Stanford | ONLINE

Stanford | ONLINE

Stanford | Center for Health Education

Stanford MEDICINE

Smarter Healthcare: Leveraging GPT-5, Cosmos, and Predictive Models for Better Outcomes

Matt Lungren
Stanford University

Justin Norden
Stanford University

Seth Hain
Epic

Stanford | ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE

Table 1: Performance on QA benchmarks (%). The blue numbers and arrows indicate changes compared to GPT-4o-2024-11-20.

Dataset	GPT-5	GPT-5-mini	GPT-5-nano	GPT-4o-2024-11-20
MedQA				
US (4-option)	**95.84** (↑4.80%)	93.48	91.44	91.04
MedXpertQA Text				
Reasoning	**56.96** (↑26.33%)	45.94	36.38	30.63
Understanding	**54.84** (↑25.30%)	43.80	33.96	29.54
MMLU				
Anatomy	**92.59** (↑1.48%)	92.59	88.15	91.11
Clinical Knowledge	**95.09** (↑2.64%)	91.32	89.81	92.45
College Biology	**99.31** (↑2.09%)	99.31	97.92	97.22
College Medicine	**91.91** (↑1.74%)	88.44	85.55	90.17
Medical Genetics	**100.00** (↑4.00%)	99.00	98.00	96.00
Professional Medicine	**97.79** (↑1.10%)	97.43	96.69	96.69

3.2 Performance of GPT-5 on USMLE Self Assessment

As shown in Table 2, GPT-5 outperformed all baselines on all three steps, with the largest margin on Step 2 (+4.17%). Step 2 focuses on clinical decision-making and management, aligning with GPT-5's improved CoT reasoning. The average score across steps reached 95.22% (+2.88% vs GPT-4o), exceeding typical human passing thresholds by a wide margin, demonstrating the model's readiness for high-stakes clinical reasoning tasks.

Table 2: USMLE Sample Exam Performance (%). The blue numbers and arrows indicate changes compared to GPT-4o-2024-11-20.

	GPT-5	GPT-5-mini	GPT-5-nano	GPT-4o-2024-11-20
Step 1	**93.28** (↑0.84%)	93.28	93.28	92.44
Step 2	**97.50** (↑4.17%)	95.83	90.00	93.33
Step 3	**94.89** (↑3.65%)	94.89	92.70	91.24
Average	**95.22** (↑2.88%)	94.67	91.99	92.34

NLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE

Stanford AI Podcast - Googl

docs.google.com/presentation/d/1YGLIHJs8Hy8kZuxDmwSXRNs7tSA6KjXpOwBxOcTUBr8/edit?slide=id.g37971c67365_1_40#slide=id.g...

Table. Model Performance on Original and None of the Other Answers (NOTA)–Modified Question

Table. Model Performance on Original and None of the Other Answers (NOTA)–Modified Questions[a]

	Accuracy, % (No./total No.)		
Model	Original	NOTA-modified	Accuracy drop, % (No./total No.) [95 % CI]
1	92.65 (63/68)	83.82 (57/68)	8.82 (6/68) [2.70-18.92]
2	95.59 (65/68)	79.41 (54/68)	16.18 (11/68) [10.81-29.73]
3	88.24 (60/68)	61.76 (42/68)	26.47 (18/68) [17.57-39.19]
4	92.65 (63/68)	58.82 (40/68)	33.82 (23/68) [24.32-47.30]
5	85.29 (58/68)	48.53 (33/68)	36.76 (25/68) [28.38-51.35]
6	80.88 (55/68)	42.65 (29/68)	38.24 (26/68) [27.03-51.35]

[a] This table compares performan
validated questions. Original ac
performance on questions in th
while NOTA-modified accuracy
when the correct answer was re
the other answers" (NOTA). Mo
increasing accuracy drop. CIs w
the McNemar test for paired no

NLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE

Stanford AI Podcast - Google

docs.google.com/presentation/d/1YGLIHJs8Hy8kZuxDmwSXRNs7tSA6KjXpOwBxOcTUBr8/edit?slide=id.g37971c67365_0_62#slide=id.g...

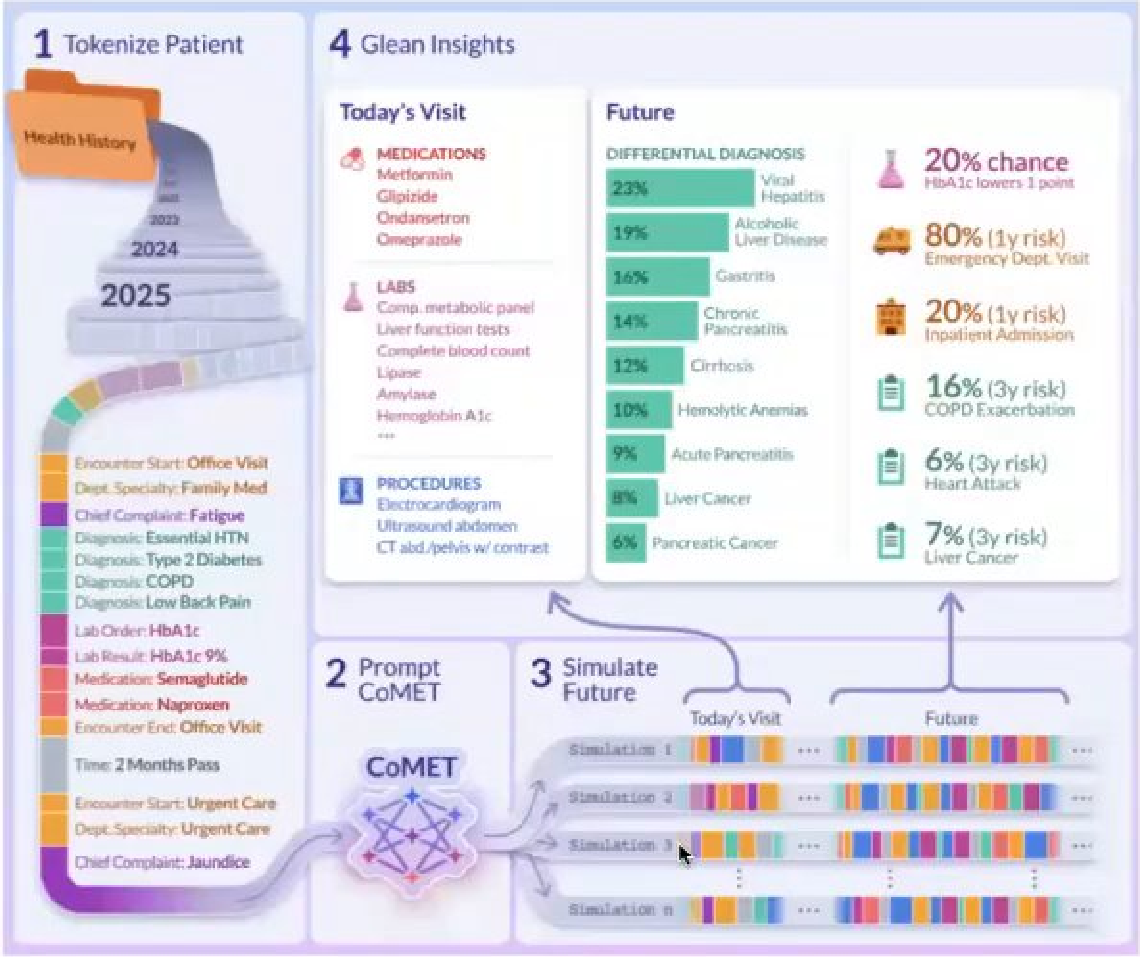


Figure 1: Overview of CoMET pretraining and inference. A patient journey is formulated as a sequence of medical events, and CoMET learns by predicting the next medical event. At inference time, CoMET is prompted with a patient's medical event history and simulates potential future trajectories by autoregressively generating the next events. Predictions for any target in CoMET's vocabulary are obtained from these simulated trajectories, enabling broad, out-of-the-box use on downstream tasks without task-specific fine-tuning or few-shot prompts.

NLINE

Stanford AI Podcast - Google

docs.google.com/presentation/d/1YGLIHJs8Hy8kZuxDmwSXRNs7tSA6KjXpOwBxOcTUBr8/edit?slide=id.g37971c67365_0_62#slide=id.g...

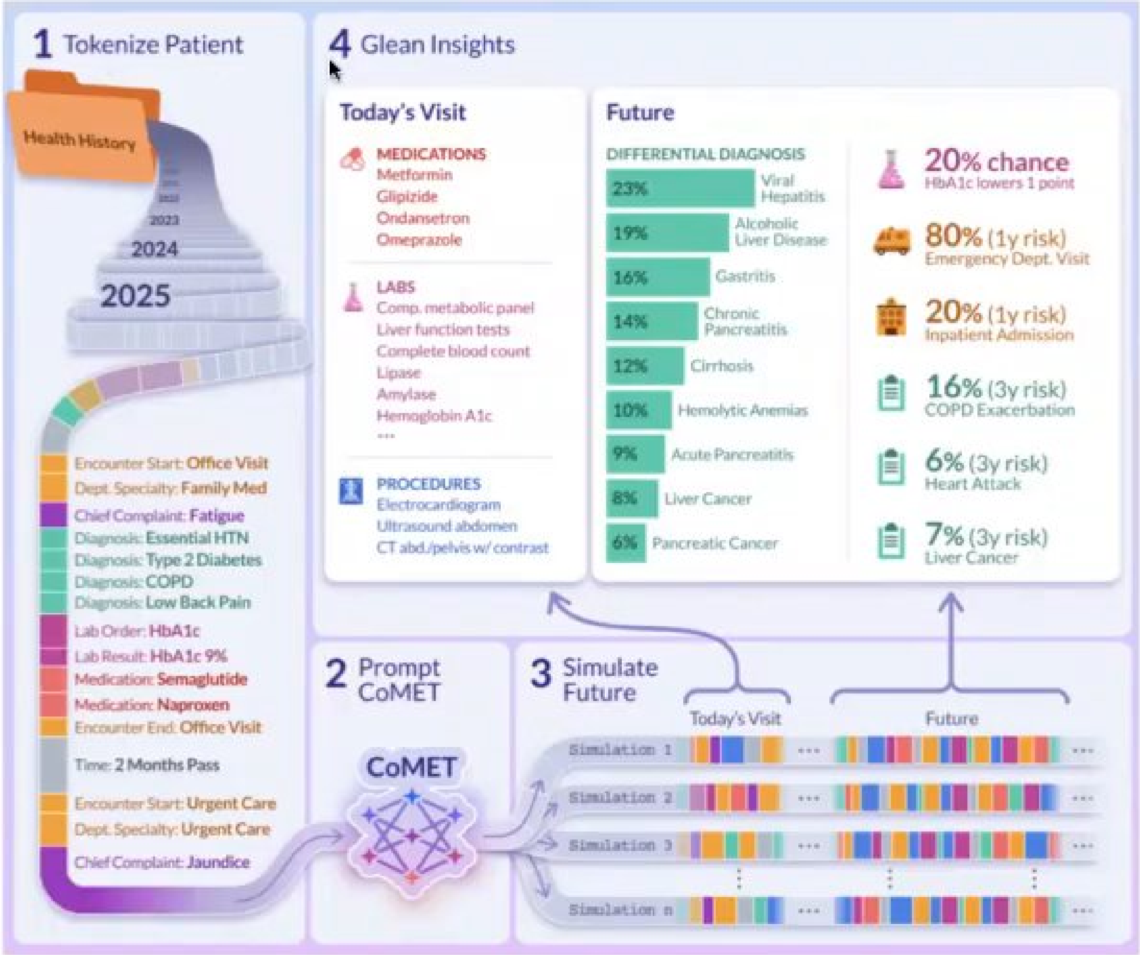


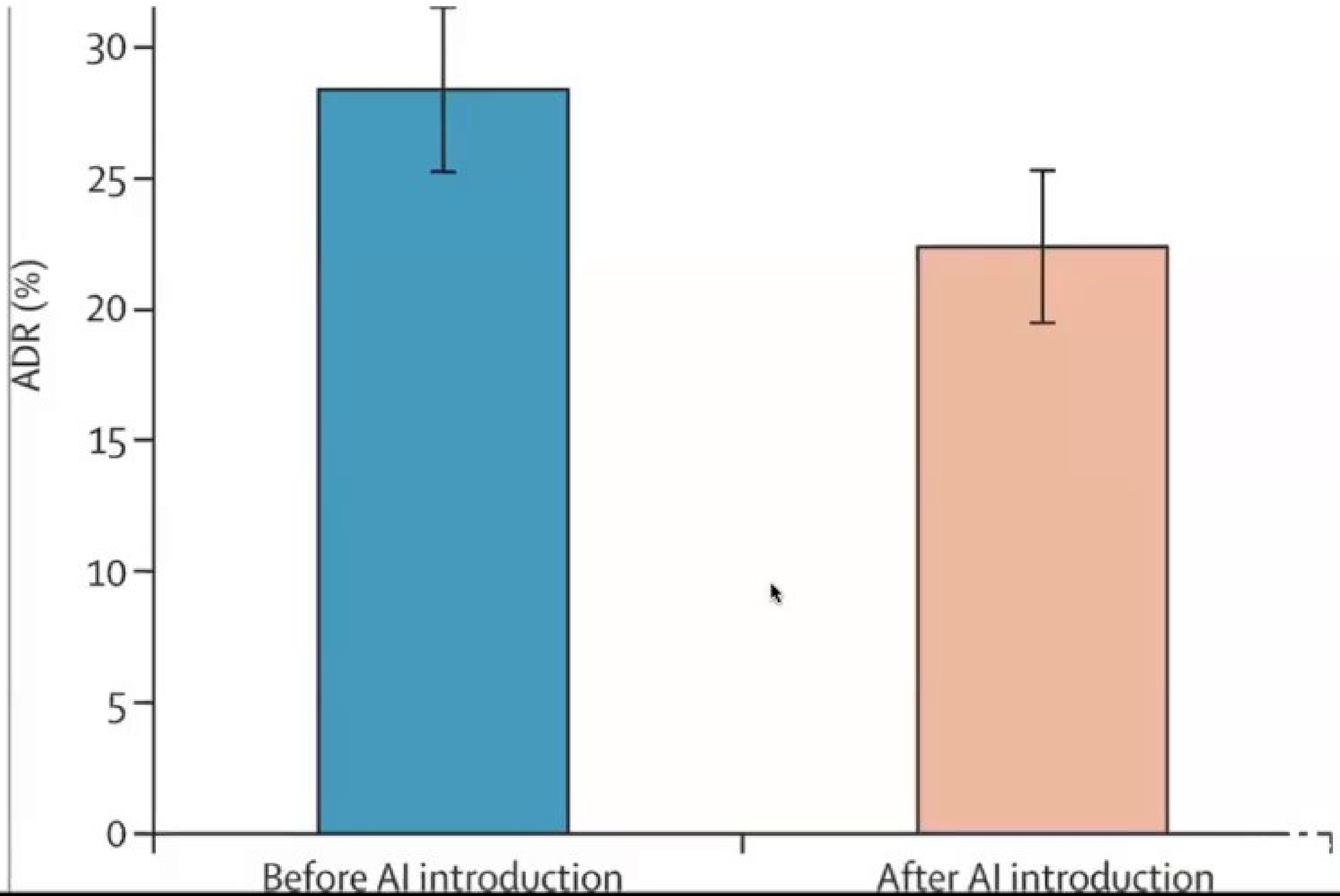
Figure 1: Overview of CoMET pretraining and inference. A patient journey is formulated as a sequence of medical events, and CoMET learns by predicting the next medical event. At inference time, CoMET is prompted with a patient's medical event history and simulates potential future trajectories by autoregressively generating the next events. Predictions for any target in CoMET's vocabulary are obtained from these simulated trajectories, enabling broad, out-of-the-box use on downstream tasks without task-specific fine-tuning or few-shot prompts.

NLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE

Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study




Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford | ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE

Stanford | ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Seth Hain
Matthew Lungren
Justin Norden
Stanford ONLINE


Stanford ONLINE