MedVAL: Toward Expert-Level Medical Text Validation with Language Models

Asad Aali **Vasiliki Bikia** **Maya Varma** **Nicole Chiou** **Sophie Ostmeier**
Arnav Singhvi **Magdalini Paschali** **Ashwin Kumar** **Andrew Johnston**
Karimar Amador Martinez **Eduardo Perez Guerrero** **Paola N Rivera**
Sergios Gatidis **Christian Bluethgen** **Eduardo Reis** **Eddy Rilland**
Poonam Hosamani **Kevin Keet** **Minjoung Go** **Evelyn Bin Ling**
David Larson **Curt Langlotz** **Roxana Daneshjou** **Jason Hom**
Sanmi Koyejo **Emily Alsentzer** **Akshay S. Chaudhari**

Stanford University

Abstract

With the growing use of language models (LMs) in clinical environments, there is an immediate need to evaluate the accuracy and safety of LM-generated medical text. However, detecting errors in LM-generated text is challenging because 1) manual review is costly and 2) expert-composed reference outputs are often unavailable in real-world settings. To address these challenges, we propose MedVAL, a novel, self-supervised, data-efficient distillation method that leverages synthetic data to train evaluator LMs to assess whether LM-generated medical outputs are factually consistent with inputs, without requiring physician labels or reference outputs. To evaluate LM performance, we introduce MedVAL-Bench, a dataset of 840 physician-annotated outputs across 6 diverse medical tasks capturing real-world challenges. Across 10 state-of-the-art LMs spanning open-source, proprietary, and medically adapted models, MedVAL distillation significantly improves ($p < 0.001$) alignment with physicians across seen and unseen tasks, increasing average F1 scores from 66% to 83%. Despite strong baseline performance, MedVAL improves the best-performing proprietary LM (GPT-4o) by 8% without training on physician-labeled data, demonstrating a performance statistically non-inferior to a single human expert ($p < 0.001$). Our benchmark provides evidence of LMs approaching expert-level ability in validating AI-generated medical text.

1 Introduction

Language models (LMs) are increasingly leveraged for documentation in medical settings, supporting tasks such as clinical text summarization, report generation, or question answering [1–7]. While LMs offer potential for reducing documentation burden [8–11], they are not infallible and frequently generate plausible text that contains subtle errors such as hallucinations, omissions, or certainty misalignments [12–14]. This issue is critical within medicine, where errors are often cloaked in jargon, misleading even experienced practitioners [15].

The adoption of LMs [16–18] for medical applications necessitates reliable risk assessment through medical text validation, which involves determining whether an output from a LM is factually consistent with the input. Currently, this process requires manual physician review both pre- and post-deployment to ensure continued safety, contributing to documentation workload and associated cognitive fatigue for physicians [19–22].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

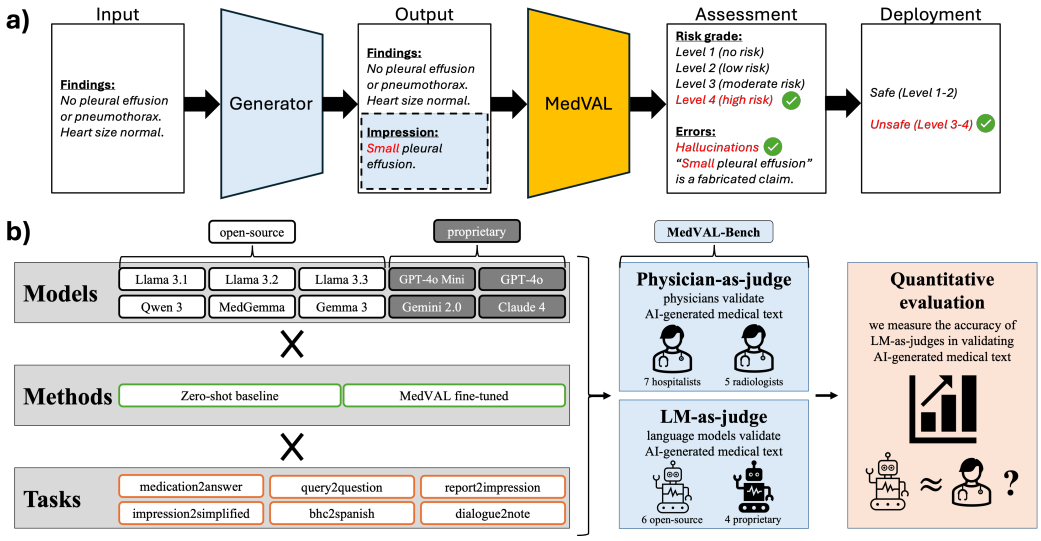

Figure 1: **a) MedVAL test-time workflow.** A generator LM produces an output, and MedVAL then assesses the output's factual consistency with the input, while assigning a risk grade and determining whether the output is safe for deployment. **b) Study framework.** 12 physicians assess 840 LM-generated medical text outputs. Using physician assessments as reference, we measure the accuracy of LMs in medical text validation across 10 LMs, 2 methods (baseline vs. MedVAL), and 6 tasks.

Despite this need, traditional NLP metrics fail to capture subtle, clinically significant errors that can impact patient care [23]. These metrics rely on expert-composed outputs (reference outputs), which are not available in many real-world scenarios [24]. Moreover, manual physician review remains the reference standard for validating medical text, but it is costly, time-consuming, and not scalable. With physician shortages and burnout at an all-time high [25–28], assigning physicians additional validation tasks is impractical. Therefore, automated and scalable strategies that minimize reliance on expert-driven data are urgently needed to adapt LMs for medical text validation.

Using a LM to evaluate another LM's output, also termed the "LM-as-judge" paradigm, has shown promise in automating the evaluation of LM-generated text. However, most general-purpose LM-as-judge methods treat LMs as static evaluators [29], often lacking the granularity required to assess nuanced, high-stakes medical text [30–34]. Recent medical-specific LM-as-judge approaches suffer from certain practical limitations; these methods often assume resources like expert-labeled training data [35], availability of reference outputs [36], or retrieval-based evidence [37], none of which are readily scalable. On the other hand, many approaches focus on subdomains such as radiology (e.g., chest X-rays), limiting generalization to broader medical tasks [38–44].

To overcome these challenges, we introduce MedVAL (medical text validator), a self-supervised, data-efficient distillation method (Figure 1), which curates high-quality synthetic examples by leveraging the agreement between a generator and a validator LM as a proxy for physician judgment [45]. Through this training, MedVAL enables LMs to assess whether an output is factually consistent with the input by assigning one of four risk levels, while flagging "deployment unsafe" outputs at near physician-level reliability. In comparison to prior methods, our framework enables: 1) development of scalable evaluators without physician-in-loop supervision, 2) evaluation without reliance on reference outputs, and 3) adaptation across diverse tasks.

We evaluate MedVAL across 6 open-source and 4 proprietary LMs on MedVAL-Bench, our newly created dataset consisting of 840 physician-annotated outputs across 6 diverse tasks [46]. Unlike prior LM-as-judge methods that rely on zero-shot inference, MedVAL distillation improves the validation capabilities of all underlying LMs across tasks. MedVAL yields significant gains ($p < 0.001$): average baseline F1 scores for four-class risk-grading improve from 36.7% to 51.0%, and for binary safe/unsafe classification, from 66.2% to 82.8%. Our benchmark provides evidence that LMs can achieve performance statistically non-inferior to a single human expert ($p < 0.001$).

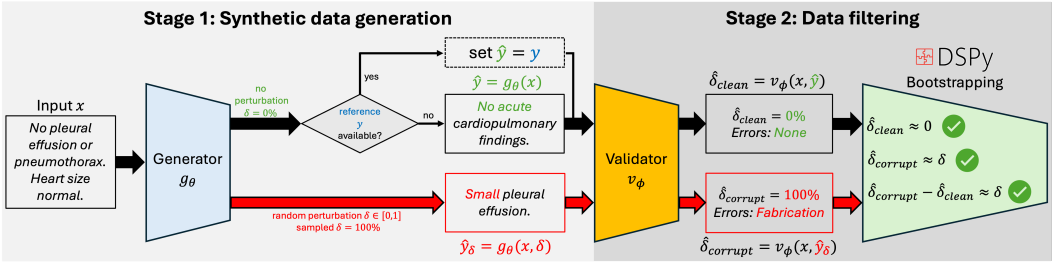

Figure 2: **MedVAL self-supervised data curation** illustrated through a radiology report summarization example. 1) A generator g_θ takes as input x, and generates a clean and a perturbed output using a random perturbation level $\delta \in [0, 1]$. 2) A validator v_ϕ then provides a detailed error assessment, predicts the factual degradation level $\hat{\delta}_{clean}$ and $\hat{\delta}_{corrupt}$ of the clean and perturbed outputs, respectively, and filters data with high generator-validator consistency for fine-tuning an arbitrary LM.

2 Methods

2.1 MedVAL Training

We propose a self-supervised distillation method for training LMs to validate LM-generated medical text, involving three stages: 1) synthetic data generation, 2) data filtering, and 3) fine-tuning. Figure 2 describes the data curation process, and Algorithm S1 summarizes the complete training process.

Definitions. We denote the generator LM as g_θ, validator LM as v_ϕ, the fine-tunable validator LM as v_α, the input as x, the corresponding reference output as y, the output as $\hat{y} = g_\theta(x)$ or $\hat{y} = y$ (if y available), and the perturbed output as $\hat{y}_\delta = g_\theta(x, \delta)$. Here, $\delta \in [0, 1]$ represents the perturbation level in terms of factual degradation, where $\delta = 0$ represents an unperturbed generation and higher δ values correspond to more factually inconsistent generations. $\hat{\delta}_{clean} = v_\phi(x, \hat{y})$ and $\hat{\delta}_{corrupt} = v_\phi(x, \hat{y}_\delta)$ denote the validator's prediciton of the factual degradation in $\hat{y}$ and $\hat{y}_\delta$.

Stage 1: Synthetic data generation. Our goal is to create a synthetic dataset $\mathcal{D} = \{x, \hat{y}, \hat{y}_\delta\}$ that contains a triplet of the input, clean unperturbed output, and the perturbed output, to enable supervision without expert labels. To achieve this, we first introduce perturbations in the outputs. The perturbations in $\hat{y}_\delta$ are introduced via additional instructions in the generator's prompt. To ensure each perturbation level corresponds to an interpretable perturbation method, we define a discrete set of degradation levels $\{\delta_1, \delta_2, \ldots, \delta_L\}$. These degradation levels were categorized into four clinically meaningful strata—no risk, low risk, moderate risk, and high risk—in conjunction with our physician team, based on their potential to influence patient safety. While we use a discrete risk-level taxonomy for implementation, the framework supports continuous degradation levels and scalar validator outputs, which can be modified based on the perturbation injection method. For each sample in the dataset, we randomly select a perturbation level to ensure coverage of a range of error severities. The perturbations, associated risk levels, and instructions are described in Table 1.

Stage 2: Data filtering. We aim to filter $\mathcal{D}$ to obtain a "high-quality" subset $\mathcal{D}_{\text{train}}$ for fine-tuning a validator LM. $\mathcal{D}_{\text{train}}$ should ideally contain examples where the validator v_ϕ and generator g_θ strongly agree on the expected factual degradation level. To achieve this, we first leverage the validator v_ϕ to predict the factual degradation levels for both the clean and perturbed outputs. For filtering the synthetic data as a "quality control" mechanism, we propose a metric $\mathcal{M}_{\text{MedVAL}}$ that quantifies consistency between the generator and validator via evaluating the agreement between expected and predicted factual consistency levels. We propose two independent components for our metric. 1) ***Absolute consistency*** ensures the validator's predicted factual degradation level for each generation should match the expected degradation level. Specifically, for clean generations, we expect $v_\phi(x, \hat{y}) \approx 0$. For perturbed generations, we expect $v_\phi(x, \hat{y}_\delta) \approx \delta$. 2) ***Relative consistency*** ensures the validator preserves the expected factual degradation between clean and perturbed generations. Specifically, we expect $v_\phi(x, \hat{y}_\delta) - v_\phi(x, \hat{y}) \approx \delta$. We define these terms as:

Table 1: **Physician-defined taxonomy of risk levels, perturbations, and error categories.**

Perturbation	Category	Risk	Safety	Action
$\delta = 0\%$	Level 1	No Risk	Safe	Expert review not required.
$\delta = 33\%$	Level 2	Low Risk	Acceptable	Expert review optional.
$\delta = 67\%$	Level 3	Moderate Risk	Potentially unsafe	Expert review required.
$\delta = 100\%$	Level 4	High Risk	Unsafe	Expert rewrite required.

Perturbation	Instructional prompt
$\delta = 0\%$	"The output should contain **no clinically meaningful factual inconsistencies**. Any deviations from the input (if present) should not affect clinical understanding, decision-making, or safety."
$\delta = 33\%$	"The output should contain **subtle or ambiguous inconsistencies that are unlikely to influence clinical decisions** or understanding. These inconsistencies should not introduce confusion or risk."
$\delta = 67\%$	"The output should contain **inconsistencies that could plausibly affect clinical interpretation, documentation, or decision-making**. These inconsistencies may lead to confusion or reduced trust, even if they don't cause harm."
$\delta = 100\%$	"The output should include one or more **inconsistencies that could result in incorrect or unsafe clinical decisions**. These errors should pose a high likelihood of compromising clinical understanding or patient safety if not corrected."

Error category	Error	Description
Hallucinations	Fabricated claim	Introduction of a claim not present in the input.
	Misleading justification	Incorrect reasoning and misleading conclusions.
	Detail misidentification	Incorrect reference to a detail in the input.
	False comparison	Mentioning a comparison unsupported by the input.
	Incorrect recommendation	Suggesting a diagnosis/follow-up outside the input.
Omissions	Missing claim	Failure to mention a claim present in the input.
	Missing comparison	Omitting a comparison detailing change over time.
	Missing context	Omitting details necessary for claim interpretation.
Certainty Misalignments	Overstating intensity	Exaggerating urgency, severity, or confidence.
	Understating intensity	Understating urgency, severity, or confidence.
Other	Other	Additional errors not covered.

$$\mathcal{M}_{\text{consistency}} = \underbrace{\|v_\phi(x, \hat{y})\|_2^2 + \|v_\phi(x, \hat{y}_\delta) - \delta\|_2^2}_{\mathcal{M}_{\text{absolute}}} + \underbrace{\|v_\phi(x, \hat{y}_\delta) - v_\phi(x, \hat{y}) - \delta\|_2^2}_{\mathcal{M}_{\text{relative}}} \tag{1}$$

The first term in $\mathcal{M}_{\text{absolute}}$ rewards clean generations that achieve validator scores close to 0, and the second term rewards perturbed generations whose validator scores closely match the expected degradation δ. $\mathcal{M}_{\text{relative}}$ ensures that the difference between the predicted scores reflects the expected degradation. To bound the score between 0 and 1 ($\uparrow$ score $= \uparrow$ generator-validator consistency), we divide it by its maximum possible value:

$$\mathcal{M}_{\text{MedVAL}} = 1 - \frac{\mathcal{M}_{\text{consistency}}}{6} \tag{2}$$

Stage 3: Fine-tuning. We create our synthetic dataset $\mathcal{D}_{\text{train}} = \{x, \hat{y}, \hat{\delta}_{\text{clean}}, \hat{y}_\delta, \hat{\delta}_{\text{corrupt}}\}$ by filtering examples from $\mathcal{D}$ where $\mathcal{M}_{\text{MedVAL}} \geq \tau$ (pre-specified threshold). Importantly, we apply a single-pass consistency filter (not iterative), and choose τ to retain examples with high generator-validator agreement. We then fine-tune an arbitrary LM $v_\alpha^* = \mathit{SFT}(v_\alpha, \mathcal{D}_{\text{train}})$ using standard parameter-efficient supervised fine-tuning (SFT). To improve robustness, we optionally ensemble multiple MedVAL fine-tuned LMs by aggregating their predictions.

Table 2: **MedVAL-Bench dataset.** Overview of task-specific data sources, relevant statistics, and generation instructions. *# physicians* denotes number of physicians involved in test-time annotation. The *train set* is self-supervised and relies on input-output pairs to create synthetic supervision labels.

Task	Dataset	# samples		Avg. # tokens		# physicians
		Train	Test	Train	Test	
`medication2answer`	MedicationQA	500	135	10 ± 4	10 ± 4	2
`query2question`	MeQSum	500	120	80 ± 57	82 ± 66	3
`report2impression`	Open-i	500	190	54 ± 21	50 ± 22	5
`report2simplified`	Open-i	500	-	54 ± 22	-	-
`impression2simplified`	MIMIC-IV	-	190	-	69 ± 61	5
`bhc2spanish`	MIMIC-IV	-	120	-	543 ± 391	3
`dialogue2note`	ACI-Bench	-	85	-	$1{,}497 \pm 445$	2

Task	Input → output	Instructional prompt
`medication2answer`	medication question → answer	"Answer the following medication-related patient health question."
`query2question`	patient query → health question	"Summarize the patient health query into one question."
`report2impression`	findings → impression	"Summarize the radiology report findings into an impression."
`report2simplified`	findings → patient-friendly	"Create a simplified, patient-friendly version of the input."
`impression2simplified`	impression → patient-friendly	"Create a simplified, patient-friendly version of the input."
`bhc2spanish`	hospital course → spanish	"Translate the brief hospital course into Spanish."
`dialogue2note`	doctor-patient dialogue → note	"Summarize the dialogue into an assessment and plan."

2.2 Data

We introduce MedVAL-Bench (Table 2), a dataset for evaluating the ability of LMs to validate LM-generated medical text. The train set is self-supervised, containing inputs (used to generate supervision labels) and reference outputs (if available). To enable LM benchmarking, the test set contains physician assessments of outputs spanning broad error severities.

3 Results

3.1 Overall Performance

We show that MedVAL distillation consistently improves the performance of all LMs. The F1 score and Cohen's κ are reported in Figure 3 and Figure S1. Across all tested LMs, MedVAL distillation leads to average baseline F1 scores improving from 36.7% to 51.0%. On average, MedVAL improved F1 scores by 113% for open-source LMs and 11% for proprietary LMs, where baseline performance was already considerably high. Despite the inherent difficulty of this four-class classification task, where even frontier LMs struggle and baseline scores never exceed 55%, MedVAL distillation improves GPT-4o's F1 score from 54.5% to 58.7%, establishing the best performance.

Among all open-source LMs, Qwen-3-4B exhibits the highest F1 score after MedVAL distillation (42.8% → 52.7%), surpassing other notably larger open-source models such as Llama-3.3-70B and even the medically fine-tuned MedGemma-27B. Among zero-shot baselines, MedGemma-27B achieves the highest performance (48.2%), slightly outperforming Llama-3.3-70B and Gemma-3-27B. Notably, MedVAL distillation substantially improves the F1 scores of Llama-3.2-3B (12.8% → 42.9%), Llama-3.1-8B (25.9% → 46.5%). Despite Llama-3.1-8B also benefiting from MedVAL, MedVAL Qwen-3-4B outperforms it. Remarkably, MedVAL Qwen-3-4B even outshines certain proprietary baselines such as GPT-4o Mini (47.4%) and Gemini 2.0 Flash (51.5%).

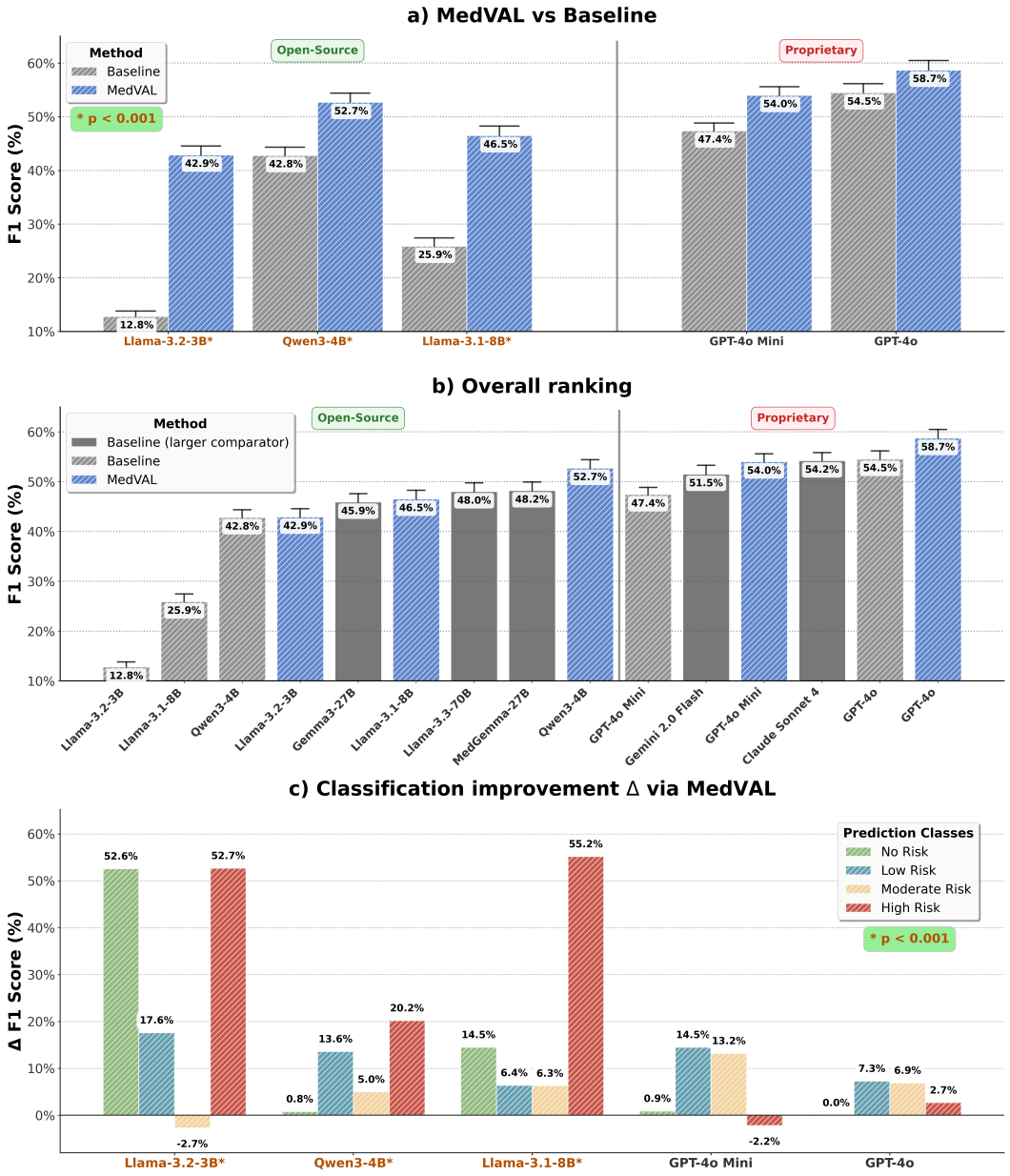

Figure 3: **Performance benchmark (F1 score)**. **a)** We report the performance of LMs before and after MedVAL distillation. **b)** We rank all LMs (low to high), grouped into three methods. **c)** We report the Δ F1 score between MedVAL and baseline LM performance across each prediction class. *Baseline* indicates zero-shot LM before distillation, *Baseline (larger comparator)* indicates a larger zero-shot LM as reference (not chosen for distillation), and *MedVAL* indicates LM after distillation. $^*p < 0.001$ indicates statistically significant difference in classification performance of MedVAL and baseline (McNemar test). Notably, smaller MedVAL LMs match or exceed the performance of much larger baseline LMs. Furthermore, MedVAL Qwen3-4B (52.7%) and GPT-4o (58.7%) achieve the highest F1 score ranking under respective categories.

Within proprietary LMs, post MedVAL distillation, GPT-4o Mini (47.4% $\rightarrow$ 54.0%) rivals the baseline performance of larger Claude Sonnet 4 (54.2%) and GPT-4o (54.5%), and surpasses Gemini 2.0 Flash (51.5%), highlighting that MedVAL can enable LMs to reach baseline performance of larger LMs.

We report Krippendorff's α to assess inter-reader variability across the six tasks in Table 3a. The overall α of 0.848 indicates substantial to almost perfect agreement, proving the reliability of the physician annotations. A value of $\alpha \geq 0.80$ indicates reliable rating quality to draw triangulated conclusions based on the rated data [47]. Notably, the highest inter-physician agreement was observed in the `bhc2spanish` and `medication2answer` tasks, with an α of 0.943 and 0.904, respectively, while the lowest was in the `query2question` task, with an α of 0.560. We also observe that physician agreement correlates strongly with MedVAL performance: we observe a Pearson correlation of $r = 0.67$ between inter-physician agreement and GPT-4o MedVAL F1 scores. This suggests that when experts agree more, MedVAL produces higher and consistent F1 scores.

3.2 Risk-Level Classification Performance

MedVAL improves classification performance across risk levels. Figure 3c demonstrates the Δ F1 score between MedVAL and baseline LM performance. Gains are most pronounced in smaller open-source LMs, where improvements are statistically significant ($p < 0.001$). For larger proprietary LMs, we also observe improvements, although they are not statistically significant ($p > 0.1$).

Among open-source models, Llama-3.2-3B showed the largest gains, with F1 score increases of over 50% at levels 1 and 4, and over 17% at level 2. Qwen3-4B exhibited improvements in level 2 (13.6%) and level 4 (20.2%). Llama-3.1-8B improved substantially at level 1 (14.5%) and level 4 (55.2%).

Proprietary models like GPT-4o Mini exhibited notable gains at intermediate levels (level 2: 14.5%; level 3: 13.2%). Even GPT-4o showed smaller but consistent improvements, especially at level 2 (7.3%) and level 3 (6.9%). These gains are clinically significant because levels 2 and 3 represent the threshold between outputs that are likely safe for use with minimal oversight (level 2) and those requiring human review or revision (level 3). Interestingly, MedVAL sharpens this decision boundary for even high-performing baseline frontier LMs. Although minor drops were observed in a few model-risk combinations, such as GPT-4o Mini at level 4 (-2.2%) and Llama-3.2-3B at level 3 (-2.7%), the overall pattern shows strong and consistent gains across clinically meaningful risk categories.

3.3 Task-Wise Performance

The task-wise F1 scores illustrated in Table 3a (standard deviations in Table S4) reveal substantial improvement in LM performance. On average across LMs, MedVAL distillation yielded a 65% F1 score improvement on seen tasks and an 84% improvement on unseen tasks. Notably, MedVAL displays strong improvements on `dialogue2note`, an unseen task with the longest context lengths.

Among open-source models, Qwen-3-4B and Llama-3.2-3B showed strong task-wise improvements. Qwen-3-4B demonstrated remarkable performance, achieving the highest F1 scores in 3 out of 6 tasks: `medication2answer`, `query2question`, and `bhc2spanish`. Although it did not lead in other tasks, it consistently maintained high performance, culminating in the highest overall F1 score of 52.7% among open-source models, and even outperforming baseline proprietary GPT-4o-Mini and Gemini 2.0 Flash LMs. Llama-3.2-3B improved substantially in each task, contributing to an overall F1 score increase from 12.8% to 42.9%. Similarly, the task-wise improvements in Llama-3.1-8B contributed to its performance increase from 25.9% to 46.5%.

In the proprietary category, MedVAL GPT-4o saw a consistent performance increase across all tasks, leading to the highest overall F1 score (58.7%). Notably, MedVAL GPT-4o Mini outperforms all LMs on two tasks (`report2impression` and `dialogue2note`). However, MedVAL GPT-4o Mini also sees minor drops in `query2question` and `bhc2spanish`. However, we observe that the `query2question` task shows consistently lower performance across models, aligning with its low inter-annotator agreement ($\alpha = 0.560$).

We also conduct an in-depth analysis to understand the implications of LM predictions. For a granular understanding, we review examples where the baseline LM incorrectly evaluates the output, but corrects the evaluation after MedVAL distillation. We present such examples in Figure S3, Figure S4, and Figure S5 (showing "reasoning" behind LM's risk-grade prediction). Together, these findings demonstrate that MedVAL enhances both the generalization and task-specific capabilities of LMs.

Table 3: **a) Task-wise performance benchmark (F1 score). Bolded** values highlight best performance under respective categories (open-source/proprietary). **Green** and **red** values indicate improvement or decline in F1 score compared to the corresponding baseline LM. On average across LMs, MedVAL distillation increases F1 scores by 65% on in-distribution tasks and 84% on out-of-distribution tasks. Krippendorff's $\alpha = 0.848$ represents high inter-physician agreement across examples evaluated by multiple physicians, indicating the reliability of physician annotations. MedVAL Qwen3-4B (referred to as MedVAL-4B) achieves the best performance among all open-source LMs, matching the baseline performance of proprietary LMs.

Method	Model	In-Distribution			Out-of-Distribution			Overall
		`medication2 answer`	`query2 question`	`report2 impression`	`impression2 simplified`	`bhc2 spanish`	`dialogue2 note`	
Open-Source								
Baseline	Llama-3.2-3B	0.091	0.110	0.174	0.096	0.120	0.146	0.128
	Qwen3-4B	0.357	0.299	0.530	0.390	0.364	0.552	0.428
	Llama3.1-8B	0.342	0.285	0.278	0.225	0.158	0.113	0.259
	Gemma3-27B	0.398	0.279	0.584	0.442	0.369	0.552	0.459
	MedGemma-27B	0.462	0.287	0.616	0.451	0.349	**0.603**	0.482
	Llama-3.3-70B	0.478	0.311	**0.633**	0.496	0.362	0.322	0.480
MedVAL	Llama-3.2-3B	0.382 +320%	0.262 +138%	0.578 +232%	0.429 +347%	0.242 +102%	0.448 +207%	0.429 +235%
	Qwen3-4B	**0.557** +56%	**0.374** +25%	0.562 +6%	0.537 +38%	**0.424** +16%	0.490 -11%	**0.527** +23%
	Llama-3.1-8B	0.456 +33%	0.372 +31%	0.480 +73%	**0.540** +140%	0.384 +143%	0.376 +233%	0.465 +80%
Proprietary								
Baseline	GPT-4o Mini	0.479	0.352	0.445	0.503	0.427	0.586	0.474
	GPT-4o	0.598	0.360	0.519	0.587	0.439	0.618	0.545
	Claude Sonnet 4	0.569	**0.413**	0.497	0.583	**0.552**	0.550	0.542
	Gemini 2.0 Flash	0.485	0.401	0.588	0.486	0.497	0.602	0.515
MedVAL	GPT-4o Mini	0.512 +7%	0.308 -13%	**0.635** +43%	0.571 +14%	0.386 -10%	**0.692** +18%	0.540 +14%
	GPT-4o	**0.695** +16%	0.361 +0%	0.564 +9%	**0.605** +3%	0.483 +10%	0.673 +9%	**0.587** +8%
Krippendorff's α								
Inter-Physician Agreement		0.904	0.560	0.861	0.872	0.943	0.830	0.848

b) Safe/unsafe (binary) performance benchmark. Bolded values represent the best-performing method under each model (MedVAL vs Baseline). **Green** values represent the best-performing model/method combination under respective categories (open-source/proprietary). *Ensemble* indicates aggregation of outputs from multiple MedVAL LMs. Across all LMs, MedVAL improves average accuracy from 71% to 81% and F1 score from 66% to 83%, demonstrating reliable discrimination between deployment-safe and unsafe responses. Notably, inter-physician agreement (Krippendorff's $\alpha = 0.754$) falls within range of the best MedVAL classifiers (Accuracy/F1 $> 80\%$), suggesting comparable consistency, albeit measured via a different metric.

#	Model	Method	Sensitivity	Specificity	F1 Score	Accuracy
1	Llama-3.2-3B	Baseline	0.086$\pm$0.01	**0.960**$\pm$0.01	0.153$\pm$0.02	0.474$\pm$0.02
		MedVAL	**0.919**$\pm$0.01	0.560$\pm$0.02	**0.809**$\pm$0.01	**0.760**$\pm$0.01
2	Llama-3.1-8B	Baseline	0.670$\pm$0.02	0.651$\pm$0.03	0.688$\pm$0.02	0.662$\pm$0.02
		MedVAL	**0.788**$\pm$0.02	**0.786**$\pm$0.01	**0.804**$\pm$0.01	**0.787**$\pm$0.01
3	Qwen3-4B	Baseline	**0.858**$\pm$0.02	0.643$\pm$0.03	0.800$\pm$0.01	0.762$\pm$0.02
		MedVAL	0.839$\pm$0.02	**0.752**$\pm$0.02	**0.823**$\pm$0.01	**0.800**$\pm$0.01
	Ensemble (1+2+3)	MedVAL	0.899$\pm$0.02	0.686$\pm$0.03	**0.837**$\pm$0.01	**0.805**$\pm$0.01
4	GPT-4o Mini	Baseline	0.784$\pm$0.02	0.807$\pm$0.02	0.809$\pm$0.02	0.794$\pm$0.02
		MedVAL	**0.848**$\pm$0.02	**0.831**$\pm$0.02	**0.855**$\pm$0.01	**0.840**$\pm$0.01
5	GPT-4o	Baseline	**0.835**$\pm$0.02	0.861$\pm$0.02	**0.858**$\pm$0.01	**0.846**$\pm$0.01
		MedVAL	0.792$\pm$0.02	**0.906**$\pm$0.02	0.849$\pm$0.01	0.843$\pm$0.01
	Ensemble (4+5)	MedVAL	**0.874**$\pm$0.02	0.815$\pm$0.02	**0.864**$\pm$0.01	**0.848**$\pm$0.01
Krippendorff's α						
Inter-Physician Agreement			0.754			

3.4 Safety (Binary) Classification Performance

We assess how well LMs can distinguish between safe (risk levels 1–2) and unsafe (levels 3–4) outputs in a binary deployment setting. As shown in Table 3b, MedVAL distillation consistently improves binary classification performance across most LMs. MedVAL improves average accuracy from 70.8% to 80.6% and F1 score from 66.2% to 82.8%, demonstrating more reliable discrimination between safe and unsafe responses. Notably, ensembling MedVAL LM outputs across all open-source models (rows 1–3) and, separately, across all proprietary models (rows 4–5) yields the highest F1/accuracy.

Among open-source LMs, the strongest gains are seen in Llama-3.2-3B (15.3% $\rightarrow$ 80.9%). Llama-3.1-8B improves from 68.8% to 80.4%, and Qwen3-4B from 80.0% to 82.3%. For proprietary models, GPT-4o Mini shows steady gains from 80.9% to 85.5%. GPT-4o, which already has a high baseline, (F1: 85.8%, accuracy: 84.6%), demonstrates only marginal changes (F1: 84.9%, accuracy: 84.3%).

3.5 MedVAL Performance Ablation

Our ablation (Figure S2) investigates the source of MedVAL's improvements, demonstrating that gains arise from the "combined effect" of frontier LM distillation and MedVAL's data filtering step. We evaluate two settings with no filtering ($\mathcal{M}_{\text{MedVAL}} \geq 0.0$) vs MedVAL filtering ($\mathcal{M}_{\text{MedVAL}} \geq 0.9$).

We evaluate whether MedVAL helps even without a frontier teacher. Across models, when training on the same LM's self-generated data, filtering consistently beats no filtering, and the self-distilled+MedVAL models also surpass their baselines. Notably, Llama-3.2-3B jumps from 12.8% $\rightarrow$ 22.1% F1 while using only 3% of the training set, whereas self-distilled without filtering moves just 12.8% $\rightarrow$ 13.6% despite using 100% of the data. This underscores that a small, high-quality subset can outperform far larger unfiltered corpora. We also see meaningful gains for Qwen3-4B (42.8% $\rightarrow$ 43.3%), Llama-3.1-8B (25.9% $\rightarrow$ 29.1%), and GPT-4o Mini (47.4% $\rightarrow$ 50.4%), indicating that a portion of the best GPT-4o-distilled performance is recovered by self-distilled+MedVAL.

Using GPT-4o as a teacher improves F1 even without filtering, as expected; adding MedVAL on top of distillation yields the largest gains while being data-efficient. With MedVAL filtering, we fine-tune on 57% examples (1,131/2,000) yet still outperform the unfiltered 100% training set, for example, Llama-3.2-3B (33.4% $\rightarrow$ 42.9%), Qwen3-4B (51.1% $\rightarrow$ 52.7%), Llama-3.1-8B (43.8% $\rightarrow$ 46.5%), and GPT-4o Mini (49.5% $\rightarrow$ 54.0%). GPT-4o Mini, self-distilled+MedVAL (50.4%) even beats vanilla GPT-4o distilled without MedVAL (49.5%), showing MedVAL's value beyond teacher strength. These results confirm that MedVAL is a data-efficient distillation method that can outperform vanilla distillation across student/teacher model variants.

3.6 External Validation and Comparison with Prior Metrics

To probe robustness beyond risk-grading, we evaluate MedVAL models on `MEDEC` [48], a benchmark for medical error detection. Despite the distribution shift, MedVAL improves over baselines: 53.3% $\rightarrow$ 54.4% for GPT-4o Mini and 58.0% $\rightarrow$ 63.3% for GPT-4o. Further, we compare MedVAL with two representative metrics from Table S1: 1) AlignScore [31], and 2) RadGraph F1 [44]. As shown in Figure S7, MedVAL correlates strongly with physicians (GPT-4o: $r = 0.825$, Qwen3-4B: $r = 0.833$), AlignScore is moderate ($r = 0.678$), and RadGraph F1 is weak ($r = 0.156$).

4 Conclusion

We introduce MedVAL, a novel, self-supervised distillation method for training LMs to validate LM-generated medical text. Our benchmark highlights substantial improvements in model performance and provides systematic evidence that LMs can achieve performance statistically non-inferior to a single human expert ($p < 0.001$). The MedVAL-Bench dataset, with its diverse tasks and physician annotations, establishes a robust benchmark for validating LM-generated medical text. Together, these contributions lay a foundation for developing and deploying safer, more trustworthy tools to automate the process of validating LM-generated medical text. The findings from this study suggest that incorporating MedVAL into clinical workflows could enhance the reliability of medical text generation, thereby improving patient safety and clinical efficiency. Future research should explore prospective studies to validate the practical benefits of MedVAL in real-world settings, potentially leading to an integrated approach that can support healthcare professionals and enhance patient care.

5 Data and Code Availability

We open-source the following resources developed during our study:

1. Codebase for fine-tuning and evaluation: github.com/StanfordMIMI/MedVAL
2. MedVAL-Bench dataset: huggingface.co/datasets/stanfordmimi/MedVAL-Bench
3. MedVAL-4B (best-performing fine-tuned Qwen3-4B): huggingface.co/stanfordmimi/MedVAL-4B

Datasets used in this study include MedicationQA [49], MeQSum [50], Open-i [51], MIMIC-IV [52], MIMIC-IV-BHC [53] and ACI-Bench [54–56]. All these datasets (except MIMIC) are fully open-source. The MIMIC-IV datasets require PhysioNet access and are currently undergoing review to be made available on PhysioNet. Further distribution of datasets is subject to the data sharing agreements stipulated by the original creators.

6 Funding

This research was, in part, funded by the Advanced Research Projects Agency for Health (ARPA-H). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Government. Furthermore, this work was supported in part by the Medical Imaging and Data Resource Center (MIDRC), which is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under contract 75N92020C00021 and through the ARPA-H.

AA is supported by NIH grant R01 HL167974 and ARPA-H contract AY2AX000045. ASC receives research support from NIH grants R01 HL167974, R01HL169345, R01 AR077604, R01 EB002524, R01 AR079431, P41 EB027060; ARPA-H contracts AY2AX000045 and 1AYSAX0000024-01; and NIH contracts 75N92020C00008 and 75N92020C00021. Unrelated to this work, ASC receives research support from GE Healthcare, Philips, Microsoft, Amazon, Google, NVIDIA, Stability; has provided consulting services to Patient Square Capital, Chondrometrics GmbH, and Elucid Bioimaging; is co-founder of Cognita; has equity interest in Cognita, Subtle Medical, LVIS Corp, and Brain Key; EA receives consulting fees from Fourier Health.

7 Acknowledgements

We acknowledge the support of the ARPA-H Chatbot Accuracy and Reliability Evaluation (CARE) project, as well as helpful discussions around stakeholder needs and technical solutions guided by program managers, Mansoor Khan, and Dr. Berkman Sahiner.

8 Author Contributions

AA collected data, developed code, ran experiments, designed studies, analyzed results, created figures, and wrote the manuscript. All authors reviewed the manuscript, providing meaningful revisions and feedback. VB, MV, NC, SO, AS, MP, and AK provided specific feedback on the manuscript and technical advice. AJ, KAM, EJPG, PNCR, SG, CB, EPR, EDZVR, PLH, KRK, MG, EL, and DBL participated in the reader study, with AJ, EPR, SG, CB, and EDZVR serving as radiologists, while the others participated as internal medicine physicians. EJPG, KAM, and PNCR participated as bilingual internal medicine physicians. DBL, CL, RD, JH, and SK advised on study design and provided feedback. EA and ASC guided the project and advised on technical details. RD and ASC obtained funding for this project and managed the overall project. No funders or third parties were involved in study design, analysis, or writing.

9 Competing Interest

No competing interests to declare.

References

[1] Joseph F Golob Jr, John J Como, and Jeffrey A Claridge. The painful truth: The documentation burden of a trauma surgeon. *Journal of Trauma and Acute Care Surgery*, 80(5):742–747, 2016.

[2] Milad Moradi and Matthias Samwald. Deep learning, natural language processing, and explainable artificial intelligence in the biomedical domain. *arXiv preprint arXiv:2202.12678*, 2022.

[3] 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. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[4] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30 (4):1134–1142, 2024.

[5] Asad Aali, Dave Van Veen, Yamin Ishraq Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash S Tehrani, et al. A dataset and benchmark for hospital course summarization with adapted large language models. *Journal of the American Medical Informatics Association*, 32(3):470–479, 2025.

[6] Khaled Saab, Jan Freyberg, Chunjong Park, Tim Strother, Yong Cheng, Wei-Hung Weng, David GT Barrett, David Stutz, Nenad Tomasev, Anil Palepu, et al. Advancing conversational diagnostic ai with multimodal reasoning. *arXiv preprint arXiv:2505.04653*, 2025.

[7] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

[8] Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760, 2016.

[9] Emily Gesner, Priscilla Gazarian, and Patricia Dykes. The burden and burnout in documenting patient care: an integrative literature review. *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 1194–1198, 2019.

[10] Raj M Ratwani, Erica Savage, Amy Will, Ryan Arnold, Saif Khairat, Kristen Miller, Rollin J Fairbanks, Michael Hodgkins, and A Zachary Hettinger. A usability and safety analysis of electronic health records: a multi-center study. *Journal of the American Medical Informatics Association*, 25(9):1197–1201, 2018.

[11] Jesse M Ehrenfeld and Jonathan P Wanderer. Technology as friend or foe? do electronic health records increase burnout? *Current Opinion in Anesthesiology*, 31(3):357–360, 2018.

[12] A Pal, LK Umapathi, and M Sankarasubbu. Med-halt: medical domain hallucination test for large language models. arxiv. *arXiv preprint arXiv:2307.15343*, 2023.

[13] Crystal T Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A Omiye, Akaash Kolluri, Akash Chaurasia, et al. Red teaming chatgpt in medicine to yield real-world insights on model behavior. *npj Digital Medicine*, 8(1):149, 2025.

[14] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. Factual consistency evaluation of summarization in the era of large language models. *Expert Systems with Applications*, 254:124456, 2024.

[15] Yubin Kim, Hyewon Jeong, Shen Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo R Gameiro, et al. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, pages 2025–02, 2025.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[17] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[18] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[19] Thomas R Yackel and Peter J Embi. Unintended errors with ehr-based result management: a case series. *Journal of the American Medical Informatics Association*, 17(1):104–107, 2010.

[20] Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2011, page 465. American Medical Informatics Association, 2011.

[21] Sue Bowman. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in health information management*, 10(Fall), 2013.

[22] M Sasseville, F Yousefi, S Ouellet, T Stefan, V Carnovale, F Bergeron, and A LeBlanc. Impacts of ai scribes on clinical outcomes, efficiency, and documentation. *Research Gate*, 2025.

[23] Oishi Banerjee, Agustina Saenz, Kay Wu, Warren Clements, Adil Zia, Dominic Buensalido, Helen Kavnoudias, Alain S Abi-Ghanem, Nour El Ghawi, Cibele Luna, et al. Rexamine-global: A framework for uncovering inconsistencies in radiology report generation metrics. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 185–198. World Scientific, 2024.

[24] Yongxin Zhou, Fabien Ringeval, and François Portet. A survey of evaluation methods of generated medical textual reports. In *The 5th Clinical Natural Language Processing Workshop*, pages 447–459. Association for Computational Linguistics, 2023.

[25] Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. Tethered to the ehr: primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426, 2017.

[26] Tait D Shanafelt, Lotte N Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P West. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. In *Mayo Clinic Proceedings*, volume 91, pages 836–848. Elsevier, 2016.

[27] Kenneth E Robinson and Joyce A Kersey. Novel electronic health record (ehr) education intervention in large healthcare organization improves quality, efficiency, time, and impact on burnout. *Medicine*, 97(38), 2018.

[28] Wiebke Toussaint, Dave Van Veen, Courtney Irwin, Yoni Nachmany, Manuel Barreiro-Perez, Elena Díaz-Peláez, Sara Guerreiro de Sousa, Liliana Millán, Pedro L Sánchez, Antonio Sánchez-Puente, et al. Design considerations for high impact, automated echocardiogram analysis. *arXiv preprint arXiv:2006.06292*, 2020.

[29] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[30] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

[31] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.

[32] Peiqi Wang, Barbara D Lam, Yingcheng Liu, Ameneh Asgari-Targhi, Rameswar Panda, William M Wells, Tina Kapur, and Polina Golland. Calibrating expressions of certainty. *arXiv preprint arXiv:2410.04315*, 2024.

[33] Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. *arXiv preprint arXiv:2404.00474*, 2024.

[34] Xin Liu, Farima Fatahi Bayat, and Lu Wang. Enhancing language model factuality via activation-based confidence calibration and guided decoding. *arXiv preprint arXiv:2406.13230*, 2024.

[35] Gaya Mehenni and Amal Zouaq. Medhal: An evaluation dataset for medical hallucination detection. *arXiv preprint arXiv:2504.08596*, 2025.

[36] Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. Doclens: Multi-aspect fine-grained evaluation for medical text generation. *arXiv preprint arXiv:2311.09581*, 2023.

[37] Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, et al. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*, 2025.

[38] Romain Hardy, Sung Eun Kim, Du Hyun Ro, and Pranav Rajpurkar. Rextrust: A model for fine-grained hallucination detection in ai-generated radiology reports. *arXiv preprint arXiv:2412.15264*, 2024.

[39] Vishwanatha M Rao, Serena Zhang, Julian N Acosta, Subathra Adithan, and Pranav Rajpurkar. Rexerr: Synthesizing clinically meaningful errors in diagnostic radiology reports. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 70–81. World Scientific, 2024.

[40] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*, 2024.

[41] Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. *arXiv preprint arXiv:2405.20613*, 2024.

[42] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

[43] 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, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

[44] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

[45] Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. Benchmarking and improving generator-validator consistency of language models. *arXiv preprint arXiv:2310.01846*, 2023.

[46] AMN Healthcare Services Inc. AMN Healthcare Study Tracks 45 Languages Spoken in Patient/Provider Encounters in U.S. Press release via GlobeNewswire, November 2023. `https://ir.amnhealthcare.com/news-releases/news-release-details/amn-healthcare-study-tracks-45-languages-spoken-patientprovider`.

[47] Giacomo Marzi, Marco Balzano, and Davide Marchiori. K-alpha calculator–krippendorff's alpha calculator: a user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12: 102545, 2024.

[48] Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. Medec: A benchmark for medical error detection and correction in clinical notes. *arXiv preprint arXiv:2412.19260*, 2024.

[49] TrueHealth. Medicationqa dataset. `https://huggingface.co/datasets/truehealth/medicationqa`, 2024. Accessed: 2025-04-29.

[50] Asma Ben Abacha and Dina Demner-Fushman. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28th - August 2*, 2019.

[51] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[52] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

[53] Asad Aali, Dave Van Veen, YI Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash Tehrani, et al. Mimic-iv-ext-bhc: labeled clinical notes dataset for hospital course summarization. *PhysioNet*, 2024.

[54] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586, 2023.

[55] Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. Overview of the mediqa-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, 2023.

[56] Wen-wai Yim, Asma Ben Abacha, Neal Snider, Griffin Adams, and Meliha Yetisgen. Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations. In *CLEF 2023 Working Notes*, CEUR Workshop Proceedings, Thessaloniki, Greece, September 18-21 2023. CEUR-WS.org.

[57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[58] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[59] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[60] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[61] OpenAI. Gpt-4 technical report, 2023.

[62] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[63] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[64] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[65] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[66] Dilara Soylu, Christopher Potts, and Omar Khattab. Fine-tuning and prompt optimization: Two great steps that work better together. *arXiv preprint arXiv:2407.10930*, 2024.

[67] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[68] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[69] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.

[70] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[71] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[72] Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*, 2025.

[73] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

[74] Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. Aligning large language models by on-policy self-judgment. *arXiv preprint arXiv:2402.11253*, 2024.

[75] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274, 2025.

[76] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.

[77] Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards evaluating and building versatile large language models for medicine. *npj Digital Medicine*, 8(1): 58, 2025.

[78] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

S1 Data

S1.1 Input Sources

Medication question answering. The `medication2answer` task involves answering a medication-related question, and utilizes the MedicationQA dataset [49], which contains real-world consumer health questions about medications and expert-written answers. Unlike other tasks, the input (a standalone question) often lacks sufficient context to validate the output's factual consistency. Hence, we aim to evaluate whether LMs can reliably perform safety-critical decisions using their knowledge base and the original question. Due to the unavailability of pre-specified train/test splits, we randomly create the splits as specified in Table 2.

Patient query summarization. The `query2question` task involves creating a concise query that encapsulates the essential details necessary to address the original question. We utilize the MeQSum dataset [50], which comprises: 1) health-related queries from patients, sourced from messages sent to the U.S. National Library of Medicine, and 2) corresponding abbreviated questions formulated by three medical experts to ensure that the summaries facilitate complete and accurate answer retrieval. For our study, we sample a subset of examples from both the train and test splits defined by Van Veen et al. [4], as detailed in Table 2.

Radiology report summarization. The `report2impression` task involves processing the "findings" section of a radiology exam report to create an "impression" section that succinctly highlights the critical information. We leverage the Open-i dataset [51], which consists of de-identified narrative chest x-ray reports from the Indiana Network for Patient Care database. We randomly sample train/test examples following the splits identified by Van Veen et al. [4], ensuring the number of train/test samples indicated in Table 2.

Radiology impression simplification. The `impression2simplified` task involves simplifying radiology report impression sections to produce patient-friendly rewrites that preserve the clinical meaning while being understandable to a layperson. For this task, we extract the "impression" sections from the MIMIC-IV dataset [52], which contains comprehensive radiology reports sourced from patient stays in critical care units of the Beth Israel Deaconess Medical Center. Our sampling process ensured an equal distribution of imaging modalities that reflect the top imaging indications at Stanford Hospital. These include chest x-rays, CT scans of the abdomen and pelvis, CT scans of the head, MR brain studies, ultrasound exams of the pelvis, digital screening mammography, and transabdominal/transvaginal pelvic ultrasounds. Test examples were selected via stratified random sampling as indicated in Table 2. Because MIMIC-IV [52] is partially open-source, we select training examples from Open-i radiology report findings on a related but distinct `report2simplified` task (i.e., simplifying the findings section). For this task, we randomly sample training examples from the training split by Van Veen et al. [4], ensuring no overlap with the samples used in `report2impression`.

Hospital course translation. The `bhc2spanish` task involves translating brief hospital course (BHC) sections from discharge summaries from English into Spanish. We employ the MIMIC-IV-BHC dataset [53], which contains curated BHC sections from discharge summaries, written by healthcare providers at the Beth Israel Deaconess Medical Center. The MIMIC-IV-BHC dataset includes diverse patient encounters, ensuring a comprehensive representation. We randomly sample testing examples as indicated in Table 2.

Doctor-patient dialogue summarization. The objective of `dialogue2note` is to summarize doctor-patient conversations into an "assessment and plan". We utilize the ACI-Bench dataset [54–56], which provides a collection of 1) 207 doctor-patient conversations and 2) patient visit notes. To ensure consistency, we randomly sample the testing examples (indicated in Table 2), from the test set defined by Van Veen et al. [4].

S1.2 Physician Reader Study

We design our reader study to enable a robust evaluation of how well LMs assess the factual consistency of outputs, compared to physicians. For each test-time input, an LM-generated output

was produced by sampling a perturbation level $\delta \in [0, 0.33, 0.67, 1.0]$ uniformly at random, and injecting the corresponding perturbation instruction into the prompt as described in Table 1. This ensured that the test set contained a balanced mix of outputs with no risk, low risk, moderate risk, and high risk errors. Since our objective was to evaluate the validator's ability to distinguish clinically meaningful degradation, we do not explicitly prompt the generator to produce only no-risk outputs, as doing so would instead assess generation quality. This design was necessary to evaluate a validator's ability to distinguish degradation across a meaningful severity spectrum. We observe that models sometimes diverge from requested perturbations; consequently, outputs often contain a blend of prompted and natural errors committed by LMs. In all cases, physician graders assign risk levels based solely on the observed text, independent of the prompt specification. Importantly, these perturbation rules apply to the test set, as the training set is curated using the pipeline in Figure 2.

To curate reference evaluations of LM outputs, we assembled a panel of 12 physicians, where 4 board-certified internal medicine physicians annotated general medical tasks (`medication2answer`, `query2question`, `dialogue2note`), 3 bilingual internal medicine residents annotated the `bhc2spanish` task, and 4 board-certified radiologists and a radiology resident annotated radiology tasks (`report2impression`, `impression2simplified`). For each study, physicians were presented with the input and LM-generated output (without indicating error injection), and requested to perform: 1) risk grading: "assign a risk level to the output, following the risk level taxonomy (between 1 and 4)", and 2) error assessment: "identify factual consistency errors in the output based on the error category taxonomy (hallucinations, omissions, or certainty misalignments)". 15 random examples from each task were annotated by multiple physicians, allowing us to measure inter-physician agreement.

We present the risk and error category distribution in Table S3 and Figure S6. The most common error categories are fabricated claim (45.7%), missing claim (14.0%), and incorrect recommendation (12.6%). We observe that error frequency rises with the risk grades (for grades 1–4, average 0.14 $\rightarrow$ 3.24 errors).

S2 Experimental Setup

S2.1 Language Models

We evaluate a diverse collection of state-of-the-art transformer-based LMs (Table S2). Our selection criteria included model license, context length, and size. For open-source models, we include Llama [57, 58] (Llama 3.1 8B, 3.2 3B, and 3.3 70B), Qwen3 [59] (4B dense version), and Gemma [60] (Gemma 3 27B and MedGemma 27B). For proprietary models, we include GPT-4o Mini, GPT-4o [61, 62], Claude Sonnet 4, and Gemini 2.0 Flash [63] for their enhanced performance capabilities [64].

S2.2 Implementation Details

Under open-source category, we select LMs under 8 billion parameters for fine-tuning, as they can be easily deployed on consumer-grade GPUs. We leverage DSPy's bootstrap fine-tuning algorithm [65, 66]. As part of our contribution, we extend DSPy's existing local parameter-efficient fine-tuning pipeline to enable Quantized Low-Rank Adaptation (QLoRA) (GitHub PR) [67]. We employ an NVIDIA A6000 GPU and fine-tune models for 5 epochs using the Adam optimizer [68]. We initialize the learning rate at 1×10^{-5} with linear decay. We set the per-device train batch size to 1 and apply 4-bit precision quantization using BitsAndBytesConfig [69].

Under proprietary LMs, we report zero-shot baseline performance and only fine-tune OpenAI models that allow fine-tuning, where DSPy collects training data and initiates a fine-tuning procedure managed by OpenAI. Since OpenAI handles the parameter-efficient fine-tuning internally, exact details are not publicly available.

Unless specified, the MedVAL framework employs GPT-4o as the teacher LM, i.e., the curator of synthetic data under stages 1 and 2 from Figure 2, and filters synthetic data using $\mathcal{M}_{\text{MedVAL}} \geq \tau$, where $\tau = 0.9$. This value was selected heuristically based on inspection of agreement scores to reflect a high-consistency cutoff.

S2.3 Statistical Analysis

We compute the F1 score for each risk level and task independently, comparing LM-predicted risk classifications to physician labels. The F1 score is a measure of a model's accuracy that considers both precision (the correctness of the positive predictions) and recall (the ability to identify all positive instances). F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates the worst performance with no correct positive predictions. We report F1 score using the macro average, ensuring equal weighting for risk levels and tasks to help mitigate class imbalance. For evaluating binary safety classification, we additionally report sensitivity (unsafe recall) and specificity (safe recall) to make class balance explicit.

We also report Cohen's κ, allowing us to assess a LM's agreement with physicians. Cohen's κ [70] is a statistical measure of agreement for categorical items, which takes into account the possibility of agreement occurring by chance. Given the ordinal nature of the four-class validation task, we compute Cohen's κ with linear weighting, comparing model-predicted risk levels to physician-assigned levels (1–4). The κ score ranges from -1 to 1, where 1 represents perfect agreement, 0 indicates agreement no greater than chance, and -1 indicates complete disagreement. We use Krippendorff's α [71] to assess inter-physician agreement over a subset of examples annotated by multiple physicians, evaluating the robustness of the physician annotations. Krippendorff's α ranges from 0 to 1, where 1 indicates perfect agreement among annotators.

To assess whether MedVAL significantly improves performance over baseline models, we apply the McNemar test, a non-parametric test for paired nominal data that detects differences in classification outcomes. We evaluate whether MedVAL and the baseline model differ in their ability to correctly classify examples. To account for multiple comparisons (10 models $\times$ 4 risk levels), we apply a Bonferroni correction with a Type I error rate of $\alpha = 0.05$. All statistical analyses were conducted using Python's `statsmodels` package.

To test the non-inferiority (NI) of a LM against a single human expert, we use the 90/840 test cases with multiple-physician ratings. To identify a reference label, we use majority consensus. A "single expert" is defined by randomly sampling one physician's label. We compute macro-F1 against the reference for both the model and sampled expert, form the NI contrast Δ = F1 (model) - F1 (single expert), and use a paired item-level bootstrap (B=10,000). Non-inferiority is declared if $\text{LCB}_{95\%}(\Delta) > -0.05$ (clinically negligible difference). We did not apply a multiplicity adjustment as our NI comparison was GPT-4o vs a single expert.

S3 Discussion

We introduce MedVAL, a novel, self-supervised, and data-efficient distillation method for training LMs to validate LM-generated medical text following a physician-defined taxonomy. By combining synthetic data generation, data filtering, and fine-tuning, MedVAL improves a LM's baseline capacity across diverse medical tasks. We validate MedVAL on MedVAL-Bench, a benchmark of 840 physician-annotated examples spanning 6 medical tasks. MedVAL extends prior methods by providing a framework that: 1) trains scalable evaluators without physician-in-loop supervision, 2) assesses medical text in the absence of reference outputs or retrieval, 3) supports multilingual evaluation, and 4) offers interpretable, expert-aligned assessments.

Similar to MedVAL, there exist automated methods for evaluating LM-generated medical text that have surfaced. However, these methods are either not medical domain-specific or require physicians-in-the-loop, the availability of reference outputs, or external knowledge bases for retrieval. MedVAL does not require either of these constraints. FActScore [30] uses domain-specific extractors for checking claims, though it relies on structured knowledge bases. AlignScore [31] proposes a unified alignment function for factuality assessment across tasks such as NLI, QA, and summarization. These methods, including model confidence calibration approaches [32–34], often lack the necessary nuance for clinically focused error assessment. In medical settings, there is often a continuum of right and wrong that requires a very nuanced understanding, making it uniquely challenging compared to general domains. Several studies have focused on error detection for LM-generated medical text. MedHAL [35], a benchmark for hallucination detection in medical text, relies on physician error annotations limited to specific curated medical tasks. DocLens [36] introduces multi-aspect metrics—completeness, conciseness, and attribution—tailored for medical text generation. While they

report higher agreement with physician assessments than existing metrics, their evaluation relies on the availability of reference outputs. VeriFact [37] uses retrieval-based evidence to verify statements and is oriented towards multi-document summarization without the capability to train LMs. Recent studies have also focused on error assessment for text generation in radiology: ReXTrust [38] presents a fine-grained hallucination detector, ReXErr [39] injects clinical errors into radiology reports, GREEN [40] identifies significant errors in radiology reports, and FineRadScore [41] evaluates radiology report generation. While these efforts are crucial for identifying errors, their focus on radiology tasks (e.g., chest X-rays [42–44]) and reliance on reference outputs limits their generalizability. A detailed comparison of these methods is summarized in Table S1.

Beyond methodological differences, real-world impact hinges on cost, latency, and privacy constraints. Hence, improving small/open models is directly relevant for real-world deployment as hospital systems face barriers to using proprietary APIs at scale. In agentic workflows, validation is a high-frequency step (per section, per note, per agent action), so efficient models are a practical path for routine validation, with frontier models reserved for escalations [72]. Furthermore, prior work has shown that modest-sized judges can track stronger teachers effectively [73] and on-policy self-judgment provides a privacy-preserving, label-efficient route to reliable evaluators [74]. MedVAL directly addresses these constraints by enabling data-efficient distillation, enabling small/open models to closely track frontier performance at a fraction of the cost and latency.

S4 Limitations

Despite MedVAL's broad applicability, our study has limitations. Our prompts for perturbation injection were manually designed, potentially yielding partly "simulated" perturbations. Our goal was to enrich the test set with error families commonly observed by physicians [75], not replace natural errors. Nevertheless, we acknowledge that in-the-wild evaluations remain future work.

While our input-based validation approach enables reference-free evaluation, this can limit effectiveness in tasks where the input lacks sufficient information, such as question-answering (`medication2answer`). For QA tasks, we aimed to explore whether safety-critical decisions can be performed, regardless of whether the knowledge originates from the input or the LM's knowledge base. Our framework demonstrated improved alignment with physicians in the `medication2answer` task, suggesting that MedVAL can enhance a LM's ability to assess outputs leveraging its knowledge base.

Our task scope is not exhaustive. We evaluate six medical tasks that directly support risk-based triage. We acknowledge that this suite does not encompass the full spectrum of clinical document types and scenarios, such as EHR tabular reasoning, complex free-text reasoning, and consultation dialogues. Expanding MedVAL-Bench to these settings requires additional physician adjudication under our risk schema, which is a practical bottleneck; we therefore leave broader evaluation to future work, including representative tasks from MedHELM [76], MedS-Bench [77], and HealthBench [78]. Another limitation is that we use a single-pass consistency filter to select high-confidence training pairs, limiting early-stage noise amplification. While distillation using one-shot filtered data consistently outperforms distillation using all data unfiltered, iterative refinement schemes where the validator improves progressively with the filtered dataset could further improve data selection. However, we leave a systematic study of such strategies to future work.

Furthermore, a reasonable concern regarding our evaluation is that improvements could stem from simply learning the perturbation distribution. However, while supervision is synthetic (teacher signals filtered by generator-validator consistency), all evaluation uses physician labels. Moreover, gains persist on held-out tasks excluded from distillation (`impression2simplified`, `bhc2spanish`, and `dialogue2note`), which introduce distribution shifts. Together, these results suggest generalization beyond the perturbation style. Regardless, we leave a controlled study isolating perturbation design effects from error generalization to future works. Additionally, inter-physician agreement in our study was estimated on a stratified multi-annotated subset (15 examples $\times$ 6 tasks; 90/840) due to physician time constraints. While this design provides per-task coverage, it is a sparse sample; reliability estimates using Krippendorff's α should be interpreted with this caveat. Finally, while we selected representative open-source and proprietary models for fine-tuning under practical resource settings, results may differ with larger models or other closed platforms. However, we open-source our code and trained models to allow the community to validate on broader tasks and models.

Table S1: **MedVAL comparison with prior methods** on LM-generated text evaluation.

Method	Focus	Train -able	Physician -free training	Reference -free evaluation	Retrieval -free evaluation	Multi -lingual evaluation
FActScore	General	✓	✓	✓	✗	✗
AlignScore	General	✓	✓	✓	✓	✗
FineRadScore	Radiology	✗	✓	✗	✓	✗
ReXTrust	Radiology	✓	✗	✗	✓	✗
GREEN	Radiology	✓	✓	✗	✓	✗
RadGraph	Radiology	✓	✗	✗	✓	✗
VeriFact	BHC	✗	✓	✓	✗	✗
DocLens	Medical	✗	✓	✗	✓	✗
MedHAL	Medical	✓	✗	✗	✓	✗
MedVAL	Medical	✓	✓	✓	✓	✓

Algorithm S1 MedVAL self-supervised training

Require: Frozen generator g_θ, frozen validator v_ϕ, fine-tunable validator v_α, inputs $\mathcal{D} = \{x_i\}$, threshold τ

Ensure: Trained validator v_α^*

Initialize training dataset $\mathcal{D}_{\text{train}} \leftarrow \emptyset$
for $x \in \mathcal{D}$ **do**
  $\delta \leftarrow \textit{RandomChoice}(\{\delta_1, \delta_2, \ldots, \delta_L\} \mid \delta \in [0, 1])$
  $\hat{y} \leftarrow y$ if available, else $g_\theta(x)$ ▷ Unperturbed output
  $\hat{y}_\delta \leftarrow g_\theta(x_\delta)$ ▷ Perturbed output
  $\hat{\delta}_{\text{clean}} \leftarrow v_\phi(x, \hat{y})$ ▷ Factual degradation of $\hat{y}$ in comparison to x
  $\hat{\delta}_{\text{corrupt}} \leftarrow v_\phi(x, \hat{y}_\delta)$ ▷ Factual degradation of $\hat{y}_\delta$ in comparison to x
  Compute $\mathcal{M}_{\text{absolute}} \leftarrow \|\hat{\delta}_{\text{clean}}\|_2^2 + \|\hat{\delta}_{\text{corrupt}} - \delta\|_2^2$ ▷ Absolute consistency
  Compute $\mathcal{M}_{\text{relative}} \leftarrow \|\hat{\delta}_{\text{corrupt}} - \hat{\delta}_{\text{clean}} - \delta\|_2^2$ ▷ Relative consistency
  $\mathcal{M}_{\text{MedVAL}} \leftarrow 1 - \frac{1}{6}(\mathcal{M}_{\text{absolute}} + \mathcal{M}_{\text{relative}})$ ▷ Generator-validator consistency score (0-1)
  if $\mathcal{M}_{\text{MedVAL}} \geq \tau$ **then**
    $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{x, \hat{y}, \hat{\delta}_{\text{clean}}\}$
    $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{x, \hat{y}_\delta, \hat{\delta}_{\text{corrupt}}\}$
  end if
end for
$v_\alpha^* = \textit{SFT}(v_\alpha, \mathcal{D}_{\text{train}})$ ▷ Supervised fine-tuning
return v_α^*

Table S2: **Language models** included in our study for benchmarking and evaluation.

Language model	Context	Parameters	Open-source?	Medical adapted?
Llama 3.2	128k	3B	✓	✗
Qwen3	128k	4B	✓	✗
Llama 3.1	128k	8B	✓	✗
Gemma 3	128k	27B	✓	✗
MedGemma	128k	27B	✓	✓
Llama 3.3	128k	70B	✓	✗
GPT-4o Mini	128k	unknown	✗	✗
GPT-4o	128k	unknown	✗	✗
Claude Sonnet 4	200k	unknown	✗	✗
Gemini 2.0 Flash	1000k	unknown	✗	✗

Table S3: **MedVAL-Bench test set and physician risk grading.** For each task, cells report the percentage of test examples assigned to each risk level by physicians.

Task	Level 1 (%)	Level 2 (%)	Level 3 (%)	Level 4 (%)	N
`medication2answer`	36.3%	11.1%	22.2%	30.4%	135
`query2question`	38.3%	20.8%	28.3%	12.5%	120
`report2impression`	34.2%	19.5%	21.1%	25.3%	190
`impression2simplified`	32.6%	30.0%	14.7%	22.6%	190
`bhc2spanish`	38.3%	11.7%	9.2%	40.8%	120
`dialogue2note`	47.1%	12.9%	15.3%	24.7%	85
Overall	**36.7%**	**18.9%**	**18.6%**	**25.8%**	**840**

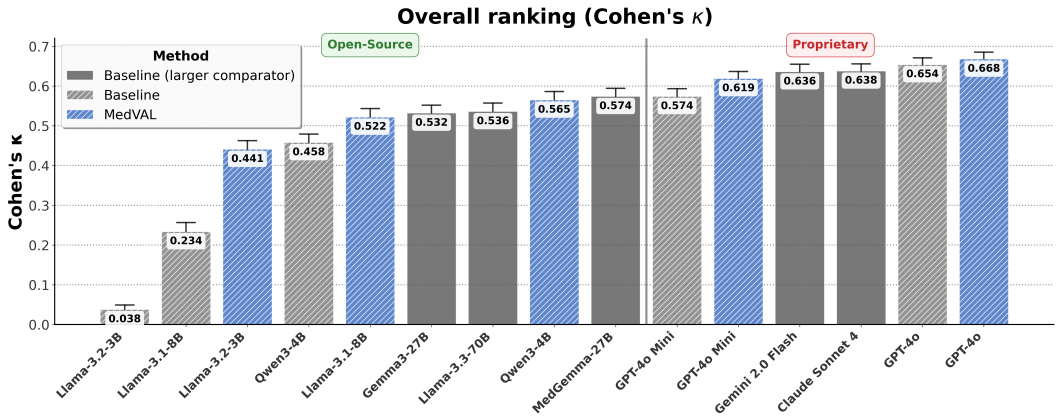

Figure S1: **Performance benchmark (Cohen's κ)**. We rank all LMs (low to high). *Baseline* indicates zero-shot LM before distillation, *Baseline (larger comparator)* indicates a larger zero-shot LM as reference (not chosen for distillation), and *MedVAL* indicates LM after distillation. Notably, smaller MedVAL LMs match or exceed the performance of much larger baselines.

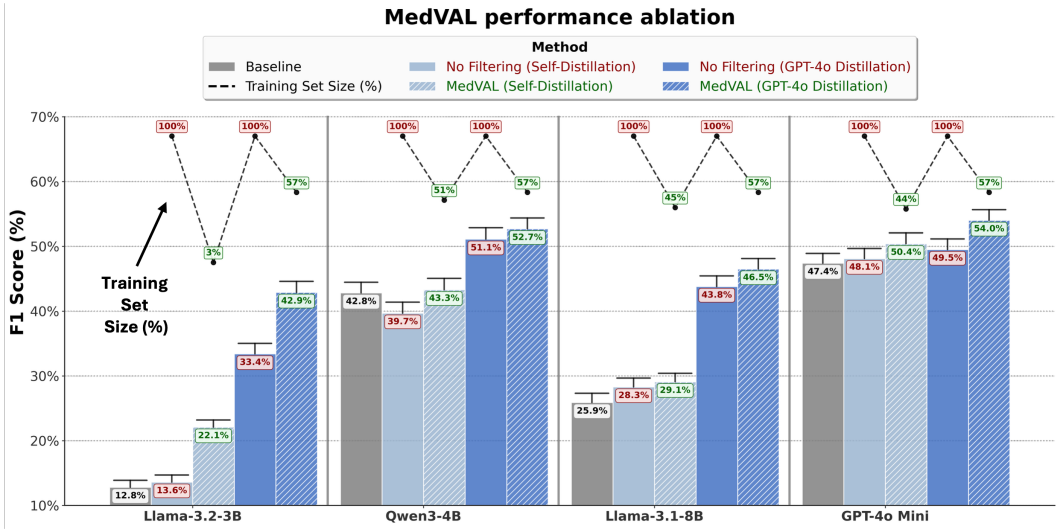

Figure S2: **MedVAL performance ablation**. We report the performance with: **a)** self-distillation (for synthetic data generation) without MedVAL filtering $\mathcal{M}_{\text{MedVAL}} \geq 0.0$, **b)** self-distillation with MedVAL filtering $\mathcal{M}_{\text{MedVAL}} \geq 0.9$, **c)** GPT-4o distillation without MedVAL filtering $\mathcal{M}_{\text{MedVAL}} \geq 0.0$, and **d)** GPT-4o distillation with MedVAL filtering $\mathcal{M}_{\text{MedVAL}} \geq 0.9$. The ablation confirms that MedVAL is a data-efficient distillation method that can outperform vanilla distillation across models.

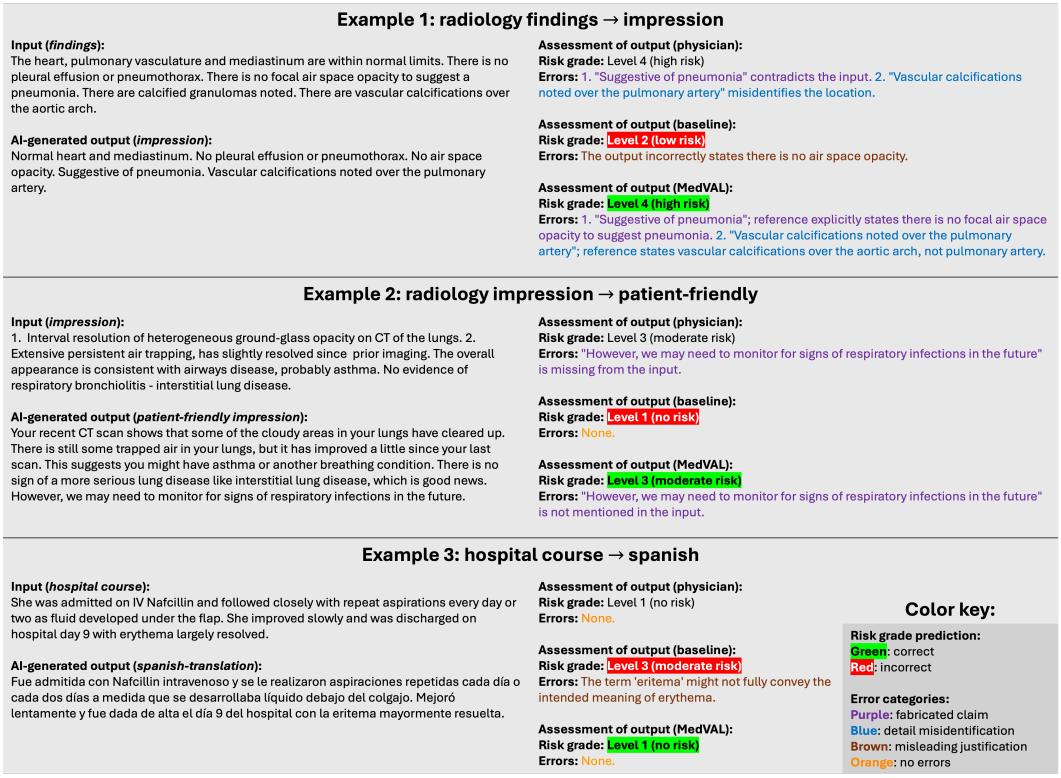

Figure S3: **Representative examples of validation of LM-generated medical text** by 1) the physician, 2) baseline GPT-4o, and 3) MedVAL fine-tuned GPT-4o. Under each example, MedVAL demonstrates higher agreement with the physician.

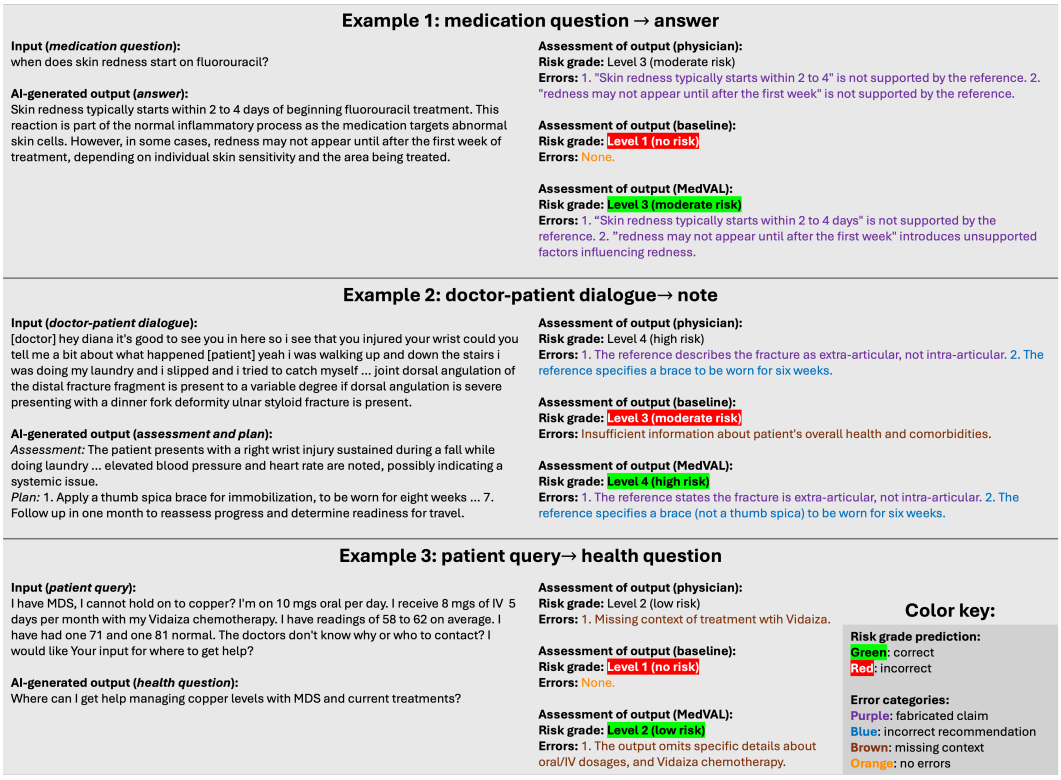

Figure S4: **Representative examples of validation of LM-generated medical text** by 1) the physician, 2) baseline GPT-4o, and 3) MedVAL GPT-4o. Under each example, MedVAL demonstrates higher agreement with the physician.

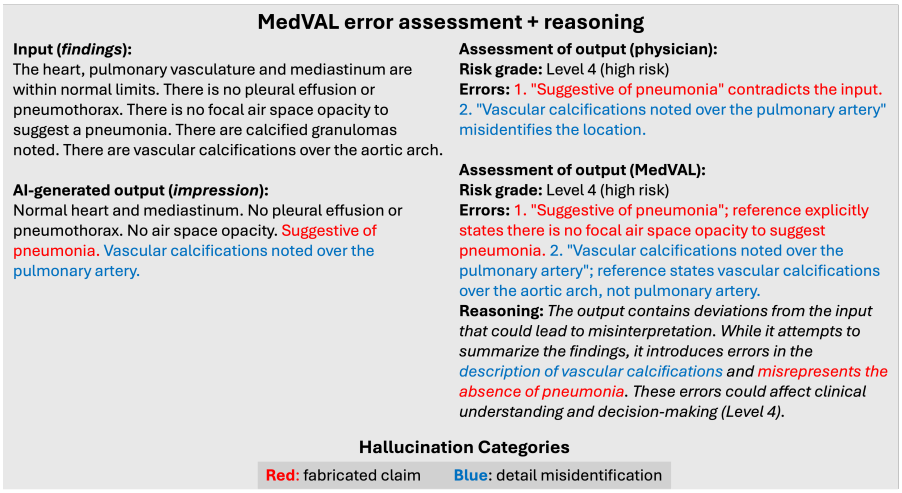

Figure S5: **Representative example of validation of LM-generated medical text** by 1) the physician, and 2) MedVAL GPT-4o. MedVAL demonstrates full agreement with the physician while also providing a "reasoning" for its risk grading.

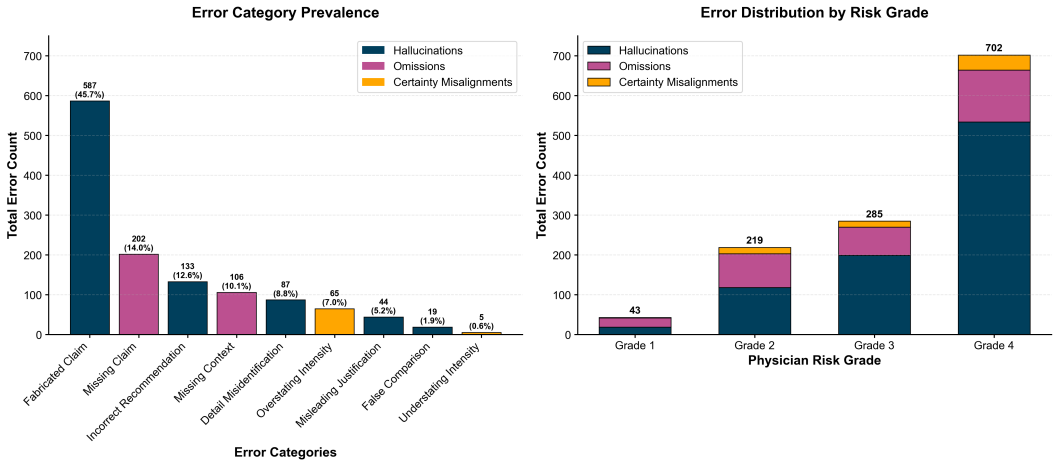

Figure S6: **Error distribution in the test set.** (Left) Total count and item prevalence of each physician-annotated error category. (Right) Stacked counts by physician risk grade, where error burden increases sharply with higher risk grades.

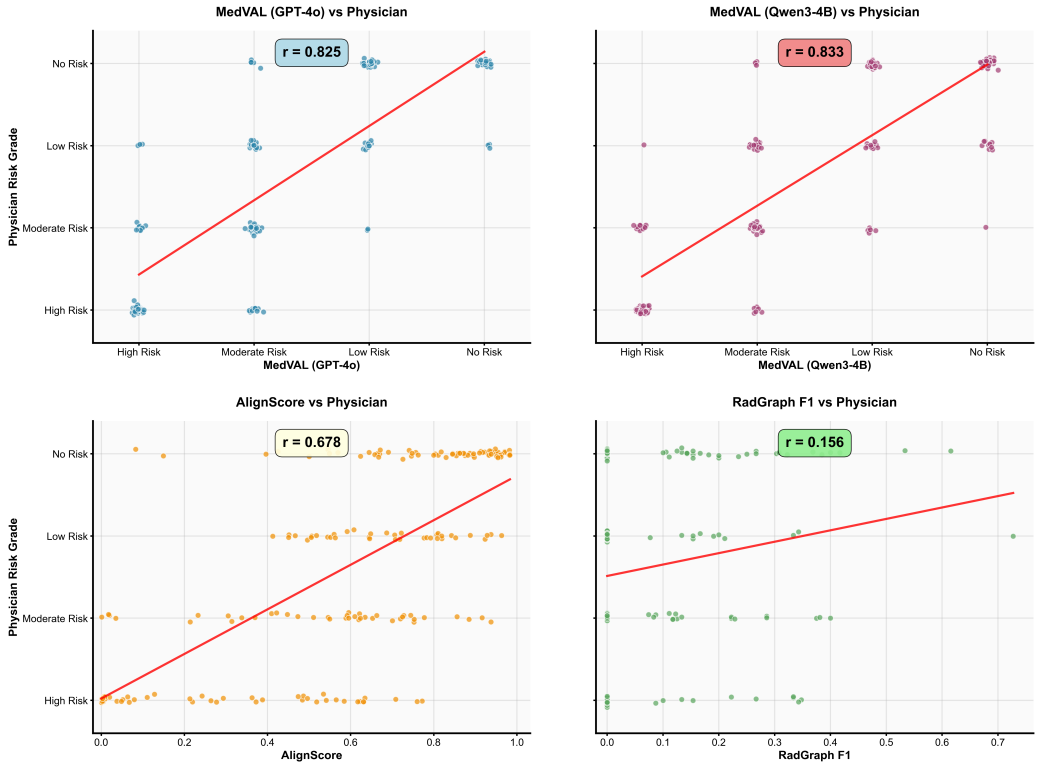

Figure S7: **Metric correlation (Pearson r) with physician risk grades.** Across 190 examples from the `report2impression` subset, MedVAL strongly correlates with physicians, AlignScore is moderate, and RadGraph F1 (a reference-based score) is weak.

Table S4: **Task-wise standard deviation (F1 score).** Bootstrapped standard deviations reflect variability across predictions.

Method	Model	In-Distribution			Out-of-Distribution			Overall
		medication2 answer	query2 question	report2 impression	impression2 simplified	bhc2 spanish	dialogue2 note	
Open-Source								
Baseline	Llama-3.2-3B	0.011	0.014	0.022	0.016	0.033	0.039	0.011
	Qwen3-4B	0.032	0.046	0.034	0.035	0.050	0.055	0.015
	Llama3.1-8B	0.042	0.037	0.033	0.024	0.028	0.030	0.013
	Gemma3-27B	0.031	0.036	0.036	0.027	0.041	0.054	0.014
	MedGemma-27B	0.042	0.041	0.034	0.042	0.034	0.050	0.015
	Llama-3.3-70B	0.032	0.047	0.035	0.038	0.039	0.056	0.017
MedVAL	Llama-3.2-3B	0.040	0.033	0.036	0.034	0.037	0.047	0.017
	Qwen3-4B	0.047	0.045	0.032	0.032	0.048	0.046	0.017
	Llama-3.1-8B	0.046	0.048	0.029	0.030	0.042	0.045	0.017
Proprietary								
Baseline	GPT-4o Mini	0.034	0.045	0.035	0.032	0.033	0.056	0.015
	GPT-4o	0.042	0.042	0.036	0.035	0.033	0.054	0.017
	Claude Sonnet 4	0.041	0.051	0.035	0.035	0.048	0.057	0.017
	Gemini 2.0 Flash	0.032	0.047	0.034	0.030	0.045	0.048	0.016
MedVAL	GPT-4o Mini	0.039	0.028	0.032	0.036	0.041	0.063	0.018
	GPT-4o	0.047	0.031	0.031	0.033	0.039	0.059	0.018