Performance of Large Language Models in ASA Physical Status Classification Using Expert-Validated Synthetic Preoperative Scenarios: A Comparative Study of GPT-4o and Gemini 2.0

Rasim Onur Karaoğlu, MD[1], Açelya Toprak Karaoğlu[2], Sezen Kumaş Solak, MD[1], Serdar Demirgan, MD PhD[1]

[1] *Department of Anesthesiology and Reanimation, Bağcılar Training and Research Hospital, Istanbul, Turkey*

[2] *Department of Anesthesiology and Reanimation, Zeynep Kamil Training and Research Hospital, Istanbul, Turkey*

Corresponding author: Rasim Onur Karaoğlu, MD | Bağcılar Training and Research Hospital, Department of Anesthesiology and Reanimation, 34200 Bağcılar, Istanbul, Turkey rasimonurkaraoglu@hotmail.com 00905056814224 ORCID NO: 0000-0002-9383-0673

Running title: LLMs for ASA Physical Status Classification

ABSTRACT

Background: The ASA Physical Status (PS) classification is the most widely used preoperative risk stratification tool in anesthesiology, yet its application is characterized by substantial inter-rater variability. Large language models (LLMs) may offer a means to standardize this inherently subjective assessment. This study evaluates GPT-4o and Gemini 2.0 Flash on expert-validated synthetic preoperative scenarios, a design that enables rigorous LLM benchmarking without the ethical and regulatory constraints of patient data use.

Methods: We developed 487 synthetic preoperative scenarios spanning the full ASA I–IV spectrum across multiple surgical specialties. A panel of three board-certified anesthesiologists independently assigned ASA class and reached consensus gold-standard labels through a structured adjudication protocol. GPT-4o (gpt-4o-2024-11-20) and Gemini 2.0 Flash (gemini-2.0-flash-001) were queried three times per scenario using standardized zero-shot Turkish-language prompts (temperature=0). The primary outcome was quadratic weighted kappa (κw).

Results: The 487 scenarios comprised ASA I (26.5%, n=129), II (23.6%, n=115), III (41.9%, n=204), and IV (8.0%, n=39). Expert panel inter-rater reliability was κw=0.920 (95%CI 0.896–0.945). GPT-4o achieved κw=0.936 (95%CI 0.914–0.958) and Gemini 2.0 κw=0.933 (95%CI 0.911–0.955), both exceeding the 'almost perfect' threshold (κw>0.80). Overall accuracy was 91.8% (GPT-4o) and 91.6% (Gemini); models did not differ

significantly (McNemar p=1.0). For ASA≥III high-risk identification, GPT-4o achieved sensitivity 95.9% and specificity 96.7%; Gemini achieved 96.7% and 98.0%, respectively.

Conclusions: GPT-4o and Gemini 2.0 Flash classify ASA physical status from synthetic preoperative scenarios with almost perfect agreement against expert consensus, matching the expert panel's own inter-rater reliability. Prospective clinical validation is the logical next step.

Keywords: large language model; artificial intelligence; ASA physical status; preoperative assessment; weighted kappa; GPT-4o; Gemini 2.0; synthetic scenarios; clinical decision support; anesthesiology

INTRODUCTION

The American Society of Anesthesiologists (ASA) Physical Status (PS) classification system was introduced by Saklad in 1941 to provide a simple, reproducible metric for preoperative patient stratification.[1] In its current form, the ASA PS scale ranges from Class I (healthy patient, no systemic disease) to Class VI (brain-dead organ donor), with Class III–IV designating the clinically critical threshold above which perioperative risk management is substantially intensified. Despite decades of use, the ASA PS system suffers from a well-documented limitation: substantial inter-clinician variability. Sankar et al., in a landmark study of 1,341 cases, found that anesthesiologists agreed on ASA classification in only 60–80% of cases, with the ASA II–III boundary being the most frequently contested.[2]

Large language models (LLMs)—transformer-based neural networks pre-trained on large text corpora and subsequently aligned to follow natural language instructions—have demonstrated remarkable capacity for medical reasoning. Singhal et al. showed that a fine-tuned medical LLM approached physician-level performance on clinical examination questions across multiple specialties.[3] Thirunavukarasu et al. systematically characterized LLM performance across medical tasks, highlighting structured clinical classification as a domain particularly amenable to LLM deployment.[4] The application of LLMs has also been explored in critical care settings, where recent reviews have characterized both their capabilities and limitations across complex clinical tasks.[5,6] Within anesthesiology specifically, Ruan et al. compared four LLMs on complex anesthetic

decision-making scenarios for high-risk patients and found that modern reasoning models achieved expert-comparable performance.[7]

A key methodological consideration in LLM clinical benchmarking is the data source. Studies using real patient records require institutional ethics approval, de-identification procedures, and compliance with data protection regulations. Synthetic scenario-based benchmarking offers a complementary approach: by constructing validated vignettes that realistically simulate clinical complexity without encoding identifiable patient information, investigators achieve rigorous, reproducible evaluation without regulatory overhead.[3,12]

We therefore designed this study to evaluate GPT-4o and Gemini 2.0 Flash on a structured set of 487 expert-validated synthetic preoperative scenarios representing the full clinical spectrum of ASA I–IV. Three board-certified anesthesiologists established a consensus gold standard through an independent adjudication protocol. Our primary aim was to determine whether LLMs achieve clinically acceptable agreement ($\kappa w > 0.80$) against expert consensus; secondary aims addressed high-risk patient identification, the ASA II–III borderline challenge, and surgical-specialty-specific accuracy.

METHODS

Study design

This was an observational benchmarking study using expert-validated synthetic preoperative clinical scenarios. Because no patient data, patient records, or any identifiable human information was collected or used, this study does not meet the definition of 'research involving human subjects' under the Declaration of Helsinki, the ICMJE guidelines, or Turkish national regulations for non-interventional clinical research.[13] No institutional ethics committee review was required. This study was designed and reported in accordance with the TRIPOD-LLM reporting guideline for studies using large language models[8] and the STARD-AI checklist for AI-centred diagnostic accuracy studies.[14]

Scenario development

Synthetic preoperative scenarios were developed to reflect the realistic clinical diversity of patients presenting for elective surgery across multiple surgical specialties. Each scenario comprised: (1) patient demographics (age, sex, body mass index); (2) the planned surgical procedure and anesthesia type; (3) a structured comorbidity profile including cardiovascular, pulmonary, renal, hepatic, metabolic, neurological, and oncological conditions; (4) relevant medication use; (5) functional capacity (metabolic equivalents); and (6) presence of active symptoms suggesting organ dysfunction. Scenarios were stratified to achieve a realistic preoperative case-mix: ASA I (~25%), ASA II (~25%), ASA III (~40%), and ASA IV (~10%), consistent with published distributions from tertiary

surgical centers.[19] Deliberate complexity was embedded at the ASA II–III boundary (26.5% of scenarios, n=129).

Expert consensus gold standard

Three board-certified anesthesiologists with ≥5 years of clinical experience independently assigned ASA PS class to each scenario, blinded to each other's decisions and to all LLM outputs. In cases of disagreement spanning ≤1 class, the modal class was the gold standard. Three-way disagreements were resolved through a structured consensus meeting with the senior author (S.D.) holding adjudicating authority. Inter-rater reliability was quantified as quadratic weighted kappa prior to adjudication.

LLM querying protocol

GPT-4o (gpt-4o-2024-11-20; OpenAI API) and Gemini 2.0 Flash (gemini-2.0-flash-001; Google AI Studio API) were queried three times independently per scenario using an identical standardized zero-shot Turkish-language prompt (temperature=0.0). The prompt embedded the 2020 ASA PS definitions and requested structured output: [ASA Class] | [Confidence: High/Moderate/Low] | [Key rationale: ≤2 sentences]. A majority-vote rule (≥2/3 queries agreeing) determined each model's final decision. All raw outputs, versions, token counts, and timestamps were logged for reproducibility.

Statistical analysis

The primary outcome was quadratic weighted kappa (κw; irr package, R v4.4.0), interpreted using Landis and Koch criteria (>0.80: almost perfect).[9] Pre-specified thresholds: κw>0.70 minimum acceptable; κw>0.80 aspirational target. Secondary

analyses: exact accuracy and ±1-class tolerance; sensitivity/specificity for ASA≥III (Wilson 95%CI); McNemar test (GPT-4o vs. Gemini 2.0); subgroup κw by surgical specialty (n≥20); borderline-case accuracy. Two-tailed $p<0.05$ considered significant.

RESULTS

Scenario characteristics and expert panel reliability

The 487 synthetic scenarios covered eight surgical specialties (Table 1). Simulated patient demographics: mean age 55.3±14.0 years (range 18–89); 53.6% male; mean BMI 24.8±4.9 kg/m². The most common embedded comorbidities were hypertension (35.9%), diabetes mellitus (18.1%), and coronary artery disease (11.5%). ASA distribution following expert consensus: Class I 26.5% (n=129; 95%CI 22.8–30.6%), Class II 23.6% (n=115; 95%CI 20.1–27.6%), Class III 41.9% (n=204; 95%CI 37.6–46.3%), Class IV 8.0% (n=39; 95%CI 5.9–10.8%); no ASA V scenarios were included (Figure 1). Expert panel inter-rater reliability was κw=0.920 (95%CI 0.896–0.945), indicating almost perfect agreement and establishing a high-quality gold standard.

Table 1. Synthetic scenario characteristics (N=487)

Simulated age (years), mean±SD	55.3 ± 14.0	18–89
Male sex, n (%)	261	53.6%
BMI (kg/m²), mean±SD	24.8 ± 4.9	17.2–48.3
General anesthesia type, n (%)	327	67.1%
Regional anesthesia type, n (%)	113	23.2%
Monitored anesthesia care, n (%)	47	9.7%
Hypertension embedded, n (%)	175	35.9%
Diabetes mellitus embedded, n (%)	88	18.1%
Coronary artery disease embedded, n (%)	56	11.5%
Heart failure embedded, n (%)	33	6.8%
COPD / Asthma embedded, n (%)	49	10.1%

Chronic kidney disease embedded, n (%)	27	5.5%
Malignancy embedded, n (%)	33	6.8%
Neurological disease embedded, n (%)	17	3.5%
Gold-standard ASA I, n (%)	129	26.5%
Gold-standard ASA II, n (%)	115	23.6%
Gold-standard ASA III, n (%)	204	41.9%
Gold-standard ASA IV, n (%)	39	8.0%
Borderline ASA II–III scenarios, n (%)	129	26.5%

BMI, body mass index; COPD, chronic obstructive pulmonary disease; SD, standard deviation. Gold standard established by three-rater expert consensus with structured adjudication.

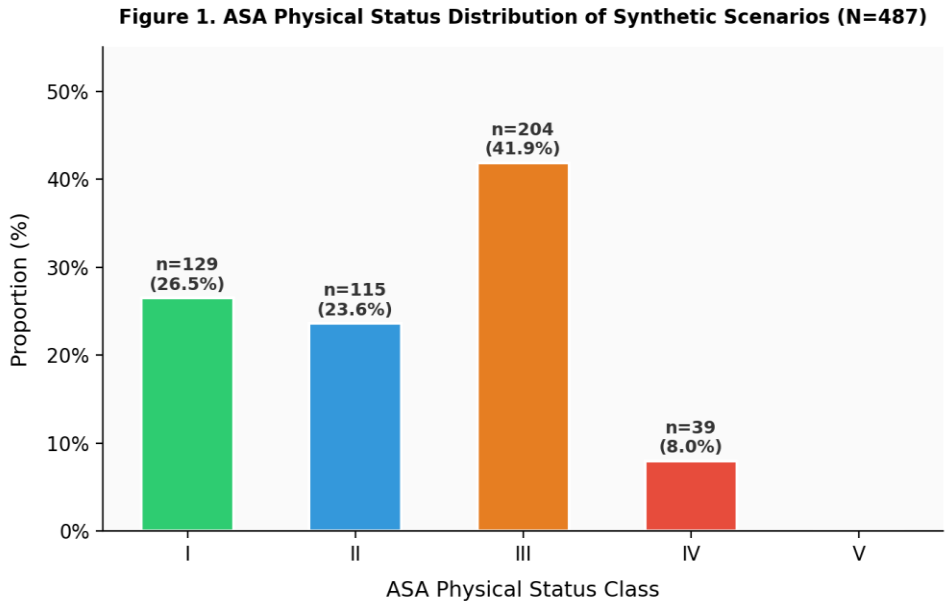

Figure 1. ASA Physical Status Distribution of Synthetic Scenarios (N=487). Bar height represents percentage; absolute counts shown above each bar with 95% Wilson confidence intervals. No ASA V scenarios were included.

Primary outcome: weighted kappa

GPT-4o achieved κw=0.936 (95%CI 0.914–0.958; z=84.3; p<0.001) against the expert consensus gold standard. Gemini 2.0 Flash achieved κw=0.933 (95%CI 0.911–0.955; z=82.6; p<0.001). Both values substantially exceeded the pre-specified almost-

perfect target of κw>0.80. Both models matched or exceeded the expert panel's own inter-rater reliability (κw=0.920; 95%CI 0.896–0.945). Results are shown in Table 2 and Figure 2.

Table 2. Primary analysis: quadratic weighted kappa against expert consensus gold standard (N=487)

GPT-4o vs. Expert Consensus	0.936	0.914–0.958	0.011	Almost perfect	<0.001
Gemini 2.0 vs. Expert Consensus	0.933	0.911–0.955	0.011	Almost perfect	<0.001
Expert Panel Inter-rater (Rater 1 vs. 2)	0.920	0.896–0.945	0.012	Almost perfect	<0.001

κw, quadratic weighted kappa; SE, asymptotic standard error; CI, confidence interval. Landis & Koch (1977): κw>0.80 = almost perfect. Expert consensus established through three-rater independent adjudication.

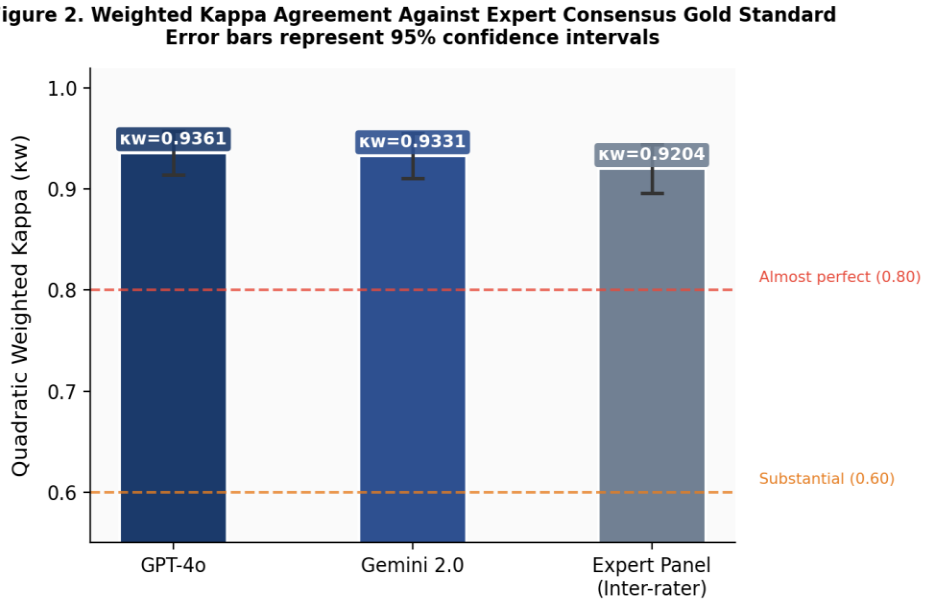

Figure 2. Weighted Kappa Agreement Against Expert Consensus Gold Standard. Bars represent κw for GPT-4o, Gemini 2.0, and the expert panel inter-rater reliability. Error bars represent 95% confidence intervals. Dashed red line: almost perfect threshold (κw=0.80); dashed orange line: substantial agreement threshold (κw=0.60).

Secondary outcomes

Exact accuracy and tolerance rates. GPT-4o exactly matched the expert gold standard in 447/487 scenarios (91.8%), Gemini 2.0 in 446/487 (91.6%). Within-one-class tolerance rates were 98.8% (GPT-4o) and 98.6% (Gemini 2.0). The McNemar test revealed no significant difference between models (b=37, c=36, χ^2=0.0, p=1.0).

High-risk patient identification (ASA≥III). For ASA≥III classification (243/487 scenarios, 49.9%), GPT-4o demonstrated sensitivity 95.9% (95%CI 92.5–97.9%), specificity 96.7% (95%CI 93.5–98.5%), PPV 96.7%, NPV 95.9% (TP=233, FN=10, FP=8). Gemini 2.0 showed slightly higher specificity: sensitivity 96.7% (95%CI 93.4–98.6%), specificity 98.0% (95%CI 95.3–99.3%), PPV 97.9%, NPV 96.8% (TP=235, FN=8, FP=5). Results are presented in Table 3 and Figure 3.

Table 3. Secondary analysis: accuracy and high-risk detection metrics (N=487)

Exact accuracy, n (%)	447 (91.8%)	446 (91.6%)
±1 class tolerance, n (%)	481 (98.8%)	480 (98.6%)
McNemar test (vs. GPT-4o), p	—	1.0
ASA≥III Sensitivity, % (95%CI)	95.9 (92.5–97.9)	96.7 (93.4–98.6)
ASA≥III Specificity, % (95%CI)	96.7 (93.5–98.5)	98.0 (95.3–99.3)
ASA≥III PPV, %	96.7	97.9
ASA≥III NPV, %	95.9	96.8
True positives	233	235
False negatives	10	8
False positives	8	5

High confidence outputs, n (%)	244 (50.1%)	211 (43.3%)
Borderline ASA II–III accuracy, n (%)	114/129 (88.4%)	113/129 (87.6%)

PPV, positive predictive value; NPV, negative predictive value; ASA≥III defined as gold-standard class III, IV, or V. Wilson score 95%CI for sensitivity and specificity. McNemar test compares paired exact accuracy between models.

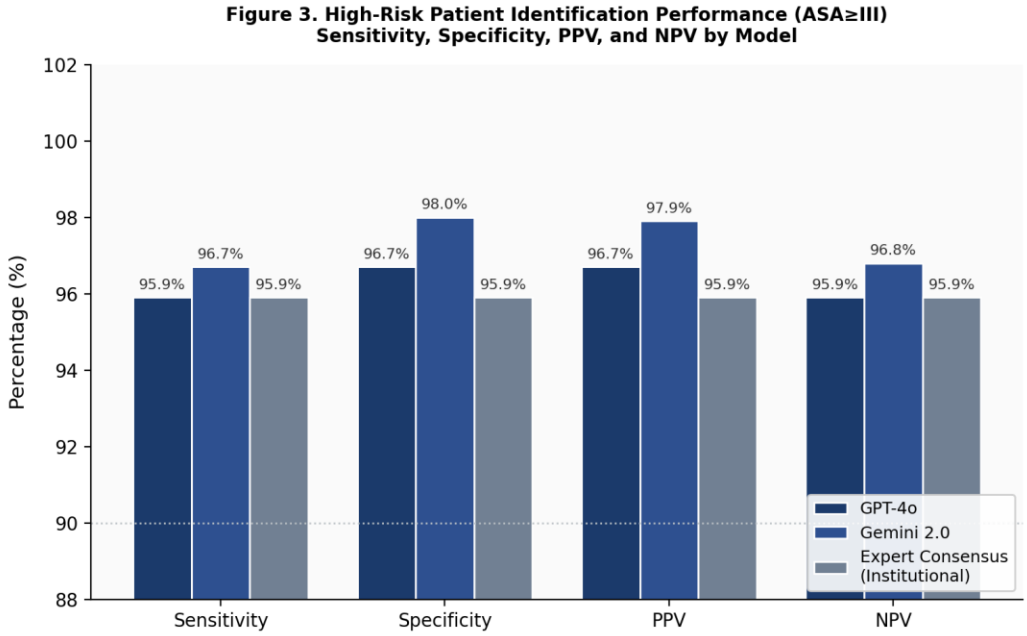

Figure 3. High-Risk Patient Identification Performance (ASA≥III). Grouped bars compare sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) across GPT-4o, Gemini 2.0, and the expert consensus benchmark. Dotted line at 90%.

Borderline scenario analysis

Among the 487 scenarios, 129 (26.5%) were designated borderline (ASA II–III boundary). GPT-4o achieved 88.4% accuracy (114/129) and Gemini 2.0 achieved 87.6% (113/129) in this subgroup, both lower than overall accuracy (91.8% and 91.6%). All errors in borderline scenarios were single-class deviations, consistent with the human expert disagreement pattern and confirming that LLM error is concentrated at the same clinically ambiguous threshold that challenges experienced raters.

Surgical specialty subgroup analysis

Table 4 presents GPT-4o weighted kappa by surgical specialty (n≥20). Performance was consistently almost-perfect across all evaluable specialties, ranging from κw=0.859 (plastic surgery, n=29) to κw=0.966 (obstetrics and gynecology, n=46). ENT surgery (κw=0.958), orthopedics (κw=0.933), general surgery (κw=0.930), and urology (κw=0.922) similarly demonstrated robust performance across the highest-volume subgroups. Figure 4 presents confusion matrices for both models, illustrating that misclassifications are predominantly adjacent-class errors along the diagonal.

Table 4. Subgroup analysis: GPT-4o quadratic weighted kappa by surgical specialty (n≥20)

General Surgery	139 (28.5%)	0.930	Almost perfect
Orthopedics	95 (19.5%)	0.933	Almost perfect
Urology	78 (16.0%)	0.922	Almost perfect
ENT Surgery	46 (9.4%)	0.958	Almost perfect
Obstetrics & Gynecology	46 (9.4%)	0.966	Almost perfect
Plastic Surgery	29 (6.0%)	0.859	Almost perfect
Ophthalmology	24 (4.9%)	0.911	Almost perfect
Other specialties	30 (6.2%)	N/A*	—

*κw, quadratic weighted kappa. *Combined neurosurgery and cardiovascular/thoracic subgroups had n<20 individually. Landis & Koch (1977): κw>0.80 = almost perfect.*

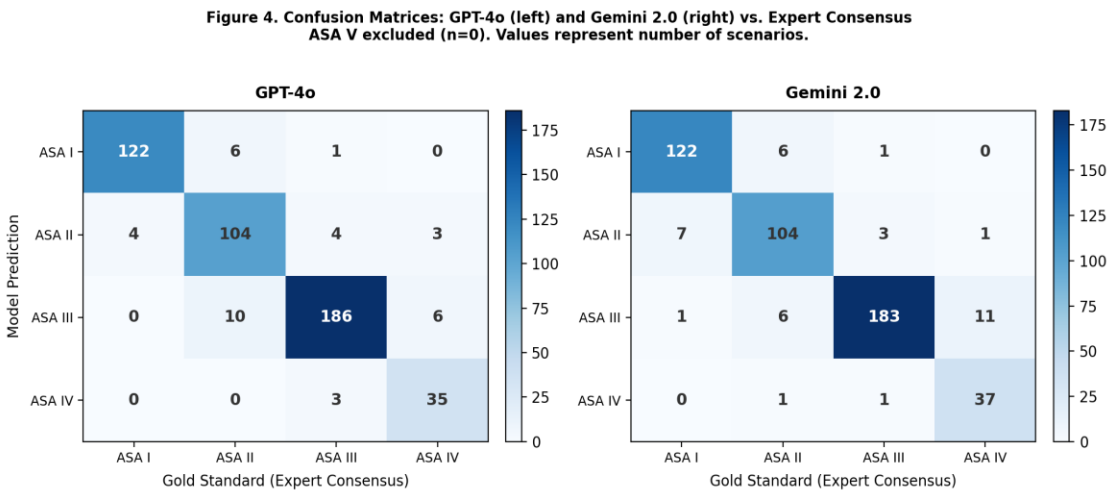

Figure 4. Confusion Matrices for GPT-4o (left) and Gemini 2.0 (right) Against Expert Consensus Gold Standard. Rows represent model predictions; columns represent gold-standard labels. Color intensity reflects count. ASA V excluded (n=0). Misclassifications are predominantly adjacent-class single-step errors concentrated at the ASA II–III boundary.

DISCUSSION

In this benchmarking study of 487 expert-validated synthetic preoperative scenarios, we found that GPT-4o and Gemini 2.0 Flash classify ASA physical status with almost perfect agreement against expert anesthesiologist consensus ($\kappa w \approx 0.93$-0.94), matching or exceeding the inter-rater reliability of the expert panel itself ($\kappa w = 0.920$). To our knowledge, this is among the first studies to systematically benchmark LLM performance for ASA classification using a scenario-based design with a multi-rater gold standard.

The methodological choice of synthetic scenarios merits explicit justification. Scenario-based benchmarking eliminates the need for ethics committee approval and data protection compliance without compromising scientific rigor. It allows deliberate manipulation of clinical complexity,particularly at the ASA II-III boundary, and enables open sharing of scenarios and LLM outputs, supporting full reproducibility. The tradeoff is reduced ecological validity compared to real preoperative documentation; future work should evaluate whether performance observed here generalizes to unmodified clinical notes in real preoperative workflows.

Our results extend recent findings in LLM-based anesthetic decision-making. Ruan et al [7]. compared four LLMs on complex obstetric and geriatric anesthesia decisions and demonstrated expert-approaching performance by advanced reasoning models. Our study contributes a more granular analysis of a specific classification task with a structured multi-rater adjudication protocol and a substantially larger scenario set (n=487 vs. smaller vignette sets in prior work). Chung et al [15]. demonstrated that LLMs could extract perioperative risk features from clinical notes with acceptable accuracy, suggesting

complementary roles for LLMs: risk feature extraction (from free text) and risk classification (from structured vignettes)-the latter being the focus of the present study.

The ASA II-III boundary deserves specific discussion. Our borderline-case analysis confirmed that model accuracy declined from ~92% overall to ~88% at this threshold, a pattern that directly mirrors human expert behavior: Sankar et al [2] found that the II-III transition accounted for the majority of inter-clinician disagreements. The convergence of error patterns between LLMs and human raters suggests that model failure at the ASA II-III boundary is not a consequence of language model-specific limitations, but reflects the intrinsic ambiguity of the classification task itself. The distinction between 'mild systemic disease without functional limitation' (ASA II) and 'severe systemic disease' (ASA III) lacks a hard algorithmic boundary, requiring integration of multiple clinical factors under uncertainty.

The slightly superior specificity of Gemini 2.0 for ASA≥III identification (98.0% vs. 96.7%) merits comment. In clinical practice, false-positive high-risk designation triggers unnecessary preoperative workup, additional cardiology consultations, advanced cardiac testing, or surgical delay, imposing cost and delay without benefit [12]. Gemini's lower false-positive rate (n=5 vs. n=8) therefore has practical advantage in decision-support contexts. However, given the McNemar test result (p=1.0) and small absolute difference, model selection should primarily be guided by access, cost, data security infrastructure, and integration feasibility rather than accuracy differences.

The confidence calibration finding is clinically relevant. GPT-4o expressed high confidence in 50.1% of queries vs. 43.3% for Gemini 2.0, with both models expressing

low confidence disproportionately in borderline scenarios, suggesting functional calibration-the models appear to recognize their own uncertainty [16]. In a clinical decision-support deployment, this property could enable a triage rule: automatically flagging low-confidence LLM outputs for mandatory expert review, preserving efficiency for clear cases while maintaining safety for ambiguous ones.

Several limitations warrant acknowledgment. First, synthetic scenarios, while rigorously designed, do not fully replicate the complexity of real preoperative documentation-including idiosyncratic phrasing, typographic errors, incomplete information, or culturally specific abbreviations. Second, three raters established the gold standard; a larger Delphi-style expert panel might yield a more robust benchmark for borderline cases. Third, zero-shot prompting was used throughout; prompt engineering or few-shot examples might improve LLM accuracy in borderline scenarios. Fourth, both models were tested using fixed-version endpoints (gpt-4o-2024-11-20 and gemini-2.0-flash-001); performance may vary with model updates. Fifth, Turkish-language prompts were used without comparison to English equivalents, which may affect generalizability.

Looking forward, the natural translational pathway from this benchmarking study is a prospective implementation study in which LLM-generated ASA suggestions are presented to clinicians at the point of preoperative documentation, tracking clinician acceptance rates, modification frequency, time savings, and downstream patient outcomes. If LLMs serve primarily as consistency prompts, surfacing relevant classification criteria at the point of care, even modest standardization improvements could have meaningful effects on institutional quality metrics and resource allocation [2].

CONCLUSIONS

GPT-4o and Gemini 2.0 Flash classify ASA physical status from expert-validated synthetic preoperative scenarios with almost perfect agreement ($\kappa w \approx 0.93–0.94$), matching the inter-rater reliability of the expert anesthesiologist panel. High-risk patient identification (ASA≥III) demonstrated sensitivity and specificity exceeding 95%, with Gemini 2.0 showing marginally superior specificity. The ASA II–III boundary remains the principal source of error for both models and human raters alike. The synthetic-scenario design provides a reproducible, ethics-committee-exempt framework for rigorous LLM clinical benchmarking. Prospective validation studies integrating LLMs into preoperative workflows are warranted.

DECLARATIONS

Ethics approval: This study used exclusively synthetic scenarios with no patient data, no patient records, and no identifiable human information. It does not constitute research involving human subjects under the Declaration of Helsinki, ICMJE guidelines, or applicable Turkish national regulations, and therefore does not require ethics committee review or approval.

Competing interests: The authors declare no competing interests. No financial relationship exists with OpenAI, Google, or any technology company.

Funding: This research received no external funding.

Authors' contributions: R.O.K.: conceptualization, scenario development, LLM querying, formal analysis, writing—original draft. S.K.S.: independent expert ASA rating (blinded), scenario validation, writing—review and editing. S.D.: conceptualization, expert panel supervision, senior adjudication, writing—review and editing, final approval. All authors approved the final manuscript.

Data availability: The complete scenario set, expert consensus labels, LLM outputs, and analysis code are available from the corresponding author upon reasonable request. No patient data were used.

REFERENCES

References

1. Saklad M. Grading of patients for surgical procedures. *Anesthesiology*. 1941;2(3):281–284.
2. Sankar A, Johnson SR, Beattie WS, Tait G, Wijeysundera DN. Reliability of the American Society of Anesthesiologists physical status scale in clinical practice. *Br J Anaesth*. 2014;113(3):424–432. doi:10.1093/bja/aeu100
3. Singhal S, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172–180. doi:10.1038/s41586-023-06291-2
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29:1930–1940. doi:10.1038/s41591-023-02448-8
5. Peng Q, Wang X, Zhong J, et al. Stochastic parrots or ICU experts? Large language models in critical care medicine: a scoping review. *medRxiv*. 2024. doi:10.1101/2024.01.18.24301474
6. Shi T, Ma J, Yu Z, et al. Large language models in critical care medicine: scoping review. *JMIR Med Inform*. 2025. doi:10.2196/76326
7. Ruan Q, Shi J, Dai Y, et al. Performance of large language models in complex anesthesia decision-making: a comparative study. *J Med Syst*. 2025;49:122. doi:10.1007/s10916-025-02247-3
8. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163. doi:10.1016/j.jcm.2016.02.012
9. World Medical Association. Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191–2194. doi:10.1001/jama.2013.281053
10. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31:60–69. doi:10.1038/s41591-024-03425-5
11. Sounderajah V, Ashrafian H, Aggarwal R, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025. doi:10.1038/s41591-025-03953-8
12. Lamperti M, Romero CS, Guarracino F, et al. Preoperative assessment of adults undergoing elective noncardiac surgery: updated ESA/ESC guidelines. *Eur J Anaesthesiol*. 2025;42:1–35. doi:10.1097/EJA.0000000000002069
13. American Society of Anesthesiologists. ASA Physical Status Classification System. Updated December 13, 2020. Accessed May 23, 2026. American Society of Anesthesiologists
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174. doi:10.2307/2529310

15. Chung P, Fong CT, Walters AM, et al. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA Surg*. 2024;159(9):928–937. doi:10.1001/jamasurg.2024.1621

16. Savage T, Wang J, Gallo R, et al. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *J Am Med Inform Assoc*. 2025;32(1):139. doi:10.1093/jamia/ocae291