This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

A Supplementary Material

A.1 Synthetic data generation

To generate synthetic reports we first randomly generate categorical distributions over demographics (i.e. age and sex) and the conditions defined in Table A.1. Multiple conditions per patient could be sampled (without replacement) based on the aforementioned randomly generated categorical distribution, but the chance of subsequent conditions being sampled decreases based on how many have already been assigned, with a maximum of 3 co-occurring conditions. The symptoms, patient history and medications are then repeatedly (according to a minimum and maximum number of samples) sampled dependent on the condition that has already been assigned (illustrated in Table A.1) according to a uniform distribution. All of the samples are then used to form a latent description of the patient that is included in the prompt of the report generating LLM. The LLM is instructed to determine for itself what type of visit each report corresponds to, for example, a GP visit, a neurologist consultation, etc.

Table A.1: Condition templates used for synthetic patient generation. Each template defines pools of symptoms, relevant history items, and medications from which subsets are sampled.

Condition	Symptom Pool	History Pool	Medication Pool
Community-acquired pneumonia	Fever, productive cough, pleuritic chest pain, fatigue, shortness of breath	Recent URI, mild COPD, former smoker	Oral antibiotics, bronchodilator inhaler, acetaminophen PRN
Congestive heart failure	Dyspnea on exertion, orthopnea, bilateral leg swelling, fatigue, rapid weight gain	Ischemic cardiomyopathy, long-standing HTN, prior NSTEMI	Loop diuretic, ACE inhibitor, beta blocker
Type 2 diabetes (poor control)	Polyuria, polydipsia, fatigue, blurred vision, distal numbness	Metabolic syndrome, diabetic neuropathy, microalbuminuria	Metformin, basal insulin, dietary counseling
COPD exacerbation	Wheezing, increased sputum, shortness of breath, chest tightness, exercise intolerance	Chronic tobacco exposure, prior steroid bursts, baseline home inhaler	Short-acting bronchodilator, inhaled corticosteroid, prednisone taper
Chronic kidney disease	Fatigue, lower-extremity edema, poor appetite, pruritus, nocturia	Proteinuria, long-standing diabetes, hypertension	Renal diet, ARB therapy, erythropoietin support
Rheumatoid arthritis flare	Morning stiffness, small-joint pain, joint swelling, fatigue, reduced grip strength	Autoimmune disease, previous steroid-responsive flares, chronic methotrexate	Methotrexate, short steroid taper, NSAID as tolerated
Migraine disorder	Unilateral headache, photophobia, nausea, visual aura, sound sensitivity	Episodic migraines since early adulthood, stress-related worsening, family history	Triptan PRN, topiramate, hydration advice
Major depressive disorder	Low mood, insomnia, poor concentration, anhedonia, low appetite	Prior depressive episodes, work-related stress, anxiety symptoms	SSRI therapy, psychotherapy, sleep hygiene counseling
Peptic ulcer disease	Epigastric pain, nausea, early satiety, melena, fatigue	Frequent NSAID use, history of gastritis, prior H. pylori infection	PPI, H. pylori eradication, avoid NSAIDs
Breast cancer (active treatment)	Fatigue, unintentional weight loss, localized pain, nausea, treatment-related neuropathy	Recent chemotherapy, family history of malignancy, post-surgical follow-up	Ongoing chemotherapy, antiemetic regimen, oncology follow-up
Inflammatory bowel disease flare	Abdominal cramping, bloody diarrhea, weight loss, fatigue, urgency	Ulcerative colitis, prior steroid-treated flares, iron-deficiency anemia	Mesalamine, steroid taper, GI follow-up
Parkinson disease progression	Resting tremor, bradykinesia, gait instability, constipation, sleep disturbance	Several years of PD, prior falls, mild cognitive impairment	Carbidopa-levodopa, physical therapy, fall-prevention plan

A.2 Examples of queries and answers

Table A.2: Examples of queries and answers from the Synthetic Medical Reports test set

Question	Reference Answer
Numeric Measurement	
What is the patient's heart rate expected to be 239 days from now?	68 bpm
What is the patient's platelet count going to be in 298 days?	230 k/uL.
In two weeks, what will the patient's eGFR value show regarding their renal status?	The eGFR is stable at 44 mL/min/1.73m^2 (baseline was 43), indicating stable Stage 3 CKD without acute decline after NSAID suspension.
Where will the patient's hemoglobin level stand in six weeks?	11.5 g/dL
Treatment Action	
What lab tests are scheduled for the patient in three days and also in a week?	Stool studies will be performed, including Calprotectin, C. difficile toxin, and culture.
In 38 days, what will the patient's Duloxetine dosage and scheduled duration look like?	Duloxetine 60mg PO daily, prescribed for 42 days.
What daily dose of Sertraline will the patient be taking under the updated tapering protocol in 12 months?	Sertraline 50mg orally once daily for 90 days.
In 24 days' time, when is the Gastroenterology follow-up scheduled and what is it for?	The follow-up is scheduled in 16 days to evaluate completion of the steroid taper and tolerance to Mesalamine.
Clinical Status	
What will be the status of the patient's tobacco use disorder in six months?	In sustained remission.
What are the patient's supervision requirements during complex movements going to be 40 days from now?	Supervision is still required for safety during complex movements, although balance is slightly improved.
What will the patient's tobacco use status look like in six months?	The patient has maintained complete tobacco abstinence.

Question	Reference Answer
What will the status of the patient's episodic migraine with aura be in 13 months?	Stable and inactive
Structural Observation	
How will the patient's surgical wound look on local exam one week from now?	Well-healed without signs of cancer recurrence, skin breakdown, or infection.
When they get the EGD in two days, what will it show regarding the size and location of the ulcer?	A single 1.5 cm ulcer with a clean base (Forrest Class III) in the gastric antrum.
What will chest auscultation reveal when the patient is examined in 14 days?	Clear breath sounds bilaterally with very mild, distant expiratory wheezing at the bases.
What are the brain MRI results expected to show regarding any secondary causes of the patient's migraines fifty-one days from now?	The brain MRI will show age-related white matter changes but no evidence of acute infarction, hemorrhage, mass lesion, or other structural abnormalities to explain the symptoms.
Future Clinical Presentation	
What will we find on the patient's chest auscultation in two weeks?	Significantly reduced, faint expiratory wheezing at the bases, clear upper lobes
What is the purpose of the planned Nephrology follow-up going to be in 173 days?	To re-evaluate CKD status, monitor kidney function, and review any long-term impact of medications.
In thirty-seven days, what will the scheduled General Practitioner routine follow-up investigate?	Long-term glycemic trends by checking HbA1c post-steroid taper, and confirmation of complete resolution of anemia following the completion of the 90-day iron course.
What will we find on the patient's lung auscultation in two weeks?	Normal vesicular breath sounds bilaterally, with no audible wheezing, rhonchi, or crackles.
Procedural Technique	
What will the scheduled imaging for further characterizing the L3 lesion entail four days from now?	A PET-CT scan or CT scan of Thorax/Abdomen/Pelvis with bone windows to evaluate the L3 lesion and check for other metastatic sites.

Question	Reference Answer
In one week, what will the purpose and timeline of the planned stool studies be?	Stool studies, including Calprotectin, C. difficile toxin, and Routine Culture, will be performed immediately to assess inflammatory markers and rule out infectious triggers, with results expected in 3-5 days.
What diagnostic procedure are we performing in two weeks to confirm the patient's IBD diagnosis and check her disease activity?	A colonoscopy with biopsies, scheduled in 7 days from now.
What type of CPAP mask interface will the patient be transitioning to in two months?	A simpler interface with magnetic clips

Table A.3: Examples of queries and answers from the MIMIC ICU Notes test set

Question	Reference Answer
Numeric Measurement	
In 2 days, what will be the distance of the ETT above the carina?	5.8 cm
What will the patient's ascending aorta measurements be in 4 years and 1 month from now?	Ascending aorta is 3.8 cm, minimally enlarged.
Six days from now, what will be the dimensions of the patient's right testis?	4.2 x 2.8 x 1.4 cm
What will the patient's initial vital signs be in the ED in 108 days?	100.6 95 173/98 20 95% on RA
Treatment Action	
Sixteen days from now, what major surgical or invasive procedures will be performed?	OR debridement, PEG placement
What will the patient's CHF medication regimen look like in 359 days?	60 mg PO Lasix BID

Question	Reference Answer
What major surgical or invasive procedure will be performed two weeks from now?	Removal of infected aortic graft, aorto left femoral bypass using femoral vein, right iliofemoral bypass using thigh femoral vein, sartorius flap myoplasty of the left femoral anastomosis.
What will be removed from the patient in 1 day?	left-sided chest tube and mediastinal drains.
Clinical Status	
In 6 days, what will the right carotid stenosis be consistent with?	This is consistent with less than 40% stenosis.
How visible will the pre-existing pneumopericardium be in 2 days?	The pre-existing pneumopericardium is still clearly visible.
What will the status of generalized demineralization be in 1 week?	There is generalized demineralization.
What will be the status of the patient's atelectasis by end of day?	There is moderate bilateral atelectasis.
Structural Observation	
What will the bibasilar consolidations look like in six days?	Bibasilar consolidations, larger on the left side, have minimally increased on the left.
What will be the position of the percutaneous cholecystostomy drain 59 days from now?	A right lateral approach drainage catheter terminates within the gallbladder.
In 100 days, what will be the caliber and configuration of the ventricles and sulci?	The ventricles and sulci are normal in caliber and configuration.
What will the patient's diagnosis be 88 days from now?	20 x 20 mm right petrous-cavernous segment aneurysm.
Future Clinical Presentation	
What will the patient's pain level be on arrival to the floor 2 years, 5 months, and 4 days from now?	On arrival to the floor pt reports severe ankle pain.
What will be the patient's chief complaint two days from now?	altered mental status
What will the patient's clinical history be tomorrow?	___ woman with pedestrian struck with swollen ankle.

Question	Reference Answer
What will be the reason for the exam in four years, two months, and 28 days?	Assess for pneumothorax. The chest tube removal.
Procedural Technique	
What will the dose acquisition sequence be in four days?	1) Spiral Acquisition 6.5 s, 41.9 cm; CTDIvol = 8.0 mGy (Body) DLP = 330.2 mGy-cm.
What technique will be used for the chest imaging in four days?	Chest PA and lateral
What technique will be used for the procedure 57 days from now?	Ultrasound guided therapeutic paracentesis
Can you tell me about the details of the patient's major surgical or invasive procedure tomorrow?	None

A.3 Prompt design

Building datasets for future querying The boxes below show the prompts given to the LLMs involved in synthesizing future queries and answers, as described in Section 4. For the Synthetic Medical Reports dataset, the chosen LLM was Gemini 3.5 Flash. For the larger MIMIC ICU Notes dataset we used Gemini 2.5 Flash for cost reasons.

Predicate Extraction

You are a medical expert reviewing clinical notes from a patient's hospital stay.

Generate meaningful clinical questions about this report whose answers can be found in the report text.

Context: This report is one entry in a chronological sequence of clinical documents describing a patient's hospital course over time.

Structure ALL questions in future tense so that a time window can be easily appended later. Format questions like "What will the patient's [finding] be..." or "What is the status of the patient's [condition]..." so that time phrases like "in 3 days" or "in 2 weeks" can be naturally added.

Focus on clinically meaningful information that a physician would want to track over time:
- Symptoms, complaints, or subjective patient reports
- Vital signs, lab values, imaging findings, or physical exam results
- Active diagnoses, new findings, or changes in clinical status
- Medications, dosages, or treatment changes
- Procedures performed and their outcomes
- Clinical assessments, differential diagnoses, or care plan decisions

Avoid questions about administrative details, report metadata, or scanning technique unless they are clinically significant.

Do NOT use ordinal references (e.g. "first", "second", "third") to identify items within a list or sequence of planned actions, visits, or recommendations. Instead, refer to the item by its clinical content (e.g. "What will the follow-up cardiology appointment involve..." rather than "What will the second planned visit be..."). The ordering of bullet points or list items in a report is not clinically meaningful and cannot be reliably predicted.

Each question should be answerable directly from this report's content.

Return a list of (question, answer) pairs.

Fig. A.1: Predicate extraction prompt.

Query Formulation

You are given a numbered list of questions about future events or states. Each question already contains a specific time reference (e.g. "in 3 days").

Rephrase each question to sound more natural and varied, as a human would ask it. You may:

- Restructure the sentence freely (move the time reference, reword the question)
- Rephrase the time reference in different ways (e.g. "in 3 days" → "three days from now", "after three days", "on day three", etc.). It's acceptable to reuse similar time expressions when the input list is long.
- Sacrifice some variety in order to maintain a natural phrasing if necessary
- Convert days to larger units when exact: 7 days = 1 week, 30 days = 1 month, 365 days = 1 year. Combine units naturally (e.g. 382 days → "in 1 year and 17 days")
- For 0 days, phrase as "later today" or "by end of day"
- For 1 day, you may say "tomorrow"
- Randomly alternate between digit and spelled-out forms for small numbers (up to about 100). For larger numbers always use digits (e.g. "in 158 days", NOT "one hundred and fifty-eight days").

You must NOT:

- Change what the question is asking — a hypothetical answer must remain the same
- Add or remove information
- Change the time horizon (e.g. don't say "next week" for 3 days)
- Spell out large numbers in words — write "229 days", not "two hundred twenty-nine days"
- Use archaic, literary, or unusual time phrasing (e.g. "hence", "X days hence", "Come X days", "upon the passing of", "in the span of")
- Say "X days out" — use "in X days" or "X days from now" instead

Use modern conversational English, as a doctor would type in a chat message to a clinical assistant.

Return exactly one rephrased question per input, in the SAME ORDER.

Fig. A.2: Query formulation prompt.

Validation

You are a clinical QA validator reviewing question-answer pairs generated from a patient's hospital records.

You will receive:
1. PAST CONTEXT: prior clinical notes from the patient's stay
2. A QUESTION about a future clinical event or finding
3. An ANSWER to that question

Imagine a physician has a crystal ball they can ask questions about their patient's future. Your task: determine whether the question is WORTH ASKING the crystal ball, or whether it would be a waste because the physician already knows the answer with certainty from the past context.

A question is VALID (is_valid=true) when:
- The physician would genuinely want to know the answer because the outcome is uncertain (even if somewhat predictable or likely)
- The answer describes how a known condition has evolved, progressed, or resolved — even if the condition itself was already documented
- The answer contains updated measurements, imaging findings, lab values, or clinical status from a later point in time

A question is INVALID (is_valid=false) only when:
- The answer is already stated as fact in the past context and the question is redundant (e.g. asking about a scheduled procedure that is already confirmed in the notes)
- The physician would NOT bother asking the crystal ball because they already know the answer with certainty — not just likelihood

Example: A tumor was found on day 0. Asking "what will imaging show on day 168?" is VALID — the tumor could be stable, growing, shrinking, or gone. The physician would want to ask the crystal ball.

Example: A surgery is scheduled for next Tuesday. Asking "what procedure will be performed next week?" is INVALID — it's already confirmed.

Return JSON with:
- is_valid: true/false as described above

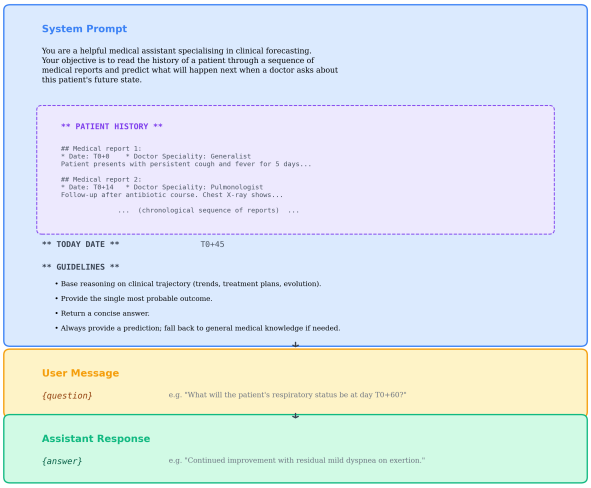
Fig. A.3: Validation prompt.

Evaluation Our evaluation pipeline uses two prompt types: inference prompts, which generate clinical predictions, and judge prompts, which evaluate those predictions against reference answers.

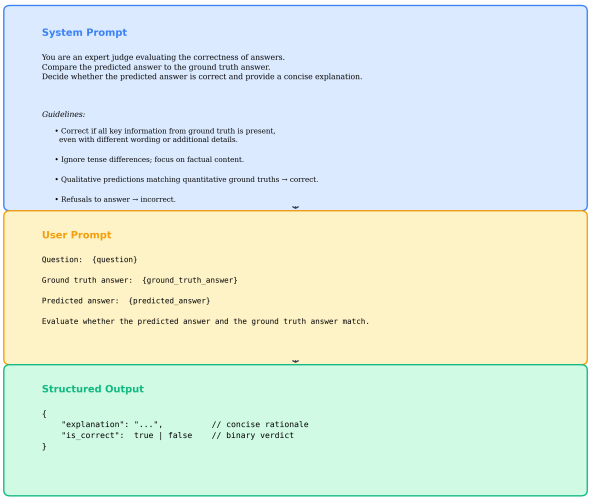
Inference prompt: During both training and inference, the model receives the patient's complete longitudinal clinical history, including report dates, physician specialties, and the current date, as a system message. It is instructed to act as a clinical forecasting assistant, reasoning over the patient's trajectory to predict the single most likely outcome. The user message contains only the question. To ensure consistency, training and inference share the same prompt-construction function. We further optimized the prompt using DSPy's GEPA (Grounded Example-based Prompt Adaptation), which iteratively refines instructions based on training-set performance and failure-case feedback. The optimized prompt preserves the original structure while adding clarifications for challenging cases, such as medication duration, stable vital signs, and projecting recent clinical values (Fig. A.4a).

LLM-as-a-judge: A separate LLM evaluates whether a prediction matches the reference answer, focusing on factual correctness rather than wording. Semantically equivalent answers, including qualitative descriptions of quantitative

findings, are accepted, while refusals are always marked incorrect. The judge returns a binary correctness label and a brief explanation (Fig. A.4b).



(a)



(b)

Fig. A.4: Overview of the QA inference prompt (a) and the prompt for the LLM-as-a-judge (b).

A.4 Training details

For each model, the optimal hyperparameters were defined through a hyperparameter optimization run on a subset of the data, trained for a shorter time. The best hyperparameters were chosen given the lowest validation loss, i.e. token cross-entropy. Final training was performed for a full epoch given the best hyperparameters, indicated in Table A.4

Table A.4: Training details for the smaller weight models. Hyperparameter optimization study was applied to a subset of the data

Model	Training type	GPU	Hyperparameters
Medgemma-1.5-4b-it	Lora (r=16)	H100	lr=$1e^{-5}$, warm-up-ratio=0.005, lr-scheduler-type=polynomial (power=2, lr-end=$1e^{-7}$), batch-size=16, lora-alpha=32, weight-decay=$2e^{-4}$
Medgemma-1.5-4b-it	Full	H100x2	lr=$1e^{-5}$, warm-up-ratio=0.005, lr-scheduler-type=polynomial (power=2, lr-end=$1e^{-7}$), batch-size=8, weight-decay=$2e^{-4}$
Gemma-3-4b-it	Lora (r=16)	H100	lr=$5e^{-5}$, warm-up-ratio=0.005, lr-scheduler-type=polynomial (power=2, lr-end=$1e^{-7}$), batch-size=16, lora-alpha=32, weight-decay=$2e^{-4}$
Gemma-3-4b-it	Full	H100x2	lr=$5e^{-5}$, warm-up-ratio=0.005, lr-scheduler-type=polynomial (power=2, lr-end=$1e^{-7}$), batch-size=8, weight-decay=$2e^{-4}$
Gemma-3-12b-it	Lora (r=16)	H100x2	lr=$5e^{-5}$, warm-up-ratio=0.005, lr-scheduler-type=polynomial (power=2, lr-end=$1e^{-7}$), batch-size=8, lora-alpha=32, weight-decay=$2e^{-4}$

A.5 LLM-as-a-judge analysis

We manually assessed the decisions of the judge for 100 randomised (over all model evaluations) questions for both the MIMIC and synthetic datasets. Below are the results for both datasets, including the Cohen's κ value and confusion matrices. We also include examples of questions with their respective predicted answers and reference answers demonstrating examples were the humans and judges both agree and disagree.

MIMIC ($n = 100$)

- Agreement with judge: 89 (89.0%)
- Disagreement with judge: 11 (11.0%)
- Judge labeled correct: 41.0%
- Human labeled correct: 46.0%
- Expected chance agreement: 50.7%
- Cohen's κ: 0.7768

Table A.5: Confusion matrix for MIMIC (rows = judge, columns = human).

	Human correct	Human incorrect
Judge correct	38	3
Judge incorrect	8	51

SYNTHETIC ($n = 100$)

- Agreement with judge: 89 (89.0%)
- Disagreement with judge: 11 (11.0%)
- Judge labeled correct: 46.0%
- Human labeled correct: 55.0%
- Expected chance agreement: 49.6%
- Cohen's κ: 0.7817

Table A.6: Confusion matrix for SYNTHETIC (rows = judge, columns = human).

	Human correct	Human incorrect
Judge correct	45	1
Judge incorrect	10	44

Example 1

Question What will the patient's mental status be in three days?
Reference Answer Clear and coherent.
Predicted Answer The patient's mental status is likely to be normal or return to baseline, given the absence of acute intracranial injury on initial CT and no reported neurological deficits.
Judge Decision: Correct
Human Decision: Correct
Judge Reasoning The predicted answer indicates "normal or return to baseline," which aligns with "clear and coherent" in the reference, suggesting an absence of mental status issues. The additional context provided in the predicted answer supports this conclusion.
Human Reasoning Answers are semantically similar.

Example 2

Question What will the patient's oxygen saturation be at admission from the transferring facility in 2 years, 4 months, and 10 days?
Reference Answer 100% RA
Predicted Answer Based on the initial report, the patient has a history of old granulomatous disease and aortic unfolding. These findings suggest a chronic, potentially progressive cardiopulmonary condition. Without further information about treatment or current stability, it is difficult to predict precise oxygen saturation levels. However, given the existing issues, I predict a chronically low oxygen saturation. Prediction: - Admission: Likely below 95%. - Future oxygen saturation: Likely between 88–92%.
Judge Decision: Incorrect
Human Decision: Incorrect
Judge Reasoning The predicted answer did not match the reference. The reference states a specific percentage (100% RA), while the predicted answer gives a range that is significantly lower than the reference and expresses uncertainty.
Human Reasoning The predicted answer incorrectly assumes a deteriorating state and hence predicts an oxygen saturation significantly below the reference.

Example 3

Question What will the findings of the targeted scanning over the left deltoid area be in two days?
Reference Answer No fluid collection, mass, or other abnormality is identified.
Predicted Answer No findings. The initial report indicates no acute chest disease and a normal chest radiograph. There is no information suggesting a problem with the left deltoid area.
Judge Decision: Incorrect
Human Decision: Correct
Judge Reasoning The predicted answer is vague and also includes information not related to the question (chest disease and chest radiograph). The reference answer is specific and directly addresses the question about the left deltoid area.
Human Reasoning The answers convey the same semantic information; namely, that there are no abnormalities shown in the scan. We felt the judge was too harsh.

Example 4

Question What will be the status of the previously seen orogastric tube 26 days from now?
Reference Answer not visualized on the current study.
Predicted Answer The provided medical report does not mention an orogastric tube. Therefore, I cannot predict its status 26 days from now.
Judge Decision: Correct
Human Decision: Incorrect
Judge Reasoning The predicted answer correctly identifies that the OGT is not mentioned in the report, aligning with the reference that it is not visualized on the current study.
Human Reasoning The judge is directed to mark any refusal to answer as incorrect.

A.6 Accuracy as function of forecast horizon and query type

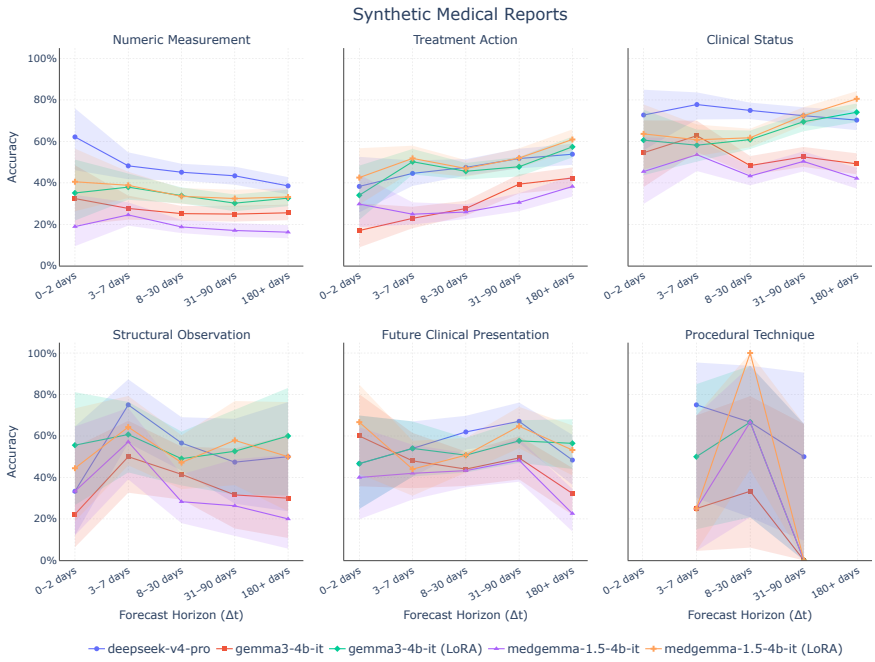


Fig. A.5: Model accuracy by forecast horizon (Δt) and query type for five models on the Synthetic Medical Reports test set. Δt was not experimentally controlled; different bins contain different questions, hampering strong causal conclusions.

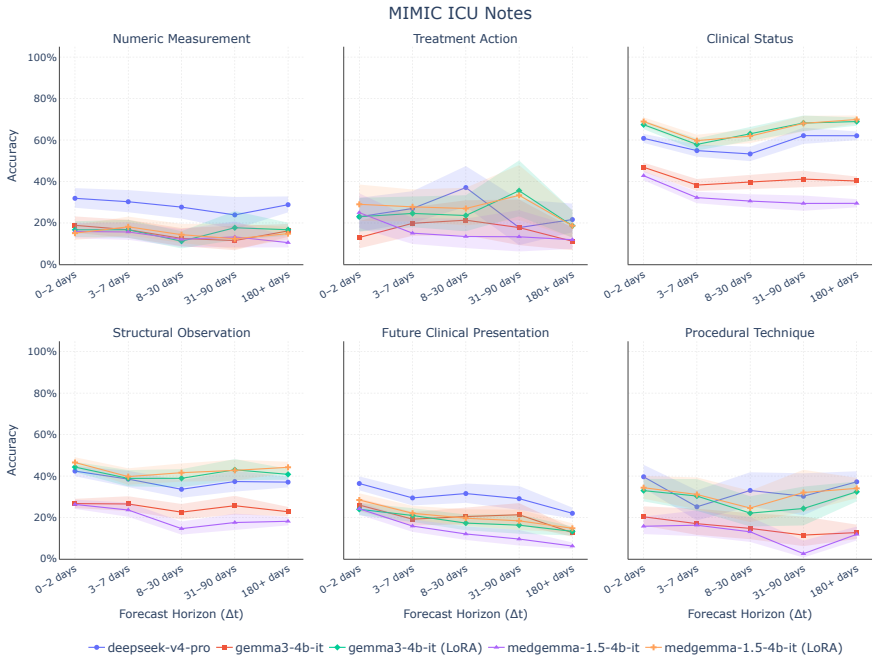


Fig. A.6: Model accuracy by forecast horizon (Δt) and query type for five models on the MIMIC ICU Notes test set. Δt was not experimentally controlled; different bins contain different questions, hampering strong causal conclusions.