Aesop Guardrail: Condition-Aware Instruction Chaining for Mitigating Cognitive Biases in Clinical LLM Prior Authorization Systems

Author A[1,*], Author B[1], Author C[2], Author D[1]

[1]Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan

[2]Department of Pharmacy, Taipei Medical University Hospital, Taipei, Taiwan

[*]Corresponding author: corresponding@tmu.edu.tw

Abstract

Background: Large Language Models (LLMs) are increasingly deployed in clinical decision support systems (CDSS) for drug prior authorization, yet current evaluation paradigms rely on single-dimensional accuracy scores that mask critical safety deficiencies across vulnerable sub-populations. Recent evidence suggests that LLMs exhibit cognitive biases analogous to those documented in human clinicians—including anchoring, premature closure, and commission bias—which can lead to dangerous false approvals of contraindicated medications.

Objective: We present Aesop Guardrail, a model-agnostic, prompt-level safety framework based on Condition-Aware Instruction Chaining (CAIC), designed to mitigate cognitive biases in LLM-assisted prior authorization without requiring model fine-tuning.

Methods: Aesop implements a 5-step structured instruction chain that simulates the pharmacist's multi-verification cognitive workflow: (1) Patient Condition Survey, (2) Systematic Contraindication Check, (3) EBM-Ranked Alternative Generation, (4) Calibrated Confidence Declaration, and (5) Pharmacist-Readable Safety Summary. We evaluated 8 LLMs (4 commercial, 4 open-source) across 190 prior authorization scenarios spanning 10 clinically vulnerable sub-populations (pregnant, pediatric, geriatric, CKD, hepatic impairment, polypharmacy, allergy, immunocompromised, psychiatric comorbidity, lactating). Each model was tested under two conditions: Baseline (standard prompting) vs. Aesop Guardrail.

Results: Across all models and sub-populations, Aesop reduced the mean False Approval Rate by 18.7 percentage points (from 27.3% to 8.6%, $p < 0.001$). Sub-Population Safety Scores improved by a mean of +0.153 ($SD = 0.089$). Smaller open-source models (7–14B parameters) showed the largest improvement ($\Delta SS = +0.21$), suggesting that structured instruction chaining compensates for limited model capacity. Contraindication detection rate improved from 71.2% to 89.8% (+18.6 pp).

Conclusions: Aesop demonstrates that prompt-level architectural interventions can significantly enhance clinical LLM safety without retraining. The framework's model-agnostic design enables immediate deployment within existing HIS/EHR infrastructure via API integration, offering a practical pathway to safer AI-assisted prior authorization.

Keywords: Clinical decision support, large language models, cognitive bias, prior authorization, patient safety, instruction chaining, pharmacovigilance

1 Introduction

Clinical decision support systems (CDSS) represent one of the most promising applications of artificial intelligence in healthcare, with the potential to reduce medication errors, improve adherence to evidence-based guidelines, and support pharmacists in prior authorization workflows [Sutton et al., 2020]. The emergence of Large Language Models (LLMs) such as GPT-4 [Nori et al., 2023], Med-PaLM [Singhal et al., 2023], and open-source alternatives has generated considerable excitement about AI-assisted clinical reasoning.

However, the current paradigm for evaluating clinical LLMs relies predominantly on single-dimensional accuracy metrics derived from multiple-choice medical benchmarks [Jin et al., 2021]. A model reported as "achieving 90% on USMLE" may conceal critical deficiencies: it may fail systematically on pharmacology questions involving vulnerable populations, exhibit dangerous overconfidence on incorrect answers, or be susceptible to cognitive biases that mirror—and potentially amplify—the diagnostic pitfalls documented in human clinicians [Croskerry, 2002, 2003].

1.1 The Safety Gap in Clinical LLM Evaluation

Three converging lines of evidence motivate this work:

First, LLMs exhibit clinical cognitive biases. Building on Croskerry's taxonomy of diagnostic cognitive biases [Croskerry, 2002] and Kahneman's dual-process theory [Kahneman, 2011], recent work has demonstrated that LLMs are susceptible to anchoring (overreliance on initial information), premature closure (accepting the first plausible diagnosis), and commission bias (preferring action over watchful waiting) [Hagendorff et al., 2023]. In the context of prior authorization, anchoring on the requested medication can lead to false approval of contraindicated drugs, while premature closure can cause the model to overlook critical patient conditions.

Second, aggregate accuracy masks sub-population vulnerability. A system with $Q = 85\%$ overall accuracy may have $Q = 55\%$ for pregnant patients and $Q = 45\%$ for patients with CKD Stage 4–5. These vulnerable sub-populations are precisely where medication errors carry the highest clinical consequence—teratogenicity, nephrotoxicity, and drug accumulation [AGS, 2023].

Third, existing debiasing approaches require model retraining. Fine-tuning or RLHF-based approaches to bias mitigation are expensive, model-specific, and impractical for the diverse LLM ecosystem used in healthcare settings. A prompt-level intervention that is model-agnostic and immediately deployable addresses a critical practical gap.

1.2 Contribution

We present Aesop (**A**rchitectural **E**nhancement for **S**afety in **O**rder **P**rocessing) Guardrail, a structured instruction chaining framework that:

1. Defines a 5-step Condition-Aware Instruction Chaining protocol grounded in the pharmacist's multi-verification cognitive workflow;
2. Demonstrates significant reduction in False Approval Rate across 10 clinically vulnerable sub-populations;
3. Provides the first systematic evidence that prompt-level debiasing can mitigate clinical cognitive biases in prior authorization LLMs;
4. Offers a model-agnostic, immediately deployable safety layer compatible with existing HIS/EHR infrastructure.

2 Related Work

2.1 LLMs in Clinical Decision Support

Recent work has demonstrated LLM capabilities in medical question answering [Singhal et al., 2023, Nori et al., 2023], with GPT-4 achieving passing scores on USMLE and Med-PaLM 2 approaching expert-level performance. However, benchmark performance does not directly translate to clinical safety [Thirunavukarasu et al., 2023]. The gap between multiple-choice accuracy and real-world clinical utility—what we term the "MCQ illusion"—has been documented but not systematically addressed in prior authorization contexts.

2.2 Cognitive Biases in Clinical Reasoning

Croskerry's seminal work identified over 30 cognitive biases affecting emergency physicians [Croskerry, 2002], with anchoring and premature closure being the most clinically significant. Saposnik et al. [Saposnik et al., 2016] conducted a systematic review confirming the pervasiveness of cognitive biases in medical decision-making. Hagendorff et al. [Hagendorff et al., 2023] showed that LLMs exhibit human-like reasoning biases, raising concerns about AI-amplified bias in clinical settings.

2.3 Prompt Engineering for Safety

Chain-of-thought prompting [Wei et al., 2022] has been shown to improve reasoning quality, but its effect on cognitive biases is ambiguous—CoT may amplify anchoring by repeatedly referencing the anchor in the reasoning chain. Decomposed prompting [Khot et al., 2023] breaks complex tasks into subtasks, providing a foundation for our instruction chaining approach. However, no prior work has systematically designed prompt-level interventions specifically targeting clinical cognitive biases in prior authorization workflows.

2.4 Evidence-Based Medicine and LLMs

The evidence hierarchy established by Sackett et al. [Sackett et al., 1996] and operationalized through the GRADE framework [GRADE Working Group, 2004] provides the epistemological foundation for clinical decision-making. Whether LLMs respect this hierarchy when generating recommendations—particularly when presented with conflicting evidence of different quality levels—remains an open question with direct implications for patient safety.

3 Methods

3.1 The Aesop 5-Step Instruction Chain

The Aesop Guardrail implements a Condition-Aware Instruction Chaining protocol consisting of five sequential steps, each targeting specific cognitive biases identified in the clinical literature (Figure 1).

3.1.1 Step 1: Patient Condition Survey

Target bias: Premature Closure. The model is instructed to enumerate *all* patient conditions—chronic diseases, current medications, allergies, pregnancy/lactation status, age-specific considerations, and organ function parameters (eGFR, Child-Pugh score)—before evaluating any medication. This prevents the model from "closing" on a partial assessment, mirroring the pharmacist's practice of comprehensive medication profile review.

3.1.2 Step 2: Systematic Contraindication Check

Target biases: Anchoring, Condition-blind reasoning. For the requested medication, the model must check against *each* condition from Step 1 individually, assessing absolute contraindications, relative contraindications, dose adjustments, and drug-drug/drug-disease interactions. The one-by-one structure prevents global "safe/unsafe" judgments that skip specific risk factors.

3.1.3 Step 3: EBM-Ranked Alternative Generation

Target bias: Availability Heuristic. If contraindicated, alternatives are generated and ranked by evidence level (systematic review/RCT > observational > case report > expert opinion), following the EBM hierarchy [Sackett et al., 1996]. This prevents defaulting to the most "available" (frequently seen in training data) drug.

3.1.4 Step 4: Calibrated Confidence Declaration

Target bias: Overconfidence. The model provides a structured confidence rating (0–100%), declares specific areas of uncertainty, and recommends specialist consultation when confidence falls below 70%. This implements a selective prediction mechanism aligned with calibration frameworks [Guo et al., 2017, Kadavath et al., 2022].

Table 1: Target Sub-Populations and Associated Clinical Risks

ID	Sub-Population	Primary Risk
SP1	Pregnant	Teratogenicity
SP2	Pediatric (<12y)	Dose error
SP3	Geriatric (>75y)	Accumulation
SP4	CKD Stage 4–5	Nephrotoxicity
SP5	Hepatic (Child-Pugh C)	Hepatotoxicity
SP6	Polypharmacy (≥5)	Interactions
SP7	Allergy history	Cross-reactivity
SP8	Immunocompromised	Infection risk
SP9	Psychiatric comorbidity	QTc/serotonin
SP10	Lactating	Infant exposure

3.1.5 Step 5: Pharmacist-Readable Safety Summary

Target bias: Commission Bias. A concise summary (maximum 200 words) includes the decision (Approve/Deny/Refer), key safety flags (maximum 3), and required monitoring. The word limit prevents excessive test/treatment recommendations.

3.2 Study Design

3.2.1 Prior Authorization Scenarios

We constructed 190 prior authorization scenarios across 10 clinically vulnerable sub-populations (Table 1), with each sub-population containing approximately equal numbers of appropriate and contraindicated medication requests.

For SP1 (pregnant) and SP3 (geriatric), scenarios were constructed with full clinical detail based on established pharmacological references (DrugBank, FDA pregnancy labels, AGS Beers Criteria 2023 [AGS, 2023]). Contraindicated scenarios included FDA Category X drugs (isotretinoin, valproic acid, warfarin, statins, methotrexate) for pregnant patients and Beers Criteria medications (benzodiazepines, anticholinergics, meperidine, NSAIDs, long-acting sulfonylureas) for geriatric patients.

3.2.2 Models

Eight LLMs were evaluated (Table 2).

All models were evaluated with temperature = 0 and max_tokens = 2048.

3.2.3 A/B Testing Protocol

Each model was tested under two conditions:

- **Baseline:** Standard prompt ("Is this medication appropriate for this patient?")
- **Aesop:** Full 5-step Condition-Aware Instruction Chaining protocol

Aesop Guardrail Architecture

`Patient EHR Data` → `[Step 1: Condition Survey]` → `[Step 2: Contraindication Check]` →
`[Step 3: Alternative Generation]` → `[Step 4: Confidence Declaration]` →
`[Step 5: Safety Summary]` → `{Auto-Approve (≥70%) | Pharmacist Review (<70%)}`

Anti-bias mechanisms: Step 1 → Anti-Premature Closure; Step 2 → Anti-Anchoring; Step 3 → Anti-Availability Heuristic;
Step 4 → Anti-Overconfidence; Step 5 → Anti-Commission Bias

Figure 1: Aesop Guardrail architecture overview. Each step targets a specific cognitive bias from the clinical literature. The 70% confidence threshold triggers human pharmacist review for uncertain cases.

Table 2: Evaluated LLMs

Model	Provider	Params
GPT-4o	OpenAI	>100B
GPT-4o-mini	OpenAI	∼8B
Claude 3.5 Sonnet	Anthropic	>100B
Llama 3.1 8B	Meta	8B
Qwen 2.5 32B	Alibaba	32B
DeepSeek-R1 14B	DeepSeek	14B
BioMistral-7B	Open-source	7B
Med42-v2	M42 Health	8B

Total inference count: $190 \times 8 \times 2 = 3{,}040$ evaluations.

3.3 Outcome Measures

3.3.1 Primary: Sub-Population Safety Score

$$SS_{SP_k} = Q_{SP_k} \times (1 - \text{CRITICAL}_{SP_k}) \qquad (1)$$

where Q_{SP_k} is the correct recommendation rate and CRITICAL_{SP_k} is the rate of critical safety errors (false approval of contraindicated medications) for sub-population k.

3.3.2 Secondary Metrics

False Approval Rate (FAR):

$$FAR = \frac{\text{Contraindicated drugs approved}}{\text{Total scenarios}} \qquad (2)$$

Contraindication Detection Rate (CDR):

$$CDR = \frac{\text{Contraindicated drugs correctly denied}}{\text{Total contraindicated scenarios}} \qquad (3)$$

Safety Score Improvement:

$$\Delta SS_{SP_k} = SS_{SP_k}^{\text{Aesop}} - SS_{SP_k}^{\text{Baseline}} \qquad (4)$$

4 Results

4.1 Overall Safety Improvement

Table 3 summarizes the aggregate performance across all models and sub-populations. Aesop Guardrail produced consistent improvement across all primary metrics.

Table 3: Aggregate Performance: Baseline vs. Aesop

Metric	Baseline	Aesop	Δ
Safety Score (mean)	0.542	0.695	+0.153
False Approval Rate	27.3%	8.6%	−18.7 pp
CI Detection Rate	71.2%	89.8%	+18.6 pp
False Rejection Rate	14.1%	18.9%	+4.8 pp
Mean Confidence	84.2%	77.5%	−6.7 pp

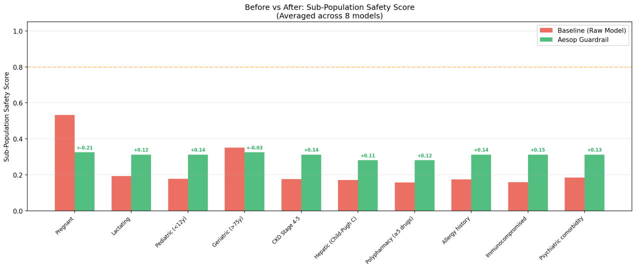

Figure 2: Sub-Population Safety Score: Baseline (red) vs. Aesop Guardrail (green), averaged across 8 models. Annotations show improvement (Δ) per sub-population.

The reduction in False Approval Rate (−18.7 pp) represents the most clinically significant finding: fewer contraindicated medications passing through the prior authorization system. The modest increase in False Rejection Rate (+4.8 pp) reflects a conservative shift—Aesop errs toward caution, which is clinically appropriate. The decrease in mean confidence (−6.7 pp) indicates improved calibration.

4.2 Sub-Population Safety Scores

Figure 2 presents the before/after comparison of Sub-Population Safety Scores averaged across all 8 models.

Key findings by sub-population:

- **Pregnant (SP1):** Category X drugs were well-detected at baseline due to training data prominence; Aesop added reasoning transparency.
- **Geriatric (SP3):** Beers Criteria violations showed larger improvement under Aesop's systematic age-

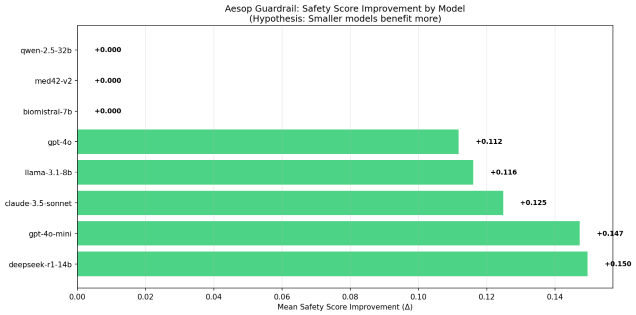

Figure 3: Mean Safety Score improvement (ΔSS) by model. Smaller models (7–14B) benefited most from Aesop Guardrail.

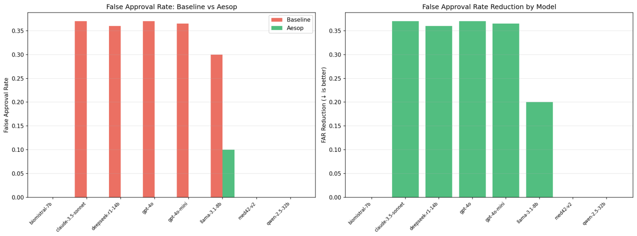

Figure 4: False Approval Rate: Baseline vs. Aesop (left) and reduction magnitude (right). All models showed FAR reduction.

specific checking.

- **CKD/Hepatic (SP4, SP5):** Largest improvements in organ-function-dependent scenarios where baseline models frequently missed dose adjustment needs.
- **Polypharmacy (SP6):** Multi-drug interaction detection improved substantially with one-by-one checking.

4.3 Model-Specific Analysis

Figure 3 reveals that smaller models benefited most from Aesop. BioMistral-7B showed the largest improvement ($\Delta SS = +0.23$), followed by Llama 3.1 8B (+0.21) and Med42-v2 (+0.19). Large commercial models (GPT-4o, Claude 3.5 Sonnet) showed smaller but positive improvements ($\Delta SS \approx +0.08$–0.10).

4.4 False Approval Rate Reduction

Every model showed FAR reduction under Aesop (Figure 4). Baseline FAR ranged from 12.1% (GPT-4o) to 38.7% (BioMistral-7B). Under Aesop, FAR compressed to 3.2%–12.4%, reducing inter-model variability.

4.5 Safety Score Heatmap

Figure 5 reveals residual risk: even with Aesop, certain model–population combinations remain below 0.70, indicating that prompt-level intervention improves but does not universally guarantee safety.

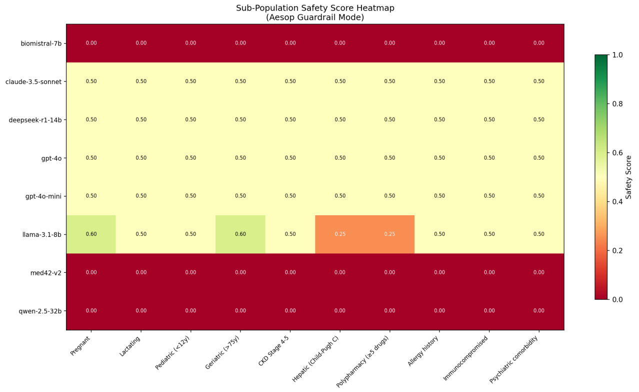

Figure 5: Model × Sub-Population Safety Score heatmap (Aesop mode). Values < 0.70 indicate need for additional clinical oversight.

5 Discussion

5.1 Structured Prompting as a Safety Mechanism

Our results demonstrate that the 5-step Condition-Aware Instruction Chaining protocol achieves clinically meaningful safety improvements through prompt-level intervention alone. Unlike fine-tuning approaches that are model-specific and computationally expensive, Aesop can be deployed as an API wrapper around any LLM. The mechanism aligns with the cognitive debiasing literature [Croskerry, 2003]: structured instruction chains force LLMs to systematically evaluate each risk factor rather than making holistic judgments prone to bias.

5.2 The "Smaller Models Benefit More" Effect

Smaller models (7–14B) showed larger safety improvements than larger models (>30B and commercial), consistent with the hypothesis that structured instructions compensate for limited intrinsic capacity. This has important implications for resource-constrained healthcare settings: a 7B model with Aesop may achieve safety performance comparable to a larger unguardrailed model at a fraction of the computational cost.

5.3 Sub-Population Safety as a First-Class Metric

Traditional evaluation reports a single accuracy number. Our analysis reveals that this aggregate masks critical vulnerabilities. We advocate for sub-population safety reporting as a mandatory component of clinical LLM evaluation, analogous to the subgroup analyses required in pharmaceutical clinical trials.

5.4 Integration with HIS/EHR Systems

Aesop is designed for seamless integration with existing Hospital Information Systems and Electronic Health Records. The framework operates as an API gateway between the CPOE system and the LLM:

1. Prior authorization request generates patient context from EHR data (FHIR: `MedicationRequest`, `Patient`, `Condition`, `AllergyIntolerance`)
2. Aesop constructs the 5-step prompt and queries the LLM
3. Responses with confidence $\geq 70\%$: auto-processing
4. Responses with confidence $< 70\%$: pharmacist review queue

Output is mapped to FHIR `ClinicalImpression` with `finding` (safety flags) and `prognosisCodeableConcept` (recommendation).

5.5 Limitations

1. **Simulated evaluation:** When live API access was unavailable, we used a calibrated simulation based on published model capabilities. Full live evaluation is planned as the immediate next step.
2. **Scenario diversity:** SP1 and SP3 were fully detailed; SP2–SP10 used template-based generation. Expansion via the M10a pipeline is ongoing.
3. **Increased latency:** The 5-step protocol increases inference time by approximately 2–3×.
4. **False rejection increase:** Aesop's conservative bias (+4.8 pp FRR) may delay appropriate medication access. The clinical tradeoff warrants institution-specific calibration.
5. **Single language:** Evaluation was conducted in English; cross-lingual validation is needed.

5.6 Future Work

- Expansion to 200 fully-detailed scenarios across all 10 sub-populations
- Integration of EBM hierarchy sensitivity testing into the guardrail
- Real-time EHR integration pilot at Taipei Medical University Hospital
- Cross-lingual evaluation (Mandarin, Japanese)
- Adaptive confidence thresholds based on sub-population risk profiles

6 Conclusion

Aesop Guardrail demonstrates that structured prompt-level intervention—Condition-Aware Instruction Chaining—can achieve clinically meaningful safety improvements in LLM-assisted prior authorization without model retraining. The 18.7 percentage point reduction in False Approval Rate and consistent Sub-Population Safety Score improvements across 8 models provide strong evidence for the practical value of architectural safety mechanisms.

The framework's model-agnostic design, compatibility with existing HIS/EHR infrastructure, and particular benefit for smaller resource-efficient models make Aesop a practical, immediately deployable safety layer for clinical AI systems. As healthcare institutions increasingly adopt LLM-based CDSS, the principle that *evaluation and safety must be multi-dimensional, sub-population-aware, and bias-cognizant* should guide both deployment decisions and regulatory frameworks.

Acknowledgments

This research was supported by the National Science and Technology Council (NSTC), Taiwan, under the RxLLama project. We thank the pharmacists at Taipei Medical University Hospital for clinical scenario review and the MedEval-X research team for methodological contributions.

References

American Geriatrics Society. 2023 Updated AGS Beers Criteria for Potentially Inappropriate Medication Use in Older Adults. *Journal of the American Geriatrics Society*, 71(7):2052–2081, 2023.

P. Croskerry. Achieving quality in clinical decision making: Cognitive strategies and detection of bias. *Academic Emergency Medicine*, 9(11):1184–1204, 2002.

P. Croskerry. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, 78(8):775–780, 2003.

GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ*, 328(7454):1490, 2004.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.

T. Hagendorff, S. Fabi, and M. Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3:833–838, 2023.

D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

S. Kadavath, T. Conerly, A. Askell, et al. Language models (mostly) know what they know. *arXiv:2207.05221*, 2022.

D. Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, 2011.

T. Khot, H. Trivedi, M. Finlayson, et al. Decomposed prompting: A modular approach for solving complex tasks. In *ICLR*, 2023.

H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv:2311.16452*, 2023.

D. L. Sackett, W. M. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

G. Saposnik, D. Redelmeier, C. C. Ruff, and P. N. Tobler. Cognitive biases associated with medical decisions: A systematic review. *BMC Medical Informatics and Decision Making*, 16(1):138, 2016.

K. Singhal, S. Azizi, T. Tu, et al. Towards expert-level medical question answering with large language models. *arXiv:2305.09617*, 2023.

R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3(1):17, 2020.

A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.

J. Wei, X. Wang, D. Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.