DischargeSim: A Simulation Benchmark for Educational Doctor–Patient Communication at Discharge

Zonghai Yao [*1,2], **Michael Sun** [*2], **Won Seok Jang**[1,3]
Sunjae Kwon[1,2], **Soie Kwon**[4], **Hong Yu**[1,2,3]
[1]Center for Healthcare Organization and Implementation Research, VA Bedford Health Care
[2]Manning College of Information and Computer Sciences, UMass Amherst, MA, USA
[3]Miner School of Computer and Information Sciences, UMass Lowell, MA, USA
[4]Department of Internal Medicine, Chung-Ang University, Seoul, Republic of Korea
{zonghaiyao, mssun}@umass.edu

Abstract

Discharge communication is a critical yet underexplored component of patient care, where the goal shifts from diagnosis to education. While recent large language model (LLM) benchmarks emphasize in-visit diagnostic reasoning, they fail to evaluate models' ability to support patients after the visit. We introduce DischargeSim, a novel benchmark that evaluates LLMs on their ability to act as personalized discharge educators. DischargeSim simulates post-visit, multi-turn conversations between LLM-driven DoctorAgents and PatientAgents with diverse psychosocial profiles (e.g., health literacy, education, emotion). Interactions are structured across six clinically grounded discharge topics and assessed along three axes: (1) dialogue quality via automatic and LLM-as-judge evaluation, (2) personalized document generation including free-text summaries and structured AHRQ checklists, and (3) patient comprehension through a downstream multiple-choice exam. Experiments across 18 LLMs reveal significant gaps in discharge education capability, with performance varying widely across patient profiles. Notably, model size does not always yield better education outcomes, highlighting trade-offs in strategy use and content prioritization. DischargeSim offers a first step toward benchmarking LLMs in post-visit clinical education and promoting equitable, personalized patient support. [1].

1 Introduction

Discharge communication plays a crucial role in ensuring patient safety and promoting long-term recovery. Upon leaving the hospital, patients must understand a range of instructions, including medications, follow-up appointments, return precautions, and post-discharge care. Yet studies show that 40% to 80% of patients forget or misunderstand this information shortly after discharge (Kessels, 2003; Richard et al., 2017; Federman et al., 2018). Miscommunication at this stage contributes to poor adherence, increased readmissions, and preventable complications (Zhao et al., 2018; Weerahandi et al., 2018).

While recent clinical NLP benchmarks, such as AgentClinic (Schmidgall et al., 2024), AMIE (Tu et al., 2025), and HealthBench (Arora et al., 2025), focus on diagnostic reasoning during visits, they largely overlook the discharge phase, a stage that prioritizes education over inference. Some recent work (Cai et al., 2023) tried to address discharge understanding but is limited to single-turn QA, lacking dialogue structure, patient diversity, and downstream evaluation.

To fill this gap, we introduce DischargeSim, a comprehensive benchmark that evaluates large language models (LLMs) in the role of discharge educators. As shown in Figure 1, DischargeSim simulates multi-turn conversations between a DoctorAgent and a PatientAgent, structured around six clinically validated discharge topics: diagnosis, tests and treatments, return-to-hospital indicators, medication, post-discharge treatment, and follow-up. Each topic functions as a stage-specific education goal. The DoctorAgent must not only deliver accurate information but also apply appropriate strategies, such as simplification, emotional support, and shared decision-making, while adapting to the PatientAgent's psychosocial profile (e.g., health literacy, education, emotional state).

Specifically, DischargeSim supports three interconnected evaluation layers:

1. **Dialogue quality:** assessed using an LLM-as-Judge framework across dimensions of language and delivery, and human-centered communication. We further include evaluations of content completeness and strategy use, mea-

[*]indicates equal contribution
[1]The source code is released at: https://github.com/michaels6060/DischargeSim with CC-BY-NC 4.0 license.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10784–10810
November 4-9, 2025

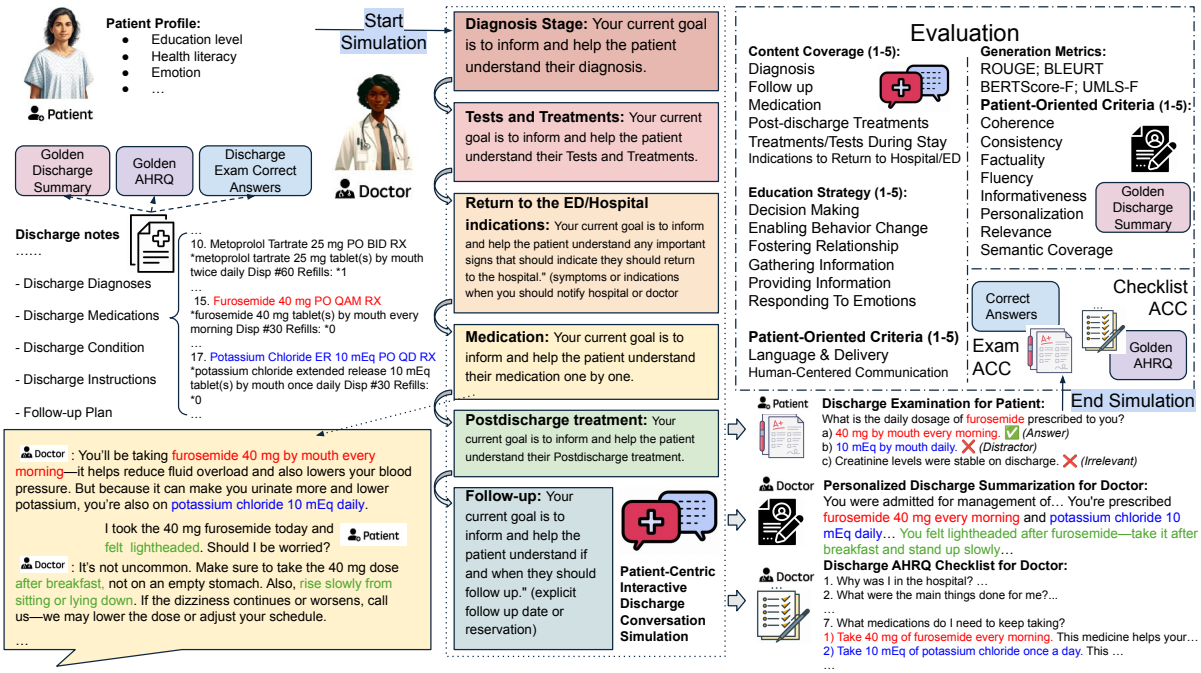

Figure 1: **DischargeSim Overview.** This figure illustrates the full workflow of DischargeSim, a simulation framework for evaluating post-visit doctor-patient discharge conversations. The simulation begins with a *PatientAgent* defined by its education level, health literacy, and emotional state. The conversation progresses through six content stages: Diagnosis, Tests and Treatments, Return-to-Hospital Indicators, Medication, Post-discharge Treatment, and Follow-up, each with stage-specific education goals. The central panel illustrates example turns where the DoctorAgent employs communication strategies such as simplification and emotional support. The right panel depicts several evaluation components that enable comprehensive simulation and evaluation of LLM-based discharge communication.

suring whether the DoctorAgent delivers comprehensive information and applies appropriate communication strategies tailored to the patient profile.

2. **Discharge summary generation:** evaluated using lexical metrics (ROUGE, BLEURT), medical metrics (UMLS-F1), and LLM-based criteria including fluency, personalization, and factuality.

3. **Patient comprehension:** measured through expert-authored multiple-choice exams based on either the conversation or the generated summary.

We evaluate 18 LLMs using DischargeSim and find that while larger models often produce more fluent responses, they do not consistently enhance patient comprehension, particularly when interacting with patients who have low health literacy or avoidant communication styles. Interestingly, some medium-sized models strike a better balance between strategy use and content completeness. Our key contributions include:

1. introducing DischargeSim, the first benchmark for personalized discharge education through structured, multi-turn simulations;

2. proposing a clinically grounded, three-stage evaluation framework covering dialogue quality, summary generation, and patient comprehension;

3. benchmarking 18 LLMs and revealing unexpected behavioral trends such as strategy overuse and profile-sensitive performance gaps;

4. presenting the first systematic analysis of how patient characteristics (e.g., health literacy, education, emotional state) affect model performance in discharge phase simulations, highlighting challenges in achieving equitable clinical AI communication.

DischargeSim thus offers a novel testbed for advancing personalized, context-aware patient educa-

tion beyond diagnostic reasoning.

2 Methodology

Our framework, DischargeSim, simulates the discharge process through a multi-agent interaction system that models post-visit patient–provider conversations. Unlike traditional diagnostic benchmarks which use static, single-turn QA formats, DischargeSim treats discharge communication as a sequential decision-making and educational task. The goal is not to infer a diagnosis, but to verify and enhance a patient's understanding of discharge instructions, including medications, follow-up care, activity restrictions, and warning signs.

2.1 PatientAgent Profiles

We configure PatientAgents with controlled variations in: (1) Health Literacy — low (limited health knowledge) vs. high (informed self-care); (2) Education Level — from no high school to bachelor's degree; (3) Emotional Style — neutral, anxious, or deflective (avoidant) behavior. These settings enable systematic evaluation of LLM adaptability, reasoning, and communication style across diverse patient profiles. In addition, each PatientAgent is initialized from a Gold dataset that comprises 49 discharge notes sampled from the MIMIC-IV database (Johnson et al., 2023)[2]. For each note, domain experts manually created between 5 and 10 multiple-choice questions and answers. The instructions and the detailed procedure for QA generation are included in the appendix A.

2.2 Data Preprocessing and Scenario Construction

We utilize raw clinical data (free-text discharge notes and associated documentation) from the MIMIC-IV dataset. Each discharge note is parsed into a structured schema inspired by the OSCE framework (Schmidgall et al., 2024) and AHRQ discharge guidelines (Agency for Healthcare Research and Quality, 2020), extracting: 1) Patient history and clinical course; 2) In-hospital test results and treatments; 3) Discharge diagnoses and medication plans; 4) Follow-up and return-to-hospital instructions. To simulate patient memory constraints, we generate a parallel structured discharge representation that can be partially masked in PatientAgent profiles. The DoctorAgent has full access to the structured summary throughout the conversation.

Medical Content Stages DischargeSim dialogues are structured into six content stages (DeSai et al., 2021): 1) Return-to-Hospital Indicators: signs/symptoms that the patient should be aware of when they should contact or return to the hospital/Emergency Department. 2) Post-discharge Medications: the medications that the patient takes post-discharge. 3) Diagnosis Information: the chief complaint of the patient, the main and sub diagnoses of the patient. This should be in Unified Medical Language System (UMLS) vocabulary. 4) Activity Restrictions: what kind of actions or activities the patient should or should not be doing post-discharge. 5) In-hospital Treatments/Tests: what type of treatments/tests were done during their stay, and what the results were. 6) Follow-up Plans: when and where the patient should follow up the patient's health issues post-discharge. Each stage can span multiple turns and is initiated by the DoctorAgent.

Conversational Strategies DoctorAgents employ six strategies (King and Hoppe, 2013): 1) Fostering relationship: build rapport and connection, respect patient statements, privacy, autonomy, engage in partnership building. Express caring and commitment. Use appropriate language. Encourage patient participation. Show interest in the patient as a person. 2) Gathering information: attempt to understand the patient's needs for the encounter. Elicit full description of major reason for visit from biological and physiological perspectives. Ask open-ended questions. Allow patient to complete responses. Listen actively. Elicit patient's full set of concerns. Elicit patient's perspective on the problem/illness. Explore full effect of the illness. Clarify information. Inquire about additional concerns. 3) Providing information: seek to understand patient's informational needs. Share information. Overcome barriers to patient understanding. Facilitate understanding. Explain nature of the problem and approach to diagnosis, treatment. Give uncomplicated explanations and instructions. Avoid jargon and complexity. Encourage questions and check understanding. Emphasize key messages. 4) Decision making: outline collaborative action plan. Identify and enlist resources and support. Discuss follow-up and plan for unexpected outcomes. 5) Enabling disease and treatment-related behavior: assess patient's inter-

[2]Because this is MIMIC-IV data, only credentialed users can contact us to obtain the annotated Gold dataset. See https://physionet.org/content/mimiciv/3.1/.

est in and capacity for self-management. Provide advice (information needs, coping skills, strategies for success). Agree on next steps. Assist patient to optimize autonomy and self-management. Arrange for needed support. Advocate for, and assist patient with, health system navigation. Assess patient's readiness to change health behaviors. Elicit goals, ideas, and decisions. 6) Responding to emotions: facilitate patient expression of emotional consequences of illness. Acknowledge and explore emotions. Express empathy, sympathy, and reassurance. Provide help in dealing with emotions. Assess psychological distress.

2.3 Simulation Workflow

In DischargeSim, the discharge conversation is modeled as a multi-turn, multi-stage simulation between a DoctorAgent and a PatientAgent. Unlike traditional diagnosis-focused benchmarks that rely on single-turn QA formats, our framework reproduces post-visit clinical education in a structured and dynamic manner. This process comprises two core stages: interactive simulation and output generation.

Interactive Simulation Phase. DoctorAgent leads a multi-turn conversation with the PatientAgent across six clinically critical content stages (as defined in § 2.2): return-to-hospital indicators, post-discharge medications, diagnosis information, activity restrictions, in-hospital treatments/tests, and follow-up plans. Each stage is initiated sequentially by the DoctorAgent and capped at 5 dialogue turns—except for the medications stage, which has no upper limit, to ensure complete coverage of all prescribed drugs. Responses by PatientAgent are influenced by its profile (education, health literacy, emotional state), and DoctorAgent is expected to adapt its language style and conversational strategy accordingly, referencing six strategy categories derived from AHRQ and medical communication literature (see § 2.2). We explicitly monitor whether DoctorAgent appropriately selects and applies these strategies—such as empathy, simplification, or clarification—to improve communication outcomes.

Post-conversation Output Generation. After the dialogue, DoctorAgent is tasked with producing two personalized documents based on both the original discharge note and the full interaction history: 1) Personalized Discharge Summary: a free-text summary tailored not only to reflect accurate clinical information but also to emphasize content based on the PatientAgent's cognitive and emotional profile. This extends traditional discharge summarization (Cai et al., 2022) by incorporating patient-aware customization through conversational history. 2) Structured AHRQ-style Checklist: a JSON-formatted structured discharge plan, modeled after the AHRQ guide (Agency for Healthcare Research and Quality, 2020). This checklist organizes the six discharge content areas into a machine-readable summary, allowing standardized comparisons with expert-written references.

Comprehension Exam for PatientAgent. To assess patient education effectiveness, the PatientAgent takes a 10-question multiple-choice exam covering all six content categories. The questions are authored by clinical experts using the gold-standard discharge record and are designed to reflect the most essential information a patient should retain. We evaluate comprehension under two input conditions to evaluate the support that patients might receive from Dialogue and generated Discharge Summary: A) Dialogue Only: tests whether the conversation alone sufficiently conveys necessary information. B) Discharge Summary: mimics real-world scenarios where patients receive a written summary post-consultation.

3 Experimental Design

DischargeSim's evaluation framework focuses on three core components: (1) multi-turn dialogue quality, (2) personalized discharge summarization, and (3) patient comprehension outcomes.

3.1 Dialogue Evaluation

We adopt LLM-as-Judge, specifically using DeepSeek-V3 (Liu et al., 2024), to evaluate each conversation across two main dimensions: Language and Delivery, and Human-Centered Communication. For Language and Delivery, we assess whether the language used is simple, fluent, and appropriate for the patient's comprehension level (linguistic clarity), whether the response is logically structured and maintains a coherent, understandable flow (coherence), and whether it avoids unnecessary repetition that may hinder educational effectiveness (repetitiveness). For Human-Centered Communication, we evaluate the extent to which the response demonstrates empathy and personalization by considering the patient's emotional state, health literacy, and context (personalization and

empathy), as well as how appropriately the system responds to the patient's prior utterances, maintaining a natural and responsive conversational flow (interaction appropriateness). Check Table 14 for more details.

3.2 Discharge Summarization Evaluation

After each interaction, the DoctorAgent generates a personalized discharge summary, which is evaluated using both automatic metrics and LLM-based judgment. We report ROUGE (Lin, 2004) to assess lexical overlap, BLEURT (Sellam et al., 2020) to measure fluency, and UMLS-F1 to quantify clinical concept coverage. In addition, we employ DeepSeek-V3-Judge to evaluate each summary across eight dimensions: semantic coverage (COV), which assesses how well the generated summary captures key semantic units from the reference; factuality (FAC), which checks whether the summary maintains factual accuracy without hallucination; consistency (CON), which ensures internal coherence; informativeness (INF), which measures how concisely key ideas are conveyed; coherence (COH), which evaluates logical flow and readability; relevance (REL), which verifies that all content pertains to discharge instructions and follow-up care; fluency (FLU), which assesses grammatical and stylistic quality; and personalization (PER), which examines how well the summary is tailored to the patient's literacy, emotional state, and specific clinical context. Check Table 17 and 18 for more details.

3.3 Validation of LLM-as-Judge Metrics

To assess the reliability of our automatic evaluation framework, we conducted a human–LLM agreement study on 49 simulated discharge cases. Two licensed clinicians (one nurse and one physician) independently provided pairwise preferences between two nurse-agent outputs per case along five criteria: (i) Language Delivery, (ii) Human-Centered Communication, (iii) Content Coverage, (iv) Education Strategy, and (v) Overall Preference. Experts selected a preferred output and highlighted supporting evidence spans (green for positive, red for negative), which we used for qualitative error analysis.

Human Inter-rater Reliability. Agreement between the two experts was high: Cohen's $\kappa = 0.715$, percent agreement $= 85.7\%$, and Spearman's $\rho = 0.720$ ($p < 0.001$), indicating consistent expert judgments.

Alignment with LLM-as-Judge. We compared aggregated LLM-as-Judge ratings with expert preferences and observed substantial alignment: Cohen's $\kappa = 0.591$, percent agreement $= 79.6\%$, and Spearman's $\rho = 0.593$ ($p < 0.001$). These results support the use of LLM-as-Judge as a scalable proxy for expert assessment in discharge communication tasks.

3.4 Models

We evaluate 18 LLMs, spanning commercial, open-source, and biomedical-specific variants: (1) OpenAI Series: GPT-3.5-turbo, GPT-4o [3], GPT-4.1 (incl. nano and mini), and GPT-5 (incl. nano and mini) (2) Qwen 2.5 Series: 0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B (3) LLaMA3 Series: LLaMA3.1-8B/70B, LLaMA3.2-1B/3B, LLaMA3.3-70B All models operate under unified prompting structures and profile settings aligned with DischargeSim's agent framework.

4 Results

We evaluate 18 large language models on both discharge conversation (Table 1) and discharge summary generation tasks (Table 2), spanning language fluency, human-centeredness, medical coverage, strategy use, and downstream patient comprehension.

4.1 Language & Communication.

We observe clear scaling trends. The highest Language & Delivery / Human-Centered Communication scores appear in LLaMA3.3-70B (4.13 / 4.26) and GPT-5-mini (4.07 / 4.28), with Qwen2.5-72B (4.03 / 4.26) and GPT-4.1-mini (3.92 / 4.27) close behind. GPT-5-nano is also strong (3.73 / 4.01), though slightly below the top tier. In contrast, Qwen2.5-0.5B is markedly lower (1.47 / 1.50), reflecting limited fluency and engagement.

4.2 Content Coverage.

Top models deliver broad and consistent coverage across discharge topics. The best average coverage scores are achieved by GPT-5-mini (4.96), GPT-5-nano (4.82), GPT-4.1-mini (4.68), LLaMA3.1-70B (4.64), and LLaMA3.3-70B (4.62). On key categories, several models reach high marks on MED and DX—for example, GPT-4.1-mini MED 4.96 / DX 4.86 and LLaMA3.3-70B MED 4.94

[3] We don't include gpt-4o-mini-as-nurse because we use it as patient agents in our experiments.

Model	Language Delivery	Human Comm	Content Coverage							Education Strategy							Exam ACC	AHRQ ACC
			IRH	MED	DX	PDT	TDS	FU	avg.	FR	GI	PI	DM	EBC	RE	avg.		
qwen2.5-0.5b	1.47	1.5	2.79	1.66	2.44	2.74	1.4	2.38	2.23	2.24	1.83	2.53	1.91	2.31	1.74	2.1	52.87	4.18
qwen2.5-1.5b	2.67	2.63	3.3	2.04	4.24	2.99	2.36	2.98	2.98	3.11	2.45	3.58	2.76	3.14	2.27	2.88	62.54	13.01
qwen2.5-3b	3.09	3.38	4.05	2.56	4.47	3.57	3.12	3.92	3.62	3.55	2.63	3.9	3.2	3.64	2.87	3.3	64.95	20.25
qwen2.5-7b	3.7	3.94	4.09	3.2	4.35	3.76	3.43	3.89	3.79	3.76	2.9	4.07	3.38	3.71	2.89	3.45	66.77	31.95
qwen2.5-14b	3.51	3.8	4.31	2.99	4.49	3.74	3.57	3.91	3.83	4.0	3.09	4.21	3.54	3.91	3.26	3.66	66.47	31.55
qwen2.5-32b	3.8	4.03	4.23	3.26	4.59	3.86	3.75	4.0	3.95	4.01	3.31	4.33	3.73	3.99	3.15	3.75	69.79	23.23
qwen2.5-72b	4.03	4.26	4.27	4.01	4.77	4.29	4.35	4.24	4.32	4.16	3.49	4.43	3.89	4.16	3.5	3.93	74.02	55.47
llama3.2-1b	2.64	2.78	2.98	3.11	3.59	3.06	2.82	3.51	3.18	3.48	2.88	3.6	3.24	3.43	3.03	3.28	69.70	25.13
llama3.2-3b	3.12	3.47	4.08	4.17	4.60	4.06	3.93	4.49	4.21	3.63	3.38	4.08	3.65	3.93	3.06	3.62	71.00	51.05
llama3.1-8b	3.32	3.66	4.36	4.83	4.81	4.51	4.18	4.67	4.56	3.61	3.17	4.08	3.58	3.86	2.86	3.52	77.34	77.50
llama3.1-70b	3.69	3.94	4.48	4.99	4.84	4.50	4.24	4.83	4.64	3.92	3.35	4.30	3.79	4.02	3.11	3.74	84.29	86.87
llama3.3-70b	4.13	4.26	4.46	4.94	4.83	4.42	4.37	4.69	4.62	3.97	3.54	4.38	3.88	4.11	3.17	3.83	82.18	86.51
gpt3.5	3.59	3.72	4.11	3.18	4.14	3.39	2.82	3.69	3.56	3.62	2.88	4.02	3.30	3.65	2.74	3.37	61.03	32.30
gpt4o	3.98	4.16	4.15	3.91	4.61	3.93	3.56	4.06	4.03	4.01	3.43	4.36	3.76	4.04	3.13	3.78	71.90	58.22
gpt4.1-nano	3.48	3.80	4.37	2.58	4.17	3.62	2.80	3.63	3.53	4.01	3.27	4.25	3.67	3.99	3.19	3.72	61.03	26.90
gpt4.1-mini	3.92	4.27	4.60	4.96	4.86	4.62	4.39	4.66	4.68	4.24	3.50	4.54	3.94	4.27	3.62	4.01	82.78	87.43
gpt5-nano	3.73	4.01	4.93	4.75	4.91	4.90	4.49	4.95	4.82	3.78	3.18	4.35	3.70	4.08	2.82	3.65	79.15	73.80
gpt5-mini	4.07	4.28	4.98	4.97	4.99	4.99	4.85	4.99	4.96	3.77	3.06	4.38	3.64	4.03	2.82	3.62	84.89	93.80

Table 1: Evaluation of simulated discharge conversations across multiple dimensions: **(1) Language & Delivery** assesses linguistic clarity, coherence, and avoidance of repetitiveness. **(2) Human-Centered Communication** evaluates personalization, empathy, and interaction appropriateness. (Note: detailed scores for (1) and (2) are reported in the Appendix.) **(3) Content Coverage** measures how well the conversation addresses key discharge information, including Indications to Return to Hospital/ED (IRH), Medication Information (MED), Diagnosis (DX), Post-Discharge Treatments (PDT), Treatments/Tests During Stay (TDS), and Follow-Up Instructions (FU). The average of these is reported as **Content Coverage avg.**. **(4) Education Strategy** evaluates the use of effective communication strategies: Fostering Relationship (FR), Gathering Information (GI), Providing Information (PI), Decision Making (DM), Enabling Behavior Change (EBC), and Responding to Emotions (RE). The average is reported as **Strategy avg.**. **(5) Exam ACC** measures the Patient Agent's accuracy on comprehension questions when only the simulated conversation is used as input. **(6) AHRQ ACC** compares the Doctor Agent-generated AHRQ checklist against the ground truth checklist based on the conversation. Higher scores indicate better performance across all metrics.

/ DX 4.83—while the GPT-5 series approaches ceiling across most items (e.g., GPT-5-mini IRH 4.98 / MED 4.97 / DX 4.99 / FU 4.99).

4.3 Education Strategy Use.

Aggregate strategy use peaks with GPT-4.1-mini (strategy avg. 4.01), followed by Qwen2.5-72B (3.93), LLaMA3.3-70B (3.83), and LLaMA3.1-70B (3.74). The GPT-5 models are slightly lower on aggregate (mini 3.62; nano 3.65), with relatively weaker scores in Responding to Emotions and Gathering Information (e.g., mini RE 2.82, GI 3.06). Lower-tier models underuse Responding to Emotions and Enabling Behavior Change (e.g., Qwen2.5-0.5B RE 2.10, EBC 2.31), indicating challenges in nuanced patient communication.

4.4 Patient Comprehension Outcomes (Conversations).

Exam accuracy (PatientAgent using only the conversation) rises with model quality: GPT-5-mini 84.89%, LLaMA3.1-70B 84.29%, GPT-4.1-mini 82.78%, LLaMA3.3-70B 82.18%, with GPT-5-nano also high at 79.15%. In the Qwen series, accuracy generally increases with size (52.87% → 74.02%), with a slight dip at 14B (66.47%). AHRQ checklist accuracy shows a similar trend, topping out at GPT-5-mini 93.80% and strong open-source results for LLaMA3.1-70B (86.87%) and LLaMA3.3-70B (86.51%).

4.5 Discharge Summary Generation.

For summaries, LLaMA3.1-70B leads the lexical/clinical overlap metrics (ROUGE-L 0.3419, UMLS-F 0.4251). GPT-5-mini attains the highest LLM-judged quality and factuality (L&A avg. 4.19; F&C avg. 4.84) and the best downstream comprehension (83.38%), with GPT-5-nano also strong (L&A avg. 4.03; F&C avg. 4.61; 77.04%). LLaMA3.3-70B is close on judged factuality/completeness (avg. 4.50), and GPT-4.1-mini shows excellent judged quality (L&A avg. 4.04; F&C avg. 4.72) despite a lower ROUGE-L (0.2653). Lower-capacity models, such as Qwen2.5-3B and LLaMA3.2-1B, yield lower summary-only comprehension (63–64%).

4.6 Overall Ranking and Trends.

Among open-source models, LLaMA3.1-70B and LLaMA3.3-70B provide the strongest overall bal-

Model	Generation Metrics			Language & Appropriateness					Factuality & Completeness					Exam
	RougeL	BLEURT	UMLS-F	FL	CO	INF	PER	avg.	SC	FAC	REL	CON	avg.	ACC
qwen2.5-0.5b	0.1530	0.3163	0.1524	3.09	2.74	2.79	2.65	2.82	1.96	1.63	2.93	1.59	2.02	52.87%
qwen2.5-1.5b	0.1152	0.3188	0.1764	3.28	2.97	2.86	2.87	3.00	2.48	2.54	3.56	2.60	2.79	57.10%
qwen2.5-3b	0.1897	0.3608	0.2409	3.73	3.73	3.58	3.50	3.64	3.11	2.94	4.27	3.26	3.39	63.44%
qwen2.5-7b	0.2018	0.3718	0.2834	3.94	3.87	3.50	3.50	3.70	3.16	2.93	4.43	3.18	3.41	63.14%
qwen2.5-14b	0.1868	0.3880	0.2767	4.02	3.90	3.51	3.54	3.74	3.24	3.50	4.73	3.76	3.80	62.84%
qwen2.5-32b	0.1830	0.3980	0.2674	3.98	3.89	3.48	3.51	3.72	3.28	3.62	4.71	3.75	3.83	64.35%
qwen2.5-72b	0.2545	0.4089	0.3709	4.04	3.97	3.62	3.57	3.80	3.77	4.06	4.86	4.22	4.22	73.41%
llama3.2-1b	0.2059	0.3325	0.1909	3.51	3.01	3.28	3.19	3.25	2.24	1.88	3.19	1.79	2.27	63.64%
llama3.2-3b	0.2225	0.4025	0.3166	3.76	3.53	3.63	3.60	3.63	3.49	3.68	4.80	3.93	3.96	65.26%
llama3.1-8b	0.2741	0.4285	0.3818	3.98	3.80	3.80	3.67	3.81	4.02	4.30	4.98	4.36	4.41	76.44%
llama3.1-70b	0.3419	0.4535	0.4251	4.09	3.92	3.86	3.69	3.89	4.33	4.62	5.00	4.66	4.65	82.48%
llama3.3-70b	0.2865	0.4463	0.4198	4.01	3.91	3.68	3.61	3.80	4.05	4.46	4.99	4.53	4.50	80.36%
gpt3.5	0.1504	0.3874	0.2790	3.92	3.74	3.29	3.36	3.58	2.99	3.33	4.40	3.40	3.52	59.52%
gpt4o	0.2076	0.4149	0.3598	4.01	3.93	3.56	3.54	3.76	3.55	4.12	4.83	4.18	4.16	67.98%
gpt4.1-nano	0.1299	0.3738	0.2322	4.01	3.92	3.52	3.52	3.74	3.09	3.33	4.51	3.55	3.61	56.19%
gpt4.1-mini	0.2653	0.4445	0.3954	4.18	4.07	4.02	3.88	4.04	4.27	4.82	4.98	4.81	4.72	79.76%
gpt5-nano	0.2341	0.4076	0.3638	4.12	4.08	4.00	3.91	4.03	4.24	4.56	4.98	4.66	4.61	77.04%
gpt5-mini	0.2451	0.3919	0.3663	4.24	4.08	4.31	4.13	4.19	4.48	4.97	4.99	4.93	4.84	83.38%

Table 2: Evaluation of LLM-generated discharge summaries. The table reports results across three metric categories: (1) **Traditional Generation Evaluation Metrics** including ROUGE-L, BLEURT, UMLS-F, and Exam Accuracy; (2) LLM-Judge Evaluation Criteria for **Language Quality & Appropriateness** covering Fluency (FL), Coherence (CO), Informativeness (INF), and Personalization (PER); and (3) LLM-Judge Evaluation Criteria for **Evidence-Based Factuality & Completeness** covering Semantic Coverage (SC), Factuality (FAC), Relevance (REL), and Consistency (CON). We also report Exam ACC, which measures the accuracy of the Patient Agent on comprehension questions when only the discharge summary is provided as input. Higher scores indicate better performance.

ance across conversation and summary tasks. Qwen2.5-72B is competitive on language/human scores (4.03 / 4.26) but trails on coverage (avg. 4.32), strategy (3.93), and conversation exam accuracy (74.02%) relative to GPT-4.1-mini (4.68 / 4.01 / 82.78%). GPT-5-mini generally sets the upper bound on coverage, exam, and AHRQ in conversations (4.96 / 84.89% / 93.80%) and leads judged quality/factuality in summaries (4.19 / 4.84), while GPT-5-nano offers a competitive but slightly lower profile.

5 Discussion

5.1 Impact of Patient Health Literacy

To investigate how patient health literacy affects dialogue agent behavior, we compare model responses in two scenarios: HL1 (interacting with patients of low health literacy) and HL2 (interacting with patients of high health literacy). Figure 4 visualizes the relative difference across six evaluation dimensions, computed as $(\text{HL2} - \text{HL1})/\text{HL1}$.

Overall, high-literacy patients (HL2) consistently elicit better responses across most models and dimensions. Improvements are especially pronounced in Language and Delivery, Human-Centered Communication, and Strategy, indicating that models are more capable of producing fluent, coherent, and empathetic responses when provided with clearer, more structured patient input. In contrast, factuality, completeness, language, and Appropriateness remain stable, indicating robustness to patient comprehension levels.

As shown in Figure 7, these expressive dimensions also exhibit the highest variability across models, indicating that they are more sensitive to the user's literacy level. We further investigate this effect within the Qwen model family (excluding the outlier Qwen-0.5B). Figure 8 shows a strong positive correlation between model size and the average performance gap between HL2 and HL1 ($r = 0.80$, $p = 0.053$), suggesting that larger models are better at tailoring their responses to high-literacy patients. This effect is most significant in Language and Delivery ($r = 0.97$, $p < 0.01$) and Human-Centered Communication ($r = 0.86$, $p < 0.05$), as shown in Figure 9. These findings underscore the need for future research to enhance model robustness in low-literacy settings. In particular, methods to enhance strategy generation and communication clarity should be prioritized to ensure equitable support for diverse patient populations.

5.2 Impact of Patient Education Level

To investigate how patient education level affects dialogue agent performance, we compare responses

under three conditions: Education 1 (no high school education), Education 2 (high school/GED), and Education 3 (college graduate). For each model and evaluation dimension, we compute the relative score difference (e.g., $(\text{Edu2} - \text{Edu1})/\text{Edu1}$) to quantify sensitivity to patient educational background.

We observe a consistent improvement in agent performance as patient education level increases. Compared to Education 1, both Education 2 and Education 3 users elicit notably better responses across expressive dimensions, such as Strategy, Human-Centered Communication, and Language and Delivery. These improvements suggest that higher education levels, which likely correlate with more precise phrasing and more structured conversational behavior, enable agents to deliver more coherent and empathetic responses. In contrast, dimensions such as Factuality, Completeness, Language, and Appropriateness exhibit minimal variation, indicating that these aspects are more dependent on the model's internal knowledge and less influenced by the quality of user input. Figure 5 shows the relative differences across all models for three education level comparisons.

We also measure the variability of these differences across models for each dimension. As shown in Figure 10, expressive and interaction-heavy aspects such as Strategy and Content demonstrate the highest variation across models, while factual and linguistic dimensions remain more stable. Focusing on the Qwen model family (excluding the 0.5B outlier), we find significant correlations between model size and education sensitivity across dimensions. Larger models, such as Qwen-14B and Qwen-72B, are more responsive to users with higher education, especially in expressive dimensions like Language and Delivery, and Human-Centered Communication (Figure 11).

These findings emphasize the importance of education-aware agent design. While large models are generally more capable of tailoring responses to educated users, performance equity must be ensured for users with limited formal education. Future work should explore controllable generation techniques and robust fine-tuning strategies to enable high-quality communication regardless of educational background.

5.3 Impact of Patient Emotional Style

To assess how patient emotional style influences dialogue agent performance, we analyze responses under three simulated conditions: deflective, neutral, and anxious. For each model and evaluation dimension, we compute the relative score difference (e.g., (anxious - deflective)/deflective) to quantify the variation in agent responses across emotional inputs.

Overall, we observe that anxious patients consistently elicit more supportive and strategic responses, particularly in dimensions such as Human-Centered Communication, Language and Delivery, and Strategy. This suggests that LLM-based agents often adapt their tone and conversational strategies when confronted with emotionally charged user behavior. By contrast, dimensions such as Language, Appropriateness, Factuality, and Completeness remain relatively stable across emotional styles, indicating that factual correctness and linguistic formality are governed more by the model's internal representations than by emotional cues from the patient. The trend across models is visualized in Figure 6.

We further compute the standard deviation of these score differences across models to assess robustness. As shown in Figure 12, Language and Delivery and Human-Centered Communication exhibit the highest variability, suggesting they are most sensitive to patient emotional style. In contrast, Factuality and Appropriateness remain stable, reinforcing their input-agnostic nature. Zooming in on the Qwen model family (excluding the 0.5B outlier), we find that larger models are more robust and less reactive to variations in emotional style. In the Anxious vs. Neutral comparison, model size is strongly negatively correlated with performance sensitivity in expressive dimensions: Strategy ($r = -0.91$), Language and Appropriateness ($r = -0.73$), and Factuality and Completeness ($r = -0.64$), as shown in Figure 13.

This suggests that small models tend to overreact to emotional input, while larger models exhibit more calibrated responses. In the Neutral vs. Deflective contrast, correlations turn moderately positive in Human-Centered Communication ($r = 0.45$) and Content ($r = 0.59$), indicating that larger models are more effective at recognizing and adapting to avoidant conversational behavior.

6 Related Work

Effective discharge communication is a well-established determinant of patient adherence and recovery, yet it remains underexplored in NLP bench-

marks (Yao and Yu, 2025). Prior studies have highlighted that patients, particularly those with low health literacy or emotional distress, often struggle to retain critical discharge information (Richard et al., 2017; Weerahandi et al., 2018). Although clinical best practices such as teach-back, simplified language, and checklist-based summaries (e.g., AHRQ's "Going Home" guide) are known to enhance patient understanding, these strategies are seldom modeled or evaluated in current LLM-based systems. Studies consistently indicate that between 40% and 80% of patients forget or misunderstand critical discharge instructions shortly after hospital visits (Kessels, 2003; Jack et al., 2009), underscoring the need for communication-centered evaluation.

Existing clinical dialogue benchmarks, such as AgentClinic (Schmidgall et al., 2024), AMIE (Tu et al., 2025), and HealthBench (Arora et al., 2025), primarily focus on in-visit diagnostic conversations, emphasizing uncertainty reasoning, evidence integration, and stepwise clinical inference. Benchmarks like MedAgents (Tang et al., 2023) further simulate multimodal physician–patient interactions for sequential decision-making. However, these works leave the crucial post-visit discharge phase unaddressed, where the primary goal shifts from diagnostic accuracy to communication clarity and patient-centered education. Similarly, domain-specific medical LLM systems such as ChatDoctor (Li et al., 2023), NoteChat (Wang et al., 2023), PMC-LLaMA (Wu et al., 2024), MedAlpaca (Han et al., 2023), BioInstruct (Tran et al., 2024a), and Huatuogpt (Zhang et al., 2023) have primarily focused on diagnostic QA, knowledge recall, and retrieval-augmented dialogue, typically evaluating factual correctness or reasoning accuracy in single-turn settings. In contrast, DischargeSim targets multi-turn, post-visit educational conversations grounded in patient-specific profiles (literacy, emotion, comprehension level), using downstream comprehension testing as a core evaluation signal, dimensions underrepresented in prior medical AI evaluation frameworks.

While PaniniQA (Cai et al., 2023) marks initial progress by evaluating patient comprehension through QA on discharge summaries, it remains limited to static, single-turn question-answer formats and overlooks crucial discharge tasks such as personalized summary generation and structured AHRQ checklist completion. Dialogue-based educational interventions have long shown advantages in enhancing comprehension and learning outcomes, particularly among populations with lower health literacy (Whitehurst, 2002; Golinkoff et al., 2019). However, NLP approaches to dialogue modeling and educational QA (Du and Cardie, 2017; Shwartz et al., 2020) have yet to be systematically applied to discharge education in a way that accounts for patient diversity and dynamic adaptation. DischargeSim systematically integrates multi-turn, personalized educational dialogues into clinical discharge settings, extending beyond QA to encompass personalized summary generation and checklist-based instruction, while explicitly modeling psychosocial and cognitive patient factors.

In addition, inspired by advances in LLM-as-Judge research, DischargeSim incorporates tailored evaluation rubrics for distinct patient-centered communication tasks. In the broader NLP community, LLMs such as GPT-4 (Achiam et al., 2023; Liu et al., 2023; Fu et al., 2023) and critique-tuned variants (Ke et al., 2023) have been applied as automated evaluators for summarization (Chen et al., 2023), dialogue (Zheng et al., 2024; Zhang et al., 2024), and translation (Kocmi and Federmann, 2023). This paradigm has recently been extended to the medical domain for evaluating clinical conversations (Tu et al., 2025; Arora et al., 2025), medical documentation (Croxford et al., 2025; Chung et al., 2025; Brake and Schaaf, 2024), exam QA (Yao et al., 2024a,b), and reasoning (Jeong et al., 2024; Tran et al., 2024b). DischargeSim builds upon this body of work by offering a structured, task-specific judging framework tailored to discharge communication, enabling robust, scalable, and context-aware evaluation of LLMs in patient education.

7 Conclusion

DischargeSim presents a simulation framework for evaluating LLMs in personalized, post-visit patient education. Our results show significant variation in performance across communication strategies, content coverage, and comprehension support, reflecting sensitivity to patient literacy, education, and emotional state. This work lays the foundation for more equitable, reliable, and patient-centered AI systems in clinical communication.

8 Limitations

While DischargeSim introduces a novel framework for evaluating LLMs in post-visit patient education, it has several limitations.

First, our PatientAgent simulation relies on scripted profiles and LLM-generated responses, which, despite being carefully constructed, may not fully capture the variability and unpredictability of real-world patient behavior. Incorporating human-in-the-loop evaluations or standardized patient actors could yield more realistic interactions.

Second, although we evaluate over 18 LLMs, our benchmark focuses primarily on English-language discharge scenarios derived from the MIMIC-IV dataset. The generalizability of our findings to other languages, health systems, and discharge formats (e.g., pediatrics, oncology) remains an open question.

Third, our evaluation pipeline, while comprehensive, still relies on LLM-as-Judge assessments and multiple-choice comprehension tests. While these offer scalable and interpretable insights, they may miss nuanced qualitative aspects of patient understanding and trust. Future work should integrate real patient feedback and long-term behavioral outcomes.

Finally, our current framework emphasizes structured discharge topics and may not fully account for free-form or emotion-driven interactions that commonly occur during discharge discussions. Our benchmark currently structures conversations around six predefined discharge topics derived from widely recognized clinical guidelines (diagnosis, medications, post-discharge activity, return indicators, in-hospital tests, and follow-up). While this ensures consistent coverage of essential safety information, it may constrain the natural, patient-driven flow of real-world discharge dialogues. Actual interactions often evolve dynamically, with patients steering the conversation toward personally salient concerns. In future iterations, we plan to incorporate flexible topic prioritization and adaptive dialogue management strategies that preserve critical safety coverage while enabling models to respond more fluidly to individual patient cues and evolving needs.

9 Ethics Statement

This study presents a simulation framework designed to evaluate and improve the patient education capabilities of large language models in discharge communication scenarios. All data used in this study are derived from the publicly available MIMIC-IV dataset, which is fully de-identified and widely adopted in clinical NLP research. No personally identifiable information is included.

The comprehension exam dataset was created through a two-step process. First, multiple-choice questions were drafted by PhD students in Computer Science based in the United States, with each questionnaire containing 5–10 questions comprising an answer, a distractor, and an irrelevant option. These questions were then reviewed and revised by three medical experts, including two nursing professors from the United States and one physician from South Korea, who validated the quality of the questions and provided comments or edits to ensure medical accuracy and clarity.

All medical experts were compensated for their time at a rate of $40 per hour, in line with academic ethics standards. The annotation and review process was conducted under an approved academic protocol to ensure transparency and quality control.

DischargeSim is intended solely for research and evaluation purposes and has not been clinically validated for real-world deployment. Given the sensitivity of discharge communication, we caution against using LLM-based systems in clinical settings without expert oversight. We strongly advocate for continued research into the safety, fairness, and interpretability of AI systems in healthcare to support equitable and responsible integration into clinical workflows.

Acknowledgments

This material is the result of work supported with resources and the use of facilities at the Center for Healthcare Organization and Implementation Research, VA Bedford Health Care.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agency for Healthcare Research and Quality. 2020. Going home: Questions to ask after the hospital. Accessed: 2025-05-20.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Nathan Brake and Thomas Schaaf. 2024. Comparing two model designs for clinical note generation; is an llm a useful evaluator of consistency? *arXiv preprint arXiv:2404.06503*.

Pengshan Cai, Fei Liu, Adarsha Bajracharya, Joe Sills, Alok Kapoor, Weisong Liu, Dan Berlowitz, David Levy, Richeek Pradhan, and Hong Yu. 2022. Generation of patient after-visit summaries to support physicians. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6234–6247.

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

Hong Chen, Duc Minh Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2023. Storyer: Automatic story evaluation via ranking, rating and reasoning. *Journal of Natural Language Processing*, 30(1):243–249.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, et al. 2025. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. 2025. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*, pages 2025–04.

Charisma DeSai, Keri Janowiak, Beatrice Secheli, Eleanor Phelps, Sam McDonald, Gary Reed, and Andra Blomkalns. 2021. Empowering patients: simplifying discharge instructions. *BMJ Open Quality*, 10(3). Publisher: British Medical Journal Publishing Group.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2067–2073.

Alex Federman, Erin Sarzynski, Cindy Brach, Paul Francaviglia, Jessica Jacques, Lina Jandorf, Angela Sanchez Munoz, Michael Wolf, and Joseph Kannry. 2018. Challenges optimizing the after visit summary. *International journal of medical informatics*, 120:14–19.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Roberta Michnick Golinkoff, Erika Hoff, Meredith L Rowe, Catherine S Tamis-LeMonda, and Kathy Hirsh-Pasek. 2019. Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child development*, 90(3):985–992.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Brian W Jack, Veerappa K Chetty, David Anthony, Jeffrey L Greenwald, Gail M Sanchez, Anna E Johnson, Shaula R Forsythe, Julie K O'Donnell, Michael K Paasche-Orlow, Christopher Manasseh, et al. 2009. A reengineered hospital discharge program to decrease rehospitalization: a randomized trial. *Annals of internal medicine*, 150(3):178–187.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1. Number: 1 Publisher: Nature Publishing Group.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. 2023. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation. *arXiv preprint arXiv:2311.18702*.

Roy PC Kessels. 2003. Patients' memory for medical information. *Journal of the royal society of medicine*, 96(5):219–222.

Ann King and Ruth B. Hoppe. 2013. "Best Practice" for Patient-Centered Communication: A Narrative Review. *Journal of Graduate Medical Education*, 5(3):385–393.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Claude Richard, Emma Glaser, and Marie-Thérèse Lussier. 2017. Communication and patient participation influencing patient recall of treatment discussions. *Health Expectations*, 20(4):760–770.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024a. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832.

Hieu Tran, Zonghai Yao, Junda Wang, Yifan Zhang, Zhichao Yang, and Hong Yu. 2024b. Rare: Retrieval-augmented reasoning enhancement for large language models. *arXiv preprint arXiv:2412.02830*.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2023. Notechat: a dataset of synthetic doctor-patient conversations conditioned on clinical notes. *arXiv preprint arXiv:2310.15959*.

Himali Weerahandi, Boback Ziaeian, Robert L Fogerty, Grace Y Jenq, and Leora I Horwitz. 2018. Predictors for patients understanding reason for hospitalization. *Plos one*, 13(4):e0196479.

Grover J Whitehurst. 2002. Dialogic reading: An effective way to read aloud with young children. *Reading Rockets*, pages 1–7.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. 2024a. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. *arXiv preprint arXiv:2410.13191*.

Zonghai Yao and Hong Yu. 2025. A survey on llm-based multi-agent ai hospital.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, et al. 2024b. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553*.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19515–19524.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Jane Y Zhao, Buer Song, Edwin Anand, Diane Schwartz, Mandip Panesar, Gretchen P Jackson, and Peter L Elkin. 2018. Barriers, facilitators, and solutions to optimal patient portal and personal health record use: a systematic review of the literature. In *AMIA annual symposium proceedings*, volume 2017, page 1913.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Data Creation

We took two steps to generate and evaluate the quality of Comprehension Exam questions. First, we asked students to annotate 5-10 questionnaires. Then, 3 experts went through the generated datasets and evaluated, commented on, or modified the questionnaires. The students were all PhD students majoring in Computer Science in the United States. The expert annotators were 2 nurse professors in the United States and 1 doctor from South Korea. For each question, we instructed the annotators to generate 5 to 10 multiple-choice questions with three choices: answer, distractor, and irrelevant, as shown in Figure 2. We asked three medical experts to go through the questions and questionnaires to validate the quality of the annotations. We asked them to modify or leave comments and made the changes according to their comments (Figure 3).

B Factuality Metrics: UMLS-F1

The assessment of factual accuracy in LLM outputs leverages the UMLS concept overlap metric. The Unified Medical Language System (UMLS), established by Bodenreider (Bodenreider, 2004), significantly contributes to interoperability in the biomedical domain. It aggregates and distributes a wide array of biomedical terminologies, classification systems, and coding standards from diverse sources, thus helping to reconcile semantic inconsistencies and representational differences across various biomedical concept repositories.

To identify and align medical named entities within texts to their corresponding UMLS concepts, we utilize the `scispaCy` library (Neumann et al., 2019), which is well-suited for biomedical entity recognition. This enables us to accurately map named entities in LLM-generated outputs to the relevant UMLS concepts—a crucial capability for evaluating factual accuracy.

We quantify factuality using precision, recall, and F1 score between the concept set of the reference summary (C_{ref}) and the generated summary (C_{gen}). Specifically:

$$\text{Recall} = \frac{|C_{\text{ref}} \cap C_{\text{gen}}|}{|C_{\text{ref}}|},$$
$$\text{Precision} = \frac{|C_{\text{ref}} \cap C_{\text{gen}}|}{|C_{\text{gen}}|}.$$

The F1 score is computed as the harmonic mean of precision and recall, providing a balanced measure of both correctness and coverage in the LLM output.

C Hardware Settings & Compute Time

For closed-source models (e.g., GPT-4.1, GPT-4o, GPT-3.5, DeepSeek-V3), we use their respective API endpoints with default parameters for inference. For all open-source models—including the full LLaMA3 and Qwen2.5 families—we conduct inference locally using two NVIDIA A100 80GB GPUs. The experiments are run on a server equipped with an Intel(R) Xeon Gold 6226R CPU @ 2.90GHz.

Guidelines for Annotation

1. You are going to create 5-10 questions for each discharge note.
2. These questions are going to be clinically "relevant" and also important for the patient.
3. What is concerned "relevant" is as follows :
i) It has to be acknowledged in the discharge note
ii) It has to be concerned with the current health issues for that particular stays
iii) It has to be concerned with instructions from the medical doctor
iv) The categories that you could consider. The questions could be asked from in such categories :
Diagnosis during hospital stay
Procedure(interventions/tests) during hospital stay
Medication during hospital stay
Diagnosis in discharge
Procedure(follow up/tests/interventions) after discharge
Medication after discharge

Example questions :
Q. Why were you admitted to the hospital?
Q. What is the medication that the doctor recommended you to take?
Q. To treat your <illness/symptom> what drug did the doctor prescribe you?
Q. During your stay, the staff found you had <illness/symptom>. What was the name of that illness?
Q. The Doctor warns about your danger of <illness/symptom>. What kind of treatment/intervention did he recommend?
Q. What was your diagnosis during your stay?
Q. What is the cause of your symptoms?
Q. What is the correct dose of Gabapentin?
Q. What is the purpose of taking Benzonatate 100 mg three times a day as needed for cough?
Q. What procedure was performed during your hospital stay?
Q. What is the dosage of Lantus at night?

4. What is NOT considered "relevant" is as follows :
i) It does not appear in the discharge notes and cannot be inferred from the discharge notes
ii) If it has less issues with the current health state of the patient or if it's something that happened in the past that does not affect current health related concerns

5. How to comprise the choices i) you will come up with 3 choices for each questions ii) each choices will be either answer, distractor and irrelevant choice iii) distractor can be defined as something similar to the answer that causes confusion but not the actual answer that the question is looking for. E.g. distractors that are opposite to the answer would be one example. iv) irrelevant choice should be something that is bizarre, out of context. It should appear in the discharge note, but a totally irrelevant answer to the question.

Figure 2: Guidelines for initial questionnaire generation for Q

Guidelines for Annotation

1. You are going to evaluate 5-10 questions for each discharge note.

2. These questions are going to be clinically "relevant" and also important for the patient.

3. What is concerned "relevant" is as follows :
i) It has to be acknowledged in the discharge note
ii) It has to be concerned with the current health issues for that particular stays
iii) It has to be concerned with instructions from the medical doctor
iv) The categories that you could consider. The questions could be asked from in such categories :
Diagnosis during hospital stay
Procedure(interventions/tests) during hospital stay
Medication during hospital stay
Diagnosis in discharge
Procedure(follow up/tests/interventions) after discharge
Medication after discharge

4. How to
i) If you think the question is okay, please check relevant.
ii) If you consider that the question itself needs to be totally removed or changed please check irrelevant.
iii) if you consider the question is okay but needs some modification please check modify and leave a comment below how we should change the questions
iv) if you checked irrelevant or modify please write what should be changed and guidance on how to fix the text or the question.

Figure 3: Guidelines for questionnaire modification for Q

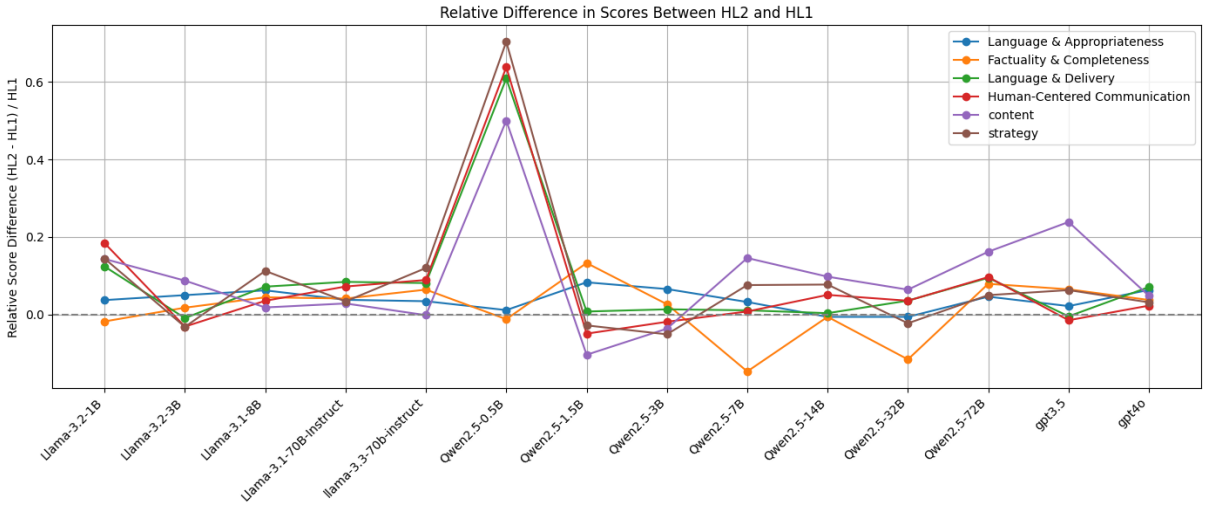

Figure 4: Relative score difference between HL2 and HL1 across all models and evaluation dimensions.

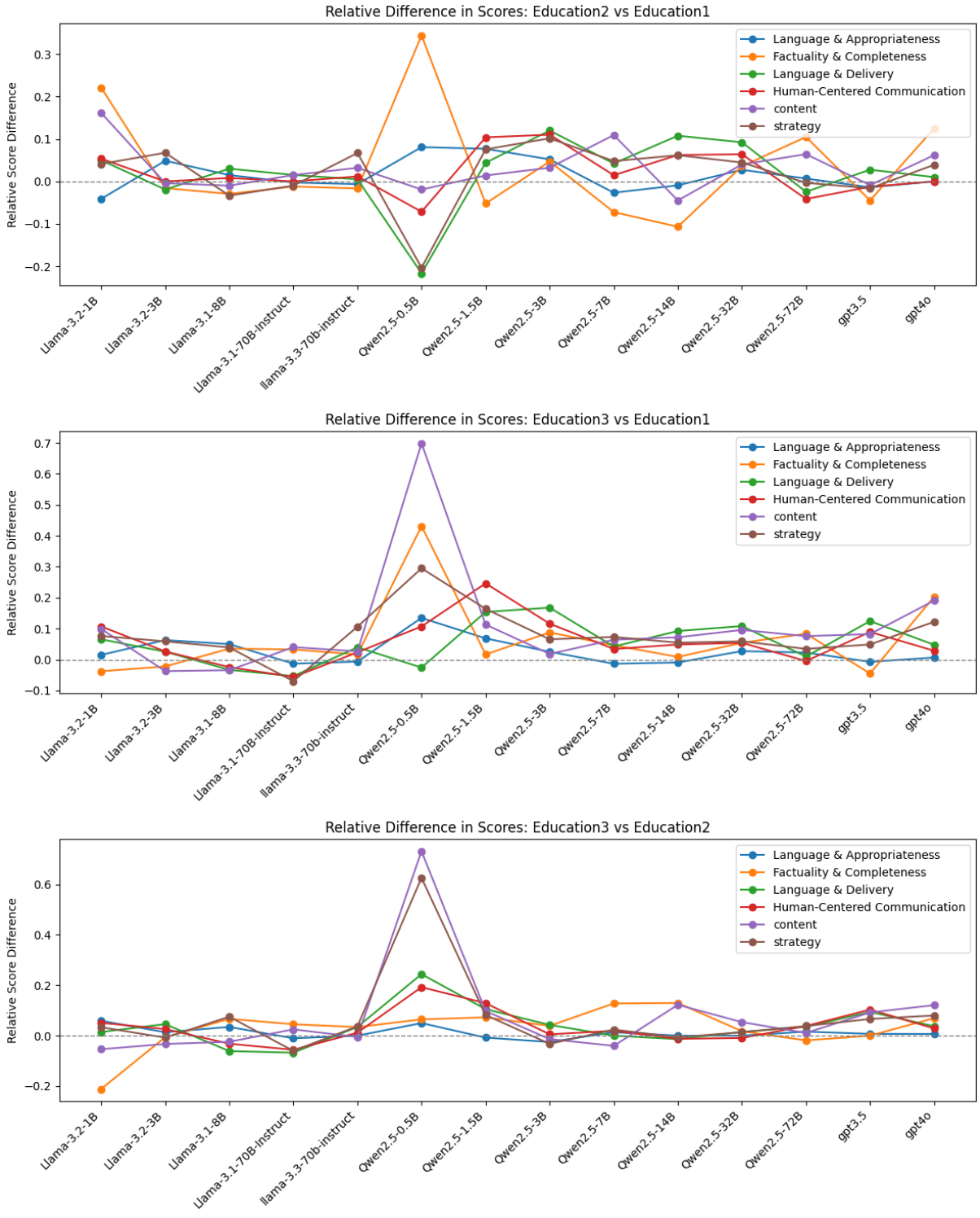

Figure 5: Relative score difference across education levels for each model and evaluation dimension. From top to bottom: Education2 vs. Education1, Education3 vs. Education1, and Education3 vs. Education2.

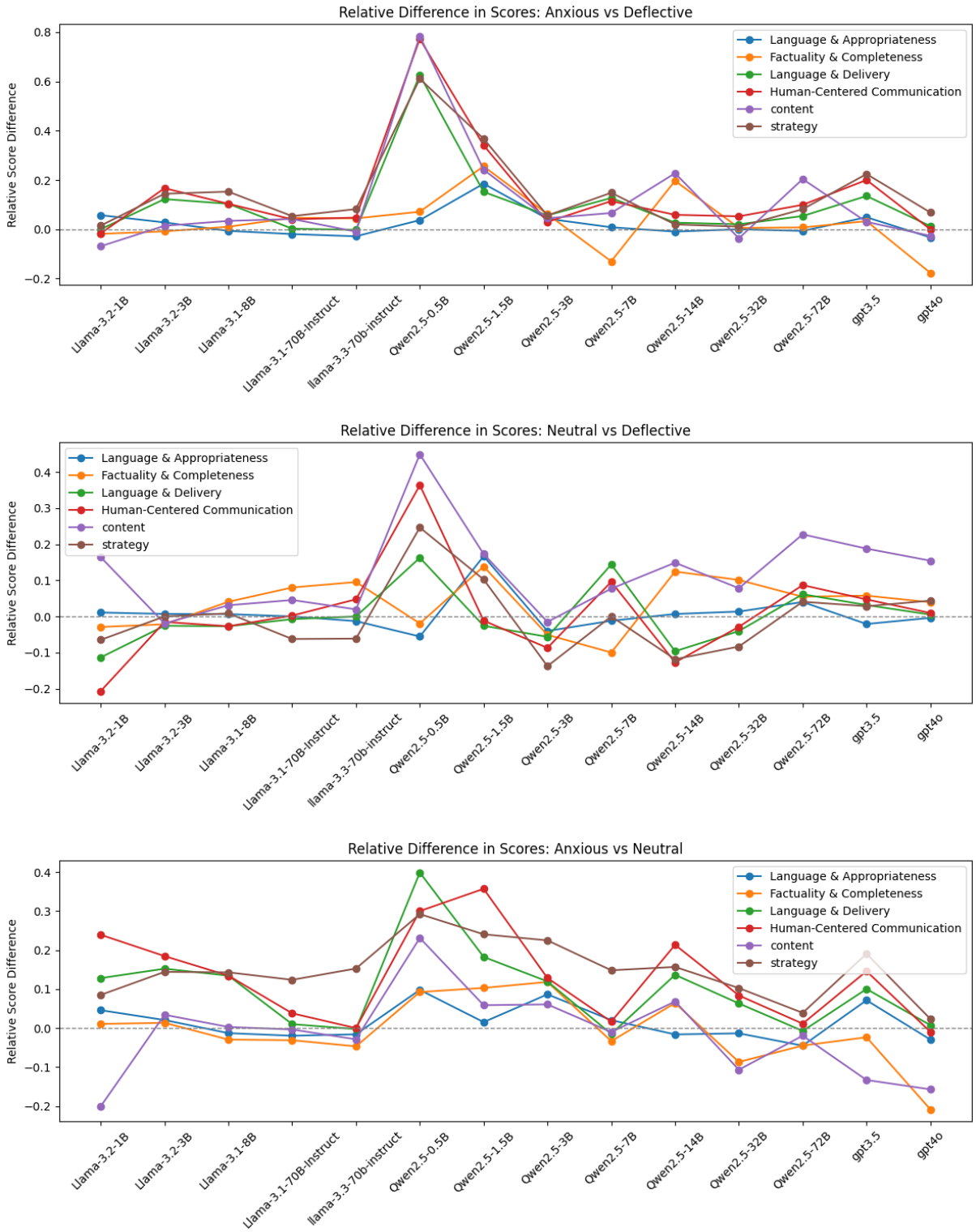

Figure 6: Relative score differences across models by emotion style. Top: Anxious vs. Deflective; Middle: Neutral vs. Deflective; Bottom: Anxious vs. Neutral.

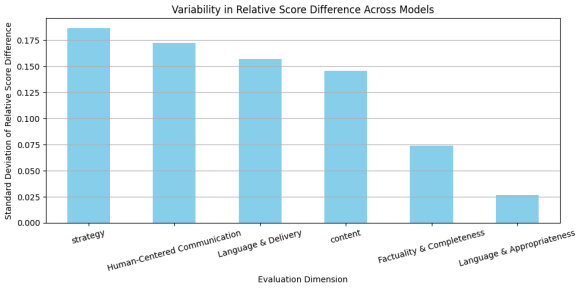

Figure 7: Standard deviation of relative difference across models for each evaluation dimension.

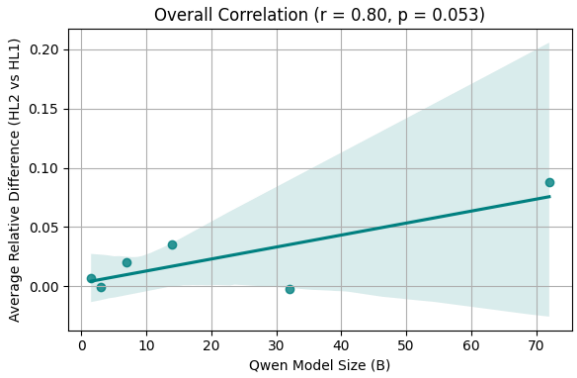

Figure 8: Correlation between Qwen model size and average HL2 vs. HL1 difference.

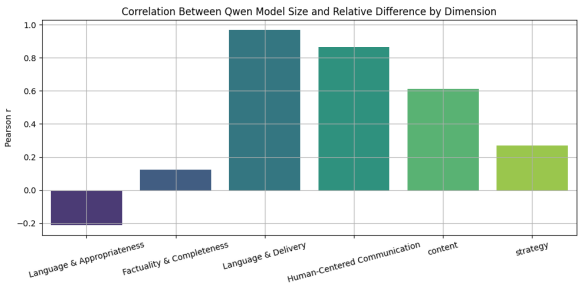

Figure 9: Correlation between Qwen model size and HL2-vs-HL1 relative difference across dimensions.

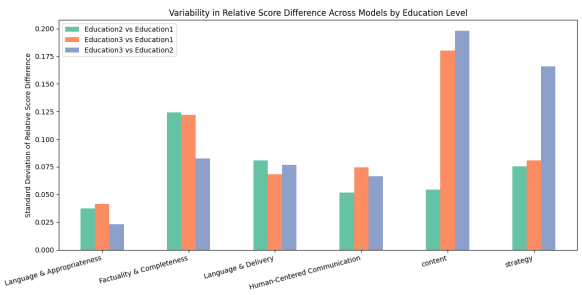

Figure 10: Standard deviation of relative score differences across models by evaluation dimension. Dimensions such as *Strategy* and *Content* show the highest variability across patient education levels.

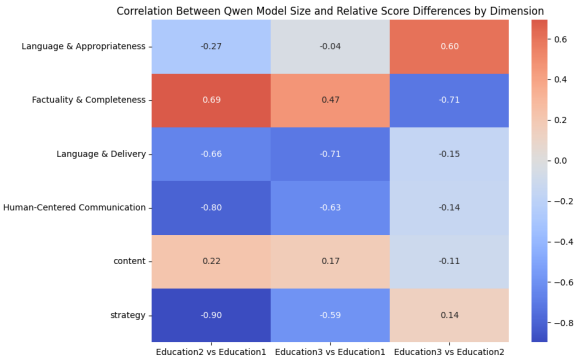

Figure 11: Pearson correlation (r) between Qwen model size and relative score difference, across evaluation dimensions and education level comparisons. Negative correlation indicates higher performance boost with increasing education level in larger models.

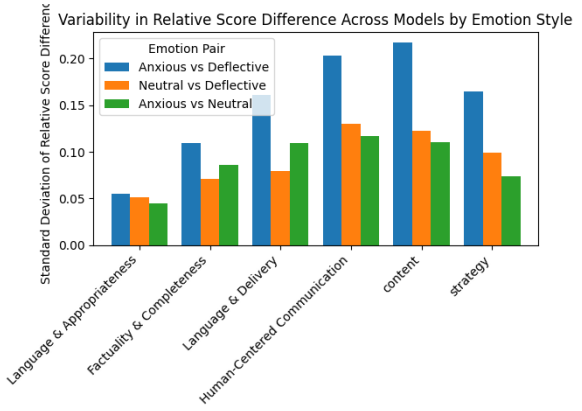

Figure 12: Variability (standard deviation) in relative score difference across models by emotion style and dimension.

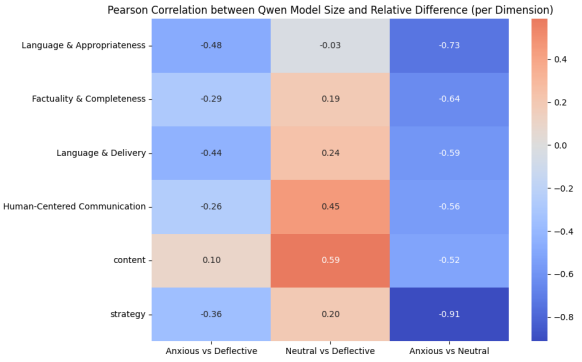

Figure 13: Pearson correlation between Qwen model size and emotional sensitivity, computed as score difference per dimension.

Language Quality and Patient-Centeredness Evaluation Criteria

System Prompt: You are an expert evaluator trained to assess the *language quality* and *patient-centeredness* of doctor responses in clinical discharge conversations.
You will be provided with:

- A **Doctor–Patient Conversation** transcript
- A **Ground Truth Discharge Note**, which contains the verified medical facts, diagnoses, and discharge instructions for the case

STRICT SCORING INSTRUCTIONS
Be **strict and conservative** in your evaluation:

- **Do not assign scores higher than 3.8** unless the language and communication quality are *clearly exceptional and consistent across the entire conversation*
- A score of **4.0 or above should be extremely rare**, and only permitted if all responses are:
  - Free from vague or overly general phrasing
  - Free from repetition or template-like delivery
  - Highly empathetic, clear, and personalized to the patient's situation
- **Most conversations should score between 2.5 and 3.8**
- Penalize even subtle weaknesses:
  - Redundant phrases
  - Overuse of generic reassurances
  - Slightly confusing transitions or non-specific encouragement

Language & Delivery

Linguistic Clarity: Is the response fluent, readable, and easy to understand for the patient's stated education level? Penalize convoluted syntax, awkward phrasing, or vague language.

Coherence: Does the conversation maintain logical structure and flow from turn to turn? Penalize digressions, abrupt transitions, or unclear logical progression.

Repetitiveness: Does the response avoid repeating phrases, ideas, or sentence templates? Penalize overuse of phrases like "Take it one day at a time" or "Please reach out if needed" unless clearly personalized.

Human-Centered Communication

Personalization & Empathy: Does the doctor show awareness of the patient's emotional state, literacy, and personal concerns? Penalize robotic tone, missing emotional cues, or excessive use of generic affirmations.

Interaction Appropriateness: Does the doctor's response clearly address each patient concern in a natural, relevant way? Penalize vague deflections or educational monologues that ignore patient questions.

Criteria & Scoring

- **5.0 = Perfect:** Every sentence is linguistically fluent, empathetic, and natural; fully tailored to the patient's literacy, tone, and needs. *Extremely rare.*
- **4.0–4.9 = Exceptional:** Very strong across all criteria; minimal imperfections; still *not generic*.
- **3.0–3.9 = Competent:** Generally human-like; minor repetition or vagueness; tone sometimes robotic; most model outputs fall here if decent.
- **2.0–2.9 = Weak:** Multiple coherence/clarity flaws; generic language; emotionally disengaged.
- **1.0–1.9 = Very Poor:** Disorganized, vague, or robotic; major patient questions missed or deflected.

Output Format
For each criterion, return:

- A score between 1.0 and 5.0 (decimals allowed)
- A brief justification (1–2 sentences)

Return your output in this JSON format:

```
{
  "Linguistic Clarity": {"score": ..., "justification": "..."},
  "Coherence": {"score": ..., "justification": "..."},
  "Repetitiveness": {"score": ..., "justification": "..."},
  "Language & Delivery score": ... // average of the 3 criteria
}

{
  "Personalization & Empathy": {"score": ..., "justification": "..."},
  "Interaction Appropriateness": {"score": ..., "justification": "..."},
  "Human-Centered Communication score": ... // average of the 2 criteria
}
```

Figure 14: Strict evaluation rubric for language quality and patient-centeredness in discharge conversations.

Content Quality Evaluation of Doctor Responses

System Prompt

You are an expert evaluator trained to assess the **content quality** of doctor responses in clinical discharge conversations.
You will be provided with:

- A **Doctor–Patient Conversation** transcript
- A **Ground Truth Discharge Note**, which contains the verified medical facts, diagnoses, and discharge instructions for the case

STRICT SCORING INSTRUCTIONS

- **Do not assign scores higher than 3.8** unless the coverage, clarity, and personalization are **exceptional and consistent**.
- Scores of **4.0 or above should be extremely rare**, and only allowed if:
  - The topic is addressed **accurately, clearly, and with detail**
  - **No key item is omitted**
  - **No vague or generic phrasing** is present
  - The response is **highly personalized**, not template-like
- **Most conversations should score between 2.5 and 3.8**
- Penalize subtle issues:
  - Mentioning a topic without explanation
  - Using vague language (e.g., "take care")
  - Partial or superficial content
  - Very short conversations that skip essentials

Criteria & Format

You will evaluate **six clinical criteria**. For each, provide:

- A **score** (1.0–5.0, decimals allowed)
- A **justification** (more than 10 words)
- A list of **evidence spans**:
  - Conversation snippet
  - Whether it addresses the criterion
  - Optional `issue` if omitted, vague, or hallucinated

Criteria Definitions

1. Indications to Return to Hospital/ED — Are specific red flags discussed (e.g., chest pain, fever)?
2. Medication Information — Are medication names, doses, and purposes clearly explained?
3. Diagnosis — Are chief complaint and final diagnoses clearly stated and contextualized?
4. Post-discharge Treatments — Are home care instructions, activity restrictions, or wound care clearly discussed?
5. Treatments/Tests During Stay — Are tests/treatments and outcomes during the stay described?
6. Follow-Up — Are follow-up provider, timing, and reason specified?
Output Format (JSON)

```
{
  "Indications to Return to Hospital": {
    "score": ...,
    "justification": "Red flags vaguely mentioned; lacked detail and personalization.",
    "evidence": [
      {"conversation": "Come back if anything gets worse", "addresses": true},
      {"conversation": "N/A", "addresses": false, "issue": "Omitted clear symptoms"}
    ]
  },
  ...
  "score": ...
}
```

Figure 15: Strict content quality evaluation prompt used to assess discharge conversations across six core clinical criteria.

Conversation Strategy Evaluation Guidelines

You are an expert evaluator trained to assess the *conversation strategy* of doctor responses in clinical discharge conversations.
You will be provided with:

- A **Doctor–Patient Conversation** transcript
- A **Ground Truth Discharge Note**, which contains the verified medical facts, diagnoses, and discharge instructions for the case

STRICT SCORING INSTRUCTIONS

- Be **strict and conservative** in your evaluation
- **Do not assign scores higher than 3.8** unless the conversation strategy is **exceptional and consistent** throughout the entire dialogue
- A score of **4.0 or above** should be **extremely rare**, and only permitted if:
  - The agent shows **clear empathy and rapport**
  - Questions are open-ended and invite patient participation
  - Responses are **tailored, emotionally sensitive, and supportive**
  - There is **evidence of shared decision making** or support for patient self-management
- **Most conversations should score between 2.5 and 3.8**
- Penalize even subtle weaknesses:
  - Responses feel generic, overly scripted, or impersonal
  - Patient emotions are ignored or dismissed
  - No attempt to verify patient understanding or autonomy
  - Missed opportunity to clarify or encourage participation

Strategy Dimensions
You will rate each conversation across six key communication strategy criteria:
For each criterion, provide:

- A **score from 1.0 to 5.0** (decimals allowed)
- A **justification** (more than 10 words)
- A list of **evidence spans**, each with:
  - the conversation snippet
  - a flag `addresses` (true/false)
  - an optional `issue` field for missing, vague, or problematic behavior

1. Fostering Relationship: Build rapport and connection, respect patient statements, privacy, and autonomy. Engage in partnership building. Express caring and commitment. Use appropriate language. Encourage patient participation. Show interest in the patient as a person.

- 5.0 = Consistent, warm, personalized interaction and active partnership-building
- 4.0–4.9 = Mostly caring and respectful with some generic language
- 3.0–3.9 = Neutral tone, minor signs of engagement
- < 3.0 = Cold, impersonal, or disrespectful tone

2. Gathering Information: Attempt to understand the patient's needs. Ask open-ended questions. Listen actively. Elicit concerns and perspectives. Clarify unclear information. Explore the full effect of illness.

- 5.0 = Multiple open-ended questions, follow-ups, and attentive listening
- 4.0–4.9 = Some engagement with patient perspective, mostly on-topic
- 3.0–3.9 = Minimal elicitation of patient input
- < 3.0 = No active effort to explore patient views

3. Providing Information: Understand the patient's informational needs. Share clear, jargon-free explanations. Facilitate understanding. Check comprehension and emphasize key messages.

- 5.0 = Clear, empathetic, responsive explanations
- 4.0–4.9 = Mostly clear, some assumptions about understanding
- 3.0–3.9 = Some helpful info, but generic or not well explained
- < 3.0 = Vague, unhelpful, or overly complex

4. Decision Making: Outline collaborative action plan. Discuss follow-up and plan for unexpected outcomes. Enlist support and check agreement.

- 5.0 = Joint decision-making, clear plan, checks patient agreement
- 4.0–4.9 = Presents options or plan, some patient involvement
- 3.0–3.9 = States plan without patient input
- < 3.0 = No mention of planning or collaboration

5. Enabling Behavior Change: Assess readiness for self-management. Provide coping strategies. Encourage autonomy. Arrange support.

- 5.0 = Encourages change with specific strategies and supports
- 4.0–4.9 = Some support for self-management
- 3.0–3.9 = Generic advice or no discussion of behavior
- < 3.0 = No guidance or autonomy support

6. Responding to Emotions: Explore emotional consequences of illness. Express empathy. Reassure or assist with emotional distress.

- 5.0 = Recognizes emotional tone, responds with empathy and care
- 4.0–4.9 = Attempts reassurance or brief acknowledgment
- 3.0–3.9 = Missed opportunities for emotional connection
- < 3.0 = No response to expressed or implied emotion

Output Format (JSON)

```
{
  "Fostering Relationship": {
    "score": ...,
    "justification": "...",
    "evidence": [
      {"conversation": "...", "addresses": true},
      {"conversation": "...", "addresses": false, "issue": "..."}
    ]
  },
  ...
  "score": ... // average of the 6 criteria
}
```

Figure 16: Evaluation protocol for assessing conversation strategies in discharge dialogue settings.

Discharge Summary Group A: Subjective Evaluation: Language Quality & Appropriateness

] You are an expert evaluator trained to assess the high-level writing quality and general appropriateness of AI-generated clinical discharge summaries. Your task is to evaluate the summary by reading it holistically and assigning scores for the following four subjective criteria.
Be strict and conservative:

- Do not give scores higher than 4.0 unless the performance is clearly exceptional.
- Most summaries should score between 2.5 and 3.8 unless outstanding.

For each of the following criteria, return:

- A score between 1.0 and 5.0 (decimals allowed)
- A brief justification (1–2 sentences)

Criteria:

1. **Fluency** – Is the summary grammatically correct, natural in tone, and professionally written?
2. **Coherence** – Is the structure of the summary logical and easy to follow?
3. **Informativeness** – Does the summary cover key points without being verbose or vague?
4. **Personalization** – Does the summary reflect the patient's specific context, including their diagnosis and background?

Return your output in the following JSON format:

```
{
  "Fluency": {"score": ..., "justification": "..."},
  "Coherence": {"score": ..., "justification": "..."},
  "Informativeness": {"score": ..., "justification": "..."},
  "Personalization": {"score": ..., "justification": "..."},
  "score": ... // average of the 4 scores
}
```

Figure 17: Evaluation criteria for quality of generated discharge summaries. System Prompt for Group A: Subjective Evaluation: Language Quality & Appropriateness.

Discharge Summary Group B: Objective Evaluation: Evidence-Based Factuality & Completeness

You are an expert evaluator trained to assess the **faithfulness and completeness** of AI-generated clinical discharge summaries by comparing them to:

1. A human-written reference summary
2. A doctor–patient conversation transcript

General Instructions

- Extract as much evidence as possible from the generated summary.
- Do not stop at a few examples — go line by line.
- Penalize sparse evidence when the reference contains more.
- If a score exceeds 4.0, justification must cite specific evidence.

Criteria:

1. **Semantic Coverage** — Are the major content units in the reference summary covered in the generated summary?
2. **Factuality** — Are all claims accurate and supported by the reference or conversation?
3. **Relevance** — Is all information directly related to discharge or follow-up?
4. **Consistency** — Is the summary free from contradictions with the source materials?

Output Format:

```
{
  "Semantic Coverage": {
    "score": ...,
    "justification": "... (more than 10 words)",
    "evidence": [
      {"summary": "...", "reference": "...", "match": true},
      {"summary": "...", "reference": "N/A", "match": false, "issue": "..."}
    ]
  },
  "Factuality": {
    "score": ...,
    "justification": "...",
    "evidence": [...]
  },
  "Relevance": {
    "score": ...,
    "justification": "...",
    "evidence": [...]
  },
  "Consistency": {
    "score": ...,
    "justification": "...",
    "evidence": [...]
  },
  "score": ...
}
```

Figure 18: Evaluation criteria for quality of generated discharge summaries. System Prompt for Group B: Objective Evaluation: Evidence-Based Factuality & Completeness.

Example Conversation — Qwen2.5-72B

Diagnosis Stage
[DOCTOR] Hello, Mr. ___. I want to go over your diagnosis and the next steps. You have L4-S1 spondylolisthesis, which means one of your vertebrae has slipped out of place. This has been causing you significant mechanical back pain. We've tried conservative treatments like physical therapy, injections, and medications, but these haven't provided enough relief. Given this, we've decided that a posterior lumbar interbody fusion and posterolateral fixation is the best option to stabilize your spine and reduce your pain. The surgery went well, and there were no complications. You've been doing well postoperatively, and your pain is now better managed. Do you have any questions about your diagnosis or the surgery?

Tests and Treatments Stage
[PATIENT] Thank you, Doctor. I understand that my vertebrae have slipped and that surgery was necessary. What should I expect during my recovery?
[DOCTOR] Great question, Mr. ___. During your recovery, it's important to follow the prescribed pain management regimen, including taking gabapentin, acetaminophen, and oxycodone as needed for pain. You'll also need to use a lumbar support brace to help stabilize your spine and prevent excessive movement. Physical therapy will be an important part of your recovery to help you regain strength and mobility. You should avoid heavy lifting and strenuous activities for several weeks.

Follow-up Stage
[DOCTOR] Regular follow-up appointments will be scheduled to monitor your progress. If you experience any unusual symptoms or have concerns, don't hesitate to contact us. How does that sound?
[PATIENT] That sounds manageable, Doctor. I'll make sure to follow the pain management plan and attend physical therapy. When should I schedule my first follow-up appointment?
[DOCTOR] That's great to hear, Mr. ___. For your first follow-up appointment, I recommend scheduling it about 2 weeks after your discharge. This will allow us to check on your incision site, assess your pain levels, and ensure that you're progressing well with your recovery. If you experience any severe pain, increased swelling, redness, or drainage from the incision site before your appointment, or if you have any other concerning symptoms, please contact us immediately. These could be signs that you need to return to the hospital. Does that sound good to you?
[PATIENT] That sounds good, Doctor. I'll make sure to schedule the follow-up for two weeks after discharge. Thank you for the information.

Medication Stage
[DOCTOR] You're welcome, Mr. ___. I'm glad you're feeling confident about the next steps. Now, let's go over the medications you'll be taking after discharge. It's important to understand each one and how to use them correctly. 1. **Gabapentin 400 mg PO TID**: This medication is used to manage nerve pain. You should take 400 mg three times a day, which means every 8 hours. It can help reduce the discomfort you might feel from your surgery. 2. **Acetaminophen 325-650 mg PO Q6H:PRN fever or pain**: This is for general pain and fever. You can take 325 to 650 mg every 6 hours as needed. However, do not exceed 4 grams (4000 mg) of acetaminophen in a 24-hour period to avoid liver damage. 3. **Diazepam
[PATIENT] Thank you, Doctor. I understand the medications and their purposes. What about the diazepam?
[DOCTOR] Of course, Mr. ___. **Diazepam 5 mg PO Q8H:PRN muscle spasm** is a muscle relaxant. You should take 5 mg every 8 hours as needed for muscle spasms. Muscle spasms can sometimes occur after surgery, and this medication can help alleviate them. Only take it when you feel a muscle spasm, and be aware that it can cause drowsiness, so avoid driving or operating heavy machinery while taking it. Next, we have **Docusate Sodium 100 mg PO BID**: This is a stool softener to help prevent constipation, which can be a side effect of the pain medications you're taking. Take 100 mg twice a day, once in the morning and once in the evening. Lastly, **OxycoDONE (Immediate Release) 5 mg PO Q4H:PRN pain**: This is a stronger pain medication for more severe pain. You can
[PATIENT] Thank you, Doctor. I understand the purpose of the diazepam and the stool softener. Can you tell me more about the oxycodone?
[DOCTOR] Certainly, Mr. ___. **OxycoDONE (Immediate Release) 5 mg PO Q4H:PRN pain** is a potent pain reliever that you can take every 4 hours as needed for severe pain. It's important to use this medication only when your pain is not adequately controlled by the other pain medications. Oxycodone can cause drowsiness, so you should avoid driving or operating heavy machinery while taking it. Additionally, if you feel overly sedated or drowsy, you should hold the dose and contact us for further instructions. Remember to always take it as directed and never exceed the recommended dosage. Do you have any other questions about the oxycodone or any of the other medications?

Return to the ED/Hospital indications Stage
[PATIENT] Thank you, Doctor. I understand how to use the oxycodone safely. Are there any specific signs I should watch for that might indicate a problem with my recovery?
[DOCTOR] Absolutely, Mr. ___. It's important to be aware of certain signs that might indicate a problem with your recovery. Here are the key things to watch for: 1. **Increased Pain**: If you experience severe or worsening pain that isn't relieved by your medications, this could be a sign of a complication. 2. **Fever**: A temperature of 100.4°F (38°C) or higher could indicate an infection. 3. **Redness, Swelling, or Drainage**: If you notice redness, swelling, warmth, or any unusual drainage from your incision site, this could be a sign of an infection. 4. **Numbness or Weakness**: If you experience new numbness, weakness, or loss of sensation in your legs, this could indicate a problem with the nerves. 5. **Difficulty Urinating or Bowel Movements**: If you have trouble urinating or if you become constipated despite using the stool softener,

---

Postdischarge treatment Stage
[PATIENT] Thank you for the information, Doctor. I will keep an eye out for those signs during my recovery. Is there anything else I should know before I go home?
[DOCTOR] You're welcome, Mr. ___. Just a few more points to keep in mind: 1. **Activity Level**: Start with light activities and gradually increase as you feel comfortable. Avoid heavy lifting, bending, or twisting for at least 6 weeks. Use the straight cane to help with balance and stability when walking. 2. **Follow-Up Appointments**: As we discussed, schedule your first follow-up appointment about 2 weeks after discharge. This will help us ensure that everything is healing properly and address any concerns you might have. 3. **Pain Management**: Continue to take your pain medications as prescribed. If you find that the pain is not well-controlled, don't hesitate to contact us. 4. **Physical Therapy**: Once cleared by your surgeon, start attending physical therapy sessions to help strengthen your back and improve your overall function. If you have any questions or concerns at any point, please don't hesitate to reach out to us. We're here to support you throughout your recovery.

Follow-up Stage
[PATIENT] Thank you, Doctor. I will follow your advice on activity levels and keep track of my pain management. I appreciate your support during my recovery.
[DOCTOR] You're very welcome, Mr. ___. I'm glad to hear that you're committed to following the plan. Remember, we're here to support you every step of the way. If you have any questions or concerns, no matter how small, don't hesitate to call us. Take care, and I look forward to seeing you at your follow-up appointment. Have a safe journey home.
[PATIENT] Thank you, Doctor. I appreciate your support and will reach out if I have any questions. See you at the follow-up appointment.
[DOCTOR] You're welcome, Mr. ___. I'm here for you, and I'll see you at your follow-up appointment in two weeks. Take care and have a smooth recovery. Safe travels home.
[PATIENT] Thank you, Doctor. I will take care and follow your instructions. See you soon.
[DOCTOR] You're very welcome, Mr. ___. I'm glad to hear that you're committed to following the instructions. I'll see you at your follow-up appointment, and I'm here for you whenever you need me. Have a safe trip home and a smooth recovery. Take care.

Figure 19: Example discharge conversation generated by Qwen2.5-72B.

Example Conversation — Qwen2.5-32B

Diagnosis Stage
[DOCTOR] Hello, Mr. ___. I wanted to make sure you understood that your back pain is due to spondylolisthesis at L4-S1, which means one of your vertebrae has slipped forward over the one below it. We've performed a posterior lumbar interbody fusion and posterolateral fixation at L5-S1 to address this issue. The procedure went well, and there are no complications noted with the hardware used. The most important thing for you now is to follow the post-operative care instructions carefully. This includes using your back brace as directed and gradually increasing your activity level under the guidance of your physical therapist. If you have any concerns about pain or recovery, don't hesitate to reach out to us.

Tests and Treatments Stage
[PATIENT] Thank you, Doctor. I understand that my vertebrae have slipped and that the surgery went well. What should I expect during my recovery?
[DOCTOR] Great question. During your recovery, you'll likely experience some discomfort and limited mobility initially. It's crucial to wear your back brace as prescribed to support your spine while it heals. You'll also start physical therapy to regain strength and flexibility in your back and legs. Expect gradual improvement over several weeks, but full recovery can take up to six months. Keep an eye on any unusual symptoms such as severe pain, numbness, or weakness, and report them to us immediately. Stick to the medication regimen we've provided to manage pain and prevent complications. How do you feel about starting your recovery process? Do you have any other questions or concerns?

Return to the ED/Hospital indications Stage
[PATIENT] I feel okay about starting the recovery process. I just want to make sure I'm doing everything right. What should I do if I experience any of those unusual symptoms?
[DOCTOR] It's great to hear that you're feeling ready to start your recovery. If you experience any unusual symptoms such as severe pain, numbness, or weakness, especially if these symptoms are worsening or spreading, it's very important that you contact us right away or go to the nearest emergency room. These could be signs of complications that need immediate attention. Also, if you notice any signs of infection around your incision site, such as increased redness, swelling, warmth, or drainage, seek medical care promptly.

Follow-up Stage
[DOCTOR] Remember, staying in touch with us and following up as scheduled is key to a successful recovery. Do you have any other questions about what to watch for or how to manage your recovery at home?
[PATIENT] Thank you for the information, Doctor. I will keep an eye on those symptoms and reach out if I notice anything unusual. What follow-up appointments should I expect?
[DOCTOR] You're welcome. For follow-up, you should schedule an appointment with me or the neurosurgery team within 2 to 4 weeks after discharge to check on your progress and the healing of your surgical site. We may also adjust your medications based on your recovery. Additionally, you'll continue to see your physical therapist regularly to monitor your rehabilitation progress. Make sure to bring any questions or concerns you might have to these appointments. Is there anything else you'd like to know about your medications or recovery process before we wrap up today?

Medication Stage
[PATIENT] I understand about the follow-up appointments. I'll make sure to bring my questions. What medications will I be taking during my recovery?
[DOCTOR] During your recovery, you'll be taking several medications to manage pain and prevent complications. You'll continue with Gabapentin 400 mg three times a day, which helps with nerve pain. For general pain, you can take Acetaminophen 325-650 mg every 6 hours as needed, but don't exceed 4 grams in a 24-hour period. If you experience muscle spasms, you can use Diazepam 5 mg every 8 hours as needed. To prevent constipation, which is common with pain medications, you'll take Docusate Sodium 100 mg twice a day. Lastly, you have Oxycontin (Immediate Release) available for breakthrough pain; take it every 4 hours as needed, but hold it if you feel overly sedated or drowsy. Make sure to follow these guidelines closely and let us know if you have any side effects or if the pain isn't well-controlled.

Postdischarge treatment Stage
[PATIENT] Thank you, Doctor. I understand the medications I need to take and their purposes. What should I do if I find the pain isn't well-controlled?
[DOCTOR] If you find that the pain isn't well-controlled, it's important to contact us so we can adjust your medication regimen. Don't hesitate to reach out if you feel the current medications aren't managing your pain effectively. In the meantime, ensure you're taking the medications as prescribed and consider keeping a log of your pain levels and when you take your medications. This can help us better understand your needs and make appropriate adjustments. How does that sound? Anything else you're concerned about?

Follow-up Stage
[PATIENT] That sounds good, Doctor. I'll keep a log of my pain levels and reach out if I need adjustments. Is there anything else I should be aware of as I start my recovery?
[DOCTOR] It's great to hear that you're planning to keep a log of your pain levels. That will be very helpful. Another important point is to avoid heavy lifting and strenuous activities for at least 6 to 12 weeks post-surgery to allow proper healing. Follow the advice of your physical therapist regarding activity levels and exercises. If you have any concerns about your recovery or if you experience any unexpected symptoms, don't hesitate to contact us. We want to ensure you have a smooth and successful recovery. Do you have any final questions or concerns before you head home?
[PATIENT] I understand about avoiding heavy lifting and following my physical therapist's advice. I don't have any final questions right now, but thank you for all the information, Doctor.
[DOCTOR] You're welcome, Mr. ___. I'm glad to hear that you feel prepared. Remember, your recovery is a journey, and we're here to support you every step of the way. If anything comes up or if you change your mind and have more questions later, please don't hesitate to call our office. Have a safe trip home and take care.

Figure 20: Example discharge conversation generated by Qwen2.5-32B.

Patient #1

Discharge Diagnosis
PRIMARY: Post-operative recovery from brain tumor removal in the fourth ventricle.
SECONDARY: —

Discharge Medications
- Bisacodyl 5 mg (EC): 2 tablets PO daily PRN constipation.
- Senna 8.6 mg: BID PRN constipation as prescribed.
- Docusate Sodium 100 mg: 1 tablet PO BID while on pain medications.
- Nifedipine SR 30 mg: 1 tablet PO HS.
- Oxycodone–Acetaminophen: q4–6h PRN pain as prescribed.
- Camphor–Menthol 0.5%/0.5% lotion: Apply TID PRN.
- Sodium Chloride 0.65% aerosol spray: PRN up to QID.
- Saline sensitive eye drops: 5 times/day PRN.

Discharge Condition
Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive. Activity: Ambulatory—independent.

Discharge Instructions
Gentle rehabilitative exercises per neurologist; maintain balanced diet and hydration; avoid strenuous activity until cleared. Seek care for severe headache, confusion, difficulty speaking, or new neurologic deficits.

Follow-up Instructions
Follow up with neurologist for rehabilitation and ongoing care.

Patient #2

Discharge Diagnosis
PRIMARY: Lower GI bleed (likely liver-disease related); Hepatic encephalopathy.
SECONDARY: Liver disease with portal hypertension and prior variceal bleeding.

Discharge Medications
- Rifaximin 550 mg: 1 tablet PO BID.
- Pantoprazole 40 mg: 1 tablet PO q24h.
- Thiamine 100 mg: 1 tablet PO daily.
- Folic Acid 1 mg: 1 tablet PO daily.
- Calcium Carbonate 500 mg: 1 tablet PO BID (chew).
- Multivitamin: 1 tablet PO daily with food.
- Lactulose 10 g/15 mL: 30 mL PO q4h.
- Furosemide 20 mg: 3 tablets PO AM and 2 tablets PO PM.
- Spironolactone 50 mg: 3 tablets PO BID.
- Lidocaine HCl 2% solution: 2 mL oral swish PRN discomfort.
- Acetaminophen 500 mg: 1 tablet PO q6h PRN pain (max 4 tablets/day).
- Ferrous Sulfate 300 mg (60 mg iron): 1 tablet PO daily on empty stomach.

Discharge Condition
Mental Status: Clear; understands instructions. Level of Consciousness: Alert and interactive. Activity: Light activity as tolerated.

Discharge Instructions
Strict alcohol avoidance; follow medication plan for encephalopathy and fluid retention; low-sodium diet; light activity. Seek care if symptoms worsen.

Follow-up Instructions
Follow up with PCP as advised; complete recommended labs; close monitoring of progress.

Patient #3

Discharge Diagnosis
PRIMARY: Liver Abscess; Gram-negative bacteremia.
SECONDARY: —

Discharge Medications
- Metronidazole 500 mg: 1 tablet PO TID for 4 weeks (Klebsiella infection).
- Aspirin 81 mg: 1 tablet PO daily.
- Dorzolamide 2% eye drops: 1 drop OU BID.
- Hydrocodone–Acetaminophen: PRN pain.
- Latanoprost 0.005% eye drops: 1 drop OU nightly.
- Levothyroxine Sodium 25 mcg: 1 tablet PO daily.
- Metformin 1000 mg: 1 tablet PO BID.
- Metoprolol Succinate XL 12.5 mg: 1 tablet PO nightly.
- Nitroglycerin SL 0.3 mg: PRN chest pain.
- Ondansetron: PRN nausea.
- Pantoprazole 40 mg: 1 tablet PO daily.
- Timolol Maleate 0.25% eye drops: 1 drop OU BID.

Discharge Condition
Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive. Activity: Ambulatory—independent; avoid alcohol.

Discharge Instructions
Admitted with liver abscess and gram-negative bacteremia; treated with antibiotics and abscess drainage. Continue medication regimen, monitor for signs of infection, and keep follow-up appointments.

Follow-up Instructions
Follow up with PCP and our office for evaluation of underlying liver cancer and other conditions; reassess abscess and remove pigtail catheter once infection is fully cleared.

Figure 21: Three example discharge summaries generated by Qwen2.5-72B.