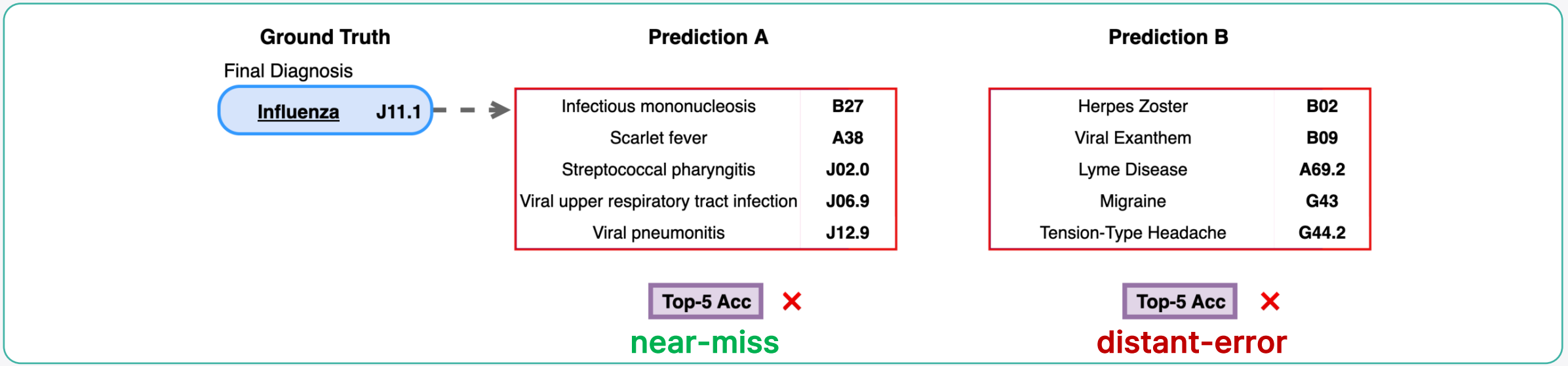
NEURAL INFORMATION PROCESSING SYSTEMS

H-DDx: A Hierarchical Evaluation Framework for Differential Diagnosis

Seungseop Lim*, Gibaeg Kim*, Hyunkyung Lee, Wooseok Han, Jean Seo, Jaehyo Yoo, Eunho Yang

AITRICS KAIST

Are All Diagnostic Errors Equal?



Limitations of Existing Evaluation Methods

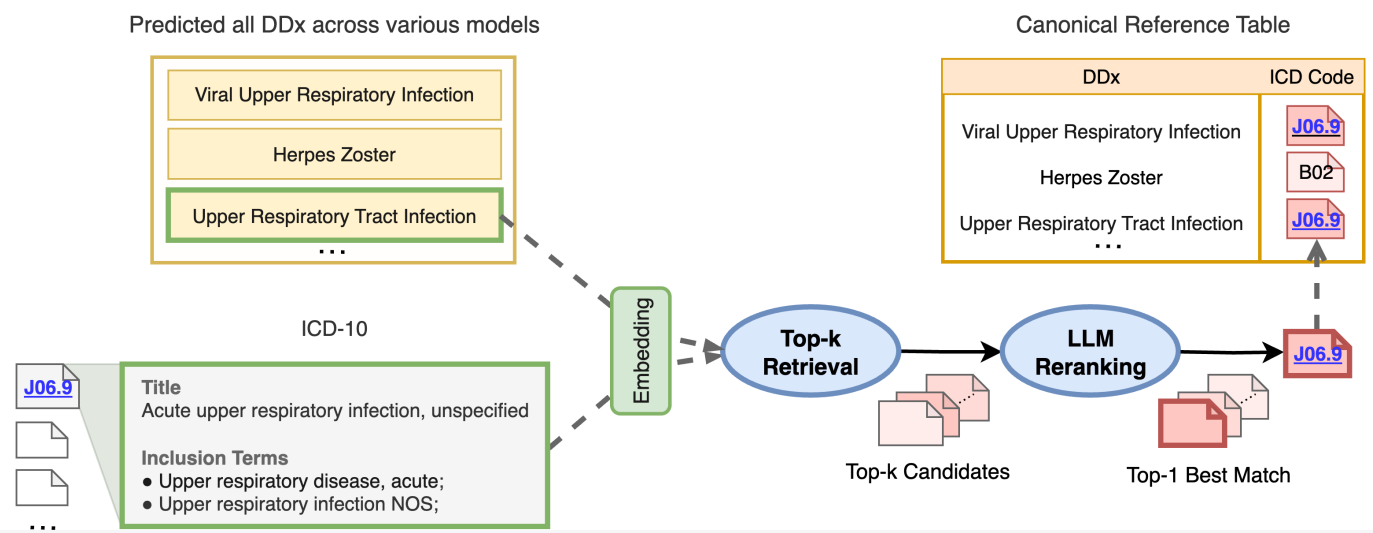
- Dependence on flat metrics (e.g., Top-k Accuracy).
- Limited to a binary determination of diagnosis inclusion.
- Verification relies heavily on LLM-based evaluators.

The H-DDx Framework

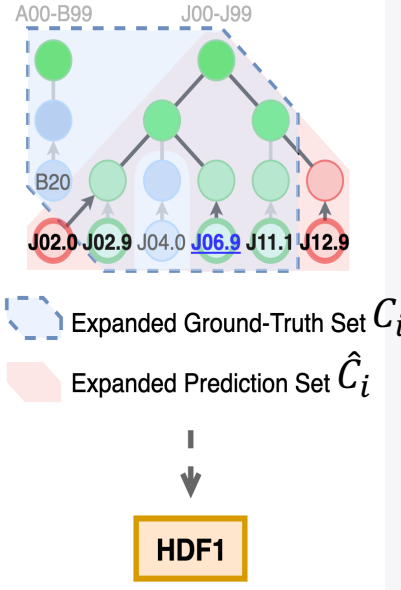
Our framework consists of two steps:

(i) a mapping pipeline that combines embedding-based retrieval with LLM reranking for high-accuracy conversion from free-text diagnoses to the ICD-10 codes.

(ii) the Hierarchical DDx F1 (HDF1) metric that p erforms set-based comparisons using ancestral code expansion within the ICD-10 taxonomy.



Mapping Free-Text Diagnoses to ICD-10 Codes



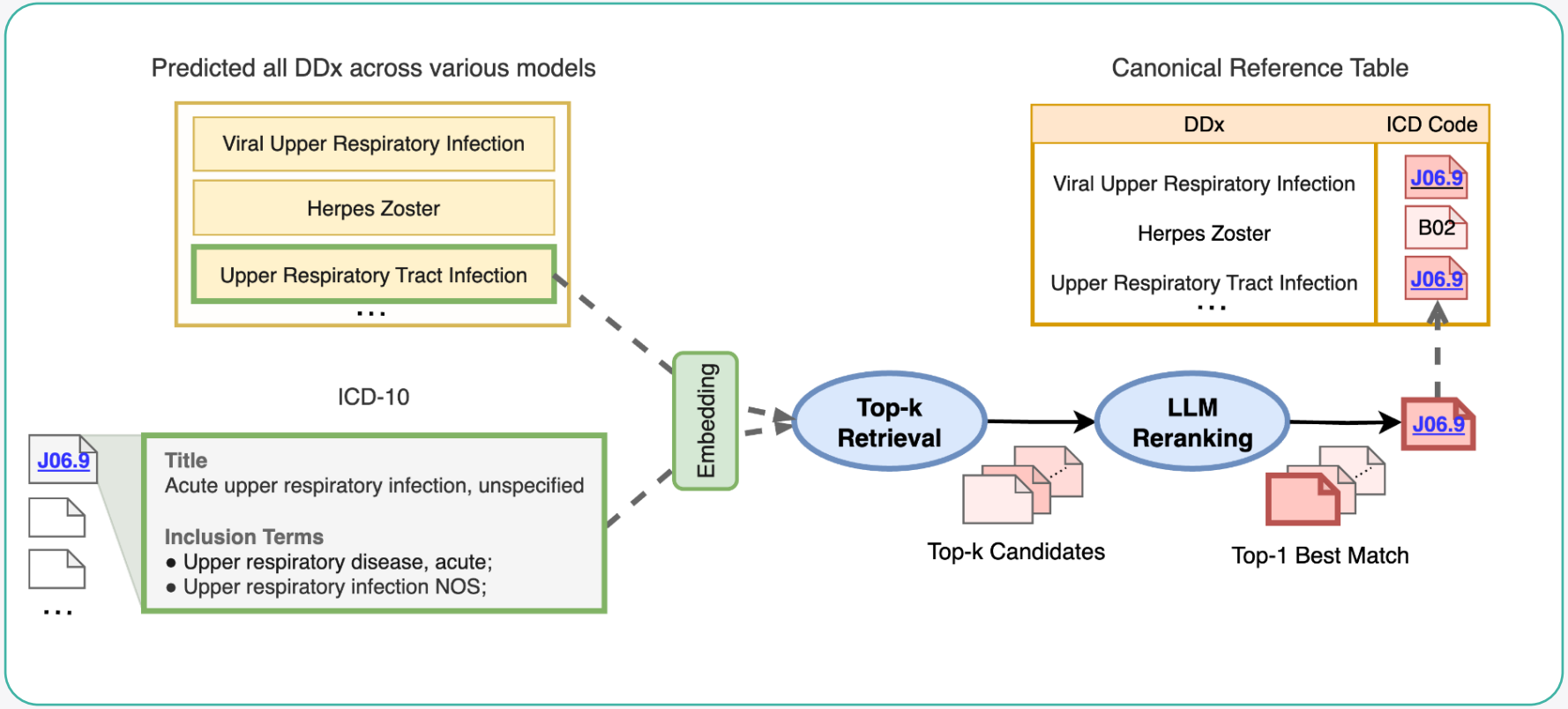
Hierarchical Evaluation

$$\text{HDP} = \frac{1}{N}\sum_{i=1}^{N}\frac{|C_i \cap \hat{C}_i|}{|\hat{C}_i|}$$

$$\text{HDR} = \frac{1}{N}\sum_{i=1}^{N}\frac{|C_i \cap \hat{C}_i|}{|C_i|}$$

$$\text{HDF1} = \frac{2 \times \text{HDP} \times \text{HDR}}{\text{HDP} + \text{HDR}}$$

Mapping Pipeline

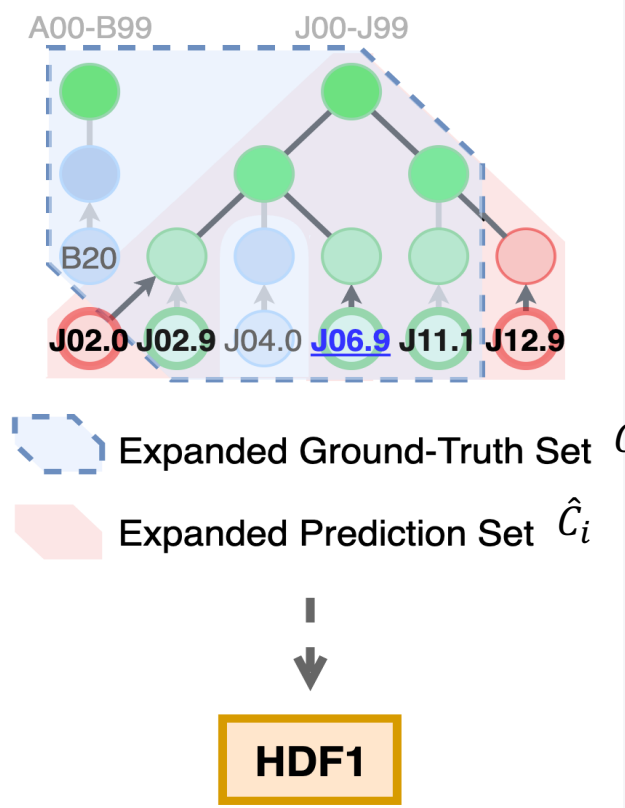


LLM-generated diagnosis	ICD-10 Code
Viral Upper Respiratory Infection	J06.9
Upper Respiratory Tract Infection	J06.9

	Model	Top-1
Top-k Retrieval	biobert-v1.1	0.4653
	BioSimCSE-BioLinkBERT-BASE	0.5446
	Qwen3-Embedding-0.6B	0.5842
	pubmedbert-base-embeddings	0.6436
	text-embedding-3-large	0.7129
LLM Reranking *(on best retriever)*	+ gemini-2.5-flash-lite	0.8317
	+ gpt-4o-mini	0.8614
	+ gemini-2.5-flash	0.8713
	+ gpt-4o	**0.9307**

Hierarchical Metric (HDF1)

The Hierarchical DDx F1 (HDF1) metric that performs set-based comparisons using ancestral code expansion wit hin the ICD-10 taxonomy.



Hierarchical Evaluation

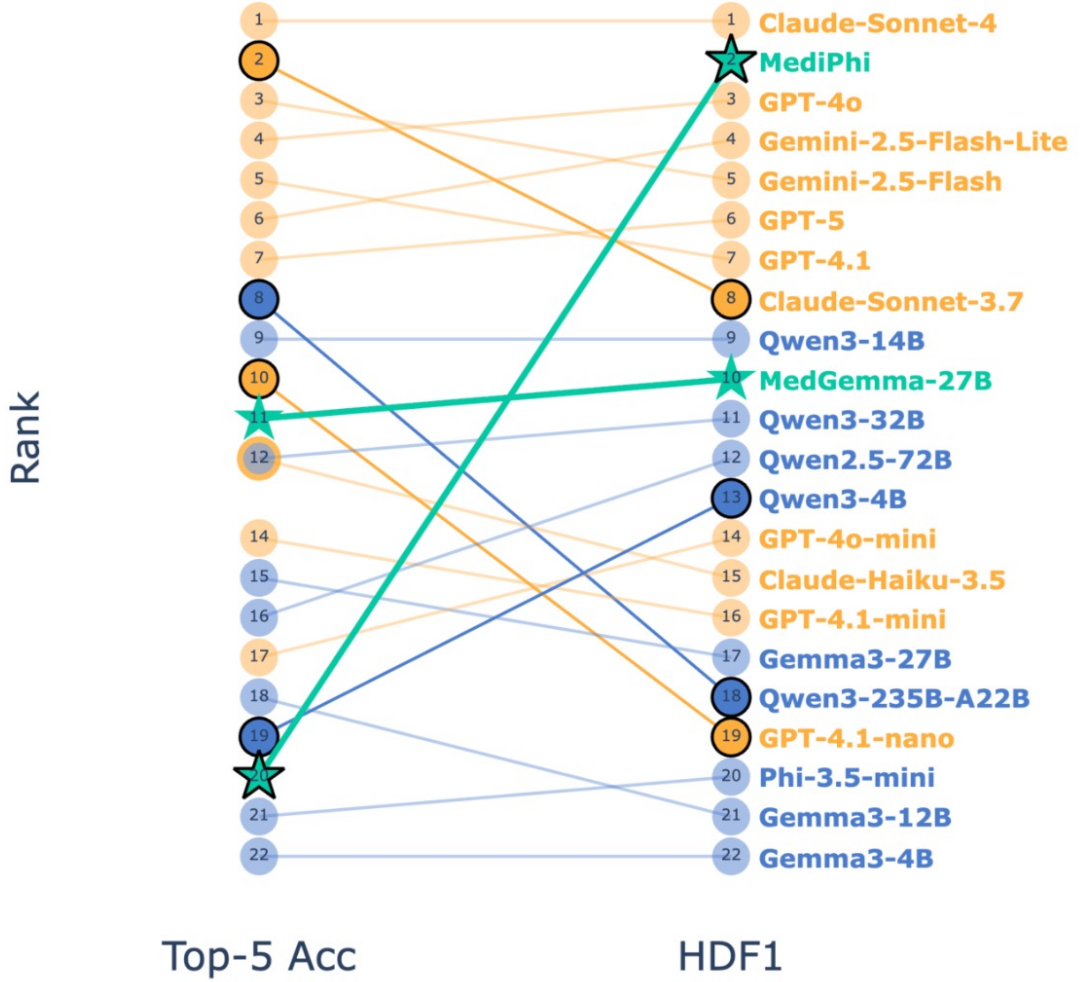
$$\mathrm{HDP} = \frac{1}{N}\sum_{i=1}^{N}\frac{|C_i \cap \hat{C}_i|}{|\hat{C}_i|}$$

$$\mathrm{HDR} = \frac{1}{N}\sum_{i=1}^{N}\frac{|C_i \cap \hat{C}_i|}{|C_i|}$$

$$\mathrm{HDF1} = \frac{2 \times \mathrm{HDP} \times \mathrm{HDR}}{\mathrm{HDP} + \mathrm{HDR}}$$

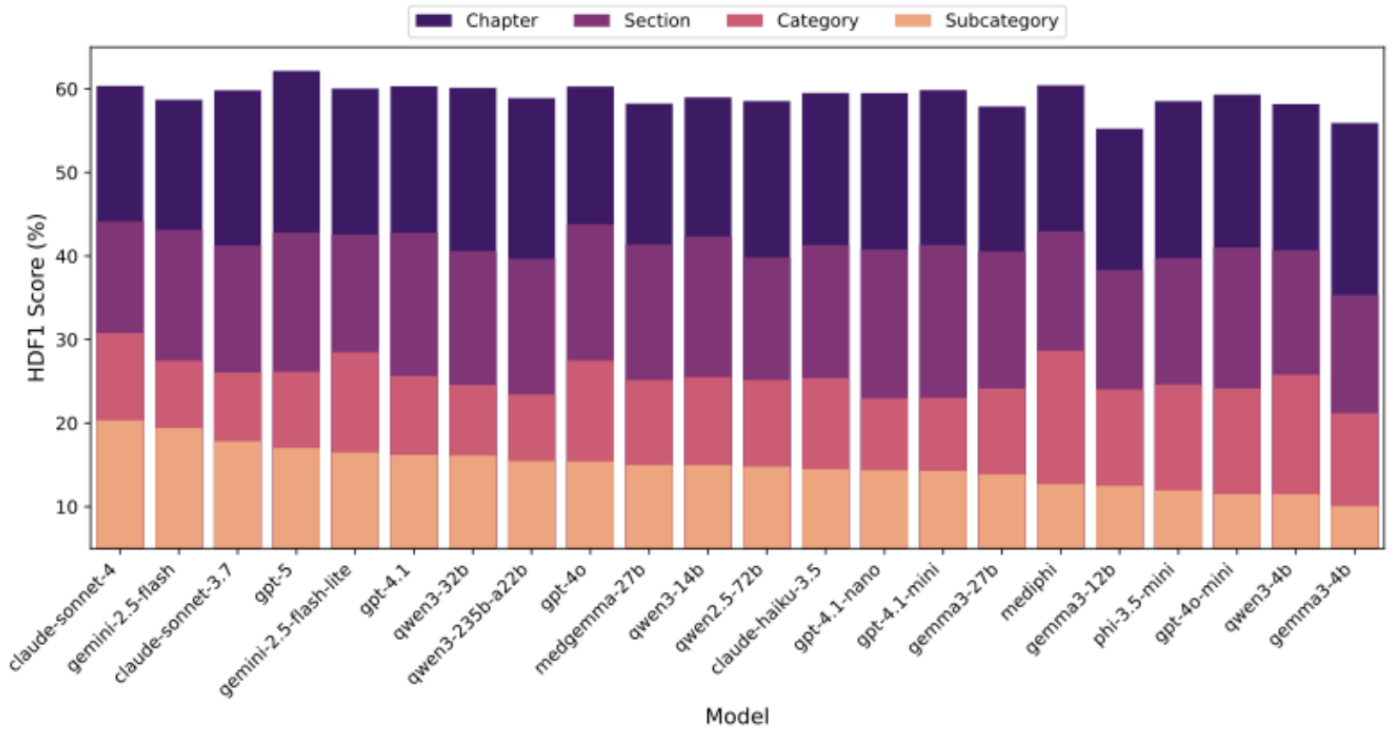
Ranking Shift

Model	Top-5	HDF1	ΔRank
Proprietary Models			
Claude-Haiku-3.5	0.7630	0.3237	↓ 3
Claude-Sonnet-3.7	0.8360	0.3380	↓ 6
Claude-Sonnet-4	**0.8390**	**0.3673**	–
Gemini-2.5-Flash-Lite	0.7890	0.3496	↑ 2
Gemini-2.5-Flash	0.8320	0.3483	↓ 2
GPT-4o-mini	0.7240	0.3276	↑ 3
GPT-4o	0.8040	0.3499	↑ 1
GPT-4.1-nano	0.7660	0.3213	↓ 9
GPT-4.1-mini	0.7590	0.3232	↓ 2
GPT-4.1	0.8010	0.3387	↓ 2
GPT-5	0.7830	0.3448	↑ 1
Open-source Models			
Phi-3.5-mini	0.6550	0.3187	↑ 1
Gemma3-4B	0.6080	0.2891	–
Gemma3-12B	0.7180	0.3075	↓ 3
Gemma3-27B	0.7460	0.3225	↓ 2
Qwen2.5-72B	0.7420	0.3299	↑ 4
Qwen3-4B	0.6720	0.3291	↑ 6
Qwen3-14B	0.7750	0.3367	–
Qwen3-32B	0.7630	0.3300	↑ 2
Qwen3-235B-A22B	0.7770	0.3215	↓ 10
Medical Fine-tuned Models			
MedGemma-27B	0.7650	0.3310	↑ 1
MediPhi	0.6660	0.3526	↑ 18



Hierarchical Error Analysis

Performance Analysis Leveraging ICD-10 Hierarchy



Model	Overall	J00-J99 (n=532)	I00-I99 (n=454)	A00-B99 (n=248)	G00-G99 (n=189)	S00-T88 (n=168)	D50-D89 (n=149)	K00-K95 (n=121)	F01-F99 (n=114)	C00-D49 (n=111)
Gemma-3-27B	32.25	33.31	30.44	10.68	23.75	19.50	10.03	31.36	18.90	33.93
MedGemma-27B	33.10 (+0.85)	30.81 (-2.50)	31.13 (+0.69)	10.66 (-0.02)	20.94 (-2.81)	20.56 (+1.06)	13.58 (+3.55)	27.21 (-4.15)	21.16 (+2.26)	33.47 (-0.46)
Phi-3.5-mini	31.87	37.72	24.33	11.45	13.86	17.18	10.82	21.58	16.24	28.64
MediPhi	35.26 (+3.39)	43.38 (+5.66)	25.17 (+0.84)	8.01 (-3.44)	13.09 (-0.77)	15.90 (-1.28)	20.97 (+10.15)	29.65 (+8.07)	18.76 (+2.52)	30.55 (+1.91)

Conclusion & Future Works

Conclusion

- Moves beyond binary accuracy by capturing clinically meaningful, hierarchical correctness.
- Highlights near-miss diagnoses that traditional metrics overlook.

Future Works

- Validation on real-world EHRs
- Integration of richer clinical ontologies (e.g., SNOMED-CT)
- Extension to multi-turn diagnostic agents
- Correlation with physician evaluations

Thank you