A prospectively specified test of a small-model colorectal screening assistant on synthetic charts

Brian McDonough, DHA, MSN, FNP-BC

October 2026. Working paper. All patient data are synthetic.

Abstract

Background. Deciding whether a patient is due for colorectal cancer screening, so that outreach reaches the right patients, requires reading heterogeneous chart documents (procedure letters, laboratory results, care-gap reports, notes) and applying interval rules. We tested a design in which a small language model only reads, deterministic code applies the USPSTF 2021 recommendation, and a score derived from the model's answer probabilities decides which charts go to a person.

Methods. A 0.8-billion-parameter model (Qwen3.5-0.8B with a LoRA adapter, "r5") answers two multiple-choice questions per document: the kind of evidence and the test date. Rules code maps the readings to one of nine screening lanes (output categories). Decision stability is a score computed from the reader's answer probabilities: for each document, the probability of readings that leave the lane unchanged, multiplied across documents. Charts below a threshold go to a person; the rest are automated, meaning the lane is accepted without review. A hand-off policy (a main threshold and a stricter threshold for "up to date" calls) was selected from 42 pre-listed policies on a 10,000-chart calibration set written by three language-model families. The test set (test4), 10,000 charts written by four other model families not involved in training, development or calibration, was read blind, frozen by hash and analysed once by a script. Primary endpoints were chart accuracy, the exact one-sided 95% upper bounds on errors (target 2%) and serious errors (target 0.5%) among automated charts with at least 50% of charts automated, and throughput. A serious error misses needed clinical action or wrongly calls a patient up to date; the taxonomy is an engineering one, not a clinical harm scale. A follow-up round retrained the reader ("r6") on the main evaluation's failure categories, calibrated each trained model on its own, and tested it once on a 10,000-chart challenge set (test5) written by the calibration authors.

Results. r5 agreed with the generator's labels on 98.78% of charts (Wilson 95% CI 98.55% to 98.98%), more than its predecessor (+0.70 points, Holm p 6e-08) and a much larger zero-shot model (+1.15 points, Holm p 2.8e-09); without a hand-off, its serious-error rate did not differ detectably from the larger model's (1.17% against 0.96%). Its answer probabilities were well calibrated (Q1 Brier score 0.0176, expected calibration error 0.0102). Under the selected policy it automated 82.3% of charts with 11 errors, 7 of them serious (upper bounds 0.22% and 0.16%), meeting both targets. These bounds treat charts as independent draws from the tested wording families; they say nothing about a new author, and all 7 automated serious errors came from one author's charts. All 11 automated errors came from two templates by that author, each a history read as an order: 7, all serious, were care-gap recall reports listing a prior adenoma, and 4 were nurse notes recording a patient-reported stool DNA test. In analyses run after unblinding, seed 0's policy did not meet the calibration criterion for any of the four models retrained with other seeds; under the policy the same rule selected for each, each had no automated serious errors on test4. r5 read 22.405 charts per second on one GPU

(late five-pass timing; 22.44 in the pre-specified single pass), about 19 times the larger model's throughput as deployed for interactive use. In the follow-up round r6 read the target categories better than r5 (+5.31 points) and met both targets (3 automated serious errors, upper bound 0.10%), but automated serious errors did not fall (3 against 3) and r6 automated fewer charts (78.8% against 89.0%). On the oldest held-out set (test2), r5 and r6 were less accurate than the earlier reader r3, and neither met the serious-error target under its policy (upper bounds 1.72% and 3.60%). In a post hoc sensitivity analysis that follows the author's clinical ruling and counts records-needed charts read as up to date as not serious, no automated count changed, and r5 had fewer raw serious errors than the larger model (0.27% against 0.91%).

Conclusions. On synthetic charts, a small reader with rules code and a pre-specified hand-off met its pre-specified error bounds on 10,000 charts from four authors not used in development, and each automated error traced to a single misread document in one of two templates. The bounds hold for charts like those tested, not for a new author. Thresholds selected for one trained model did not transfer to models retrained with other seeds, which supports calibrating each trained model. Retraining on the observed failures improved reading but did not reduce automated serious errors; because the follow-up round's calibration and test sets were written by the same authors, whose calibration errors had informed two training categories, it could not show whether the improvement carries over to new authors. The system is meant to route charts for screening outreach, not to make clinical decisions. The study does not establish clinical safety: the documents and labels are synthetic and the guideline encoding has not been independently reviewed by clinicians.

1. Introduction

Screening outreach for colorectal cancer depends on knowing who is up to date. In most records that status is not a field; it has to be read out of faxed colonoscopy letters, stool-test results, outside reports, care-gap lists and notes, and then combined with the patient's age and the intervals in the USPSTF 2021 recommendation [1]. The reading is the hard part. The rules are short enough to write as code.

Large language models can read such documents, but asking a model for the final answer mixes reading errors with rule errors and gives no principled way to tell which answers to trust. We tested a narrower design, which we call a calibrated classification engine: the model answers small, closed questions about each document; code applies the guideline; and the model's own probabilities decide which charts it may finish alone. Because the model never outputs a lane, every lane can be traced to specific readings of specific documents.

The question was whether this design, with a model small enough to run quickly on one GPU, can meet a pre-stated error bound on documents written in wording it was not trained on, and what its failures look like when it does not. This paper reports the main evaluation of a prospectively specified study (protocol amendment 6, test set test4) and a follow-up round that retrained the reader on that evaluation's failures and tested it once more (amendment 7, test set test5; section 4). Two earlier rounds developed the design; their test sets (test2 and test3) appear here only as regression checks.

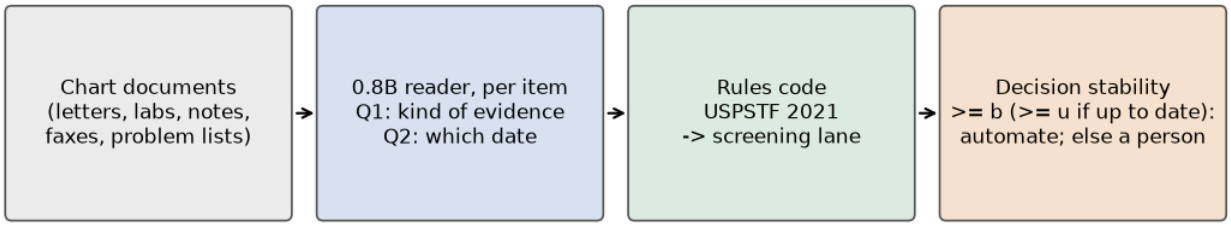

Figure 1. The pipeline. The reader answers two questions per document; code applies the guideline; decision stability decides who finishes the chart.

1.1 Related work

Abstention and hand-off. Letting a classifier decline some inputs is an old idea. Chow described the optimum trade-off between errors and rejections [2]; selective classification formalised the trade-off between risk and coverage [3] and applied it to deep networks [4, 5], often scoring confidence by the model's largest output probability [6]. Learning to defer trains a model to hand a decision to a person or another model [7, 8]. Our design keeps the reader as it is and puts the hand-off in code: the score comes from the reader's answer probabilities passed through the rules, and the threshold is chosen afterwards on a separate set.

Choosing a threshold with a guarantee. Distribution-free risk control chooses a model's parameters on calibration data so that a stated risk holds with finite-sample guarantees [9, 10]; Learn then Test frames that choice as multiple hypothesis testing [11]. Our policy selection is a simpler relative: it takes the policy that automates the most calibration charts while its exact upper bounds meet both targets, then tests that single policy once on a separate set. Those guarantees assume that calibration and test data are exchangeable. Text from a different author breaks that assumption, which is why our test wording came from authors not used for calibration, and why the bounds we report describe charts like the ones tested.

Calibration of language models. Modern neural networks are often overconfident, and temperature scaling is a simple remedy [12]. Pre-trained transformers can be fairly well calibrated within their training domain [13], while language models answering questions have been found poorly calibrated [14]. Large models can be well calibrated on multiple-choice questions posed in the right format [15], which is how our reader answers. Calibration needs to be checked under dataset shift, not only within the training distribution [16], and the expected calibration error is an imperfect estimator [17, 18], so we report the Brier score and negative log-likelihood alongside it.

General-purpose decision models. Open "System One" decision models score a document against typed questions and return a probability for each predefined option in one forward pass, without generating text [19, 20]. An evaluation of one such model on medical benchmarks found that its calibration and discrimination varied by task, and called for task-specific validation before clinical use [21]. The closest work to ours fine-tuned a 4-billion-parameter decision model on the records of a multi-hospital health system. It answers questions posed at request time about notes, reports and structured timelines, and its largest

gains over open models came on that system's own conventions [22]. It reports calibration error and the share of answers each model could decide at a fixed precision, and illustrates accepting confident answers automatically and sending the rest to a reviewer, but it does not test a hand-off policy chosen in advance against an error target. That system is far broader than ours and is trained and tested on real records. Ours is narrow and synthetic, and tests that last step: a hand-off policy selected on calibration charts and tested once on charts from other authors, with thresholds calibrated for each trained model.

Reading colonoscopy status from records. Earlier systems used natural language processing on real records to find completed colonoscopies and their timing in clinical notes [23] and to extract findings from colonoscopy and pathology reports for quality measures and adenoma detection [24, 25, 26]. They worked on real documents; this study uses synthetic ones, its main limitation, but tests across wording authors and adds a calibrated hand-off.

Outreach and guidelines. Mailed stool-test and colonoscopy outreach have been compared with usual care in randomized trials in safety-net health systems [27, 28], and health-plan quality measures count the share of adults aged 45 to 75 who are appropriately screened [29]. Both depend on knowing who is already up to date, which is the decision this system makes. We encode USPSTF 2021 [1]; other guidelines differ from it in places [30, 31], and surveillance after polypectomy follows separate intervals [32].

Reporting. Reporting guidance exists for prediction models that use machine learning [33], for studies of large language models [34] and for early clinical evaluation of AI decision support [35]. This is a methods study on synthetic data and does not claim to follow those checklists; they set out what a study on real documents would need to report.

2. Methods

2.1 Task and guideline encoding

Each chart has a patient age and sex, a visit date, and zero to six documents ("items"). The output is one of nine lanes (categories), assigned by the first rule that applies, in this order: symptoms (GI symptoms documented within six months, a convention, needing work-up); not average risk (a high-risk history, including a prior adenoma); follow-up colonoscopy (an abnormal stool or imaging result not yet followed up, at any age); below 45; above 85; up to date; obtain records (a test is reported but not documented, or documented without a date); shared decision (ages 76 to 85 and not up to date); and due.

The rules engine encodes USPSTF 2021 [1] with stated conventions where the recommendation is silent (`docs/SPEC.md`): a FIT or high-sensitivity gFOBT counts for one year, stool DNA (sDNA-FIT) for three (the upper end of USPSTF's one to three), CT colonography and flexible sigmoidoscopy for five, a complete colonoscopy for ten; an incomplete colonoscopy does not count; a patient's report without documentation, or a qualifying test without a date, goes to "obtain records"; a year-only date means January 1. Flexible sigmoidoscopy every ten years with a yearly FIT is not coded separately, because the FIT alone keeps the patient up to date. The rules engine has unit tests at every cutoff, plus and minus one day and one year.

An error is serious when the true lane needs clinical action (symptoms, follow-up colonoscopy, not average risk) and the system gives any other lane, or when the system calls a patient up to date whose true lane is due, shared decision or obtain records. The taxonomy is an engineering one, written before the results, and is not a clinical harm scale. After the editorial review of this paper, the author ruled that one pre-specified serious pair, a records-needed chart called up to date, is not clinically serious: most such charts hold a letter reporting a normal colonoscopy whose only date is the letter's (section 3.6). The pre-specified definition governs every primary result; section 3.11 reports a post hoc sensitivity analysis with that pair reclassified. Table 1 gives every pair. The taxonomy does not count unnecessary outreach (for example a patient under 45 or over 85 routed to "due") or a due patient sent to "obtain records", which delays screening while records are sought.

True lane \ assigned	UTD	Due	FU	OR	SD	NAR	Sx	<45	>85
Up to date (UTD)	–	·	·	·	·	·	·	·	·
Due (Due)	**S**	–	·	·	·	·	·	·	·
Follow-up colonoscopy (FU)	**S**	**S**	–	**S**	**S**	**S**	**S**	**S**	**S**
Obtain records (OR)	**S**	·	·	–	·	·	·	·	·
Shared decision (76-85) (SD)	**S**	·	·	·	–	·	·	·	·
Not average risk (NAR)	**S**	**S**	**S**	**S**	**S**	–	**S**	**S**	**S**
Symptoms (Sx)	**S**	**S**	**S**	**S**	**S**	**S**	–	**S**	**S**
Below 45 (<45)	·	·	·	·	·	·	·	–	·
Above 85 (>85)	·	·	·	·	·	·	·	·	–

Table 1. Serious errors (S) by true and assigned lane, as scored by the analysis code (`serious()` in `uspstf_cce/analyze.py`). UTD up to date, FU follow-up colonoscopy, OR obtain records, SD shared decision, NAR not average risk, Sx symptoms.

2.2 Synthetic charts, wording authors and study rounds

A generator draws a patient archetype, true events and dates, weighted toward guideline boundaries (30% of ages are 44, 45, 75, 76, 85 or 86 with the birthday near the visit, and 30% of dated tests fall at the edge of their interval). It renders each event as a document and sometimes degrades fax pages with OCR-style noise. The lane is computed from the true events, never from the text.

The study ran in four rounds, each under a protocol amendment committed before its data existed (Table 2). Readers are numbered in training order and data sets by the amendment that created them (train6, dev6 and cal6 come from amendment 6); only the readers and sets used in this paper are listed.

Round	Protocol	Reader	Training wording added	Hand-off threshold chosen on	Test set	Test wording written by
1	protocol, amendments 1-4	r1, r3	the project's templates (r1); a lexicon and 3,709 paraphrases written by Flash (r3)	validation	test2 (2,000)	held-out lexicon and layouts of the project's templates
2	amendment 5	r3 (frozen), r4 (exploratory)	none	r3's validation threshold, unchanged	test3 (5,000)	DeepSeek-V4-Pro, GPT-6-astra
3, main evaluation	amendment 6	r5	train6 (DeepSeek-V4-Pro, GPT-6-astra)	cal6 (10,000; Gemma, Muse, Nemotron)	test4 (10,000)	GLM-5.3, Kimi K3, MiniMax M3, Mistral Large 4
4, follow-up	amendment 7	r6	train7 (the four test4 authors)	cal6, reused	test5 (10,000)	Gemma 4 31B, Muse Spark, Nemotron 3 Ultra

Table 2. Study rounds. Every reader's temperatures were fitted on the validation set (2,000 charts of held-out templates), which shares no charts with cal6 or any test set.

Document wording was written by language models from a fixed brief, one wording family (template) per author, kind of evidence and style; in this paper an "author" is the language model that wrote the wording. Every document passed mechanical checks; every label was audited by the author of this paper under written rules; each author model blind-labelled another author model's documents as a cross-check (`docs/DATA_CARD.md`). Four sets were built for this round (Table 3).

Set	Size	Wording written by	Purpose
train6	1,392 items	DeepSeek-V4-Pro, GPT-6-astra (the test3 authors)	new training wording for earlier error categories
dev6	5,000	the same two; each training run leaves one out	choosing the training configuration
cal6	10,000	Gemma, Muse, Nemotron	selecting the hand-off policy, read once
test4	10,000	GLM-5.3, Kimi K3, MiniMax M3, Mistral Large 4	the test; read blind, analysed once

Table 3. Data sets of protocol amendment 6.

test4's 10,000 charts contain 22,028 items rendered from 838 audited documents in 284 wording families. The blind cross-check agreed with the audit on 833 of the 838; in the five disagreements the audit label stood under the written rules, and no audit label was changed after the cross-check. Its items share 33.2% of their word 4-grams with r5's training text, and no test4 document occurs verbatim in training. The calibration authors were chosen after a pilot because they broke more of the brief's code checks; the protocol records that the direction of any effect on the hand-off thresholds is unknown.

2.3 The reader

For each item the reader sees the document and answers two questions as a single letter: Q1, which of 17 kinds of evidence it shows (for example normal colonoscopy, adenoma found, incomplete colonoscopy, negative or positive FIT, negative or positive stool DNA, normal or abnormal CT colonography or sigmoidoscopy, ordered but not done, patient report, high-risk history, symptoms, a non-qualifying test, irrelevant); and Q2, for tests, which of the date

strings code found in the text is the test date (or the document's own date, or none). The generator re-renders any document whose true date did not survive the simulated fax noise, so the true date is always among the options; real faxes with unreadable dates are not represented. Answer probabilities are read from the first-token distribution over the option letters and temperature-scaled on the validation set (Table 2) [12].

r5 is Qwen3.5-0.8B [36] with a rank-16 LoRA adapter [37], trained for one epoch on the earlier rounds' template data, paraphrases written and filtered by Flash, and the new training wording (Table 4). Its configuration was chosen by a pre-ordered development loop on dev6: four candidate configurations, each trained twice with one training author's new wording left out and scored on that author's dev6 charts. The selected configuration uses the last training step rather than the lowest validation loss.

Setting	Value
Base model	Qwen/Qwen3.5-0.8B
Adapter	LoRA, rank 16
Trainable parameters	10,223,616
Training examples	56,921
Batch size	16
Learning rate	0.0002
Warm-up steps	50
Epochs	1.0
Steps	3,558
Training seed	0
Checkpoint	`runs/r5/step-003558` (last saved step (amendment 6 round 3))
Temperatures (fitted on validation NLL)	Q1 1.25, Q2 2.05

Table 4. r5's training and calibration configuration.

Comparators: r3, the previous round's reader; Flash, a much larger local model (Qwen3.8-Flash-Next with 4-bit floating-point weights, NVFP4) reading the same prompts zero-shot, which also wrote r5's paraphrase training data, so the comparison with it is not fully independent; r1, the first adapter, trained on the original templates only; a keyword baseline; and, as a secondary arm, a 2B model (Qwen3.5-2B) trained by the same recipe. r5 and the 2B arm were each trained with three seeds; seed 0 of each was fixed in advance as the system model.

2.4 Decision stability and the hand-off

Decision stability is computed by code from the reader's probabilities. For each item, every alternative reading (kind crossed with date) is run through the rules with the other items kept at their most likely reading; the probability of the readings that leave the lane unchanged is multiplied across items. A doubt that cannot change the lane (for example, irrelevant versus a non-qualifying test) does not lower the score. The score is an approximation: it ignores changes on several items at once, and it skips alternatives with probability below 1e-4. Its calibration was checked empirically (Table 6, Figure 2), not assumed.

A hand-off policy has a main threshold b, an optional stricter threshold u that applies when the lane would be "up to date", and an optional agreement check with Flash. A chart is

automated when its decision stability reaches the applicable threshold; otherwise it goes to a person. Automated means the lane is accepted without a person reviewing the chart. In a practice the lanes would drive screening outreach, for example a reminder or mailed stool test for "due", or closing the care gap for "up to date"; the system routes charts for outreach and does not make clinical decisions. That workflow was not evaluated. The menu had 42 policies (b and u from 0.9, 0.95, 0.99, 0.995, 0.997 and 0.999, with u above b, with and without requiring Flash's lane to agree). The selection rule, fixed in advance, takes the policy with the most automated calibration charts among those whose exact upper bounds meet both targets with at least 50% coverage; ties go to no Flash, then lower b, then no u. For r5 it selected b = 0.9, u = 0.99.

2.5 Prospective specification and blinding

The protocol and each amendment were committed to the project's git history before the data they govern existed. The commits are not registered with a third party; the full history will be released with the code. Amendment 6 fixed the training candidates and their selection rule, the policy menu and its selection rule, the endpoints, targets and tests, and the failure handling. Every reader read test4 with outputs suppressed except timing. The hashes of the data, audit, adapters, temperatures, policies, calibration readings, the larger model's serving identity and the analysis code were committed (`docs/test4/freeze.json`); the analysis script then refused to run unless every hash matched, and it completed once (after one unchanged restart following a tool timeout). Three language-model reviewers, each prompted as a skeptical referee and combined as a mixture of agents, reviewed the design before it ran and the results after; every number the panel computed was recomputed by script before anything was changed. Analyses added after unblinding are labelled post hoc.

Two pre-specified pieces did not run before unblinding and were run afterwards, labelled as late: the calibration reads of the retrained seeds, and the full timing method (five timed passes after a warm-up and 200 single-chart latencies). The late timing replaces the single-pass measurement made before unblinding (22.44 charts per second; 22.405 late); the late seed reads change no test4 result.

2.6 Endpoints and statistics

P1 is chart accuracy on raw readings with a Wilson 95% interval [38]. P2 holds when, under the selected policy, the exact (Clopper-Pearson) one-sided 95% upper bounds on the error rate and the serious-error rate among automated charts are at most 2% and 0.5%, and at least half the charts are automated [39]. Requiring both bounds is an intersection-union test, so each at 95% keeps the joint claim's error rate at 5%. P3 is end-to-end throughput.

Readers were compared on the same charts with exact McNemar tests [40] (one Holm family [41] for r5: against r3, the larger model, and the keyword baseline on correctness, and against the larger model on serious errors; a separate secondary family for the 2B arm), with paired chart-level bootstrap 95% intervals for the differences (2,000 replicates). Reading quality is reported per question as accuracy, Brier score [42], negative log-likelihood and expected calibration error (ten equal-width bins), and selective-prediction quality as the area under the risk-coverage curve for decision stability [4]. Per-lane precision and recall are given with Wilson intervals. Results by author are descriptive. Each round's primary claim is its own confirmatory family; statements comparing rounds are descriptive.

3. Results

3.1 Primary endpoints

r5 agreed with the generator's labels on 9,878 of 10,000 charts, 98.78% (95% CI 98.55% to 98.98%). Under b = 0.9, u = 0.99 it automated 8,229 charts (82.3%) with 11 errors and 7 serious errors; the exact one-sided 95% upper bounds were 0.22% (target 2%) and 0.16% (target 0.5%). P2 was demonstrated. These bounds treat charts as independent; errors cluster by template and author (sections 3.5 and 3.7), so they describe charts like those tested, not wording from a new author. The 1,771 charts sent to a person contained 110 of r5's 117 serious errors. Timed over five passes, r5 read 22.405 charts per second (P3; 22.44 in the pre-specified single pass, section 2.5).

Table 5 gives the hand-off results for every system and test set. For comparison it includes r3 at the single threshold chosen for it in round 1 (computed after unblinding): on test4 that threshold automated 99.2% of charts with 135 serious errors (upper bound 1.57%), so the earlier design would have missed the target that r5's calibrated two-threshold policy met.

Set	System	Policy	Raw accuracy	Automated	Errors / serious	Upper 95%, errors	Upper 95%, serious	To a person	Targets met
test4	r5 (main evaluation)	b=0.9,u=0.99	98.78%	8,229 (82.3%)	11 / 7	0.22%	0.16%	1,771	yes
test4	r3 at its earlier fixed threshold (late)	b=0.9	98.08%	9,923 (99.2%)	162 / 135	1.86%	1.57%	77	no
test5	r6 (follow-up round)	b=0.9,u=0.99	98.91%	7,876 (78.8%)	5 / 3	0.13%	0.10%	2,124	yes
test5	r5	b=0.9,u=0.99	98.19%	8,897 (89.0%)	14 / 3	0.25%	0.09%	1,103	yes
test2	r5 (regression check)	b=0.9,u=0.99	97.00%	1,405 (70.2%)	17 / 16	1.81%	1.72%	595	no
test2	r6 (regression check)	b=0.9,u=0.99	97.05%	1,532 (76.6%)	45 / 43	3.75%	3.60%	468	no

Table 5. Hand-off results by system and test set: raw chart accuracy, charts automated, automated errors and serious errors, exact one-sided 95% upper bounds (targets 2% and 0.5%), and charts sent to a person. Each system uses the policy selected for it on calibration; r3 uses its earlier fixed threshold. test2 is a regression check.

Against the comparators on raw readings (exact McNemar, Holm-adjusted; Table 6), r5 was more often right than r3 (only r5 right on 115 charts, only r3 on 45; difference +0.70 points, 0.45% to 0.95%, Holm p 6e-08) and than Flash (234 against 119; +1.15 points, 0.79% to 1.52%, Holm p 2.8e-09) and than the keyword baseline (1,744 against 67; +16.77 points, 16.05% to 17.53%; the exact p is below floating-point range). On serious errors without a hand-off, we could not detect a difference from Flash (r5 1.17%, Flash 0.96%; 0.21 points, -0.07% to 0.49%, Holm p 0.16); that comparison reverses when records-needed charts read as up to date are not counted as serious (section 3.11).

3.2 Reading quality and calibration

Reader	Q1 accuracy	Q1 Brier	Q1 NLL	Q1 ECE	Q2 accuracy	Q2 Brier	Q2 NLL	Q2 ECE	Decision-stability ECE	AURC (decision stability)
r5 (0.8B, system)	98.82% (98.67–98.96)	0.0176 (0.0158–0.0195)	0.0389 (0.0354–0.0425)	0.0102 (0.0092–0.0114)	98.67% (98.41–98.92)	0.0022 (0.0014–0.0032)	0.0233 (0.0175–0.0303)	0.0119 (0.0103–0.0143)	0.0114 (0.0098–0.0136)	0.0010 (0.0006–0.0016)
2B arm	97.66% (97.45–97.86)	0.0359 (0.0330–0.0391)	0.0852 (0.0785–0.0923)	0.0122 (0.0106–0.0138)	97.27% (96.90–97.61)	0.0028 (0.0016–0.0041)	0.0300 (0.0199–0.0412)	0.0121 (0.0091–0.0157)	0.0110 (0.0095–0.0140)	0.0016 (0.0013–0.0019)
r3 (previous)	98.04% (97.87–98.22)	0.0384 (0.0349–0.0418)	0.1475 (0.1334–0.1612)	0.0187 (0.0169–0.0204)	97.69% (97.36–98.02)	0.0123 (0.0095–0.0152)	0.0435 (0.0317–0.0570)	0.0122 (0.0095–0.0153)	0.0137 (0.0111–0.0164)	0.0025 (0.0016–0.0035)
Flash (large, zero-shot)	81.65% (81.18–82.10)	0.3252 (0.3171–0.3333)	0.8236 (0.7993–0.8467)	0.1495 (0.1450–0.1541)	97.22% (96.86–97.57)	0.0466 (0.0418–0.0516)	0.0861 (0.0780–0.0938)	0.0103 (0.0073–0.0136)	0.0139 (0.0115–0.0166)	0.0012 (0.0010–0.0015)
r1	91.27% (90.89–91.66)	0.1753 (0.1678–0.1825)	0.9570 (0.9113–1.0009)	0.0827 (0.0789–0.0866)	90.39% (89.72–91.05)	0.1353 (0.1245–0.1467)	1.0416 (0.9469–1.1365)	0.0918 (0.0854–0.0983)	0.0682 (0.0628–0.0735)	0.0453 (0.0406–0.0504)
keyword baseline	84.01% (83.54–84.48)	n/a	n/a	n/a	82.26% (81.36–83.17)	n/a	n/a	n/a	n/a	n/a

Table 6. Reading quality on all test4 items (Q1, Q2) and charts (decision stability), raw readings, with chart-level bootstrap 95% intervals. Q1 and Q2 accuracy are per item. The keyword baseline gives no probabilities, so its calibration columns do not apply. AURC: area under the risk-coverage curve.

r5 had the lowest Brier score and negative log-likelihood on both questions. Its stated Q1 confidence tracked its accuracy (Figure 2, left); Flash, reading zero-shot, was confident and often wrong on Q1 (81.65% per item), but most of its misreads swapped kinds that lead to the same lane, so its chart accuracy (97.63%) was much higher than its item accuracy. Decision stability was close to calibrated for r5, r3, the 2B arm and Flash (Figure 2, right), and r5 had the lowest area under its risk-coverage curve (Figure 3).

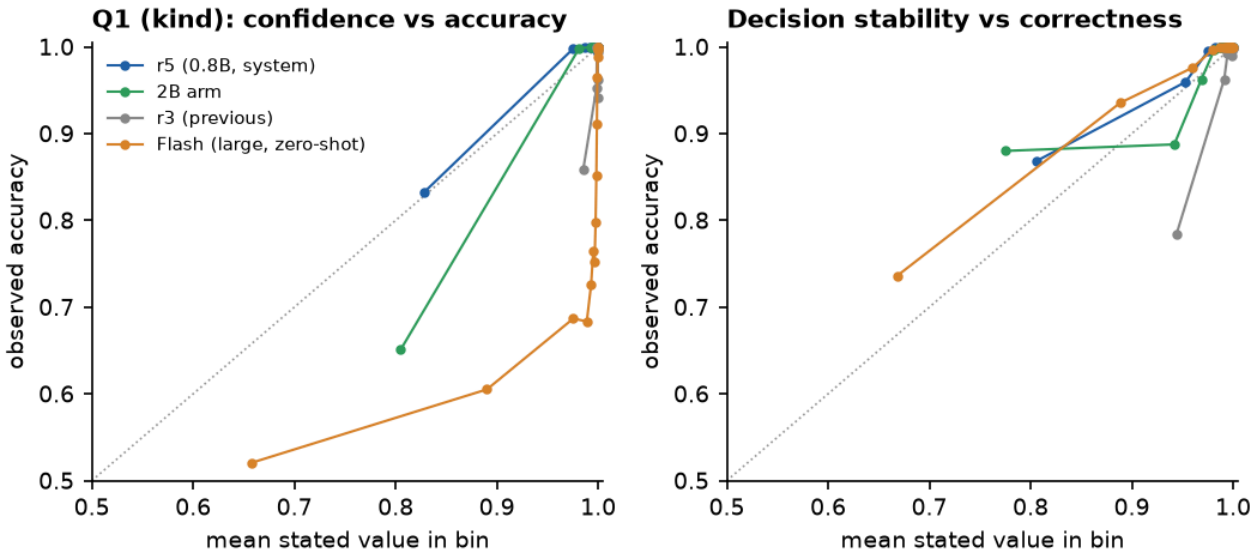

Figure 2. Reliability of Q1 confidence (left, items) and of decision stability (right, charts) on test4, 15 equal-mass bins (the ECE in Table 6 uses ten equal-width bins). Points on the dotted diagonal are calibrated. Most mass lies above 0.99.

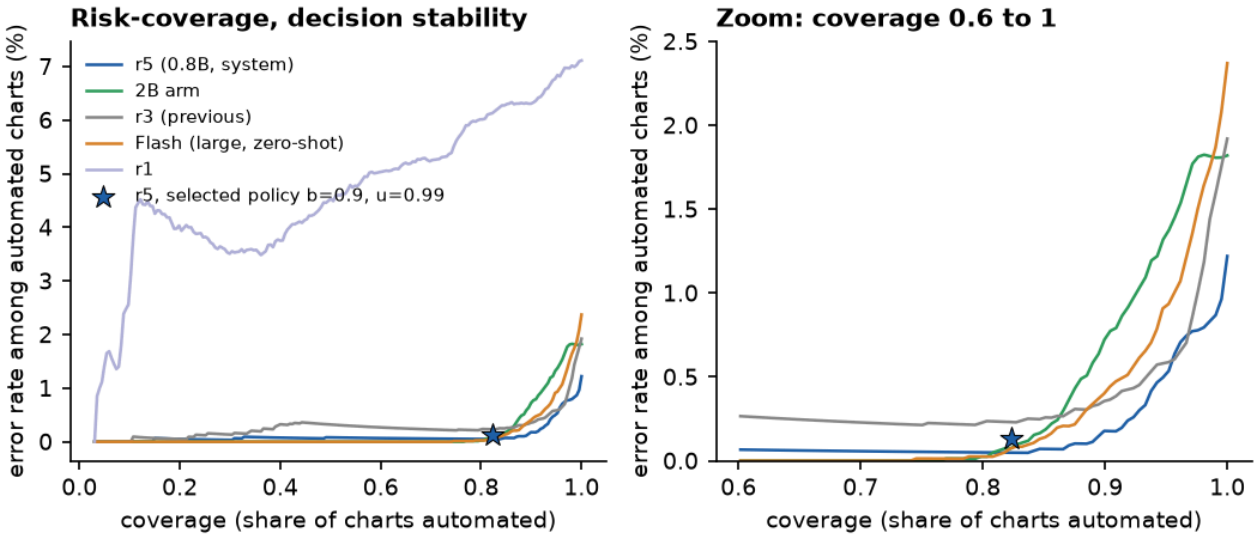

Figure 3. Risk-coverage curves on test4: error rate among automated charts as charts are added in order of decision stability (single threshold). The star is the selected two-threshold policy; because the curve uses a single threshold, the star need not lie on it. Right: coverage 0.6 to 1.

3.3 Lanes

Lane	True charts	r5 recall	r5 precision	Flash recall	Flash precision	2B recall	2B precision
Up to date	2,313	99.96% (99.8–100.0)	95.81% (94.9–96.5)	95.46% (94.5–96.2)	99.55% (99.2–99.8)	99.96% (99.8–100.0)	92.96% (91.9–93.9)
Due	2,152	99.58% (99.2–99.8)	99.40% (99.0–99.6)	98.75% (98.2–99.1)	97.97% (97.3–98.5)	99.91% (99.7–100.0)	100.00% (99.8–100.0)
Follow-up colonoscopy	835	100.00% (99.5–100.0)	99.64% (99.0–99.9)	98.56% (97.5–99.2)	100.00% (99.5–100.0)	100.00% (99.5–100.0)	100.00% (99.5–100.0)
Obtain records	666	85.89% (83.0–88.3)	100.00% (99.3–100.0)	99.25% (98.3–99.7)	85.62% (83.0–87.9)	73.72% (70.3–76.9)	100.00% (99.2–100.0)
Shared decision (76-85)	831	99.76% (99.1–99.9)	99.88% (99.3–100.0)	99.40% (98.6–99.7)	97.64% (96.4–98.5)	99.88% (99.3–100.0)	100.00% (99.5–100.0)
Not average risk	972	99.07% (98.2–99.5)	100.00% (99.6–100.0)	100.00% (99.6–100.0)	97.20% (96.0–98.1)	99.79% (99.3–99.9)	100.00% (99.6–100.0)
Symptoms	510	98.63% (97.2–99.3)	100.00% (99.2–100.0)	84.51% (81.1–87.4)	99.54% (98.3–99.9)	100.00% (99.3–100.0)	98.65% (97.2–99.3)
Below 45	986	100.00% (99.6–100.0)	99.80% (99.3–99.9)	99.80% (99.3–99.9)	98.50% (97.5–99.1)	99.90% (99.4–100.0)	100.00% (99.6–100.0)
Above 85	735	100.00% (99.5–100.0)	99.73% (99.0–99.9)	99.73% (99.0–99.9)	99.05% (98.1–99.5)	100.00% (99.5–100.0)	100.00% (99.5–100.0)

Table 7. Precision and recall by lane on raw readings, all test4 charts.

r5's weakest lane was obtain records (recall 85.89%): 90 records-needed charts were read as up to date, which also lowered up-to-date precision to 95.81% (Figure 4, Figure 5). The larger model's weakest lane was symptoms (recall 84.51%). Every other r5 recall and precision was at least 98.63%.

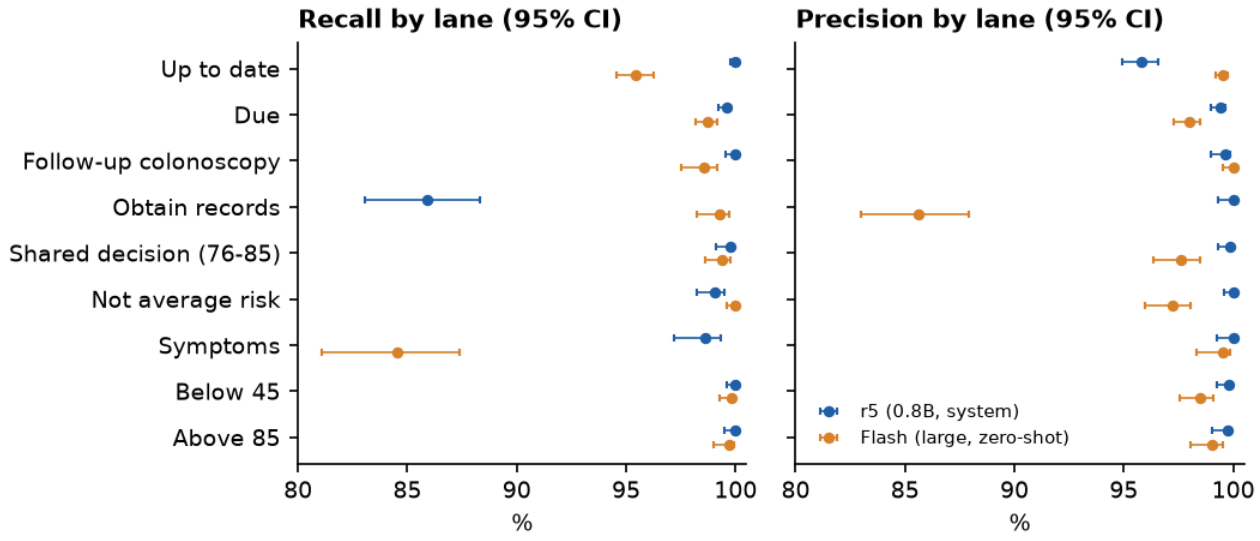

Figure 4. Recall and precision by lane for r5 and Flash, with Wilson 95% intervals.

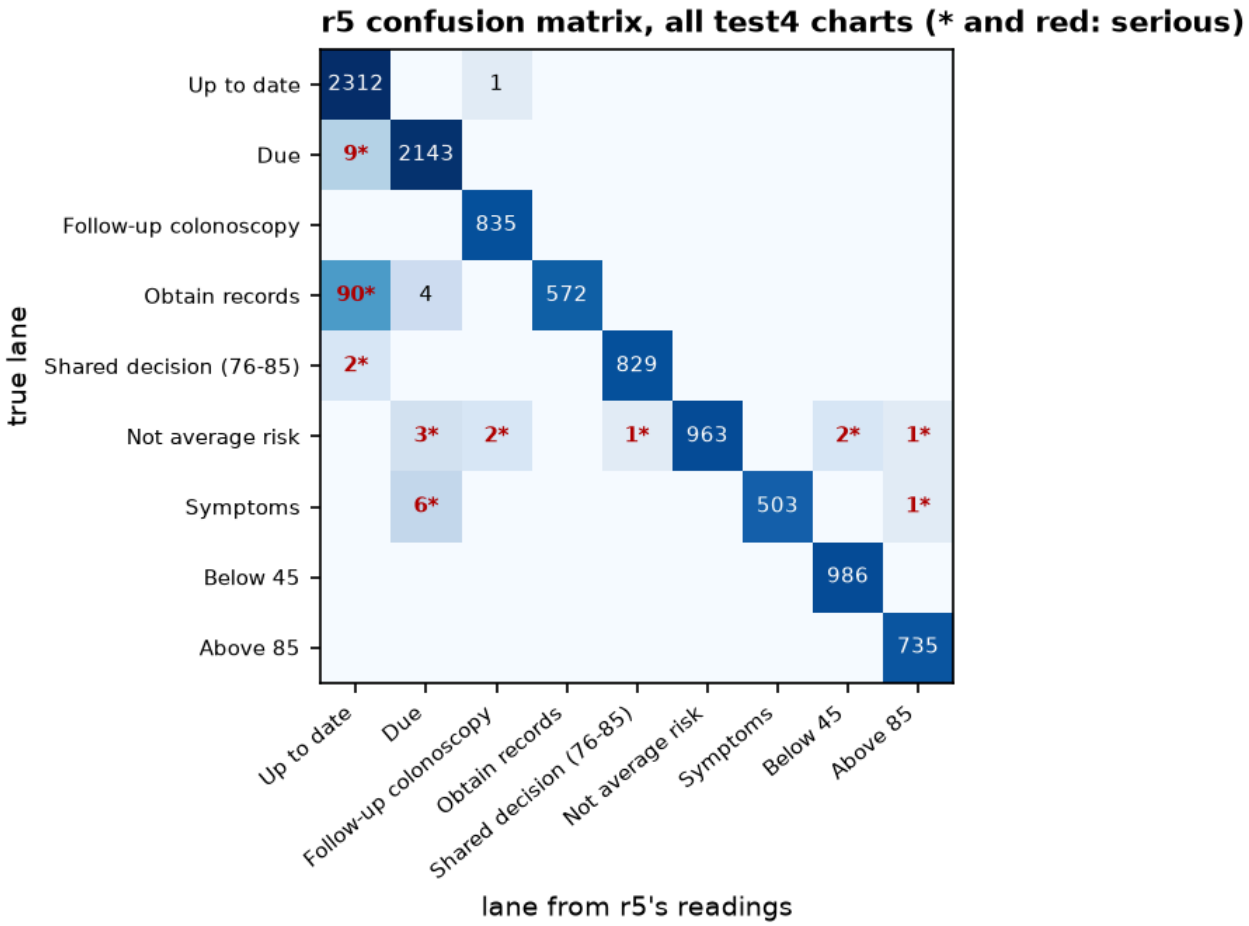

Figure 5. r5's confusion matrix on all test4 charts. Serious errors are marked with an asterisk (and red).

3.4 The hand-off and the policy menu

Figure 6 shows every policy on the menu on calibration and, in hindsight, on test4. The selected policy's calibration bounds (1.44% and 0.27%) are nominal because it was selected on them; test4 is independent of the selection, so its bounds are valid for the one frozen policy. Most policies gave similar results on the two sets; the policies that rely on the main threshold alone did worse on calibration than on test4. In hindsight, b = 0.9 without the up-to-date threshold would also have met the target on test4 (35 serious errors, upper bound 0.49%), but not on calibration (79 serious errors, upper bound 1.05%).

The hand-off is not precise: r5 had read 1,660 of the 1,771 charts it sent to a person correctly (93.7%). How much work those charts make for the person, and how accurately the person reads them, were not measured.

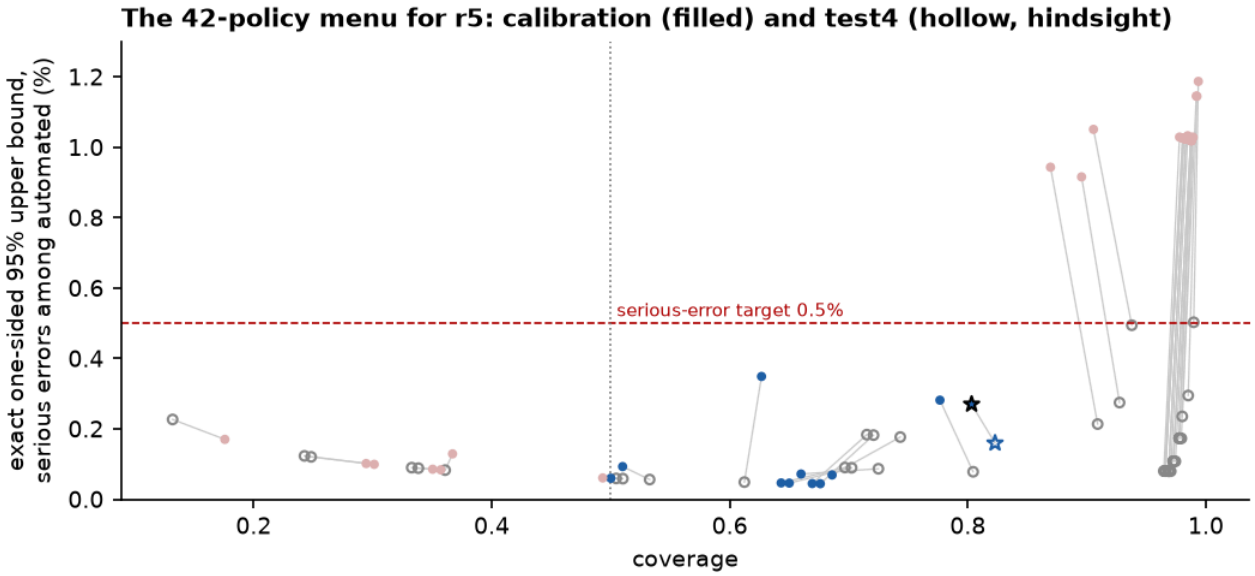

Figure 6. The 42-policy menu for r5: exact upper bound on serious errors among automated charts against coverage, on calibration (filled) and on test4 (hollow, hindsight). Filled blue policies qualified on calibration (both upper bounds within target and at least 50% of charts automated); the plot shows only the serious-error bound. The star marks the selected policy.

3.5 What got through

All 11 automated errors came from two MiniMax templates, and each was a history read as an order.

Seven were care-gap recall reports for patients whose last colonoscopy found an adenoma ("PATIENTS DUE FOR SURVEILLANCE COLONOSCOPY ... Reason flagged: adenoma on screening colonoscopy"). r5 read the adenoma history as an order, so patients who belong in the not-average-risk lane were assigned ordinary screening lanes: due (2, aged 65 and 69), below 45 (2, aged 40 and 44), above 85 (1, aged 94) and follow-up colonoscopy (2, aged 50 and 83). These are the 7 serious errors. We did not grade their clinical consequences, which differ by age and lane: a patient with an adenoma history routed to "due" could be sent a stool test instead of a surveillance colonoscopy, while the two routed to follow-up colonoscopy would reach colonoscopy anyway. Adenoma surveillance intervals [32] are outside USPSTF and outside this encoding, which stops at "not average risk". r5 was fairly confident in these readings (decision stability 0.9100 to 0.9634), and the stricter up-to-date threshold does not apply to them because none was read as up to date.

Four were nurse notes recording a patient's report of a home stool DNA test with no documentation. The correct lane is obtain records; r5 read the report as an order and assigned "due", which is not serious under our taxonomy.

The same kind of failure, a history read as an order, explains r5's regression on the oldest held-out set, test2 (section 3.9).

3.6 Undated letters

Most of r5's wrong charts were date errors. Of the 121 wrong charts that one corrected answer would fix, 87 were dates and 34 were kinds, and all of them came from nine wording

families. The date errors came from one letter style per author: a letter reporting a normal colonoscopy whose only date is the date of the letter itself (for example "This letter follows up on your recent procedure"). Our encoding treats the colonoscopy as undated and the lane as obtain records; r5 often took the letter's date as the exam date and called the patient up to date. After the analysis the author reviewed the four letter styles (`docs/letter-date-examples.md`) and confirmed the rule: the date of a letter is not the date of the exam. Under our encoding these are reading errors; the author's clinical ruling is that calling such a patient up to date is not clinically serious (section 3.11). r5 was inconsistent on them, reading 43 such items as undated and 103 with the letter's date, and sometimes reading two copies of the same template differently. The hand-off sent all 90 records-needed charts that r5 read as up to date to a person, because none reached decision stability 0.99.

3.7 Results by author

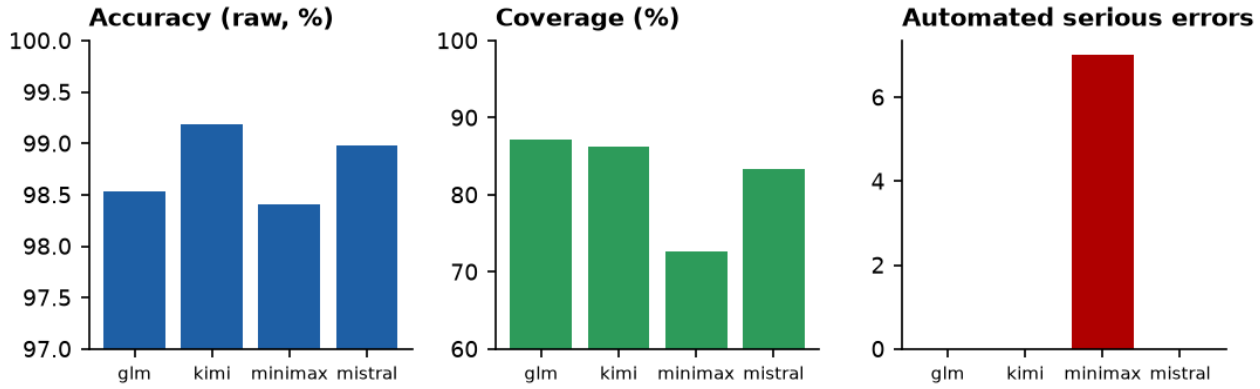

Figure 7. r5 by test4 author: raw accuracy, coverage under the selected policy, and automated serious errors.

All 7 automated serious errors were on MiniMax's charts: 0.38% of its automated charts (exact one-sided upper bound 0.72%). Its charts also had the lowest coverage (72.6%). GLM, Kimi and Mistral had none among 2,140, 2,133 and 2,128 automated charts (one-sided upper bound 0.14% each). With MiniMax's charts removed, 6,401 charts were automated with no serious error. Results by author are descriptive; no per-author pass or fail was pre-specified. Leaving out any one of GLM, Kimi or Mistral, the upper bound on serious errors was 0.22%; leaving out MiniMax, 0.05% (post hoc). With one of four authors responsible for every automated serious error, the share of new authors whose charts would produce such errors cannot usefully be estimated (one of four: exact 95% CI 0.6% to 80.6%). On calibration the same pattern appeared: 13 of r5's 14 automated serious errors were on one author's charts.

3.8 Training seeds and per-model calibration

Under seed 0's policy, r5's seed 1 automated 9,391 charts with 3 serious errors, but seed 2 automated 9,867 with 41 (upper bound 0.54%) and would have failed. The seeds' calibration reads, run late, explain this (Figure 8, Table 8): seed 0's policy failed on calibration for every retrained seed, so calibration would have caught seed 2, and the same pre-specified rule selected a stricter policy for each. Under its own policy each of the four retrained seeds had no automated serious errors on test4 (r5 seed 2: 0 among 9,405 automated). Retraining with the same recipe was enough to change which thresholds met the calibration criterion.

Run	Seed 0's policy on calibration: automated, serious	Its own selected policy	On test4: automated, serious
r5 seed 1	9,231, 41 (fails)	b=0.9, u=0.995	8,490, 0
r5 seed 2	9,877, 87 (fails)	b=0.9, u=0.999	9,405, 0
2B seed 1	8,770, 50 (fails)	b=0.95, u=0.99	8,959, 0
2B seed 2	9,525, 51 (fails)	b=0.9, u=0.995	8,449, 0

Table 8. Retrained seeds: seed 0's policy on calibration, and each seed's own calibrated policy on test4 (late reads, after unblinding).

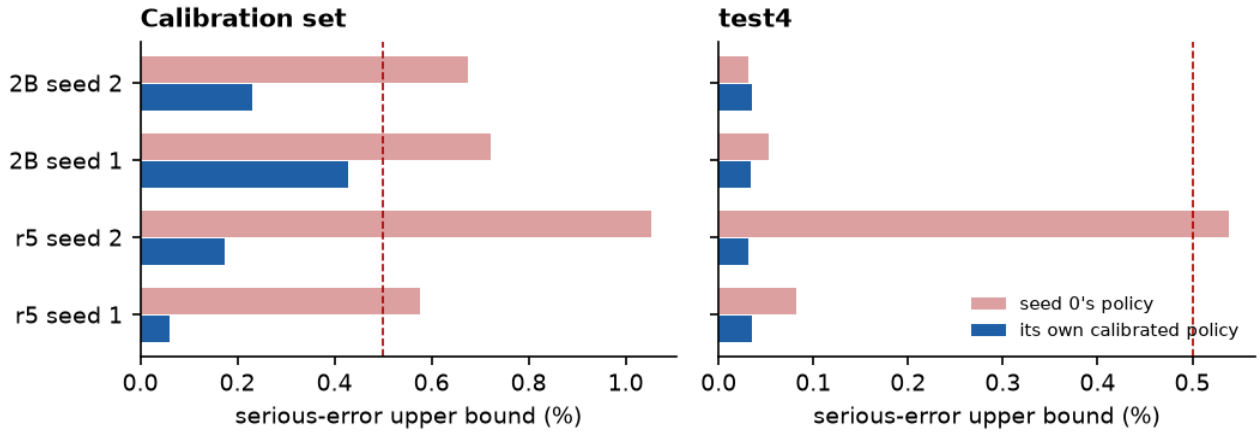

Figure 8. Upper bound on serious errors among automated charts for each retrained seed, under seed 0's policy and under its own calibrated policy, on calibration and test4. Dashed: the 0.5% target.

3.9 Secondary results

The 2B arm, under its own selected policy (also b = 0.9, u = 0.99), automated 8,379 charts with 1 serious error (upper bound 0.06%). Its seed-0 raw readings were less safe than the larger model's (1.77% against 0.96% serious, Holm p 2.7e-06), its other two seeds read better than any r5 seed (99.11% and 99.48%), and on calibration its selected policy automated more serious errors than r5's (19 against 14). We treat it as a candidate for a later round.

On the oldest held-out set (test2, held-out wording by the original authors), r5 scored 97.00% (r3: 98.9%) and under its policy automated 1,405 charts with 16 serious errors (upper bound 1.72%), so the criterion was not met on that older wording. Fifteen of those were high-risk histories in one template read as orders.

3.10 Speed

Reader	Run	Charts	Timed passes	Charts/s	SD	Median latency (ms)	p90 latency (ms)
r5	`p3-test4-r5`	10,000	5	22.405	0.02	59.9	90.5
2B arm	`p3-test4-r5-2b`	10,000	5	13.542	0.007	91.3	145.7
r3	`p3-test4-r3`	10,000	5	22.383	0.017	59.0	91.5
Flash (300-chart pass, 200 latencies)	`p3-flash-test4`	300	1	1.218	–	1187.4	1695.0
Flash (10,000-chart pass)	`flash-test4`	10,000	1	1.199	–	–	–

Table 9. End-to-end throughput (prompts, date extraction, model calls, rules and decision stability) on test4. The small models ran on one NVIDIA GB10 with vLLM [43]; Flash ran as deployed for interactive use (max-num-seqs 4, MTP speculative decoding). Ratios depend on these serving configurations.

The selected policy does not use Flash, so the system reads at r5's speed, about 19 times the larger model's throughput as deployed, with a median single-chart latency of 59.9 ms against 1187.4 ms. These are machine figures and exclude human review. With Flash in a batch serving configuration, the ratio on test2's charts was 13.8 times (19.9 as deployed), so the ratio depends heavily on how the larger model is served.

3.11 Sensitivity: records-needed charts read as up to date

The pre-specified taxonomy counts a records-needed chart called up to date as serious; the author's later ruling is that it is not (section 2.1). Reclassifying that one pair (post hoc; `results/paper/sensitivity-records-up-to-date.md`) changed no automated count: r5 still automated 7 serious errors on test4, r6 and r5 3 each on test5, and r5 and r6 16 and 43 on test2, because no chart of this kind was automated. It did change the raw comparisons, which have no hand-off. On test4, 90 of r5's 117 raw serious errors were this pair; reclassified, r5's raw serious-error rate was 0.27% against the larger model's 0.91% (exact McNemar p 2.7e-09, unadjusted). On test5, r6's rate fell from 0.98% to 0.26%, against Flash's 0.24% (p 0.87) and r5's 0.35% (p 0.26). The raw serious-error comparisons in this paper therefore rest on this one convention; the primary endpoints do not.

4. Follow-up round: training on the failures (amendment 7)

4.1 What changed

Amendment 7 was written and reviewed before any of its wording or models existed. Training wording for seven categories, each paired with the kind it was confused with, was written by the four test4 authors (care-gap and recall reports, letters after a procedure, patient-reported tests, high-risk history in structured notes, and three smaller ones). A new label rule (A8, the author's ruling) states that a document's own date does not date a test it mentions, except a document written on the procedure day that says so. The new reader r6 is r5's recipe on the enlarged data, chosen in one development round scored on wording left out of training; there r6's serious-error rate was 0.18% against r5's 2.16% on the same charts. Every trained model, including each seed, was calibrated on its own on the existing calibration set.

The one unused author family that was available failed the pilot (6 of 24 documents passed the code checks), and to avoid paid authors, test5 (10,000 charts) was written by the three calibration authors in new documents, enriched with letters and care-gap reports; it is a challenge set, and it tests new documents from authors the system had already seen, not a new author. The pre-specified primary comparison with r5 was chart correctness on the target stratum S (charts with an undated mention in a letter or registry, an adenoma or high-risk history in a registry, or a patient report), chart correctness on all charts, and the automated-serious burden (automated serious errors per incoming chart, each system under its own policy). The 2B arm was dropped before training. The freeze was written after the blind reads rather than before them as the amendment ordered; every frozen file other than the readings was already fixed when the reads began.

4.2 Results

Measure	r6	r5
Chart accuracy, all charts	98.91%	98.19%
Raw serious-error rate	0.98%	1.59%
Accuracy, stratum S (1,223 charts)	93.46%	88.14%
Accuracy, M1 (322 charts)	83.54%	81.37%
Accuracy, M2 (213 charts)	100.00%	100.00%
Accuracy, M3 (688 charts)	96.08%	87.65%
Hand-off policy	b=0.9, u=0.99	b=0.9, u=0.99
Automated	7,876 (78.8%)	8,897 (89.0%)
Automated errors / serious	5 / 3	14 / 3
Upper bound, serious (target 0.5%)	0.10%	0.09%

Table 10. r6 and r5 on test5 (raw readings and each under its own policy). M1: an undated later mention of a qualifying test in a letter or registry style; M2: an adenoma or high-risk history in a registry style; M3: a patient-reported test.

Test	Only r6	Only r5	Holm p	Difference, r6 minus r5 (95% CI)
(a) chart correctness, stratum S	85	20	3e-10	5.31% (3.76% to 6.95%)
(b) chart correctness, all charts	115	43	1.8e-08	0.72% (0.48% to 0.96%)
(c) automated-serious burden, all charts	2	2	1	+0.0 per 10,000 (-4.0 per 10,000 to +4.0 per 10,000)

Table 11. The pre-specified comparison with r5 (exact McNemar; chart-level bootstrap intervals; the burden per 10,000 incoming charts).

r6 met P2 (7,876 charts automated, 3 serious errors, upper bound 0.10%), and so did r5 under its frozen policy. r6 read the target stratum better (+5.31 points), almost entirely through patient-reported tests; undated mentions improved little, and the adenoma recall reports that drove test4's errors were read correctly by both models, so test5 could not test that fix. No difference in automated serious errors was detected: each system automated 3, with 1 chart in common, and none of these six errors (on five charts) was in the target stratum. r6 automated fewer charts, sending 2,124 to a person against r5's 1,103. Two of r6's three automated serious errors came from a pathology fax family whose label the audit had marked borderline and the cross-check had disputed; both readings are serious under either

label. All three of r6's seeds met P2 under their own policies, though seed 1 would also have passed under seed 0's.

Training also shifted errors to new places. r6 read 93 orders as irrelevant (r5 none), 14 normal sigmoidoscopies as normal colonoscopies and 7 abnormal sigmoidoscopies as normal. On the oldest held-out wording (test2) its policy automated 43 serious errors (upper bound 3.60%, r5 16), mostly symptoms in structured notes read as irrelevant or as an order; on test4 it automated 10 against r5's 7.

5. Discussion

The design met its pre-specified error bounds on charts from four authors whose wording the reader had not been trained on, and each automated error traced to one of two document templates. Those bounds describe charts like the ones tested; they cannot say how the system would do on a new author's wording.

First, calibration is per model. Seed 0's policy failed on calibration for all four models retrained with other seeds, and on test4 for one of the two r5 ones; with each calibrated on its own, none of the four retrained models had an automated serious error on test4. Any change to the reader, including retraining on the same data, needs its own calibration before use.

Second, the hand-off caught a systematic reading failure. The undated-letter errors were common, but r5 was rarely sure of them, so the up-to-date threshold sent all of them to a person. The hand-off can only catch errors the reader is unsure of, though. The seven automated serious errors were confident misreadings of an unfamiliar document type, and a higher threshold on up-to-date calls could not have caught them. Catching them in advance needs either training on that document type or a fresh calibration set that contains it.

Third, because code applies the rules, the analysis could say which question on which document went wrong, and the fixes were concrete: teach the reader that a recall or care-gap report describes history, and that a letter's date is not an exam date. These are conventions about documents rather than general reading skill; a decision model trained on one health system's records also gained most over open models on that system's own conventions [22].

Fourth, the follow-up round improved reading on the categories it targeted, but it did not show a reduction in automated serious errors (3 against 3). The previous model already met the target on the new set, the new model's calibrated policy handed off more charts, and both models regressed on the oldest held-out wording (test2). Each round passed the set it was built against. Because the follow-up round's calibration and test sets were written by the same three authors, whose calibration errors had informed two of its training categories, it could not show whether the change generalises. That design choice, more than the training objective, is what should change next.

5.1 Limitations

The charts are synthetic, written by language models from our brief; real documents are messier and more varied. The labels follow our encoding of USPSTF 2021 and our conventions (for example, family history without a known syndrome does not change the lane, and an undated report goes to obtain records); no independent clinician has reviewed the

rules or the labels, and the undated-letter rule was confirmed by the author. The generator and the rules share one encoding, so agreement with the labels measures reading, not whether the encoding is clinically right; the label audit was done by the author, and cross-checks by other language models are not independent validation. The serious-error taxonomy is unvalidated, and one convention in it accounts for most raw serious errors (section 3.11). Clinicians may also disagree with other conventions: stool DNA counts for three years, the long end of USPSTF's one to three; family history of colorectal cancer without a known syndrome stays average risk, as in USPSTF, though other guidelines screen such patients earlier; adenoma history stops at "not average risk" without surveillance intervals; and conflicting documents over time, and comorbidity in the shared-decision lane, are not modelled. Lane proportions come from the generator, which is weighted toward boundaries, so per-lane precision does not describe any real population. Errors cluster by template and author, so intervals over charts understate the uncertainty for new wording; the family-level bootstrap we pre-specified failed its own transfer check (for r5 its interval, 98.94% to 99.83%, excluded its own point estimate of 98.78%) and is not used. Four authors cannot support an interval for a new author. The reader answers two fixed questions it was trained for; unlike a general-purpose decision model [22], it cannot take a new question without new training, and each topic needs its own. The calibration authors were chosen for breaking more of the brief's rules. r5's training categories came from the earlier test sets' errors, so its scores on those sets are optimistic, and Flash wrote and filtered r5's paraphrase training data, so the comparison with Flash is not fully independent. In the follow-up round, calibration and test shared their authors, and those authors' calibration errors informed two training categories; the freeze followed the reads (with every input already fixed); and the 2B arm was dropped. Sending a chart to a person was only simulated: nobody reviewed those charts, so the accuracy and workload of that step were not measured. The panel reviews were run with language models, and the code and analysis were written with an AI coding assistant (Claude, Anthropic), which also wrote the document templates behind most of the training text; every number in this paper is generated from the run outputs and checked by script.

5.2 Next steps

The next round should separate the hand-off policy and the test from the error-mining loop: calibration and test sets that share no authors or wording families with each other and are never a source of training categories, with every spent wording set kept as a fixed regression check. Within that design, cost-sensitive training (weighting items whose misreading would be serious) is the first candidate objective, with GRPO a later arm judged on calibration. The same design will be applied to other USPSTF screening topics, with the author split built in from the start. An open general-purpose decision model of the same size, reading the same questions without training and calibrated under the same selection rule, will show what training on the topic adds [20]. Independent clinician review of the encoding and labels, and de-identified real documents, remain the main steps toward any clinical claim.

6. Conclusion

On 10,000 synthetic charts written by four language-model authors whose wording was not used in development, a 0.8B reader with rules code and a pre-specified hand-off agreed with the generator's labels on 98.78% of charts and automated 82.3% of them with 7 serious

errors (upper bound 0.16%), meeting its pre-specified targets at about 19 times the throughput of a much larger model as deployed. The bounds hold for charts like those tested, not for a new author. Each automated error came from one of two templates and traced to a single misread document. Hand-off thresholds had to be calibrated per trained model. A follow-up round that retrained the reader on those failures read them better and again met the targets, but did not reduce automated serious errors, automated fewer charts and regressed on the oldest held-out wording.

Data, code and reproducibility

The paper's home is https://www.clinical-si.com. The code and data will be released later as a tagged version, announced there; until then they are available from the author on request (brian@clinical-si.com). The release will contain:

- the protocol and every amendment with the full commit history (`docs/PROTOCOL.md`, `docs/amendment6/`, `docs/amendment7/`), the decision logs, the recorded deviations and the freeze manifests. Before release the history was rewritten only to replace the author's local file paths, machine names and personal email address; `docs/history-rewrite.md` maps every original commit hash to its published one and lists the five frozen files whose hashes changed as a result, with both hashes;
- the chart generator, the rules engine with its boundary tests, the serious-error definition and the data card;
- every synthetic chart, the wording libraries, the generation briefs and the authors' raw replies, and the audits, cross-checks and dropped documents;
- the training code and configurations, and the r5 and r6 adapters (LoRA weights on Qwen3.5-0.8B, CC BY-NC 4.0, for research only and not for clinical use), byte-identical to the frozen adapters;
- a credits file naming every model that wrote training, calibration or test text, how it was accessed, and the license or terms it was used under;
- the frozen per-item readings, probabilities and hand-off decisions of the 34 runs that the confirmatory analyses, figures and sensitivity analyses read: the scored test4 and test5 readers, the calibration readings that set the hand-off thresholds, the throughput timing runs, r5 and r6 on test2 and test3, and r6 on test4. Each frozen reading is byte-identical to the hash recorded before it was scored. The test4 and test5 readings before temperature scaling, the validation readings that fitted the temperatures, the development ablations and the other calibration variants are available from the author on request;
- the scripts that rebuild every table, figure, interval and test in this paper from those frozen outputs (`tools/paper_figures.py`, `tools/paper_sensitivity.py` and the analysis scripts), with the generated results (`docs/RESULTS.md`, `results/`) and the interpretation (`docs/FINDINGS.md`);
- the review panels' requests and reports, and our responses (`docs/reviews/`).

Reproduction works at three levels. Re-running the analysis from the frozen outputs is exact and needs no GPU; `tools/release_verify.py` checks every released output against the freeze manifests and re-runs each frozen analysis at its own commit. Retraining needs a GPU (we

used one NVIDIA GB10) and reproduces the models only approximately. The wording cannot be regenerated: hosted language models wrote it and they change over time, so the release keeps their raw replies, and the sampled paraphrase file is archived with its SHA-256. Each chart's patient, events, dates and label are regenerable from its seed. All documents and labels are synthetic and contain no patient data.

References

1. US Preventive Services Task Force. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. JAMA. 2021;325(19):1965-1977.
2. Chow C. On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory. 1970;16(1):41-46.
3. El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. Journal of Machine Learning Research. 2010;11:1605-1641.
4. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. Advances in Neural Information Processing Systems (NeurIPS). 2017.
5. Geifman Y, El-Yaniv R. SelectiveNet: a deep neural network with an integrated reject option. Proceedings of the International Conference on Machine Learning (ICML), PMLR. 2019;97:2151-2159.
6. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. International Conference on Learning Representations (ICLR). 2017.
7. Madras D, Pitassi T, Zemel R. Predict responsibly: improving fairness and accuracy by learning to defer. Advances in Neural Information Processing Systems 31 (NeurIPS). 2018.
8. Mozannar H, Sontag D. Consistent estimators for learning to defer to an expert. Proceedings of the International Conference on Machine Learning (ICML), PMLR. 2020;119:7076-7087.
9. Bates S, Angelopoulos A, Lei L, Malik J, Jordan M. Distribution-free, risk-controlling prediction sets. Journal of the ACM. 2021;68(6):1-34.
10. Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. Conformal risk control. International Conference on Learning Representations (ICLR). 2024:55198-55218.
11. Angelopoulos AN, Bates S, Candès EJ, Jordan MI, Lei L. Learn then test: calibrating predictive algorithms to achieve risk control. The Annals of Applied Statistics. 2025;19(2):1641-1662.
12. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML). 2017.
13. Desai S, Durrett G. Calibration of pre-trained transformers. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:295-302.
14. Jiang Z, Araki J, Ding H, Neubig G. How can we know when language models know? On the calibration of language models for question answering. Transactions of the Association for Computational Linguistics. 2021;9:962-977.
15. Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, et al. Language models (mostly) know what they know. arXiv:2207.05221. 2022.
16. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems 32 (NeurIPS). 2019.
17. Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D. Measuring calibration in deep learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2019:38-41.
18. Kumar A, Liang P, Ma T. Verified uncertainty calibration. Advances in Neural Information Processing Systems 32 (NeurIPS). 2019.

19. Almeida D. Introducing System One models & Jev. TypeSafe AI blog. September 15, 2026. https://typesafe.ai/blog/introducing-system-one-models-and-jev (accessed October 21, 2026).
20. vLLM Semantic Router Team. Decision 2.0 (Kai-0.6B, Eos-0.8B, Sol-2B, Nox-4B, Lux-9B, Vega-27B) [model collection]. Hugging Face. 2026. https://huggingface.co/collections/vllm-sr/decision-20-6ab7cf7bdfb506bf8269cb00 (accessed October 21, 2026).
21. Madrid-García A, Merino-Barbancho B. Jev in medicine: a benchmark evaluation. arXiv:2609.34024. 2026.
22. Zhang A, Vuong TTL, Wagner SJ, Bang D, Ding T, Su Y, et al. A general-purpose System One decision model for clinical records. arXiv:2610.27099. 2026.
23. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. Journal of the American Medical Informatics Association. 2010;17(4):383-388.
24. Harkema H, Chapman WW, Saul M, Dellon ES, Schoen RE, Mehrotra A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. Journal of the American Medical Informatics Association. 2011;18(Suppl 1):i150-i156.
25. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clinical Gastroenterology and Hepatology. 2013;11(6):689-694.
26. Raju GS, Lum PJ, Slack RS, Thirumurthi S, Lynch PM, Miller E, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. Gastrointestinal Endoscopy. 2015;82(3):512-519.
27. Gupta S, Halm EA, Rockey DC, Hammons M, Koch M, Carter E, et al. Comparative effectiveness of fecal immunochemical test outreach, colonoscopy outreach, and usual care for boosting colorectal cancer screening among the underserved: a randomized clinical trial. JAMA Internal Medicine. 2013;173(18):1725-1732.
28. Singal AG, Gupta S, Skinner CS, Ahn C, Santini NO, Agrawal D, et al. Effect of colonoscopy outreach vs fecal immunochemical test outreach on colorectal cancer screening completion: a randomized clinical trial. JAMA. 2017;318(9):806-815.
29. National Committee for Quality Assurance. Colorectal Cancer Screening (COL, COL-E). HEDIS measure. https://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality-report/colorectal-cancer-screening-col-e/ (accessed October 20, 2026).
30. Rex DK, Boland CR, Dominitz JA, Giardiello FM, Johnson DA, Kaltenbach T, et al. Colorectal cancer screening: recommendations for physicians and patients from the U.S. Multi-Society Task Force on Colorectal Cancer. Gastroenterology. 2017;153(1):307-323.
31. Wolf AMD, Fontham ETH, Church TR, Flowers CR, Guerra CE, LaMonte SJ, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. CA: A Cancer Journal for Clinicians. 2018;68(4):250-281.
32. Gupta S, Lieberman D, Anderson JC, Burke CA, Dominitz JA, Kaltenbach T, et al. Recommendations for follow-up after colonoscopy and polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology. 2020;158(4):1131-1153.e5.
33. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
34. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nature Medicine. 2025;31(1):60-69.
35. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904.
36. Qwen Team. Qwen3.5-0.8B [model card]. Hugging Face. 2026. https://huggingface.co/Qwen/Qwen3.5-0.8B (accessed October 20, 2026).

37. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. International Conference on Learning Representations (ICLR). 2022.
38. Wilson EB. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association. 1927;22(158):209-212.
39. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404-413.
40. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-157.
41. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6(2):65-70.
42. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950;78(1):1-3.
43. Kwon W, Li Z, Zhuang S, Sheng Y, Zheng L, Yu CH, et al. Efficient memory management for large language model serving with PagedAttention. Proceedings of the 29th Symposium on Operating Systems Principles (SOSP). 2023:611-626.