"From Triage To Discharge": Do New LLM Versions Handle the ED Better? Skyer Answers.

Shadi Nourbakhsh, Mehrdad Majzoobi
Skye Insights, LTD
Toronto, Ontario, Canada
shadi@skyeinsights.net, mehrdad@skyeinsights.net

Abstract

Generative AI models undergo frequent updates, but for clinical implementers, each release triggers a significant and costly dilemma that standard benchmarks fail to address: does the new version truly enhance performance for their specific workflows, and does it justify the associated costs and risks? We previously introduced Skyer, a 55-scenario, clinician-validated benchmark for pediatric Emergency Department (ED) triage. Here we generalize it into a reusable framework for safety-aware, version-over-version evaluation of LLMs across diagnosis, triage, and decision-making performance, plus their consistency. In Skyer, rather than rewarding raw accuracy, we applied a weighting system to reward correct scores while penalizing both over- and under-scoring errors to capture the real-world patient-safety impact of mistriage, paired with a consistency criterion for reliability. Applying Skyer to fifteen models (eight GPT, five Gemini, and two Gemma versions) as an empirical demonstration, we find that newer is frequently not better: Gemini improves on a gradual steady path, Gemma remains on a low and inconsistent plateau, and GPT improves erratically, with an older version outperforming its successor. These findings show why upgrade decisions need a per-release instrument rather than vendor benchmarks, and we distil them into a deployment protocol that weighs each release's safety-weighted gain against its incremental cost. Because Skyer is contamination-resistant and version-agnostic, it extends to future releases and, by re-parameterization, to other clinical tasks.

Keywords: LLM, Emergency departments (ED) tasks, ML models iteration and increments, LLM version-over-version framework, version upgrade decision, cost-benefit, patient safety, model deployment

1 Introduction

The capabilities of Large Language Models (LLMs) have advanced rapidly, and vendors release upgraded versions continually to improve their performance. Each release raises a recurring question: has the model genuinely improved for a given use, or quietly regressed? Benchmarking answers this by testing models across tasks and metrics to detect regressions, quantify performance, and guide development (Chang et al., 2024; Rudd et al., 2025).

Assessing whether updated model versions offer chronological functional improvements is key to determining their cost-effectiveness and safety, especially in healthcare application. Users such as hospitals and healthcare networks must identify the most suitable model and continually evaluate the impact of newest updates. Our results suggest that seeking the absolute latest version may not consistently be the most advantageous approach for a specific performance task. Consequently, they may be better served by a stable, local model or a long-term version freeze than by chasing every release. Crucially, vendor release notes and headline benchmarks cannot tell a deployer whether their specific task improved: in our evaluation the older GPT-4.5-preview outperforms the newer GPT-5, which is precisely why each release must be re-tested on the target task.

Skyer was introduced as a benchmark for pediatric Emergency Department (ED) triage: 55 clinician-validated scenarios scored by a safety-aware weighting rubric, validated against human experts (Nourbakhsh, 2026). We use "the Skyer benchmark" for that fixed scenario set and rubric, and "the Skyer framework," or "Skyer" alone, for the version-over-version evaluation protocol this paper builds on top of it. In this paper we generalize the benchmark into that framework by extending its weighted scoring from triage alone to diagnosis and decision-making, and by re-applying it chronologically within each vendor family rather than as a single cross-vendor snapshot, adding the statistical validation (Cohen's d, Monte Carlo weight-sensitivity) and deployment protocol that supports that use. We compare successive versions of the Gemini, Gemma, and GPT families under identical conditions, scoring each against a shared human-expert reference standard to determine whether newer updates yield statistically significant improvements on core ED metrics. Strong benchmark accuracy can mask the failures that matter at the bedside. Skyer is therefore organized around three deployment concerns: whether a version is *safe* (penalizing harmful under- and over-triage rather than rewarding raw accuracy), whether it is fit for an *underserved* pediatric population largely absent from the coding- and adult-centric benchmarks that define the state of the art, and whether it actually *works* under the operational reality of version selection and upgrade decisions.

This paper makes three contributions: (i) *a reusable framework* for safety-aware, version-over-version upgrade decisions that transfers to future releases and to other clinical tasks; (ii) *an extension of Skyer's safety-weighted scoring* from triage alone to diagnosis and decision-making, so the same patient-safety cost accounting applies across all three ED tasks; and (iii) *an empirical evaluation* across fifteen models (eight GPT, five Gemini, and two Gemma versions) showing that newer is frequently not better.

2 Related Work

Developing models for ED triage is an active area of research. Several studies, such as those by Sax et al. (2023) and Williams et al. (2025), aim to create models that accurately predict triage outcomes. Sitthiprawiat et al. (2025)'s retrospective analysis of 163,452 ED visits reported superior results against the Canadian Triage and Acuity Scale (CTAS) CTAS National Working Group (2013). Taylor et al. (2025) performed a study of an AI-informed triage clinical decision support (CDS) intervention of 174,648 ED visits. The AI triage CDS intervention reduced median times from arrival to initial care area (33.0%; 12.0 to 8.0 minutes), ED disposition (4.2%; 190.0 to 182.0 minutes), and ED departure (6.1%; 311.0 to 292.0 minutes).

Newer Generative AI models show varied effectiveness for diagnostic support when compared to older AI models as well as human physicians Takita et al. (2025). AlDahoul & Zaki (2025) benchmarked GPT-4o, GPT-4o-mini, o3, Gemini 2.5 Pro, Gemini 2.5 Flash, and other frontier and open-source models on Arabic medical reasoning tasks. A majority-voting ensemble of o3, Gemini 2.5 Flash, and Gemini 2.5 Pro achieved the best overall accuracy (77%). Wang et al. (2025b) studied GPT-5 and reported improved accuracy and dependability and fewer hallucinations, making it a more trustworthy tool for clinical practice. Vishwanath et al. (2026) even showed how general purpose LLMs such as Gemini and GPT can perform better than specialized clinical AI tools such as OpenEvidence (OpenEvidence Inc., 2026).

Although newer model updates show significant advancements in healthcare applications, older model versions may retain value. Wang et al. (2025b) report GPT-4-Turbo is a better fit for prompt-amenable triage and GPT-4o excels in responsiveness and emotional engagement in EDs. For triaging patient urgency from raw dialogue, Lee et al. (2025) observed that Gemini 2.5 Flash excels in accuracy, specificity, and Positive Predictive Value (PPV); Gemini 2.5 Pro led in sensitivity and F1-score; and GPT-4.1 showed high, balanced accuracy and sensitivity. Wang et al. (2025a) noted that GPT-4o and o1 improved leading diagnosis prioritization but had inconsistent top-three differential diagnoses accuracy; thus, older models like GPT-3.5 remain valuable for broader coverage. Mykhalko et al. (2025) also observed that Gemini 2.0 Flash surpassed 1.5 Flash in differential diagnosis, though both had limitations in complex cases.

Version instability is not unique to the ED domain. Chen et al. (2023) tracked GPT-3.5 and GPT-4 across two 2023 snapshots and found large behavioral swings on math, code, and USMLE-style medical questions, including outright regressions between updates on the same task. This cross-domain finding motivates evaluating each release on its own rather than trusting a vendor's aggregate benchmark once. Two recent suites apply comparable rubric-based grading to broad clinical use: HealthBench (Arora et al., 2025) scores responses against physician-written rubrics across thousands of general-health conversations, and MedHELM (Bedi et al., 2026) evaluates LLMs on a clinician-validated, 121-task taxonomy spanning categories including clinical decision support, documentation, patient communication, and administration. Both are general-purpose and adult-oriented. Neither targets the pediatric ED population or separates the safety cost of over- from under-triage, which is the gap Skyer addresses.

Improving the development of new ED-specific models or fine tuning existing ones requires an ED-specific benchmark. It also requires a standardized method for evaluating incremental model improvements. Currently, assessments are limited and fail to offer a long-term system for this evaluation. Skyer is uniquely designed for this, evaluating LLM version updates for use as a pediatric ED assistant. Unlike prior, limited triage assessments, Skyer uses a weighted scoring system to reflect patient safety consequences of each aspect of triage.

3 Method

Our Skyer benchmark (Nourbakhsh, 2026) incorporated 55 distinct and realistic pediatric clinical scenarios. These scenarios deliberately covered a range of conditions, both common and severe, and were carefully chosen by our tenured medical staff to avoid overlap. We deliberately restricted the number of scenarios due to the complexity of evaluating each response and our limited resources.

This scale matches comparable clinician-validated vignette benchmarks in emergency medicine. Recent LLM evaluations in this class range from 12 physician-reviewed cases (Naderi et al., 2026) to 60 clinician-authored vignettes (Ramaswamy et al., 2026), 80 pediatric ED cases (Del Monte et al., 2025), and 124 consensus-adjudicated vignettes (Masanneck et al., 2024). Skyer's 55 scenarios fall within that range rather than below it. This is a different design point from large-scale, corpus- or crowd-sourced clinical suites such as HealthBench (5,000 conversations, Arora et al., 2025) and MedHELM (121 tasks across 35 benchmarks, Bedi et al., 2026). Those suites trade individual case validation for breadth. Skyer instead prioritizes per-case clinical fidelity. The same expert-curation trade-off caps scale in flagship benchmarks well beyond medicine: the Winograd Schema Challenge (273 items, Levesque et al., 2012), HumanEval (164 items, Chen et al., 2021), GPQA (198–448 items, Rein et al., 2023), and FrontierMath (300 items, Glazer et al., 2024) all evaluate the same GPT/Gemini/Gemma model families at a similar or larger scale, and remain among the most widely cited LLM benchmarks despite that.

This enables thorough human assessment of responses at this phase. Furthermore, we excluded publicly available standardized tests and datasets to prevent contamination of the models' training data. For the same reason we do not publicly release the complete 55-scenario set. Instead, we have released a gated, canary-protected 5-scenario prompt sample (SkyeInsights, 2026b) and the full per-case model predictions and scores underlying Table 2 (SkyeInsights, 2026a) on Hugging Face; decoding parameters, query dates, and per-run outputs are not yet included. To keep the interactions close to real life, we established a baseline for evaluation using only the patient description and context, without extra prompts. Each model's performance was assessed across three key ED tasks: diagnosis, triage and decision-making.

The reference standard for all 55 scenarios was established by a board-certified physician and evaluated by a panel of four experienced ED triage experts (one physician and three triage nurses). These experts independently specified the expected initial diagnosis and differential, triage levels (CTAS or ESI), and disposition for each case. Inter-rater reliability among the experts was good (ICC 0.75–0.90 across measures; ICC = 0.86 for CTAS and 0.82 for ESI triage), indicating consistent scoring of the reference answers and yielding a

single agreed ground truth against which every model response was subsequently scored. As a human reference point, the experts' own triage accuracy ranged from 56% to 69% (mean 64%), consistent with previously reported adult and pediatric triage performance and serving as the baseline against which model versions are compared (Nourbakhsh, 2026).

Evaluating LLMs solely on accuracy fails to capture true real-world performance. In healthcare, distinguishing over- from under-scoring errors matters because their consequences differ. In triage, overtriage is preferable to undertriage, despite leading to unnecessary costs and staff fatigue. Undertriage can result in disastrous patient outcomes due to increased waiting times and inadequate care. Similarly, in decision making, over-decisions are generally preferred to under-decisions. Over-decisions can lead to incorrect patient admission, resource over-allocation, and staff fatigue. Under-decisions can lead to life-threatening complications, premature discharging, and undercaring of a patient in poor condition. To better reflect this preference, Skyer implemented a weighted evaluation system for our models. This system assigns different weights to various positive and negative variables, rewarding accurate scores while separately penalizing both over- and under-scoring errors. These weighting scores facilitated a comparison that goes beyond simple correctness of results, and evaluated the impact of decisions on patient outcome.

Diagnostic accuracy was measured using three metrics: accuracy of initial diagnosis, top three LLM proposals, and full differential diagnosis. Top three accuracy was the acceptable standard for evaluating LLM performance due to its clinical relevance, with initial accuracy as a secondary consideration. Full differential diagnosis accuracy has limited practical value (Sorich et al. (2024),Wang et al. (2025a)). Both initial diagnosis and top three diagnosis were scored in our weighting system as follows: 10 points for an accurate, completely correct diagnosis; 5 points for a partial diagnosis (correct subgroup but not the definitive diagnosis); and 0 points for an incorrect diagnosis.

Triage was evaluated using five-level Canadian Triage and Acuity Scale (CTAS) (CTAS National Working Group, 2013) and Emergency Severity Index (ESI) (Gilboy et al., 2012) systems. Since no minor instances are assumed in the pediatric field, levels 4 and 5 are consolidated. Whenever the model yielded two scores, the higher score determined the official triage level. Disposition decisions included Intensive Care Unit or ward admission, observation or treatment in ED, and self-care or treatment without further evaluation. The evaluation of triage and decision-making weights was conducted using several parameters. Each parameter was assigned a base weight range, reflecting its influence on patient outcomes and the efficiency of ED operations.

- Patient outcome, prognosis and improved patient care [-5,5]
- Patient safety [-3,3]
- Patient satisfaction and confidence [-2,2]
- Resource consumption or healthcare utilization [-2,0]
- Staff fatigue [-2,0]
- Impact on operational ED flow and other patients' waiting times (during triage), or the effects of over-hospitalization and bed occupancy (during decision-making) [-2,0]

Model versions were assessed in triage and decision-making by categorizing their scores based on accuracy, over-scoring, or under-scoring. These categories were grouped into 11 for triage ($T_1 - T_{11}$) and 7 for decision-making ($D_1 - D_7$), collectively referred to as A_i. For example, a correct triage score was group T_1, or an under-scoring error in decision-making (e.g., suggesting self-care when admission was correct) was group D_2. Each A_i group was assigned a unique multiplier, listed in Table 1, reflecting the impact severity of the over- or under-scoring error on the respective parameters within its category. The scoring engine is fixed, whereas these multiplier tables form the configurable layer of the framework: re-parameterizing them retargets Skyer to a different task or clinical domain without altering the underlying method.

Within each category (T_i or D_i based on the task), the weight score is calculated using six basic weights (W_1 to W_6) and their corresponding six multipliers, denoted as "j". Therefore,

Triage		p1	p2	p3	p4	p5	p6	Decision	p1	p2	p3	p4	p5	p6
Correct Score		1	1	1	0	0	0	Correct Score	1	1	1	0	0	0
Under score	T2: 1,2 to 4,5	1	1	1	0	0	0	D2: Admit to self-care	1	1	1	0	0	0
	T3: 1 to 2	0.6	0.6	0.5	0	0	0	D3: Admit to ED observation	0.6	0.3	0.5	0	0	0
	T4: 1 to 3	1	1	0.5	0	0	0							
	T5: 2 to 3	0.6	0.3	0.5	0	0	0	D4: ED observation to self-care	0.6	0.6	0.5	0	0	0
	T6: 3 to 4,5	0.6	0.6	0.5	0	0	0							
Over score	T7: 4,5 to 1,2	0	0	0	1	1	1	D5: Self-care to admit	0	0	0	1	1	1
	T8: 4,5 to 3	0	0	0	0.5	0.5	0.5	D6: Self-care to ED observation	0	0	0	1	1	0
	T9: 3 to 1	0	0	0	0.5	0.5	1							
	T10: 3 to 2	0	0	0	0	0.5	0.5	D7: ED observation to admit	0	0	0	0	0.5	1
	T11: 2 to 1	0	0	0	0.5	0.5	0							

Table 1: Multipliers (j) for weighted parameters (P_1-P_6) in Decision-making and Triage categories (A_i)

the total weight score (S) for any given question is the sum of the product of each basic weight (W) and its unique multiplier in the specific category, collectively named A_i ($M[A_{i,j}]$).

$$S_i = \Sigma_{j=1}^{6} W_j * M[A_{i,j}] \tag{1}$$

The total weight score used to evaluate an LLM's performance for a given task, such as triage, is calculated by adding the scores assigned to all 55 questions within that task.

Formally, the per-task weight is $\Sigma_{i=1}^{55}(\Sigma_{j=1}^{6} W_j * M[A_{i,j}])$ and a model's overall weight is the sum of its task scores, $S_{triage} + S_{decision\ making} + S_{diagnosis}$, hereafter referred to as its **weights**.

Consistent LLM operation is vital for accurate performance, alongside the total performance weights score. Inconsistency is a major issue, especially for general ED tasks. A model is considered consistent if it returns the same response to repeated queries 80% of the time (Patwardhan et al., 2024). To assess this, in our study, models were tested three times on identical cases. For routine comparisons, only the first result was utilized, as repeated queries and averaging are impractical in a real-world setting. In contrast, all three results were included in the consistency evaluation. A reliable model achieves a consistency exceeding 80% across all three performances. Crucially, a model's consistency was based on its lowest score if any single task fell below the 80% threshold.

Skyer assessed the weighted models' performance scores across the following parameter weights: initial and top three diagnosis, ESI and CTAS triage, and decision-making. One tailored T-test was conducted to statistically evaluate if the updated model outperformed its preceding version. However, small sample sizes resulted in high and unstable p-values, even in cases with real differences. Therefore, instead of relying solely on potentially unreliable statistical measures, we adopted Cohen's D for a more efficient practical evaluation of the upgraded model versions' performance (Cohen, 1969). A large Cohen's D alongside a non-significant p-value indicated a strong practical effect but insufficient data for statistical certainty. Since equal variance was confirmed by the F-Test (p-value $>$0.05), the pooled Standard Deviation was used. To verify that our conclusions do not depend on the specific weight values in Table 1, we additionally ran a Monte Carlo sensitivity analysis, independently perturbing every base weight W_1–W_6, every non-zero multiplier, and the diagnosis point values by $\pm$30% across 2,000 samples. Of the 120 pairwise model comparisons, 110 (92%) never changed order in any sample. Eleven of the twelve pairwise comparisons between chronologically consecutive versions within a family retained their baseline order in at least 97% of samples; the exception is GPT-4.1_2025-04-14 vs. GPT-5, whose near-tied baseline scores flip order in 12% of samples, consistent with the diagnosis-task plateau we report between these two versions in Section 4. All comparisons underlying our Gemini-improvement findings, and the GPT-4.5-preview-vs-GPT-5 regression finding, held in at least 97% of samples.

4 Results

As an empirical demonstration of Skyer, the performance of various iterations/increments of Gemini, Gemma and GPT models was assessed to pinpoint areas for enhancement across the models' updates (Table 2). Since neither of the models was fine-tuned for healthcare applications, their comparison remained fair. Our Skyer framework aimed to determine the impact of model upgrades on the accuracy and reliability of LLMs in overall ED performance. As noted before, Skyer uses several key metrics: initial and top three diagnoses (D_x) accuracy, mean triage scores (CTAS and ESI), and decision-making capabilities and total performance. Given the serious consequences of both over- and under-scoring, the assessment incorporated the total weight assigned to these factors as noted earlier. Since models' diagnostic outputs were nearly identical, the combined weight of triage and decision-making proved an effective comparative parameter.

4.1 Gemini models

The Gemini updates, from the initial version to the most recent, show marked improvements across all performances, with the exception of Gemini-2.5_03-25. Compared to the baseline of Gemini-1.5-pro, the most recent version (Gemini-3.0-pro-preview) shows significant growth across tasks. Specifically, there is a 20% weight increase (110 out of 550 points) in initial diagnosis. Furthermore, top-three diagnosis improved by 14% (75 out of 550 points), and triage scores grew by 13% (69.5 out of 550 points). The chronological decision making improvement is also notable, rising 15% (81 out of 550 points) in Gemini-2.5_03-25, before slightly decreasing in subsequent upgrades. Furthermore, Gemini versions demonstrated high consistency, exceeding 80%, with the exception of Gemini-2.5_03-25, which was deemed unreliable (Table 2 and Figure 1). Gemini-2.0-pro-exp version was unavailable for re-evaluation due to deprecation.

Weights	Initial Dx accuracy	Top three Dx accuracy	Mean triage	Decision making	Mean triage and Decision	Total performance	Consistency
Gemini-1.5-pro	360	440	295	303	598	1038	84.6%
Gemini-2.0-pro-exp	390	485	290	340	630	1115	N/A
Gemini-2.5_03-25	430	495	259	384	643	1138	66%
Gemini-2.5_05-06	450	505	332	349	681	1186	83.6%
Gemini-3.0-pro-preview	470	515	364.5	348	712.5	1227.5	89%
Gemma-2-27b-it	190	345	221	230	451	796	44%
Gemma-3-27b-it	290	400	279	213	492	892	59%
GPT-4_2023-03-14	310	450	227.5	265	492.5	942.5	47%
GPT-4o_2024-05-13	420	510	232	355	587	1097	69%
GPT-4.5-preview	430	510	344.5	381	725.5	1235.5	91%
GPT-4.1_2025-04-14	440	490	270	383	653	1143	76%
GPT-5	460	520	236.5	379	615.5	1135.5	71%
GPT-5_2025-08-07	440	505	294.5	351	645.5	1150.5	55.5%
GPT-5.1_2025-11-13	390	500	196.5	382	578.5	1078.5	90%
GPT-5.2_2025-12-11	410	495	293	416	709	1204	89%
Theoretical perfect score		550	550	550		1650	100%

Table 2: Overall performance scores in ED tasks

Overall, Gemini model updates showed continuous but slow improved performance and consistency across versions (Table 3). This improvement was initially slight and only practically better. Gemini-2.5_05-06 and Gemini-3.0-pro-preview, the latest updates, represent significant advancements both in statistical and practical evaluation and consistency. In contrast, Gemini-2.5_03-25 version is underperforming and appears unsuitable for use in ED environments.

4.2 Gemma models

Gemma model iterations showed significant practical improvement across diagnosis (18% or 100 points in initial diagnosis and 10% or 55 out of 550 points in top three diagnosis weight) and huge practical improvements (10%) in triage weight. In contrast, Gemma models experienced a slight regression (3%; 17 out of 550 points) in decision making. However, despite these advancements, their performance remains substantially lower than the Gemini

and GPT models. The low consistency of both Gemma versions (<60%) renders them unreliable (Tables 2 and 1).

Overall, Gemma represents statistically significant and practically meaningful improvements across all performance metrics with the update from 2-27b-it to 3-27b-it (Table 3). A key advantage of this model is its offline availability. However, even the upgraded version's performance remains far from reliable and acceptable, indicating a long path of improvement is needed before being deemed acceptable.

Both Gemma versions tested are the general-purpose release, not MedGemma, the medically fine-tuned variant of the same 27B backbone (Sellergren et al., 2025). Skyer excludes healthcare fine-tunes across all vendors at this stage to keep the version-over-version comparison fair. These results describe base Gemma's off-the-shelf ED performance, not open-weight models generally. Evaluating MedGemma with Skyer is a natural next step.

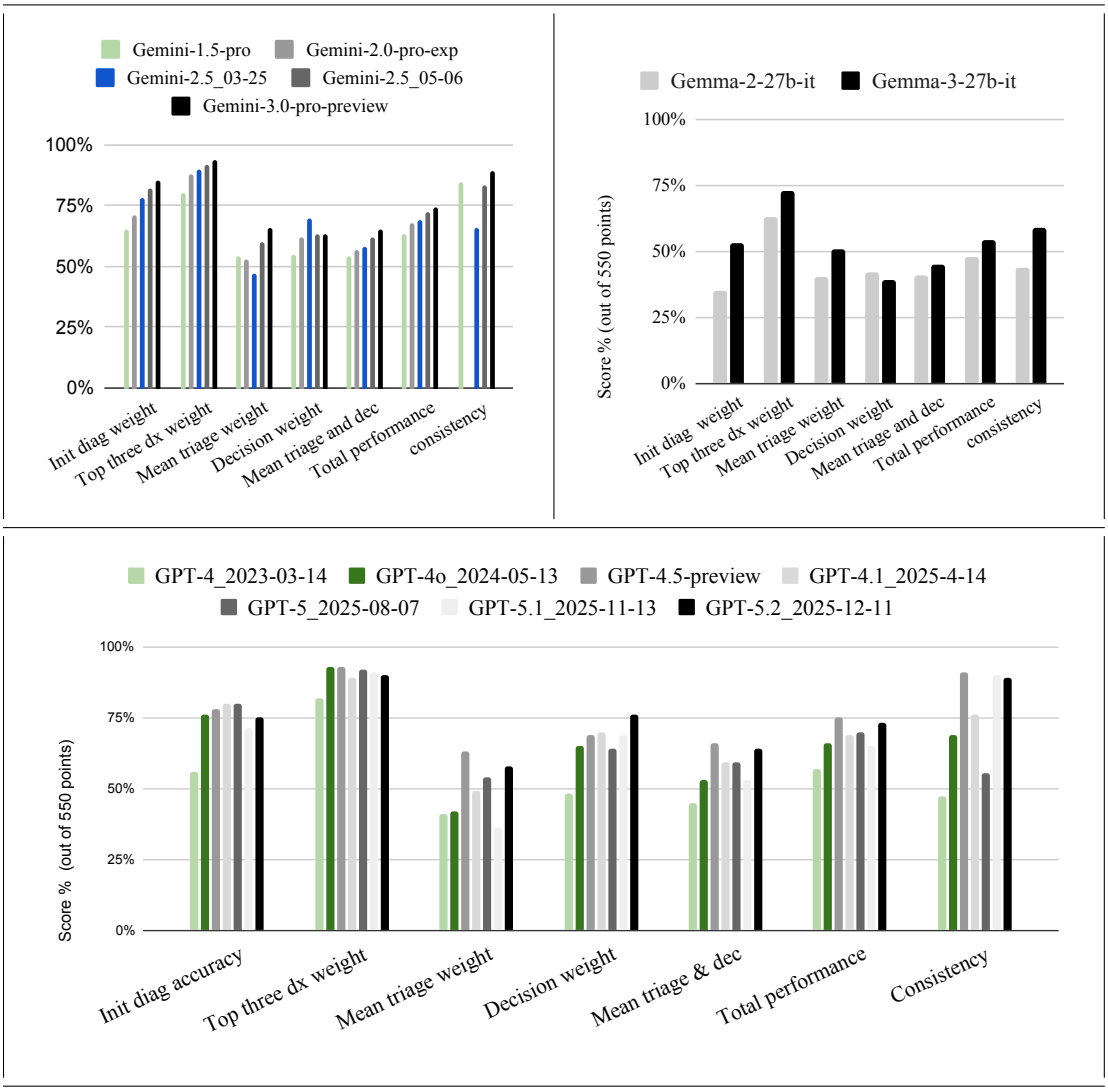

Figure 1: Chronological comparison of LLM versions overall performance versus task

4.3 GPT models

The ED performance of GPT models showed an unsteady pattern, experiencing both improvements and regressions. The diagnosis task of GPT models demonstrated a significant

		Gemini-2.0-pro-exp	Gemini-2.5_03-25	Gemini-2.5_05-06	Gemini-3.0-pro-preview
Gemini-1.5-pro	Cohen's D	0.25 slight diff	0.28, slight diff*	0.69 noticeable diff	0.95 meaningful diff
	P.value	0.07 not sig**	0.19 not sig	0.002 significant	0.00005 significant
Gemini-2.0-pro-exp	Cohen's D		0.05 relatively similar	0.40 slight diff	0.63 noticeable diff
	P.value		0.38 not sig	0.008 significant	0.01 significant
Gemini-2.5_03-25	Cohen's D			0.29 slight diff	0.49 noticeable diff
	P.value			0.12 not sig	0.08 not sig
Gemini-2.5_05-06	Cohen's D				0.22 slight diff
	P.value				0.03 significant

		Gemma-3-27b-it
Gemma-2-27b-it	Cohen's D	0.03 significant
	P.value	0.81 meaningful diff

		GPT-4o_2024-05-13	GPT-4.5-preview	GPT-4.1_2025-04-14	GPT-5	GPT-5_2025-08-07	GPT-5.1_2025-11-13	GPT-5.2_2025-12-11
GPT-4_2023-03-14	Cohen's D	0.53 noticeable diff	1.25 huge diff	0.74 noticeable diff	0.59 noticeable diff	0.83 meaningful diff	0.32 slight diff	0.92 meaningful diff
	P.value	0.02 sig	0.0006 sig	0.01 sig	0.04 sig	0.002 sig	0.14 not sig	0.005 sig
GPT-4o_2024-05-13	Cohen's D		0.48 slight diff	0.13 relatively similar	0.08 relatively similar	0.19 slight diff	-0.15 negative diff	0.25 slight diff
	P.value		0.052 not sig	0.095 not sig	0.21 not sig	0.09 not sig	negative	0.10 not sig
GPT-4.5-preview	Cohen's D			-0.37 negative diff	-0.34 negative diff	-0.31 negative diff	-0.62 negative diff	-0.26 negative diff
	P.value			negative	negative	negative	negative	negative
GPT-4.1_2025-04-14	Cohen's D				-0.03 negative diff	0.06 relatively similar	-0.29 negative diff	0.12 relatively similar
	P.value				negative	0.29 not sig	negative	0.16 not sig
GPT-5	Cohen's D					0.09 relatively similar	-0.22 negative diff	0.14 relatively similar
	P.value					0.32 not sig	negative	0.27 not sig
GPT-5_2025-08-07	Cohen's D						-0.35 negative diff	0.06 relatively similar
	P.value						negative	0.38 not sig
GPT-5.1_2025-11-13	Cohen's D							0.40 noticeable diff
	P.value							0.05 not sig

Table 3: Chronological practical and statistical assessment of upgrading model versions in overall performances. Green cells denote a positive, practically and/or statistically meaningful improvement; grey cells denote a negative change (regression). *Diff = difference; **sig = significant.

practical leap of up to 27% (150 out of 550 score points) increase at the beginning with no specific further improvement. Alarmingly, this family sees a regression of 9% (50 out of 550 points) decrease in the most recent versions. The triage and decision making performance of GPT models showed a similar unsteady pattern, too. Triage weight is increased 22% in GPT-4.5-preview, however each subsequent update initially regressed (27% to 10% decrease) before showing improvement. The decision making initially demonstrates some regression in updates and then a 27% (151 out of 550 points) weight increase in the latest version. Meanwhile, except for GPT-4.5-preview, GPT-5.1_2025-11-13 and GPT-5.2_2025-12-11 (over 80% consistency), and partly GPT-4.1_2025-04-14, the remaining GPT model versions were unreliable (Table 2 and Figure 1).

Overall, the GPT versions performance trend is unsteady, possibly due to frequent rapid updates (Table 3). Initial updates generally yielded significant improvements, both statistically significant and practically noticeable. The best-performing model in this research, exhibiting both high performance and consistency, was the now-deprecated GPT-4.5-preview. Following that, the second-best result was the GPT-5.2_2025-12-11 model. It demonstrated a successful and consistent version improvement. These two GPT iterations can be considered the only successful incremental improvements. Notably, despite over 80% consistency, the GPT-5.1_2025-11-13 version performed poorly in ED tasks, achieving only 65%.

5 Discussion

Within the healthcare context, while the primary goal is to identify an LLM capable of performing the task accurately, achieving a flawless model is unrealistic. Meanwhile, inaccurate performance for ED tasks has varying and distinct implications for patient safety. Therefore, simply relying on an accuracy evaluation is insufficient and a weighted scoring system is essential, as we have achieved in Skyer. For instance, our results indicate that while GPT-5.2_2025-12-11 has a lower triage accuracy than GPT-5_2025-08-07, its performance excels when the triage weight is considered. The key advantage for GPT-5.2_2025-12-11 lies in its reduced underscoring, which makes it the safer and more acceptable option for ED tasks. While Gemini-1.5 and Gemini-2.5_05-06 show comparable triage accuracy, Gemini-2.5_05-06 achieves a better weighted triage score due to more overscoring.

Skyer's per-version weights also speak to model developers, for whom the chronological trend differs sharply by vendor: Gemini exemplifies a persistent positive trend, GPT improves erratically (likely linked to the high frequency of its updates), and Gemma improves on all three ED tasks yet remains well below the others and inconsistent. These chrono-

logical, per-vendor trends are robust to weight choice: our sensitivity analysis (Section 3) confirmed that every version-succession comparison underlying them held in at least 97% of Monte Carlo samples across the twelve chronological-succession pairs, with one exception (GPT-4.1_2025-04-14 vs. GPT-5, a near-tied plateau rather than a directional trend). The complementary end-user question of whether to adopt each new release is addressed in Section 5.1.

5.1 Deployment implications

For end users such as hospitals and healthcare networks, the most actionable takeaway from Skyer is that upgrading to the newest model version is not always beneficial for a given ED task. In some cases an organization is better served by mandating a stable, locally-hosted model or freezing a cloud-based version: in our study GPT-4.5-preview outperformed the newer GPT-5, and Gemini-1.5 outperformed the updated Gemini-2.5_03-25, so retaining the older and often cheaper version would have been the financially and clinically prudent choice. These observations argue for treating model-version selection and version pinning as an explicit, continuously re-evaluated deployment decision rather than a default upgrade to the latest release.

Skyer operationalizes this decision as a repeatable upgrade protocol that a deployer applies on each release:

1. Run Skyer on the candidate version against the incumbent for the target task.
2. Require a safety-weighted gain over the incumbent, not merely higher accuracy.
3. Require consistency of at least 80%.
4. Weigh any gain against the incremental cost of the new version.
5. Emit a verdict: upgrade, hold, freeze, or roll back.

Read this way, Skyer estimates incremental safety-weighted gain per incremental cost; an upgrade that misses that bar should be deferred regardless of headline accuracy. Because the protocol is contamination-resistant and version-agnostic, it applies unchanged to releases postdating the benchmark's design: we already scored Gemini-3.0-pro-preview and GPT-5.2_2025-12-11, both released after Skyer was built.

6 Conclusion

We applied our Skyer framework to commonly available general-purpose models in the ED settings. Our results indicate that LLM performance and consistency do not necessarily follow a steady upward trend. Notably, instability is observed in some models, despite claims that internal benchmarking is performed prior to public release. We even observed the deprecation of the version, which is still among the top-performing updates. In contrast, other comparable general-purpose models demonstrate measurable improvement over time. This suggests the variability reflects model-specific characteristics rather than the ED domain or question design.

These results underscore the need for transparent third-party benchmarking and independent, task-specific validation before deploying or upgrading AI-driven systems in healthcare. In sum, Skyer offers a practical lens on three dimensions that raw accuracy benchmarks leave unmeasured: whether successive model versions are safe and appropriate for an underserved pediatric population, and operationally fit for emergency-department deployment.

References

Nouar AlDahoul and Yasir Zaki. Benchmarking the medical understanding and reasoning of large language models in Arabic healthcare tasks. *arXiv preprint arXiv:2508.15797*, 2025.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quinonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nature Medicine*, 32:943–951, 2026. doi: 10.1038/s41591-025-04151-2.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, 1969.

CTAS National Working Group. *The Canadian Triage and Acuity Scale Combined Adult/Pediatric Educational Program, Participant's Manual*, 2013. URL https://ctas-phctas.ca/wp-content/uploads/2018/05/participant_manual_v2.5b_november_2013_0.pdf.

Francesco Del Monte, Roberta Barolo, Maria Circhetta, Angelo Giovanni Delmonaco, Emanuele Castagno, Emanuele Pivetta, Letizia Bergamasco, Matteo Franco, Gabriella Olmo, and Claudia Bondone. Diagnostic efficacy of large language models in the pediatric emergency department: a pilot study. *Frontiers in Digital Health*, pp. 1624786, 2025. doi: 10.3389/fdgth.2025.1624786.

Nicki Gilboy, Paula Tanabe, Debbie Travers, Alexander M Rosenau, et al. Emergency Severity Index (ESI): a triage tool for emergency department care, version 4. *Implementation handbook*, 2012:12–0014, 2012.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Sukyo Lee, Sumin Jung, Jong-Hak Park, Hanjin Cho, Sungwoo Moon, and Sejoong Ahn. Performance of ChatGPT, Gemini and DeepSeek for non-critical triage support using real-world conversations in emergency department. *BMC Emergency Medicine*, 25(1):176, 2025.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference (KR2012)*, pp. 552–561, Palo Alto, CA, 2012. AAAI Press.

Lars Masanneck, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven G Meuth, Lennert Böhm, and Marc Pawlitzki. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study. *Journal of Medical Internet Research*, 26:e53297, 2024. doi: 10.2196/53297.

Yaroslav Mykhalko, Andrii Mykhalko, and Heorii Mykhalko. Evaluating the performance of Google's Gemini 1.5 flash and gemini 2.0 flash large language models in differential diagnosis. In *Materials of the 79th Final Scientific Conference of Uzhhorod National University*, Uzhhorod, Ukraine, 2025.

Borna Naderi, Longsha Liu, Anita Ghandehari, Darius Khoshons, R Andrew Taylor, Neil Bhavsar, Shriman Balasubramanian, Robert Tanouye, Nancy Creech, Christian Davidson, Justin Norden, Rahul Sharma, and Alexander Fortenko. The role of large language models in emergency care: a comprehensive benchmarking study. *npj Artificial Intelligence*, 2:24, 2026. doi: 10.1038/s44387-026-00078-2.

Shadi Nourbakhsh. Skyer: a novel benchmark for evaluating the effectiveness of large language models in emergency department triage. *Canadian Journal of Emergency Medicine*, pp. 1–9, 2026.

OpenEvidence Inc. OpenEvidence: An AI platform for clinical decision-making. https://www.openevidence.com/, 2026. Accessed: 2026-06-20.

Aditya Patwardhan, Vivek Vaidya, and Ashish Kundu. Automated consistency analysis of LLMs. In *2024 IEEE 6th international conference on trust, privacy and security in intelligent systems, and applications (TPS-ISA)*, pp. 118–127. IEEE, 2024.

Ashwin Ramaswamy, Alvira Tyagi, Hannah Hugo, Joy Jiang, Pushkala Jayaraman, Mateen Jangda, Alexis E Te, Steven A Kaplan, Joshua Lampert, Robert Freeman, et al. ChatGPT health performance in a structured test of triage recommendations. *Nature Medicine*, 32(5): 1671–1675, 2026. doi: 10.1038/s41591-026-04297-7.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Ethan M Rudd, Christopher Andrews, and Philip Tully. A practical guide for evaluating LLMs and LLM-reliant systems. *arXiv preprint arXiv:2506.13023*, 2025.

Dana R Sax, E Margaret Warton, Oleg Sofrygin, Dustin G Mark, Dustin W Ballard, Mamata V Kene, David R Vinson, and Mary E Reed. Automated analysis of unstructured clinical assessments improves emergency department triage performance: A retrospective deep learning analysis. *JACEP Open*, 4(4):e13003, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Patipan Sitthiprawiat, Borwon Wittayachamnankul, Wachiranun Sirikul, and Korsin Laohavisudhi. Development and internal validation of an ai-based emergency triage model for predicting critical outcomes in emergency department. *Scientific Reports*, 15(1):31212, 2025.

SkyeInsights. Skyer results: Per-case model predictions and reference-standard scores across 16 LLM versions. https://huggingface.co/datasets/SkyeInsights/skyer-results, 2026a. Accessed: 2026-08-18.

SkyeInsights. Skyer sample prompts: A gated, evaluation-only subset of pediatric ED scenarios. https://huggingface.co/datasets/SkyeInsights/skyer-sample-prompts, 2026b. Accessed: 2026-08-18.

Michael Joseph Sorich, Arduino Aleksander Mangoni, Stephen Bacchi, Bradley Douglas Menz, and Ashley Mark Hopkins. The triage and diagnostic accuracy of frontier large language models: updated comparison to physician performance. *Journal of Medical Internet Research*, 26:e67409, 2024.

Hirotaka Takita, Daijiro Kabata, Shannon L Walston, Hiroyuki Tatekawa, Kenichi Saito, Yasushi Tsujimoto, Yukio Miki, and Daiju Ueda. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *npj Digital Medicine*, 8(1):175, 2025.

R Andrew Taylor, Chris Chmura, Jeremiah Hinson, Benjamin Steinhart, Rohit Sangal, Arjun K Venkatesh, Haipeng Xu, Inessa Cohen, Isaac V Faustino, and Scott Levin. Impact of artificial intelligence–based triage decision support on emergency department care. *NEJM AI*, 2(3):AIoa2400296, 2025.

Krithik Vishwanath, Anton Alyakin, Mrigayu Ghosh, Ali Hage, Sean N Neifert, Cordelia Orillac, Nataniel J Mandelberg, Hammad A Khan, Jin Vivian Lee, Jie J Yao, et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nature Medicine*, pp. 1–5, 2026.

Jinge Wang, Kenneth Shue, Li Liu, and Gangqing Hu. Preliminary evaluation of ChatGPT model iterations in emergency department diagnostics. *Scientific reports*, 15(1):10426, 2025a.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of GPT-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*, 2025b.

Ethan L Williams, Daniel Huynh, Mohamed Estai, Toshi Sinha, Matthew Summerscales, and Yogesan Kanagasingam. Predicting inpatient admissions from emergency department triage using machine learning: a systematic review. *Mayo Clinic Proceedings: Digital Health*, 3(1):100197, 2025.