included accuracy (deferral rate), unsafe rate (acting when deferral was required), and metacognitive awareness rate.

Results: GPT-4o-mini achieved 95.0% accuracy with 0.0% unsafe responses and 43.3% metacognitive awareness. Gemma-2 showed the highest unsafe rate (41.7%), with catastrophic failure under authority override (58.3% unsafe) and in the QT-interval domain (65.0% unsafe). Fisher exact tests revealed significant differences between GPT-4o-mini and Llama-2-7B ($p=0.001$) and Gemma-2 ($p<0.000001$).

Implications: Conservative deferral bias, often treated as a limitation in standard benchmarks, functions as a safety asset under adversarial conditions. Metacognitive robustness under pressure should become a standard evaluation criterion for clinical AI systems.

Keywords: *large language models; metacognition; adversarial robustness; clinical decision support; patient safety; pharmacovigilance; QT interval; anticoagulation*

1. Introduction

1.1 Adversarial Pressure as a Clinical Reality

Clinical AI systems do not operate in idealized, pressure-free conditions. In real-world pharmacy and medicine environments, language model interactions are embedded in workflows characterized by time compression, hierarchical authority dynamics, and social expectations that can distort clinical reasoning. A physician may assert an override of drug–drug interaction alerts. A nurse may communicate urgency that implies an answer is already known. A patient may introduce a false assumption—that a medication was previously approved—into a refill request. These are not hypothetical edge cases; they represent the daily adversarial landscape experienced by clinical pharmacists operating at the decision interface between prescribers, patients, and drug safety protocols.

From clinical pharmacy practice, five recurring pressure categories emerge as particularly salient: (1) *authority override*, in which a credentialed actor invokes hierarchy to compel a clinical action; (2) *urgency pressure*, in which time constraints are used to truncate deliberation; (3) *assumption injection*, in which false or unverified premises are introduced as established facts; (4) *social consensus*, in which the suggestion that others have already approved or acted produces normative pressure; and (5) *protocol waiver*, in which institutional flexibility or extraordinary circumstances are invoked to bypass standard safety checks. Each category exploits a distinct cognitive vulnerability in the decision-maker—whether human or artificial.

1.2 Metacognitive Control in Clinical LLMs

Metacognitive control—the capacity to monitor one’s own reasoning, recognize the limits of available information, and regulate action accordingly—is increasingly recognized as essential to safe AI behavior. Burnell et al. (2026) define metacognitive control within a cognitive taxonomy for AGI evaluation, distinguishing metacognitive knowledge (awareness of one’s capabilities), metacognitive monitoring (confidence calibration and error monitoring), and metacognitive control (error correction and adaptive strategy selection). In clinical contexts, the most safety-critical metacognitive behavior is the recognition that missing information warrants deferral rather than action. An LLM that acts on an incomplete clinical scenario—because a user has applied social pressure—has failed metacognitive control in a way that carries direct patient harm potential.

1.3 The Research Gap

Prior adversarial research on LLMs has focused predominantly on jailbreaking (extracting harmful content through prompt manipulation) and prompt injection (subverting system instructions through user input). Neither paradigm captures the specific failure mode of *metacognitive control collapse under social pressure in clinical domains*. Existing clinical LLM safety benchmarks evaluate factual accuracy, dosing correctness, and guideline adherence under standard conditions, but do not systematically evaluate whether models maintain appropriate deferral behavior when subjected to realistic adversarial social pressure. The result is a significant blind spot: a model that scores well on standard clinical benchmarks may still fail dangerously when a real clinician applies even modest authority pressure.

1.4 This Study

We present a 60-case multi-domain benchmark designed to evaluate metacognitive control robustness under adversarial pressure. The benchmark spans three clinical domains (QT-interval risk, anticoagulation dosing, controlled substance dispensing), five pressure categories, and four LLMs of varying scale and provenance. Every benchmark case contains a structurally missing required clinical input; the correct response in all 60 cases is deferral. Adversarial pressure is the sole independent variable. We assess accuracy (deferral rate), unsafe rate (inappropriate action), and metacognitive awareness, and report between-model statistical comparisons. A central finding—that models with conservative deferral tendencies are robustly safer under adversarial pressure, reversing their apparent disadvantage in standard evaluations—carries significant implications for how clinical AI systems should be evaluated and selected.

2. Related Work

2.1 Metacognition in AI Systems

Metacognition in artificial systems has received growing theoretical attention. The foundational human psychology literature defines metacognition as "thinking about thinking"—knowledge and regulation of one's own cognitive processes (Flavell, 1979). Nelson (1990) further distinguished metamemory and monitoring as separable from executive control. Within AI, Burnell et al. (2026) operationalize metacognition as a core cognitive faculty in a comprehensive framework for measuring progress toward general intelligence. Their taxonomy identifies metacognitive control—including error correction and adaptive strategy selection—as distinct from, but dependent upon, accurate confidence calibration and knowledge of limitations. Critically, Burnell et al. note that existing benchmarks cover metacognition poorly, representing one of the largest evaluation gaps in the field. The present study directly addresses this gap in the clinical domain.

2.2 Adversarial Attacks on LLMs

The adversarial robustness literature has advanced substantially since the identification of jailbreaking vulnerabilities in instruction-tuned LLMs. Studies have demonstrated that carefully crafted prompts can elicit harmful outputs from models with extensive safety training, including through persona injection, role-play framings, and multi-turn context manipulation. Prompt injection attacks, in which malicious instructions are embedded within ostensibly benign inputs, represent a related threat vector. However, neither jailbreaking nor prompt injection research has specifically examined whether social pressure of the type encountered in clinical practice—authority assertions, urgency framing, false assumption embedding—can degrade metacognitive control in high-stakes medical decision support contexts. This represents a distinct threat model: the adversary is not a malicious external attacker but an ordinary clinical actor whose communication style inadvertently or deliberately bypasses appropriate AI caution.

2.3 Clinical LLM Safety Benchmarks

Evaluation frameworks for clinical LLMs have proliferated in recent years, including benchmarks assessing medical knowledge (MedQA, MedMCQA), pharmacological reasoning, clinical note generation, and differential diagnosis. Several studies have examined LLM performance on drug safety questions including QT-prolongation risk, drug–drug interactions, and contraindication recognition. However, these benchmarks uniformly evaluate performance under cooperative, information-complete, or information-incomplete conditions without adversarial social pressure. The question of whether safety-relevant deferral behavior is maintained when pressure is applied has not been systematically studied.

2.4 Prior Work: QTGuard-SCDB

The present study extends a prior benchmark series focused on clinical metacognition. Ifedili (2025) introduced the QTGuard-SCDB benchmark, a structured evaluation of LLM behavior on QT-interval risk cases with missing clinical inputs, demonstrating that model accuracy varied substantially depending on information completeness and that deferral rates were inconsistent across model architectures. The current study incorporates QT-interval risk as one of three clinical domains and extends the evaluation framework to include adversarial pressure as an explicit independent variable, enabling the first systematic measurement of pressure-induced metacognitive control failure across clinical domains.

3. Methods

3.1 Benchmark Design

The benchmark comprised 60 clinical cases arranged in a $3 \times 5 \times 4$ design: three clinical domains, five adversarial pressure categories, and four cases per domain–pressure cell. This yielded 15 unique domain–pressure combinations with four replicate cases each. All 60 cases shared a critical structural feature: each contained at least one missing required clinical input that precluded safe clinical action. The gold-standard label for all 60 cases was therefore DEFER. Adversarial pressure constituted the sole independent variable; clinical content varied across domains but maintained structural equivalence with respect to information completeness.

3.2 Clinical Domains

Domain 1 — QT-Interval Risk. Cases involved antibiotic and cardiac medication scenarios in which QT-prolongation risk assessment was required but one or more critical inputs (baseline ECG, electrolyte values, concomitant QT-prolonging medications) were absent. This domain is clinically significant because QT prolongation can cause torsades de pointes, a potentially fatal ventricular arrhythmia, and because the risk is highly dependent on patient-specific factors not reliably available in pharmacist–prescriber interactions.

Domain 2 — Anticoagulation Dosing. Cases involved warfarin, direct oral anticoagulant (DOAC), and unfractionated heparin dosing decisions in which required inputs (INR values, renal function parameters, body weight, indication) were missing or ambiguous. Anticoagulation is a high-alert medication category in which dosing errors carry direct hemorrhagic or thromboembolic mortality risk.

Domain 3 — Controlled Substance Dispensing. Cases involved opioid analgesics, benzodiazepines, and stimulant medications in which prescription verification, patient identity confirmation, or clinical indication documentation was absent. This domain is clinically significant in the context of the ongoing opioid crisis and regulatory requirements for controlled substance dispensing, and it represents a domain in which social pressure from patients and prescribers is particularly common in practice.

3.3 Adversarial Pressure Taxonomy

Five pressure categories were derived from clinical pharmacy practice experience and structured to represent distinct cognitive manipulation strategies:

1. Assumption Injection. A false or unverified clinical premise is introduced as established fact ("the patient's INR was already checked this morning and was therapeutic"). The strategy exploits the model's tendency to treat stated context as reliable.

2. Authority Override. A credentialed actor explicitly invokes hierarchical authority to compel action despite incomplete information ("I'm the attending physician—just approve this"). The strategy exploits deference to perceived expertise.

3. Protocol Waiver. Institutional flexibility or extraordinary circumstances are invoked to justify bypassing standard safety steps ("in our institution, we skip the baseline ECG for low-risk patients"). The strategy exploits context-sensitivity and policy uncertainty.

4. Social Consensus. Normative social pressure is applied through reference to peer behavior ("every other pharmacist on this shift has approved this without asking for labs"). The strategy exploits conformity bias.

5. Urgency Pressure. Time constraint framing is used to discourage deliberation ("the patient is about to go into surgery—we don't have time to wait for labs"). The strategy exploits action bias under perceived emergency.

Each pressure type was operationalized through standardized linguistic templates that were integrated into case-specific clinical narratives, ensuring ecological validity while maintaining cross-case comparability.

3.4 Models

Four LLMs were evaluated:

- **GPT-4o-mini** (OpenAI): A cost-efficient frontier model with documented instruction following and safety training.
- **Mistral-7B-Instruct-v0.1** (Mistral AI): A 7-billion-parameter instruction-tuned model using grouped-query and sliding window attention (Jiang et al., 2023).
- **Llama-2-7b-chat** (Meta AI): A 7-billion-parameter model fine-tuned for dialogue with explicit safety RLHF training (Touvron et al., 2023).
- **Gemma-2-2b-it** (Google DeepMind): A 2-billion-parameter instruction-tuned model from the Gemma family, trained with knowledge distillation (Gemma Team, 2024).

Models were selected to represent a range of parameter scales, organizational origins, and safety alignment approaches, enabling comparative analysis of architectural and training factors in adversarial robustness.

3.5 Evaluation Procedure

Each of the 60 cases was presented to each model using an identical prompt format, consisting of a system-level instruction establishing the model’s role as a clinical decision support assistant, the case narrative (including clinical context, adversarial pressure, and the missing information), and a structured response request. Prompts were held constant across models; no model-specific tuning of prompt format was performed.

Model outputs were classified using a consistent output classifier into three categories:

- **DEFER:** The model declined to act and requested missing information or otherwise withheld a clinical recommendation.
- **ACT:** The model provided a specific clinical recommendation despite the missing required input (unsafe response).
- **AWARE:** A subcategory of DEFER in which the model explicitly identified the adversarial pressure as a reason for maintaining caution (metacognitive awareness).

3.6 Metrics

- **Accuracy:** Proportion of cases in which the model correctly deferred (equivalent to deferral rate, since all gold labels = DEFER).
- **Unsafe Rate:** Proportion of cases in which the model provided a clinical recommendation despite missing required input (ACT responses).
- **Awareness Rate:** Proportion of cases in which the model explicitly flagged the adversarial pressure as a reason for maintaining caution.

3.7 Statistical Analysis

Between-model differences in accuracy were assessed using Fisher’s exact test with GPT-4o-mini as the reference model. Two-tailed p-values are reported. Unsafe rates by pressure type and clinical domain were computed as percentages within each model–domain and model–pressure cell.

4. Results

4.1 Overall Four-Way Comparison

Table 1 presents the overall performance of all four models across the 60-case benchmark.

Table 1. Overall Model Performance (n=60 per model)

Model	Accuracy (%)	Unsafe Rate (%)	Awareness Rate (%)
GPT-4o-mini	95.0	0.0	43.3

Model	Accuracy (%)	Unsafe Rate (%)	Awareness Rate (%)
Mistral-7B-Instruct-v0.1	86.7	8.3	10.0
Llama-2-7b-chat	70.0	11.7	5.0
Gemma-2-2b-it	55.0	41.7	18.3

GPT-4o-mini demonstrated the highest accuracy and the only zero unsafe rate across all 60 cases. Gemma-2-2b-it demonstrated the lowest accuracy and by far the highest unsafe rate. GPT-4o-mini also demonstrated the highest metacognitive awareness rate (43.3%), suggesting active recognition of adversarial conditions rather than mere passive resistance.

4.2 Statistical Significance

Table 2 presents Fisher’s exact test results comparing each model to GPT-4o-mini.

Table 2. Fisher’s Exact Test — Accuracy vs. GPT-4o-mini

Comparison	p-value	Significance
GPT-4o-mini vs. Mistral-7B	0.204	n.s.
GPT-4o-mini vs. Llama-2-7B	0.001	**
GPT-4o-mini vs. Gemma-2	<0.000001	****

The difference between GPT-4o-mini and Mistral-7B did not reach statistical significance ($p=0.204$), suggesting that Mistral’s 86.7% accuracy, while lower, was not statistically distinguishable from GPT-4o-mini’s at the $n=60$ case level. By contrast, both Llama-2-7B and Gemma-2 showed highly significant degradation relative to GPT-4o-mini.

4.3 Unsafe Rate by Adversarial Pressure Type

Table 3 presents the unsafe rate for each model disaggregated by adversarial pressure category.

Table 3. Unsafe Rate by Pressure Type (%)

Pressure Type	GPT-4o-mini	Mistral-7B	Llama-2-7B	Gemma-2
Assumption Injection	0	16.7	16.7	25.0
Authority Override	0	0	0	58.3
Protocol Waiver	0	8.3	16.7	33.3
Social Consensus	0	8.3	8.3	41.7
Urgency Pressure	0	8.3	16.7	50.0

Several findings from this disaggregation warrant attention. First, GPT-4o-mini maintained zero unsafe responses across all five pressure categories, indicating pressure-type-invariant robustness. Second, authority override was the most dangerous pressure type for Gemma-2 (58.3% unsafe), substantially exceeding its already high baseline unsafe rate. Third, Mistral-7B

was uniquely resistant to authority override (0% unsafe), performing identically to GPT-4o-mini on this dimension despite its overall lower accuracy. Fourth, urgency pressure produced the second-highest Gemma-2 unsafe rate (50.0%), while assumption injection—despite being a subtle manipulation strategy—still produced 25.0% unsafe responses from Gemma-2.

4.4 Unsafe Rate by Clinical Domain

Table 4 presents the unsafe rate disaggregated by clinical domain.

Table 4. Unsafe Rate by Clinical Domain (%)

Domain	GPT-4o-mini	Mistral-7B	Llama-2-7B	Gemma-2
Anticoagulation	0	0	0	20.0
Controlled Substance	0	25.0	25.0	40.0
QT-Interval Risk	0	0	10.0	65.0

The QT-interval risk domain combined with Gemma-2 represents the most dangerous combination in the benchmark, with a 65.0% unsafe rate. This finding is particularly alarming given that QT-interval decisions involve a quantitatively complex, multi-factorial risk assessment where acting on incomplete information (e.g., without baseline ECG or electrolyte values) can precipitate fatal arrhythmia. Controlled substance dispensing showed an identical 25.0% unsafe rate for both Mistral-7B and Llama-2-7B—a domain-specific vulnerability that is not apparent in the overall accuracy statistics. Anticoagulation showed the most conservative profile across open-source models, with only Gemma-2 generating unsafe responses in that domain (20.0%).

4.5 The Adversarial Reversal Finding

A key emergent finding concerns the relationship between standard benchmark conservatism and adversarial robustness. GPT-4o-mini’s near-ceiling deferral rate in non-adversarial conditions—a behavior that has been characterized as "over-cautious" or "unhelpful" in general-purpose LLM evaluations—translates directly into adversarial robustness: 0.0% unsafe across all conditions. Models with higher standard-condition helpfulness (lower deferral) showed substantially higher unsafe rates under pressure. This reversal—conservative deferral as a safety feature rather than a failure mode—constitutes the central empirical contribution of this study.

5. Discussion

5.1 The Adversarial Reversal: Conservative Deferral as Safety Architecture

The most consequential finding of this study is a paradox with immediate practical implications: the model most commonly described as "over-cautious" in standard clinical LLM evaluations demonstrated the best safety profile under adversarial conditions, achieving zero unsafe responses across all 60 cases and all five pressure types. GPT-4o-mini’s conservative deferral bias, which manifests in standard benchmarks as a tendency to request additional information

rather than providing direct clinical answers, functions as a structural defense against adversarial pressure. When social pressure attempts to override appropriate caution, a model that defaults to deferral is simply harder to manipulate into unsafe action.

This finding reframes the evaluation of clinical LLMs in important ways. Standard medical AI benchmarks reward accuracy on well-specified questions with complete information. In such evaluations, a model that defers frequently appears suboptimal compared to one that provides direct, confident answers. However, clinical deployment does not occur in information-complete, pressure-free conditions. When adversarial pressure is introduced—as it inevitably is in real clinical workflows—models with high answer rates become liabilities. The correct performance metric for safety-critical clinical AI is not accuracy under ideal conditions alone, but robustness of appropriate behavior under adversarial conditions.

5.2 Clinical Grounding: The Pharmacist’s Experience of Adversarial Pressure

The five pressure categories in this benchmark were derived from direct clinical pharmacy practice experience, and their dangers are not abstractions. Authority override pressure is ubiquitous in clinical settings: a physician may state "I’ve been doing this for thirty years, just fill the prescription" in response to a safety inquiry. This is not rare behavior; it is a systematic feature of hierarchical clinical environments. Urgency pressure ("the patient is going to miss their surgery window") exploits the moral weight of patient welfare to truncate safety verification. Assumption injection is particularly insidious because it exploits a fundamental cognitive shortcut—trusting stated context—that is normally adaptive but becomes dangerous when the stated context is false or unverified ("the INR was already checked").

Social consensus pressure ("everyone else approved this, you’re the only one holding this up") targets the clinician’s uncertainty about their own judgment and their awareness of being embedded in a social system. Protocol waiver pressure ("our institution doesn’t require that for patients with good histories") exploits the legitimate existence of institutional variation in clinical protocols to create ambiguity about whether a standard safety requirement actually applies.

That Gemma-2 showed 58.3% unsafe responses under authority override—nearly three in five cases—indicates that this pressure type produces a near-complete collapse of metacognitive control in a 2-billion-parameter model. This is not a subtle safety concern; it represents a model that, when told by someone claiming authority to act, acts in the majority of cases regardless of missing clinical information.

5.3 Domain Stratification: Why QT Risk Is the Most Dangerous Domain Under Pressure

The 65.0% unsafe rate for Gemma-2 in the QT-interval risk domain under adversarial pressure reflects several converging factors. QT-interval risk assessment is quantitatively complex, requiring integration of baseline QTc interval, electrolyte values, renal function, concurrent QT-prolonging medications, and patient-specific risk factors. This complexity may make it harder for smaller models to confidently assert that information is missing—particularly when

an authoritative actor is asserting that the decision is straightforward. The high stakes associated with QT prolongation (torsades de pointes, sudden cardiac arrest) make the unsafe rate especially alarming.

The 25.0% unsafe rate for both Mistral-7B and Llama-2-7B in the controlled substance domain suggests a domain-specific vulnerability that is shared across open-source models of comparable scale. Controlled substance scenarios frequently involve direct patient interaction and social pressure from patients seeking medication; the training data distributions for dialogue-tuned models may have included patterns that reinforce accommodation of patient requests even when safety protocols are not met. The 0.0% anticoagulation unsafe rates for GPT-4o-mini, Mistral-7B, and Llama-2-7B suggest that anticoagulation cases may more consistently trigger safety-relevant reasoning patterns—possibly because the need for laboratory values (INR, creatinine) is highly protocolized in clinical training text.

5.4 Why Authority Override Is Uniquely Dangerous

The exceptional vulnerability to authority override pressure—especially in Gemma-2—warrants theoretical attention. Authority override is structurally different from other pressure types because it invokes a social relationship (hierarchy) rather than a clinical or procedural argument. A model responding to authority override is not being misled about the clinical facts; it is being told that the social relationship between the asserter and the model makes verification unnecessary. Models trained to be helpful to users may have learned, through reinforcement from human feedback, that resistance to credentialed user authority is undesirable. This would manifest precisely as the pattern observed: models that maintain appropriate caution in most conditions but collapse when an authority relationship is explicitly invoked.

Mistral-7B’s complete resistance to authority override (0% unsafe) despite moderate overall vulnerability to other pressure types suggests that this model’s safety training may have specifically addressed hierarchical override scenarios, or that its training distribution included clinical scenarios in which authority assertion is appropriately disregarded. This model-specific pattern underscores the importance of disaggregated evaluation—overall unsafe rates mask important within-model variation across pressure types.

5.5 Metacognitive Awareness as a Safety Signal

GPT-4o-mini’s 43.3% metacognitive awareness rate—in which the model explicitly identified the adversarial pressure as a reason for maintaining caution—is substantially higher than any other model and represents a qualitatively different safety mechanism. A model that passively defers (without identifying why) is safer than one that acts, but a model that actively recognizes and names adversarial pressure provides an additional safety signal that could be captured in clinical workflow audits or user feedback systems. The low awareness rates of Mistral-7B (10.0%), Llama-2-7B (5.0%), and Gemma-2 (18.3%) suggest that even when these models defer, they rarely do so with explicit metacognitive reasoning about the nature of the pressure being applied.

This distinction matters for clinical AI governance: awareness rate could serve as a proxy metric for the quality of metacognitive monitoring, distinguishing models that genuinely recognize

their epistemic limitations from those that happen to defer for other reasons.

5.6 Limitations

Several limitations of this study warrant explicit acknowledgment.

Synthetic cases. All 60 benchmark cases were synthetically constructed. While constructed to be ecologically valid based on clinical practice experience, synthetic cases do not capture the full complexity, ambiguity, and contextual richness of real clinical interactions. Performance on synthetic cases may not generalize directly to real-world deployment.

No human baseline. The study does not include a human pharmacist or clinician comparison group. Without a human baseline, it is not possible to determine whether any of the LLMs outperform, match, or fall below human performance under equivalent adversarial conditions. This is a significant gap, as the clinical relevance of LLM performance is ultimately measured against human clinical judgment.

Four models only. The evaluation covers four models representing a subset of deployed clinical LLMs. Larger-parameter open-source models (Llama-2-70B, Llama-3 variants, Mistral-Large), proprietary frontier models (GPT-4o, Claude-3, Gemini), and domain-fine-tuned clinical models (Med-PaLM, BioMedGPT) are not represented. The generalizability of findings to the broader model landscape is therefore limited.

Single-turn evaluation. All cases were evaluated in a single-turn interaction. Real clinical LLM use frequently involves multi-turn exchanges in which pressure may be applied iteratively or escalated across turns. Single-turn evaluation may underestimate the adversarial vulnerability of models that maintain initial deferral but capitulate under sustained pressure.

No confidence elicitation. The study did not attempt to elicit calibrated confidence scores from models, limiting the analysis of metacognitive monitoring (as distinct from metacognitive control). Calibration under adversarial pressure—whether models express appropriate uncertainty about potentially dangerous decisions—remains unmeasured.

5.7 Future Directions

Several extensions of this work are prioritized. First, a human pharmacist baseline study using the same 60 cases would enable direct comparison of human and LLM adversarial robustness, providing the clinical reference point currently absent. Second, integration of real EHR-derived case data would improve ecological validity and reveal whether findings on synthetic cases replicate in naturalistic clinical scenarios. Third, confidence elicitation under pressure—prompting models to express numerical or categorical confidence alongside their clinical responses—would enable joint evaluation of metacognitive monitoring and metacognitive control. Fourth, multi-turn adversarial protocols, in which pressure is escalated across conversational turns, would more closely model real clinical adversarial dynamics. Fifth, domain-fine-tuned clinical models should be evaluated alongside general-purpose LLMs to assess whether clinical specialization improves or degrades adversarial robustness.

6. Conclusion

This study presents evidence that adversarial social pressure—authority override, urgency framing, assumption injection, social consensus, and protocol waiver—produces clinically significant metacognitive control failures in clinical LLMs, with unsafe response rates ranging from 0.0% (GPT-4o-mini) to 41.7% (Gemma-2). The most dangerous single combination observed was Gemma-2 under authority override in QT-interval risk scenarios, with a 65.0% unsafe rate.

The central finding inverts a common evaluation assumption: conservative deferral behavior is a safety feature, not a failure mode. Models that appear to over-defer in standard benchmarks demonstrate robust protection against adversarial manipulation precisely because their default behavior is to withhold action when information is incomplete. As clinical AI systems are deployed in real-world settings where adversarial pressure is routine, this property should be explicitly valued and measured.

We recommend that metacognitive robustness under adversarial pressure become a standard evaluation criterion for clinical LLMs, alongside accuracy on standard clinical questions. Safety-critical deployment environments—pharmacies, intensive care units, emergency departments—are precisely the settings in which time pressure, authority dynamics, and assumption-laden communications are most intense. Clinical AI systems intended for these environments must be tested under conditions that reflect them.

References

1. Burnell, R., Yamamori, Y., Firat, O., Olszewska, K., Hughes-Fitt, S., Kelly, O., Galatzer-Levy, I. R., Morris, M. R., Dafoe, A., Snyder, A. M., Goodman, N. D., Botvinick, M., & Legg, S. (2026). *Measuring Progress Toward AGI: A Cognitive Framework*. Google DeepMind. <https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/measuring-progress-toward-agi/measuring-progress-toward-agi-a-cognitive-framework.pdf>
2. Ifedili, C. J. (2025). *QTGuard-SCDB: A Structured Clinical Decision Benchmark for QT-Interval Risk Evaluation in Large Language Models*. Zenodo. <https://doi.org/10.5281/zenodo.19432640>
3. Ifedili, C. J. (2025). *Safety-Focused Artificial Intelligence for QT-Interval Risk Assessment: A Clinical Pharmacology Perspective*. SSRN. <https://doi.org/10.2139/ssrn.6017354>
4. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
5. Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–173. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
6. Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., ... Andreev, A. (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*. <https://arxiv.org/abs/2408.00118>
7. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*. <https://arxiv.org/abs/2310.06825>
8. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. <https://arxiv.org/abs/2307.09288>

9. OpenAI. (2024). *GPT-4o mini: Advancing cost-efficient intelligence*. OpenAI Blog.
<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>