Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Performance and Safety of Three Large Language Models in Pediatric Paracetamol Poisoning Management

Alper Kacar
Sağlık Bilimleri Üniversitesi
Belen Ates
belenterlemez@gmail.com

Sağlık Bilimleri Üniversitesi
Sevilay Kok
Sağlık Bilimleri Üniversitesi

Research Article

Keywords: Large language models, ChatGPT, Claude, Gemini 3 Flash, Paracetamol poisoning, Pediatric emergency, Clinical decision support, Patient safety

Posted Date: June 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10018707/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background.

Paracetamol overdose is a common reason for pediatric poisoning consultations and emergency department visits, and its management is highly time and dose-dependent. Large language models (LLMs) are increasingly being explored for clinical decision support, but evidence on pediatric poisoning is scarce. We compared the performance and safety of three contemporary LLMs for managing pediatric paracetamol poisoning using guideline-based simulated scenarios.

Methods.

In this cross-sectional, comparative, scenario-based study, 30 hypothetical pediatric paracetamol poisoning scenarios were presented to ChatGPT 5.5, Claude Sonnet 4.6, and Gemini 3 Flash via their public browser-based interfaces, with a standardized single-turn prompt (90 responses). Blinded to model identity, a three-clinician expert panel scored each response by consensus using a study-specific Artificial Intelligence Evaluation Score for Clinical Poisoning Vignettes (AIES-CPV; range 5–25), an eight-item key-action checklist (0–8), the Global Quality Scale (GQS; 1–5), reference-standard concordance, and the presence of critical safety errors, against a predefined reference algorithm derived primarily from the Australia and New Zealand guidelines. Models were compared using the Friedman test (with Wilcoxon signed-rank tests and a Bonferroni correction for pairwise comparisons) and the Cochran Q test for critical error rates.

Results.

Claude Sonnet 4.6 achieved significantly higher total AIES-CPV (24.50 ± 1.43), key-action (7.93 ± 0.25), and GQS (4.90 ± 0.31) scores than ChatGPT 5.5 and Gemini 3 Flash (Friedman $p < 0.001$ for all); ChatGPT 5.5 and Gemini 3 Flash did not differ significantly. Claude Sonnet 4.6 produced no critical safety errors and was concordant with the reference standard in all responses, whereas one critical error occurred with ChatGPT 5.5 (1/30) and three with Gemini 3 Flash (3/30); the difference in critical-error rates was not statistically significant (Cochran $Q = 4.67$, $p = 0.097$).

Conclusions.

Among three LLMs evaluated on pediatric paracetamol poisoning scenarios, Claude Sonnet 4.6 showed the highest guideline-based accuracy, key-action performance, and overall quality, with no critical safety errors. Contemporary LLMs may serve as adjunctive decision-support tools in pediatric toxicology, but given inter-model safety differences and the risk of confident errors in high-stakes decisions, they should be used only under expert supervision and with clinical verification.

Background

Paracetamol is among the most commonly used analgesics and antipyretics in children and is readily available in many households. Although it has a wide safety margin at recommended doses, overdose remains a common cause of pediatric poisoning consultations and emergency department visits. In children, paracetamol toxicity may result from a single accidental ingestion, caregiver dosing errors, repeated supratherapeutic use during febrile illnesses, intentional self-poisoning in adolescents, modified-release preparations, or combination products containing other active substances. Each of these situations requires a different clinical approach, making it essential to evaluate the ingested dose, mg/kg exposure, formulation, time since ingestion, symptoms, and relevant laboratory findings together.

The management of paracetamol poisoning is highly time- and dose-dependent. Important decision points include calculating the total ingested dose and mg/kg exposure, distinguishing acute single ingestion from repeated supratherapeutic ingestion, timing the measurement of serum paracetamol concentration appropriately, using the Rumack–Matthew nomogram only in appropriate clinical circumstances, recognizing indications for N-acetylcysteine (NAC), and being aware of situations requiring extended treatment or consultation with a toxicologist [1–3]. Risk stratification and management may be further complicated by delayed presentation, unknown time of ingestion, modified-release preparations, malnutrition, decreased oral intake, abnormal aminotransferases or INR, metabolic acidosis, lactate elevation, and co-ingestions [1–3].

The Australia and New Zealand paracetamol poisoning guidelines provide a comprehensive framework for the assessment and management of acute, repeated supratherapeutic, modified-release, and massive paracetamol ingestions, including guidance on the administration of NAC and extended treatment in selected patients [1]. Practical, child-focused guidance on liquid paracetamol ingestion, timing of serum paracetamol concentration sampling, use of nomograms, NAC treatment, laboratory assessment, and follow-up is available in pediatric emergency care guidelines, including those from the Royal Children’s Hospital Melbourne and Queensland Paediatric Emergency Care [2, 3]. Thresholds and protocols may vary across countries and institutions. Therefore, evaluation of clinical decision support tools for paracetamol poisoning should be based on a predefined reference standard and structured key-action criteria rather than on isolated protocol details alone.

Large language models (LLMs) such as ChatGPT, Claude, and Gemini 3 Flash can generate coherent responses to complex clinical questions and are increasingly being studied as potential tools for health information retrieval, clinical reasoning support, and medical education [4, 5]. Their ability to provide unstructured clinical information makes them attractive in emergency medicine settings, where rapid analysis of history, dosage, timing, symptoms, and laboratory findings is often necessary. Nonetheless, LLMs may generate suggestions that are inaccurate, obsolete, excessively reassuring, or fabricated, and their seemingly fluent output might conceal clinically significant inaccuracies [4–6]. These factors are especially pertinent in pediatric toxicology, where inaccurate dose, misguided reassurance, delays in NAC delivery, or failure to identify high-risk characteristics can result in avoidable injury.

Assessments of clinical reasoning and safety of LLMs in medical decision-making are often scenario-based. In medical toxicology, simulated poisoning cases have been used to evaluate LLM responses across domains such as initial stabilization, diagnostic assessment, treatment recommendations, and follow-up [7]. Such vignette-based designs are ideal for paracetamol poisoning, as many of the decisions can be mapped to key actions derived from the guidelines, such as dose calculation, timing of serum levels, application of the nomogram, indication for NAC, laboratory monitoring, and recognition of special risk situations.

Although there is growing interest in LLMs as tools for clinical decision support, information on their efficacy in managing pediatric poisoning remains scarce. To our knowledge, comparative assessments of contemporary LLMs using pediatric paracetamol poisoning scenarios and a standardized, guideline-based expert consensus scoring framework are limited. This study sought to evaluate the efficacy and safety of three large language models—ChatGPT 5.5, Claude Sonnet 4.6, and Gemini 3 Flash—in managing pediatric paracetamol toxicity through clinician-focused simulated scenarios. Model responses were assessed against a predetermined reference algorithm and a key-action checklist primarily drawn from the Australia and New Zealand guidelines, supplemented by pediatric emergency guidance from the Royal Children's Hospital Melbourne and Queensland Paediatric Emergency Care.

Methods

This was a cross-sectional, comparative, scenario-based evaluation study designed to assess the clinical decision-support performance of three large language models in the management of pediatric paracetamol poisoning. The responses of ChatGPT, Claude, and Gemini 3 Flash to hypothetical pediatric paracetamol poisoning scenarios were compared in terms of accuracy, guideline concordance, clinical safety, applicability, and overall response quality.

Construction of case scenarios

The scenarios were developed by experienced pediatricians and contained no real patient data. Their content and the clinical decision points they were designed to probe were derived from the assessment and management pathways defined in the Australia and New Zealand paracetamol poisoning guidelines and the pediatric paracetamol guidance of the Royal Children's Hospital Melbourne and Queensland Paediatric Emergency Care [1–3]. On the basis of these sources, the scenarios were constructed to span the full range of clinically relevant situations, including low-risk ingestions, acute toxic ingestion, delayed presentation, cases with unknown time or dose of ingestion, repeated supratherapeutic ingestion, modified-release preparation ingestion, malnutrition, termination/extension of N-acetylcysteine therapy, N-acetylcysteine reactions, laboratory discrepancies, metabolic acidosis/lactate elevation, and co-ingestion of other drugs. A total of 30 clinical scenarios were created, each written as a realistic presentation that clinicians might encounter in a pediatric emergency department. Draft scenarios were reviewed iteratively by the study pediatricians; the wording, clinical parameters, and the specific decision point targeted by each case were discussed, and the final set of 30 scenarios was agreed upon by expert

consensus to ensure clinical realism, guideline relevance, and coverage of distinct decision points without redundancy.

For the purposes of an exploratory category-based subgroup analysis, the 30 scenarios were grouped a priori into four thematic categories based on the predominant clinical decision type they were designed to probe: (1) Acute ingestion (scenarios 1, 2, 3, 7–12; n = 9), comprising acute presentations with a known ingestion time requiring toxicity risk classification, nomogram interpretation, and NAC indication decision; (2) Atypical ingestion (scenarios 4, 6, 13–17; n = 7), encompassing cases with an unknown ingestion time or dose and presentations with delayed arrival; (3) Special clinical situations (scenarios 5, 18–24; n = 8), including repeated supratherapeutic ingestion, modified-release preparations, and malnutrition; and (4) Advanced clinical decision-making (scenarios 25–30; n = 6), covering NAC termination or extension decisions, NAC adverse reactions, laboratory discrepancies, metabolic acidosis or lactate elevation, and co-ingestion of other substances. Within-category model comparisons were performed using the Friedman test with Bonferroni-adjusted Wilcoxon signed-rank post-hoc tests (adjusted threshold $p < 0.017$). Given the small number of scenarios per category (n = 6–9), this analysis was pre-specified as exploratory and descriptive.

Evaluated large language models

Three contemporary large language models were evaluated: ChatGPT, Claude, and Gemini 3 Flash. Responses were obtained using the models' publicly available browser-based interfaces. During data collection, the model version, access date, interface used, and—where possible—the account/plan type were recorded. To reflect the default user experience of the models, no API-based custom configurations were used. Web browsing, external plug-ins, file upload, source citation, and similar additional tools were disabled as far as possible.

Data collection and prompting strategy

Each scenario was presented separately to each model using the same standardized prompt. A total of 90 LLM responses were obtained for 30 scenarios across 3 models. Each scenario was asked in a single turn; no follow-up questions, corrections, re-prompting, or iterative prompt refinement were performed.

The standardized prompt used for all scenarios was: "As a healthcare professional, you are evaluating the following patient in a pediatric emergency department. How would you assess the toxicity risk and what should the management steps be? Please state the total ingested paracetamol dose, the mg/kg dose calculation, whether and when a serum paracetamol concentration is needed, whether the Rumack–Matthew nomogram should be used, whether N-acetylcysteine should be started, and the necessary laboratory/monitoring recommendations."

Responses were recorded verbatim immediately after generation and archived with the model name, scenario number, and date and time. To minimize the influence of prior conversations, each scenario

was administered in an independent session; a temporary or session-limited chat mode was used, and when this option was unavailable, a new incognito/private browser session was started for each scenario.

Blinding and anonymization

Before evaluation, the LLM responses were anonymized by an independent researcher. Information that could reveal model identity was removed, and the responses were presented to the evaluators in random order. The evaluators did not know which response had been generated by which LLM.

Reference standard and guideline concordance

The accuracy and guideline concordance of the responses were assessed against a predefined reference algorithm and key-action checklist established before the study. The Australia and New Zealand paracetamol poisoning guidelines were used as the primary reference standard in constructing this algorithm. For pediatric emergency practice, the Royal Children's Hospital Melbourne and Queensland Paediatric Emergency Care paracetamol poisoning guidelines were considered as supporting sources.

Because minor differences in dose thresholds and management protocols may exist among guidelines, such differences alone were not considered errors. In contrast, omissions or actions that could affect patient safety—omitting an N-acetylcysteine indication, recommending home monitoring for a patient requiring emergency evaluation, inappropriate use of the Rumack–Matthew nomogram, omitting liver function tests or INR assessment in delayed presentation, disregarding a modified-release preparation or co-ingestion, and failing to account for risk-increasing conditions—were classified as critical safety errors.

Evaluation of responses

Clinical response quality was assessed using the Artificial Intelligence Evaluation Score for Clinical Poisoning Vignettes (AIES-CPV), adapted for this study from the Artificial Intelligence Evaluation Score for Common Patient Questions (AIES-CPQ), a structured five-domain expert rubric we recently introduced for appraising the quality of AI-generated responses to parent-oriented health questions [8]. For the present clinician-focused poisoning setting, the original parent-readability domains of the AIES-CPQ were reframed as guideline concordance and currency and as clarity and applicability for the clinician, consistent with the accuracy and completeness dimensions most commonly used to evaluate large language model outputs in clinical studies [9]. As with the AIES-CPQ, the AIES-CPV is a study-specific evaluation rubric and has not undergone formal external validation. It comprises five domains: medical accuracy, completeness of information, guideline concordance and currency, clinical appropriateness and safety, and clarity and applicability for the clinician. Each domain was scored from

1 to 5, and the total score was calculated between 5 and 25. A higher score indicated a more accurate, complete, safe, guideline-concordant, and clinically applicable response.

In addition, each response was evaluated with an eight-item scenario-based checklist: total dose calculation, mg/kg dose calculation, toxicity risk classification, ingestion type/timing, timing of serum paracetamol concentration, appropriate use of the Rumack–Matthew nomogram, N-acetylcysteine indication, and laboratory/monitoring/referral recommendations. Each item was scored as 0/1, and the total checklist score was calculated between 0 and 8. Critical safety errors were additionally recorded as present/absent. The overall quality of the responses was rated from 1 to 5 using the Global Quality Scale (GQS) [10]. Reference-standard concordance was also recorded for each response as concordant, partially concordant, or non-concordant.

Expert evaluation process

All responses were evaluated by a three-member expert panel comprising two general pediatricians and one pediatric emergency medicine subspecialist experienced in pediatric poisoning management. Blinded to model identity, the panel jointly reviewed and discussed each response and agreed on a single consensus score for the AIES-CPV, the scenario-based checklist, critical safety error, reference-standard concordance, and GQS. A single panel consensus score was generated for each response, and all analyses were conducted on these consensus scores. Because the evaluation used a structured consensus method rather than an independent multi-rater design, inter-rater reliability was not separately calculated.

Statistical analysis

Continuous variables were presented, according to their distribution, as mean ± standard deviation or as median and interquartile range. The normality of the score distributions was assessed with the Shapiro–Wilk test. The total and domain scores of the three LLMs were compared using the Friedman test. When a significant difference was detected, pairwise comparisons were performed with the Wilcoxon signed-rank test and Bonferroni correction was applied. Critical-error rates were presented as frequency and percentage by model. Where appropriate, the Cochran Q test was used to compare critical-error rates across models and the McNemar test for pairwise comparisons. Statistical significance was set at $p < 0.05$; for pairwise comparisons, the significance threshold after Bonferroni correction is additionally indicated.

Results

A total of 30 pediatric paracetamol poisoning scenarios were administered to the three LLMs, and 90 responses were evaluated using the final expert-panel consensus scores. The models were coded as LLM1, ChatGPT 5.5; LLM2, Claude Sonnet 4.6; and LLM3, Gemini 3 Flash.

Table 1
Overall performance scores of the LLM responses

Variable	ChatGPT 5.5 Mean ± SD	ChatGPT 5.5 Median (IQR)	Claude Sonnet 4.6 Mean ± SD	Claude Sonnet 4.6 Median (IQR)	Gemini 3 Flash Mean ± SD	Gemini 3 Flash Median (IQR)	Friedman χ^2	p
Total AIES-CPV	21.00 ± 3.66	21.5 (20–23)	24.50 ± 1.43	25 (25–25)	20.07 ± 4.00	21 (18.25–23)	34.98	< 0.001
Key-action checklist	7.13 ± 1.38	8 (7–8)	7.93 ± 0.25	8 (8–8)	6.90 ± 1.58	7 (6–8)	15.74	< 0.001
GQS	3.97 ± 0.76	4 (4–4)	4.90 ± 0.31	5 (5–5)	3.60 ± 0.67	4 (3–4)	37.73	< 0.001

AIES-CPV: Artificial Intelligence Evaluation Score for Clinical Poisoning Vignettes; GQS: Global Quality Scale; IQR: interquartile range; SD: standard deviation. Values were calculated from the final expert-panel consensus scores (N = 30 per model).

A significant difference was found among the three models in total AIES-CPV score (Friedman $\chi^2 = 34.98$, $p < 0.001$). Claude Sonnet 4.6 scored significantly higher than ChatGPT 5.5 and Gemini 3 Flash; no significant difference was found between ChatGPT 5.5 and Gemini 3 Flash. Significant differences among the models were also found for the key-action checklist score and GQS (Friedman $\chi^2 = 15.74$, $p < 0.001$ and $\chi^2 = 37.73$, $p < 0.001$, respectively). Claude Sonnet 4.6 showed the highest performance on both of these measures.

Table 2
Pairwise model comparisons with Bonferroni correction

Variable	Comparison	Wilcoxon W	Raw p	Bonferroni-adjusted p
Total AIES-CPV	ChatGPT 5.5 vs Claude Sonnet 4.6	31.5	< 0.001	0.001
Total AIES-CPV	ChatGPT 5.5 vs Gemini 3 Flash	116.0	0.325	0.976
Total AIES-CPV	Claude Sonnet 4.6 vs Gemini 3 Flash	14.5	< 0.001	< 0.001
Key-action checklist	ChatGPT 5.5 vs Claude Sonnet 4.6	9.0	0.003	0.009
Key-action checklist	ChatGPT 5.5 vs Gemini 3 Flash	55.0	0.480	1.000
Key-action checklist	Claude Sonnet 4.6 vs Gemini 3 Flash	4.5	0.001	0.002
GQS	ChatGPT 5.5 vs Claude Sonnet 4.6	18.0	< 0.001	< 0.001
GQS	ChatGPT 5.5 vs Gemini 3 Flash	37.5	0.048	0.144
GQS	Claude Sonnet 4.6 vs Gemini 3 Flash	0.0	< 0.001	< 0.001

The Wilcoxon signed-rank test was used. After Bonferroni correction, the significance threshold for pairwise comparisons was set at $p < 0.017$.

Table 3
Comparison of AIES-CPV subdomain scores

AIES-CPV domain	ChatGPT 5.5 Mean ± SD	Claude Sonnet 4.6 Mean ± SD	Gemini 3 Flash Mean ± SD	Friedman χ^2	p
Medical accuracy	4.27 ± 0.87	4.87 ± 0.35	4.13 ± 0.97	15.25	< 0.001
Completeness of information	3.87 ± 0.78	4.90 ± 0.31	3.60 ± 0.67	40.33	< 0.001
Guideline concordance/currency	4.37 ± 0.81	4.90 ± 0.31	4.30 ± 0.95	12.28	0.002
Clinical appropriateness/safety	4.30 ± 0.79	4.90 ± 0.31	4.37 ± 1.00	11.94	0.003
Clarity/applicability for the clinician	4.20 ± 0.81	4.93 ± 0.25	3.67 ± 0.66	37.24	< 0.001

Values are presented as mean ± standard deviation based on the final expert-panel consensus scores (N = 30 per model). Between-model comparison was performed with the Friedman test. AIES-CPV: Artificial Intelligence Evaluation Score for Clinical Poisoning Vignettes; SD: standard deviation.

Claude Sonnet 4.6 obtained the highest mean score in all AIES-CPV subdomains. The most pronounced differences were observed in completeness of information, clarity and applicability for the clinician, and overall quality.

Table 4
Critical safety errors and reference-standard concordance

Model	Critical errors n/N	Critical-error rate	Reference-standard concordant	Partially concordant	Non-concordant
ChatGPT 5.5	1/30	3.3%	26/30	3/30	1/30
Claude Sonnet 4.6	0/30	0%	30/30	0/30	0/30
Gemini 3 Flash	3/30	10.0%	27/30	2/30	1/30

Data are based on a single consensus evaluation per scenario (N = 30 per model). Critical-error rates were compared with the Cochran Q test; because the number of events was low, the statistical power of this comparison is limited. n/N: number of scenarios with a critical safety error / total number of scenarios.

No statistically significant difference was found among the critical safety error rates (Cochran Q = 4.67, p = 0.097). No critical safety error was observed for Claude Sonnet 4.6, and all of its responses were judged concordant with the reference standard.

Table 5
Number of correct responses by key-action checklist item

Checklist item	ChatGPT 5.5 n/N	Claude Sonnet 4.6 n/N	Gemini 3 Flash n/N
Correct total dose	26/30	30/30	25/30
Correct mg/kg dose	26/30	30/30	22/30
Correct toxicity risk	28/30	30/30	29/30
Correct ingestion type/timing	29/30	30/30	29/30
Correct timing of serum level	25/30	28/30	25/30
Appropriate Rumack–Matthew use	30/30	30/30	28/30
Correct NAC indication	22/30	30/30	24/30
Appropriate laboratory/monitoring/referral	28/30	30/30	25/30

Each checklist item was scored as 0/1 per scenario by single consensus evaluation. n/N: number of scenarios in which the item was correctly addressed / total number of scenarios (30 per model). NAC: N-acetylcysteine.

On examination of the checklist items, Claude Sonnet 4.6 showed the highest proportion of correct responses for total dose, mg/kg dose, toxicity risk, ingestion type/timing, Rumack–Matthew use, NAC indication, and laboratory/monitoring/referral recommendations.

Qualitative assessment of critical safety errors

For ChatGPT 5.5, one response was recorded as a critical error of the type “missing the need for laboratory evaluation in repeated supratherapeutic ingestion.” No critical error was identified for Claude Sonnet 4.6. For Gemini 3 Flash, critical errors were recorded in three responses and were classified as inadequate evaluation of an ingestion of unknown amount, omission of laboratory evaluation in repeated supratherapeutic ingestion, and risk of premature termination of NAC therapy.

Category-based performance across thematic scenarios

Category-based analysis demonstrated consistent superiority of Claude Sonnet 4.6 across all four thematic categories for both AIES-CPV and GQS scores (Table 6). Statistically significant differences among models were observed in all four categories for AIES-CPV (Friedman χ^2=7.29–13.13, p = 0.001–0.026) and GQS (χ^2=9.17–11.86, p = 0.003–0.010). Claude Sonnet 4.6 consistently achieved the highest mean ranks across all categories and both metrics.

The performance advantage of Claude Sonnet 4.6 was most pronounced in the Advanced clinical decision-making category (NAC management decisions, laboratory discrepancies, metabolic acidosis,

and co-ingestions), in which it achieved a perfect AIES-CPV score in all six scenarios (median 25.00, IQR 25.00–25.00; mean rank 3.00) compared with Gemini 3 Flash (median 22.00, IQR 20.25–23.00; mean rank 1.92) and ChatGPT 5.5 (median 18.00, IQR 15.75–18.75; mean rank 1.08; Friedman χ^2=11.565, p = 0.003). This was also the only category in which ChatGPT 5.5 ranked lower than Gemini 3 Flash. In the Atypical ingestion category, Claude Sonnet 4.6 significantly outperformed both ChatGPT 5.5 (Wilcoxon p = 0.016) and Gemini 3 Flash (p = 0.016); Gemini 3 Flash demonstrated the highest score variability in this category for AIES-CPV (mean rank 1.36), suggesting inconsistent handling of unknown-time or delayed-presentation scenarios. Post-hoc comparisons confirmed statistically significant superiority of Claude Sonnet 4.6 over Gemini 3 Flash in the Special clinical situations and Atypical ingestion categories for AIES-CPV (p = 0.008 and p = 0.016, respectively), and over Gemini 3 Flash in all four categories for GQS (p = 0.008–0.031).

Key-action checklist scores approached ceiling for Claude Sonnet 4.6 in all categories (median 8.00/8 in three of four categories), limiting between-model differentiation. Category-level Friedman tests did not reach statistical significance after Bonferroni correction (p = 0.050–0.186); key-action results are therefore presented descriptively.

Table 6
Friedman test results comparing mean ranks of LLMs across thematic scenario categories

Theme	Model	Median (IQR)	Mean Rank	Friedman χ^2 (df)	p value
AIES-CPV (5–25)					
Acute ingestion (n = 9)	ChatGPT 5.5	21.00 (21.00–25.00)	1.89	χ^2=7.294 (df = 2)	**0.026**
	Claude Sonnet 4.6	25.00 (25.00–25.00)	2.67		
	Gemini 3 Flash	23.00 (18.00–23.00)	1.44		
Atypical ingestion (n = 7)	ChatGPT 5.5	23.00 (22.00–23.00)	1.64	χ^2=13.130 (df = 2)	**0.001**
	Claude Sonnet 4.6	**25.00 (25.00–25.00)**	**3.00**		
	Gemini 3 Flash	23.00 (20.50–23.00)	1.36		
Special clinical situations (n = 8)	ChatGPT 5.5	22.50 (18.75–24.25)	2.00	χ^2=9.600 (df = 2)	**0.008**
	Claude Sonnet 4.6	25.00 (24.75–25.00)	2.75		
	Gemini 3 Flash	19.50 (17.50–20.25)	1.25		
Advanced clinical decision-making (n = 6)	ChatGPT 5.5	18.00 (15.75–18.75)	1.08	χ^2=11.565 (df = 2)	**0.003**
	Claude Sonnet 4.6	**25.00 (25.00–25.00)**	**3.00**		
	Gemini 3 Flash	22.00 (20.25–23.00)	1.92		
GQS (1–5)					
Acute ingestion (n = 9)	ChatGPT 5.5	4.00 (4.00–5.00)	2.11	χ^2=11.862 (df = 2)	**0.003**
	Claude Sonnet 4.6	5.00 (5.00–5.00)	2.67		
	Gemini 3 Flash	4.00 (3.00–4.00)	1.22		
Atypical ingestion (n = 7)	ChatGPT 5.5	4.00 (4.00–4.00)	1.64	χ^2=11.474 (df = 2)	**0.003**

Theme	Model	Median (IQR)	Mean Rank	Friedman χ^2 (df)	p value
AIES-CPV (5–25)					
	Claude Sonnet 4.6	5.00 (5.00–5.00)	2.86		
	Gemini 3 Flash	4.00 (4.00–4.00)	1.50		
Special clinical situations (n = 8)	ChatGPT 5.5	4.00 (3.75–5.00)	1.94	χ^2=9.172 (df = 2)	**0.010**
	Claude Sonnet 4.6	5.00 (5.00–5.00)	2.75		
	Gemini 3 Flash	3.50 (3.00–4.00)	1.31		
Advanced clinical decision-making (n = 6)	ChatGPT 5.5	3.00 (3.00–3.00)	1.25	χ^2=11.143 (df = 2)	**0.004**
	Claude Sonnet 4.6	**5.00 (5.00–5.00)**	**3.00**		
	Gemini 3 Flash	4.00 (3.25–4.00)	1.75		

Values are presented as median (interquartile range). Bold values denote the model with the highest mean rank within each category. Pairwise post-hoc comparisons (Wilcoxon signed-rank test with Bonferroni correction, adjusted threshold p < 0.017): AIES-CPV — Atypical ingestion: Claude Sonnet 4.6 vs ChatGPT 5.5 p = 0.016, Claude Sonnet 4.6 vs Gemini 3 Flash p = 0.016; Special clinical situations: Claude Sonnet 4.6 vs Gemini 3 Flash p = 0.008; Advanced clinical decision-making: Claude Sonnet 4.6 vs ChatGPT 5.5 p = 0.031, Claude Sonnet 4.6 vs Gemini 3 Flash p = 0.031. GQS — Acute ingestion: Claude Sonnet 4.6 vs Gemini 3 Flash p = 0.008, ChatGPT 5.5 vs Gemini 3 Flash p = 0.031; Atypical ingestion: Claude Sonnet 4.6 vs ChatGPT 5.5 p = 0.031, Claude Sonnet 4.6 vs Gemini 3 Flash p = 0.031; Special clinical situations: Claude Sonnet 4.6 vs Gemini 3 Flash p = 0.008; Advanced clinical decision-making: Claude Sonnet 4.6 vs ChatGPT 5.5 p = 0.031, Claude Sonnet 4.6 vs Gemini 3 Flash p = 0.031. Key-action checklist results are presented descriptively (Friedman p = 0.050–0.186 across categories). AIES-CPV, Artificial Intelligence Evaluation Score for Clinical Poisoning Vignettes; GQS, Global Quality Scale; IQR, interquartile range; LLM, large language model.

Discussion

In this study, three contemporary large language models were evaluated on 30 hypothetical pediatric paracetamol poisoning scenarios using a guideline-based reference standard and a key-action checklist, scored by a blinded expert-panel consensus. Claude Sonnet 4.6 performed significantly better than ChatGPT 5.5 and Gemini 3 Flash on the total AIES-CPV, key-action checklist, and GQS scores; no

significant difference was found between ChatGPT 5.5 and Gemini 3 Flash. Moreover, no critical safety error was observed for Claude Sonnet 4.6 and all of its responses were concordant with the reference standard, whereas three critical errors were recorded for Gemini 3 Flash and one for ChatGPT 5.5; however, the difference in critical-error rates did not reach statistical significance.

Our findings are consistent with the literature showing that LLM performance varies markedly by model, clinical domain, and even prompt language [4, 5]. Some studies report that Claude achieves the highest accuracy among models on standardized medical examination questions [11], whereas on an emergency medicine examination ChatGPT-4o led in overall accuracy but Claude showed the highest response consistency and Gemini 3 Flash was the most prone to hallucination [12]. In our study, Claude Sonnet 4.6 stood out particularly in completeness of information and clarity/applicability for the clinician and made no critical errors, which aligns with this picture of real, task-dependent differences among models.

The nature of the critical errors observed—inadequate evaluation of an ingestion of unknown amount, omission of laboratory evaluation in repeated supratherapeutic ingestion, and risk of premature termination of N-acetylcysteine therapy—is consistent with the “fluent but incorrect” output pattern that is of greatest concern in the clinical use of LLMs [5, 13]. In pediatric toxicology, such errors can lead directly to harm in dose calculation, NAC indication, or monitoring decisions. Altundağ et al. likewise found ChatGPT useful in simulated toxicology scenarios but reported important gaps [7]. Differences in hallucination and consistency among models [12] may partly explain why Gemini 3 Flash accumulated more critical errors.

In the pediatric population, the stand-alone diagnostic accuracy of LLMs has been shown to remain only moderate [14]; at the same time, parents increasingly trust and use these tools for health information about their children [15]. This increases the importance of accuracy in caregiver-facing pediatric advice. In a directly comparable, parent-focused evaluation of three contemporary models for childhood rickets that applied the same AIES-CPQ and GQS framework, newer-generation models (Gemini 3 Pro and DeepSeek V3.2) outperformed GPT-5.1, while inter-rater agreement ranged only from poor to fair—underscoring both the value of structured, multidisciplinary scoring and the variability inherent in expert appraisal of AI-generated content [8]. Rather than generic accuracy, a guideline-based [1] key-action evaluation such as ours is a more clinically meaningful measure. These models may be positioned as adjunctive tools for clinical decision support [6]; however, especially for critical decisions such as dose calculation, NAC indication, and serum-level timing, they should not be used as autonomous decision-makers without clinician oversight.

Category-based analysis revealed that Claude Sonnet 4.6 maintained its performance advantage across all four thematic domains, with the most pronounced margin observed in Advanced clinical decision-making scenarios. These scenarios — encompassing NAC management, laboratory interpretation, and co-ingestion cases — require the integration of multiple concurrent clinical variables and nuanced pharmacological reasoning, suggesting that performance differences between models are amplified in

settings of greater complexity. This pattern is consistent with prior observations that LLMs may produce more variable or superficial responses when faced with complex or multifactorial clinical queries, compared with straightforward presentations [13, 14]. A domain-specific pattern of performance was also reported in a parallel evaluation of three LLMs responding to parent-oriented questions about rickets, in which model rankings varied across thematic categories and newer-generation models demonstrated stronger results in clinically interpretative domains [8]. Collectively, these findings indicate that model selection in clinical decision-support contexts may be particularly consequential for high-acuity or pharmacologically complex scenarios.

Strengths of the study include the combination of a scenario-based design with a key-action and critical-safety-error framework, consensus evaluation blinded to model identity, and the reflection of the default public-interface user experience. Nevertheless, because LLMs are updated frequently, these findings are specific to the versions and the date evaluated [11, 12] and should be interpreted in light of the limitations summarized below.

Limitations

This study has several limitations. First, responses were scored by the structured consensus of a three-clinician panel rather than by an independent multi-rater design; although this approach reduces individual rater error, it does not allow inter-rater reliability to be reported quantitatively. Second, the AIES-CPV is a study-specific evaluation rubric and is not an externally validated scale. Third, a pronounced ceiling effect was observed, particularly for high-performing models (e.g., the total AIES-CPV interquartile range of 25–25 for Claude Sonnet 4.6); this may limit discrimination among models, and because of the skewed score distributions, non-parametric tests were preferred for the comparisons. Fourth, because critical safety errors were rare (0–3/30 across models), the Cochran Q test was underpowered for this comparison; therefore, the absence of a significant difference should be interpreted as a descriptive finding rather than as evidence that the models have equivalent safety. Fifth, the category-based subgroup analysis was conducted with a small number of scenarios per category (n = 6–9); within-category Friedman tests therefore have limited statistical power, and these results should be interpreted as exploratory and hypothesis-generating rather than confirmatory. Finally, because responses were obtained using specific model versions on a specific date, with single-turn prompting and a browser-based interface, the findings are time- and version-specific; external validity is also limited because the scenarios were hypothetical.

Conclusions

Among three large language models evaluated on pediatric paracetamol poisoning scenarios, Claude Sonnet 4.6 obtained the highest scores for guideline-based accuracy, key-action performance, and overall response quality and made no critical safety errors; no significant difference was found between ChatGPT 5.5 and Gemini 3 Flash. These results indicate that contemporary large language models have the potential to play an adjunctive role in pediatric toxicology decision support, but, because of inter-

model safety differences and the possibility of errors—particularly in high-risk decisions—should be used only under expert supervision and with clinical verification.

Abbreviations

AI
Artificial Intelligence
AIES-CPQ
Artificial Intelligence Evaluation Score for Common Patient Questions
AIES-CPV
Artificial Intelligence Evaluation Score for Clinical Poisoning Vignettes
API
Application Programming Interface
GQS
Global Quality Scale
INR
International Normalized Ratio
IQR
Interquartile Range
LLM
Large Language Model
NAC
N-acetylcysteine
SD
Standard Deviation

Declarations

Ethics approval and consent to participate.

This study did not involve human participants, identifiable human data, or biological samples. It evaluated responses of large language models to hypothetical pediatric paracetamol poisoning scenarios created by experienced pediatricians. Therefore, ethics approval and informed consent were not required.

Consent for publication.

Not applicable.

Competing interests.

The authors declare that they have no competing interests.

Funding.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Alper Kaçar contributed to the conceptualization and design of the study, development of the clinical scenarios, data collection, statistical analysis, interpretation of the results, and drafting of the manuscript. Belen Ateş contributed to the study design, expert evaluation of the large language model responses, interpretation of the findings, and critical revision of the manuscript from a pediatric perspective. Sevilay Kök contributed to the expert evaluation of the responses, interpretation of the clinical findings, and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgement

None

Availability of data and materials.

The datasets generated and analyzed during the current study—the scenario set, the anonymized model responses, and the consensus scoring sheets—are provided as supplementary files accompanying this article.

References

1. Chiew AL, Reith D, Pomerleau A, Wong A, Isoardi KZ, Soderstrom J, Greene SL, et al. Updated guidelines for the management of paracetamol poisoning in Australia and New Zealand. Med J Aust. 2020;212(4):175-183. doi:10.5694/mja2.50428.
2. Royal Children's Hospital Melbourne. Clinical Practice Guidelines: Paracetamol poisoning. Melbourne: Royal Children's Hospital Melbourne. https://www.rch.org.au/clinicalguide/guideline_index/paracetamol_poisoning/. Accessed 9 Jun 2026.
3. Children's Health Queensland Hospital and Health Service. Queensland Paediatric Emergency Care: Paracetamol ingestion – emergency management in children. Brisbane: Children's Health Queensland. https://www.childrens.health.qld.gov.au/for-health-professionals/queensland-

paediatric-emergency-care-qpec/queensland-paediatric-clinical-guidelines/paracetamol-ingestion. Accessed 9 Jun 2026.

4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8.
5. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls. Ann Intern Med. 2024;177(2):210-220. doi:10.7326/M23-2772.
6. Oniani D, Hilsman J, Peng Y, Poropatich RK, Pamplin JC, Legault GL, Wang Y, et al. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. NPJ Digit Med. 2024;7:105. doi:10.1038/s41746-024-01081-0.
7. Altundağ İ, Çakmak M, Uysalol E, Aksay E. Assessing the performance of ChatGPT in medical toxicology through simulated case scenarios. Glob Emerg Crit Care. 2024;3(3):132-139.
8. Çörekci AM, Ateş B, Öztürk AA, Özdemir M. Performance of three large language models in answering parent-focused questions on rickets: a dual pediatric–orthopedic specialist evaluation. BMC Pediatr. 2026;26(1):525. doi:10.1186/s12887-026-06851-1.
9. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483.
10. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102(9):2070-2077.
11. Wójcik D, Adamiak O, Czerepak G, Tokarczuk O, Szalewski L. Comparing the performance of ChatGPT, Gemini, and Claude in English and Polish on medical examinations. Sci Rep. 2025;15:33083. doi:10.1038/s41598-025-17030-0.
12. Şan İ, Akkan Öz M, Yortanlı M, Genç M, Bulut B, Gür A, et al. AI performance in emergency medicine fellowship examination: comparative analysis of ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 models. Turk J Med Sci. 2025;55(5). doi:10.55730/1300-0144.6083.
13. Schwartz IS, Link KE, Daneshjou R, Cortés-Penfield N. Black box warning: large language models and the future of infectious diseases consultation. Clin Infect Dis. 2024;78(4):860-866. doi:10.1093/cid/ciad633.
14. Barile J, Margolis A, Cason G, Kim R, Kalash S, Tchaconas A, et al. Diagnostic accuracy of a large language model in pediatric case studies. JAMA Pediatr. 2024;178(3):313-315. doi:10.1001/jamapediatrics.2023.5750.
15. Tan J, Wang L, Wang G, Yang Y, Jia F, Chi X, et al. Safety and user perception of general-purpose large language models in pediatric healthcare: evaluations of ChatGPT by doctors and parents. Digit Health. 2026;12. doi:10.1177/20552076261427505.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1ConsensusScoringSheets.xlsx
- YZ1.docx
- YZ2.docx
- YZ3.docx