

Toward Improved Evaluation of Language Models for Medicine

Asad Aali
PhD Student
Stanford University

About Me



- Education
  - **PhD in Computer Science @ Stanford University** (2026 – XXXX)
  - Master's in Electrical and Computer Engineering @ UT Austin (2022 – 2024)
  - Master's in Information Technology @ UT Austin (2021 – 2022)
  - Bachelor's in Accounting & Finance @ LUMS (2015 – 2019)
- PhD funded by **Knight-Hennessy Scholars Fellowship**

Knight-Hennessy Scholars
Stanford University

Prior Research

- Solving inverse problems with score-based generative priors learned from noisy data
  - Aali et al. **IEEE Asilomar Conference 2023**

- A dataset and benchmark for hospital course summarization with adapted large language models
  - Aali et al. **Journal of the American Medical Informatics Association 2025**

- Ambient diffusion posterior sampling
  - Aali et al. **ICLR 2025**

- Robust multi-coil MRI reconstruction via self-supervised denoising
  - Aali et al. **Magnetic Resonance in Imaging 2025**

- Structured Prompts Improve Evaluation of Language Models
  - Aali et al. **Machine Learning for Health 2025**

- MedVAL: Toward expert-level medical text validation with language models
  - Aali et al. **npj Digital Medicine 2026**

Today's Talk

Stanford MEDICINE

MedVAL
Toward Expert-Level Medical Text Validation with Language Models

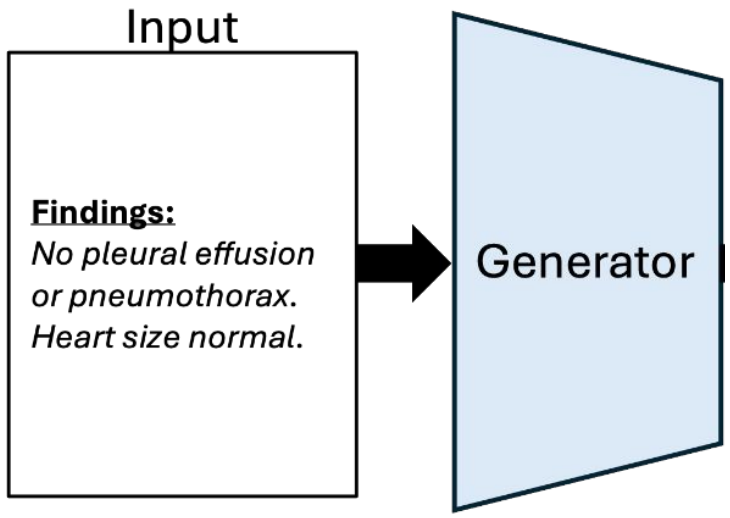
npj | digital medicine

Introduction

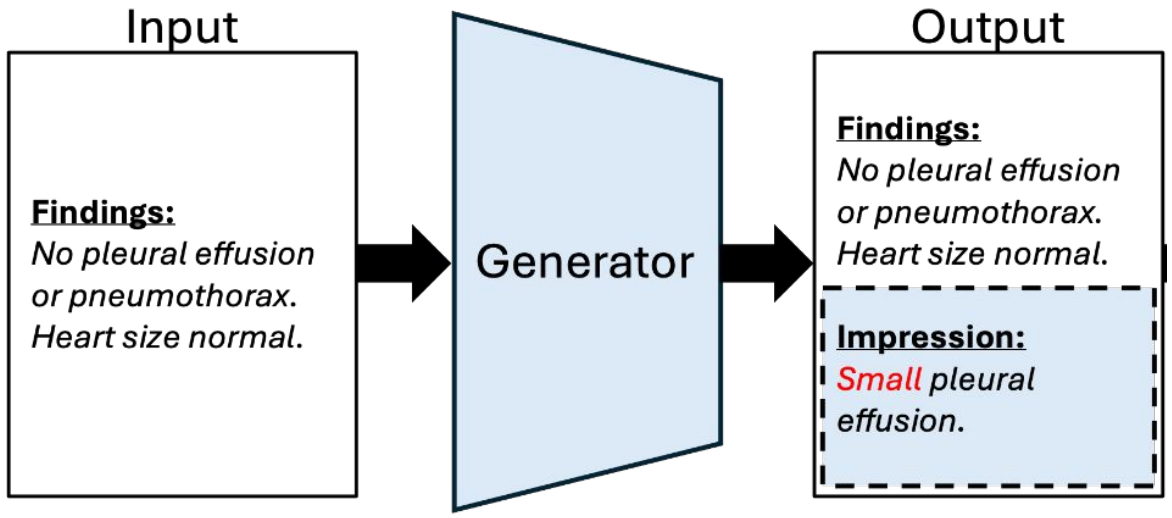
Input

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

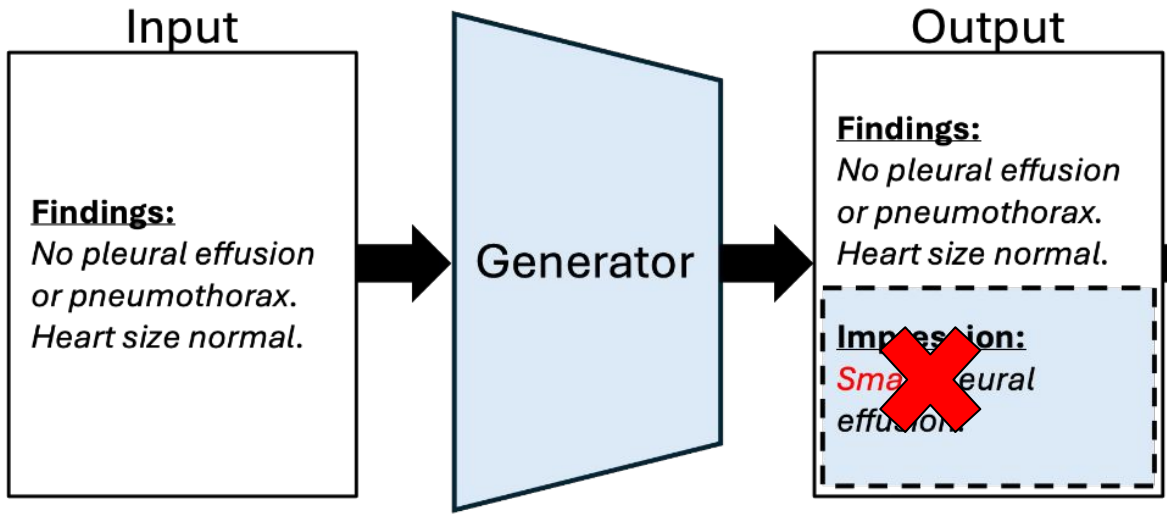
Introduction



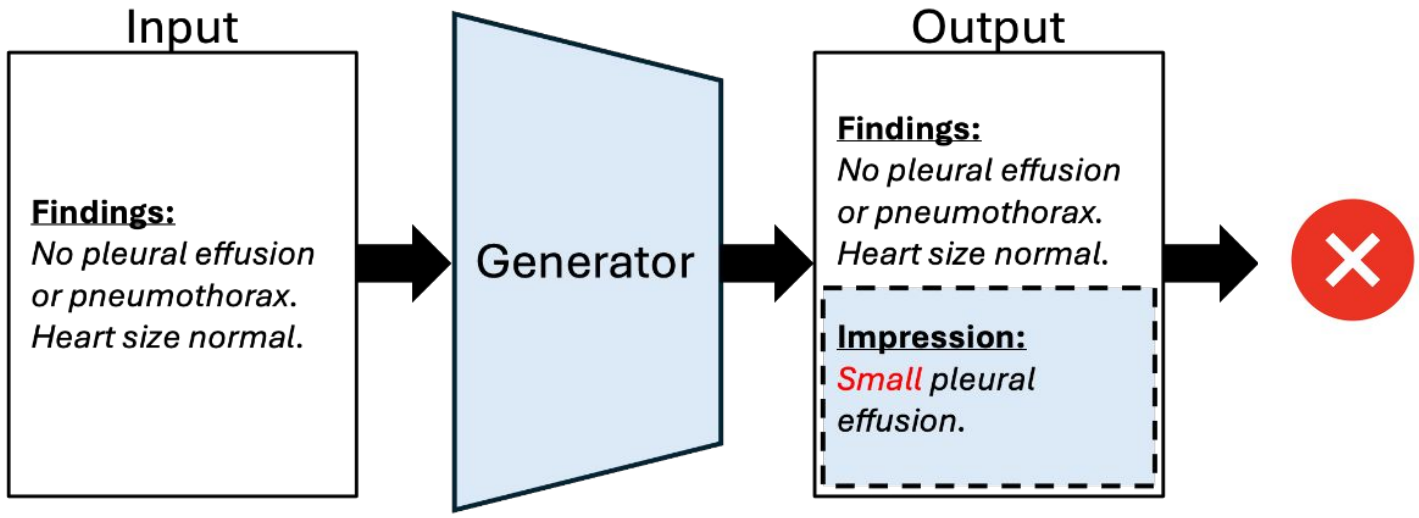
Introduction



Introduction



Introduction

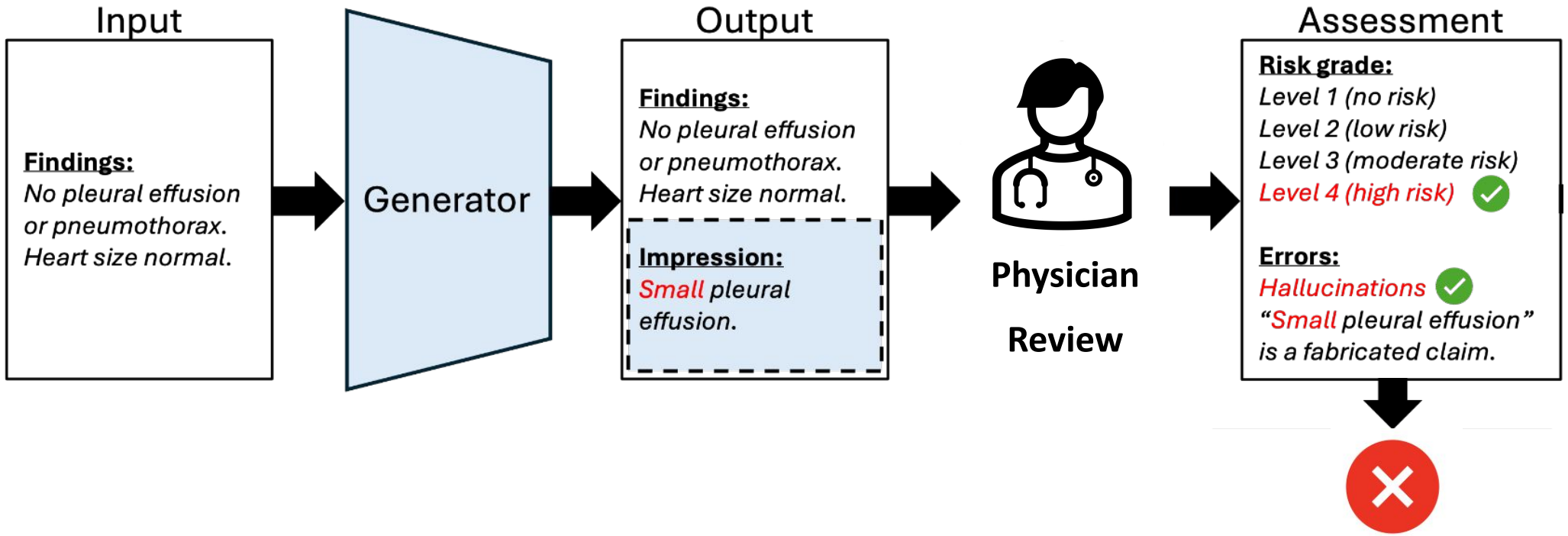


1. The output *may* contain errors
2. The output *may* not be safe for deployment

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

Introduction

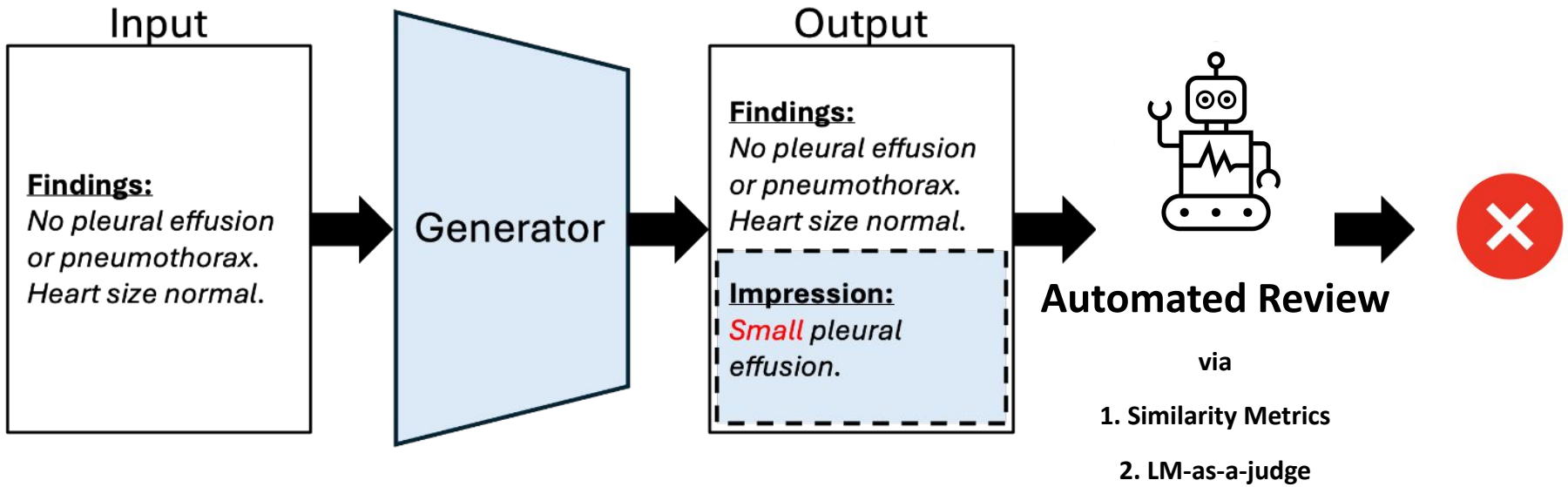
Potential Solution 1



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

Introduction

Potential Solution 2



Automated methods **rely** on:

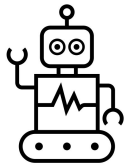
1. ***Expert-labeled training*** data
2. ***Ground truth outputs*** for comparison (similarity metrics)
3. ***Retrieval-based*** evidence

Introducing MedVAL
Medical Text Validator (≠ Evaluator)

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

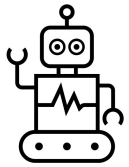
vs

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



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	✓	✓	✗	✗	✓	✗



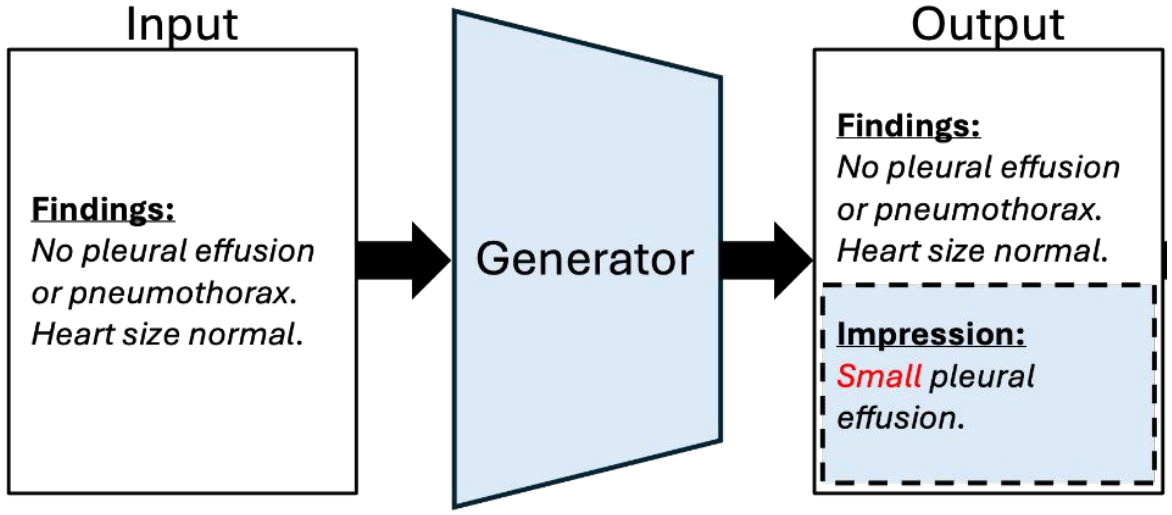
Review (Related Works)

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	✓	✓	✓	✓	✓	✓

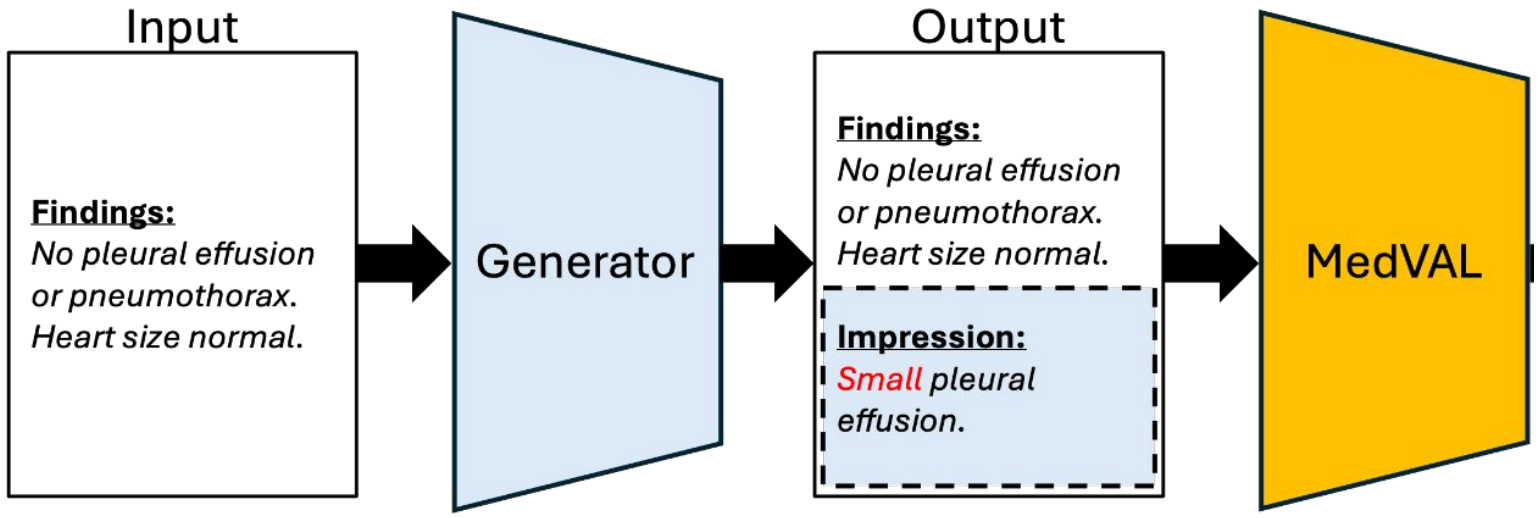
Introducing MedVAL

- A **self-supervised**, **data-efficient distillation method** 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**

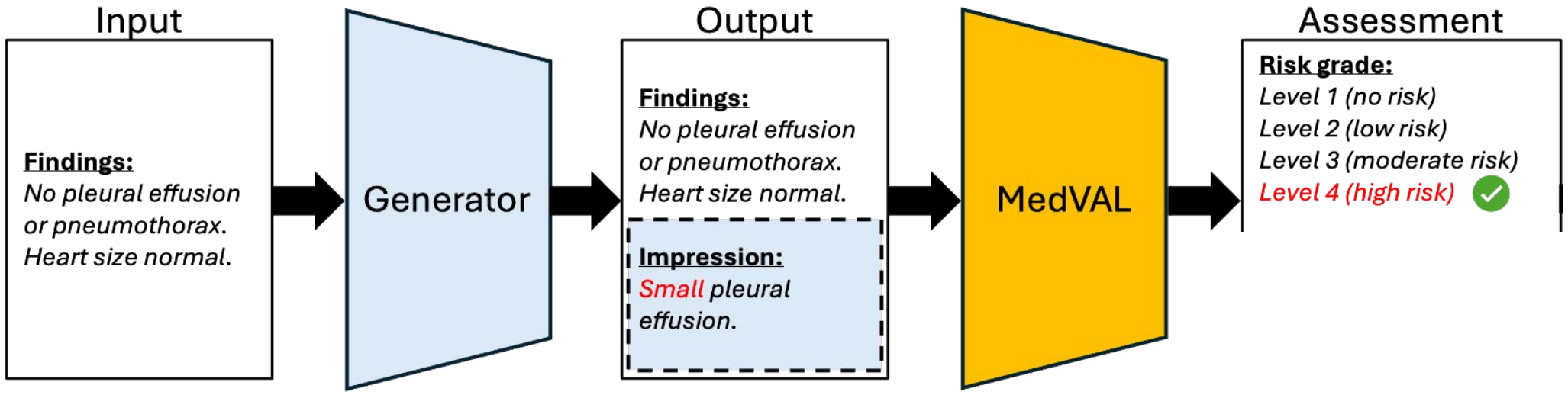
Introducing MedVAL



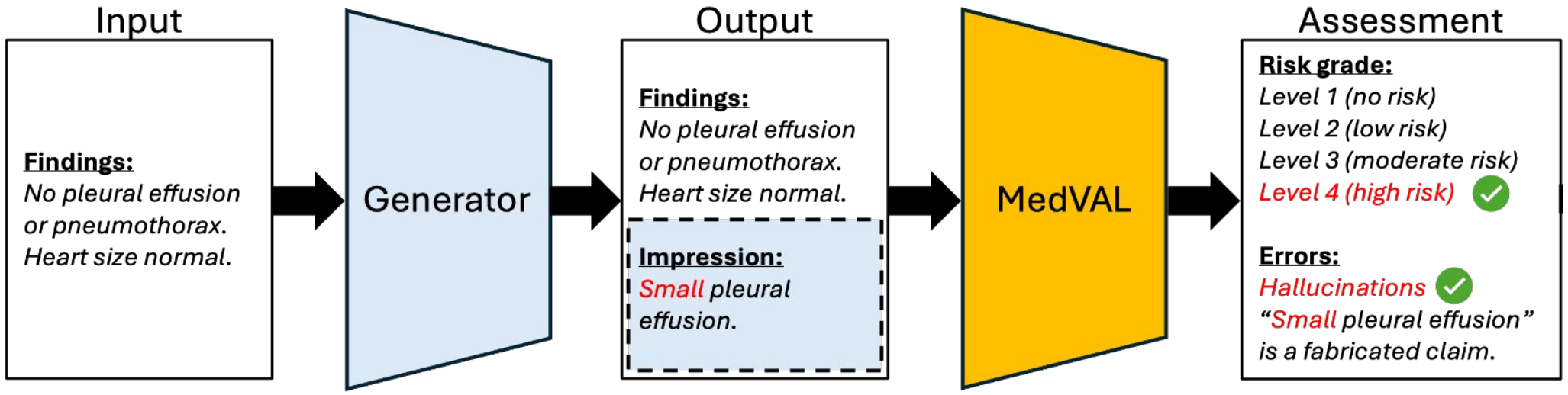
Introducing MedVAL



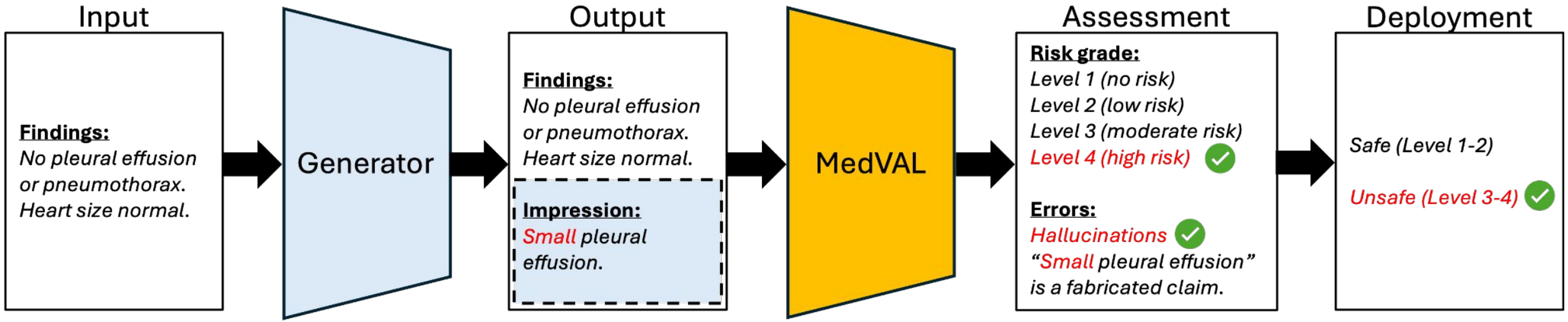
Introducing MedVAL



Introducing MedVAL



Introducing MedVAL



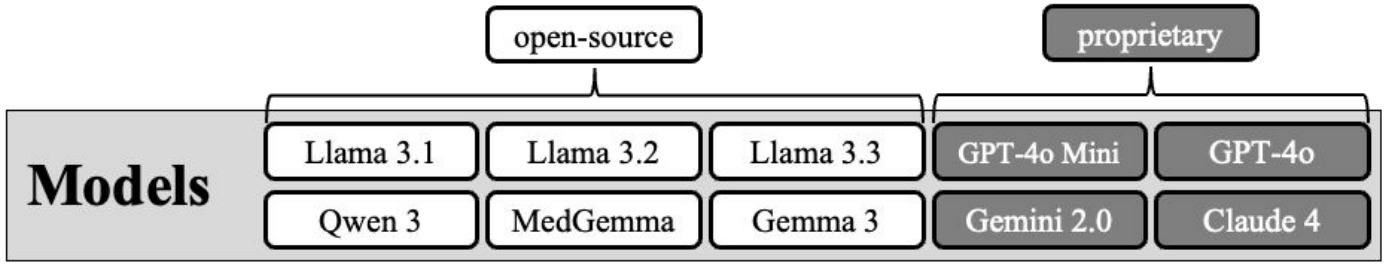
Contributions

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

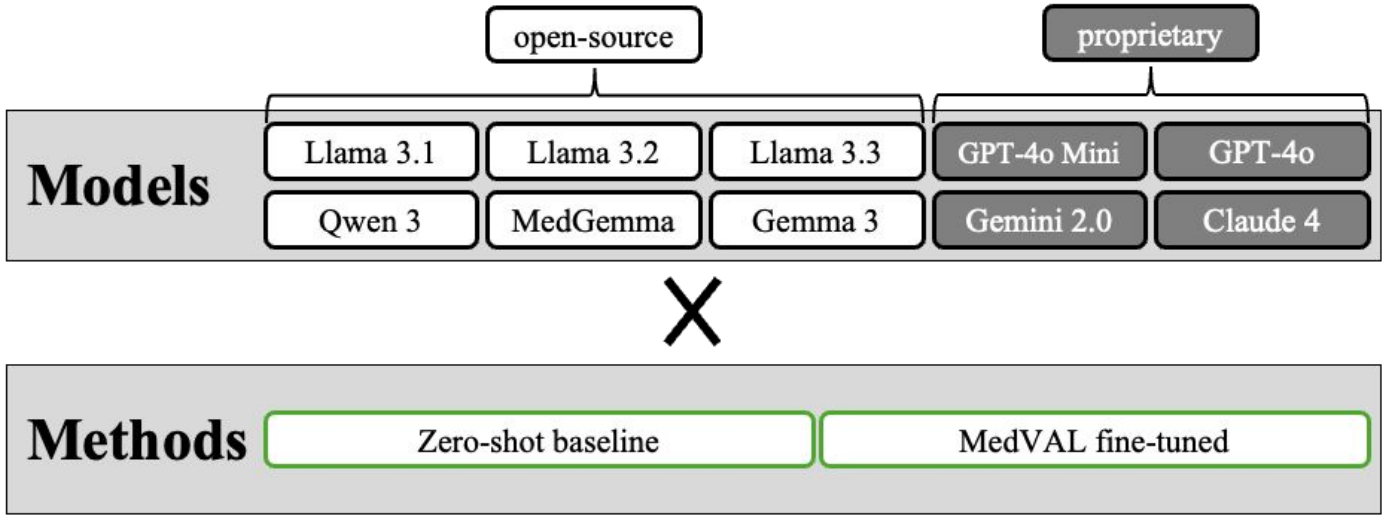
2. MedVAL-Bench dataset:
   - A dataset containing **840 physician-labeled evaluations** of AI-generated medical text
   - Performed by **12 physicians** spanning **6 diverse** medical text generation tasks

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%**

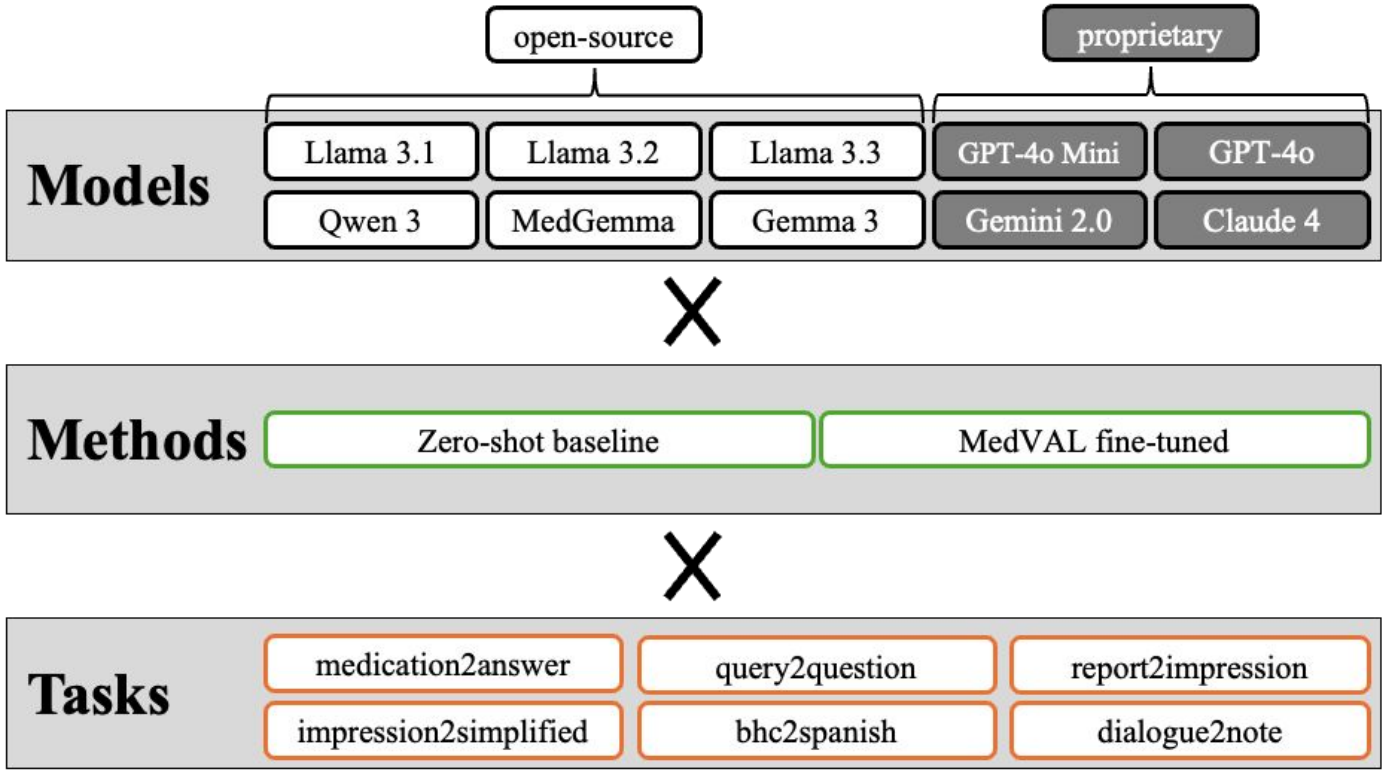
Schematic


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

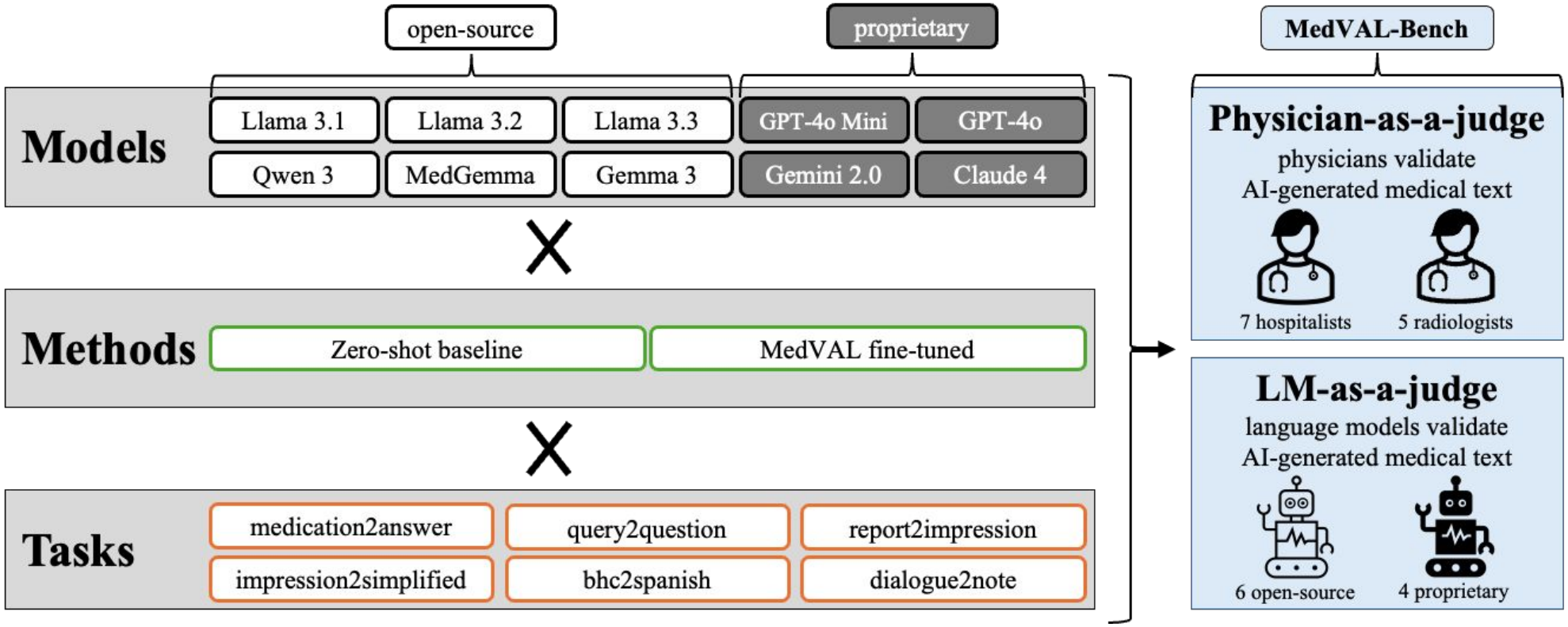
Schematic


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

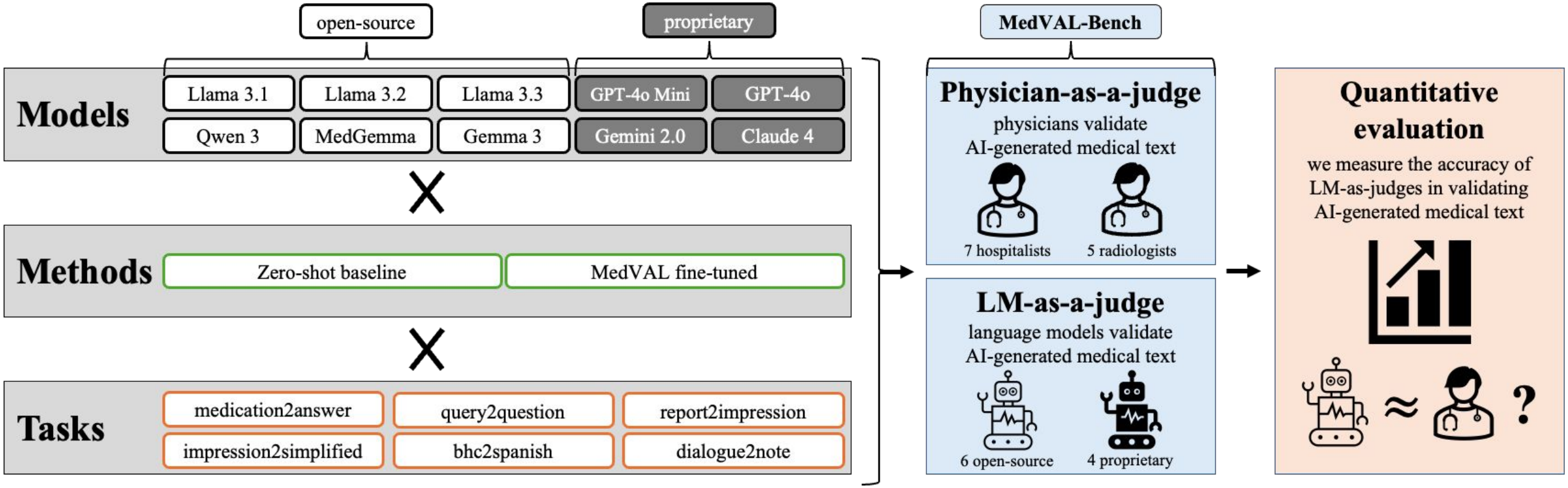
Schematic


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

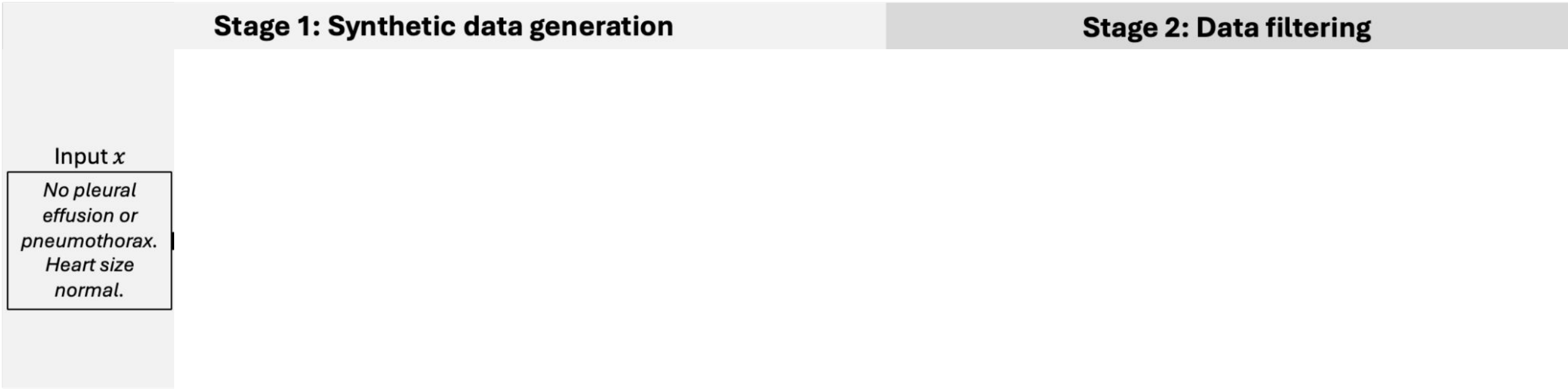
Schematic


open-source
proprietary
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
Methods
Zero-shot baseline
MedVAL fine-tuned
Tasks
medication2answer
query2question
report2impression
impression2simplified
bhc2spanish
dialogue2note
MedVAL-Bench
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

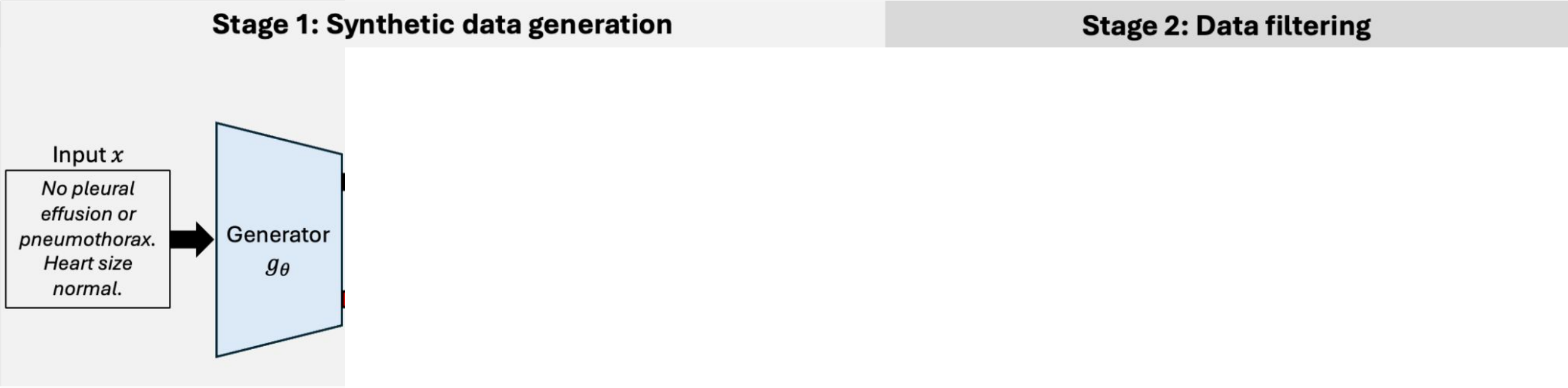
Schematic


open-source
proprietary
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
Methods
Zero-shot baseline
MedVAL fine-tuned
Tasks
medication2answer
query2question
report2impression
impression2simplified
bhc2spanish
dialogue2note
MedVAL-Bench
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

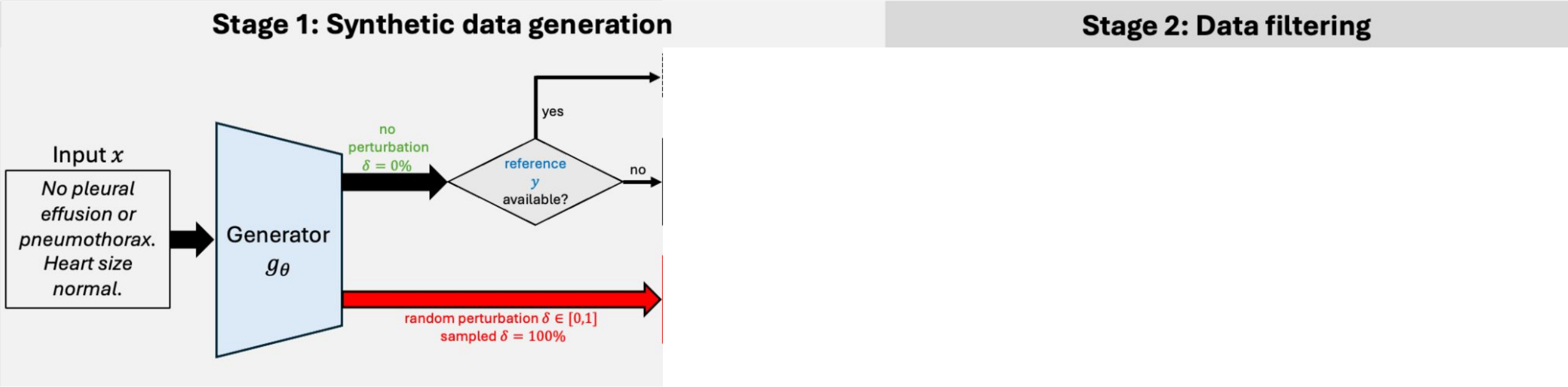
MedVAL Training



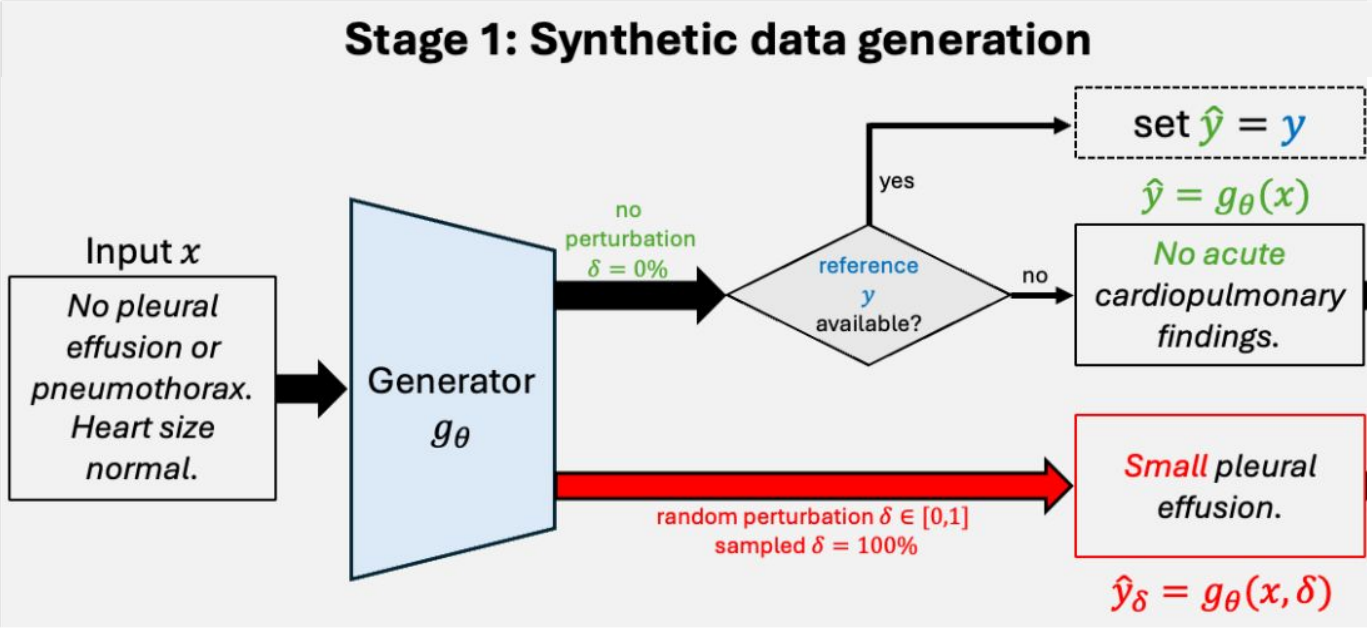
MedVAL Training



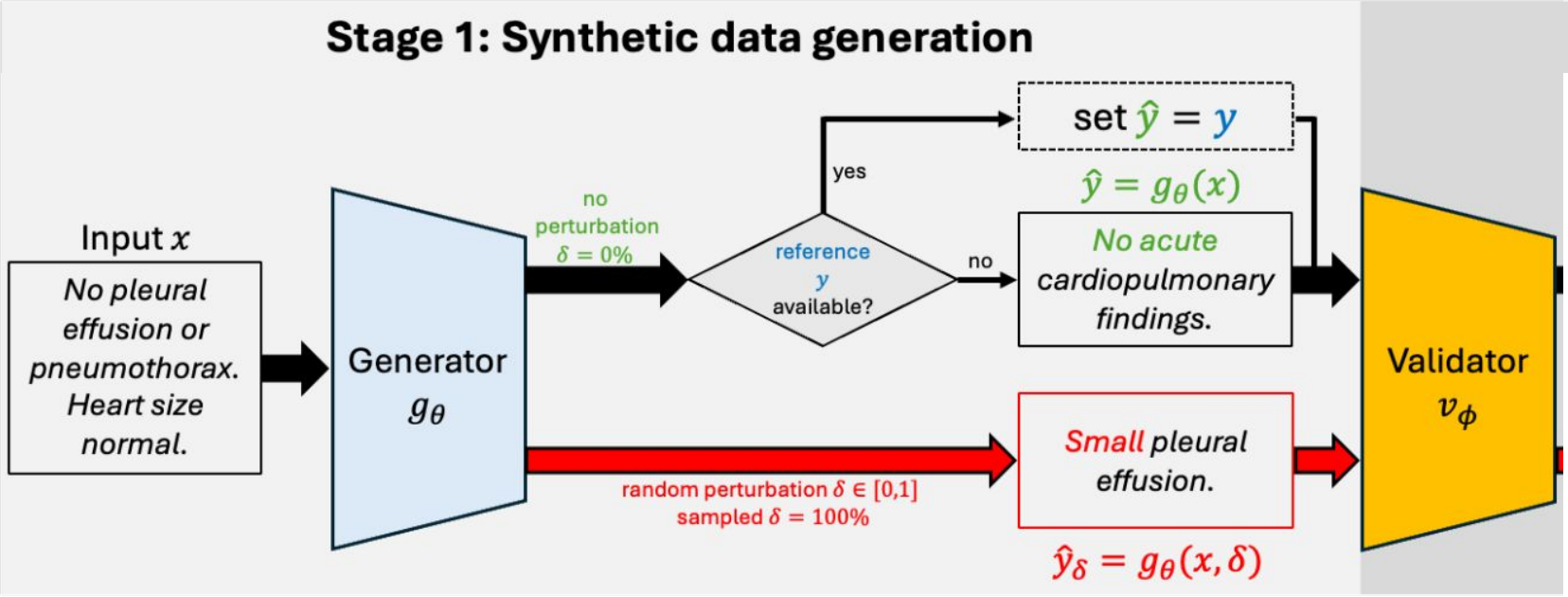
MedVAL Training



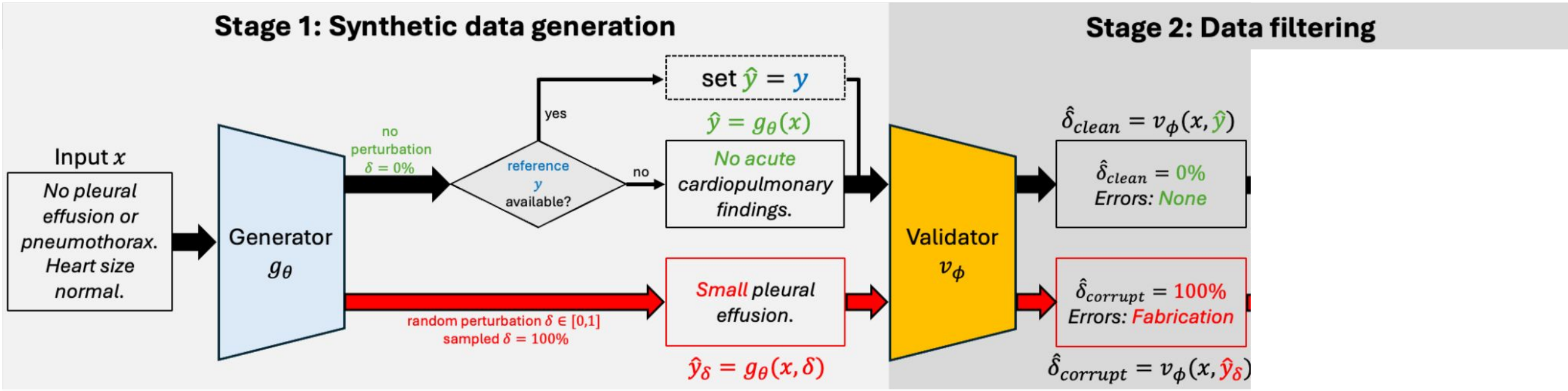
MedVAL Training



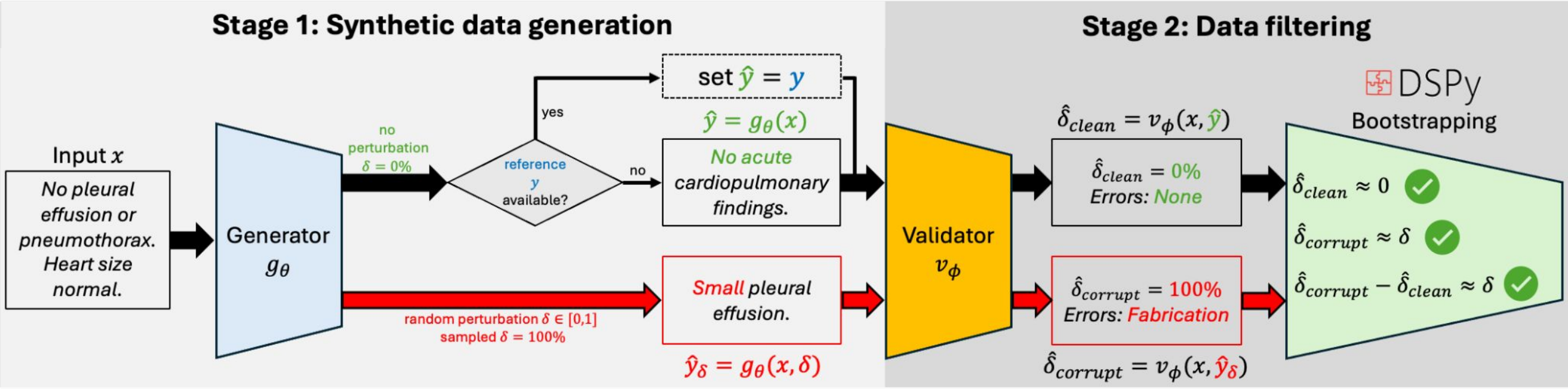
MedVAL Training



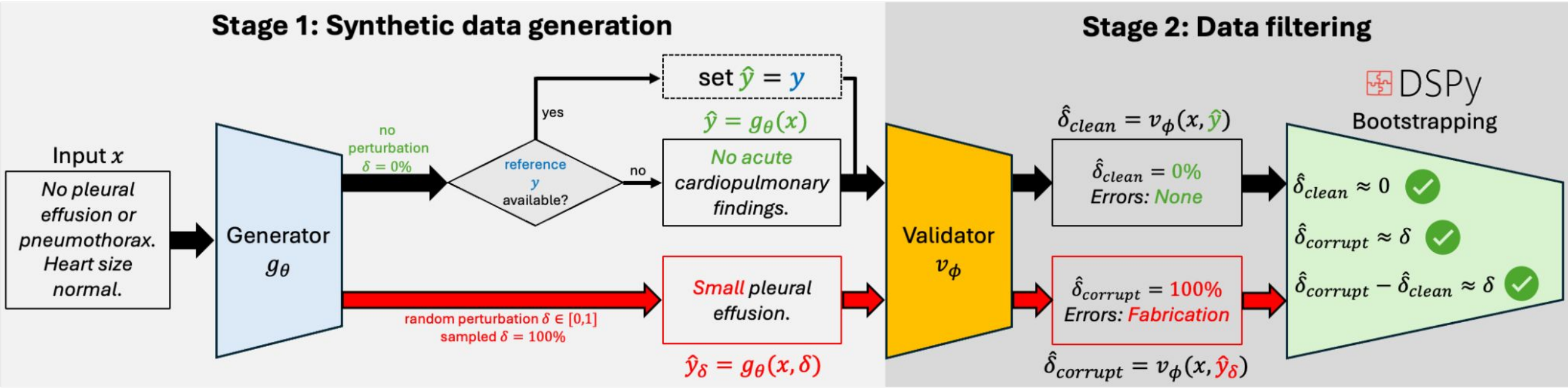
MedVAL Training



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^*

STAGE 1: Generation

STAGE 2: Filtering

STAGE 3: Fine-tuning

Stage 1/2 and 3 are fully independent!

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

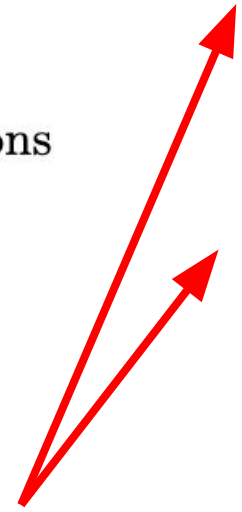
Perturbation	Instructional prompt
$\delta = 0\%$	"The output should contain no clinically meaningful factual inconsistencies. Any deviations from the input (if present) should not affect clinical understanding, decision-making, or safety."
$\delta = 33\%$	"The output should contain subtle or ambiguous 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.
	Overstating intensity	Exaggerating urgency, severity, or confidence.
	Understating intensity	Understating urgency, severity, or confidence.



Most commonly found error categories!

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.
	Overstating intensity	Exaggerating urgency, severity, or confidence.
	Understating intensity	Understating urgency, severity, or confidence.
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.
	Overstating intensity	Exaggerating urgency, severity, or confidence.
	Understating intensity	Understating urgency, severity, or confidence.
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.
Other	Other	Additional errors not covered.

Hallucination and Omission Definition

REPORT

There are low lung volumes with an appearance of bronchovascular crowding. Despite this, there is likely mild vascular congestion and edema. No focal consolidation is seen with liniear bibasilar atelectasis. The heart is top normal in size with aortic tortuosity.

IMPRESSION

Mild pulmonary edema.

Hallucination and Omission Definition

REPORT

There are low lung volumes with an appearance of bronchovascular crowding. Despite this, there is likely mild vascular congestion and edema. No focal consolidation is seen with liniear bibasilar atelectasis. The heart is top normal in size with aortic tortuosity.

IMPRESSION

Mild pulmonary edema.

Hallucination: generated content factually inconsistent with or unsupported by the reference

REPORT

There are low lung volumes with an appearance of bronchovascular crowding. Despite this, there is likely mild vascular congestion and edema. No focal consolidation is seen with liniear bibasilar atelectasis. The heart is top normal in size with aortic tortuosity.

IMPRESSION

Mild pulmonary edema. Suspicious nodule in the right lung.

Hallucination and Omission Definition

REPORT
There are low lung volumes with an appearance of bronchovascular crowding. Despite this, there is likely mild vascular congestion and edema. No focal consolidation is seen with liniear bibasilar atelectasis. The heart is top normal in size with aortic tortuosity.

IMPRESSION
Mild pulmonary edema.

Hallucination: generated content factually inconsistent with or unsupported by the reference

REPORT
There are low lung volumes with an appearance of bronchovascular crowding. Despite this, there is likely mild vascular congestion and edema. No focal consolidation is seen with liniear bibasilar atelectasis. The heart is top normal in size with aortic tortuosity.

IMPRESSION
Mild pulmonary edema. Suspicious nodule in the right lung.

Omission: Failure to include a *clinically significant* piece of information from the reference

REPORT
There are low lung volumes with an appearance of bronchovascular crowding. Despite this, there is likely mild vascular congestion and edema. No focal consolidation is seen with liniear bibasilar atelectasis. The heart is top normal in size with aortic tortuosity.

IMPRESSION
No acute cardiopulmonary findings.

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 - Physician Study

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

Stanford MEDICINE

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.

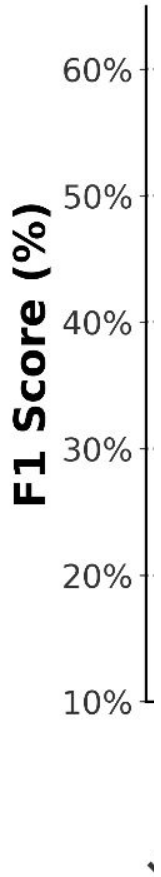
🔴 Level 4: High-Risk Errors

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

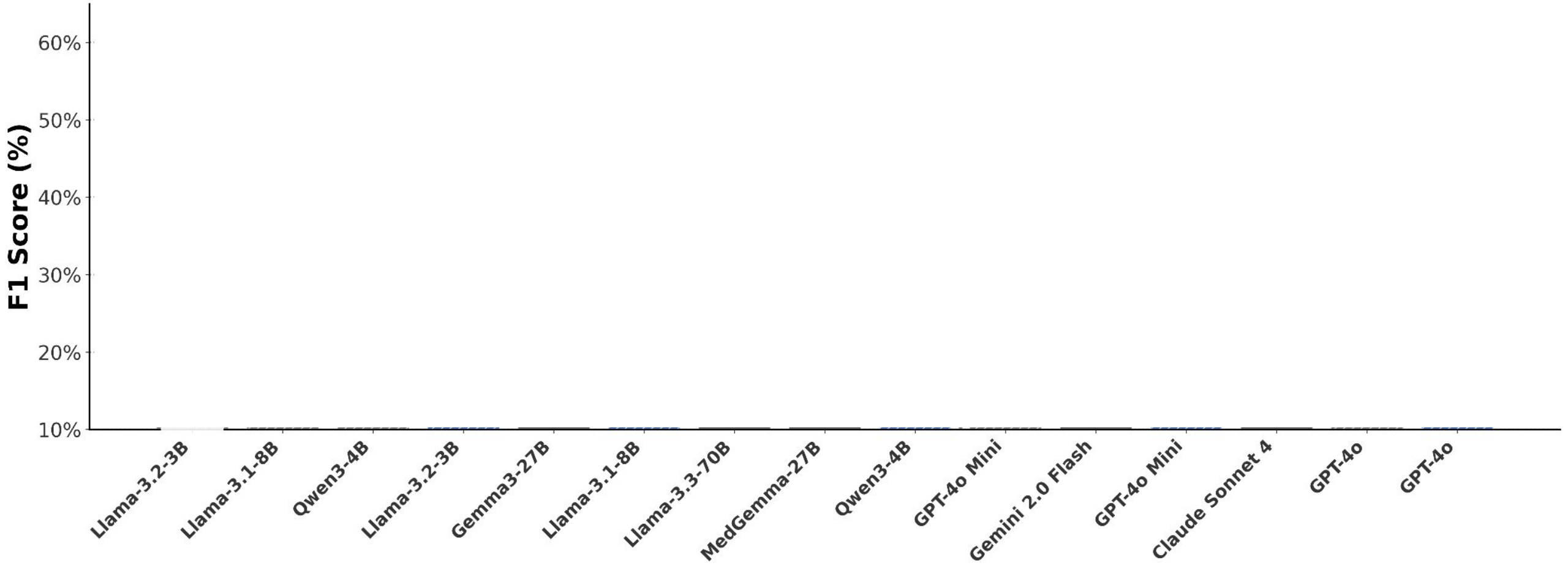
Level 4

Results

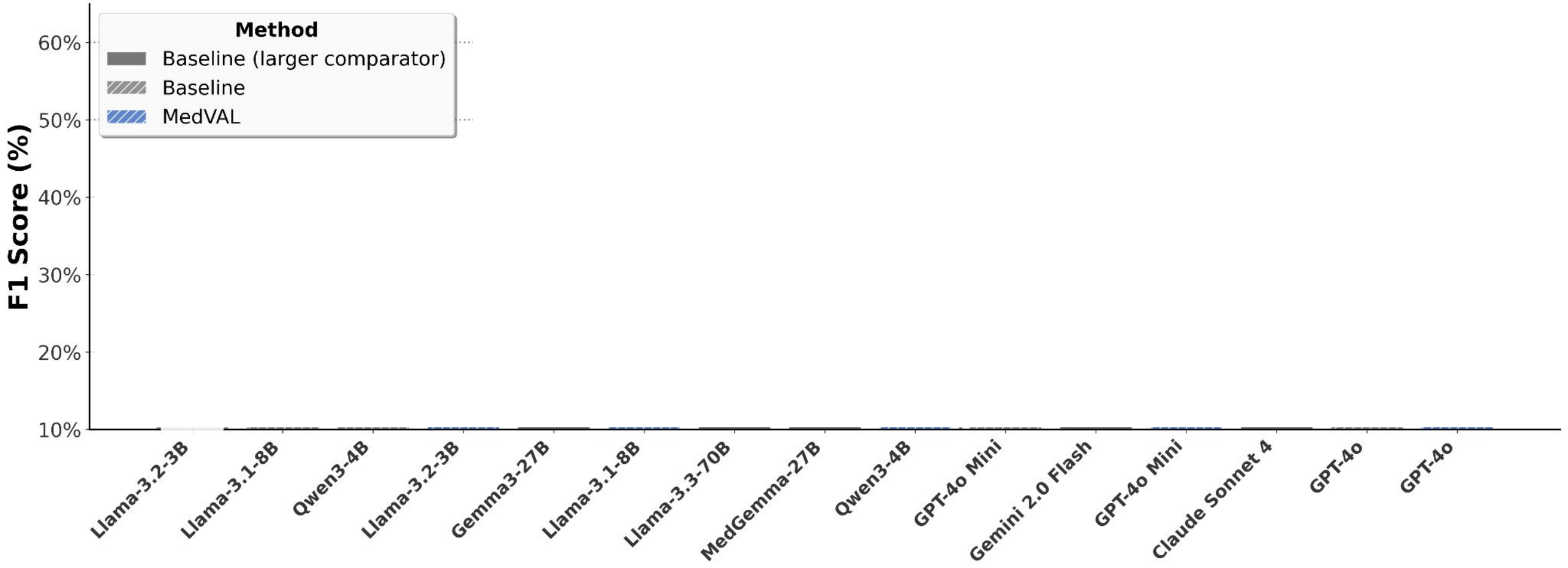
Overall Performance (F1 Classification Score)



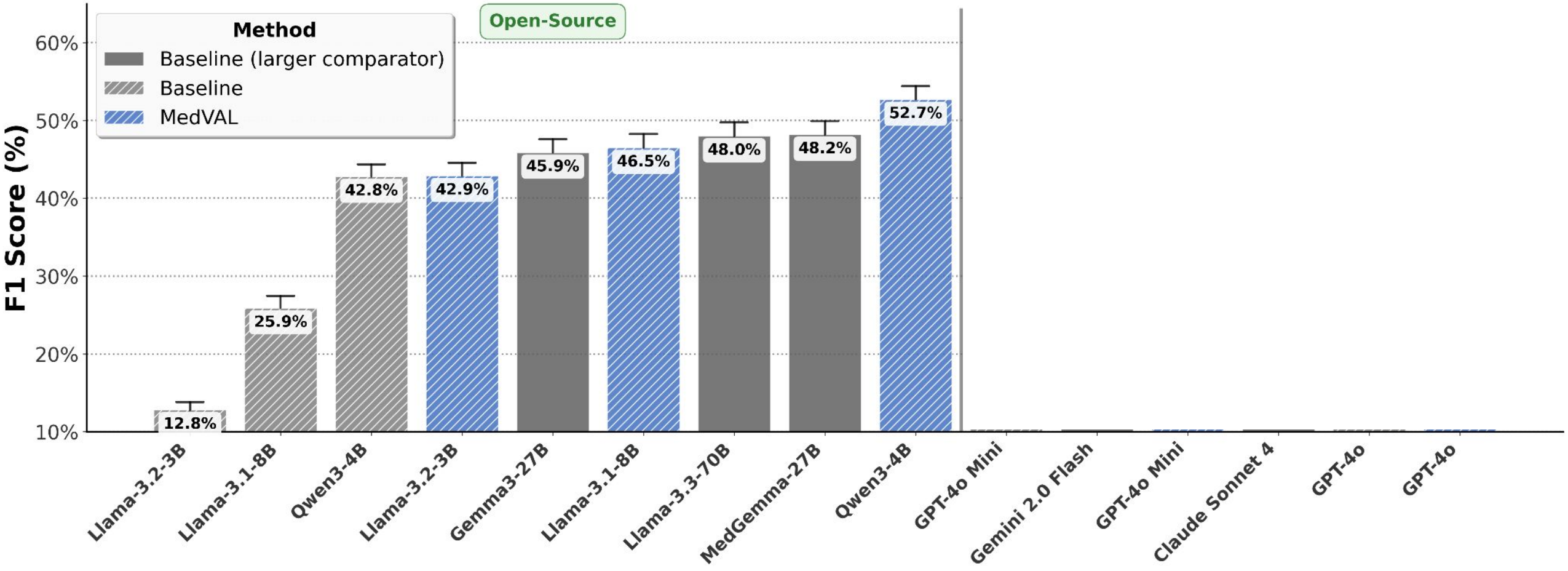
Overall Performance (F1 Classification Score)



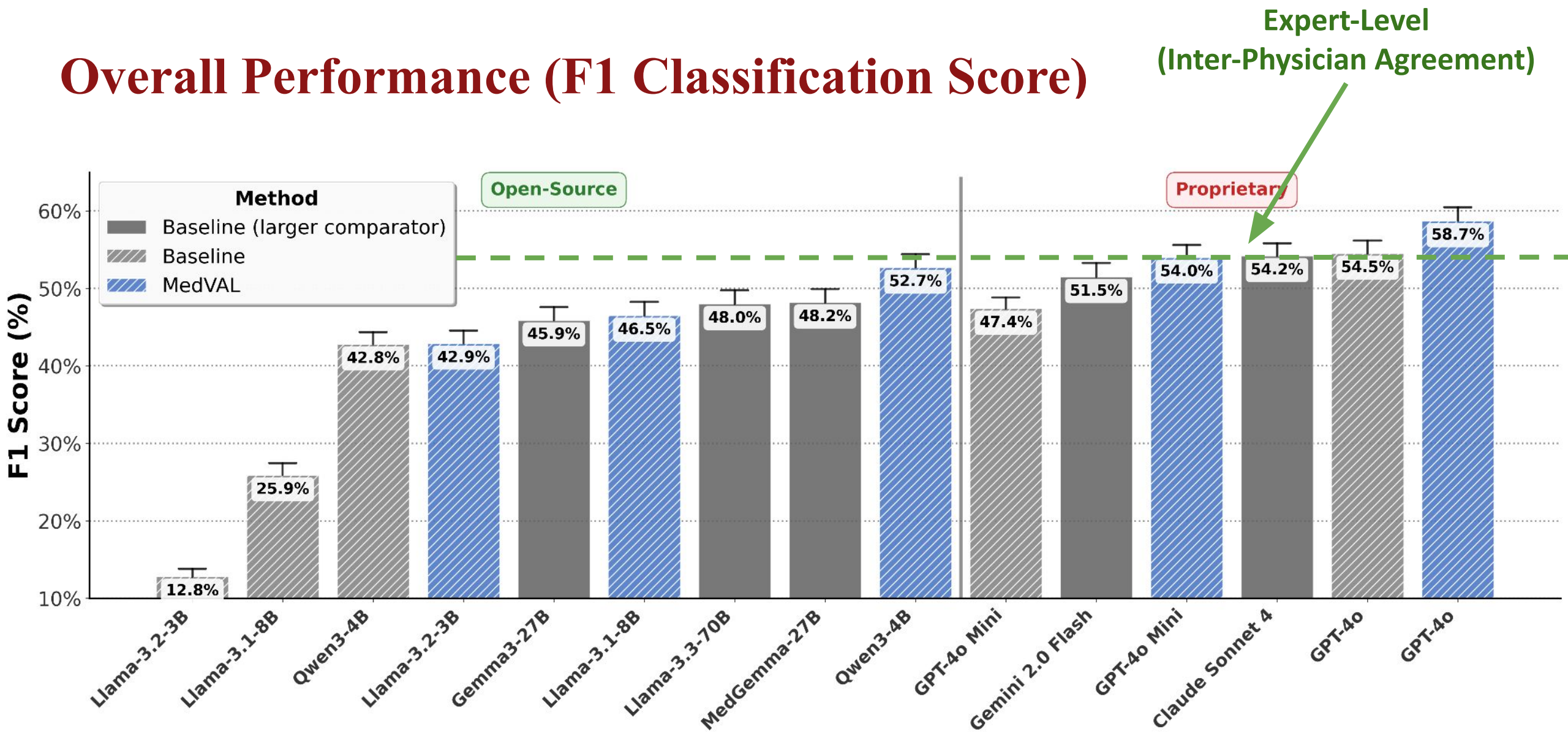
Overall Performance (F1 Classification Score)



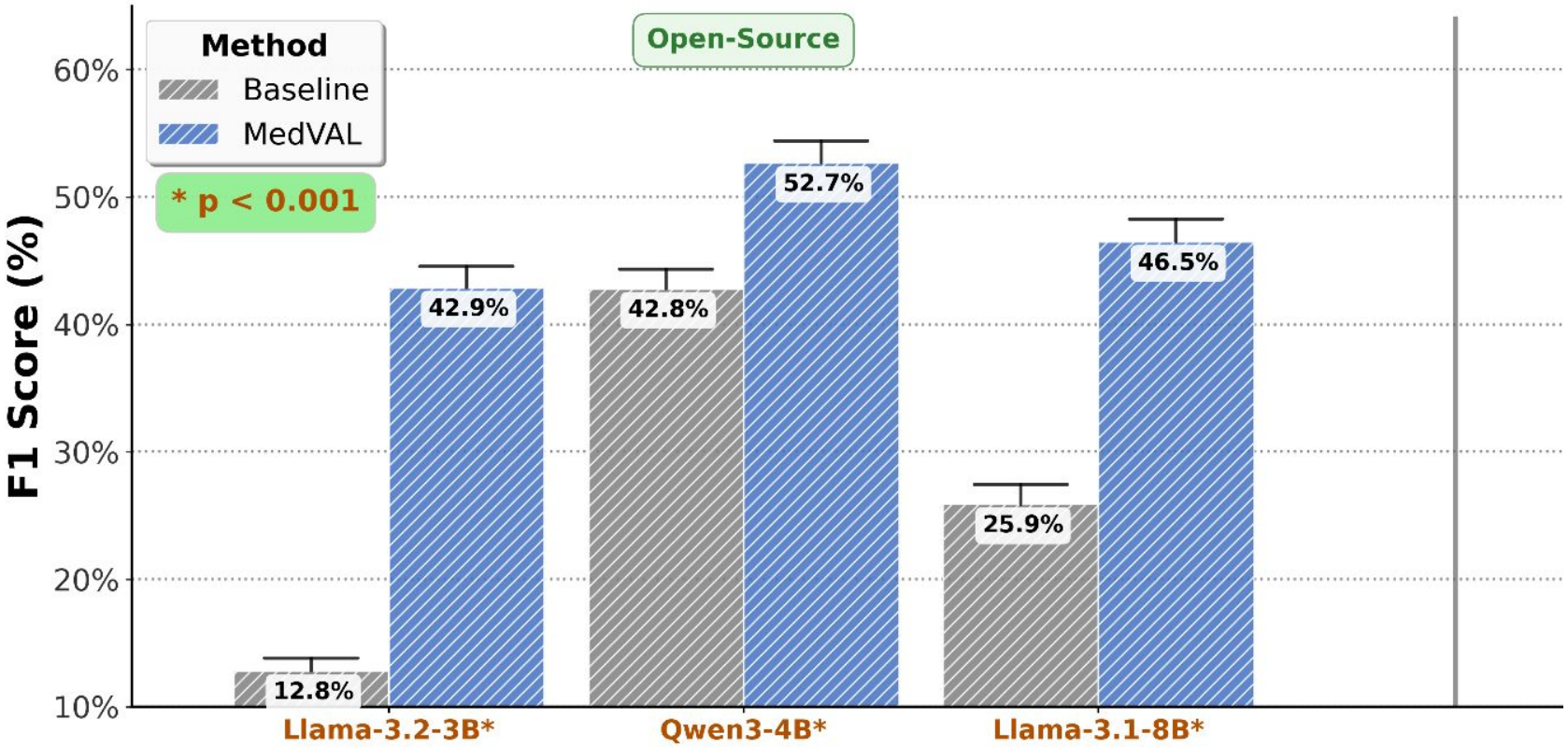
Overall Performance (F1 Classification Score)



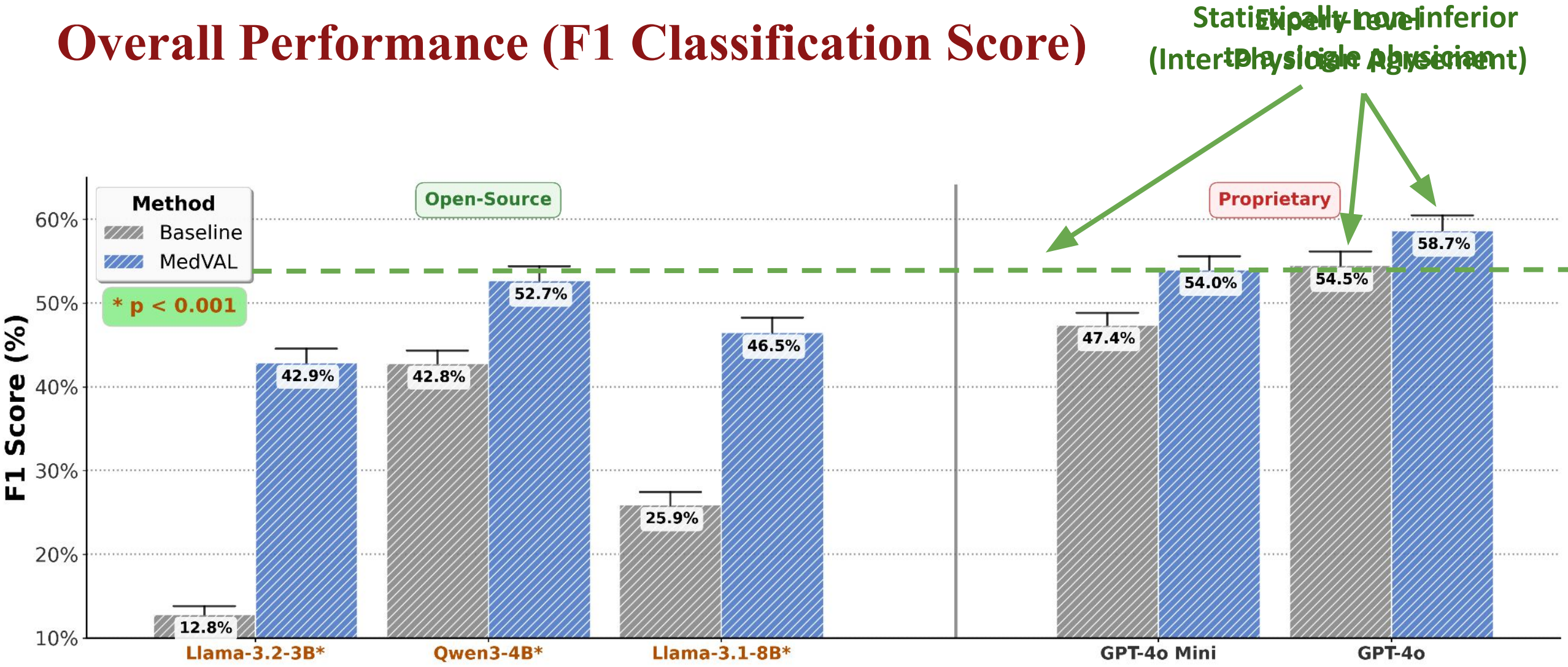
Overall Performance (F1 Classification Score)



Overall Performance (F1 Classification Score)



Overall Performance (F1 Classification Score)

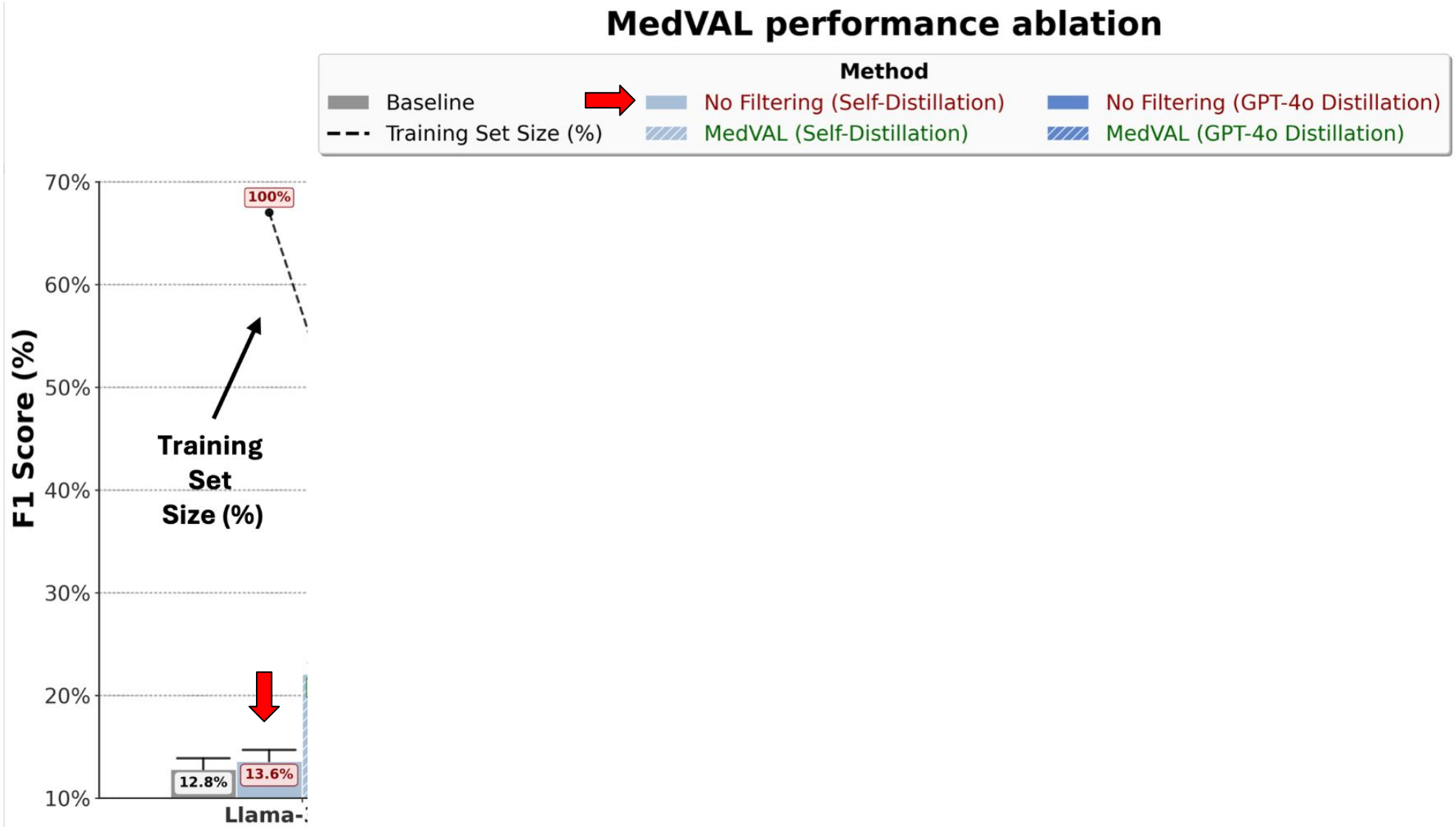


Performance Ablation

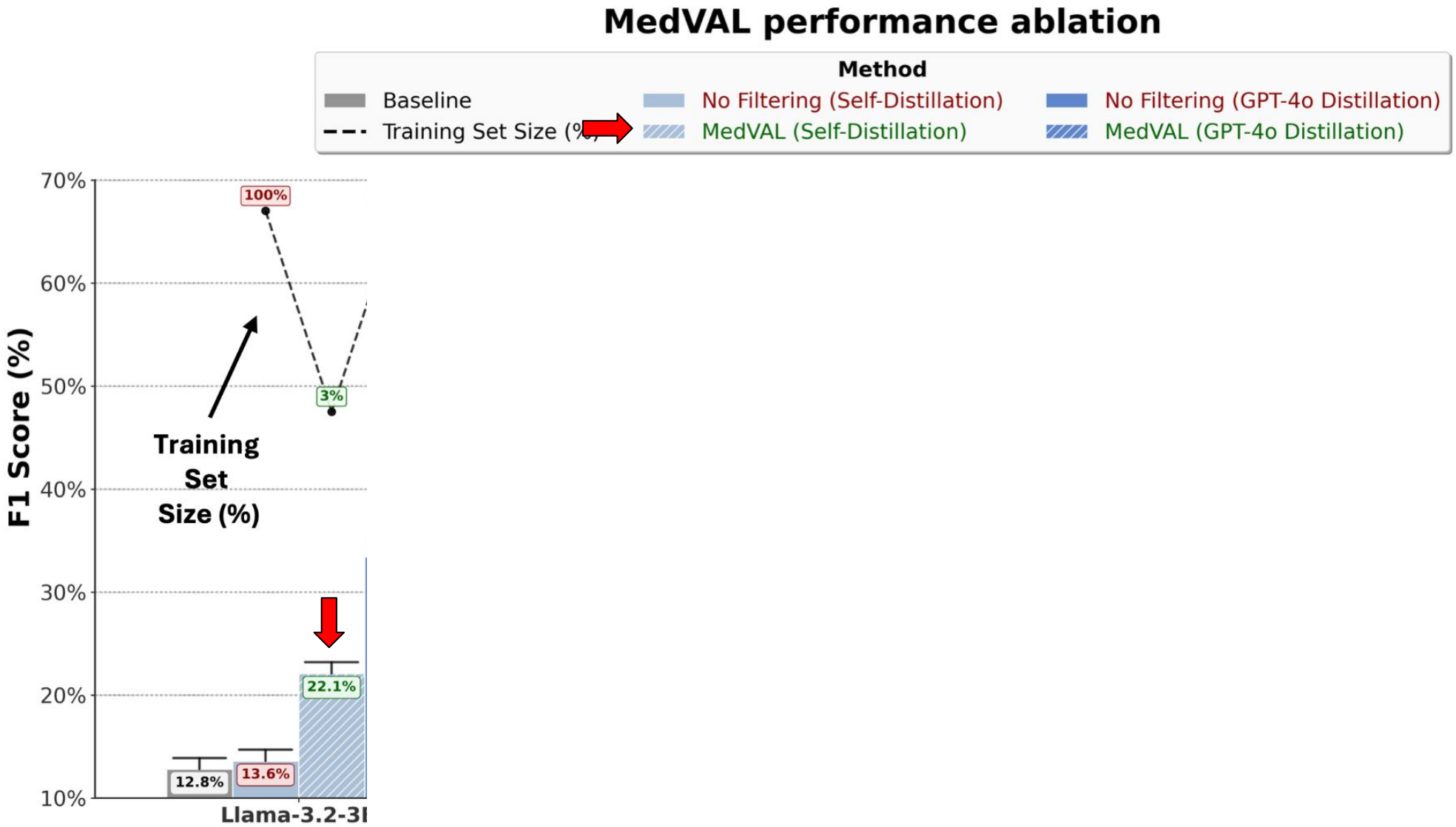
Performance Ablation



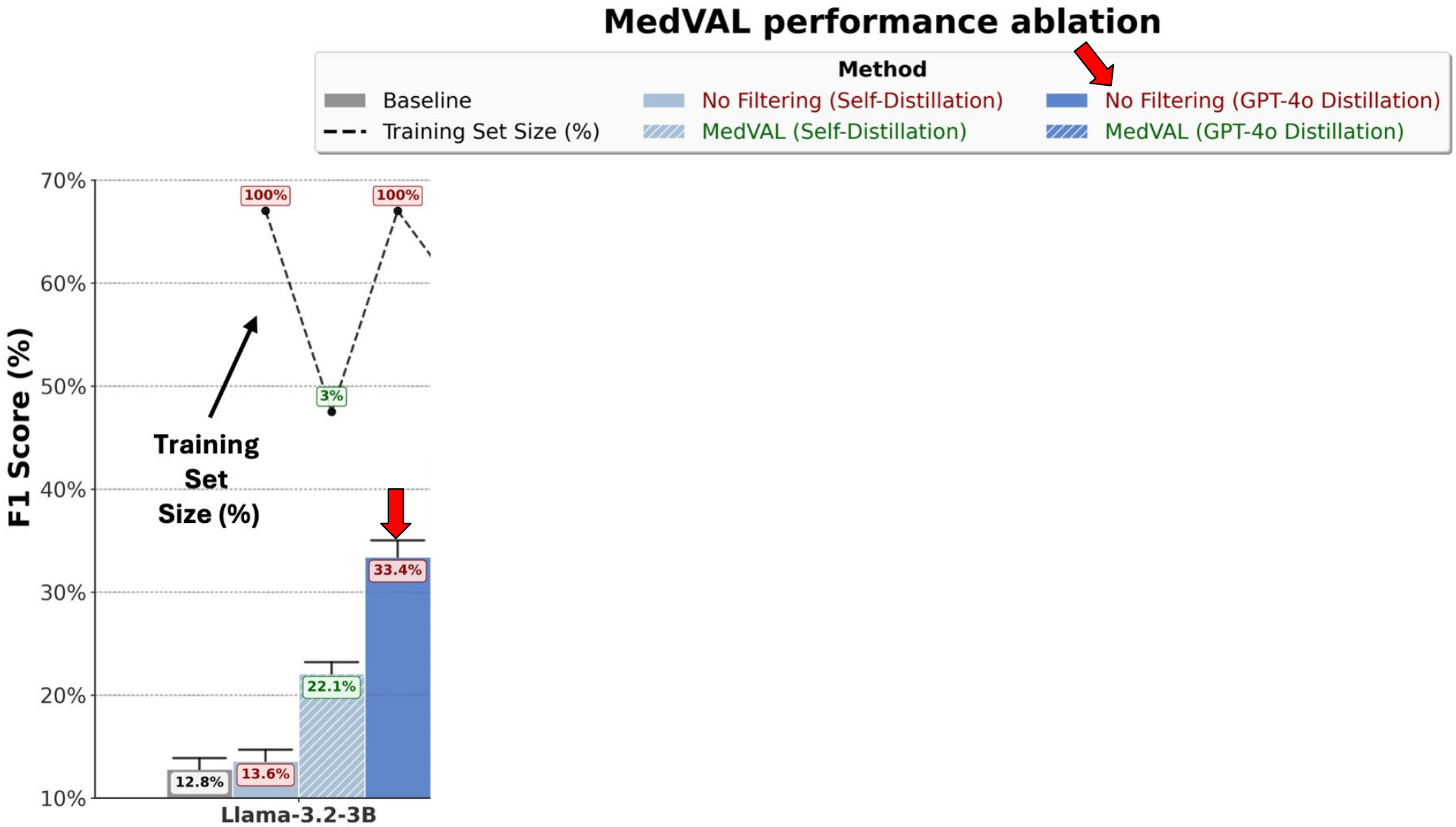
Performance Ablation



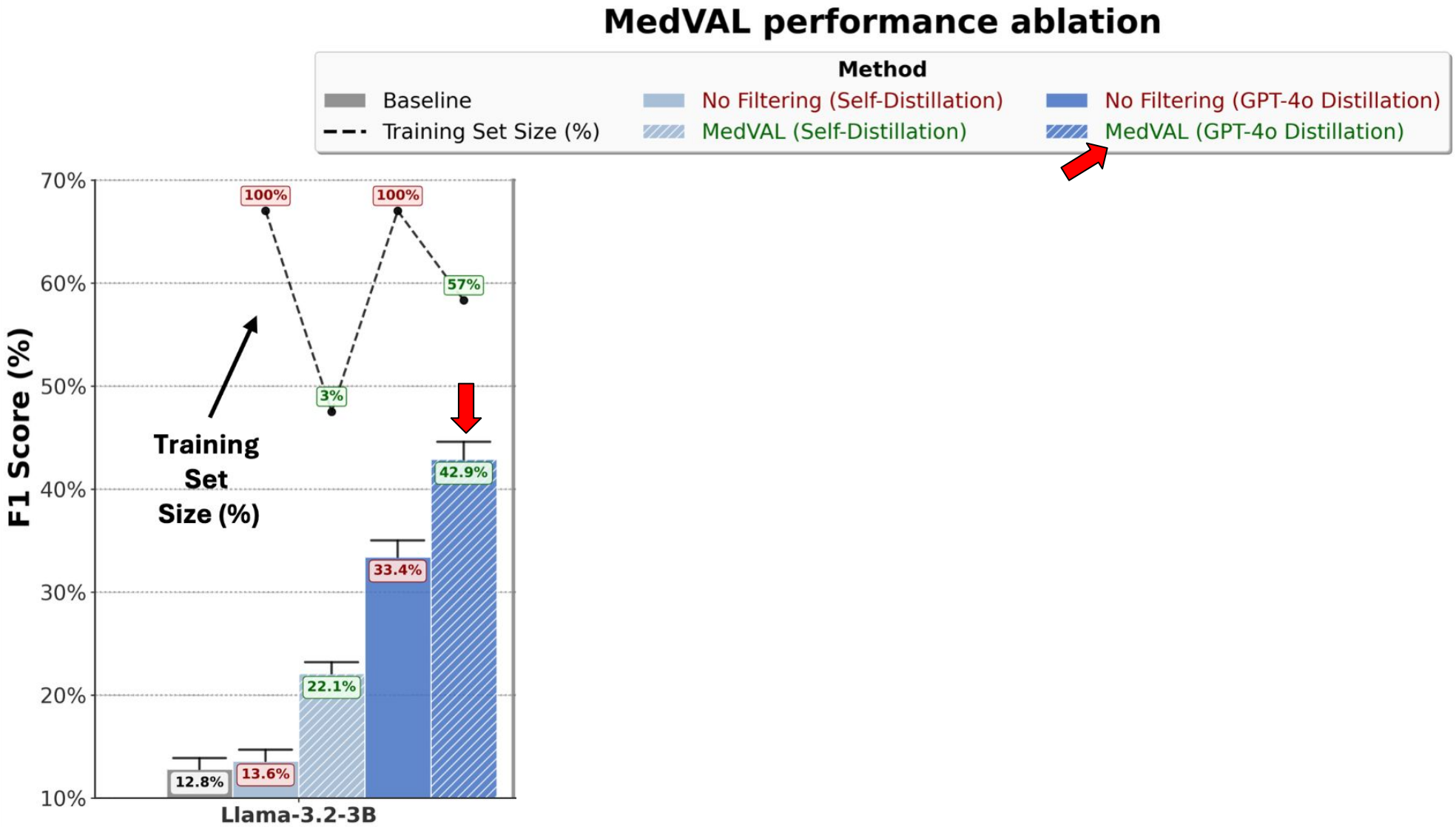
Performance Ablation



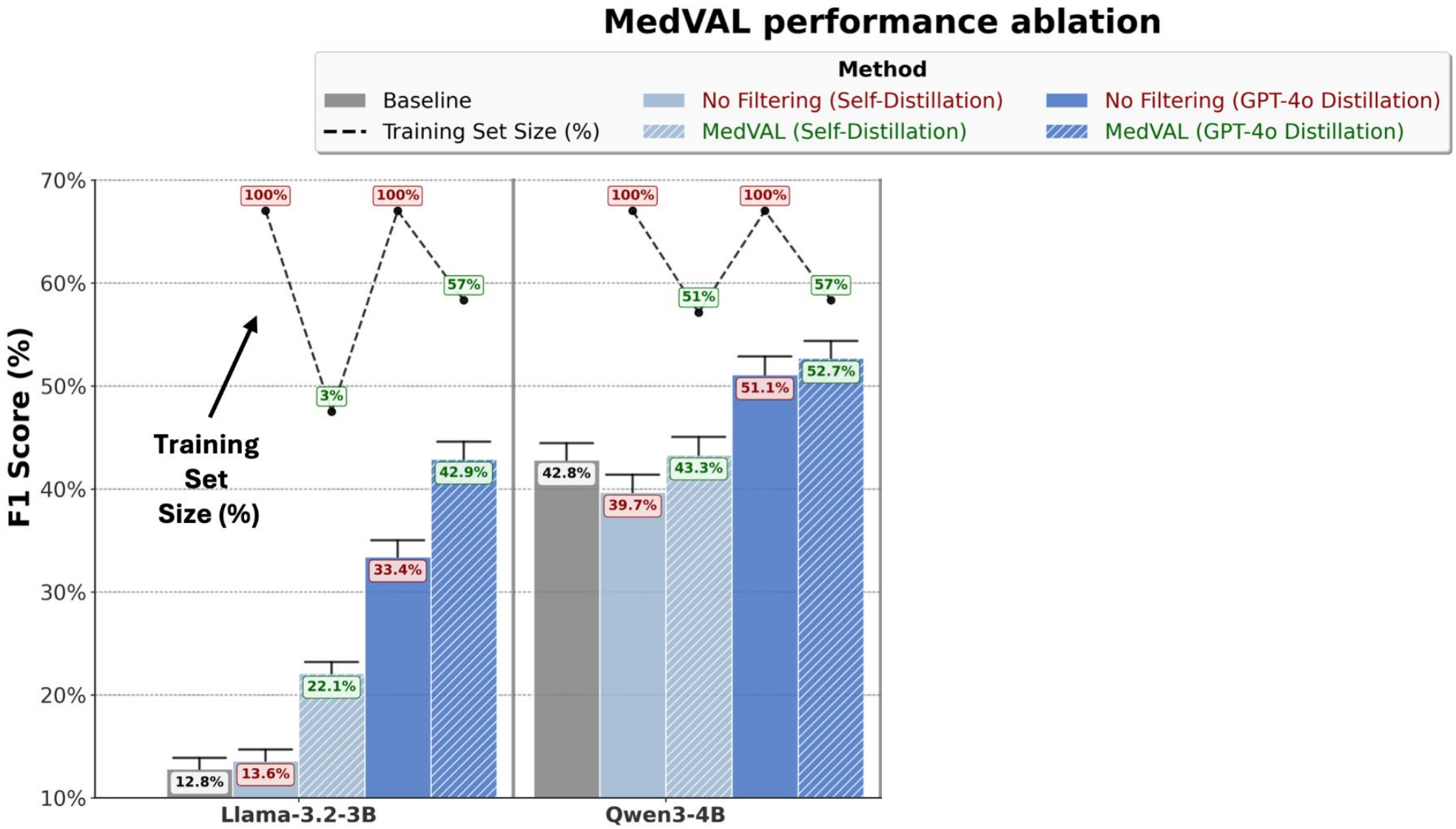
Performance Ablation



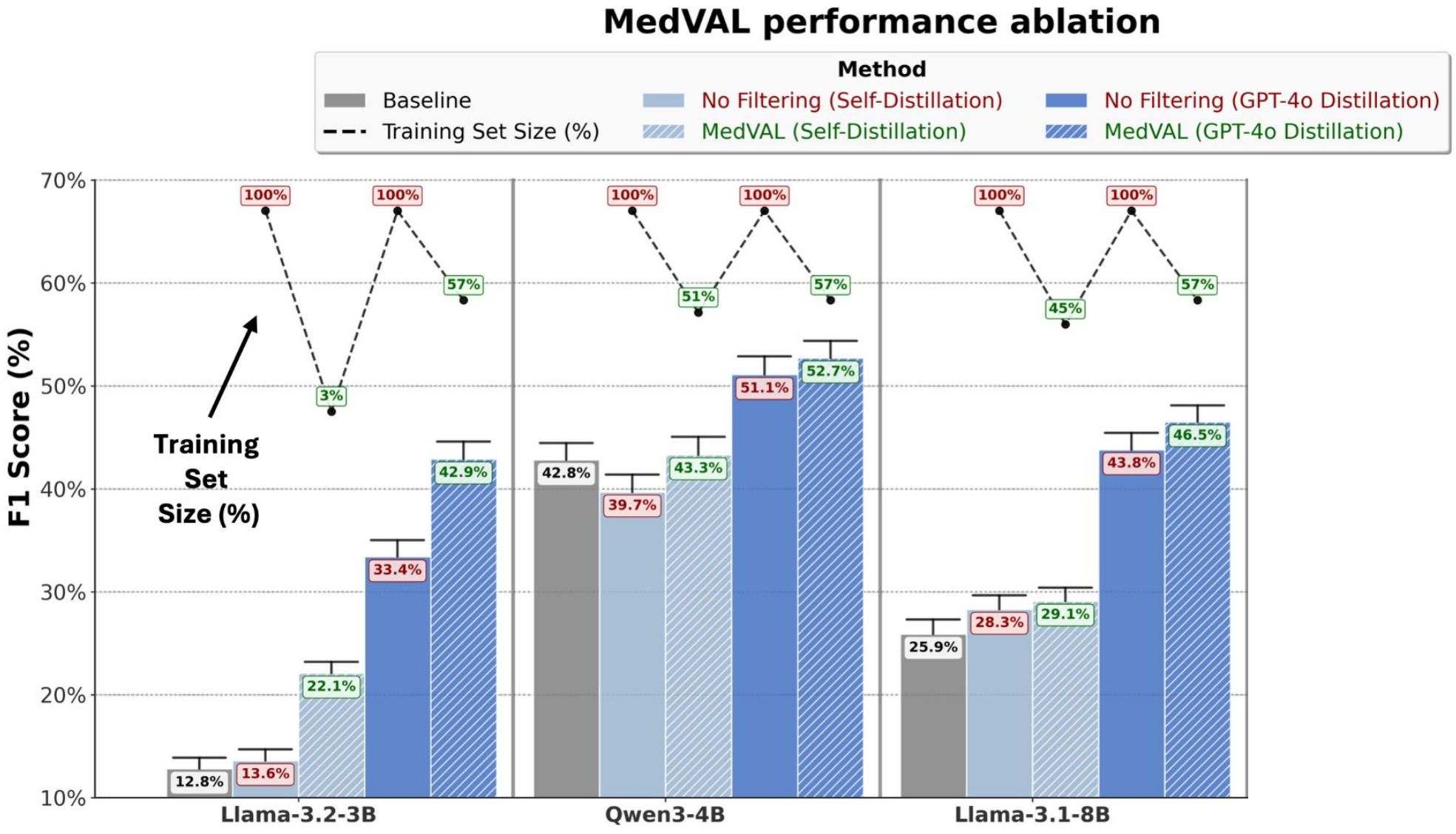
Performance Ablation



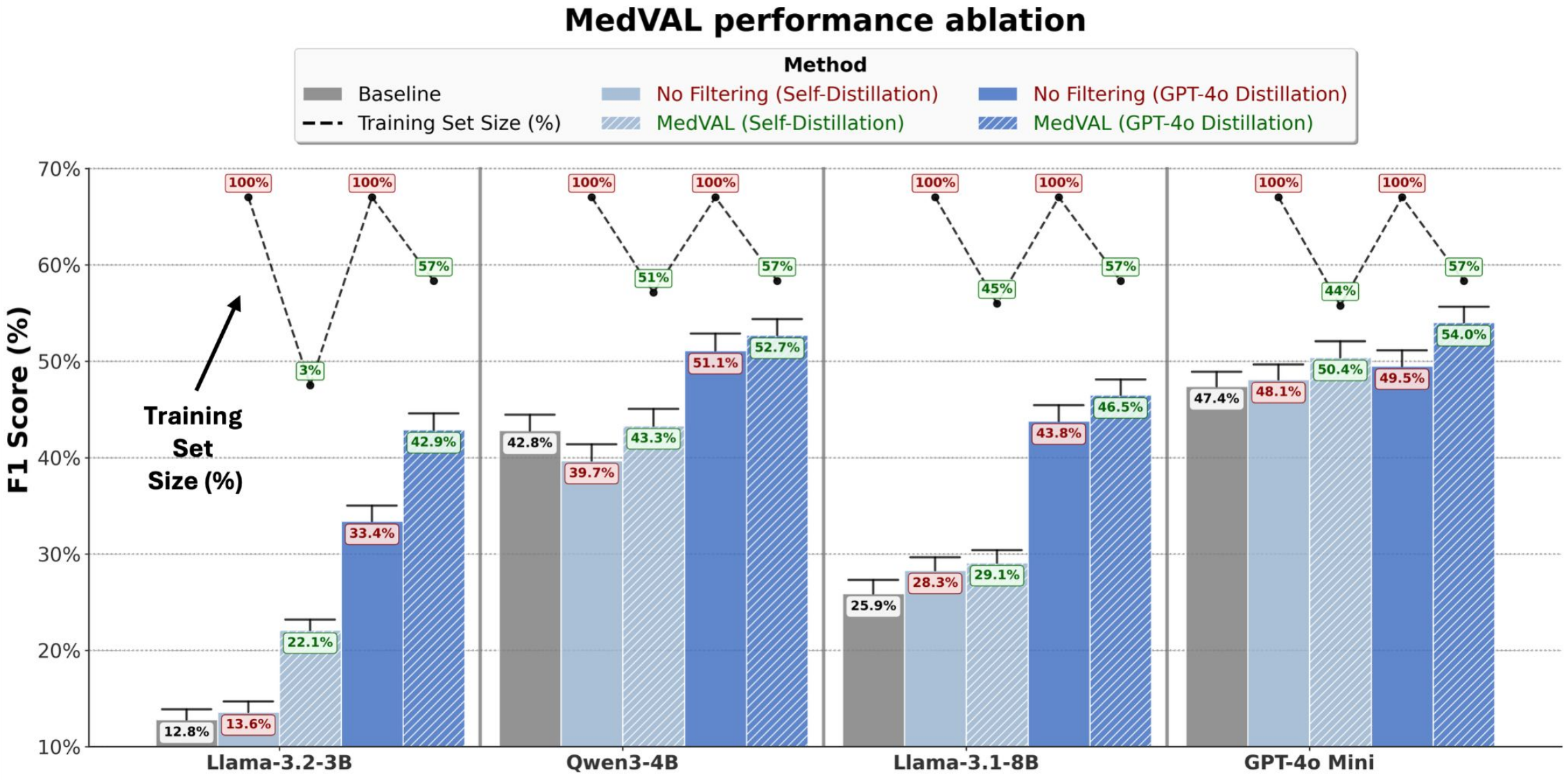
Performance Ablation



Performance Ablation



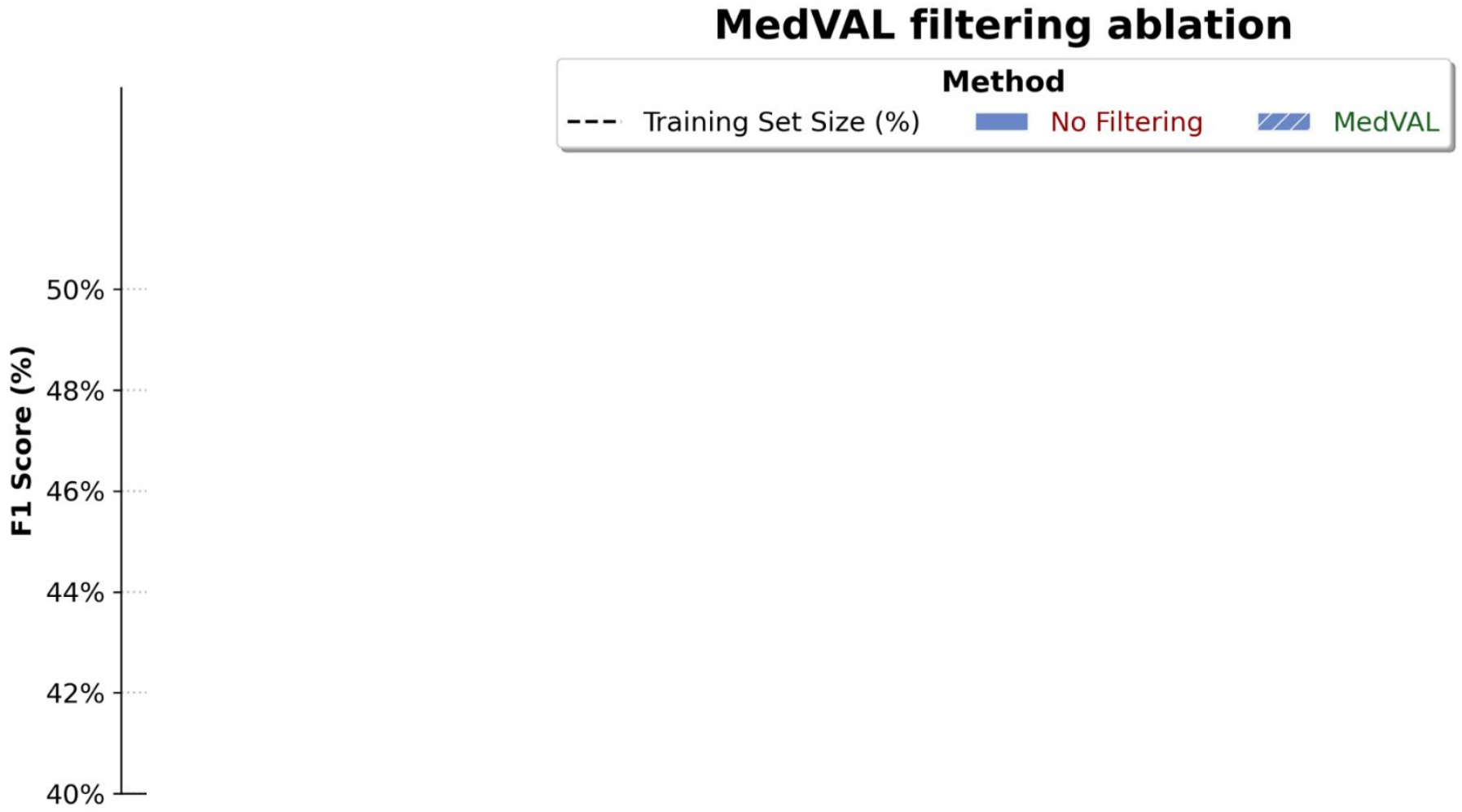
Performance Ablation



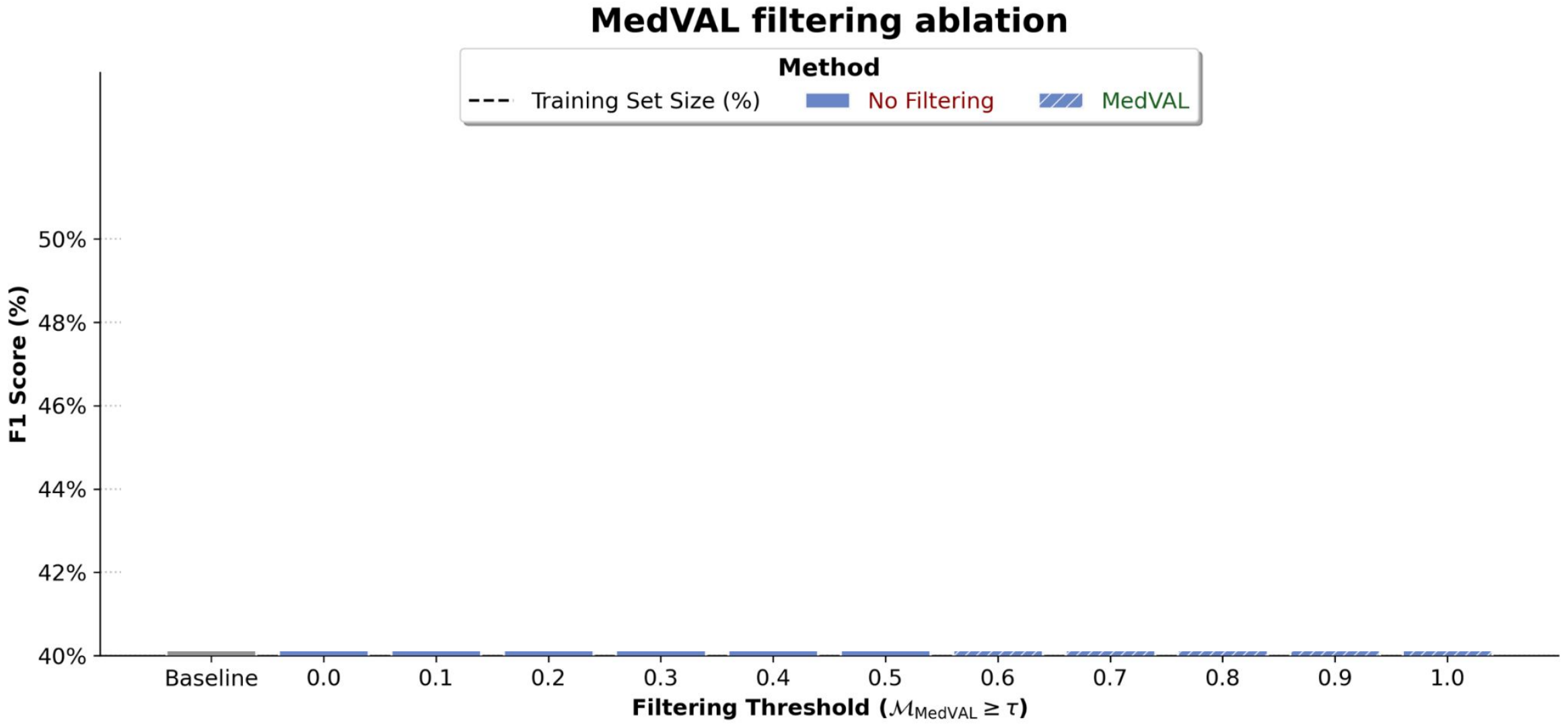
Filtering Ablation



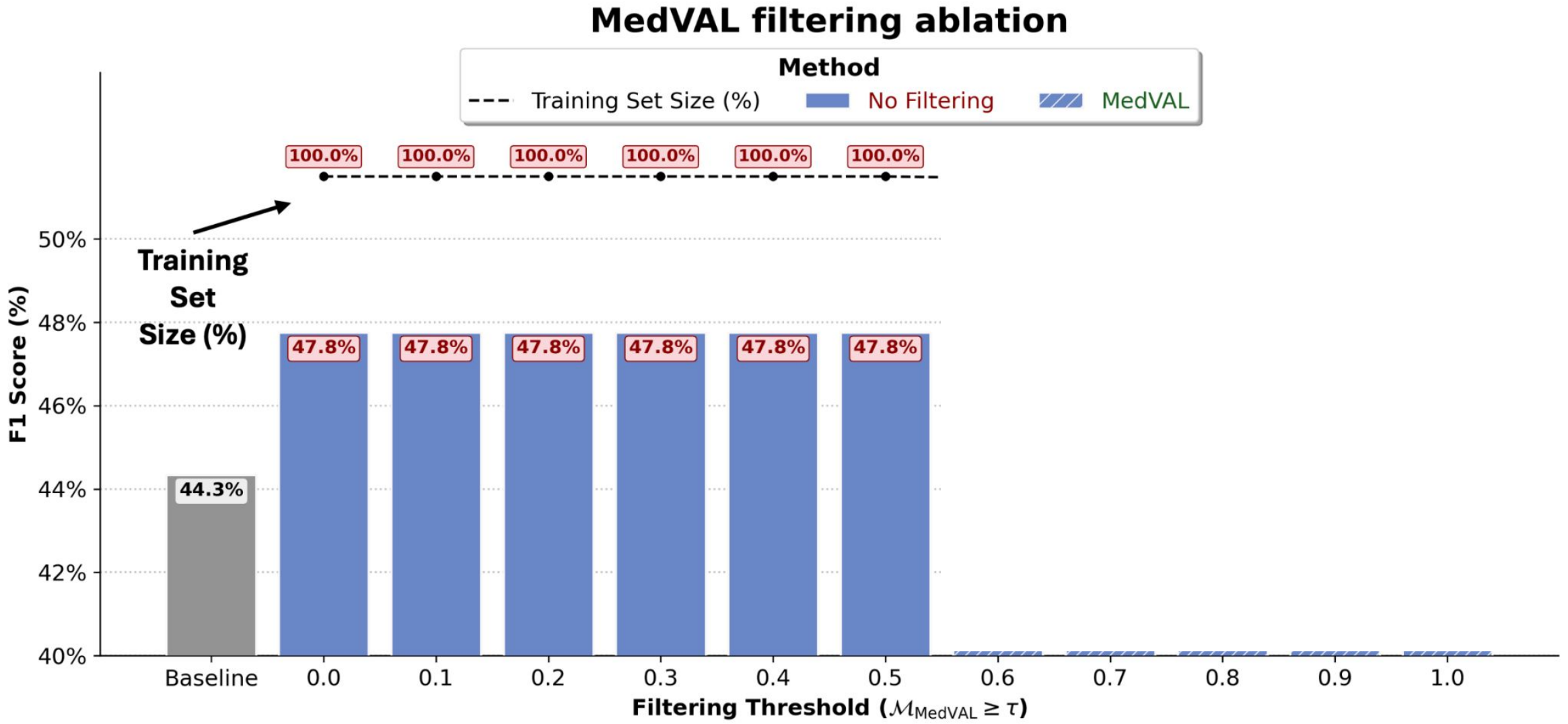
Filtering Ablation



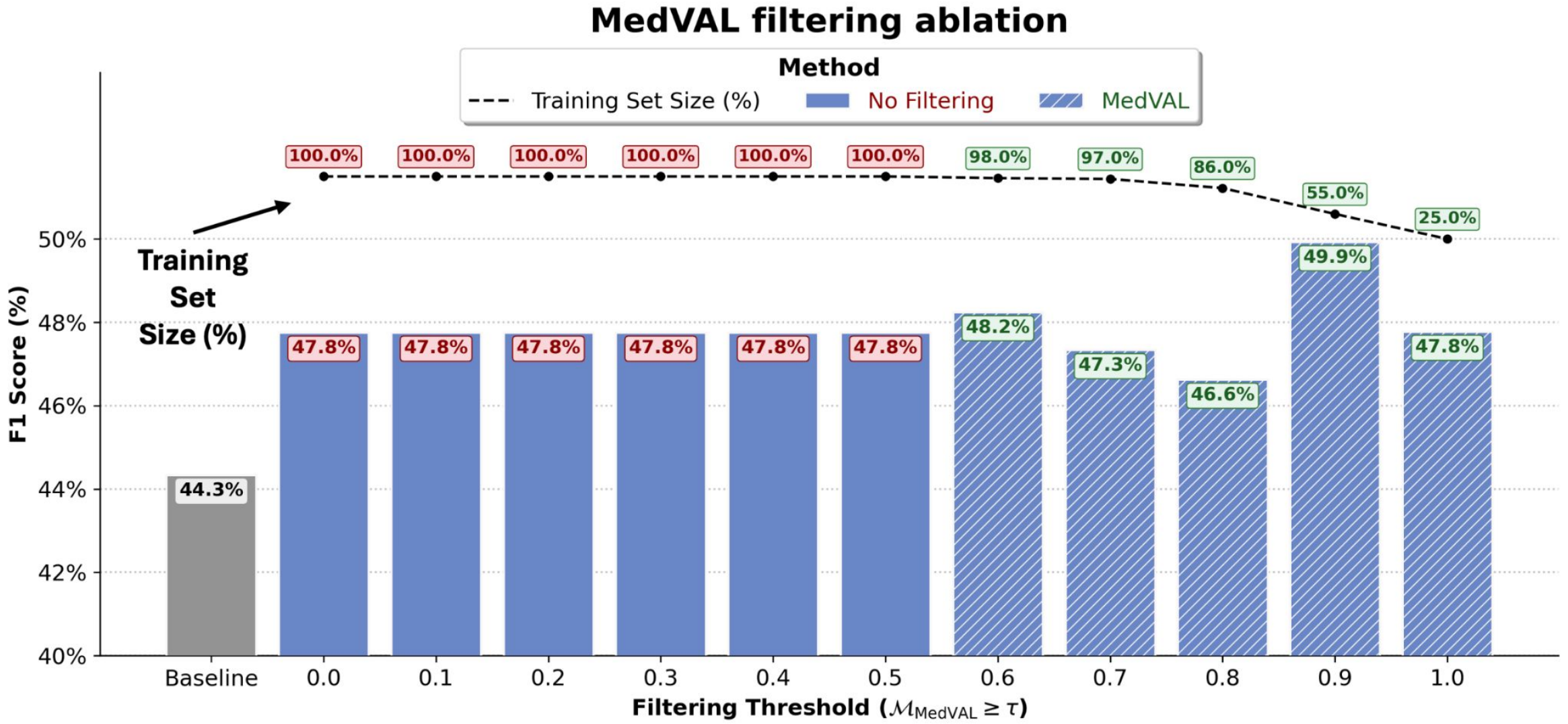
Filtering Ablation



Filtering Ablation



Filtering Ablation



Task-Wise Performance

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%

Task-Wise Performance

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

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

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.

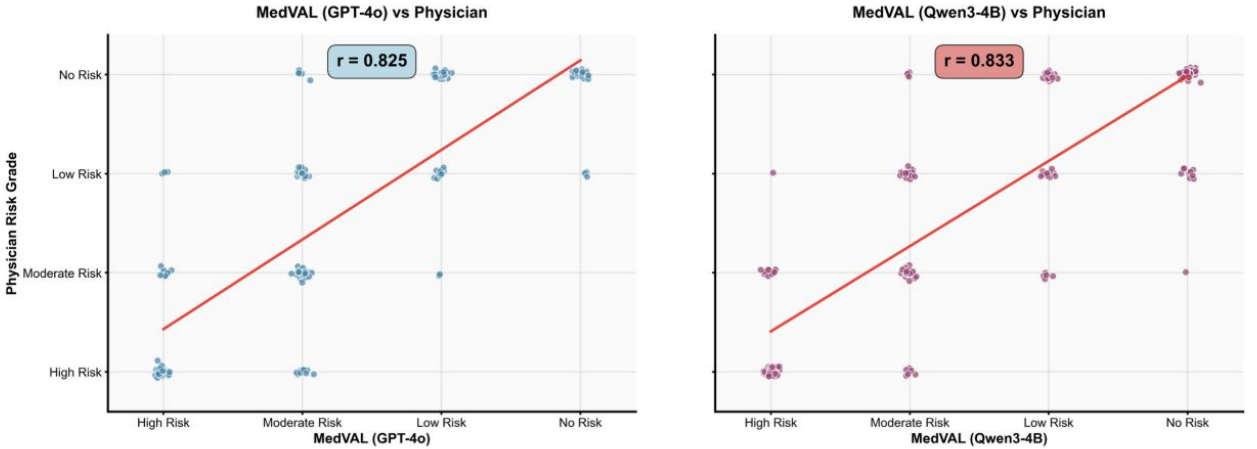
Color key:

Risk grade prediction:
Green: correct
Red: incorrect

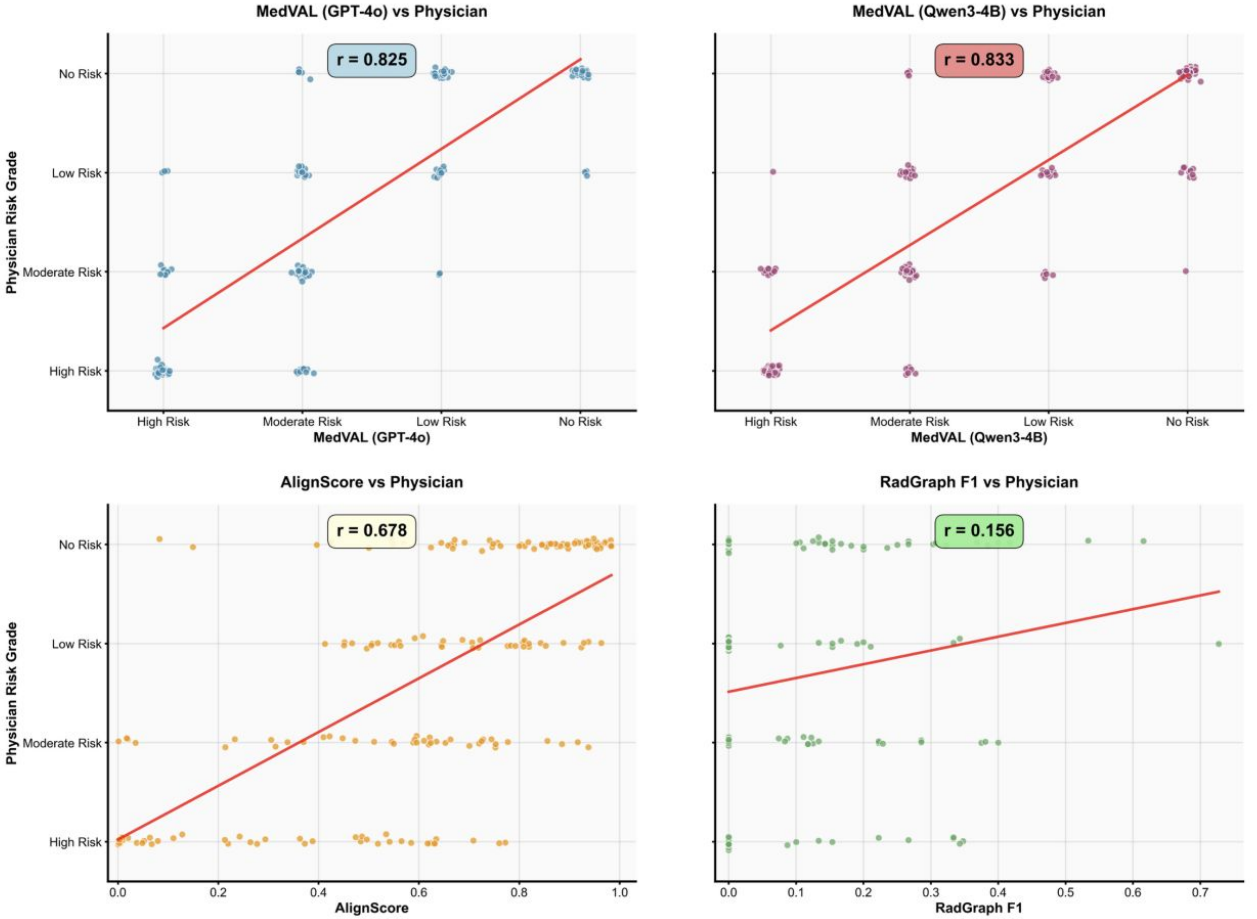
Error categories:
Purple: fabricated claim
Blue: incorrect recommendation
Brown: missing context
Orange: no errors

Comparison with other metrics

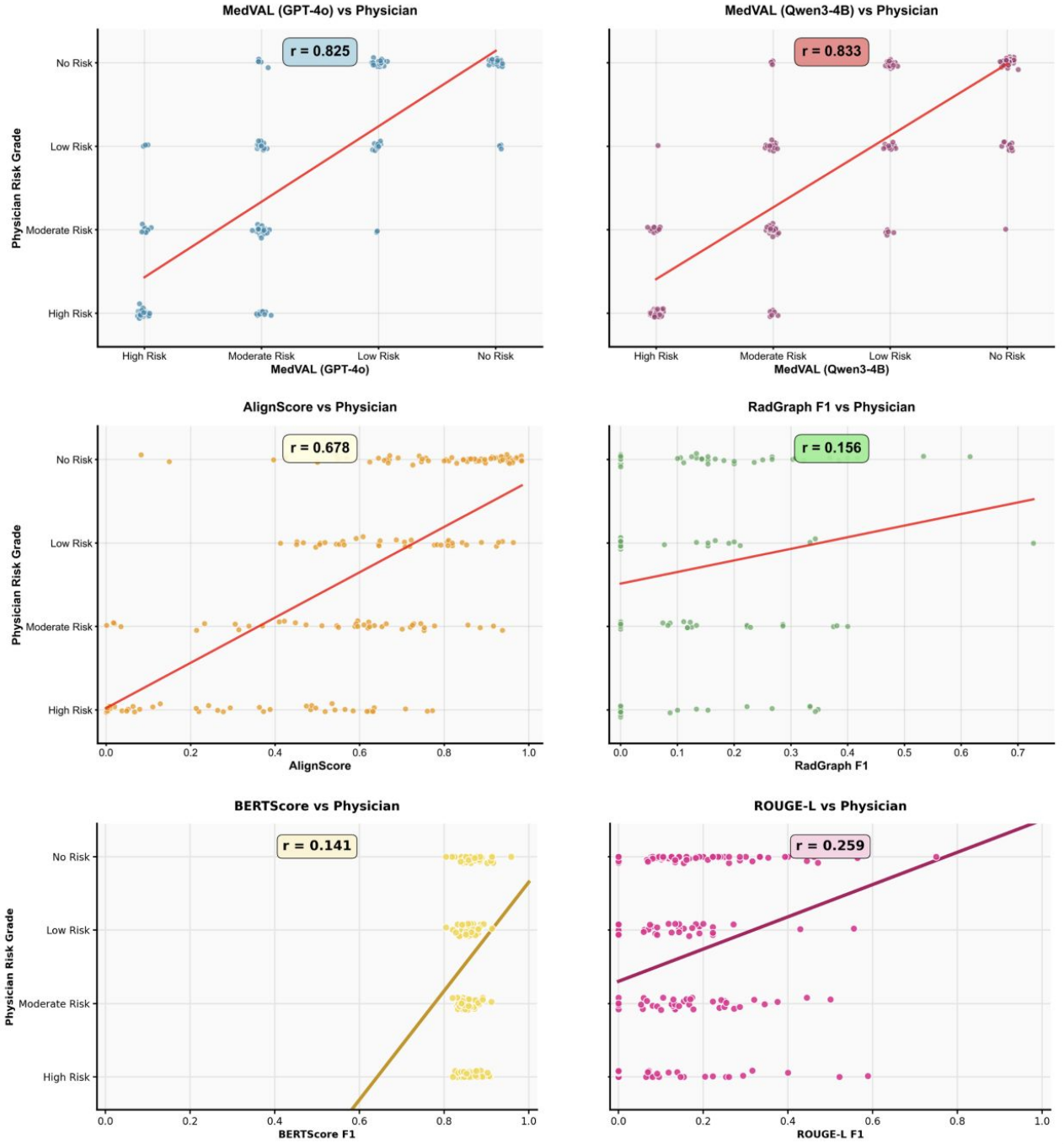
Comparison with other metrics



Comparison with other metrics



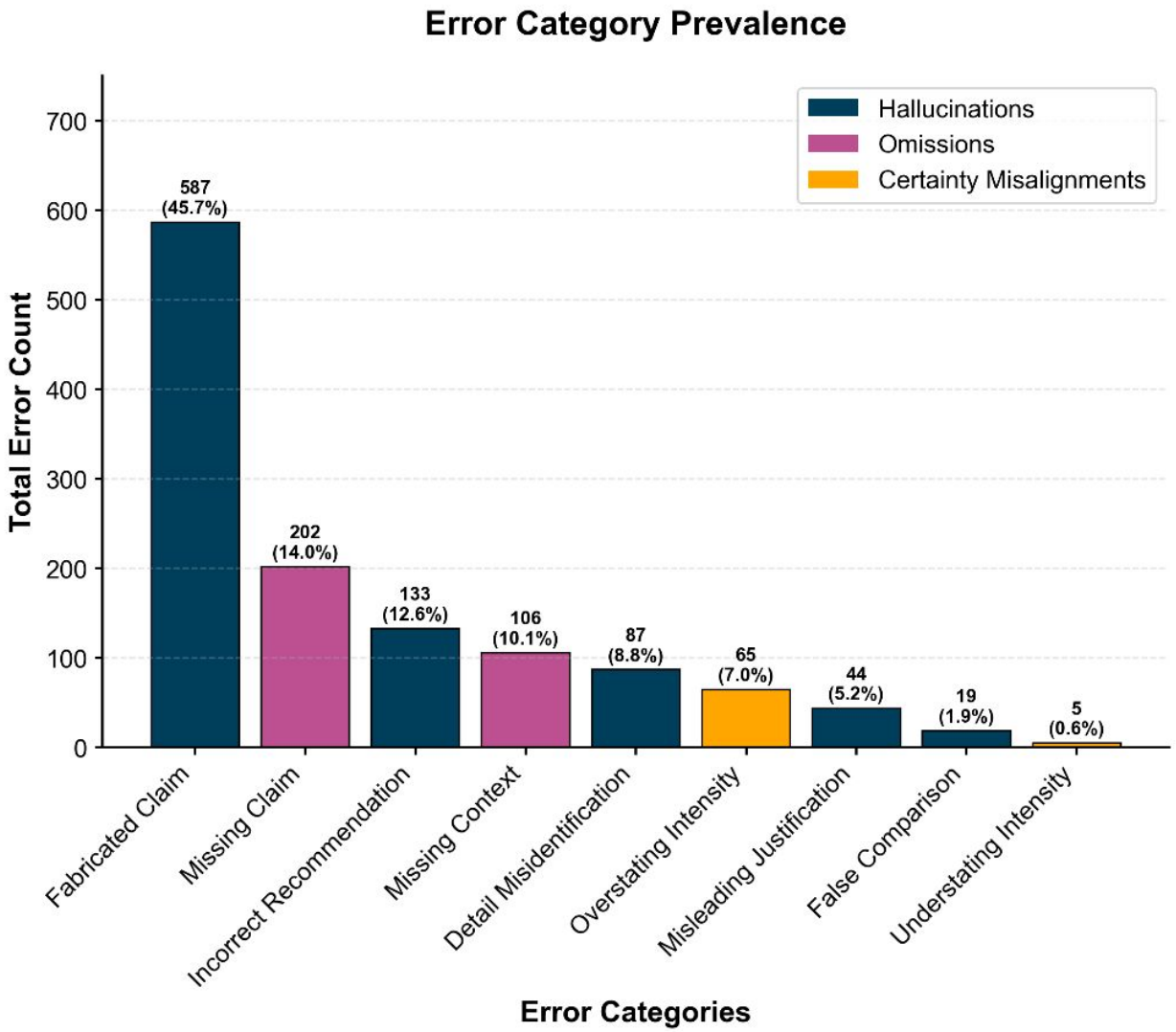
Comparison with other metrics



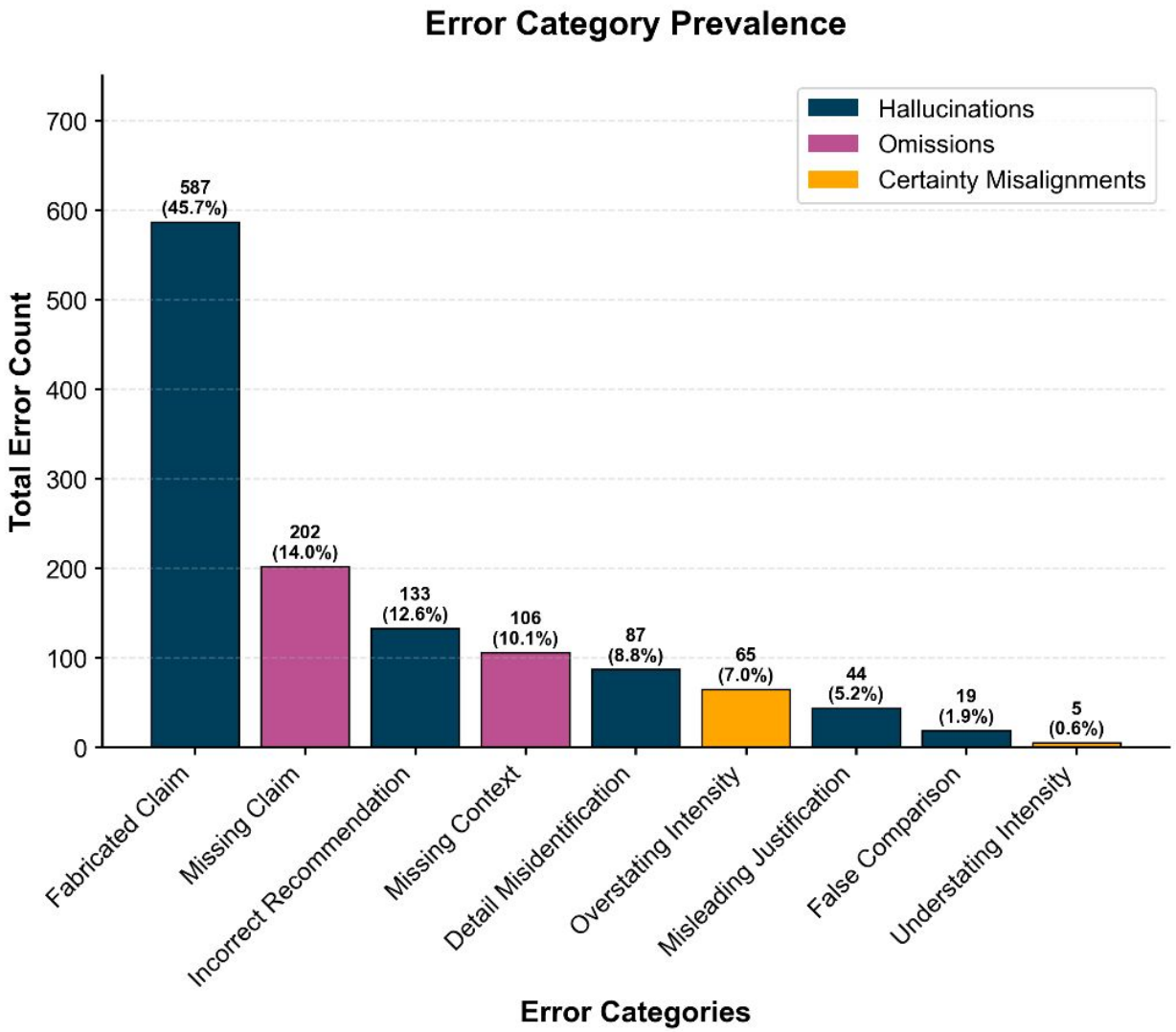
Stanford MEDICINE

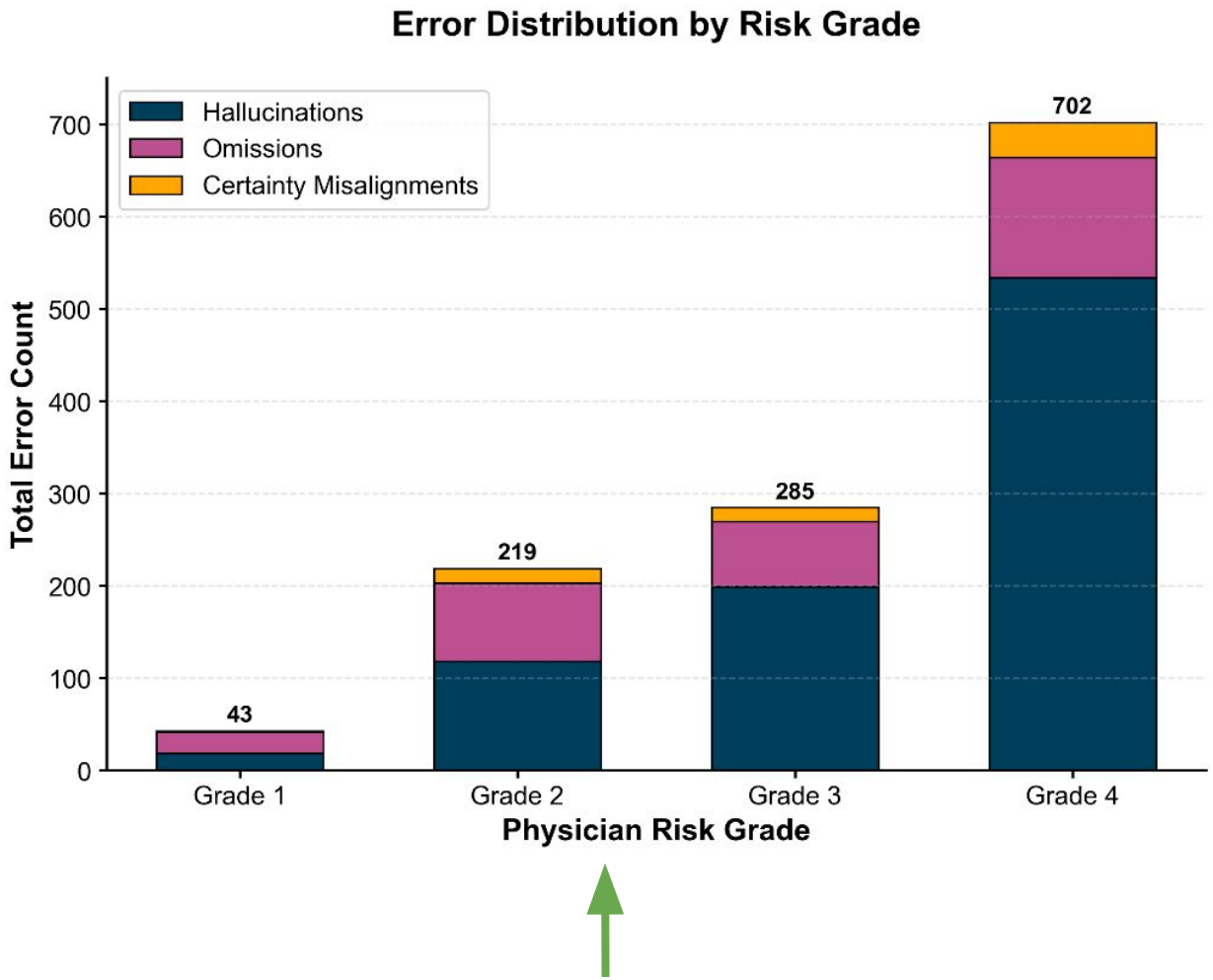
Error Distribution

Error Distribution



Error Distribution



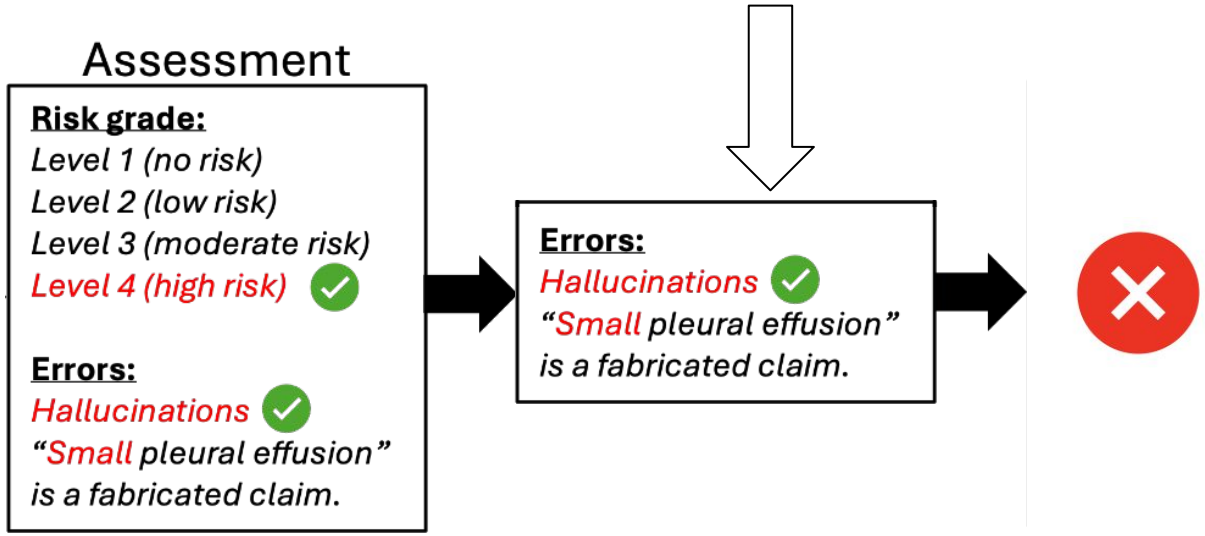


High correlation b/w error count and risk grade

HIGH ERROR COUNT -> HIGH RISK

Future Work - Beyond Risk Grade Classification

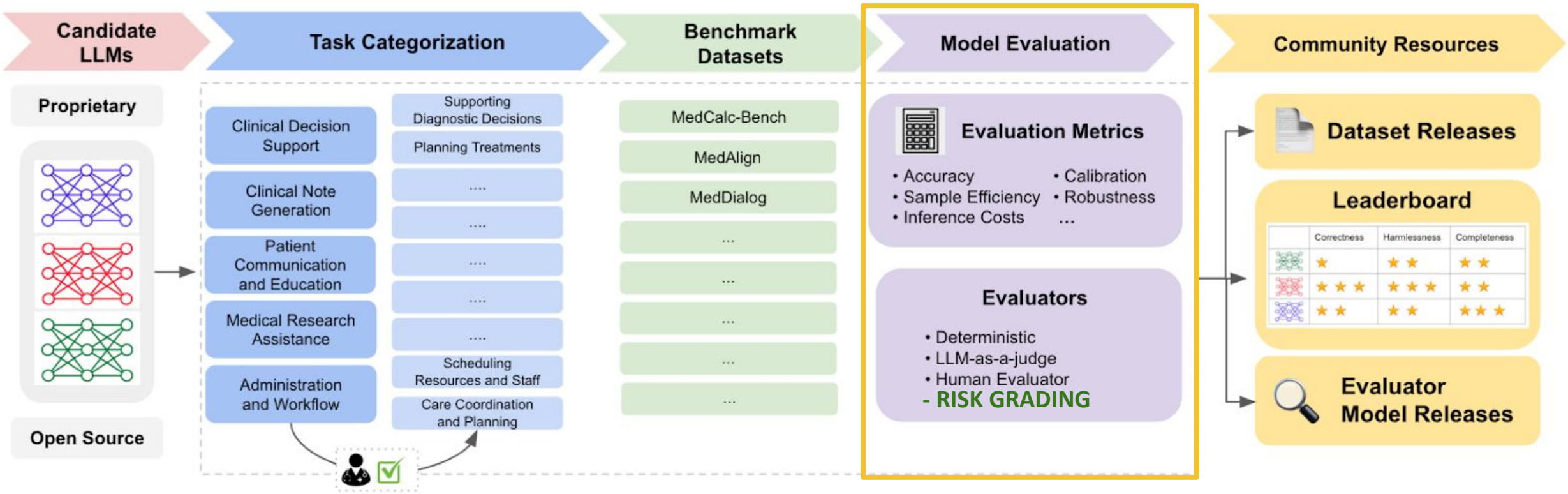
MedVAL only evaluates the risk classification, and currently ignores this field



Complete risk assessment requires answering:

1. What is the **"true" reasoning** behind a model's **risk grade prediction**?
2. Does that **reasoning** match the **physician's reasoning**?

Future Work - Integration with Medical Benchmarks



Conclusion

- We introduce MedVAL, a **self-supervised, data-efficient** distillation method for **validating 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**
- Language models can achieve a **performance statistically non-inferior to a single human expert** ($p < 0.001$).

Open-Source

- Paper: https://www.nature.com/articles/s41746-026-03084-5
- 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://www.nature.com/articles/s41746-026-03084-5
- 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

Paper

npj | digital medicine

View all journals | Saved research | Search | Log in

Explore content | About the journal | Publish with us | Sign up for alerts | RSS feed

nature > npj digital medicine > articles > article

Article | Open access | Published: 04 August 2026

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, ... Akshay S. Chaudhari + Show authors

npj Digital Medicine (2026) | Cite this article

Download PDF

Associated content

Collection

Evaluating the Real-World Clinical Performance of AI

Sections

Abstract

Accepted at NeurIPS GenAI4Health!

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 et al. (7 additional authors not shown)

Published: 12 Oct 2025, Last Modified: 28 Oct 2025 GenAI4Health 2025 Oral Everyone Revisions BibTeX CC BY 4.0 Edit

Rating: 10: Top 5% of accepted papers, seminal paper

Paper

Stanford MEDICINE

DSPy meets HELM
Structured Prompts Improve Evaluation of Language Models



Introduction

- LMs are increasingly adopted across domains, **accurate performance estimation is essential** for deployment decisions
- Frameworks like **Holistic Evaluation of Language Models (HELM) evaluate LMs under fixed prompts** that fail to generalize
- Unless we approximate the **maximum achievable by a LM via changes to the prompt**, we risk underestimating performance
- **DSPy offers a scalable alternative to manual prompt engineering** by crafting structured prompts that can be optimized per task
- We present a reproducible ***DSPy+HELM*** framework that enables more accurate LM benchmarking:
  - First study to **systematically integrate structured prompting into an established evaluation framework**
  - We demonstrate how **performance-ceiling approximation yields more decision-useful benchmarks**
- We find that without structured prompting:
  - HELM **underestimates** LM performance
  - Performance **estimates vary more** across benchmarks
  - Performance gaps are misrepresented (leaderboard **rankings flip**)
  - **Chain-of-thought reduces sensitivity** to prompt design

HELM

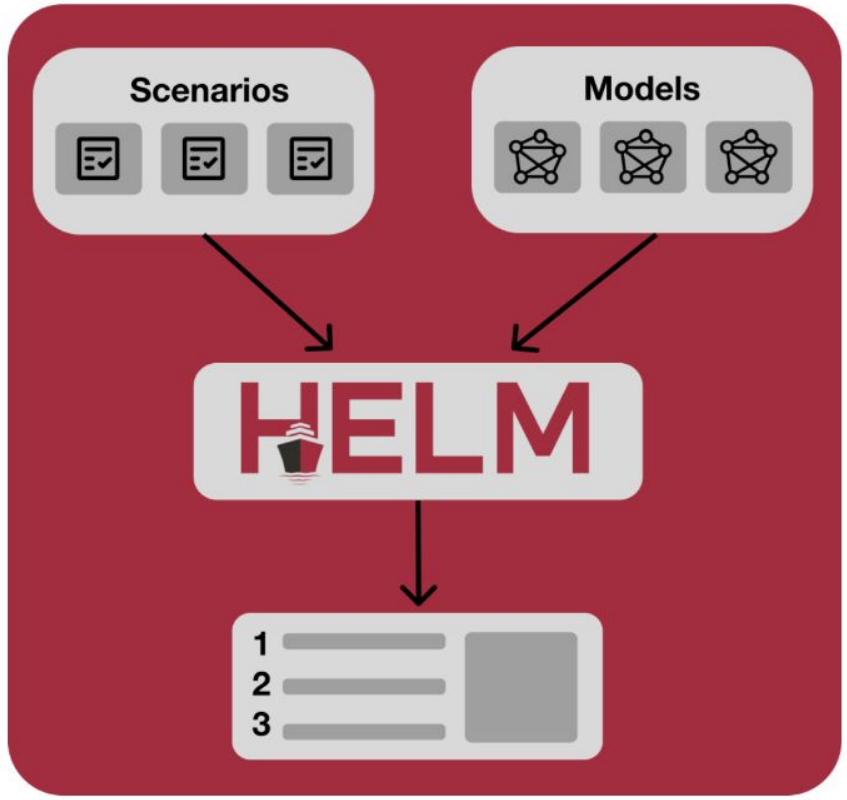
"Holistic Evaluation of Language Models"

Center for Research on Foundation Models HELM All Leaderboards

A reproducible and transparent framework for evaluating foundation models.

Find leaderboards with many scenarios, metrics, and models with support for multimodality and model-graded evaluation.

LEADERBOARDS ↓ GITHUB



Center for Research on Foundation Models

HELM All Leaderboards

Capabilities: A new leaderboard for evaluating general capabilities of language models

Audio: Holistic Evaluation of Audio-Language Models

Lite: Lightweight, broad evaluation of the capabilities of language models using in-context learning

Classic: Thorough language model evaluations based on the scenarios from the original HELM paper

HEIM: Holistic evaluation of text-to-image models

Instruct: Evaluations of instruction following models with absolute ratings

MMLU: Massive Multitask Language Understanding (MMLU) evaluations using standardized prompts

VHELM: Holistic Evaluation of Vision-Language Models

Image2Struct: Evaluations of Vision-Language Models on extracting structured information from images

AIR-Bench: Safety benchmark based on emerging government regulations and company policies

Safety: Safety benchmark that aggregates popular safety benchmarks across 6 risk vectors

CLEVA: Chinese-language benchmark for holistic evaluation of Chinese language models

ThaiExam: Thai-language evaluations of language models on standardized examinations in Thailand

SEA-HELM: Assessment of large language models across various tasks, emphasizing Southeast Asian languages

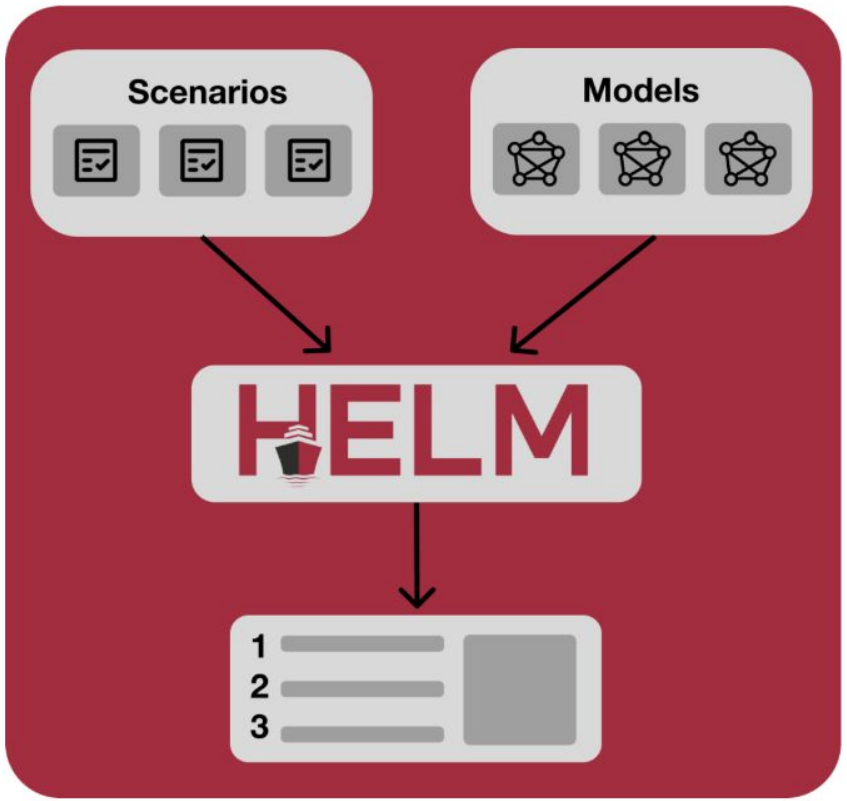
MMLU-Winogrande-Afr: Clinical MMLU and Winogrande in 11 low-resource African languages

ToRR: A benchmark for table reasoning and robustness

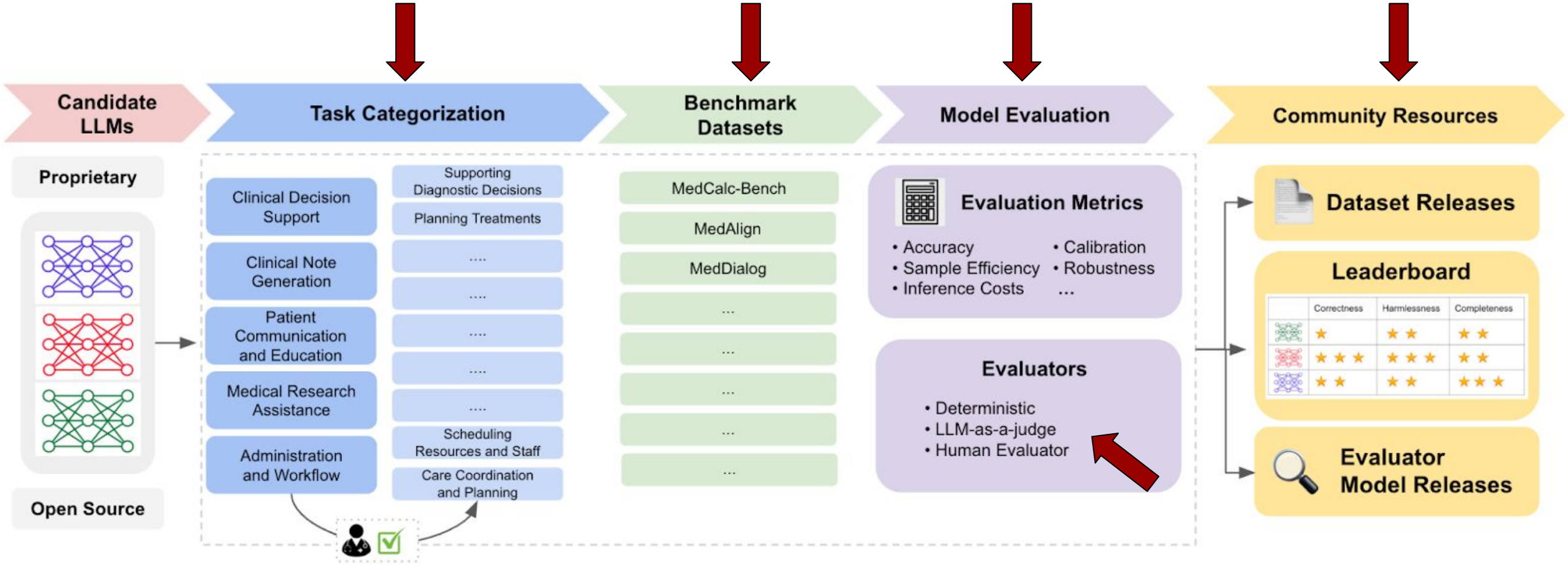
Finance: Financial-domain benchmark using real financial documents

MedHELM: A benchmark by medical experts for LLMs grounded in real-world healthcare needs

Long Context: A benchmark of LLM long context capabilities



MedHELM


Candidate LLMs
Task Categorization
Benchmark Datasets
Model Evaluation
Community Resources
Proprietary
Open Source
Clinical Decision Support
Clinical Note Generation
Patient Communication and Education
Medical Research Assistance
Administration and Workflow
Supporting Diagnostic Decisions
Planning Treatments
....
....
....
....
....
Scheduling Resources and Staff
Care Coordination and Planning
MedCalc-Bench
MedAlign
MedDialog
...
...
...
...
...
Evaluation Metrics
• Accuracy
• Calibration
• Sample Efficiency
• Robustness
• Inference Costs
...
Evaluators
• Deterministic
• LLM-as-a-judge
• Human Evaluator
Dataset Releases
Leaderboard
Correctness
Harmlessness
Completeness
Evaluator Model Releases

Leaderboard: Core Scenarios

Core Scenarios

Select a group: Core Scenarios

Accuracy Effici General information

Model	Mean score	MMLU-Pro - COT correct	GPQA - COT correct	IFEval - IFEval Strict Acc	WildBench - WB Score	Omni-MATH - Acc
GPT-5 mini (2025-08-07)	0.819	0.835	0.756	0.927	0.855	**0.722**
o4-mini (2025-04-16)	0.812	0.82	0.735	0.929	0.854	0.72
o3 (2025-04-16)	0.811	0.859	0.753	0.869	0.861	0.714
GPT-5 (2025-08-07)	0.807	0.863	0.791	0.875	0.857	0.647
Gemini 3 Pro (Preview)	0.799	**0.903**	**0.803**	0.876	0.859	0.556
Qwen3 235B A22B Instruct 2507 FP8	0.798	0.844	0.726	0.835	**0.866**	0.718
Grok 4 (0709)	0.785	0.851	0.726	0.949	0.797	0.603
Claude 4 Opus (20250514, extended thinking)	0.78	0.875	0.709	0.849	0.852	0.616
gpt-oss-120b	0.77	0.795	0.684	0.836	0.845	0.688
Kimi K2 Instruct	0.768	0.819	0.652	0.85	0.862	0.654

DSPy

“*A framework for programming language models*”

Prompt 1: HELM Baseline

Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————————————

Prompt 2: Zero-Shot CoT

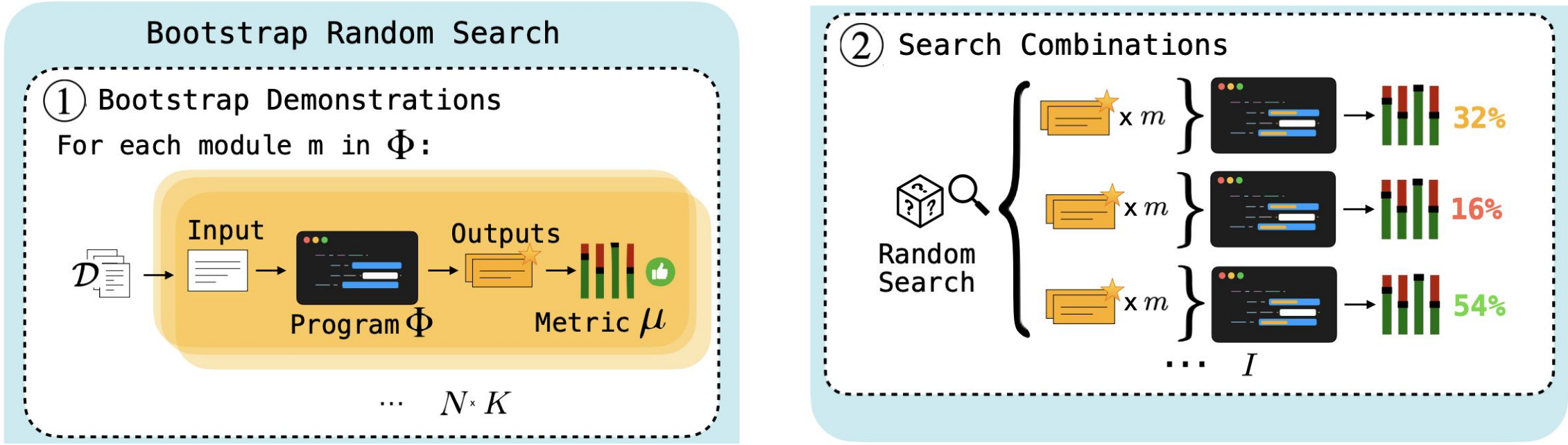
Your input fields are: "INPUTS"
Your output fields are: "REASONING" and "OUTPUT"
Your objective is: Given the fields "INPUTS", produce the fields "OUTPUT"

INPUTS:

Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————————————

Respond with the corresponding output fields, starting with "REASONING", then "OUTPUT".

Prompt Optimization - Algorithm 1



Bootstrap Few Shot with Random Search

Prompt 1: HELM Baseline

Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————————————

Prompt 3: BFRS (Few-Shot Optimized)

Your input fields are: "INPUTS"
Your output fields are: "REASONING" and "OUTPUT"
Your objective is: Given the fields "INPUTS", produce the fields "OUTPUT"

IN-CONTEXT EXAMPLES (K Demos):

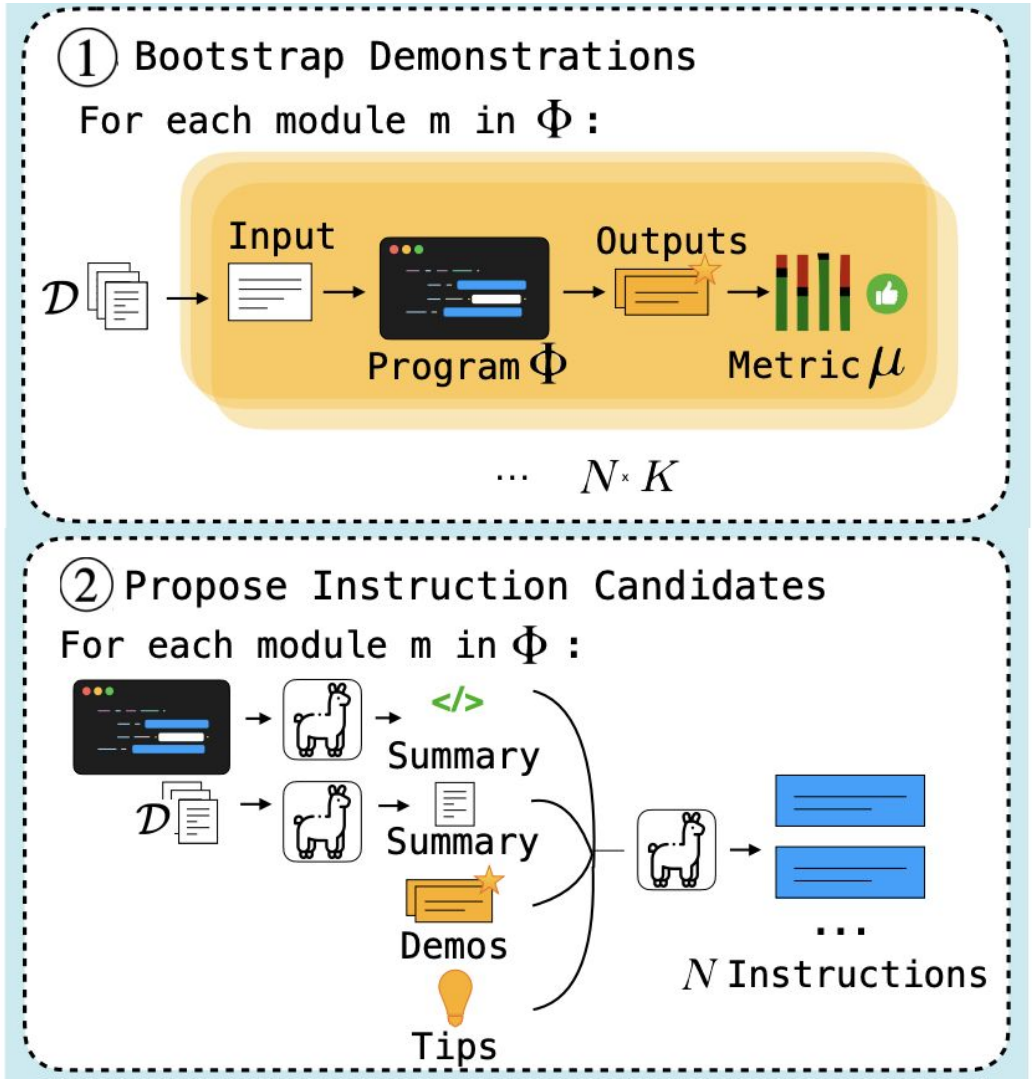
INPUTS: <input text> → REASONING: <steps>, OUTPUT: <output text>

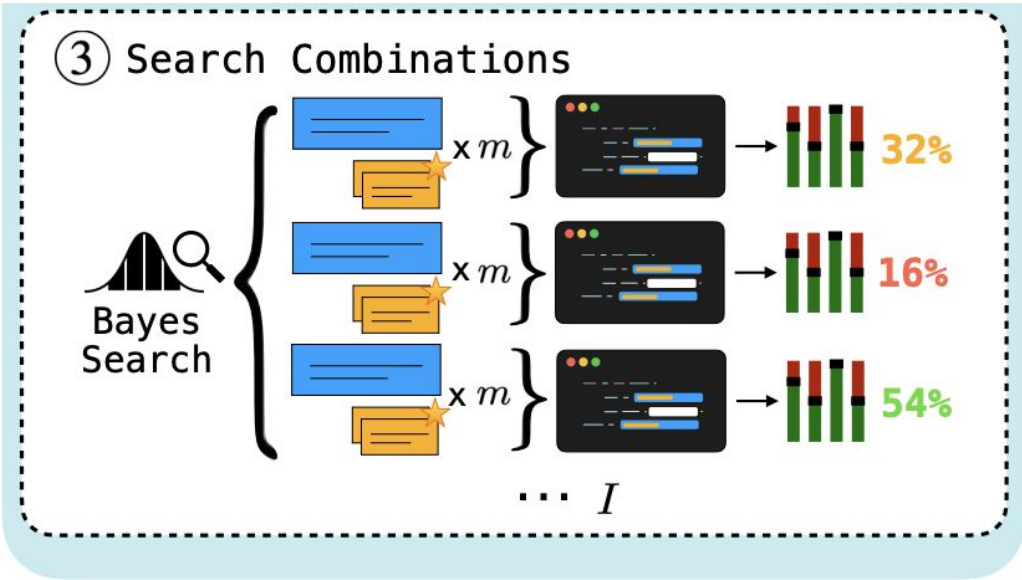
INPUTS:

Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————————————

Respond with the corresponding output fields, starting with "REASONING", then "OUTPUT".

Prompt Optimization - Algorithm 2





MIPROv2

Prompt 1: HELM Baseline

Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————————————

Prompt 4: MIPROv2 (Instruction + Few-Shot Optimized)

Your input fields are: "INPUTS"
Your output fields are: "REASONING" and "OUTPUT"
Your objective is: You are a highly skilled medical expert working in a busy emergency room. A patient presents with a complex medical history and concerning symptoms. The attending physician needs your immediate assistance in calculating a critical risk score to guide treatment decisions. The patient's life may depend on your accuracy.

IN-CONTEXT EXAMPLES (K Demos):

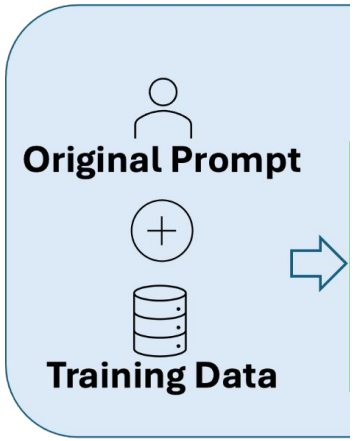
INPUTS: <input text> → REASONING: <steps>, OUTPUT: <output text>

INPUTS:

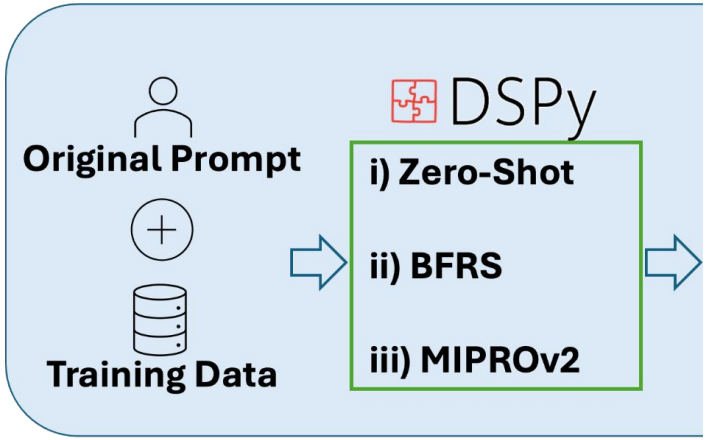
Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————————————

Respond with the corresponding output fields, starting with "REASONING", then "OUTPUT".

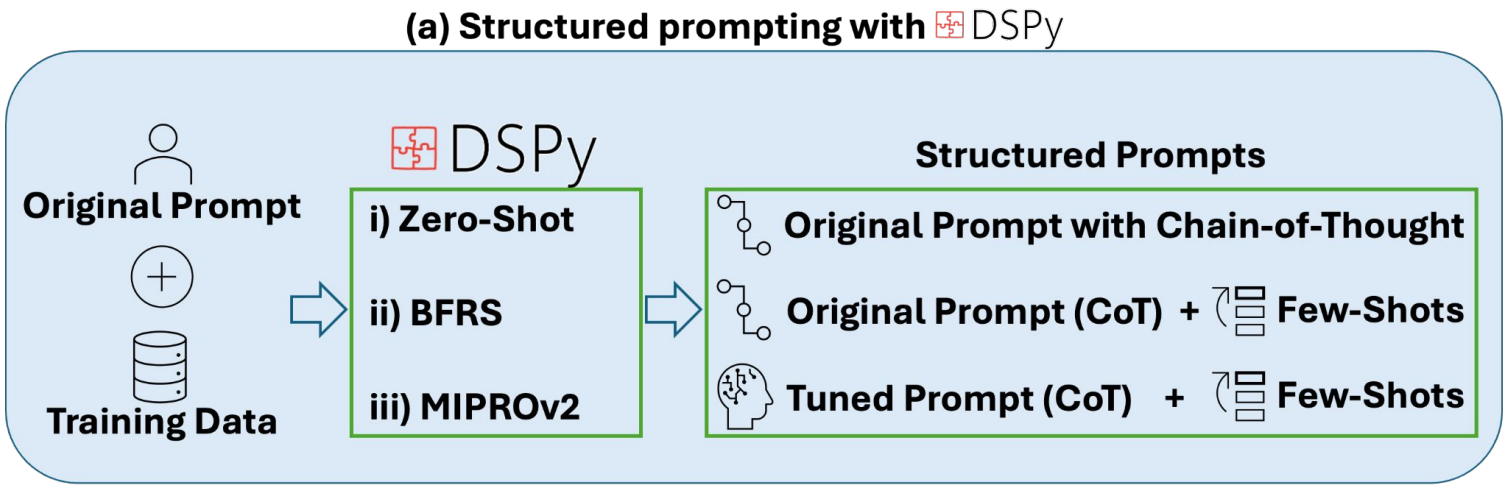
DSPy-HELM



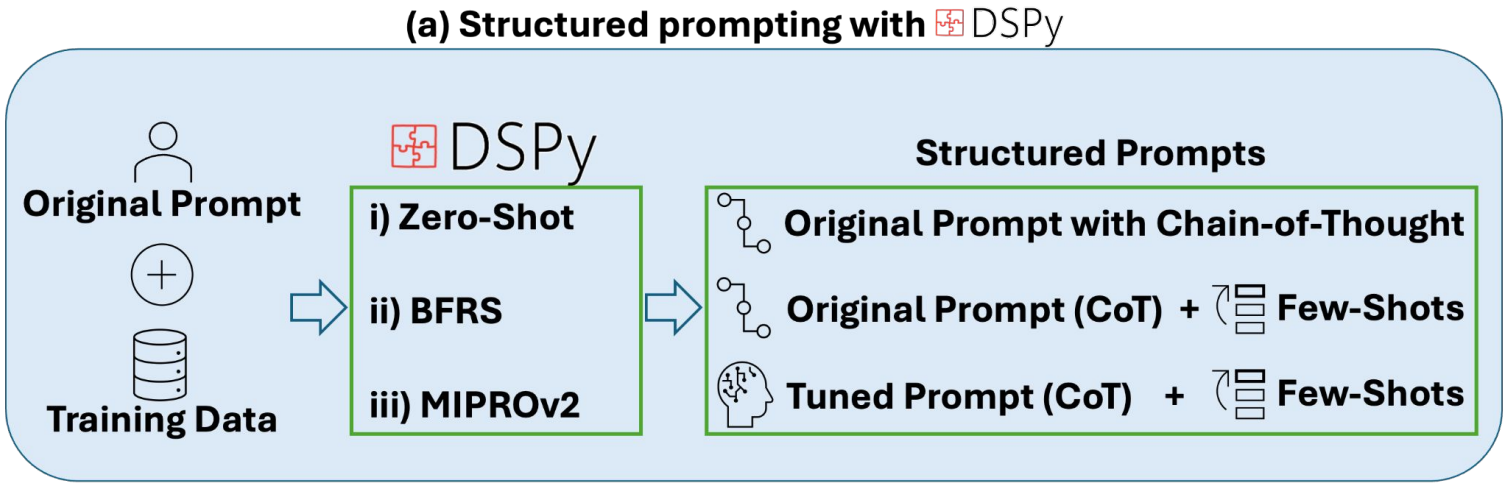
DSPy-HELM

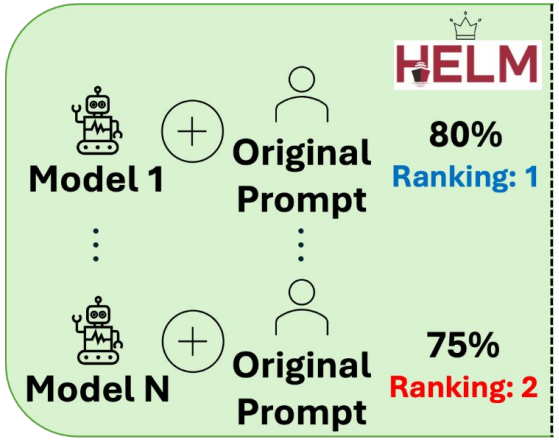

DSPy
Original Prompt
Training Data
i) Zero-Shot
ii) BFRS
iii) MIPROv2

DSPy-HELM

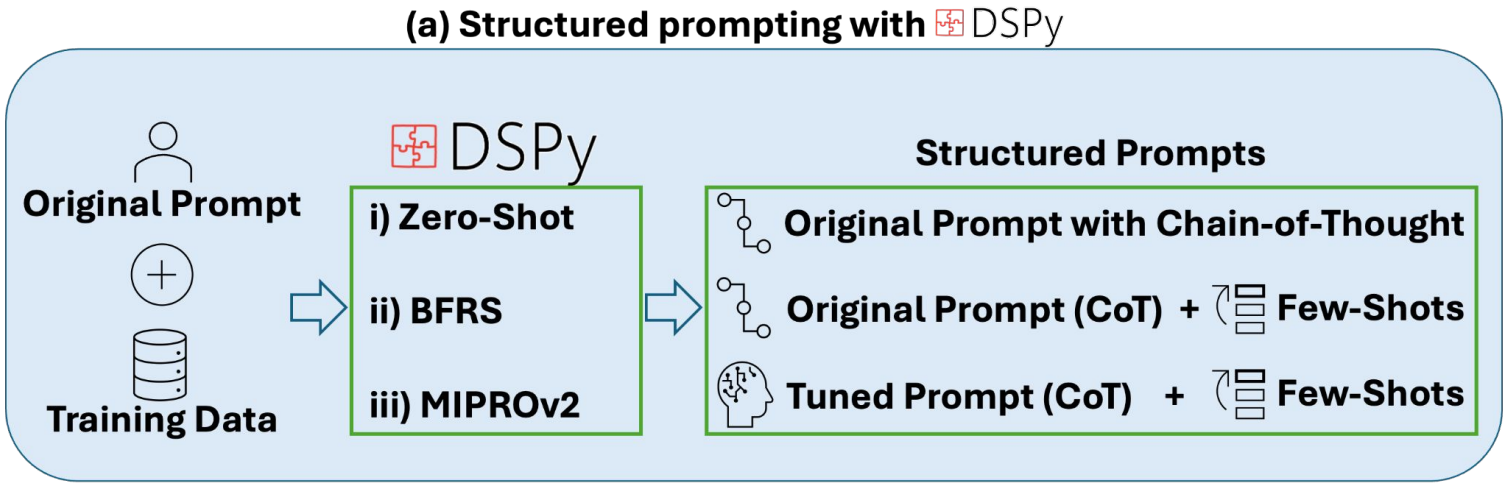

(a) Structured prompting with DSPy
Original Prompt
Training Data
DSPy
i) Zero-Shot
ii) BFRS
iii) MIPROv2
Structured Prompts
Original Prompt with Chain-of-Thought
Original Prompt (CoT) + Few-Shots
Tuned Prompt (CoT) + Few-Shots

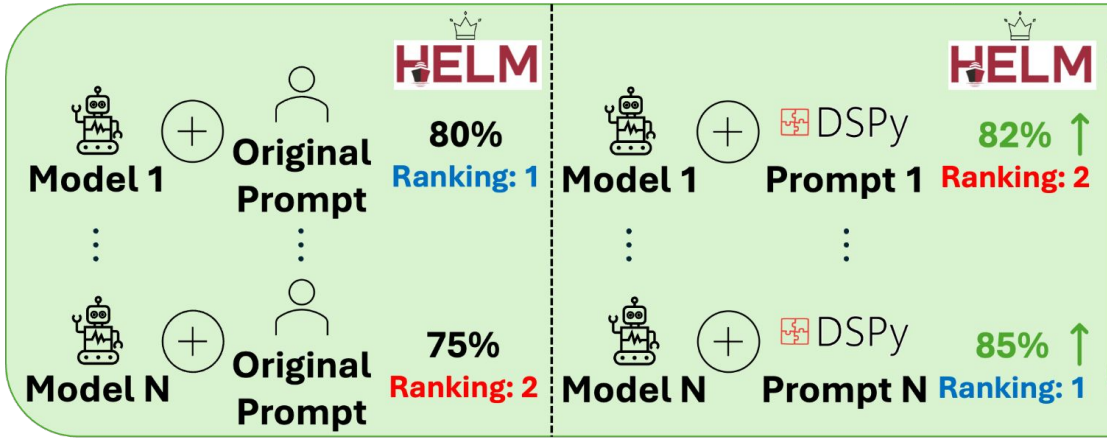
DSPy-HELM


(a) Structured prompting with DSPy
Original Prompt
Training Data
DSPy
i) Zero-Shot
ii) BFRS
iii) MIPROv2
Structured Prompts
Original Prompt with Chain-of-Thought
Original Prompt (CoT) + Few-Shots
Tuned Prompt (CoT) + Few-Shots

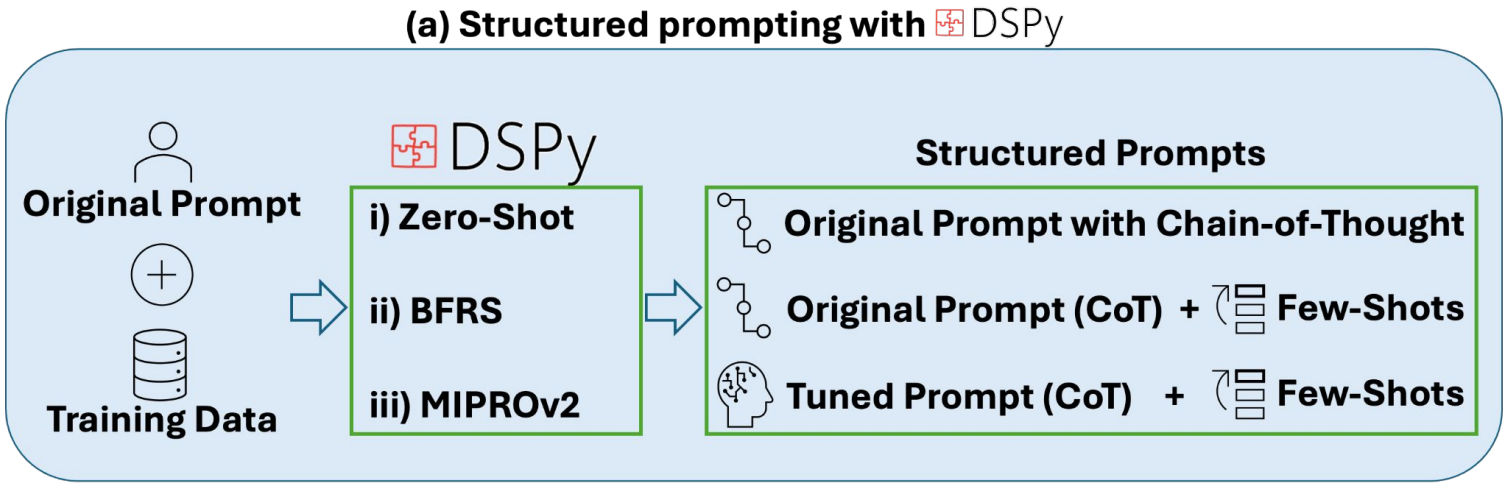

HELM
Model 1
Original Prompt
80%
Ranking: 1
Model N
Original Prompt
75%
Ranking: 2

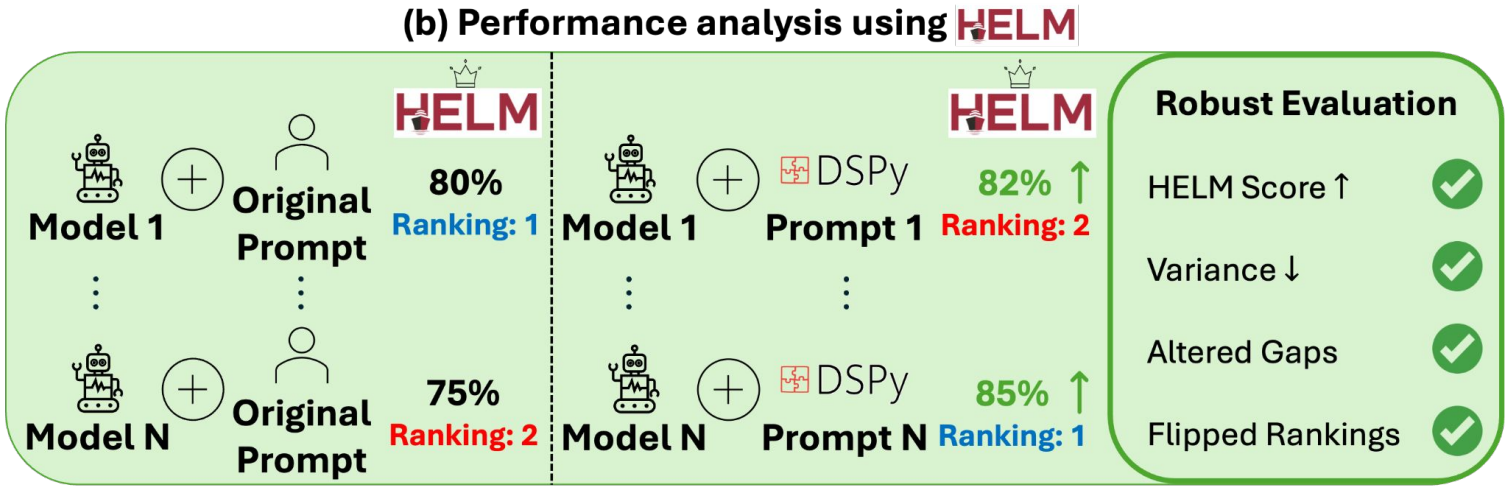
DSPy-HELM


(a) Structured prompting with DSPy
Original Prompt
Training Data
DSPy
i) Zero-Shot
ii) BFRS
iii) MIPROv2
Structured Prompts
Original Prompt with Chain-of-Thought
Original Prompt (CoT) + Few-Shots
Tuned Prompt (CoT) + Few-Shots


HELM
Model 1
Original Prompt
80%
Ranking: 1
Model N
Original Prompt
75%
Ranking: 2
HELM
Model 1
DSPy Prompt 1
82%
Ranking: 2
Model N
DSPy Prompt N
85%
Ranking: 1

DSPy-HELM


(a) Structured prompting with DSPy
Original Prompt
Training Data
DSPy
i) Zero-Shot
ii) BFRS
iii) MIPROv2
Structured Prompts
Original Prompt with Chain-of-Thought
Original Prompt (CoT) + Few-Shots
Tuned Prompt (CoT) + Few-Shots


(b) Performance analysis using HELM
HELM
Model 1
Original Prompt
80%
Ranking: 1
Model N
Original Prompt
75%
Ranking: 2
HELM
Model 1
DSPy Prompt 1
82%
Ranking: 2
Model N
DSPy Prompt N
85%
Ranking: 1
Robust Evaluation
HELM Score ↑
Variance ↓
Altered Gaps
Flipped Rankings

Benchmarks

Benchmark	Input → Output	Task	Samples
MMLU-Pro	Reasoning Question → Answer	Multi-Task Reasoning	1,000
GPQA	Graduate Question → Answer	Graduate-Level QA	446
GSM8K	Math Problem → Solution	Numeric Problem-Solving	1,000
MedCalc-Bench	Patient Note → Computed Value	Computational Reasoning	1,000
Medec	Medical Narrative → Errors	Error Classification	597
HeadQA	Medical Question → Answer	USMLE-Style QA	1,000
MedBullets	Medical Question → Answer	USMLE-Style QA	308

Overall Results

Prompting Method	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT 4o	o3 Mini	Llama 3.3 70B	Qwen3 4B
HELM Baseline	64.8% ± 3.3	61.4% ± 3.4	61.0% ± 3.3	70.9% ± 3.3	57.2% ± 3.4	47.8% ± 3.3

Overall Results

Prompting Method	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT 4o	o3 Mini	Llama 3.3 70B	Qwen3 4B
HELM Baseline	64.8% ± 3.3	61.4% ± 3.4	61.0% ± 3.3	70.9% ± 3.3	57.2% ± 3.4	47.8% ± 3.3
Zero-Shot Predict	65.1% ± 3.4	61.7% ± 3.4	59.7% ± 3.3	**73.2% ± 3.2**	59.2% ± 3.4	49.3% ± 3.4

Overall Results

Prompting Method	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT 4o	o3 Mini	Llama 3.3 70B	Qwen3 4B
HELM Baseline	64.8% ± 3.3	61.4% ± 3.4	61.0% ± 3.3	70.9% ± 3.3	57.2% ± 3.4	47.8% ± 3.3
Zero-Shot Predict	65.1% ± 3.4	61.7% ± 3.4	59.7% ± 3.3	**73.2% ± 3.2**	59.2% ± 3.4	49.3% ± 3.4
Zero-Shot CoT ★	69.4% ± 3.2	**66.2% ± 3.4**	65.7% ± 3.3	72.7% ± 3.2	64.0% ± 3.4	59.0% ± 3.4

Overall Results

Prompting Method	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT 4o	o3 Mini	Llama 3.3 70B	Qwen3 4B
HELM Baseline	64.8% ± 3.3	61.4% ± 3.4	61.0% ± 3.3	70.9% ± 3.3	57.2% ± 3.4	47.8% ± 3.3
Zero-Shot Predict	65.1% ± 3.4	61.7% ± 3.4	59.7% ± 3.3	**73.2% ± 3.2**	59.2% ± 3.4	49.3% ± 3.4
Zero-Shot CoT ★	69.4% ± 3.2	**66.2% ± 3.4**	65.7% ± 3.3	72.7% ± 3.2	64.0% ± 3.4	59.0% ± 3.4
BFRS ★	69.3% ± 3.3	**66.2% ± 3.3**	**65.9% ± 3.2**	73.1% ± 3.2	**64.2% ± 3.3**	60.3% ± 3.4
MIPROv2 ★	**69.8% ± 3.3**	**66.2% ± 3.3**	65.3% ± 3.3	73.1% ± 3.1	62.1% ± 3.4	**60.4% ± 3.4**

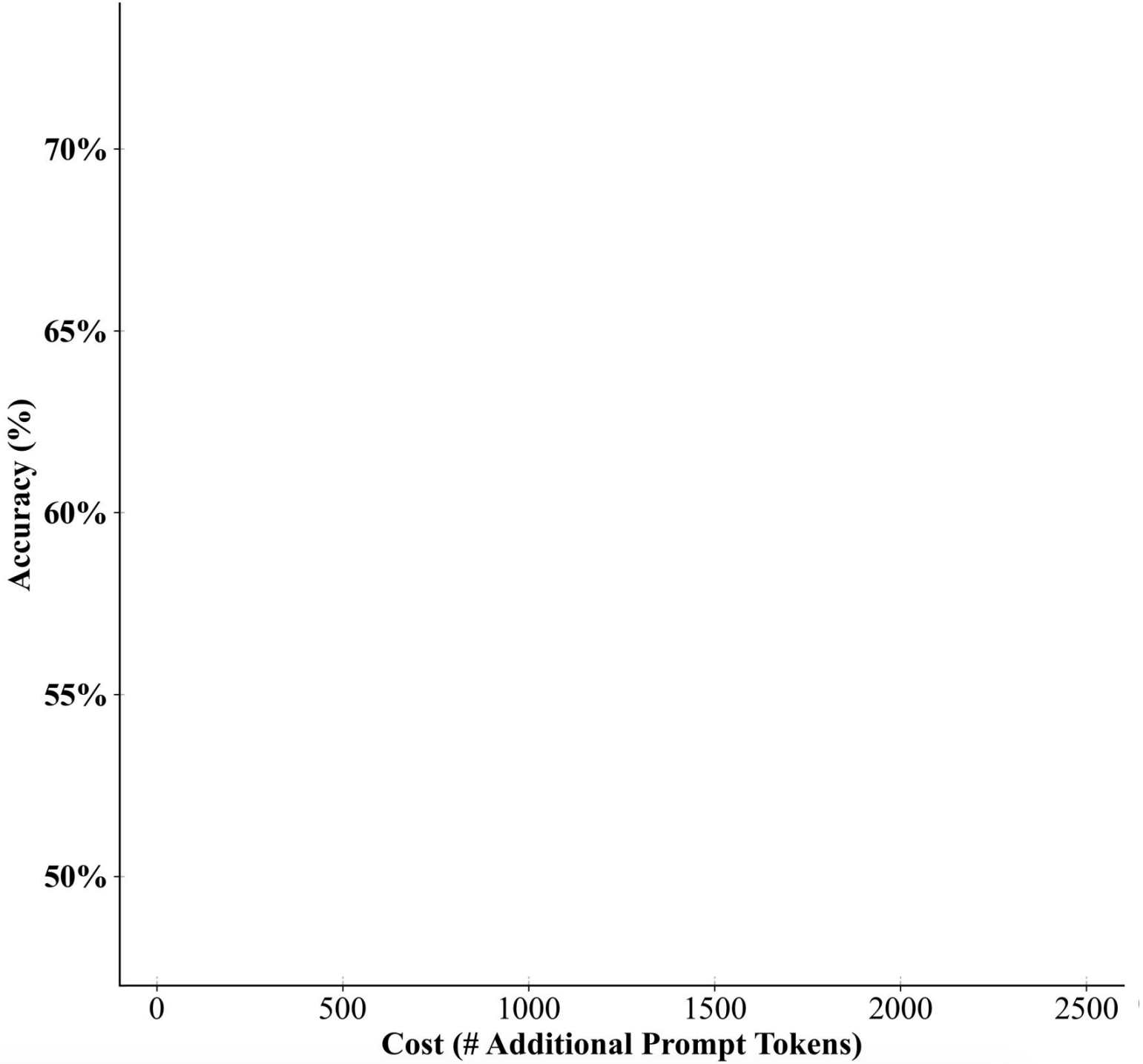
Overall Results

Prompting Method	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT 4o	o3 Mini	Llama 3.3 70B	Qwen3 4B
HELM Baseline	64.8% ± 3.3	61.4% ± 3.4	61.0% ± 3.3	70.9% ± 3.3	57.2% ± 3.4	47.8% ± 3.3
Zero-Shot Predict	65.1% ± 3.4	61.7% ± 3.4	59.7% ± 3.3	**73.2% ± 3.2**	59.2% ± 3.4	49.3% ± 3.4
Zero-Shot CoT ★	69.4% ± 3.2	**66.2% ± 3.4**	65.7% ± 3.3	72.7% ± 3.2	64.0% ± 3.4	59.0% ± 3.4
BFRS ★	69.3% ± 3.3	**66.2% ± 3.3**	**65.9% ± 3.2**	73.1% ± 3.2	**64.2% ± 3.3**	60.3% ± 3.4
MIPROv2 ★	**69.8% ± 3.3**	**66.2% ± 3.3**	65.3% ± 3.3	73.1% ± 3.1	62.1% ± 3.4	**60.4% ± 3.4**
Highest − Baseline ★	+5.0%	+4.8%	+4.9%	+2.3%	+7.0%	+12.6%

Accuracy vs Cost

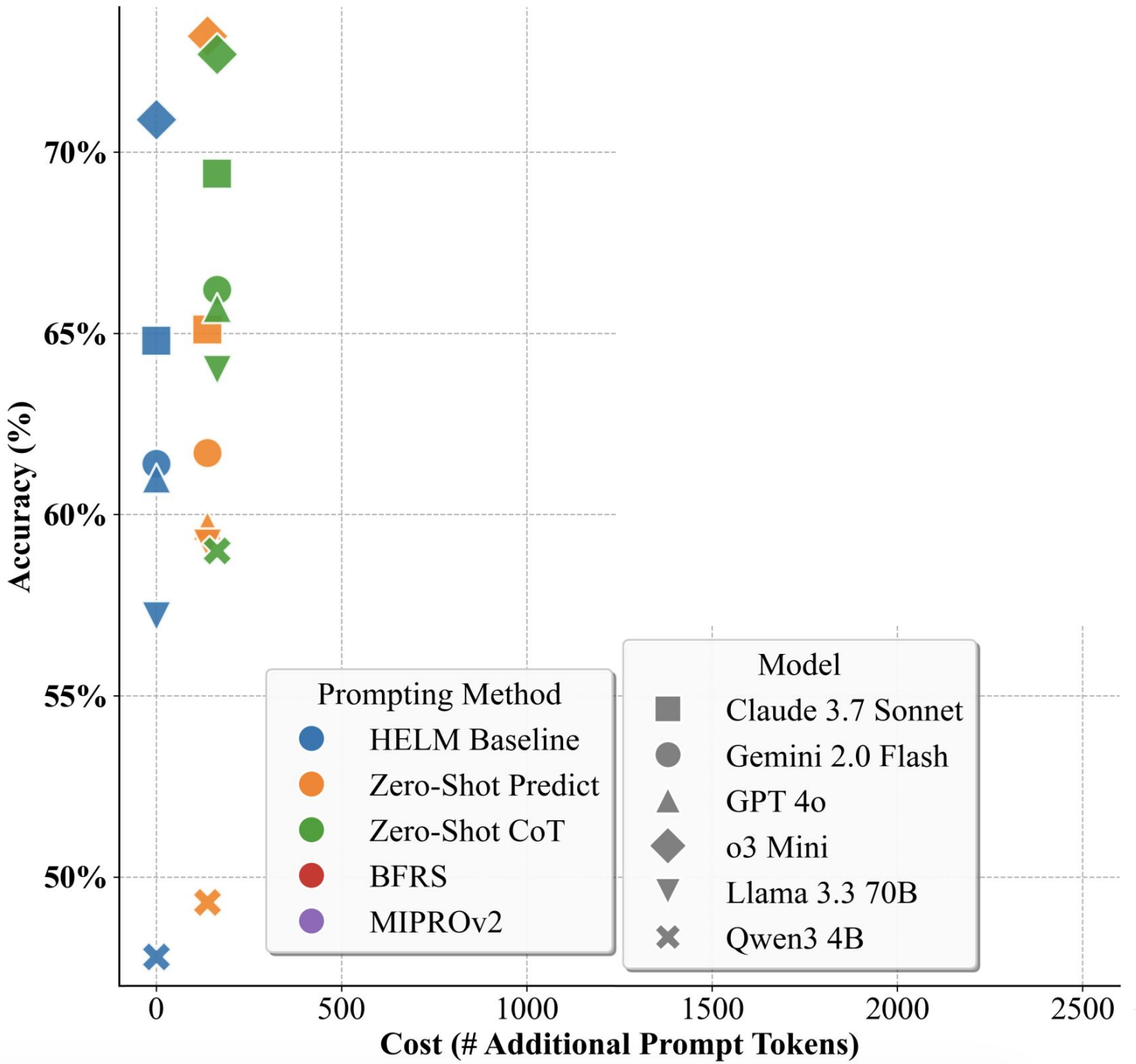


Accuracy vs Cost

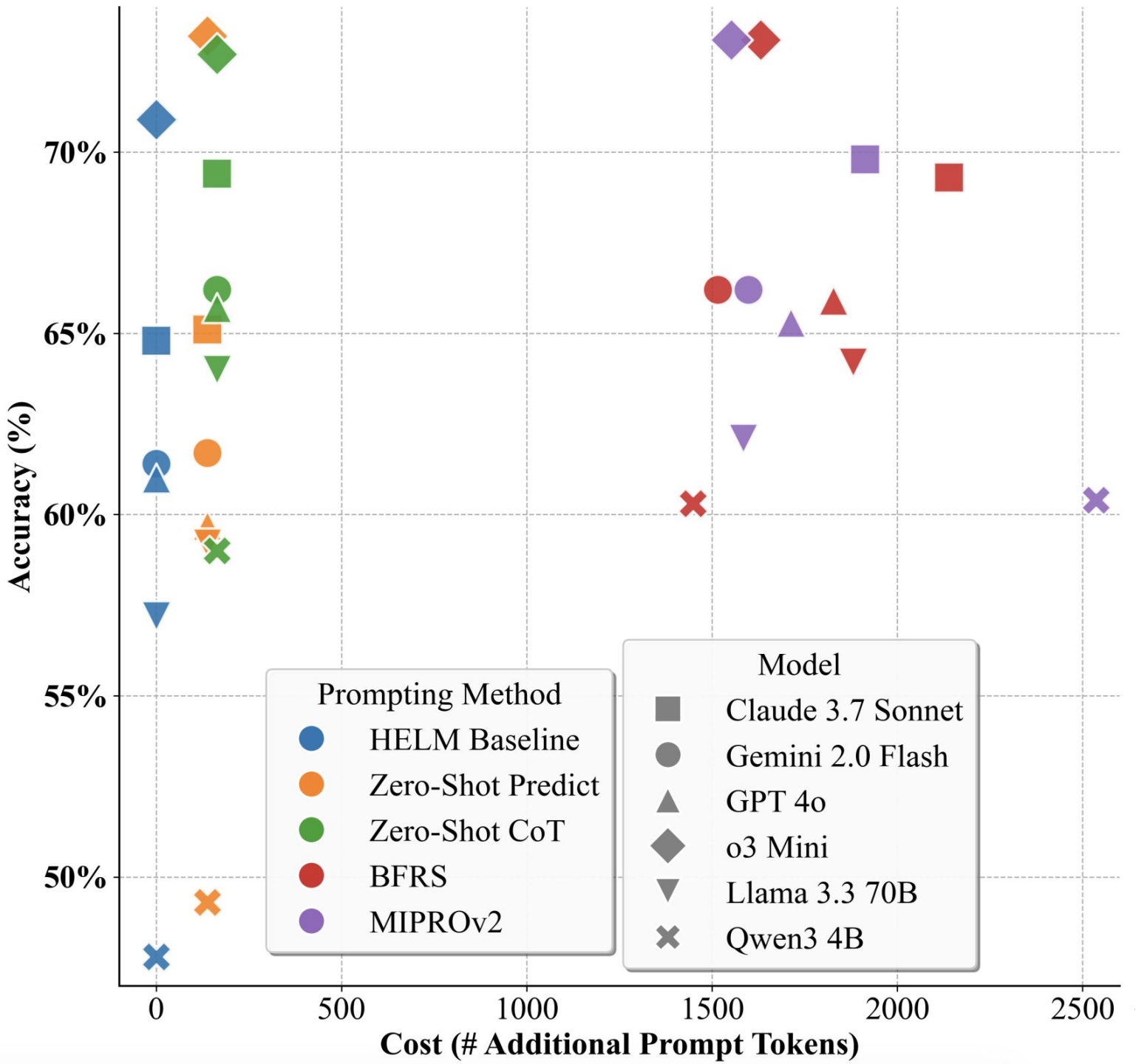


Stanford MEDICINE

Accuracy vs Cost



Accuracy vs Cost



Stanford MEDICINE

Main Findings & Conclusion

- Improved performance over HELM baseline
- Flipped leaderboard rankings
- Altered inter-model performance gaps
- Benchmark-dependent sensitivity
- CoT reduces sensitivity to prompt design



DSPy meets HELM

Asad Aali, Muhammad Ahmed Mohsin, Vasiliki Bikia, Arnav Singhvi, Richard Gaus, Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Yifan Mai, Jordan Cahoon, Mike Pfeffer, Roxana Daneshjou, Sanmi Koyejo, Emily Alsentzer, Christopher Potts, Nigam Shah, Akshay Chaudhari

Corresponding Author: ***Asad Aali*** (asadaali.com)



Accepted at ML4H!



Asad Aali, MS
Stanford University

asadaali@stanford.edu
https://asadaali.com/

Stanford MEDICINE