MedVAL
Toward Expert-Level Medical Text Validation with Language Models

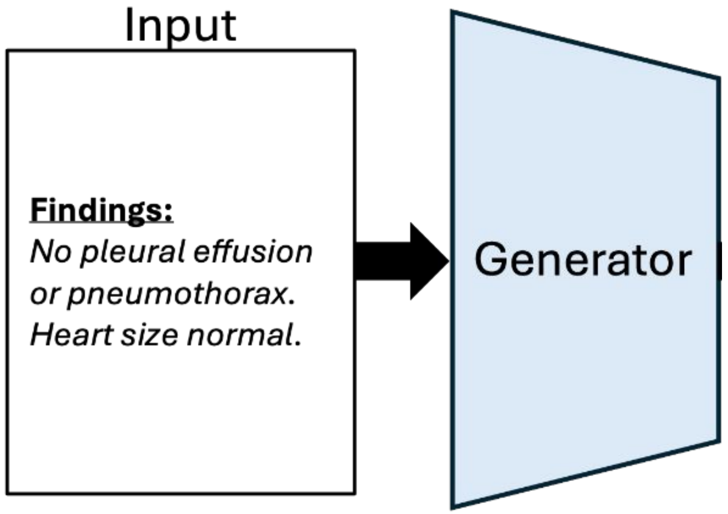
Asad Aali
Research Talk
Stanford AI+Biomedicine Seminar
August 26, 2025

Introduction

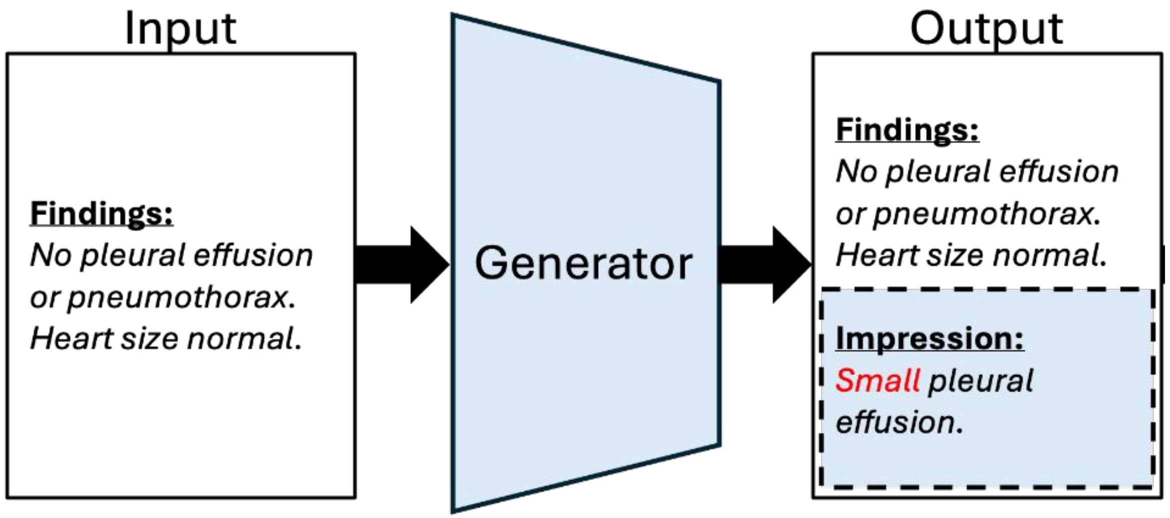
Input

Findings:
No pleural effusion or pneumothorax. Heart size normal.

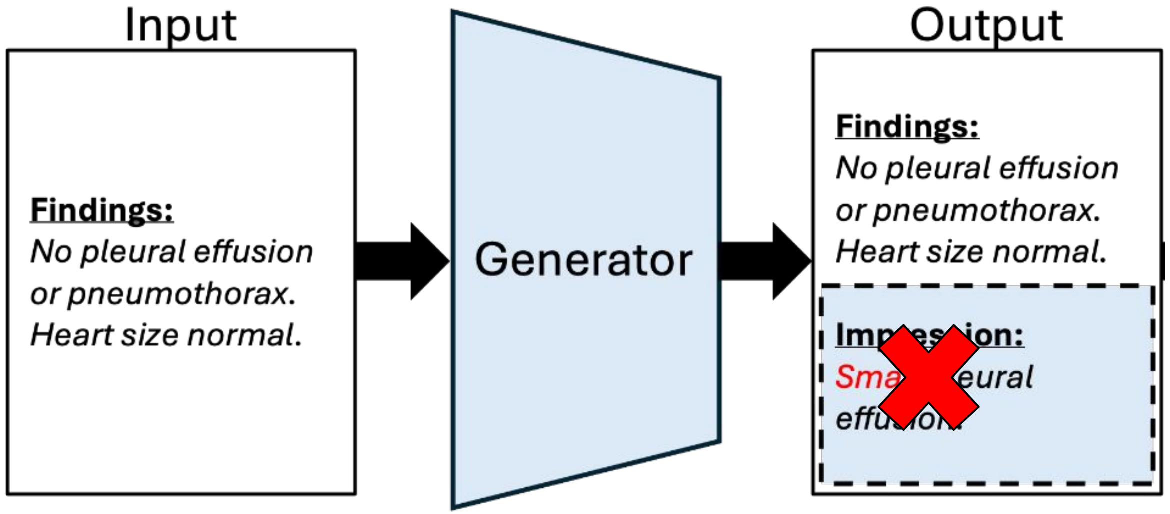
Introduction



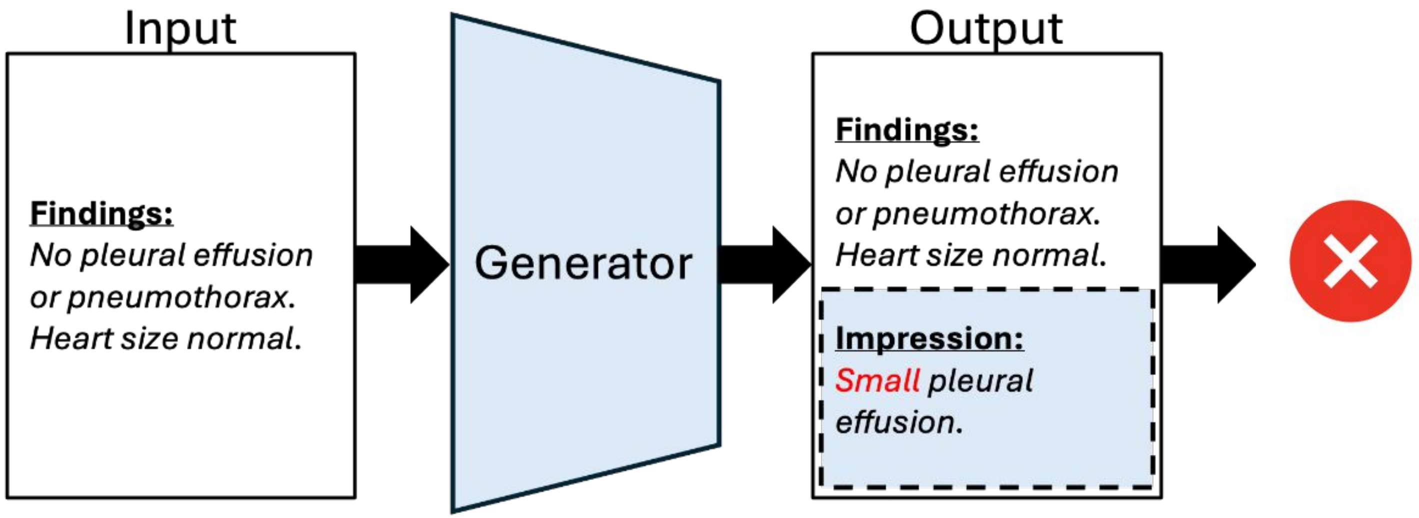
Introduction



Introduction



Introduction

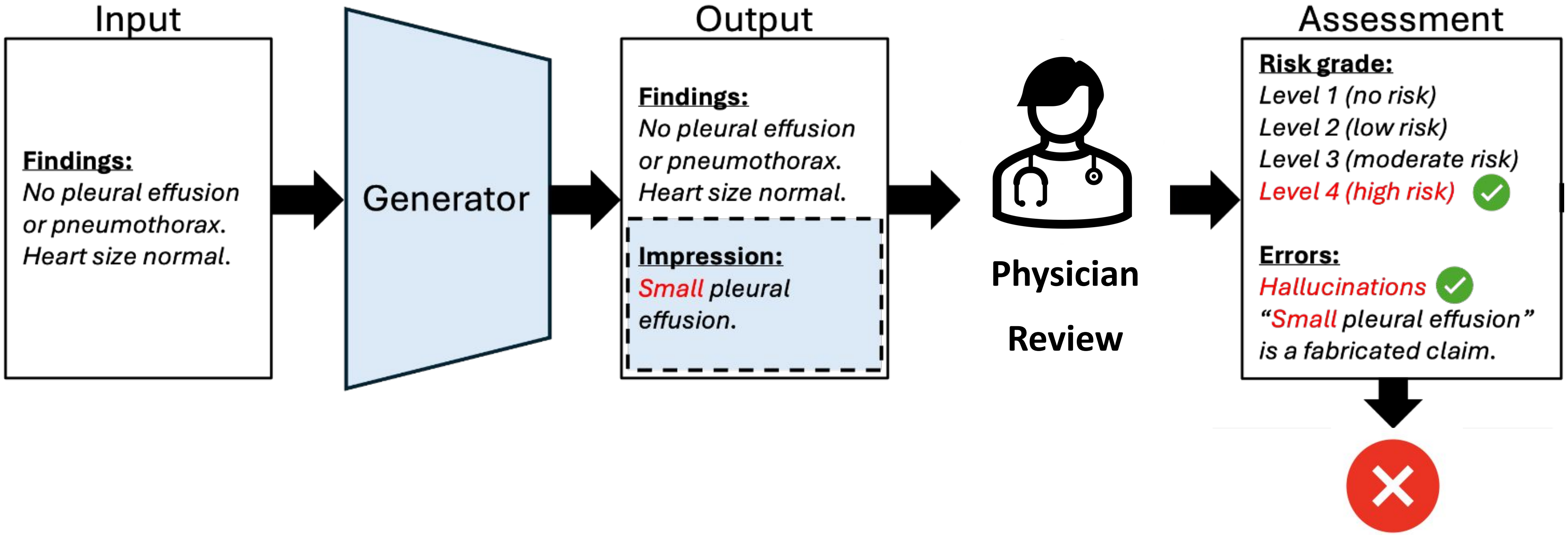


1. The output *may* contain errors
2. The output *may* not be safe for deployment

The adoption of AI for medical applications necessitates ***reliable risk assessment***

Introduction

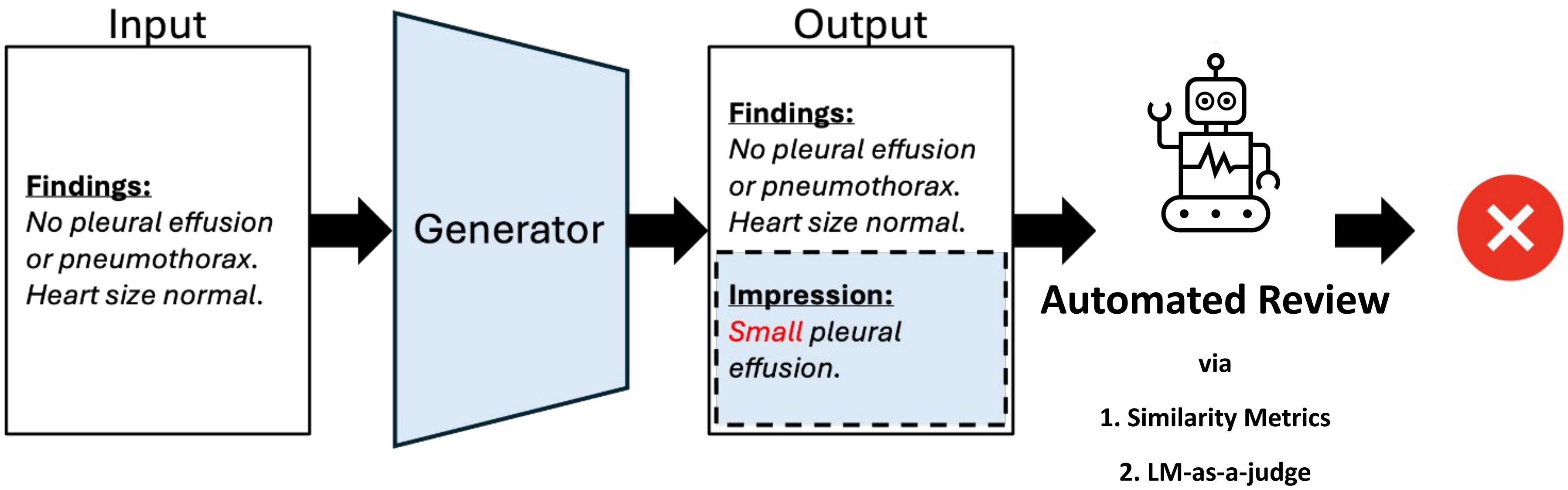
Potential Solution 1



1. *Time-consuming / expensive*
2. Not *scalable*
3. Physician *fatigue*

Introduction

Potential Solution 2



Automated methods **rely** on:

1. ***Expert-labeled training*** data
2. ***Ground truth outputs*** for comparison (similarity metrics)
3. ***Retrieval-based*** evidence

Review (Related Works)

Method
FActScore
AlignScore
FineRadScore
ReXTrust
GREEN
VeriFact
DocLens
MedHAL

Review (Related Works)

Method	Focus
FActScore	General
AlignScore	General
FineRadScore	Radiology
ReXTrust	Radiology
GREEN	Radiology
VeriFact	BHC
DocLens	Medical
MedHAL	Medical

Review (Related Works)

Method	Focus	Medical domain
FActScore	General	✗
AlignScore	General	✗
FineRadScore	Radiology	✓
ReXTrust	Radiology	✓
GREEN	Radiology	✓
VeriFact	BHC	✓
DocLens	Medical	✓
MedHAL	Medical	✓

Review (Related Works)

Method	Focus	Medical domain	Train -able
FActScore	General	✗	✓
AlignScore	General	✗	✓
FineRadScore	Radiology	✓	✗
ReXTrust	Radiology	✓	✓
GREEN	Radiology	✓	✓
VeriFact	BHC	✓	✗
DocLens	Medical	✓	✗
MedHAL	Medical	✓	✓

Review (Related Works)

Method	Focus	Medical domain	Train-able	Physician-free training
FActScore	General	✗	✓	✓
AlignScore	General	✗	✓	✓
FineRadScore	Radiology	✓	✗	✓
ReXTrust	Radiology	✓	✓	✗
GREEN	Radiology	✓	✓	✓
VeriFact	BHC	✓	✗	✓
DocLens	Medical	✓	✗	✓
MedHAL	Medical	✓	✓	✗

Review (Related Works)

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

Review (Related Works)

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

Review (Related Works)

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

Introducing MedVAL
Medical Text Validator (≠ Evaluator)

Medical text validation: Determining whether an AI's ***output*** is ***factually consistent*** with the ***input*** (binary) ✓

vs

Medical text evaluation: Assessing several attributes of an AI's output (conciseness, comprehensiveness) ✗

Review (Related Works)

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

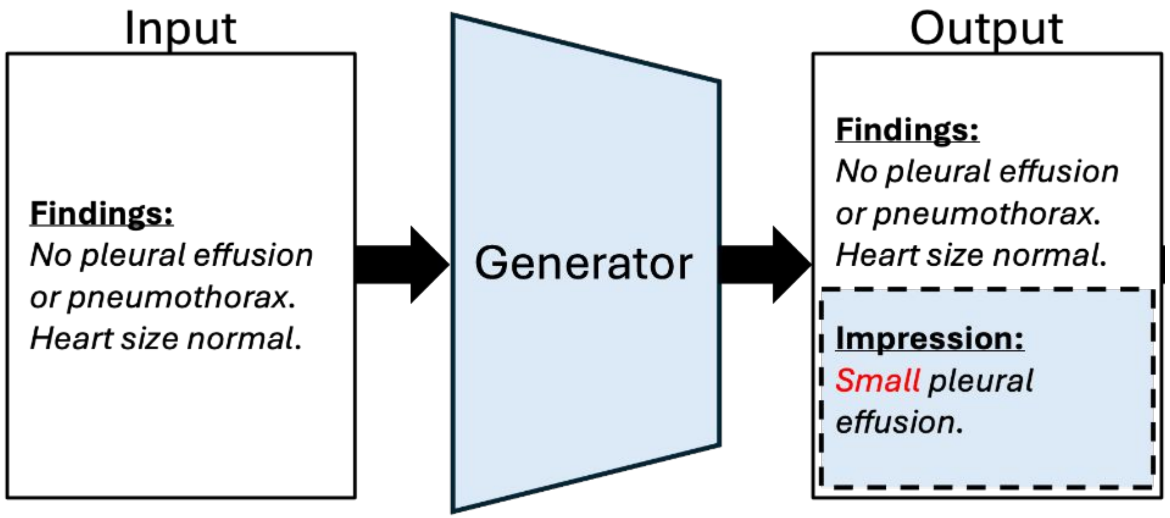
Review (Related Works)

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

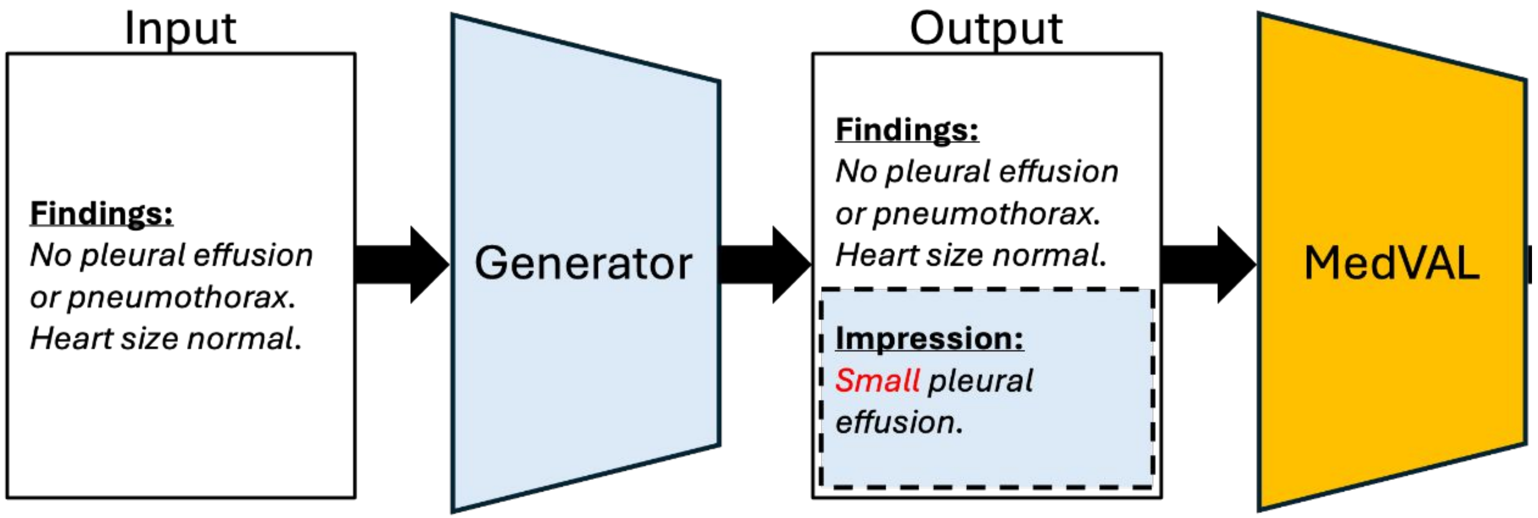
Introducing MedVAL

- A **self-supervised** framework that leverages **synthetic data** to train LMs for robust medical text validation
  - Involves curating **high-quality** synthetic training examples
  - Leverages the **agreement between a generator and a validator** LM as a **proxy for physician judgment**
- MedVAL assesses whether an **output is factually consistent** with the **input**
  - Assigns one of **four risk levels**
  - Flags **"unsafe for deployment"** outputs at **near physician-level reliability**

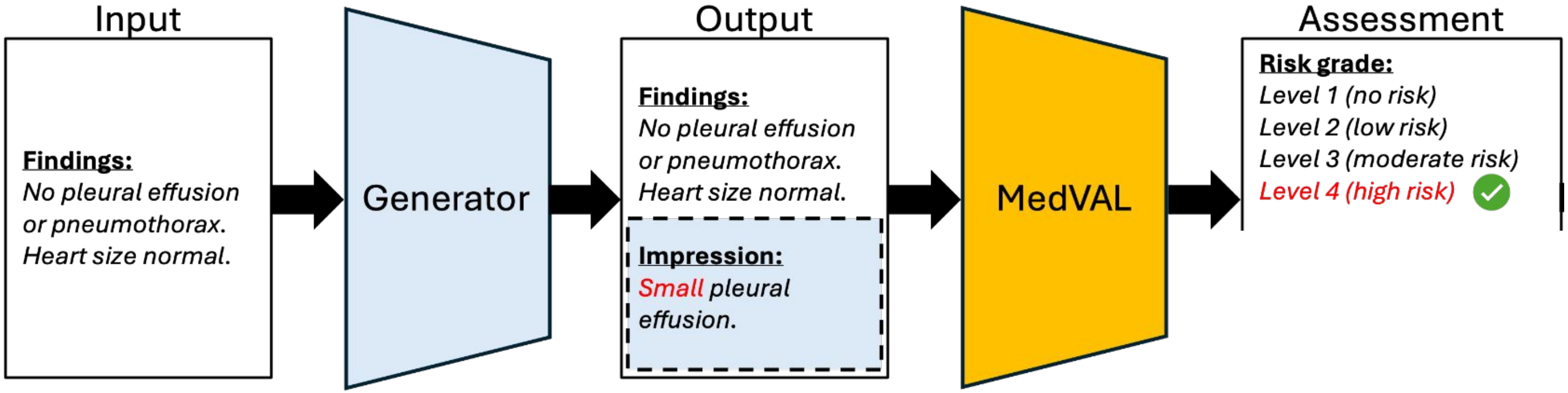
Introducing MedVAL



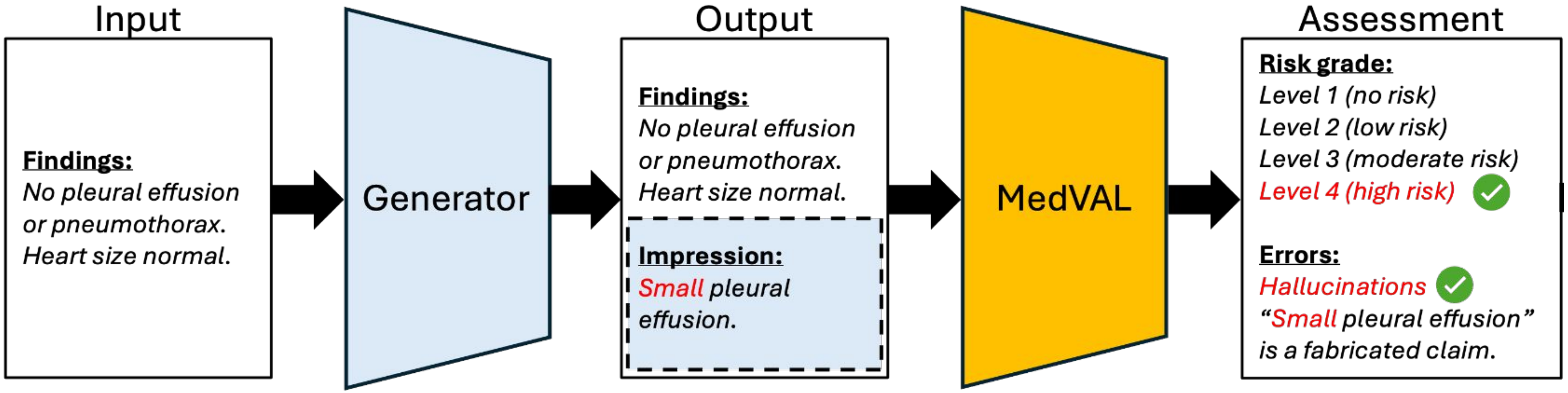
Introducing MedVAL



Introducing MedVAL

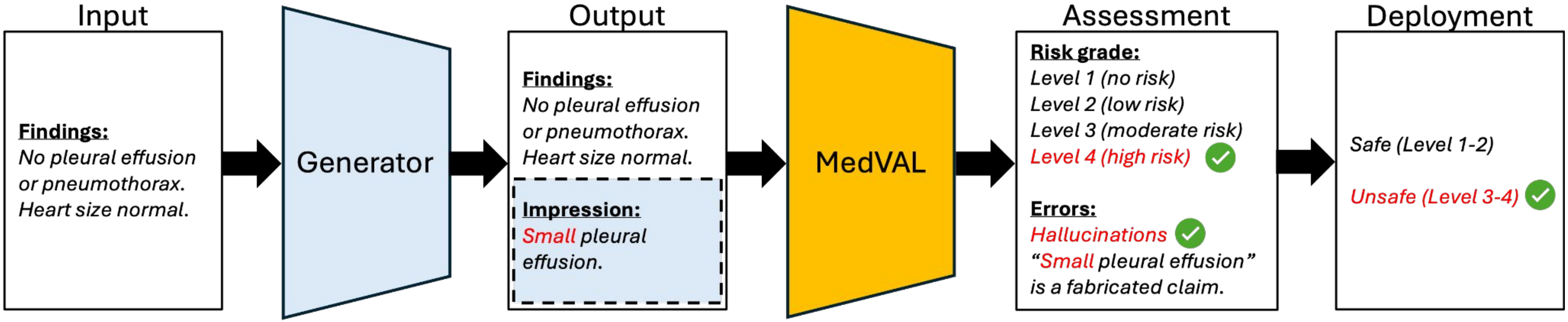


Introducing MedVAL



Introducing MedVAL

1. Scalable training **without physician-in-loop** supervision
2. Medical text assessment in the **absence of reference outputs** or **retrieval**
3. **Multilingual** evaluation
4. **Interpretable, expert-aligned** assessments



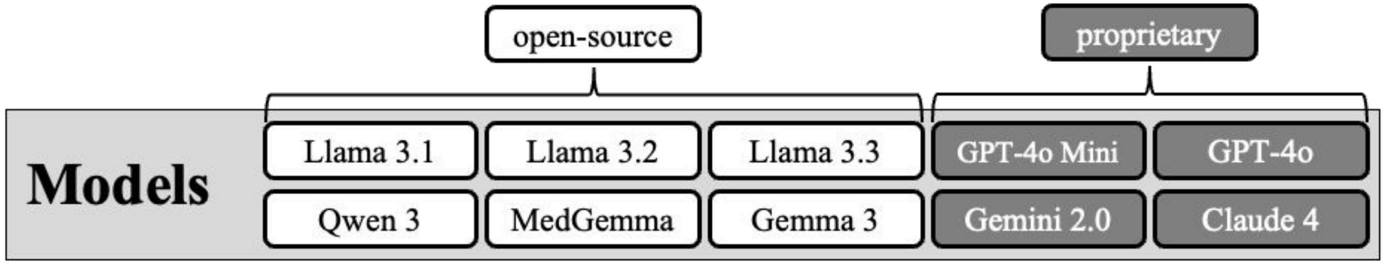
Contributions

1. A general-purpose, self-supervised framework for training LMs to **validate factual consistency**

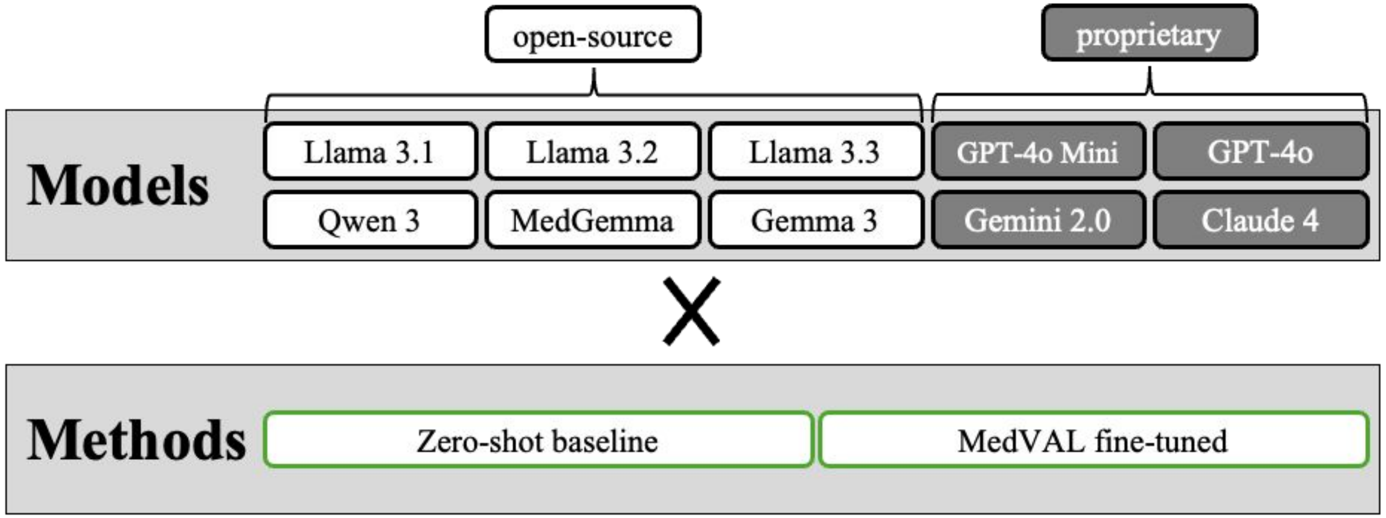
2. MedVAL-Bench dataset:
   - A dataset containing **840 physician-labeled evaluations** of AI-generated medical text
   - Performed by **12 physicians** spanning **6 diverse** medical text generation tasks

3. MedVAL performance benchmark:
   - MedVAL fine-tuning improves the validation **capabilities of all underlying LMs**
   - MedVAL yields **significant gains** ($p < 0.001$): average baseline F1 scores for:
     - **Safe/unsafe** classification improve from **66.2% to 82.8%**

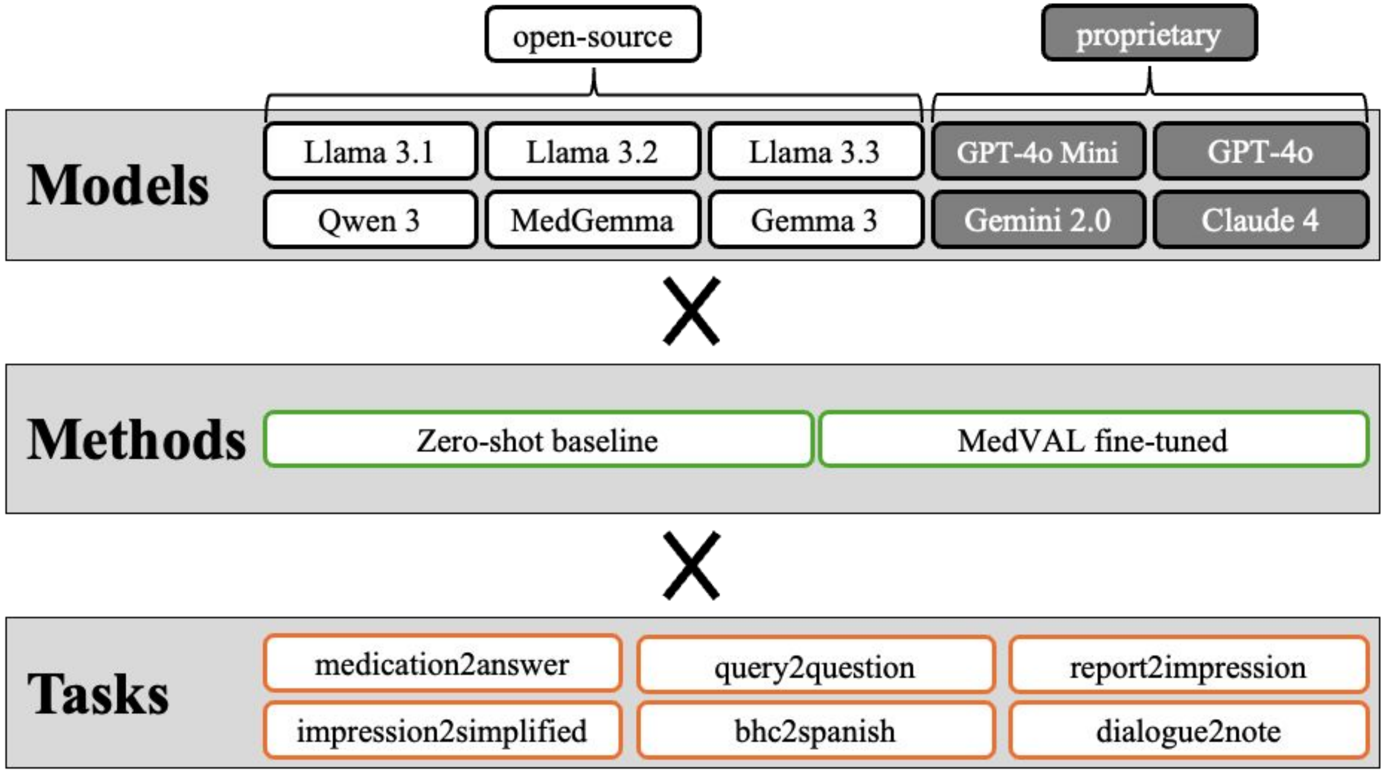
Schematic


open-source
proprietary
Models
Llama 3.1
Llama 3.2
Llama 3.3
GPT-4o Mini
GPT-4o
Qwen 3
MedGemma
Gemma 3
Gemini 2.0
Claude 4

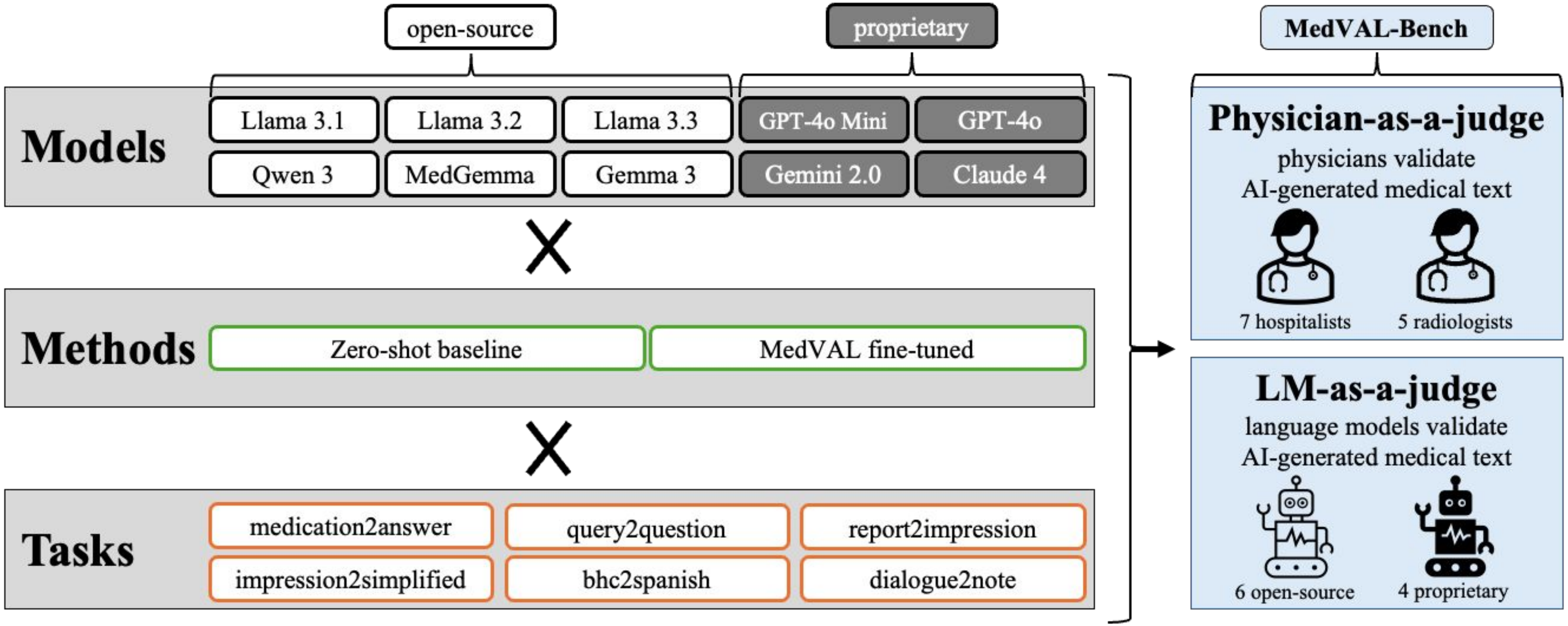
Schematic


open-source
proprietary
Models
Llama 3.1
Llama 3.2
Llama 3.3
GPT-4o Mini
GPT-4o
Qwen 3
MedGemma
Gemma 3
Gemini 2.0
Claude 4
X
Methods
Zero-shot baseline
MedVAL fine-tuned

Schematic

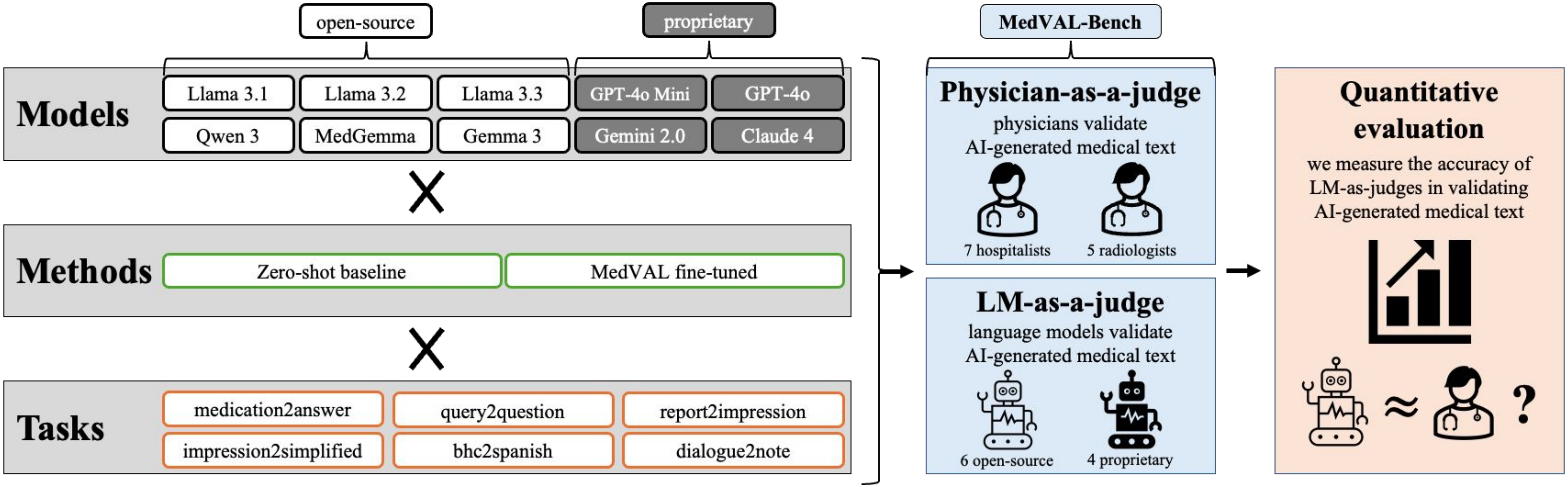

open-source
proprietary
Models
Llama 3.1
Llama 3.2
Llama 3.3
GPT-4o Mini
GPT-4o
Qwen 3
MedGemma
Gemma 3
Gemini 2.0
Claude 4
X
Methods
Zero-shot baseline
MedVAL fine-tuned
X
Tasks
medication2answer
query2question
report2impression
impression2simplified
bhc2spanish
dialogue2note

Schematic



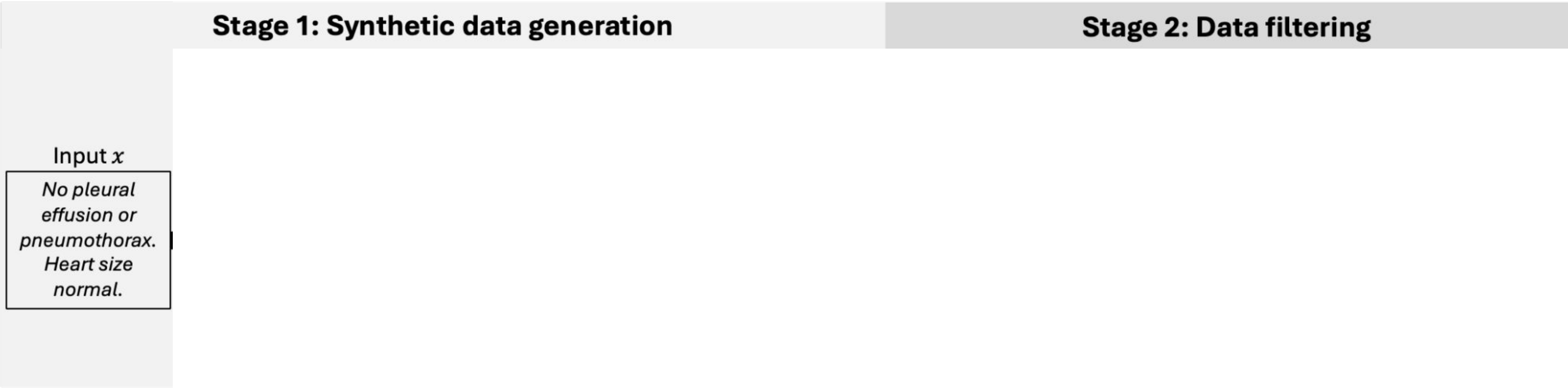
open-source
proprietary
MedVAL-Bench
Models
Llama 3.1
Llama 3.2
Llama 3.3
GPT-4o Mini
GPT-4o
Qwen 3
MedGemma
Gemma 3
Gemini 2.0
Claude 4
X
Methods
Zero-shot baseline
MedVAL fine-tuned
X
Tasks
medication2answer
query2question
report2impression
impression2simplified
bhc2spanish
dialogue2note
Physician-as-a-judge
physicians validate
AI-generated medical text
7 hospitalists
5 radiologists
LM-as-a-judge
language models validate
AI-generated medical text
6 open-source
4 proprietary

Schematic

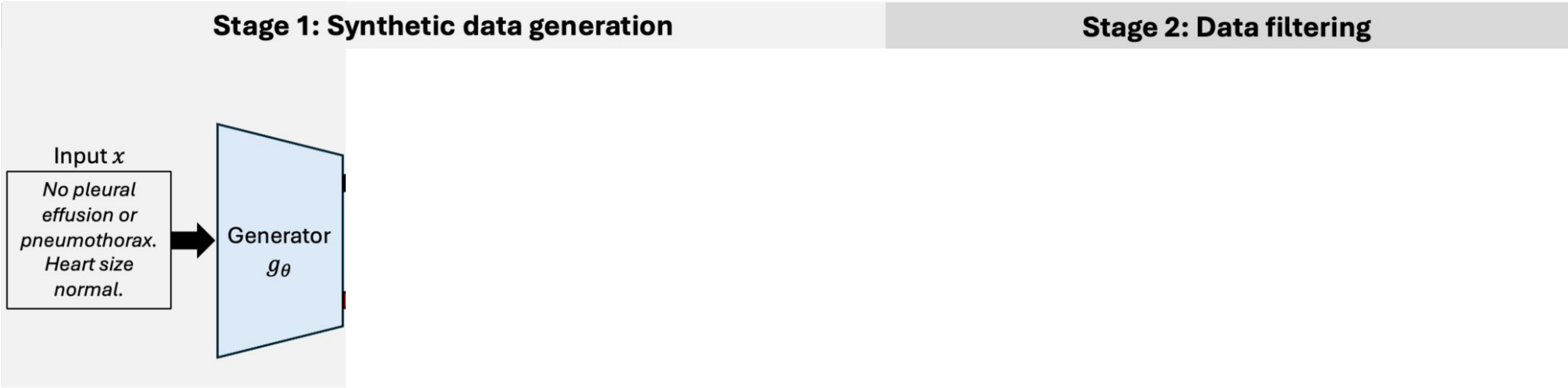


open-source
proprietary
MedVAL-Bench
Models
Llama 3.1
Llama 3.2
Llama 3.3
GPT-4o Mini
GPT-4o
Qwen 3
MedGemma
Gemma 3
Gemini 2.0
Claude 4
X
Methods
Zero-shot baseline
MedVAL fine-tuned
X
Tasks
medication2answer
query2question
report2impression
impression2simplified
bhc2spanish
dialogue2note
Physician-as-a-judge
physicians validate
AI-generated medical text
7 hospitalists
5 radiologists
LM-as-a-judge
language models validate
AI-generated medical text
6 open-source
4 proprietary
Quantitative evaluation
we measure the accuracy of
LM-as-judges in validating
AI-generated medical text
≈
?

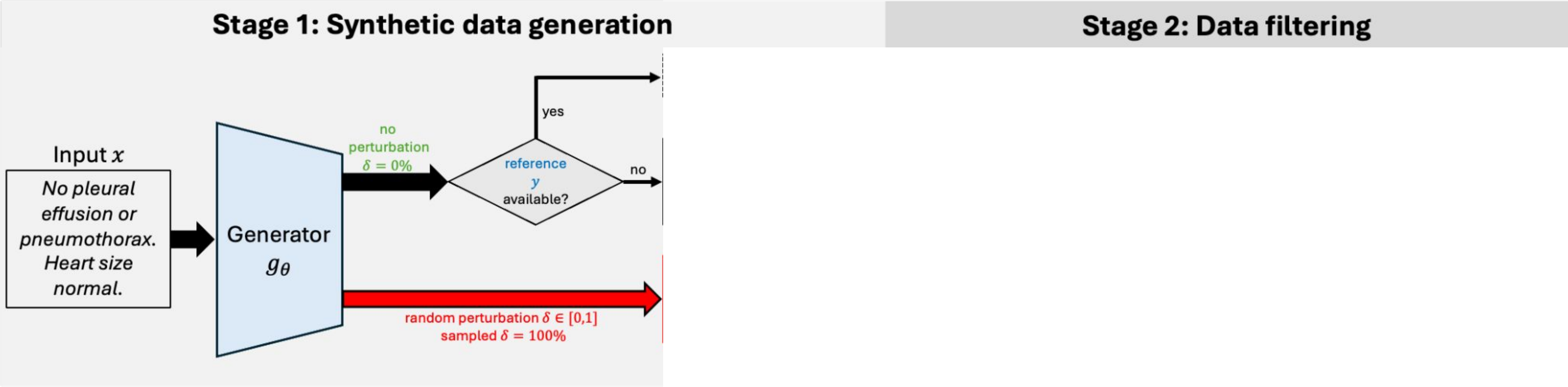
MedVAL Training



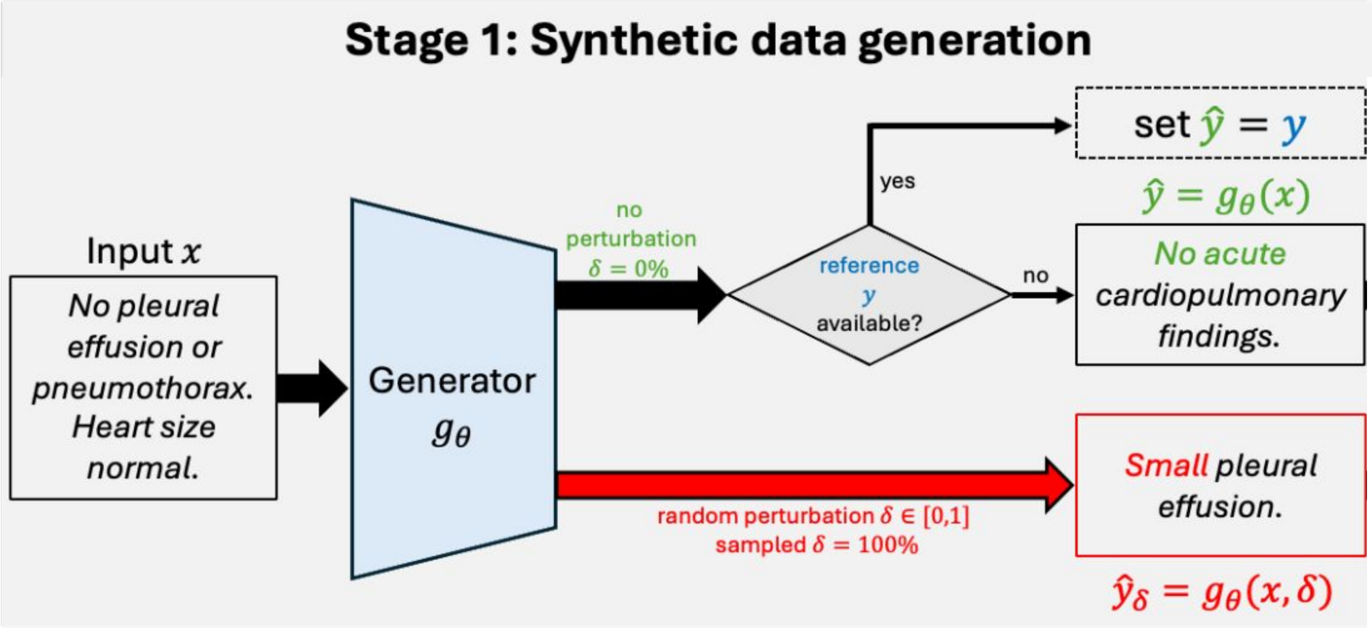
MedVAL Training



MedVAL Training

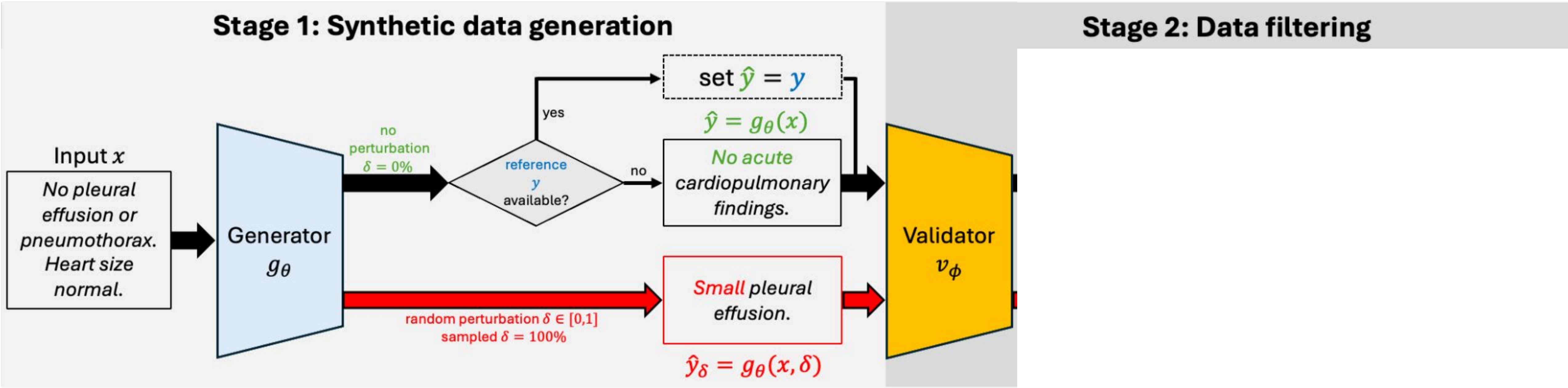


MedVAL Training

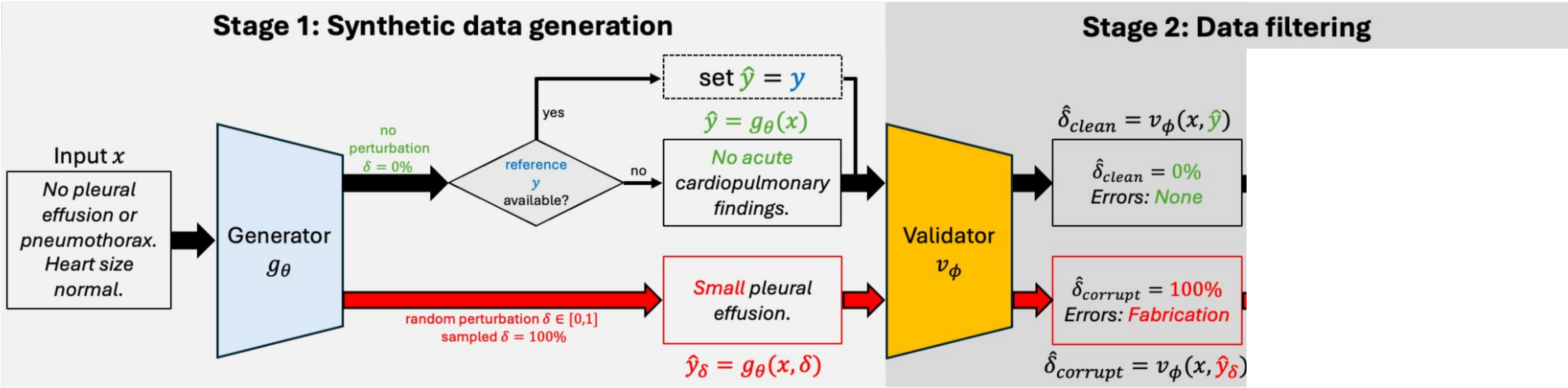


Stage 2: Data filtering

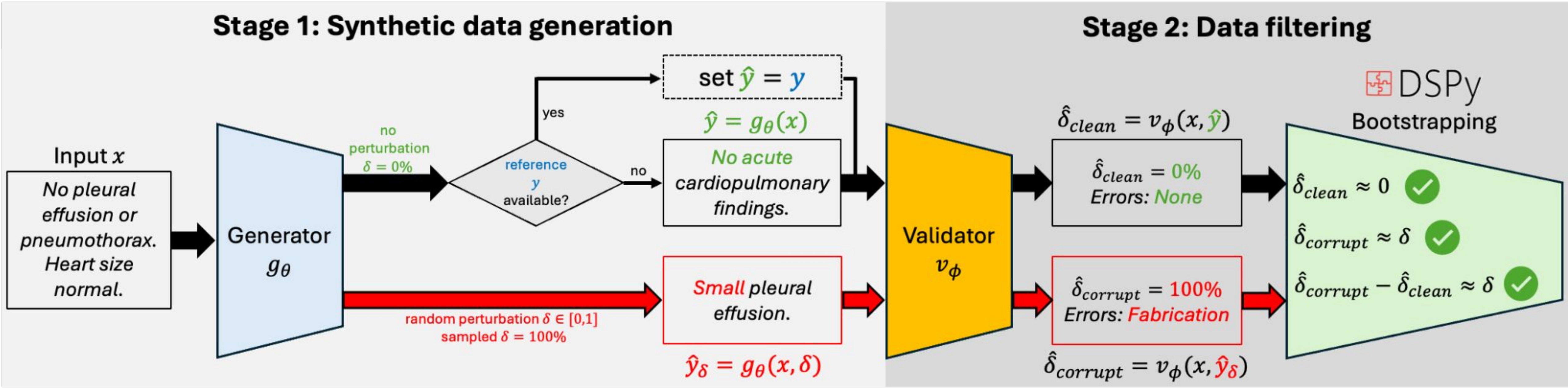
MedVAL Training



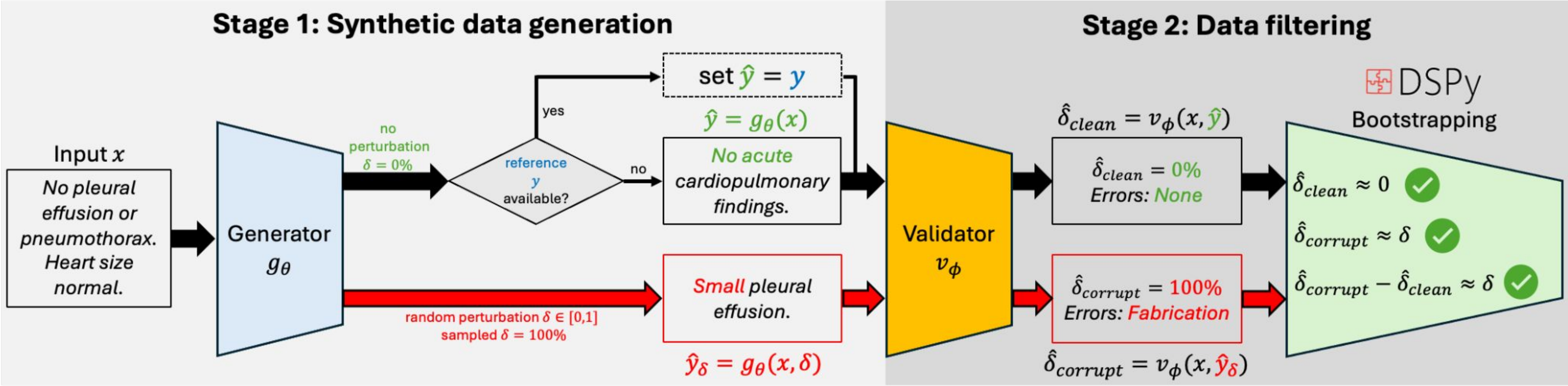
MedVAL Training



MedVAL Training



MedVAL Training



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.

MedVAL Training - Algorithm

Algorithm 1 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**
  $\delta \leftarrow \mathit{RandomChoice}(\{\delta_1, \delta_2, \ldots, \delta_L\} \mid \delta \in [0, 1])$
  $\hat{y} \leftarrow y$ 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
$v_\alpha^* = \mathit{SFT}(v_\alpha, \mathcal{D}_{\text{train}})$ ▷ Supervised fine-tuning
return v_α^*

Perturbation Strategies

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."

Perturbation Strategies

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."

Perturbation Strategies

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."

Perturbation Strategies

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 Categories

Error category	Error	Description
Hallucinations	Fabricated claim	Introduction of a claim not present in the input.
	Misleading justification	Incorrect reasoning, leading to misleading conclusions.
	Detail misidentification	Incorrect reference to a detail in the input.
	False comparison	Mentioning a comparison not supported by the input.
	Incorrect recommendation	Suggesting a diagnosis/follow-up outside the input.

Error Categories

Error category	Error	Description
Hallucinations	Fabricated claim	Introduction of a claim not present in the input.
	Misleading justification	Incorrect reasoning, leading to misleading conclusions.
	Detail misidentification	Incorrect reference to a detail in the input.
	False comparison	Mentioning a comparison not supported 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 that details change over time.
	Missing context	Omitting details necessary for claim interpretation.

Error Categories

Error category	**Error**	**Description**
Hallucinations	Fabricated claim	Introduction of a claim not present in the input.
	Misleading justification	Incorrect reasoning, leading to misleading conclusions.
	Detail misidentification	Incorrect reference to a detail in the input.
	False comparison	Mentioning a comparison not supported 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 that details 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.

Error Categories

Error category	Error	Description
Hallucinations	Fabricated claim	Introduction of a claim not present in the input.
	Misleading justification	Incorrect reasoning, leading to misleading conclusions.
	Detail misidentification	Incorrect reference to a detail in the input.
	False comparison	Mentioning a comparison not supported 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 that details 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.

MedVAL-Bench

- A dataset for **training** and **evaluation** of medical text validators
- Contains: (1) ***inputs***, (2) ***outputs***, (3) ***physician assessments (only test)***

Task Name	Dataset	Task Description	y_{ref}?	# Train	# Test
`medication2answer`	MedicationQA	medication question → answer	✓	500	135
`query2question`	MeQSum	patient query → health question	✓	500	120
`report2impression`	Open-i	findings → impression	✓	500	190
`radiology2simplified`	Open-i	findings → patient-friendly	✗	500	–
`radiology2simplified`†	MIMIC-IV	impression → patient-friendly	✗	–	190
`bhc2spanish`†	MIMIC-IV-BHC	hospital course → spanish	✗	–	120
`dialogue2note`†	ACI-Bench	doctor-patient dialogue → note	✓	–	85
Total				**2000**	**840**

1. Partially open-source
2. Out-of-distribution

1. Fully open-source
2. In-distribution

MedVAL-Bench - Tasks

Task	Input → output	Instructional prompt
`medication2answer`	medication question → answer	"Answer the following medication-related patient health question."

MedVAL-Bench - Tasks

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 of 15 words or less."

MedVAL-Bench - Tasks

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 of 15 words or less."
`report2impression`	findings → impression	"Summarize the radiology report findings into an impression with minimal text."
`report2simplified`	findings → patient-friendly	"Create a simplified, patient-friendly version of the input."

MedVAL-Bench - Tasks

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 of 15 words or less."
`report2impression`	findings → impression	"Summarize the radiology report findings into an impression with minimal text."
`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."

MedVAL-Bench - Tasks

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 of 15 words or less."
`report2impression`	findings → impression	"Summarize the radiology report findings into an impression with minimal text."
`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."

MedVAL-Bench - Tasks

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 of 15 words or less."
`report2impression`	findings → impression	"Summarize the radiology report findings into an impression with minimal text."
`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 doctor/patient dialogue into an assessment and plan."

MedVAL-Bench - Physician Study

Your goal is to verify the existing evaluation report and revise it if necessary.

Only categorize a factual claim as a clinically significant error if it affects clinical understanding, decision-making, or safety.

You may reassign, revise, add, or remove errors based on your judgment.

Impression (Expert-Written)

Claim 1: Low left lung volume with surrounding pleural thickening and calcified pleural plaques, consistent with prior asbestos exposure.

Claim 2: Recommend comparison with prior for change in pleural thickening

Patient-Friendly Impression (Model-Generated)

Claim 1: Your left lung is severely damaged, and the surrounding tissue is turning into bone due to asbestos exposure.

Claim 2: This condition is rapidly worsening and will likely require immediate surgery to remove the affected lung.

Claim 3: There's no need to compare this with previous scans because the damage is already too advanced to reverse.

❌ Clinically Significant Errors

Error 1: "Your left lung is severely damaged, and the surrounding tissue is turning into bone due to asbestos exposure." - Hallucination (fabricated claim): The reference mentions pleural thickening and calcified pleural plaques but does not state that the surrounding tissue is turning into bone or that the lung is severely damaged.
Error 2: "This condition is rapidly worsening and will likely require immediate surgery to remove the affected lung." - Hallucination (incorrect recommendation): The reference does not mention rapid worsening or suggest surgery as a necessary intervention.
Error 3: "There's no need to compare this with previous scans because the damage is already too advanced to reverse." - Certainty misalignment (overstating intensity): The reference explicitly recommends comparison with prior imaging for changes in pleural thickening, and the candidate dismisses this recommendation with an exaggerated claim

Add

MedVAL-Bench - Physician Study

📊 Overall Quality Rating

Specify the level that best matches the candidate's factual consistency with the reference.

🔷 Level 1: Fully Factually Consistent

- No hallucinations, omissions, or certainty misalignments. All factual claims match the reference.

🔶 Level 2: Low-Risk Errors

- Subtle errors such as mild overstatements or omissions with low clinical impact. The main message is unchanged.

🟠 Level 3: Moderate-Risk Errors

- Errors span at least two categories. At least one error significantly impacts clinical interpretation.

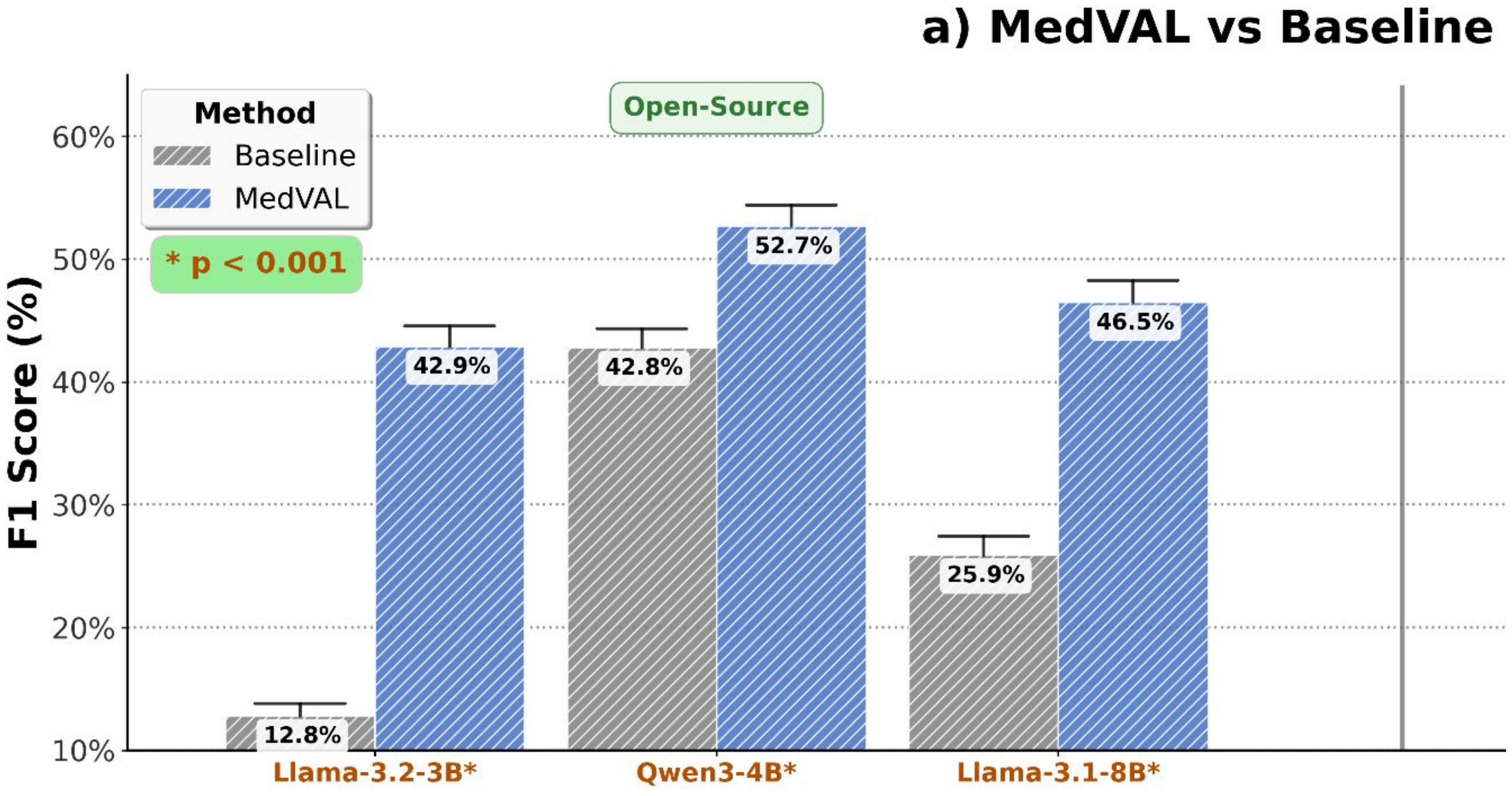
🔴 Level 4: High-Risk Errors

- Severe factual inconsistencies across all categories. At least two errors pose high clinical risk or misinterpretation.

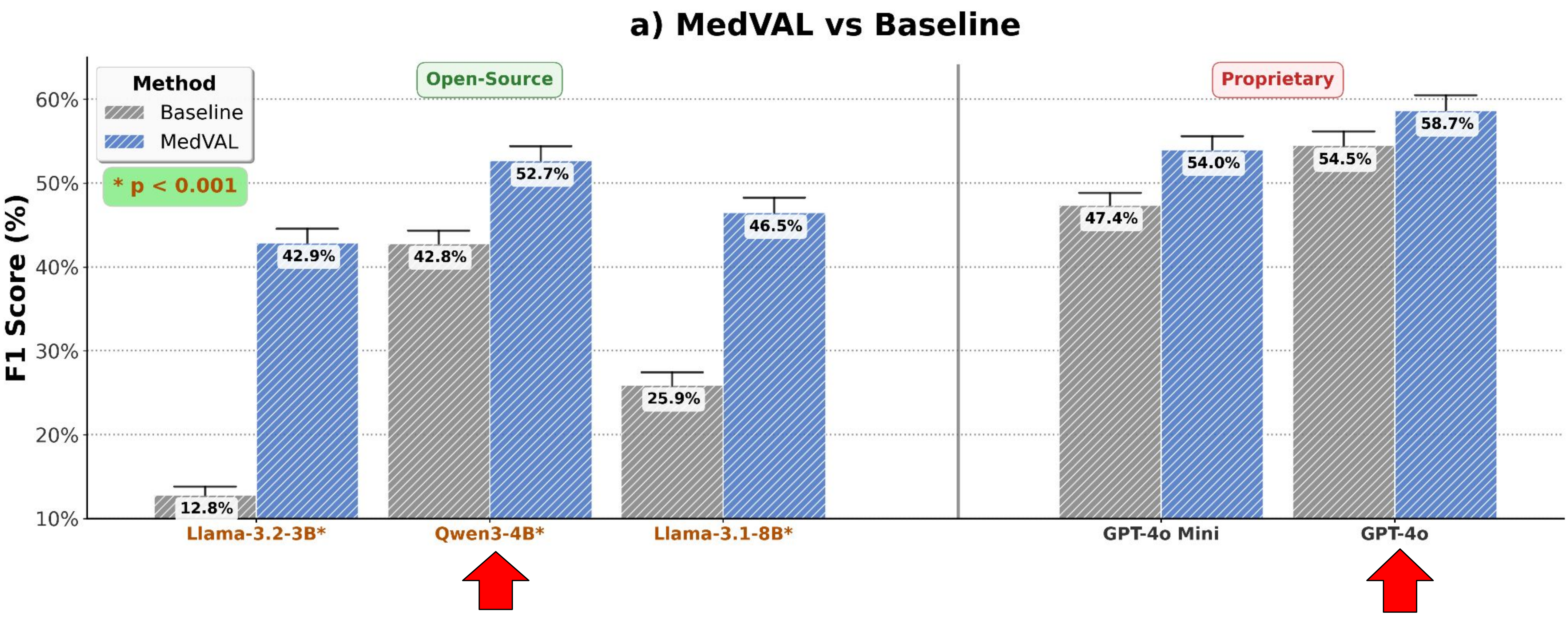
Level 4

Results

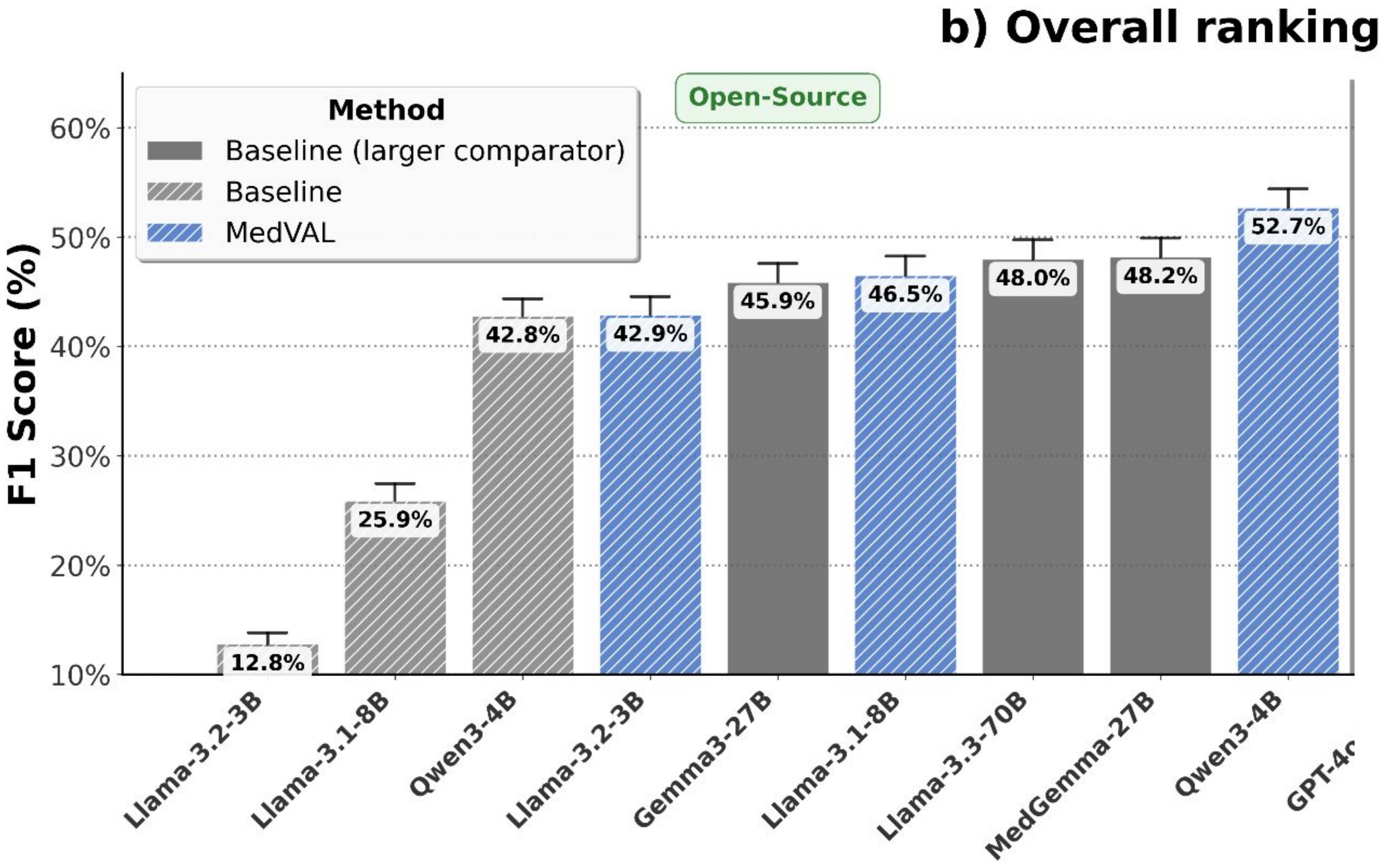
Overall Performance (F1 Classification Score)



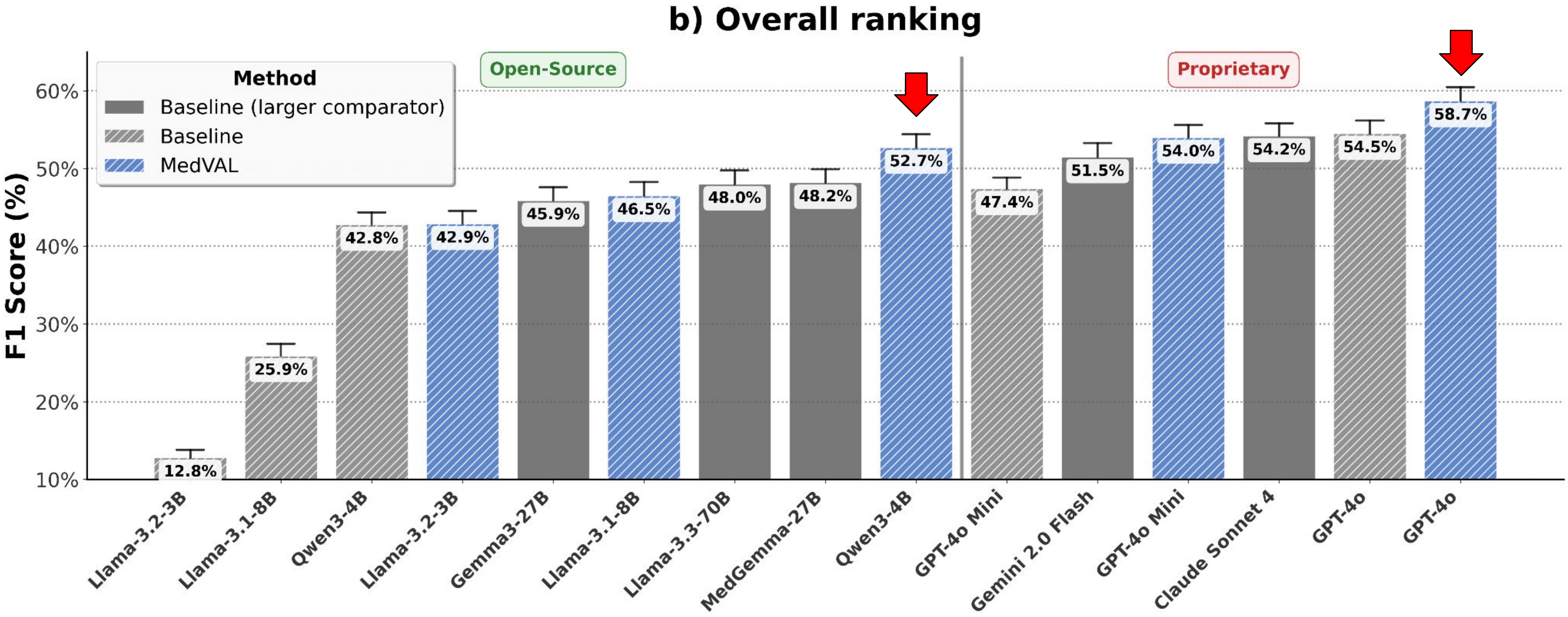
Overall Performance (F1 Classification Score)



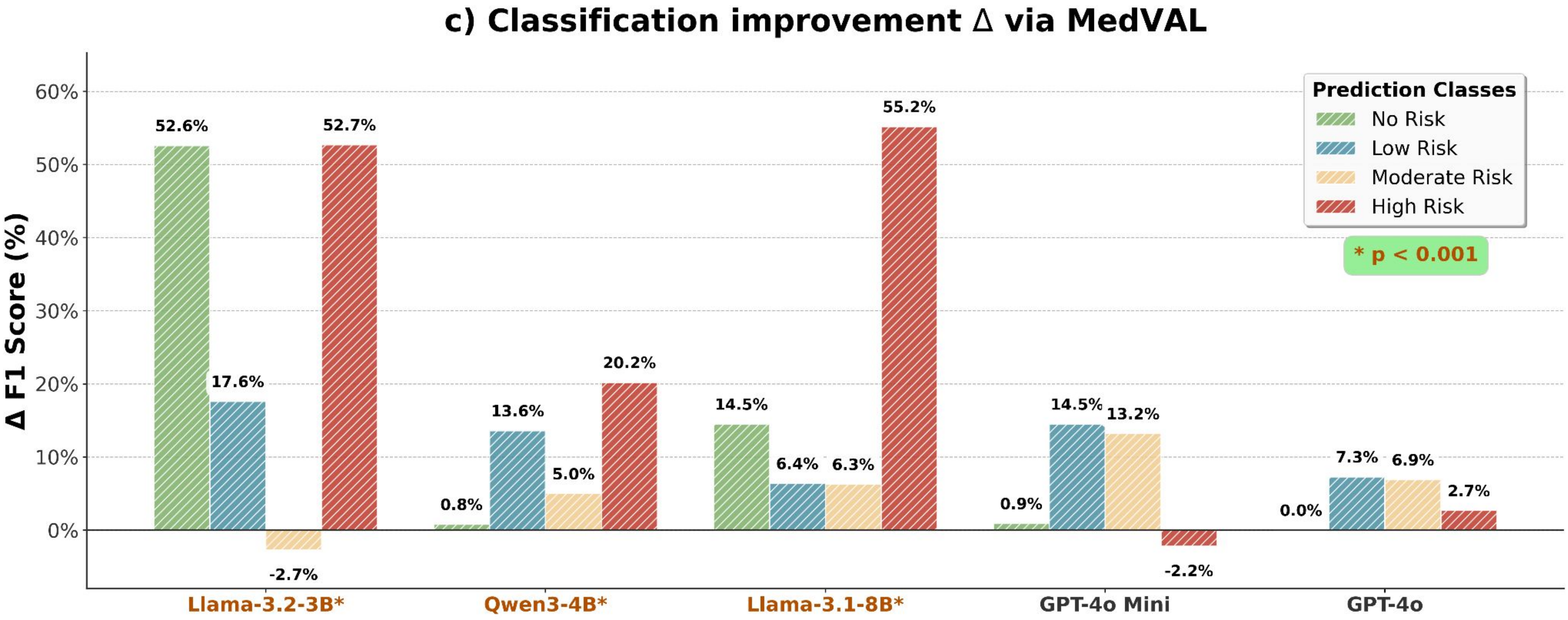
Overall Performance (F1 Classification Score)



Overall Performance (F1 Classification Score)



Risk-Level Classification Performance



Task-Wise Performance

Method	Model	In-Distribution medication2 answer	query2 question	report2 impression	Out-of-Distribution impression2 simplified	bhc2 spanish	dialogue2 note	Overall
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%

Task-Wise Performance

Method	Model	In-Distribution: medication2 answer	query2 question	report2 impression	Out-of-Distribution: impression2 simplified	bhc2 spanish	dialogue2 note	Overall
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

Safety (Binary) Classification Performance

#	Model	Method	Sensitivity	Specificity	F1 Score	Accuracy
1	Llama-3.2-3B	Baseline	0.086±0.01	**0.960**±0.01	0.153±0.02	0.474±0.02
		MedVAL	**0.919**±0.01	0.560±0.02	**0.809**±0.01	**0.760**±0.01
2	Llama-3.1-8B	Baseline	0.670±0.02	0.651±0.03	0.688±0.02	0.662±0.02
		MedVAL	**0.788**±0.02	**0.786**±0.01	**0.804**±0.01	**0.787**±0.01
3	Qwen3-4B	Baseline	**0.858**±0.02	0.643±0.03	0.800±0.01	0.762±0.02
		MedVAL	0.839±0.02	**0.752**±0.02	**0.823**±0.01	**0.800**±0.01
	Ensemble (1+2+3)	MedVAL	0.899±0.02	0.686±0.03	**0.837**±0.01	**0.805**±0.01

Safety (Binary) Classification Performance

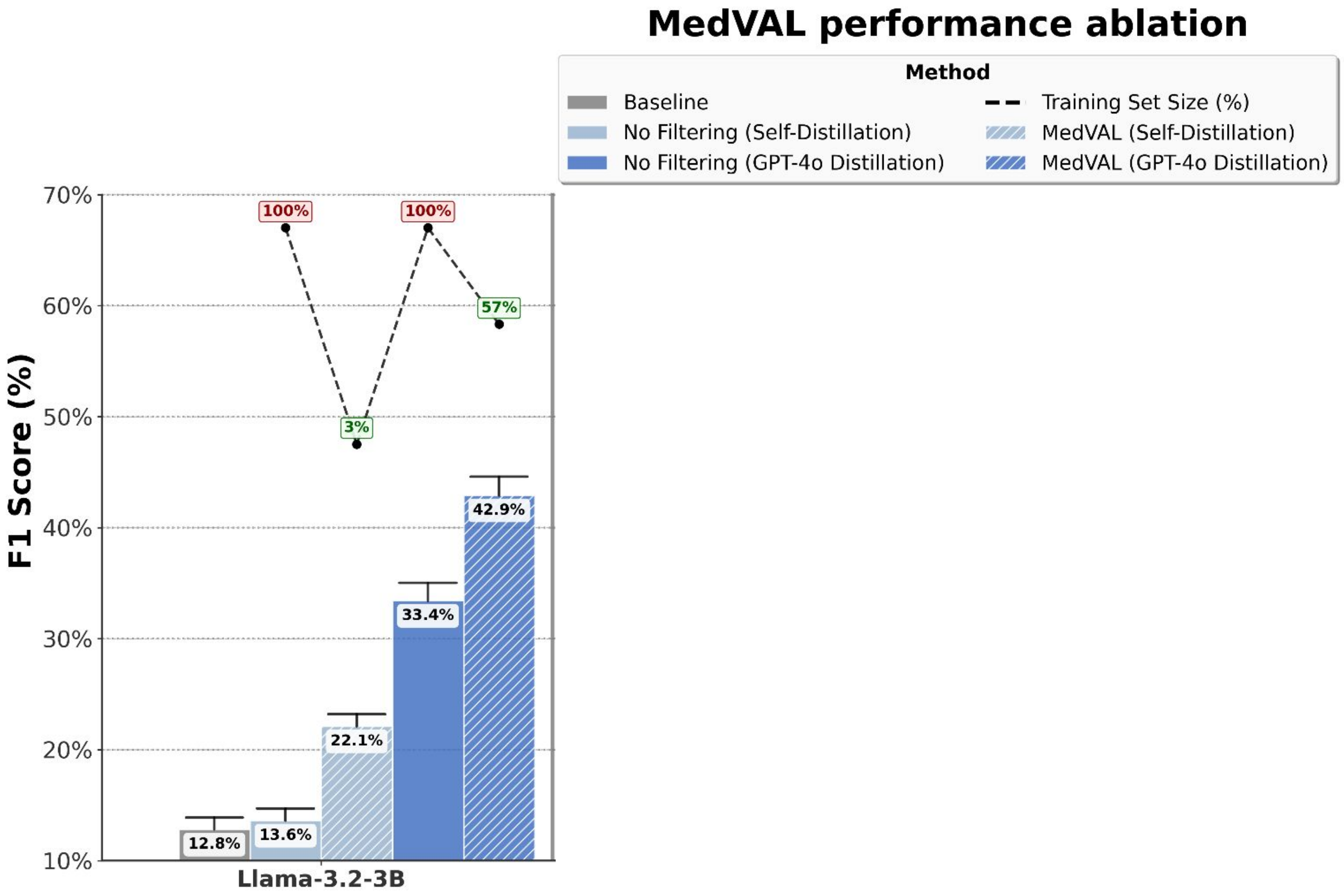
#	Model	Method	Sensitivity	Specificity	F1 Score	Accuracy
1	Llama-3.2-3B	Baseline	0.086±0.01	**0.960**±0.01	0.153±0.02	0.474±0.02
		MedVAL	**0.919**±0.01	0.560±0.02	**0.809**±0.01	**0.760**±0.01
2	Llama-3.1-8B	Baseline	0.670±0.02	0.651±0.03	0.688±0.02	0.662±0.02
		MedVAL	**0.788**±0.02	**0.786**±0.01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

	Krippendorff's α
Inter-Physician Agreement	0.754

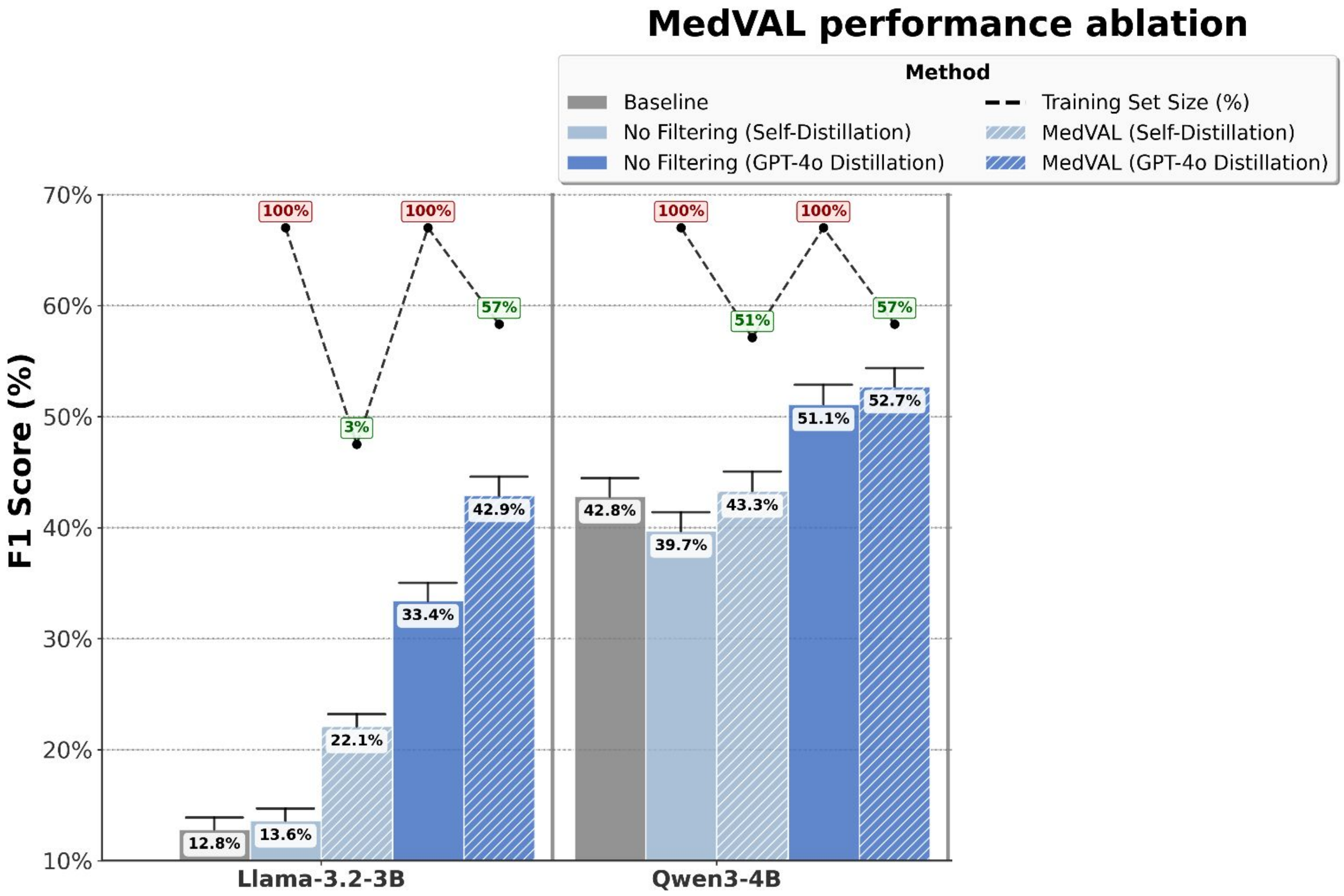
Performance Ablation

Performance Ablation

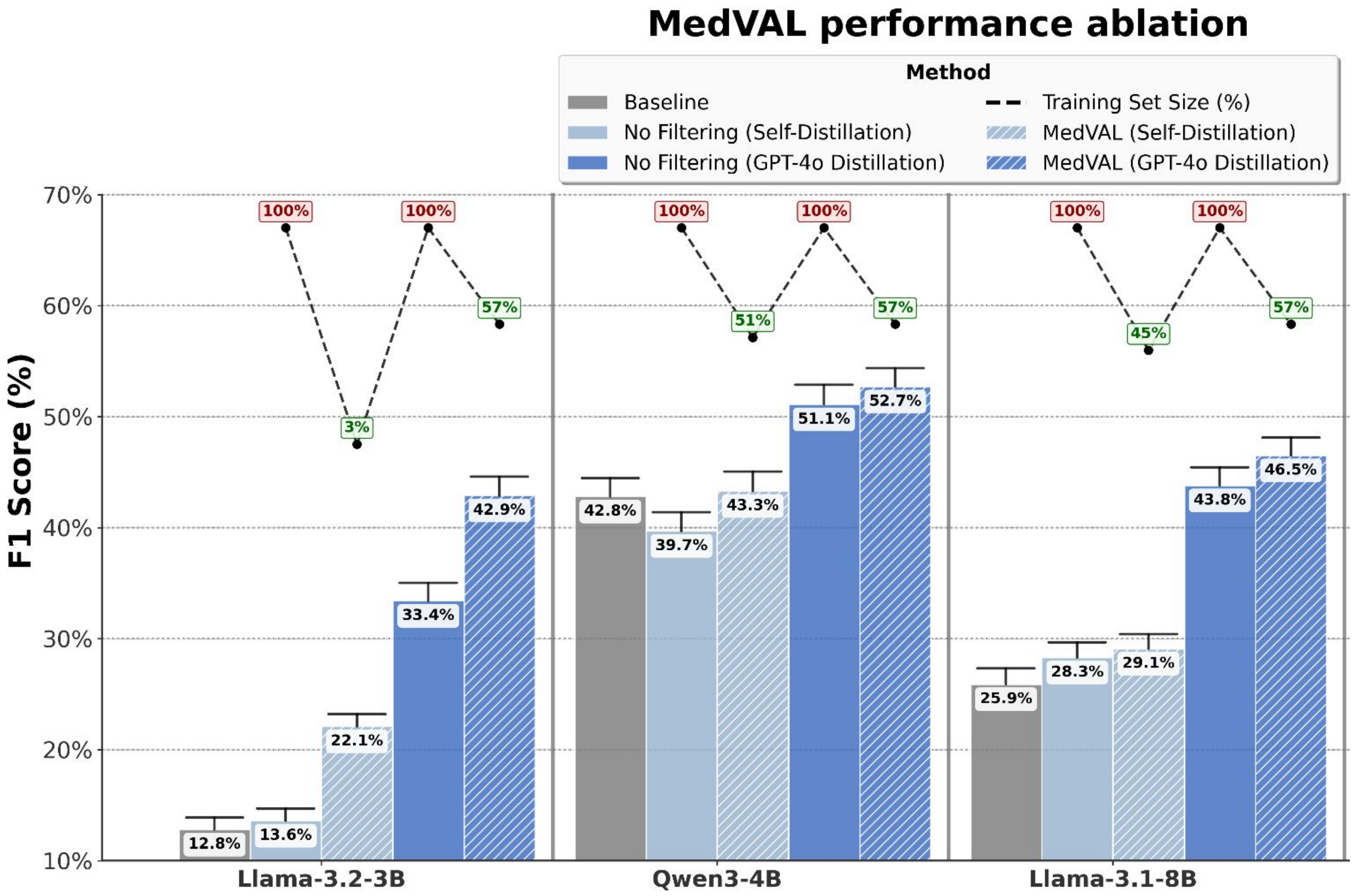
Performance Ablation



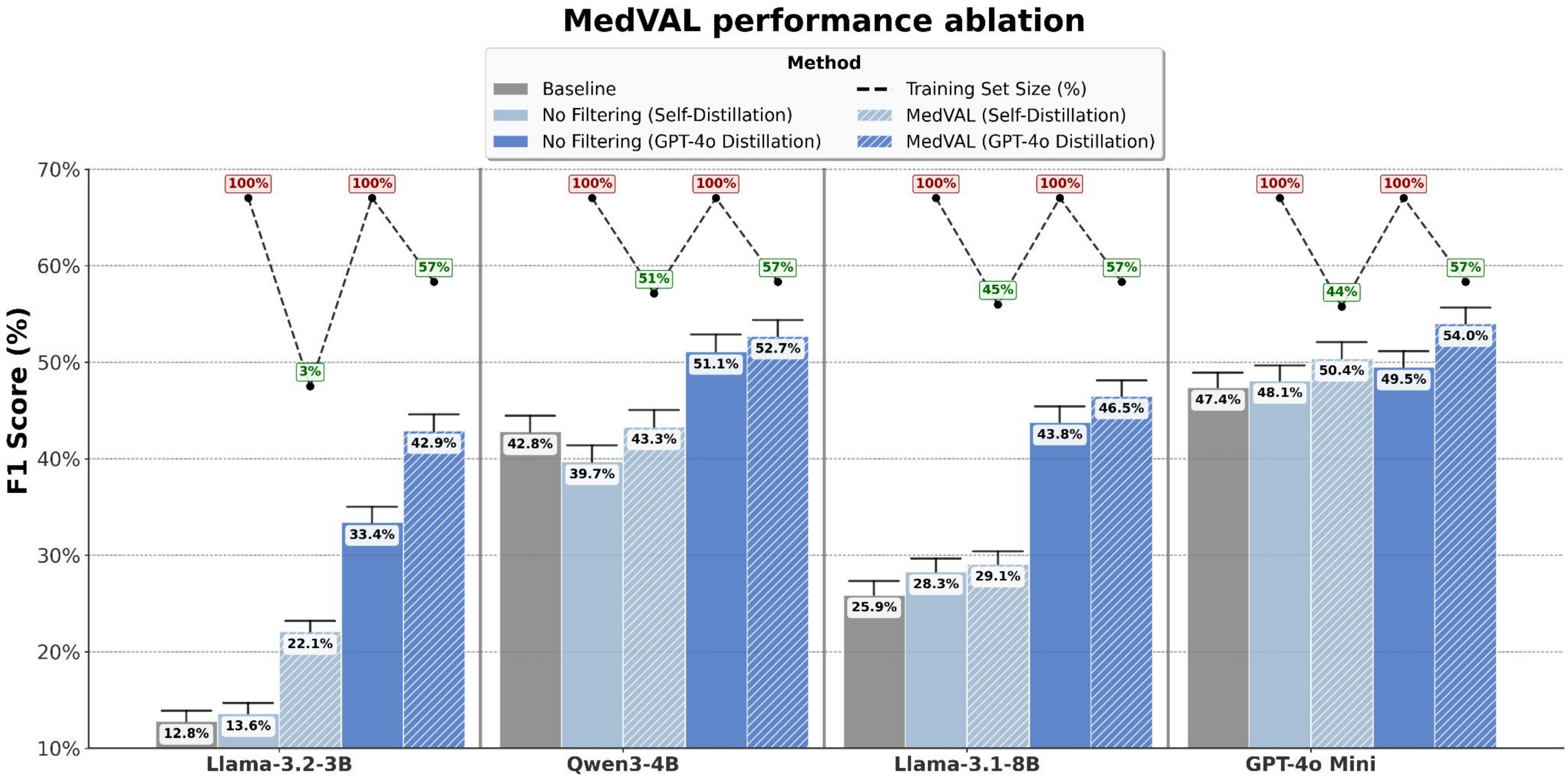
Performance Ablation



Performance Ablation



Performance Ablation



Example:
Impression (Input) -> Patient Friendly (Output)

Input (*impression*):
1. Interval resolution of heterogeneous ground-glass opacity on CT of the lungs. 2. Extensive persistent air trapping, has slightly resolved since prior imaging. The overall appearance is consistent with airways disease, probably asthma. No evidence of respiratory bronchiolitis - interstitial lung disease.

Color key:

Risk grade prediction:
Green: correct
Red: incorrect

Error categories:
Purple: fabricated claim
Blue: detail misidentification
Brown: misleading justification
Orange: no errors found.

Example:
Impression (Input) -> Patient Friendly (Output)

Input (*impression*):
1. Interval resolution of heterogeneous ground-glass opacity on CT of the lungs. 2. Extensive persistent air trapping, has slightly resolved since prior imaging. The overall appearance is consistent with airways disease, probably asthma. No evidence of respiratory bronchiolitis - interstitial lung disease.

Output (*patient-friendly impression*):
Your recent CT scan shows that some of the cloudy areas in your lungs have cleared up. There is still some trapped air in your lungs, but it has improved a little since your last scan. This suggests you might have asthma or another breathing condition. There is no sign of a more serious lung disease like interstitial lung disease, which is good news. However, we may need to monitor for signs of respiratory infections in the future.

Color key:

Risk grade prediction:
Green: correct
Red: incorrect

Error categories:
Purple: fabricated claim
Blue: detail misidentification
Brown: misleading justification
Orange: no errors found.

Example:
Impression (Input) -> Patient Friendly (Output)

Input (*impression*):
1. Interval resolution of heterogeneous ground-glass opacity on CT of the lungs. 2. Extensive persistent air trapping, has slightly resolved since prior imaging. The overall appearance is consistent with airways disease, probably asthma. No evidence of respiratory bronchiolitis - interstitial lung disease.

Output (*patient-friendly impression*):
Your recent CT scan shows that some of the cloudy areas in your lungs have cleared up. There is still some trapped air in your lungs, but it has improved a little since your last scan. This suggests you might have asthma or another breathing condition. There is no sign of a more serious lung disease like interstitial lung disease, which is good news. However, we may need to monitor for signs of respiratory infections in the future.

Assessment (physician):
Risk grade: Level 3 (moderate risk)
Errors: "However, we may need to monitor for signs of respiratory infections in the future" is missing from the input.

Color key:

Risk grade prediction:
Green: correct
Red: incorrect

Error categories:
Purple: fabricated claim
Blue: detail misidentification
Brown: misleading justification
Orange: no errors found.

Example:
Impression (Input) -> Patient Friendly (Output)

Input (*impression*):
1. Interval resolution of heterogeneous ground-glass opacity on CT of the lungs. 2. Extensive persistent air trapping, has slightly resolved since prior imaging. The overall appearance is consistent with airways disease, probably asthma. No evidence of respiratory bronchiolitis - interstitial lung disease.

Output (*patient-friendly impression*):
Your recent CT scan shows that some of the cloudy areas in your lungs have cleared up. There is still some trapped air in your lungs, but it has improved a little since your last scan. This suggests you might have asthma or another breathing condition. There is no sign of a more serious lung disease like interstitial lung disease, which is good news. However, we may need to monitor for signs of respiratory infections in the future.

Assessment (physician):
Risk grade: Level 3 (moderate risk)
Errors: "However, we may need to monitor for signs of respiratory infections in the future" is missing from the input.

Assessment (baseline):
Risk grade: Level 1 (no risk)
Errors: None.

Color key:

Risk grade prediction:
Green: correct
Red: incorrect

Error categories:
Purple: fabricated claim
Blue: detail misidentification
Brown: misleading justification
Orange: no errors found.

Example:
Impression (Input) -> Patient Friendly (Output)

Input (*impression*):
1. Interval resolution of heterogeneous ground-glass opacity on CT of the lungs. 2. Extensive persistent air trapping, has slightly resolved since prior imaging. The overall appearance is consistent with airways disease, probably asthma. No evidence of respiratory bronchiolitis - interstitial lung disease.

Output (*patient-friendly impression*):
Your recent CT scan shows that some of the cloudy areas in your lungs have cleared up. There is still some trapped air in your lungs, but it has improved a little since your last scan. This suggests you might have asthma or another breathing condition. There is no sign of a more serious lung disease like interstitial lung disease, which is good news. However, we may need to monitor for signs of respiratory infections in the future.

Assessment (physician):
Risk grade: Level 3 (moderate risk)
Errors: "However, we may need to monitor for signs of respiratory infections in the future" is missing from the input.

Assessment (baseline):
Risk grade: **Level 1 (no risk)**
Errors: None.

Assessment (MedVAL):
Risk grade: **Level 3 (moderate risk)**
Errors: "However, we may need to monitor for signs of respiratory infections in the future" is not mentioned in the input.

Color key:

Risk grade prediction:
Green: correct
Red: incorrect

Error categories:
Purple: fabricated claim
Blue: detail misidentification
Brown: misleading justification
Orange: no errors found.

Example:
Dialogue (Input) -> Assessment and Plan (Output)

Input (*doctor-patient dialogue*):
[doctor] hey diana it's good to see you in here so i see that you injured your wrist could you tell me a bit about what happened [patient] yeah i was walking up and down the stairs i was doing my laundry and i slipped and i tried to catch myself ... joint dorsal angulation of the distal fracture fragment is present to a variable degree if dorsal angulation is severe presenting with a dinner fork deformity ulnar styloid fracture is present.

Output (*assessment and plan*):
Assessment: The patient presents with a right wrist injury sustained during a fall while doing laundry ... elevated blood pressure and heart rate are noted, possibly indicating a systemic issue.
Plan: 1. Apply a thumb spica brace for immobilization, to be worn for eight weeks ... 7. Follow up in one month to reassess progress and determine readiness for travel.

Assessment (physician):
Risk grade: Level 4 (high risk)
Errors: 1. The reference describes the fracture as extra-articular, not intra-articular. 2. The reference specifies a brace to be worn for six weeks.

Assessment (baseline):
Risk grade: Level 3 (moderate risk)
Errors: Insufficient information about patient's overall health and comorbidities.

Assessment (MedVAL):
Risk grade: Level 4 (high risk)
Errors: 1. The reference states the fracture is extra-articular, not intra-articular. 2. The reference specifies a brace (not a thumb spica) to be worn for six weeks.

Color key:

Risk grade prediction:
Green: correct
Red: incorrect

Error categories:
Purple: fabricated claim
Blue: incorrect recommendation
Brown: missing context
Orange: no errors

Conclusion

- We introduce MedVAL, a **generalizable**, **self-supervised** framework for **validating LM-generated medical text**
- Across all settings, MedVAL **improved average F1 scores for all underlying models**
- Risk-level analysis revealed that **MedVAL enhances model sensitivity**
  - particularly at **intermediate risk levels (2–3)**, which are critical for deciding human review.
- Task-wise results confirmed **strong generalization across in-distribution and out-of-distribution settings**
- Notably, MedVAL **displayed strong improvements** on **dialogue2note**
  - the **longest input context** (average 1.5k tokens) **out-of-distribution task**
  - showing **robustness on challenging, real-world medical tasks**.

Open-Source

- Paper: https://arxiv.org/abs/2507.03152
- Code: https://github.com/StanfordMIMI/MedVAL
- MedVAL-Bench Dataset: https://huggingface.co/datasets/stanfordmimi/MedVAL-Bench
- MedVAL-4B Model: https://huggingface.co/stanfordmimi/MedVAL-4B

Open-Source

- Paper: https://arxiv.org/abs/2507.03152
- Code: https://github.com/StanfordMIMI/MedVAL
- MedVAL-Bench Dataset: https://huggingface.co/datasets/stanfordmimi/MedVAL-Bench
- MedVAL-4B Model: https://huggingface.co/stanfordmimi/MedVAL-4B

Computer Science > Computation and Language

[Submitted on 3 Jul 2025 (v1), last revised 2 Sep 2025 (this version, v3)]

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 Juan Perez Guerrero, Paola Naovi Cruz Rivera, Sergios Gatidis, Christian Bluethgen, Eduardo Pontes Reis, Eddy D. Zandee van Rilland, Poonam Laxmappa Hosamani, Kevin R Keet, Minjoung Go, Evelyn Ling, David B. Larson, Curtis Langlotz, Roxana Daneshjou, Jason Hom, Sanmi Koyejo, Emily Alsentzer, Akshay S. Chaudhari

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. Currently, such evaluation relies solely on manual physician review. 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. While the "LM-as-judge" paradigm (a LM evaluating another LM) offers scalable evaluation, even frontier LMs can miss subtle but clinically significant errors. To address these challenges, we propose MedVAL, a self-supervised framework 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 containing 840 outputs annotated by physicians, following a physician-defined taxonomy of risk levels and error categories. Across 6 diverse medical tasks and 10 state-of-the-art LMs spanning open-source, proprietary, and medically adapted models, MedVAL fine-tuning significantly improves ($p < 0.001$) alignment with physicians on both seen and unseen tasks, increasing average F1 scores from 66% to 83%, with per-sample safety classification scores up to 86%. MedVAL improves the performance of even the best-performing proprietary LM (GPT-4o) by 8%. To support a scalable, risk-aware pathway towards clinical integration, we open-source the 1) codebase (this https URL), 2) MedVAL-Bench (this https URL), and 3) MedVAL-4B (this https URL), the best-performing open-source LM. Our research provides the first evidence of LMs approaching expert-level validation ability for medical text.

barbiturates hypnotic drugs that were…

← **Post**

DSPy
@DSPyOSS

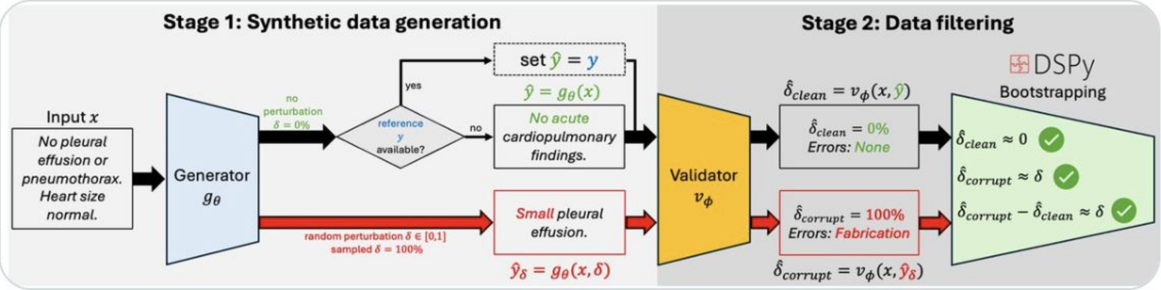
New paper from Stanford University.

"Expert-level validation of AI-generated medical text with scalable language models"

The authors use dspy.BootstrapFinetune for offline RL to update the weights of their LLMs.

They introduce MedVAL, a method to train LLMs to evaluate whether LM-generated medical outputs are factually consistent with inputs.

As part of their contribution, they wrote a PR to extend DSPy's parameter-efficient fine-tuning optimizers to enable Quantized Low-Rank Adaptation (QLoRA).



11:34 AM · Jul 10, 2025 · **10.8K** Views

3 33 254 182

Asad Aali, MS
Research Scientist
Stanford University

Stanford MEDICINE