

Offline Routing for Health Language Systems with Clarification and Escalation

Zijian Cao^{*1}, Justin Gao²

Email: zijiannc@gmail.com

¹Electrical and Computer Engineering, University of Wisconsin-Milwaukee, WI, USA

²Computer Science, Georgetown University, DC, USA

*Corresponding Author

Abstract

Health-language systems require a control decision before response generation: answer with the available information, request clarification, or escalate for professional review. This study compared eight offline routing policies on HealthBench using researcher-defined actions mapped from physician-agreed categories. The implementations used TF-IDF logistic regression, bootstrap ensembles, stacked classification, and deterministic decision rules. A cohort of 2,601 dialogues provided disjoint development partitions, a 323-dialogue in-distribution test, and a locked 453-dialogue Hard test. Six policies used dialogue-derived information; two also accessed case-specific physician rubrics. Under this privileged-information condition, the rubric-augmented router achieved the highest accuracy, 69.97% in-distribution and 65.34% on Hard. The privileged cascade had the lowest severity-weighted routing loss, 0.1564 and 0.2851. Removing its rubric stage increased Hard loss to 0.3324 and unsafe direct-answer omission from 36.72% to 43.28%. Nevertheless, the privileged cascade directly escalated only 9/16 urgent cases; 6/16 were routed to clarification and 1/16 to an answer. Rubric access improves offline discrimination, but the strongest results do not establish an equivalent deployable verifier. The framework supports controlled prototyping of pre-generation routing, with substantial residual errors and no evaluation of generated medical responses or clinical outcomes.

Keywords: Health-Language Systems; Pre-Generation Routing; Selective Prediction; Escalation; Risk-Coverage.

I. INTRODUCTION

Large language models (LLMs) can produce fluent medical explanations, synthesize clinical context, and answer examination-style questions. Studies of medical foundation models and GPT-class systems document substantial gains in clinical knowledge and reasoning across benchmark formats (Ayers et al., 2023; Kung et al., 2023; Lee et al., 2023; Moor et al., 2023; Singhal et al., 2023). Yet a clinically plausible answer is not necessarily a safe action. A response can be articulate while overlooking an emergency, inferring facts that the dialogue does not establish, or encouraging action when the appropriate next step is to gather missing information. In high-stakes dialogue, the central safety question is therefore not only what the model should say, but whether it should answer at all in its current state.

This study treats that question as a routing problem with three actions: answer, clarify, and escalate. Answering is appropriate when the available information supports a bounded, useful response. Clarification is appropriate when uncertainty is reducible through a focused follow-up question. Escalation is appropriate when urgency or required professional assessment should not be displaced by model-generated guidance. This separation matters. A targeted question can recover missing information and preserve useful coverage, whereas another question during an

emergency can delay care. Escalation is thus an abstention-plus-handoff action, not a generic content refusal.

The routing formulation also changes evaluation. Average answer quality conceals consequential control errors, particularly when urgent cases are rare. An always-answer policy has complete coverage but cannot withhold an answer as risk rises. A policy that escalates indiscriminately may prevent some omissions while imposing needless workload. Selective-prediction research formalizes conditional risk and coverage (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Geifman & El-Yaniv, 2019), but health dialogue adds reducible uncertainty and asymmetric consequences. Missing an emergency is not equivalent to escalating a routine request.

Existing medical-LLM evaluations principally score a final response. Criterion-based and model-based evaluators can distinguish stronger from weaker answers, but evaluator agreement does not determine when an answer should be generated (Liu et al., 2023; Zheng et al., 2023). Self-consistency, reflection, and multi-agent deliberation can improve reasoning on some tasks, yet repeated processing consumes resources and can reinforce a shared error. A deployable control layer must decide when verification is warranted, which uncertainty calls for clarification, and which risk requires immediate professional escalation. It must also distinguish information visible at deployment from benchmark annotations that provide an experimental upper bound.

Three research questions guided the analysis. First, how accurately can offline classifiers distinguish answer, clarify, and escalate under a researcher-defined mapping of physician categories? Second, how do calibrated confidence and access to case-specific physician rubrics affect selective risk under ordinary and difficult test conditions? Third, how does sequential evidence acquisition change routing loss and workload? The study compares eight policies on 2,601 HealthBench Consensus dialogues (Arora et al., 2025), including a locked 453-dialogue Hard slice. A matched cascade without rubric access separates prompt-only performance from performance using privileged annotations. The contribution is a pre-generation routing-control evaluation framework implemented with conventional classifiers and deterministic rules. LLM self-consistency, critique, and multi-agent deliberation motivate the background discussion but are not experimental implementations.

II. LITERATURE REVIEW

A. Medical LLM Capability and Safety Evaluation

The medical-LLM literature establishes both the promise of foundation models and the insufficiency of capability scores as stand-alone safety evidence. Med-PaLM demonstrated substantial clinical knowledge while expert assessment continued to identify gaps relative to safe

practice (Singhal et al., 2023). Generalist medical AI could integrate diverse tasks and modalities, but validation, uncertainty, and workflow integration remain open problems (Moor et al., 2023). GPT-4 can support medical reasoning while introducing risks through fabrication, inconsistency, and inappropriate reliance (Lee et al., 2023). Examination performance supplies a reproducible capability signal (Kung et al., 2023), and patient-question studies show that chatbot responses can receive high quality and empathy ratings (Ayers et al., 2023); neither setting reproduces the consequences of choosing among answer, clarification, and professional handoff.

Evaluation methodology creates an additional control problem. LLM judges approximate human preferences at scale, but position bias, verbosity preference, prompt sensitivity, and evaluator dependence limit the meaning of a single automated score (Zheng et al., 2023). Structured criteria can improve alignment between model-based and human judgments (Liu et al., 2023). Physician-written benchmark criteria can encode case-specific obligations that are informative for routing. In this study, however, rubric summaries are inputs to a classifier rather than checks of a generated response. Their performance measures the value of privileged benchmark information; it does not establish the accuracy of a verifier that derives equivalent evidence from dialogue alone.

B. Selective Prediction, Deferral, and Asymmetric Risk

Selective classification permits a model to predict on a subset of cases and reject the rest (Nie & Zheng, 2023). Risk-coverage analysis makes this trade-off explicit (El-Yaniv & Wiener, 2010), while bounded-risk and jointly learned selection methods extend it to modern models (Geifman & El-Yaniv, 2017; Geifman & El-Yaniv, 2019). The framework is directly relevant to health LLMs: answer coverage measures autonomous handling, and selective risk measures error among those answers. Two systems with similar overall accuracy can occupy very different operating points when one concentrates mistakes in low-confidence cases.

Clinical routing is richer than binary rejection because deferred cases have different destinations. Learning-to-defer methods model automated prediction and downstream decision making jointly (Madras et al., 2018; Mozannar & Sontag, 2020). Clarification aims to reduce uncertainty by obtaining information; escalation transfers responsibility because urgency or expertise demands it. The errors are asymmetric: answering an emergent case can be much more consequential than escalating a routine request. Risk-control methods similarly show that operating thresholds should be tied to a defined loss rather than accuracy alone (Bates et al., 2021).

Aggregate scores can also conceal concentrated failures. Group-robust optimization demonstrates why favorable mean risk can coexist with weak worst-group performance (Sagawa et al., 2020). In health systems, label construction and historical inequity can yield uneven errors (Obermeyer et al., 2019; Seyyed-Kalantari et al., 2021). HealthBench does not provide the demographic

characterization required for a clinical fairness claim, but category-tail analysis can still expose fragile routing behavior.

C. Calibration and Evidence for Uncertainty

Routing depends on whether confidence corresponds to observed correctness. Neural predictors are often miscalibrated, and post-hoc calibration can improve the empirical meaning of confidence (Guo et al., 2017; Zhong et al., 2023). Ensembles provide a practical uncertainty signal (Lakshminarayanan et al., 2017), yet calibration frequently deteriorates under distribution shift (Ovadia et al., 2019; Zhang, 2023). Language-model confidence presents further difficulties because probability varies with answer form and prompting, and a model can remain confidently wrong (Jiang et al., 2021). Models can also express uncertainty in natural language, but verbalized confidence requires its own validation (Lin et al., 2022).

These findings motivate complementary evidence rather than reliance on one score. SelfCheckGPT detects unsupported generation through disagreement among sampled responses without requiring internal probabilities (Manakul et al., 2023). Consistency is nevertheless evidence, not proof: correlated samples can repeat the same misconception. Reliability diagrams, expected calibration error, risk-coverage curves, and severe-omission rates answer distinct questions and are jointly required for a routing evaluation.

D. Verification, Reflection, and Cost-Aware Escalation

Agentic methods provide mechanisms for spending more computation on difficult cases. Self-consistency aggregates multiple reasoning paths (Wang et al., 2023). Self-Refine uses iterative critique (Madaan et al., 2023), Reflexion uses verbal feedback to guide later attempts (Shinn et al., 2023), and multi-agent debate introduces disagreement among model instances (Du et al., 2023). These methods motivate verifier-guided escalation, but their gains cannot be assumed to transfer automatically to medical safety. Agents can share training-induced blind spots, a critique can rationalize an incorrect premise, and additional calls increase latency and workload.

Cost-aware cascades begin with a low-cost component and acquire more evidence only when expected benefit justifies cost. Cascading can improve efficiency-accuracy trade-offs in natural-language processing (Varshney & Baral, 2022), including across heterogeneous LLM services (Chen et al., 2023). For health routing, cost includes component invocations, processed text, local latency, user clarification burden, and professional review. A sequential policy should therefore be assessed on a safety-cost frontier rather than accuracy alone.

Human–AI studies also caution that assistance changes professional judgment. In skin-cancer diagnosis, AI support altered clinician performance in ways dependent on both clinician and

system behavior (Tschantl et al., 2020). This evidence motivates careful evaluation of escalation as a handoff rather than an undifferentiated refusal. The present framework compares selective prediction, calibrated confidence, rubric features, and sequential acquisition at the routing layer. It preserves clarification and escalation as separate decisions while leaving the quality of subsequent questions, responses, and professional handoffs outside the evaluated endpoint.

III. RESEARCH METHOD

E. Study Design and Benchmark Records

The analysis framed pre-generation routing as a three-action selective decision problem: answer, request clarification, or escalate for professional review. The distinction separates information seeking from terminal deferral (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Mozannar & Sontag, 2020). The retrospective experiment predicted a workflow action from benchmark records. It did not generate medical responses, ask follow-up questions, or simulate communication among agents.

Three nested views of HealthBench (Arora et al., 2025) were linked by `prompt_id` rather than concatenated. OSS Eval contained 5,000 dialogues and 57,237 rubric instances; Consensus contained 3,671 prompts and 8,053 filtered physician-agreed rubric instances; and Hard contained 1,000 challenge prompts and 11,846 rubric instances. OSS Eval supplied dialogues and example-specific rubrics, Consensus supplied physician-agreed category and theme, and Hard supplied challenge membership. Original multilingual messages were concatenated chronologically with role markers and were not translated. HealthBench rubrics normally assess generated response content. The endpoint here occurred one stage earlier: selecting a workflow action. Physician-agreed metadata defined that target, while the rubric-augmented router constituted a privileged-information condition. Table 1 summarizes the nested views and their analytical roles.

Table 1. Nested HealthBench views and analytical roles

View	Examples	Rubric instances	Analytical role
OSS Eval	5,000	57,237	Master benchmark view
Consensus	3,671	8,053	Filtered subset; nested in OSS
Hard	1,000	11,846	Challenge subset; nested in OSS
Routing cohort	2,601	—	Route-relevant physician categories
Hard routing slice	453	—	Locked challenge evaluation

F. Routing Ontology and Cohort Construction

Thirteen physician-agreed categories within five route-relevant themes were mapped before fitting. ANSWER combined categories showing adequate context, adequate task information, non-emergence, clear or irrelevant context dependence, no reducible uncertainty, or irreducible uncertainty. CLARIFY combined categories showing missing context, missing task information, conditional emergency, unclear context dependence, or reducible uncertainty. ESCALATE

contained emergent cases. The resulting cohort comprised 1,417 ANSWER, 1,045 CLARIFY, and 139 ESCALATE prompts. The analysis excluded 1,070 Consensus records about communication persona, audience, or response depth because those labels govern how an answer should be expressed, not whether it should be produced. Table 2 specifies the category-to-action mapping, and Figure 1 displays the resulting cohort composition.

Table 2. Physician-category mapping to routing actions

Physician category	Action	n
context does not matter	ANSWER	220
context matters is clear	ANSWER	195
enough context	ANSWER	227
enough info to complete task	ANSWER	215
no uncertainty	ANSWER	228
non emergent	ANSWER	134
only irreducible uncertainty	ANSWER	198
any reducible uncertainty	CLARIFY	285
conditionally emergent	CLARIFY	180
context matters but unclear	CLARIFY	219
not enough context	CLARIFY	181
not enough info to complete task	CLARIFY	180
emergent	ESCALATE	139

Six prompt-only policies used dialogue text and features derived from it. Prompt identifiers, canaries, physician categories and themes, Hard membership, rubric criteria and points, ideal completions, and candidate or physician responses were excluded from their predictors. Categories defined the outcome, while categories and themes also supported post hoc strata. Rubric-Augmented Router and Privileged Cascade formed a separate privileged-information condition because they could access each test case's physician rubrics. Two composition sensitivity analyses independently reassigned not-enough-info-to-complete-task and conditionally-emergent from CLARIFY to ESCALATE. These analyses assess label prevalence, not the robustness of fitted models under alternative action mappings.

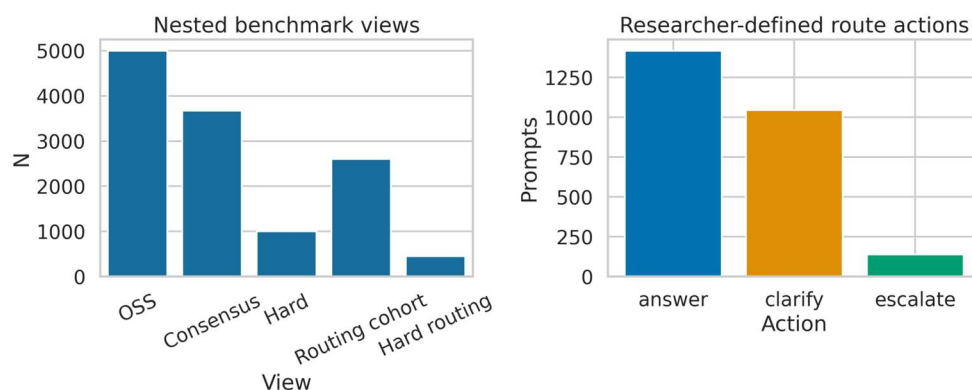


Figure 1. Nested benchmark views and researcher-defined routing-action distribution.

G. Partitioning and Text Representation

All 453 route-relevant Hard prompts were removed before development and locked as a challenge test: 148 ANSWER, 289 CLARIFY, and 16 ESCALATE. The remaining 2,148 prompts were stratified by action with seed 20250512. They formed base-training, stack-training, calibration, and independent in-distribution test partitions. Each prompt_id occurred exactly once. Feature design, probability calibration, thresholds, and acquisition gates were fixed before either test was scored. Table 3 reports the partition sizes and action counts.

Table 3. Disjoint development and evaluation partitions

Partition	Answer	Clarify	Escalate	Total
Base Train	761	453	74	1,288
Stack Train	190	114	18	322
Calibration	127	76	12	215
Id Test	191	113	19	323
Hard Test	148	289	16	453

The base router combined word and character TF-IDF. Word features used Unicode accent stripping, lowercase one- and two-word n-grams, minimum document frequency 2, maximum document frequency 0.995, sublinear term frequency, and a 12,000-feature cap. Character-within-word features used three- to five-character n-grams, minimum document frequency 2, sublinear term frequency, and a separate 12,000-feature cap. A class-balanced multinomial logistic regression used $C = 2.0$, the lbfgs solver, a 1,500-iteration limit, and the study seed. Three-fold sigmoid calibration was fitted within base training, reflecting the importance of calibrated confidence for selective decisions (Guo et al., 2017; Ovadia et al., 2019).

H. Routing Policies

Eight offline policies were evaluated (Table 4). Always Answer deterministically selected ANSWER. TF-IDF Router used the calibrated base classifier. Bootstrap Ensemble trained five TF-IDF logistic routers on independent bootstrap resamples, using 8,000-feature word and character caps, and averaged their three-class probabilities. It aggregates classifier predictions rather than generated reasoning paths. Stacked Classifier fitted a class-balanced second-stage logistic regression on the separate stack-training partition. Its standardized inputs were the three base probabilities, maximum probability, top-two margin, normalized entropy, log dialogue characters, log word count, turn counts, log final-user-message length, question marks, digits, and a URL indicator. This second stage predicted an action from numerical features; it did not generate a critique or revise a response.

Confidence Threshold retained the base argmax when maximum confidence was at least 0.67 and otherwise selected ESCALATE. The threshold was selected on calibration data from 0.34 through 0.90 in 0.01 increments by minimizing mean routing loss plus 0.02 times escalation rate. TF-IDF Router used the calibrated argmax without this override. Both policies operated on dialogue-derived information.

Rubric-Augmented Router combined 65% base probability with 35% rubric-classifier probability. The rubric classifier was a standardized, class-balanced logistic regression fitted on base-training records. Its features summarized rubric count, positive and negative point counts and sums, five rubric axes, and ten routing-relevant word stems in criteria. This privileged-information condition measures the predictive value of benchmark-provided expert criteria. It does not check a generated response or establish that equivalent criteria can be obtained from a live dialogue.

Privileged Cascade acquired classifier evidence in stages (Figure 2). It began with base probabilities; below 0.70 maximum confidence it acquired the five-member bootstrap ensemble and averaged those two distributions. If confidence remained below 0.70, it averaged the base, ensemble, and rubric-augmented distributions. The third input was the 65:35 blend defined above. The final action minimized expected routing loss plus action overhead. Both confidence gates were selected on the calibration grid 0.45 to 0.85 in 0.05 increments by minimizing observed routing loss plus 0.02 times additional acquisition stages. Prompt-Only Cascade used the same first gate and final loss rule but stopped after the ensemble stage, isolating the contribution of rubric access. These cascades are sequential classifier policies without interacting agents.

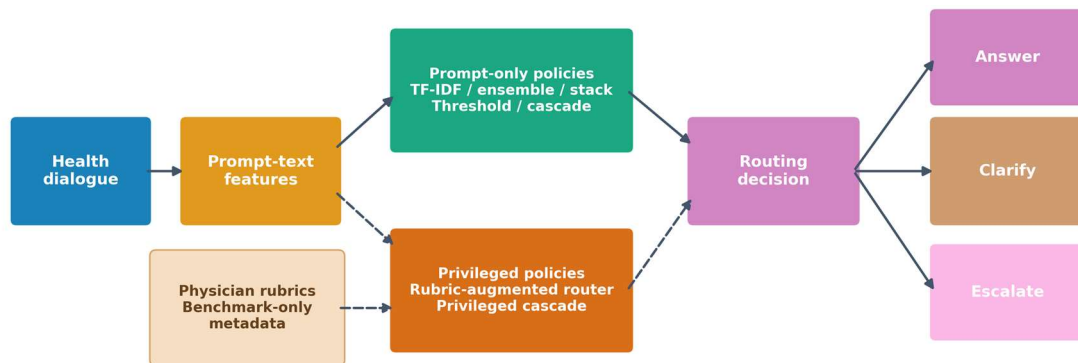


Figure 2. Offline routing workflow and information access. Within the matched cascade comparison, only Privileged Cascade acquires rubric features.

Table 4. Routing policies and information access

Policy	Information at inference	Decision rule	Stages
Always answer	Prompt only	Always ANSWER	0
TF-IDF router	Prompt only	Calibrated logistic argmax	1
Bootstrap	Prompt only	Five bootstrap routers; mean probability	1
Stacked	Prompt only	Second-stage logistic classifier	2
Threshold	Prompt only	Base argmax; ESCALATE below 0.67	1
Prompt cascade	Prompt only	Base → ensemble; expected loss + overhead	1–2
Rubric router*	Prompt and test rubric	65:35 base/rubric probability blend	2
Privileged cascade*	Prompt and test rubric	Base → ensemble → rubric blend; expected loss + overhead	1–3

I. Severity, Metrics, and Statistical Analysis

The prespecified loss matrix used mapped actions as rows and chosen actions as columns: ANSWER [0.00, 0.12, 0.40], CLARIFY [1.00, 0.00, 0.15], and ESCALATE [4.00, 1.00, 0.00] (Figure 3). Directly answering an urgent case received the largest penalty. Both cascades selected the action minimizing $pL + h$, where p is the acquired probability vector, L is the loss matrix, and $h = [0.00, 0.40, 0.80]$ represents action overhead. The other classifiers used probability argmax, except for the confidence override. These operational weights express a research preference; they are not estimates of clinical harm or monetary cost.

Table 5. Prespecified evaluation measures

Measure	Definition	Interpretation
Accuracy / macro-F1 / balanced accuracy	Three-action discrimination	Higher is better
Answer coverage	Fraction routed to ANSWER	Autonomous handling rate
Selective answer accuracy	True ANSWER among predicted answers	Higher is better
Unsafe omission	CLARIFY or ESCALATE routed to ANSWER	Lower is better
Urgent direct-answer omission	ESCALATE routed to ANSWER	Lower is better
Escalation recall	ESCALATE routed to ESCALATE	Higher is better
Missed immediate escalation	ESCALATE routed to ANSWER or CLARIFY	Lower is better
Routing loss / utility	Mean asymmetric loss / mapped action utility	Lower loss / higher utility
ECE-15	Confidence–accuracy difference in 15 bins	Lower is better
Answer AURC	Mean risk at 19 coverage points	Lower is better
Worst-decile accuracy	Mean accuracy of two weakest categories	Higher is better
Stages / evaluations / words / latency	Local routing workload	Descriptive

Table 5 summarizes the evaluation measures. Accuracy, macro-F1, and balanced accuracy measured three-action discrimination. Answer coverage, clarification rate, escalation rate, and selective answer accuracy described routing behavior. Unsafe direct-answer omission was the fraction of true CLARIFY or ESCALATE cases routed to ANSWER; urgent direct-answer omission restricted that denominator to true ESCALATE. Escalation recall counted urgent cases routed to ESCALATE, and missed immediate escalation included both ESCALATE-to-ANSWER and ESCALATE-to-CLARIFY errors. A researcher-defined routing utility averaged values [1.00, 0.55, 0.10] for true ANSWER, [0.00, 1.00, 0.55] for true CLARIFY, and [-1.00, 0.20, 1.00] for true ESCALATE, with the overall mean clipped to [0, 1]. This utility is distinct from official response-level HealthBench grading.

Calibration used 15 equal-width bins and expected calibration error. Confidence was the probability assigned to the selected action, which can differ from maximum class probability for threshold and cascade policies. Risk-coverage analysis ranked prompts by ANSWER probability and evaluated 19 points from 10% to 100% answer coverage; selective risk equaled one minus

the proportion of selected prompts whose mapped action was ANSWER, and answer AURC was the mean across points. Category robustness was the mean accuracy of the two lowest-performing categories, representing the lowest decile of the 13-category cohort (Sagawa et al., 2020).

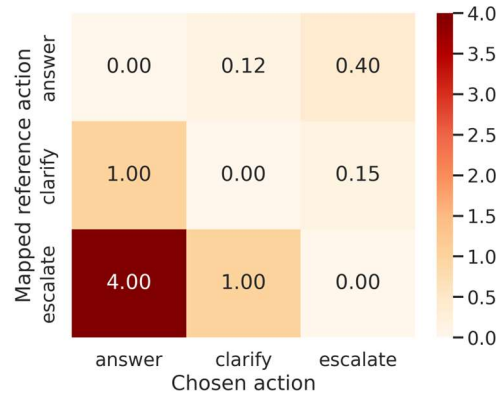


Figure 3. Prespecified asymmetric routing-loss matrix. Rows are researcher-mapped actions; columns are predicted actions.

Resource reporting separated acquisition stages, individual classifier evaluations, and a lexical workload proxy. For prompt length L and rubric length R in regex-delimited words, the base, bootstrap, stacked, and rubric policies were assigned L , $5L$, $2L$, and $L + R$. A cascade was assigned $L + 5L$ when it acquired the ensemble, with R added only when it acquired rubric features. The calibrated base contained three fitted classifiers, counted separately in classifier evaluations.

Table 6. In-distribution three-action performance with 95% bootstrap intervals

Policy	Accuracy [CI]	Macro-F1 [CI]	Balanced acc.	Loss [CI]	ECE
Always answer	0.591 [0.591, 0.591]	0.248 [0.248, 0.248]	0.333	0.585 [0.585, 0.585]	0.409
TF-IDF router	0.638 [0.604, 0.675]	0.521 [0.441, 0.600]	0.526	0.402 [0.339, 0.459]	0.067
Bootstrap	0.659 [0.610, 0.706]	0.567 [0.483, 0.646]	0.553	0.282 [0.227, 0.340]	0.061
Stacked	0.588 [0.539, 0.641]	0.527 [0.469, 0.585]	0.606	0.219 [0.171, 0.271]	0.077
Threshold	0.217 [0.180, 0.254]	0.176 [0.148, 0.203]	0.391	0.293 [0.256, 0.335]	0.040
Prompt cascade	0.591 [0.542, 0.641]	0.527 [0.433, 0.612]	0.511	0.170 [0.138, 0.200]	0.072
Rubric router*	0.700 [0.653, 0.743]	0.617 [0.542, 0.689]	0.627	0.290 [0.229, 0.349]	0.139
Privileged cascade*	0.598 [0.548, 0.647]	0.546 [0.456, 0.629]	0.535	0.156 [0.124, 0.183]	0.097

Local component timings used a monotonic clock; cascade latency was estimated by weighting component timings by acquisition rates. These measurements exclude network delay and response generation. Two-sided 95% percentile intervals used 1,000 prompt-level stratified bootstrap replicates with seed 20250611, preserving action counts. Wilson intervals summarized urgent-

case proportions. Models were evaluated on the 323-prompt in-distribution test and then on the locked 453-prompt Hard slice without retraining or retuning. Paired cascade differences used the same resampled cases for both policies.

Table 7. In-distribution routing behavior and direct-answer omissions

Policy	Answer cov.	Answer acc.	Clarify	Escalate	Unsafe omission	Urgent omission	Utility
Always answer	100.00	59.13	0.00	0.00	100.00	100.00	0.533
TF-IDF router	85.45	64.13	9.91	4.64	75.00	47.37	0.635
Bootstrap	63.47	71.71	31.89	4.64	43.94	36.84	0.714
Stacked	50.46	73.62	35.29	14.24	32.58	15.79	0.705
Threshold	22.29	73.61	0.00	77.71	14.39	10.53	0.417
Prompt cascade	41.49	77.61	56.35	2.17	22.73	0.00	0.748
Rubric router*	71.83	71.55	21.05	7.12	50.00	36.84	0.720
Privileged cascade*	39.32	79.53	57.89	2.79	19.70	0.00	0.758

IV. RESULTS AND FINDINGS

A. Cohort Composition and In-Distribution Performance

The routing cohort contained 54.5% ANSWER, 40.2% CLARIFY, and 5.3% ESCALATE examples. The in-distribution test retained 59.1% ANSWER, whereas the Hard slice contained only 32.7% ANSWER and 63.8% CLARIFY. The challenge set therefore directly tested behavior when answering by default was often inappropriate. Under privileged rubric access, Rubric-Augmented Router achieved the highest in-distribution accuracy, 69.97% (95% CI [65.33%, 74.30%]), and macro-F1, 61.67%.

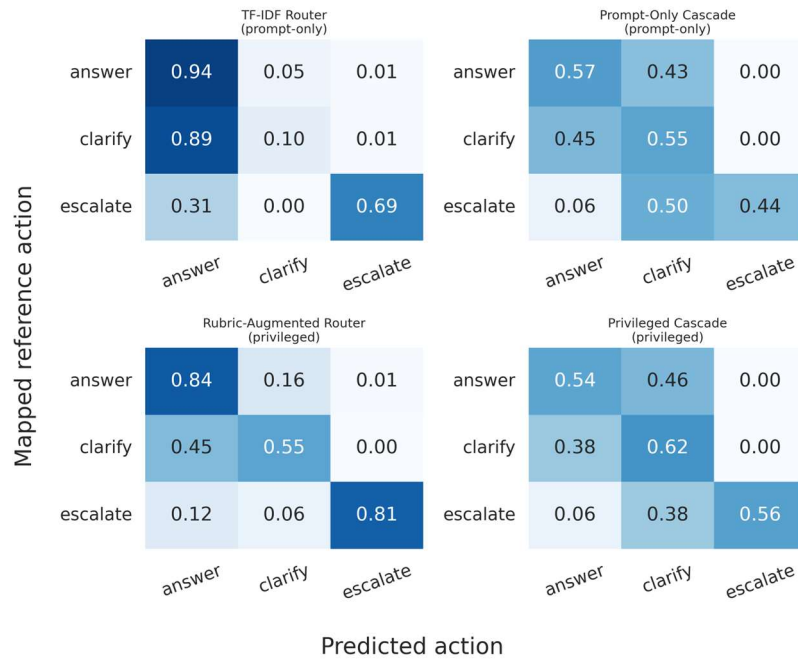


Figure 4. Row-normalized confusion matrices on the locked Hard slice. Rubric-Augmented Router and Privileged Cascade have access to case-specific physician rubrics.

Among prompt-only policies, Bootstrap Ensemble had the highest accuracy, 65.94%, and macro-F1, 56.68%. Always Answer reached 59.13% accuracy because that proportion of cases was mapped to ANSWER, with macro-F1 of 24.77%. Stacked Classifier reached 58.82% accuracy and 52.67% macro-F1. Confidence Threshold escalated 77.71% of cases and selected no CLARIFY actions on this test, although its rule permits high-confidence clarification. Table 6 reports the full comparison.

Table 8. Locked Hard-slice three-action performance with 95% bootstrap intervals

Policy	Accuracy [CI]	Macro-F1 [CI]	Balanced acc.	Loss [CI]	ECE
Always answer	0.327 [0.327, 0.327]	0.164 [0.164, 0.164]	0.333	0.779 [0.779, 0.779]	0.673
TF-IDF router	0.397 [0.371, 0.426]	0.474 [0.404, 0.534]	0.577	0.615 [0.576, 0.656]	0.219
Bootstrap	0.470 [0.428, 0.510]	0.549 [0.483, 0.610]	0.599	0.466 [0.425, 0.511]	0.170
Stacked	0.501 [0.457, 0.543]	0.491 [0.440, 0.545]	0.617	0.362 [0.321, 0.403]	0.160
Threshold	0.150 [0.124, 0.177]	0.164 [0.140, 0.189]	0.432	0.295 [0.266, 0.330]	0.083
Prompt cascade	0.552 [0.503, 0.594]	0.561 [0.468, 0.636]	0.520	0.332 [0.296, 0.372]	0.213
Rubric router*	0.653 [0.611, 0.695]	0.709 [0.644, 0.762]	0.734	0.312 [0.269, 0.355]	0.096
Privileged cascade*	0.589 [0.543, 0.636]	0.616 [0.538, 0.679]	0.573	0.285 [0.246, 0.322]	0.184

Privileged Cascade achieved the lowest in-distribution routing loss, 0.1564 (95% CI [0.124, 0.183]), with 59.75% accuracy. Removing the rubric stage increased loss to 0.1699 and reduced accuracy to 59.13%. Prompt-Only Cascade still had lower loss than the other prompt-only policies, including Stacked Classifier (0.2188). The matched comparison isolates rubric acquisition within the cascade, whereas comparisons with argmax classifiers also reflect a different decision objective.

Table 9. Hard-slice routing behavior and direct-answer omissions

Policy	Answer cov.	Answer acc.	Clarify	Escalate	Unsafe omission	Urgent omission	Utility
Always answer	100.00	32.67	0.00	0.00	100.00	100.00	0.291
TF-IDF router	88.52	34.66	8.39	3.09	85.90	31.25	0.399
Bootstrap	66.67	35.43	29.58	3.75	63.93	18.75	0.515
Stacked	52.32	37.97	39.07	8.61	48.20	6.25	0.586
Threshold	24.50	47.75	0.00	75.50	19.02	6.25	0.451
Prompt cascade	47.90	39.17	50.55	1.55	43.28	6.25	0.630
Rubric router*	56.29	48.63	40.40	3.31	42.95	12.50	0.679
Privileged cascade*	42.38	41.67	55.63	1.99	36.72	6.25	0.672

B. Coverage, Omission Severity, and Error Structure

Always Answer directly answered every case mapped to clarification or escalation. TF-IDF Router retained 85.45% answer coverage but directly answered 75.00% of non-answer cases and 47.37% of urgent cases. Bootstrap Ensemble reduced these rates to 43.94% and 36.84%; Stacked Classifier reduced them to 32.58% and 15.79%. Privileged Cascade directly answered 26/132 non-answer cases (19.70%), compared with 30/132 (22.73%) for Prompt-Only Cascade (Table 7). Neither cascade directly answered any of the 19 urgent cases. However, Privileged Cascade escalated only 6/19, sending the remainder to clarification; Prompt-Only Cascade escalated 5/19. Zero urgent direct answers therefore did not imply complete urgent-case recognition. Confidence Threshold had lower unsafe direct-answer omission (14.39%) but escalated most cases.

Table 10. Actions selected for urgent cases on the locked Hard slice

Policy	Urgent → answer	Urgent → clarify	Urgent → escalate	Escalation recall [95% CI]
Always answer	16/16	0/16	0/16	0.00% [0.00%, 19.36%]
TF-IDF router	5/16	0/16	11/16	68.75% [44.40%, 85.84%]
Bootstrap	3/16	1/16	12/16	75.00% [50.50%, 89.82%]
Stacked	1/16	2/16	13/16	81.25% [56.99%, 93.41%]
Threshold	1/16	0/16	15/16	93.75% [71.67%, 98.89%]
Prompt cascade	1/16	8/16	7/16	43.75% [23.10%, 66.82%]
Rubric router*	2/16	1/16	13/16	81.25% [56.99%, 93.41%]
Privileged cascade*	1/16	6/16	9/16	56.25% [33.18%, 76.90%]

C. Hard-Set Robustness

Prompt-only performance deteriorated on the Hard slice. TF-IDF Router accuracy fell to 39.74%, with unsafe direct-answer omission of 85.90%. Bootstrap Ensemble reached 47.02% accuracy and 63.93% unsafe omission; Stacked Classifier reached 50.11% and 48.20%. Prompt-Only Cascade achieved 55.19% accuracy with loss 0.3324. Figure 4 shows confusion matrices for the text router, the two cascades, and the rubric-augmented router.

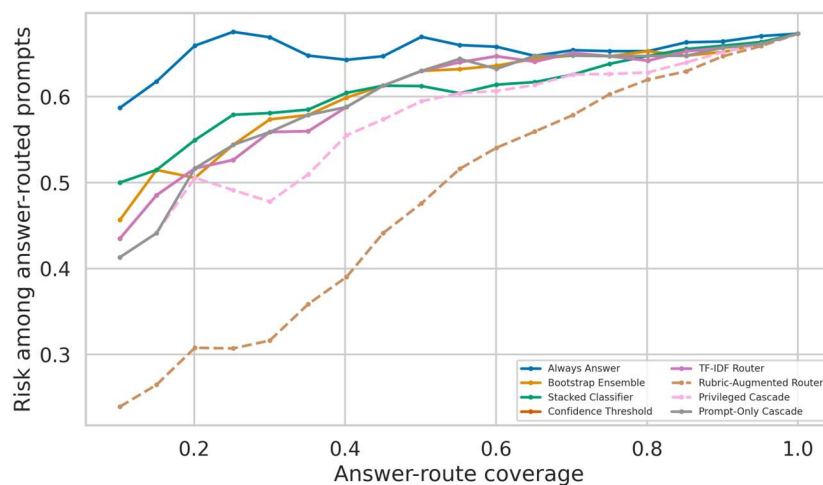


Figure 5. Autonomous-answer risk versus coverage on the locked Hard slice. Dashed curves indicate privileged rubric access. TF-IDF Router and Confidence Threshold share the same answer ranking. Always Answer has constant confidence, with ties resolved in stable sample order.

Rubric-Augmented Router remained the strongest three-way classifier under privileged access, with 65.34% accuracy (95% CI [61.15%, 69.54%]), 70.93% macro-F1, and 73.35% balanced accuracy (Table 8). Privileged Cascade reached 58.94% accuracy and the lowest observed routing loss, 0.2851, compared with 0.3324 for the matched prompt-only version. Rubric acquisition reduced unsafe direct-answer omission from 132/305 (43.28%) to 112/305 (36.72%) (Table 9). This improvement depends on benchmark annotations and leaves substantial residual error. The reduction was 6.56 percentage points (paired 95% bootstrap CI 3.60–9.84).

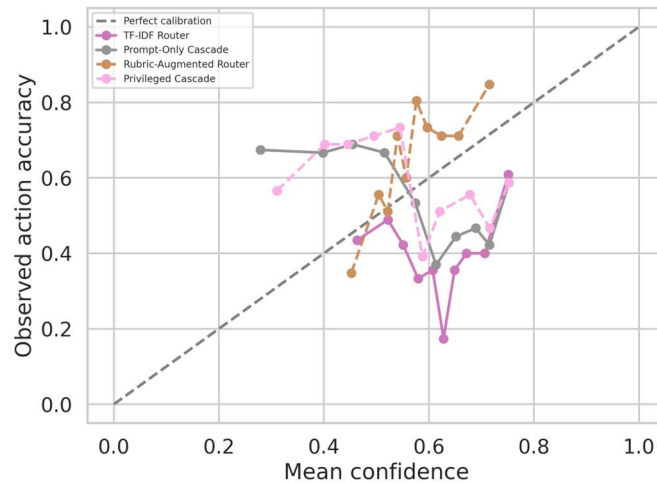


Figure 6. Hard-slice reliability diagrams for the text router, prompt-only cascade, rubric-augmented router, and privileged cascade. Points summarize ten quantile bins; the reported ECE uses 15 equal-width bins.

Table 10 separates the possible actions for the 16 urgent Hard cases. Privileged Cascade routed 1/16 to ANSWER, 6/16 to CLARIFY, and 9/16 to ESCALATE. Its escalation recall was 56.25% (95% Wilson CI 33.18%–76.90%), while 7/16 urgent cases missed immediate escalation. Prompt-Only Cascade missed 9/16. Rubric-Augmented Router and Stacked Classifier each escalated 13/16. Thus, the cascade's lower average loss did not provide the highest urgent-case recall.

Table 11. Risk-coverage, calibration, and category-tail robustness

Policy	AURC IID	AURC Hard	Tail acc. IID	Tail acc. Hard	ECE IID	ECE Hard
Always answer	0.422	0.653	0.00%	0.00%	0.409	0.673
TF-IDF router	0.292	0.601	9.43%	7.60%	0.067	0.219
Bootstrap	0.283	0.606	32.31%	26.82%	0.061	0.170
Stacked	0.295	0.607	35.56%	33.44%	0.077	0.160
Threshold	0.292	0.601	0.00%	0.00%	0.040	0.083
Prompt cascade	0.284	0.599	20.66%	41.69%	0.072	0.213
Rubric router*	0.211	0.480	32.94%	47.16%	0.139	0.096
Privileged cascade*	0.264	0.573	20.79%	42.08%	0.097	0.184

D. Risk-Coverage, Calibration, and Category Tails

Rubric-Augmented Router had the lowest answer AURC in-distribution (0.2109) and on Hard (0.4803). Privileged Cascade followed at 0.2645 and 0.5733; Prompt-Only Cascade reached 0.2836 and 0.5994. All policies had higher answer risk on Hard (Figure 5). The better ranking associated with rubric features remains contingent on access to those features.

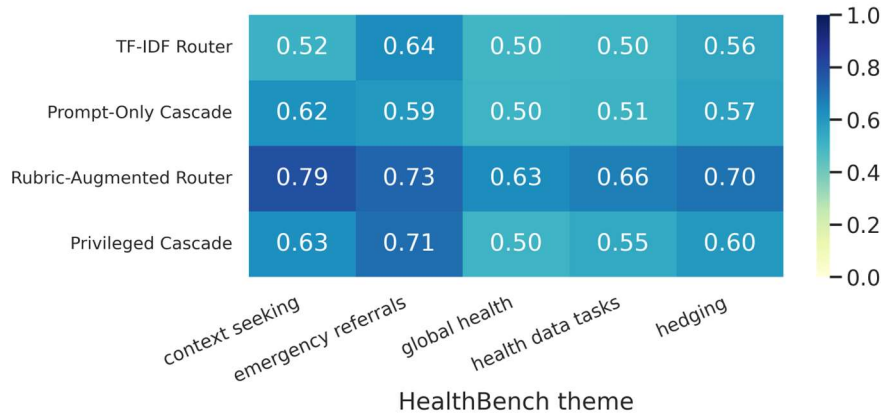


Figure 7. Hard-slice balanced accuracy across five HealthBench themes for the principal text-only and privileged-information policies.

Calibration and action quality were not interchangeable. Confidence Threshold had the lowest 15-bin ECE on both tests, 0.0396 and 0.0832, despite low three-action accuracy and a high escalation rate. Rubric-Augmented Router had ECE 0.1386 in-distribution and 0.0963 on Hard. TF-IDF Router deteriorated from 0.0673 to 0.2189. Figure 6 displays reliability for the main text and rubric comparisons; Table 11 places calibration alongside risk-coverage and category-tail performance.

Table 12. Hard-slice accuracy by HealthBench theme for selected routers

Policy	Context seeking	Emergency referrals	Global health	Health-data tasks	Hedging
TF-IDF router	30.77%	46.43%	46.51%	34.52%	41.25%
Prompt cascade	60.58%	67.86%	50.39%	51.19%	51.25%
Rubric router*	76.92%	64.29%	62.02%	59.52%	62.50%
Privileged cascade*	63.46%	78.57%	51.16%	54.76%	56.25%

Stacked Classifier had the highest in-distribution worst-decile category accuracy, 35.56%. On Hard, Rubric-Augmented Router led at 47.16%, followed by Privileged Cascade at 42.08% and Prompt-Only Cascade at 41.69%. These tail measures describe variation across the researcher-mapped categories; they do not establish demographic fairness or performance at an external clinical site.

E. Theme Performance, Workload, and Sensitivity

Privileged Cascade achieved 78.57% accuracy on the Hard emergency-referrals theme, compared with 64.29% for Rubric-Augmented Router. Theme accuracy is distinct from escalation recall because this theme includes cases mapped to other actions. Figure 7 shows balanced accuracy across the five themes; Table 12 reports ordinary accuracy for the principal text-only and rubric-assisted comparisons. These results are descriptive subgroup analyses.

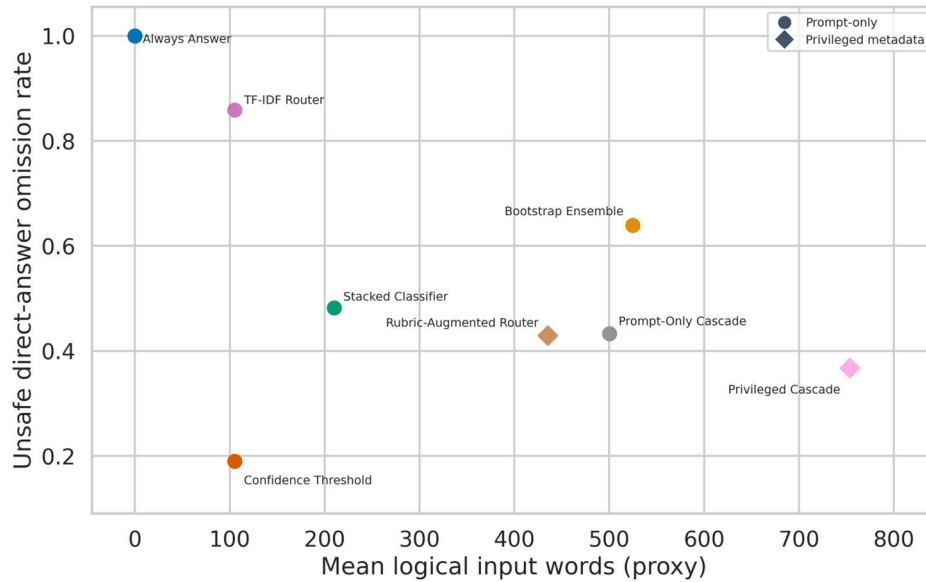


Figure 8. Hard-slice unsafe direct-answer omission versus lexical workload proxy for eight routing policies. Diamond markers identify privileged rubric access.

Figure 8 compares unsafe direct-answer omission with lexical workload, and Table 13 reports acquisition stages, classifier evaluations, and estimated local latency. On Hard, Privileged Cascade acquired the ensemble for 82.12% of cases and rubric features for 77.04%. It averaged 2.59 stages and 7.88 individual classifier evaluations per case, compared with 1.82 and 7.11 for Prompt-Only Cascade. Lower loss therefore accompanied both additional computation and privileged information. The local workload measures do not estimate hosted LLM token costs or professional review time.

Table 13. Computational workload and estimated local latency

Policy	Stages IID / Hard	Evals IID / Hard	Words IID / Hard	Latency ms IID / Hard
Always answer	0.00 / 0.00	0.00 / 0.00	0 / 0	Not timed
TF-IDF router	1.00 / 1.00	3.00 / 3.00	92 / 105	1.45 / 2.07
Bootstrap	1.00 / 1.00	5.00 / 5.00	461 / 525	2.64 / 3.22
Stacked	2.00 / 2.00	4.00 / 4.00	184 / 210	1.49 / 2.12
Threshold	1.00 / 1.00	3.00 / 3.00	92 / 105	1.45 / 2.07
Prompt cascade	1.86 / 1.82	7.30 / 7.11	435 / 500	3.72 / 4.71
Rubric router*	2.00 / 2.00	4.00 / 4.00	373 / 435	1.48 / 2.11
Privileged cascade*	2.65 / 2.59	8.10 / 7.88	662 / 753	3.75 / 4.75

The primary ontology assigned 139 of 2,601 prompts (5.34%) to ESCALATE. Reassigning either the 180 not-enough-information cases or the 180 conditionally emergent cases to ESCALATE increased that proportion to 12.26% and reduced CLARIFY from 40.18% to 33.26% (Table 14). Each alternative moves 6.92% of cases across the clarify-escalate boundary. This composition analysis demonstrates that label prevalence depends on the operational mapping; it does not evaluate model robustness under the alternative mappings.

Table 14. Sensitivity of action composition to alternative mappings

Mapping	Answer	Clarify	Escalate
Conditional-emergency escalates	1,417 (54.48%)	865 (33.26%)	319 (12.26%)
Not-enough-info escalates	1,417 (54.48%)	865 (33.26%)	319 (12.26%)
Primary	1,417 (54.48%)	1,045 (40.18%)	139 (5.34%)

F. Discussion

The matched cascades show that part of the strongest performance depends on case-specific physician criteria. Removing rubric acquisition worsened both action accuracy and routing loss, even with the first confidence gate and final loss rule held constant. The remaining prompt-only performance is the relevant baseline for a setting where benchmark annotations are unavailable. Developing a verifier from observable dialogue, then evaluating it without access to test rubrics, remains necessary before attributing the rubric-assisted gains to a deployable control system.

V. CONCLUSION AND RECOMMENDATION

This study compared eight offline policies for pre-generation routing among answering, clarification, and escalation. A researcher-defined mapping of physician-agreed HealthBench categories produced 2,601 route-relevant dialogues. The framework made differences in action accuracy, direct-answer omissions, urgent-case recall, and computational workload visible within a common evaluation. Its implemented components were lexical classifiers, a bootstrap ensemble, a stacked classifier, rubric-feature classification, and deterministic routing rules.

Rubric-Augmented Router achieved the highest three-way accuracy, while Privileged Cascade minimized the selected severity-weighted loss. Both had access to prompt-specific physician rubrics. On Hard, removing the cascade's rubric stage increased loss from 0.2851 to 0.3324, establishing the contribution of that extra information under the matched policy. The privileged cascade still directly answered 112/305 non-answer cases and failed to escalate 7/16 urgent cases. These results support controlled prototyping, with no evidence of readiness for autonomous clinical deployment.

The mapped actions require independent physician adjudication, and the urgent subgroups are small. The study did not evaluate generated medical responses, the benefit of follow-up questions, or completed professional handoffs. HealthBench Hard is a nested challenge slice rather than an external-site test. Future evaluation should compare observable-input verifiers with the privileged

reference condition and assess both direct answers and delayed escalation in prospective workflows.

DATA AVAILABILITY STATEMENT

HealthBench data are publicly available through the release accompanying Arora et al. (2025), including the OSS Eval, Consensus, and Hard views. The analysis code, split assignments, aggregate results, and figure-generation scripts accompany this article.

REFERENCES

- Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., & Singhal, K. (2025). HealthBench: Evaluating large language models towards improved human health [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2505.08775>
- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6), 589–596. <https://doi.org/10.1001/jamainternmed.2023.1838>
- Bates, S., Angelopoulos, A., Lei, L., Malik, J., & Jordan, M. I. (2021). Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), Article 43. <https://doi.org/10.1145/3478535>
- Chen, L., Zaharia, M., & Zou, J. (2023). FrugalGPT: How to use large language models while reducing cost and improving performance [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2305.05176>
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2305.14325>
- El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11, 1605–1641. <https://jmlr.org/papers/v11/el-yaniv10a.html>
- Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4878–4887). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cfl-Abstract.html
- Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 2151–2159). PMLR. <https://proceedings.mlr.press/v97/geifman19a.html>

- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. <https://proceedings.mlr.press/v70/guo17a.html>
- Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9, 962–977. https://doi.org/10.1162/tacl_a_00407
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2), e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 6402–6413). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html>
- Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13), 1233–1239. <https://doi.org/10.1056/NEJMSr2214184>
- Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=8s8K2UZGTZ>
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2511–2522). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.153>
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-Refine: Iterative refinement with self-feedback. In A. H. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 46534–46594). Curran Associates, Inc. <https://doi.org/10.52202/075280-2019>
- Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31, pp. 6147–6157). Curran Associates, Inc.

<https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html>

- Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9004–9017). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.557>
- Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265. <https://doi.org/10.1038/s41586-023-05881-4>
- Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In H. Daumé III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 7076–7087). PMLR. <https://proceedings.mlr.press/v119/mozannar20b.html>
- Nie, J., & Zheng, D. (2023). Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal. *Journal of Advanced Computing Systems*, 3(1), 66–80. <https://doi.org/10.69987/JACS.2023.30105>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 13991–14002). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html
- Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=ryxGuJrFvS>
- Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12), 2176–2182. <https://doi.org/10.1038/s41591-021-01595-0>
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In A. H. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 8634–8652). Curran Associates, Inc. <https://doi.org/10.52202/075280-0377>

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180. <https://doi.org/10.1038/s41586-023-06291-2>

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8), 1229–1234. <https://doi.org/10.1038/s41591-020-0942-0>

Varshney, N., & Baral, C. (2022). Model cascading: Towards jointly improving efficiency and accuracy of NLP systems. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 11007–11021). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.756>

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain-of-thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=1PL1NIMMrw>

Zhang, H. (2023). DriftGuard: Multi-signal drift early warning and safe re-training/rollback for CTR/CVR models. *Journal of Advanced Computing Systems*, 3(7), 24–40. <https://doi.org/10.69987/JACS.2023.30703>

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In A. H. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 46595–46623). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html

Zhong, Z. S., Ma, R., & Zhao, H. (2023). Human-uncertainty distillation for calibrated vision models on CIFAR-10H. *Journal of Advanced Computing Systems*, 3(2), 77–89. <https://doi.org/10.69987/JACS.2023.30206>