

Dissociating Answerability from Extraction in LLM Question Answering over Radiology Reports

Priyanshu Sinha
Labcorp, Durham, NC
sinhap3@labcorp.com

Kelsey McDonald
Labcorp, Durham, NC
kelsey.mcdonald2@labcorp.com

Abstract

Answering questions from radiology reports involves two decisions: whether the report contains an answer and which span gives it. A false answer presents an unsupported claim as fact; an over-abstention hides what the report says. Yet RadQA results are reported as a single F1 that merges both. We propose an evaluation method that separates them, reporting false-answer and over-abstention rates beside F1 and splitting F1 changes into extraction and answerability terms. We apply it to 13 LLMs on the RadQA test set, zero-shot and with three cumulative demonstration sets, the last adding an unanswerable example. Models with similar F1 differed more than twofold in false-answer rate. The unanswerable example changed when models answered, not how well they extracted: 12 of 13 models abstained more often, and three-quarters of the answers they withdrew had $F1 \geq 0.5$. Within models, false answers showed no clear association with extraction quality, but acting on the decision itself helped: keeping an answer only when enough other models also answered lowered the zero-shot false-answer rate by 4 points without a detectable F1 change. Because the two decisions can change independently, we suggest radiology QA evaluations report them separately, not through F1 alone.

1 Introduction

Large language models (LLMs) perform strongly on medical licensing-exam questions (Singhal et al., 2023; Nori et al., 2023) and medical evidence summarization (Tang et al., 2023), and are increasingly applied to patient records, from answering clinicians’ questions about them (Fleming et al., 2024) to extracting information from radiology reports (Le Guellec et al., 2024). Radiology reports are a natural target for LLM question answering: they are produced in large volumes, use dense domain terminology, and are read by ordering physicians who need accurate answers from them. Exam

benchmarks test medical knowledge, not the reading of a specific report, where the answer must be found in a section such as Findings or Impression and is sometimes not there at all.

Extractive question answering (QA) over reports is a controlled version of this setting: the answer is a verbatim span of the report, or the model declines because it decides the question is unanswerable. In RadQA (Soni et al., 2022), physicians wrote questions from the clinical referral of MIMIC-III radiology reports (Johnson et al., 2016a,b) without seeing the report text, so a quarter of its test questions have no answer in the report. Such a system makes two decisions per question: whether the report answers it, and if so, which span does. The two fail in clinically distinct ways. A *false answer* to an unanswerable question puts a statement the report does not support in front of a clinician; an *over-abstention* withholds information the report contains and sends the clinician back to the full text.

The standard metric, SQuAD 2.0 F1 (Rajpurkar et al., 2018), folds both decisions into one number: a correct abstention scores 1, a wrong answer/abstain decision scores 0, and otherwise the score is token overlap between words in a predicted answer and the ground-truth answer. Models with the same F1 can therefore give very different numbers of false answers, and a change in F1 between prompts can come from better spans or from a moved abstention boundary. SQuAD 2.0’s evaluation does also report F1 separately on answerable and unanswerable questions (HasAns/NoAns), and NoAns F1 is exactly one minus the false-answer rate. But RadQA results are reported only as overall F1 and EM (Soni et al., 2022; Al Nahian and Kavuluru, 2025). Even HasAns F1 mixes the two decisions: it equals the share of answerable questions answered times the mean F1 of those answers, so it changes when either one does.

The distinction matters most for in-context

demonstrations, the obvious intervention that needs no training. On RadQA, LLMs still trail task-specific fine-tuning: RadQA-DPO (Al Nahian and Kavuluru, 2025) reaches 77.48 test F1 and 55.15 exact match (EM), and the two annotators reach 78.72 and 90.31 F1 (Soni et al., 2022). In-context demonstrations (Brown et al., 2020) could narrow this gap if they improve extraction, but not if they mainly move the answer/abstain boundary. Studies of demonstration choice, order and format report effects on aggregate accuracy (Zhao et al., 2021; Lu et al., 2022; Min et al., 2022), which cannot tell which decision moved.

We propose an evaluation method that separates the two decisions, reporting the false-answer rate (FAR) and over-abstention rate (OAR) beside span F1, and apply it to 13 LLMs from six families on the RadQA test set, zero-shot and with three cumulative demonstration sets (an answerable Findings question, an answerable Impression question, and an unanswerable question). Our contributions:

1. **A decomposition of $\Delta F1$ that isolates extraction, with a selectivity test (§3.4).** Any change in F1 splits into an extraction term, computed on answerable questions answered in both conditions, and an answerability shift. A companion test asks whether new abstentions remove weak answers or good ones. Unlike a change in HasAns F1, the extraction term contains no abstentions, and both tools work with any span metric. FAR and OAR follow from the standard HasAns/NoAns breakdown, so we offer them as a reporting recommendation rather than as new metrics.
2. **The tested unanswerable demonstration moves the answer decision, not extraction (§4.2).** It raises OAR in 12 of 13 models. Of the resulting -0.0126 change in F1, -0.0127 is answerability shift, and extraction is flat under both token and RadGraph F1 (Jain et al., 2021). For answerable demonstrations, the split depends on the span metric.
3. **The added abstentions are not detectably selective (§4.3).** Three-quarters of the 272 withdrawn answers had $F1 \geq 0.5$. In exchange, models gave a net 55 fewer false answers, most of them from one model.
4. **Similar F1, different false-answer rates (§4.1, §4.4; Figure 1).** Eleven of 13 models lie within 0.669–0.738 zero-shot F1, while FAR ranges

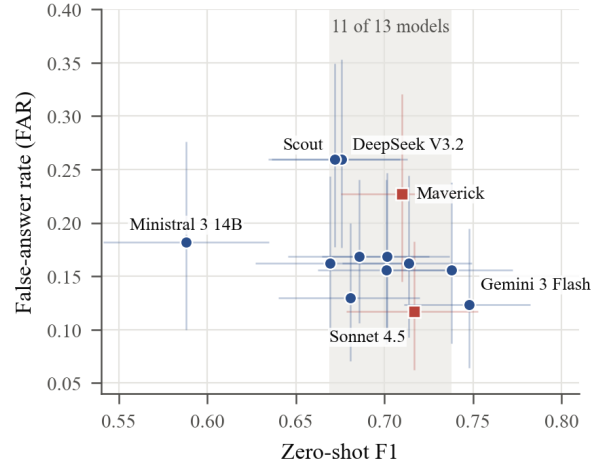


Figure 1: Zero-shot span F1 against false-answer rate (FAR) for the 13 LLMs, with report-clustered 95% CIs. The shaded band holds 11 models within 0.07 F1 that differ more than twofold in FAR. Squares mark the example pair discussed in §4.1: Claude Sonnet 4.5, the lowest-FAR model, and LLaMA 4 Maverick, the highest-F1 model whose FAR is significantly higher than Sonnet 4.5’s. They differ by 0.007 in F1 but 0.110 in FAR. Values for unlabelled models are in Table 1.

from 0.117 to 0.260. The cross-model F1–FAR correlation is largely built into the metric, and within a model we find no association between extraction quality and false answers.

5. **Evaluating a mitigation that targets the answer decision (§4.5).** A consensus gate, cross-fitted over reports, lowers zero-shot FAR by 4 points with no detectable F1 change.

We also find that no model matches the exact match of the fine-tuned RadQA-DPO, and none reaches inter-annotator agreement (§4.1). Our code is available at <https://github.com/pri2si17-1997/rad-qa-llm-eval>.

2 Related Work

Clinical and radiology QA. Domain-adapted encoders such as BioBERT, ClinicalBERT and RadBERT advanced extraction and QA over clinical text (Lee et al., 2019; Alsentzer et al., 2019; Huang et al., 2020; Yan et al., 2022), and radiology-specific models followed (Ghosh et al., 2023; Ranjit et al., 2024). emrQA (Pampari et al., 2018) and RadQA (Soni et al., 2022) provide extractive QA over clinical notes and radiology reports. In RadQA, questions were written from the clinical referral without seeing the report, so unanswerable questions arise naturally rather than by construction. Fine-tuned RadQA systems have progressed

from BERT baselines (Soni et al., 2022) to RadQA-DPO (Al Nahian and Kavuluru, 2025). General and medical LLMs perform strongly on exam-style medical QA (Singhal et al., 2023, 2025; Nori et al., 2023; Chen et al., 2023; Labrak et al., 2024; Xie et al., 2025), and have been applied to clinical extraction and radiology reporting (Busch et al., 2025; Agrawal et al., 2022). Like prior RadQA work, these evaluations report span F1 and EM; we separate the answer decision from extraction and ask which of the two demonstrations change.

Unanswerable questions and abstention. SQuAD 2.0 (Rajpurkar et al., 2018) added unanswerable questions, and its official evaluation reports F1 separately on answerable and unanswerable questions (HasAns/NoAns). Selective QA lets a model abstain when it is likely to be wrong and evaluates the resulting trade-off between coverage and accuracy (Kamath et al., 2020); related work asks when LLMs decline to answer and whether they know what they know (Wen et al., 2025; Kadavath et al., 2022). This breakdown already exposes false answers, but HasAns F1 scores an over-abstention as 0 and so mixes span quality with abstention; our extraction term excludes abstentions. We also ask the selective-prediction question of a prompting intervention: do the abstentions it adds fall on weak answers?

Sensitivity of in-context learning. In-context learning (Brown et al., 2020) is sensitive to the choice, order and label balance of demonstrations (Zhao et al., 2021; Lu et al., 2022). Min et al. (2022) find that demonstrations act largely through the label space, input distribution and format they expose, rather than through correct input-label mappings. Our results show a related pattern for extractive clinical QA: the unanswerable demonstration changes how often models use the abstain option, not how well they extract.

Evaluation reliability. Comparisons of NLP systems benefit from significance testing that respects the structure of the test data (Dror et al., 2018). For clinical content, RadGraph (Jain et al., 2021) annotates entities and relations in radiology reports (Yu et al., 2023). We use RadGraph F1 as an entity-level complement to token F1 and report-clustered inference throughout.

3 Methods

3.1 Data

We use the RadQA test split (Soni et al., 2022; Soni and Roberts, 2022): 614 questions over 104 MIMIC-III radiology reports. Each answer is either a text span within the Findings or Impression section, or the question is marked unanswerable.

Of the 614 questions, 154 (25.1%) are unanswerable, across 57 reports. The remaining 460 answerable questions are drawn from the Findings section (255 questions) and the Impression section (205 questions). The distributed file has one row per question and annotator answer (894 rows), and 280 answerable questions carry two annotators’ answers. We score one prediction per question (the first of the two rows for doubly annotated questions) and analyze at the question level (614 questions); only token usage (Table 7) is reported per row (894 rows).

3.2 Models and Inference

We evaluate 13 models from six families (Table 3): Claude Opus 5, Sonnet 5, Sonnet 4.5 and Haiku 4.5; Gemini 3.1 Pro and 3 Flash; LLaMA 4 Maverick and Scout; DeepSeek V3.2; Mistral Large 3 and Ministral 3 14B; and GPT-OSS 120B and 20B, through Google Vertex AI (Gemini) and AWS Bedrock (all others).

Each family has its own system prompt, and all share one task specification: return the exact, minimal verbatim span that answers the question, or mark it unanswerable. The output is a JSON object with the extracted spans, a Boolean answerability flag and a short rationale, enforced through tool calling (LLaMA 4 output is parsed from free text). Calls that still fail after three attempts are scored as abstentions (3 of 31,928 predictions), so failures cannot raise FAR.

All models ran with a 4,096-token output limit. All except Claude Opus 5 and Sonnet 5 (11 of 13) ran at temperature 0; those two ran at the provider-default temperature, as user-defined temperature parameters were deprecated. Each model used its provider’s default reasoning and thinking settings.

3.3 Demonstration Conditions

Demonstrations are drawn from the RadQA train split (reports of at most 4,000 characters) and are fixed across all demonstration conditions. We select one example of each type: an answerable Findings question, an answerable Impression question

and an unanswerable question, each paired with a fixed rationale.

The four conditions are cumulative: $k=0$ (none); +F (an answerable Findings question); +F+I (adding an answerable Impression question); and +F+I+U (adding an unanswerable question, placed last). The unanswerable demonstration is a near-miss case: it asks about gallbladder stones in a report describing a common bile duct stone after cholecystectomy, and its rationale states that the information is not mentioned. Every model and test question receives the same demonstrations. Because number and type change together, we attribute effects to the demonstration set, not solely to its size.

3.4 Scoring and Decomposition

Following SQuAD 2.0 (Rajpurkar et al., 2018), a question scores 0 if prediction and ground truth disagree on answerability, 1 if both say unanswerable, and otherwise the maximum token F1 (or EM) over ground truth answers and predicted spans.

A prediction’s answerability is its answerable flag, so a prediction flagged answerable with no span counts as an answer: it scores 0 on an answerable question and is a false answer on an unanswerable one.

RadGraph F1 (Jain et al., 2021) measures the overlap of clinical entities and relations between the predicted and reference answer spans, with answerability treated the same way and the maximum taken over ground truth answers. Unlike token F1, it scores a multi-span prediction as one string, with the spans joined, rather than crediting the best-matching span; about 6% of answers to answerable questions contain more than one span. It is among the automatic metrics that agree best with radiologists’ error counts on generated reports (Yu et al., 2023). We use it as an entity-level complement to token F1 that is less sensitive to span boundaries, and report each decomposition under both metrics to check whether conclusions hold for clinical content as well as for token overlap.

Answer-decision metrics. The *false-answer rate* (FAR) is the share of the 154 unanswerable questions a model answers; the *over-abstention rate* (OAR) is the share of the 460 answerable questions it declines. Both are measured against the physician annotators’ answerability labels and describe the answer/abstain decision; they are not measures of clinical harm.

Decomposition of $\Delta F1$. For one model and conditions $A \rightarrow B$, let s_q^A and s_q^B be the scores of question q , and let $\mathcal{E}$ be the answerable questions the model answers in both conditions. With $d_q = s_q^B - s_q^A$ and $n = 614$ questions,

$$\Delta = \underbrace{\frac{1}{n} \sum_{q \in \mathcal{E}} d_q}_{\text{extraction}} + \underbrace{\frac{1}{n} \sum_{q \notin \mathcal{E}} d_q}_{\text{answerability shift}}. \quad (1)$$

The extraction term captures span-level changes for questions where the model’s answer decision stays the same. The answerability shift term captures all cases where the decision itself changes, such as a new abstention, a newly answered question, or a flipped decision on an unanswerable question.

Selectivity of new abstentions. Let $\mathcal{N}$ be the answerable questions answered in A but declined in B ; each contributes $-s_q^A/n$ to the answerability term. We compare the mean prior score s_q^A over $\mathcal{N}$ with that over $\mathcal{E}$. If newly abstained questions had lower prior scores, the abstentions are *selective*: they remove weak answers; if prior scores were higher, then they removed valid answers.

Within-model predictors of false answers. To ask whether a model’s false answers follow from weak extraction, we use the 52 reports that contain both answerable and unanswerable questions (126 unanswerable questions). For each model and report, *extraction quality* is the model’s mean F1 on the report’s answerable questions it answered, and *answering propensity* is the fraction of them it answered. We relate both to whether the model answers each of the report’s unanswerable questions, controlling for report difficulty (the other 12 models’ mean F1 on the report; Appendix B).

Other measures. Because SQuAD F1 credits correct abstentions, $F1 = (\sum_{q \text{ answerable}} s_q + 154(1 - FAR))/614$, so its correlation with FAR is partly built in. We therefore also use *answerable F1* (mean over the 460 answerable questions, abstentions scored 0) and *extraction F1* (mean over the answerable questions the model answered). In SQuAD 2.0’s breakdown, NoAns $F1 = 1 - FAR$ and HasAns F1 (our answerable $F1$) = $(1 - OAR) \times \text{extraction F1}$. An error taxonomy assigns each question to exact match, EM-mismatch ($F1 \geq 0.5$, $EM = 0$), partial ($0.1 \leq F1 < 0.5$), no overlap ($F1 < 0.1$), correct abstention, over-abstention or false answer. Published human, BERT-MIMIC and RadQA-DPO results

(Soni et al., 2022; Al Nahian and Kavuluru, 2025) report token F1 and EM only, so comparisons with them are descriptive. For a like-for-like human reference under all three metrics, we score models and annotators against single annotators on the 280 doubly annotated questions.

3.5 Consensus Gate

To test whether the answer decision can be improved without changing extraction, we gate each model’s answers with the other models’ decisions. For a model in a given condition, let v_q be the number of the other 12 models, in the same condition, that answer question q . The gate keeps the model’s span if $v_q \geq t$ and otherwise replaces it with an abstention, scored 1 if q is unanswerable and 0 otherwise. Abstentions are never overridden and kept spans are unchanged, so the gate can only lower FAR and raise OAR. At $t = 0$ the model is ungated; at $t = 12$ an answer is kept only if all 12 other models also answer. Because we have predictions only for the test split, t is cross-fitted: we split the 104 reports at random into halves, choose t on one half by maximizing span F1, and measure the change from the ungated model on the other. t is chosen per model (own- t) or as one value for all models (pooled- t). We repeat this over 200 splits and report the mean model-mean change, with the 2.5–97.5% range over splits. A two-model gate instead keeps an answer only if one partner model also answers; we report the median over the 156 ordered pairs.

3.6 Statistical Analysis

Inference is clustered by report (104 clusters). Confidence intervals are report-level cluster-bootstrap intervals (10,000 replicates; 2,000 for decomposition terms). Per-model condition effects use Wilcoxon signed-rank tests on paired report means, with Benjamini–Hochberg (BH) correction (Benjamini and Hochberg, 1995) across models. For FAR, we use Cochran’s Q across models and, against the lowest-FAR model, exact McNemar and report-level permutation tests. Selectivity and within-model predictors of false answers use generalized estimating equations (GEE) (Liang and Zeger, 1986) with exchangeable within-report correlation. Across the 13 models, we report Spearman ρ with bootstrap CIs and leave-one-model-out checks, and interpret these descriptively. Full specifications are in Appendix B.

4 Results

4.1 Zero-Shot Performance

Similar F1, different false-answer rates. Zero-shot F1 ranges from 0.588 to 0.748, and 11 of 13 models lie between 0.669 and 0.738, with overlapping CIs for the top ten (Table 1). FAR ranges from 0.117 to 0.260, and the models differ overall (Cochran’s $Q = 69.2$, $df = 12$, $p < 10^{-5}$). Against the lowest-FAR model, Claude Sonnet 4.5 (0.117), three models are significantly higher under both the McNemar and the cluster permutation test: LLaMA 4 Maverick (0.227; $q = 0.0001$ and 0.004), DeepSeek V3.2 (0.260; $q = 0.0002$ and 0.011) and LLaMA 4 Scout (0.260; $q < 0.0001$); all others have $q \geq 0.06$. Maverick thus has the fifth-highest F1 but one of the three highest FARs (Figure 1). OAR ranges from 0.057 to 0.202.

Where F1 is lost. Pooled over models at $k=0$, 19.5% of questions are exact matches, 35.4% EM-mismatches, 9.4% partial, 1.9% no overlap, 20.7% correct abstentions, 8.7% over-abstentions and 4.4% false answers. 84.9% of EM-mismatches share clinical entities with the correct answer (RadGraph F1 > 0), whereas only 2.1% of the 95 non-empty no-overlap answers do. Answers that miss the correct answer entirely are therefore rare. Most lost F1 comes from span boundaries and from the answer decision.

Gap to fine-tuning and to humans. No model reaches RadQA-DPO’s F1 (0.775) in any condition; only Gemini 3 Flash at $k=0$ has a CI that includes it, and all 52 model $\times$ condition EM CIs lie below its 0.552. On the 280 doubly annotated questions, every model is below inter-annotator agreement in every condition under all three metrics (BH $q < 0.0001$). At $k=0$, models score a mean token F1 of 0.756, EM of 0.318 and RadGraph F1 of 0.625 against single annotators; the annotators score 0.935, 0.746 and 0.858 against each other. Models thus reach 81% of human agreement under token F1 but only 73% under RadGraph F1, so the shortfall is not only a matter of span boundaries.

4.2 What Demonstrations Change

Small changes in F1. Model-mean F1 is 0.692 at $k=0$, 0.698 at +F, 0.706 at +F+I and 0.693 at +F+I+U (per-model values in Table 4). Answerable demonstrations raise F1 in 10 of 13 models (Wilcoxon over model means, $p = 0.048$), but only

Model	F1 [95% CI]	EM	RadGraph F1	FAR [95% CI]	OAR
Gemini 3 Flash	0.748 [0.711, 0.782]	0.482	0.673	0.123 [0.064, 0.195]	0.100
Gemini 3.1 Pro	0.738 [0.702, 0.772]	0.451	0.654	0.156 [0.087, 0.238]	0.083
Claude Sonnet 4.5	0.717 [0.678, 0.753]	0.436	0.659	0.117 [0.062, 0.182]	0.128
Claude Opus 5	0.714 [0.676, 0.749]	0.386	0.628	0.162 [0.093, 0.245]	0.057
LLaMA 4 Maverick	0.710 [0.675, 0.742]	0.381	0.620	0.227 [0.145, 0.321]	0.061
Claude Sonnet 5	0.701 [0.665, 0.737]	0.388	0.623	0.169 [0.101, 0.247]	0.096
Claude Haiku 4.5	0.701 [0.662, 0.738]	0.409	0.639	0.156 [0.086, 0.240]	0.135
Mistral Large 3	0.686 [0.645, 0.725]	0.410	0.614	0.169 [0.107, 0.240]	0.133
GPT-OSS 120B	0.681 [0.640, 0.720]	0.383	0.602	0.130 [0.071, 0.200]	0.150
DeepSeek V3.2	0.676 [0.636, 0.713]	0.389	0.618	0.260 [0.177, 0.353]	0.130
LLaMA 4 Scout	0.672 [0.634, 0.709]	0.375	0.585	0.260 [0.178, 0.349]	0.100
GPT-OSS 20B	0.669 [0.627, 0.709]	0.376	0.585	0.162 [0.096, 0.244]	0.141
Ministral 3 14B	0.588 [0.541, 0.635]	0.358	0.535	0.182 [0.100, 0.276]	0.202
<i>Published references (RadQA test set; not re-run)</i>					
Human, best annotator ^a	0.903	0.814	—	—	—
Human, mean of two annotators ^a	0.845	0.754	—	—	—
RadQA-DPO, Flan-T5-3B ^b	0.775	0.552	—	—	—
BERT-MIMIC ^a	0.636	0.494	—	—	—

Table 1: Zero-shot ($k=0$) results on the RadQA test set (614 questions). CIs are report-clustered bootstrap intervals (10,000 replicates). FAR: share of the 154 unanswerable questions answered; OAR: share of the 460 answerable questions declined. References: ^aSoni et al. (2022); ^bAl Nahian and Kavuluru (2025). Published references use their authors’ scoring and are not directly comparable (§3.4).

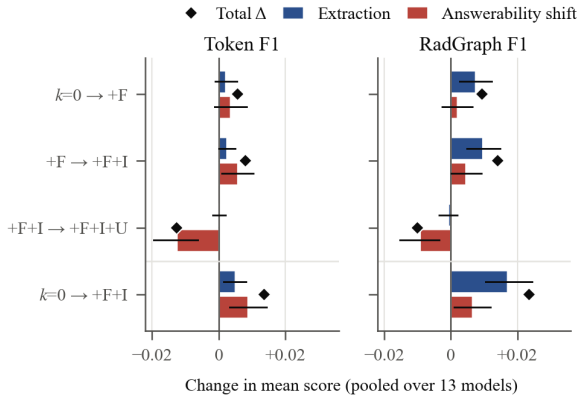


Figure 2: Decomposition of the change in mean score between conditions (Eq. 1), pooled over 13 models, with 95% report-cluster bootstrap CIs. Bars sum to the total, shown as ♦. Exact values are in Table 5.

DeepSeek V3.2 and Ministral 3 14B improve significantly under report-clustered testing. Adding the unanswerable demonstration lowers F1 in 12 of 13 models ($p = 0.0034$; significant for four models), raises mean OAR from 0.105 to 0.142 and lowers mean FAR from 0.171 to 0.143.

The unanswerable demonstration moves the answer decision. Of the -0.0126 token-F1 drop from $+F+I$ to $+F+I+U$, -0.0127 is answerability shift and $+0.0001$ extraction (Figure 2). Extraction does not change significantly in any of the 13 models (BH $q \geq 0.48$), while the answerability shift is negative in 12 of the 13. On answerable

questions, the demonstration adds 272 abstentions. On unanswerable questions, it adds 18 false answers and removes 73, a net reduction of 55. Under RadGraph F1 the picture is the same: -0.0093 of the -0.0101 drop is answerability shift, and extraction is flat ($-0.0008 [-0.0034, +0.0019]$; BH $q \geq 0.40$ in every model).

For answerable demonstrations, the split depends on the metric. Under token F1, 64% of the $+0.0136$ gain from $k=0$ to $+F+I$ is answerability shift, and per model, extraction improves significantly only for Ministral 3 14B. Under RadGraph F1, 73% of the $+0.0235$ gain is extraction, positive in all 13 models and significant in 6. The two metrics rank models almost identically (ρ 0.87–0.93) and disagree only on spans answered in both conditions ($+0.0075$ token vs. $+0.0265$ RadGraph F1). The span convention explains little of this: scoring token F1 on the joined prediction, as RadGraph F1 does (§3.4), raises its extraction term only to $+0.0063 [+0.0030, +0.0096]$, or 41% of its gain. Without radiologist adjudication we cannot say which better reflects clinical correctness, so we do not claim that answerable demonstrations act mainly through answerability.

4.3 Which Answers Do New Abstentions Remove?

The unanswerable demonstration is not detectably selective. Its 272 new abstentions had a

Transition	$\mathcal{N}$	Prior F1		Diff.	p	≥ 0.5
		$\mathcal{N}$	$\mathcal{E}$			
$k=0 \rightarrow +F$	134	0.618	0.736	-0.099 [-0.170, -0.028]	0.0065	67%
$+F \rightarrow +F+I$	96	0.676	0.738	-0.042 [-0.123, +0.038]	0.30	68%
$+F+I \rightarrow +F+I+U$	272	0.701	0.745	-0.035 [-0.084, +0.014]	0.16	76%
$k=0 \rightarrow +F+I$	144	0.604	0.737	-0.121 [-0.196, -0.047]	0.0014	61%

Table 2: Selectivity of new abstentions on answerable questions, pooled over 13 models. $\mathcal{N}$: questions newly declined; $\mathcal{E}$: questions answered in both conditions. Prior F1 is F1 in the earlier condition. Diff. (95% CI below) is from a GEE of prior F1 on a newly-declined indicator with model fixed effects and report clustering. ≥ 0.5 : share of $\mathcal{N}$ with prior F1 ≥ 0.5 .

mean prior F1 of 0.701, against 0.745 for questions still answered (Table 2). The CI of the difference includes zero, although it cannot exclude modest selectivity. This holds when any single model is left out (leave-one-out $p \geq 0.08$), under RadGraph F1 (-0.047 [-0.107, +0.013]) and in every model separately (all $q > 0.24$). Three-quarters had F1 ≥ 0.5 , so its F1 cost comes mostly from withdrawing good answers.

Answerable-demonstration selectivity comes from one model. From $k=0$ to $+F+I$, newly declined questions had a prior F1 0.12 lower than those still answered, but this rests on Ministral 3 14B (22 new abstentions with prior F1 0.083 vs. 0.676 kept; $q < 0.0001$). Without it, the pooled difference is -0.042 ($p = 0.27$), and LLaMA 4 Maverick goes the other way (+0.146, $q = 0.017$). RadGraph F1 agrees on the pooled result. We therefore treat this selectivity as model-specific.

4.4 How False Answers Relate to Accuracy

Across models (Table 8). At $k=0$, F1 and FAR correlate ($\rho = -0.59$ [-0.86, -0.06]), but largely because F1 credits correct abstentions. The FAR range alone accounts for 0.036 of the 0.160 F1 range. For 7 of the 13 models, leaving that one model out makes the correlation non-significant, and with answerable F1 in place of F1 it is only -0.28 ($p = 0.35$). Extraction F1 and FAR are only weakly associated ($\rho -0.43$ to -0.59 , significant only at $+F$). With demonstrations, models with lower FAR instead have higher OAR ($\rho -0.58$ to -0.72 , against -0.08 at $k=0$; robust to leaving out any model at $+F+I$ and $+F+I+U$). This is consistent with models differing mainly in how readily

they answer. With demonstrations, no F1-FAR correlation is distinguishable from zero ($\rho -0.23$ to $+0.25$), so F1 rank is not a dependable guide to false answers.

Within models (Table 9). On reports with both answerable and unanswerable questions, a model’s extraction quality on the answerable ones is not associated with a false answer on the unanswerable ones in any condition (OR per SD 0.94–1.04; upper CI bounds ≤ 1.18). With about four answerable questions per report, this measure is noisy, which biases the association toward null, so we cannot exclude a modest effect. A model’s answering propensity on the report is nominally associated with false answers (OR 1.21 at $k=0$, 1.31 at $+F+I$; BH $q = 0.076$), which we treat as suggestive.

4.5 Consensus Gating of the Answer Decision

With t chosen on held-in reports (§3.5), at $k=0$ the gate lowers mean FAR by 0.040 [0.026, 0.057] on held-out reports and raises OAR by 0.010, with no detectable F1 change (+0.005 [-0.000, +0.009]; Table 10). The effect is concentrated in high-FAR models (DeepSeek V3.2 -0.144, LLaMA 4 Scout -0.099, Ministral 3 14B -0.076, LLaMA 4 Maverick -0.063). It is smallest at $+F+I+U$ (-0.011), where the unanswerable demonstration has already moved models toward abstention. Two-model gates do worse: the median of the 156 ordered pairs lowers FAR by 0.049 but costs 0.013 F1.

5 Discussion

Two errors with different consequences, and F1 separates neither. A false answer puts a statement the report does not support in front of the reader; an over-abstention sends the reader back to the full report. How much each costs depends on the use. A system that pre-fills structured fields for later review (Adams et al., 2023) and one that answers a clinician’s question about a specific patient (Fleming et al., 2024) weigh them differently, and neither weighting is a property of the benchmark. When models are compared or selected for clinical QA, FAR and OAR at the operating point carry information that F1 rank does not.

The unanswerable demonstration moved the answer threshold. The largest and most consistent demonstration effect was on the answer decision. Pooled over models, it withdrew about three good answers (F1 ≥ 0.5) for each false answer it removed (§4.3). The reduction is uneven: LLaMA 4

Scout accounts for 34 of the net 55 fewer false answers, and FAR fell in only 7 of the 13 models while OAR rose in 12 of the 13. Whether that exchange is acceptable depends on the relative cost of the two errors, which only the clinical setting can fix. Prompt construction thus sets an operating point implicitly. For clinical QA, we suggest setting it explicitly against a stated FAR target and re-checking it with the decomposition and selectivity test whenever the prompt changes. With one demonstration set, this is evidence about this demonstration only. Testing several unanswerable demonstrations, from clean absences to near misses, each with its own rationale and with count-matched answerable controls, would show whether the threshold shift generalizes.

Better extraction may not reduce false answers.

Within a model, we found no association between extraction quality on a report and whether the model answers that report’s unanswerable questions, although the per-report measure is noisy and modest effects cannot be excluded. This suggests that improving extraction alone is unlikely to reduce false answers much; as a within-model result, it does not by itself show how models with different F1 differ in false answers. The answer decision may therefore need its own control, such as a verification or consensus step or a separately calibrated answerability model (Kamath et al., 2020; Wen et al., 2025). Our gate shows that information from outside the model can lower FAR without a detectable F1 change, mostly in the highest-FAR models. It shows that the answer decision can be improved independently of extraction; it is not a deployment recipe. In practice, such a gate would need a calibration set separate from the evaluation data, a FAR target chosen by the system’s clinical owners, recalibration on local data, and 13 inference calls per question.

Where the remaining gap lies. Demonstrations changed model-mean F1 by at most 0.014, while the gap to RadQA-DPO and to the annotators lies in the spans themselves: exact boundaries and, under RadGraph F1, the clinical entities extracted. We offer this as a hypothesis about what span-level supervision supplies and a few demonstrations do not. Where the exact extracted text matters, the models evaluated here do not yet substitute for a fine-tuned extractor or human review.

Reporting recommendations. For extractive clinical QA with unanswerable questions, we recommend reporting (i) FAR and OAR beside span F1; (ii) an F1 restricted to answerable questions; (iii) the decomposition of Eq. 1, with the selectivity of new abstentions, when comparing prompts or models; and (iv) at least two span metrics, since token and RadGraph F1 disagree on whether answerable demonstrations improve extraction, a disagreement that comes mostly from what the metrics measure rather than from their span conventions (§4.2). This matters when a prompt or model change is approved for deployment: under token F1 alone the answerable demonstrations barely improve extraction, whereas RadGraph F1 shows a gain in all 13 models, significant in 6. The approving metric should match how the output is used (entity-level agreement when findings populate structured fields, token F1 and EM when the extracted text is shown to clinicians), and since neither has been checked against radiologists on short spans, local clinical review is still needed.

6 Conclusion

The overall F1 reported for RadQA folds two decisions into one number; we separated them with an extraction term restricted to unchanged answer decisions and a test of which answers new abstentions remove. On RadQA, the unanswerable demonstration we tested moved the answer decision rather than extraction: it raised over-abstention in 12 of 13 LLMs, and three-quarters of the answers it withdrew had $F1 \geq 0.5$. At similar zero-shot F1, models differed more than twofold in false-answer rate, and within a model, we found no association between extraction quality and false answers. Evaluations of clinical QA systems should therefore report FAR and OAR beside span F1, and check any prompt or model change with the decomposition and selectivity test before deployment is considered.

Limitations

No clinical adjudication. FAR and OAR measure agreement with the physician annotators’ answerability labels, not whether an answer would mislead a clinician. Token and RadGraph F1 are automatic measures of span agreement, and RadGraph F1 was validated against radiologists on full reports, not short spans. The labels themselves are imperfect (Soni et al., 2022), so some questions counted as unanswerable may be answerable, or

the reverse. No radiologist reviewed correctness or clinical consequence. A stratified adjudication of EM mismatches, token/RadGraph disagreements, false answers and newly induced abstentions is the missing validation.

One dataset and one demonstration set. All results come from the RadQA test split: one institution, one report style, and 154 unanswerable questions in 57 reports. All few-shot results use one fixed demonstration set with a single, near-miss unanswerable example whose generic rationale does not describe the near miss (§3.3). Because each condition adds a new demonstration type, we cannot separate “a third demonstration” from “an unanswerable demonstration”, or this example from unanswerable examples in general.

Thirteen models. Cross-model correlations rest on 13 points, and several are sensitive to single models (§4.4).

Consensus gate. The gate threshold is cross-fitted within the test set (§3.5), not set on a separate calibration split, and the gate is built from and evaluated on the same 13 models.

External comparisons. Human, BERT-MIMIC and RadQA-DPO results are taken from their papers. Their test answers, like ours, keep both annotators’ answers where both answer the question (Soni et al., 2022), and RadQA-DPO is scored with the SQuAD evaluation script (Al Nahian and Kavuluru, 2025). We additionally credit the best of several predicted spans, which is slightly more lenient to the LLMs, and Soni et al. (2022) do not state the reference against which each annotator was scored.

Contamination. RadQA is distributed only under the PhysioNet Credentialed Health Data License, which prohibits redistribution, but the training data of most evaluated models is undisclosed and we did not test for contamination. Comparisons between conditions are less affected, because all conditions share the same test items.

Ethical Considerations

Data. RadQA (Soni and Roberts, 2022) is derived from MIMIC-III (Johnson et al., 2016a,b), which contains de-identified clinical records. All authors accessed the data through PhysioNet (Goldberger et al., 2000) credentialed access, and we used it under the PhysioNet Credentialed Health Data License and its data use agreement.

Responsible use of LLM services. We sent MIMIC-derived report text to third-party LLM services only in accordance with PhysioNet’s policy on the use of MIMIC data with large language models and online services (PhysioNet, 2023, 2025). That policy requires zero data retention and advises confirming that services do not use the data for training or subject it to human review; its earlier guidance names Amazon Bedrock and Google Vertex AI among suitable services. All inference therefore ran through AWS Bedrock and Google Vertex AI with zero data retention confirmed on both platforms: prompts and outputs are neither stored by the provider beyond processing the request, nor used to train or improve its models, nor subject to human review. No consumer chat interface was used, and no report text was entered into any service outside these configured endpoints.

Data release. Our code, including all system prompts and the demonstration-selection procedure, is publicly available at <https://github.com/pri2si17-1997/rad-qa-llm-eval>. It contains no report text: demonstrations are drawn at run time from the credentialed RadQA training split using a fixed seed and question identifiers, so reproducing them requires PhysioNet access. We do not release report text, model outputs that quote report text, or derived data that would allow reconstruction of the records.

Use of AI assistants. We used an AI assistant (Claude Code, Anthropic) to help run the experiments and write the manuscript. The authors verified all code, results and text, and take full responsibility for the content.

Intended use. This work proposes an evaluation method and applies it to 13 LLMs; it does not propose a system for clinical use. None of the evaluated configurations, including the consensus gate, has been validated for clinical decision-making, and the error rates we report (FAR up to 0.403 in one condition) show why none should be used to answer questions about patient records without review. The method is meant for the evaluation that should precede any deployment decision; passing it is not evidence of clinical safety, which would also require local validation and clinical adjudication.

References

- Lisa C. Adams, Daniel Truhn, Felix Busch, Avan Kader, Stefan M. Niehues, Marcus R. Makowski, and Ken K. Bressem. 2023. [Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study](#). *Radiology*, 307(4):e230725.
- Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. [Large language models are few-shot clinical information extractors](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Md Sultan Al Nahian and Ramakanth Kavuluru. 2025. [RadQA-DPO: A radiology question answering system with encoder-decoder models enhanced by direct preference optimization](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 101–113.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.
- Anthropic. 2025a. System Card: Claude Haiku 4.5. <https://www.anthropic.com/claude-haiku-4-5-system-card>.
- Anthropic. 2025b. System Card: Claude Sonnet 4.5. <https://www.anthropic.com/claude-sonnet-4-5-system-card>.
- Anthropic. 2026a. System Card: Claude Opus 5. <https://www.anthropic.com/claude-opus-5-system-card>.
- Anthropic. 2026b. System Card: Claude Sonnet 5. <https://www.anthropic.com/claude-sonnet-5-system-card>.
- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, Martin Hadamitzky, Nassir Navab, Jakob Nikolas Kather, Daniel Truhn, et al. 2025. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, 35(5):2589–2602.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. [MEDITRON-70B: Scaling medical pretraining for large language models](#). *arXiv preprint arXiv:2311.16079*.
- DeepSeek-AI et al. 2025. [DeepSeek-V3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, and 11 others. 2024. [MedAlign: A clinician-generated dataset for instruction following with electronic medical records](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22021–22030.
- Rikhiya Ghosh, Oladimeji Farri, Sanjeev Kumar Karn, Manuela Danu, Ramya Vunikili, and Larisa Micu. 2023. [RadLing: Towards efficient radiology report understanding](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 640–651, Toronto, Canada. Association for Computational Linguistics.
- Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. [PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals](#). *Circulation*, 101(23):e215–e220.
- Google DeepMind. 2025. [Gemini 3 Flash model card](#).
- Google DeepMind. 2026. [Gemini 3.1 Pro model card](#).
- Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. [ClinicalBERT: Modeling clinical notes and predicting hospital readmission](#). *Preprint*, arXiv:1904.05342.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. [RadGraph: Extracting clinical entities and relations from radiology reports](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

- Alistair Johnson, Tom Pollard, and Roger Mark. 2016a. [MIMIC-III Clinical Database](#). *PhysioNet*. Version 1.4.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016b. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3:160035.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Amita Kamath, Robin Jia, and Percy Liang. 2020. [Selective question answering under domain shift](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696. Association for Computational Linguistics.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [BioMistral: A collection of open-source pretrained large language models for medical domains](#). In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864.
- Bastien Le Guellec, Alexandre Lefèvre, Charlotte Geay, Lucas Shorten, Cyril Bruge, Lotfi Hachein-Bey, Philippe Amouyel, Jean-Pierre Pruvo, Gregory Kuchcinski, and Aghiles Hamroun. 2024. [Performance of an open-source large language model in extracting information from free-text radiology reports](#). *Radiology: Artificial Intelligence*, 6(4):e230364.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Kung-Yee Liang and Scott L. Zeger. 1986. [Longitudinal data analysis using generalized linear models](#). *Biometrika*, 73(1):13–22.
- Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. 2026. [Ministral 3](#). *Preprint*, arXiv:2601.08584.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Meta AI. 2025. [The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation](#).
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mistral AI. 2025. Introducing Mistral 3. <https://mistral.ai/news/mistral-3>.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- OpenAI et al. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. [emrQA: A large corpus for question answering on electronic medical records](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.
- PhysioNet. 2023. Responsible use of MIMIC data with online services like GPT. <https://physionet.org/news/post/gpt-responsible-use/>. Accessed 2026-10-07.
- PhysioNet. 2025. Use of MIMIC data with large language models and online services. <https://physionet.org/news/post/llm-responsible-use/>. Accessed 2026-10-07.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Mercy Ranjit, Shaury Srivastav, and Tanuja Ganu. 2024. [RadPhi-3: Small language models for radiology](#). *Preprint*, arXiv:2411.13604.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.
- Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, and Kirk Roberts. 2022. [RadQA: A question answering](#)

dataset to improve comprehension of radiology reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6250–6259, Marseille, France. European Language Resources Association.

Sarvesh Soni and Kirk Roberts. 2022. *RadQA: A question answering dataset to improve comprehension of radiology reports (version 1.0.0)*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. *Know your limits: A survey of abstention in large language models*. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2025. *Medical foundation large language models for comprehensive text analysis and beyond*. *npj Digital Medicine*, 8:141.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. RadBERT: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. *Evaluating progress in automatic chest X-ray radiology report generation*. *Patterns*, 4(9):100802.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706.

A Models and Inference Details

Table 3 lists the evaluated models. Bedrock models were called through the Bedrock Converse API and Gemini models through Vertex AI.

Demonstrations. Demonstrations are drawn from the RadQA training split after removing reports longer than 4,000 characters and shuffling with seed 20260719. The first answerable Findings question (879321_1_1_0), the first answerable Impression question (974127_2_1_I) and the first

unanswerable question (1061153_1_1_0) are used, in that order, in every prompt of the corresponding condition. Each is paired with a fixed rationale written for its type (answerable Findings, answerable Impression, unanswerable), not with any model’s output. Prompts and demonstration rendering are in the released code.

B Statistical Details

Condition effects and intervals. Model-level CIs use a report-level cluster bootstrap (10,000 replicates; 2,000 for decomposition terms). Per-model condition effects use a Wilcoxon signed-rank test on the 104 paired report-mean scores, with BH correction across the 13 models within each transition; per-model extraction terms are tested in the same way on questions in $\mathcal{E}$. Over model means, condition effects are tested with a Wilcoxon test across the 13 models.

False-answer rate. Cochran’s Q tests whether the 13 models differ in FAR within a condition. Each model is compared with the lowest-FAR model on the same 154 unanswerable questions by an exact McNemar test and a report-level sign-flip permutation test (10,000 permutations), BH-corrected within condition (Table 6). An unclustered McNemar screen of per-question correctness across all 78 model pairs (BH across 312 tests) finds 32 pairs that differ on $F1 \geq 0.5$ at $k=0$; Gemini 3 Flash is not separable from Gemini 3.1 Pro, Claude Opus 5, Claude Sonnet 4.5 or LLaMA 4 Maverick ($q = 0.136$ – 1.000).

Item-level models. Selectivity regresses prior score on a newly-abstained indicator (Gaussian GEE, exchangeable within-report correlation); pooled fits include model fixed effects and are checked by leaving out one model at a time, and per-model fits (models with at least 5 new abstentions) are BH-corrected across models. Within-model predictors of false answers (§3.4) use a logistic GEE over model $\times$ unanswerable-question instances (1,596–1,616 per condition), with extraction quality and answering propensity centred within model, model fixed effects and the report-difficulty control.

Correlations across models. Spearman ρ with a 10,000-replicate bootstrap over models and leave-one-model-out p -values (Table 8).

Model	Provider	Model identifier	Release
Claude Opus 5 (Anthropic, 2026a)	AWS Bedrock	us.anthropic.claude-opus-5	2026-07-24
Claude Sonnet 5 (Anthropic, 2026b)	AWS Bedrock	us.anthropic.claude-sonnet-5	2026-06-30
Claude Sonnet 4.5 (Anthropic, 2025b)	AWS Bedrock	us.anthropic.claude-sonnet-4-5-20250929-v1:0	2025-09-29
Claude Haiku 4.5 (Anthropic, 2025a)	AWS Bedrock	us.anthropic.claude-haiku-4-5-20251001-v1:0	2025-10-15
Gemini 3.1 Pro (Google DeepMind, 2026)	Vertex AI	gemini-3.1-pro-preview	2026-02-19
Gemini 3 Flash (Google DeepMind, 2025)	Vertex AI	gemini-3-flash-preview	2025-12-17
LLaMA 4 Maverick (Meta AI, 2025)	AWS Bedrock	us.meta.llama4-maverick-17b-instruct-v1:0	2025-04-05
LLaMA 4 Scout (Meta AI, 2025)	AWS Bedrock	us.meta.llama4-scout-17b-instruct-v1:0	2025-04-05
DeepSeek V3.2 (DeepSeek-AI et al., 2025)	AWS Bedrock	deepseek.v3.2	2025-12-01
Mistral Large 3 (Mistral AI, 2025)	AWS Bedrock	mistral.mistral-large-3-675b-instruct	2025-12-02
Minstral 3 14B (Liu et al., 2026)	AWS Bedrock	mistral.ministral-3-14b-instruct	2025-12-02
GPT-OSS 120B (OpenAI et al., 2025)	AWS Bedrock	openai.gpt-oss-120b-1:0	2025-08-05
GPT-OSS 20B (OpenAI et al., 2025)	AWS Bedrock	openai.gpt-oss-20b-1:0	2025-08-05

Table 3: Evaluated models. Release is the public announcement date from the developer. Mean input and output tokens per condition are in Table 7.

Model	F1				FAR				OAR			
	$k=0$	+F	+F+I	+F+I+U	$k=0$	+F	+F+I	+F+I+U	$k=0$	+F	+F+I	+F+I+U
Gemini 3 Flash	0.748	0.735	0.738	0.716	0.123	0.156	0.143	0.143	0.100	0.080	0.100	0.130
Gemini 3.1 Pro	0.738	0.730	0.728	0.705	0.156	0.162	0.136	0.143	0.083	0.078	0.087	0.122
Claude Sonnet 4.5	0.717	0.723	0.733	0.717	0.117	0.117	0.117	0.110	0.128	0.113	0.102	0.141
Claude Opus 5	0.714	0.716	0.720	0.713	0.162	0.188	0.201	0.182	0.057	0.043	0.035	0.054
LLaMA 4 Maverick	0.710	0.714	0.719	0.711	0.227	0.221	0.214	0.156	0.061	0.059	0.057	0.109
Claude Sonnet 5	0.701	0.709	0.714	0.712	0.169	0.188	0.188	0.195	0.096	0.070	0.061	0.065
Claude Haiku 4.5	0.701	0.688	0.701	0.685	0.156	0.136	0.104	0.104	0.135	0.159	0.163	0.189
Mistral Large 3	0.686	0.710	0.710	0.706	0.169	0.143	0.169	0.143	0.133	0.117	0.111	0.137
GPT-OSS 120B	0.681	0.677	0.666	0.677	0.130	0.123	0.110	0.117	0.150	0.163	0.180	0.178
DeepSeek V3.2	0.676	0.680	0.718	0.688	0.260	0.208	0.175	0.156	0.130	0.120	0.093	0.152
LLaMA 4 Scout	0.672	0.660	0.682	0.662	0.260	0.403	0.344	0.123	0.100	0.054	0.048	0.185
GPT-OSS 20B	0.669	0.685	0.692	0.669	0.162	0.169	0.149	0.162	0.141	0.137	0.137	0.163
Minstral 3 14B	0.588	0.644	0.655	0.652	0.182	0.162	0.169	0.130	0.202	0.204	0.193	0.224
Mean	0.692	0.698	0.706	0.693	0.175	0.183	0.171	0.143	0.117	0.108	0.105	0.142

Table 4: F1, false-answer rate (FAR) and over-abstention rate (OAR) by demonstration condition. Conditions are cumulative: +F adds an answerable Findings demonstration, +F+I an answerable Impression demonstration, and +F+I+U an unanswerable demonstration.

Error-taxonomy thresholds. The no-overlap rate at $k=0$ is 1.64%, 1.88% and 3.41% for F1 thresholds of 0.05, 0.10 (used) and 0.20; its decline from $k=0$ to +F+I holds at every threshold, and the EM-mismatch threshold (0.4–0.6) does not affect it.

C Additional Results

Table 4 gives per-model F1, FAR and OAR in every condition; Table 5 the exact values behind Figure 2; Table 6 the FAR tests in every condition; Tables 8 and 9 the cross-model and within-model analyses of §4.4; Table 10 the consensus gate; and Table 7 token usage.

Transition	Token F1			RadGraph F1			Counts	
	Total Δ	Extraction	Answerability	Total Δ	Extraction	Answerability	New abst.	New FA
$k=0 \rightarrow +F$	+0.0056 [−0.0007, +0.0120]	+0.0021 [−0.0011, +0.0054]	+0.0035 [−0.0013, +0.0083]	+0.0094 [+0.0028, +0.0161]	+0.0074 [+0.0026, +0.0122]	+0.0020 [−0.0025, +0.0065]	134	71
$+F \rightarrow +F+I$	+0.0080 [+0.0032, +0.0132]	+0.0025 [+0.0001, +0.0050]	+0.0056 [+0.0010, +0.0104]	+0.0141 [+0.0076, +0.0215]	+0.0095 [+0.0048, +0.0149]	+0.0046 [+0.0001, +0.0093]	96	26
$+F+I \rightarrow +F+I+U$	−0.0126 [−0.0196, −0.0058]	+0.0001 [−0.0018, +0.0021]	−0.0127 [−0.0195, −0.0063]	−0.0101 [−0.0169, −0.0032]	−0.0008 [−0.0034, +0.0019]	−0.0093 [−0.0153, −0.0035]	272	18
$k=0 \rightarrow +F+I$	+0.0136 [+0.0072, +0.0200]	+0.0048 [+0.0015, +0.0081]	+0.0087 [+0.0034, +0.0144]	+0.0235 [+0.0149, +0.0325]	+0.0171 [+0.0104, +0.0244]	+0.0064 [+0.0011, +0.0120]	144	62

Table 5: Decomposition of the change in mean score between conditions (Eq. 1), pooled over 13 models, with 95% report-cluster bootstrap CIs (2,000 replicates) below each estimate. Extraction and answerability terms sum to the total. New abst.: answerable questions newly declined; New FA: unanswerable questions newly answered (new false answers). Counts are pooled over models and are the same for both metrics.

Model	$k=0$	+F	+F+I	+F+I+U
Gemini 3 Flash	0.123	0.156	0.143	0.143
Gemini 3.1 Pro	0.156	0.162	0.136	0.143
Claude Sonnet 4.5	0.117 ^r	0.117 ^r	0.117	0.110
Claude Opus 5	0.162	0.188*	0.201*	0.182*
LLaMA 4 Maverick	0.227*	0.221*	0.214*	0.156
Claude Sonnet 5	0.169	0.188*	0.188*	0.195*
Claude Haiku 4.5	0.156	0.136	0.104 ^r	0.104 ^r
Mistral Large 3	0.169	0.143	0.169*	0.143
GPT-OSS 120B	0.130	0.123	0.110	0.117
DeepSeek V3.2	0.260*	0.208*	0.175*	0.156
LLaMA 4 Scout	0.260*	0.403*	0.344*	0.123
GPT-OSS 20B	0.162	0.169	0.149	0.162
Minstral 3 14B	0.182	0.162	0.169*	0.130
Cochran’s Q (df = 12)	69.2	177.3	134.0	36.5

Table 6: FAR by model and condition. ^rReference (lowest-FAR) model; *significantly higher than the reference under both tests ($q < 0.05$). All Cochran’s Q tests have $p \leq 0.0003$.

Model	$k=0$	+F	+F+I	+F+I+U
Gemini 3 Flash	852 / 552	1,145 / 587	1,322 / 611	1,621 / 654
Gemini 3.1 Pro	852 / 361	1,145 / 306	1,322 / 323	1,621 / 334
Claude Sonnet 4.5	1,510 / 159	1,844 / 146	2,064 / 145	2,410 / 144
Claude Opus 5	1,747 / 174	2,216 / 168	2,538 / 165	3,014 / 161
LLaMA 4 Maverick [‡]	822 / 87	1,102 / 77	1,270 / 77	1,559 / 84
Claude Sonnet 5	1,815 / 170	2,284 / 161	2,606 / 158	3,082 / 159
Claude Haiku 4.5	1,510 / 157	1,844 / 142	2,064 / 142	2,410 / 141
Mistral Large 3	1,201 / 268	1,509 / 173	1,662 / 130	1,959 / 123
GPT-OSS 120B	1,184 / 242	1,473 / 245	1,643 / 246	1,937 / 243
DeepSeek V3.2	967 / 211	1,253 / 180	1,417 / 184	1,709 / 183
LLaMA 4 Scout [‡]	822 / 89	836 / 76	1,004 / 75	1,293 / 71
GPT-OSS 20B [§]	1,184 / 251	1,549 / 295	1,832 / 319	2,271 / 372
Minstral 3 14B	1,201 / 105	1,494 / 96	1,662 / 94	1,959 / 94
Model mean	1,205 / 217	1,515 / 204	1,724 / 205	2,065 / 212

Table 7: Token usage by model and condition. Mean input / output tokens per row (all 894 rows of the test file; §3.1), as reported by each provider’s API. Counts use each model’s own tokenizer, so they are comparable across conditions within a model but not across model families. [‡]Bedrock-reported input tokens for the LLaMA 4 models are inconsistent: Scout reports about 266 fewer tokens than Maverick in every demonstration condition, and re-querying both with an identical prompt gave equal, lower counts. A probe confirmed that Scout received all demonstrations; treat LLaMA 4 input-token counts as approximate. [§]Token means are inflated by retried calls (counts exactly 2–3 $\times$ those of GPT-OSS 120B on 29/74/109 of 894 rows at +F/+F+I/+F+I+U); answers are unaffected.

Measure	$k=0$	+F	+F+I	+F+I+U
SQuAD F1	-0.59 (.036)	-0.23 (.46)	-0.04 (.89)	+0.25 (.41)
Answerable F1	-0.28 (.35)	+0.35 (.24)	+0.48 (.10)	+0.55 (.053)
Extraction F1	-0.47 (.105)	-0.59 (.033)	-0.49 (.092)	-0.43 (.14)
OAR	-0.08 (.78)	-0.58 (.037)	-0.72 (.005)	-0.69 (.009)

Table 8: Spearman correlation with FAR across the 13 models (p in parentheses). Answerable F1 scores abstentions on answerable questions as 0; extraction F1 averages over answerable questions the model answered (§3.4).

Condition	Extraction	Propensity	Difficulty
$k=0$	1.04 [0.92, 1.18]	1.21 [1.01, 1.45]	0.84 [0.57, 1.24]
+F	0.94 [0.84, 1.04]	1.23 [0.94, 1.60]	0.88 [0.55, 1.39]
+F+I	0.99 [0.86, 1.15]	1.31 [1.03, 1.65]	0.83 [0.56, 1.22]
+F+I+U	0.99 [0.89, 1.09]	1.10 [0.79, 1.53]	0.91 [0.59, 1.42]

Table 9: Within-model predictors of a false answer: odds ratio per SD [95% CI] from a logistic GEE with model fixed effects and report clustering (1,596–1,616 model $\times$ unanswerable-question instances per condition, 52 reports). Extraction: the model’s mean F1 on the report’s answerable questions it answered; Propensity: the fraction of those questions it answered; Difficulty: the other 12 models’ mean F1 on the report (§3.4).

Condition, rule	Δ F1	Δ FAR	Δ OAR
$k=0$, own- t	+0.005 [-0.000, +0.009]	-0.040 [-0.057, -0.026]	+0.010 [+0.004, +0.018]
$k=0$, pooled- t	+0.004 [-0.003, +0.009]	-0.042 [-0.063, -0.016]	+0.012 [+0.002, +0.027]
+F, own- t	+0.005 [-0.001, +0.008]	-0.032 [-0.044, -0.023]	+0.006 [+0.001, +0.015]
+F+I, own- t	+0.003 [-0.003, +0.006]	-0.031 [-0.045, -0.018]	+0.009 [+0.003, +0.023]
+F+I+U, own- t	-0.000 [-0.004, +0.001]	-0.011 [-0.018, -0.004]	+0.005 [+0.002, +0.012]

Table 10: Held-out effect of the cross-fitted consensus gate: model-mean change vs. the ungated model, with 2.5–97.5% intervals over 200 report splits below each estimate. Pooled- t uses one threshold for all models (modal $t = 4$).