frontiers in Medicine

TYPE Original Research
PUBLISHED 03 November 2026
DOI 10.3389/fmed.2026.2003897

Check for updates

OPEN ACCESS

EDITED BY
Shahab Saquib Sohail,
Jamia Hamdard University, India

REVIEWED BY
Deepak Kumar,
Chandigarh University, India
Wu Gong,
Travere Therapeutics, United States

*CORRESPONDENCE
Hsuan-Ming Lin
vierylin@gmail.com

RECEIVED 19 September 2026
REVISED 06 October 2026
ACCEPTED 13 October 2026
PUBLISHED 03 November 2026

CITATION
Yeh S-E, Lin H-J, Lai W-W and Lin H-M (2026) Boundary-stratified safety evaluation of large language models for clinical decision support: a synthetic benchmark in chronic kidney disease renoprotective therapy.
Front. Med. 13:2003897.
doi: 10.3389/fmed.2026.2003897

COPYRIGHT
© 2026 Yeh, Lin, Lai and Lin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Boundary-stratified safety evaluation of large language models for clinical decision support: a synthetic benchmark in chronic kidney disease renoprotective therapy

Shang-En Yeh[1], Hsuan-Jen Lin[2], Wu-Wei Lai[1] and Hsuan-Ming Lin[1,3]*

[1]Division of Nephrology, Department of Internal Medicine, An Nan Hospital, China Medical University, Tainan, Taiwan, [2]Division of Nephrology, Department of Internal Medicine, Asia University Hospital, Taichung, Taiwan, [3]Institute of Information Management, National Cheng Kung University, Tainan, Taiwan

Objective: Large language models (LLMs) are increasingly proposed for medication-review support, but their behaviour at guideline-uncertain decision boundaries is rarely examined. Aggregate accuracy can hide model-specific divergence at such boundaries. We characterised boundary-zone rule adherence, replicate-query reproducibility, and reasoning-error rate of four frontier LLMs reviewing renoprotective-therapy omissions across the chronic kidney disease (CKD) severity spectrum.

Materials and Methods: We constructed 100 synthetic CKD vignettes (G3a–G5D; 75 with prespecified omissions, 25 decoys). Four LLMs were queried three times each at temperature 0 (1,200 calls). Omission criteria were operational adaptations of KDIGO 2024. They included an investigator-defined rule for de novo renin–angiotensin system inhibitor (RASi) initiation at eGFR < 15, a zone not covered by KDIGO 2024 recommendations. Two nephrologists applied the criteria to a stratified 20-case subset. In a post hoc analysis, every boundary non-detection was re-read and its stated rationale coded, and sensitivity was recomputed under a KDIGO-concordant reference standard.

Results: All models detected 97%–100% of SGLT2 inhibitor and finerenone omissions. At the eGFR < 15 RASi boundary (n = 20), adherence to the operational rule was 85% (Grok 4.1 Fast), 55% (GPT-5.4), 10% (Gemini 3.1 Pro), and 10% (DeepSeek R1). All 48 non-detections stated an explicit reason. GPT-5.4 withheld RASi only in the 9 dialysis-dependent cases and detected all 11 non-dialysis cases. Gemini and DeepSeek also withheld RASi in non-dialysis G5, citing hyperkalaemia or acute kidney injury risk. In a post hoc analysis excluding the 20 boundary cases, sensitivity was 96.5%–100% for every model. Rubric-application agreement was perfect (kappa = 1.000). Reasoning-defect rates ranged from 0% (GPT-5.4) to 27% (DeepSeek R1), and replicate Jaccard similarity from 0.74 to 0.93.

Conclusion: Between-model differences were concentrated where guidelines are silent. There, models applied distinct and explicitly stated decision policies rather than overlooking the drug class. Lower adherence to the study rule should not be read as lower clinical safety. Pre-deployment evaluation of clinical LLMs should report boundary-stratified behaviour, replicate reproducibility, and reasoning errors alongside aggregate accuracy, and should be validated against real-world prescribing.

KEYWORDS

artificial intelligence, chronic kidney disease, clinical decision support, large language model, medication review, patient safety, renoprotective therapy

1 Introduction

Clinical artificial-intelligence tools are being adopted for medication-review support faster than evaluation standards are developing. A large language model (LLM) with high aggregate accuracy on a benchmark may be treated as safe by downstream quality-assurance processes. Aggregate measures, however, can conceal behaviour that is concentrated in narrow clinical subgroups.

Subgroup-concentrated failures have repeatedly been hidden inside apparently well-performing clinical AI. Racial bias in pulse oximetry produced hypoxaemia mismeasurement in Black patients that went unrecognised in aggregate device performance for decades (1). A widely deployed sepsis-prediction model missed the majority of septic patients on external validation despite in-vendor performance claims (2). Chest-radiograph deep-learning models generalised poorly across hospitals because they had latched onto site-specific shortcuts rather than pathology (3). A commercial health-management algorithm systematically under-referred Black patients to high-risk care programs while performing well by its own designed metric (4). In each case an aggregate metric certified safety while the harm concentrated in a subpopulation the metric could not detect—classical latent errors in Reason's Swiss-Cheese model of accident causation (5).

Multi-model evaluations of LLMs in other clinical domains show the same pattern. In 2,400 real patient cases with four common abdominal pathologies, LLMs followed neither diagnostic nor treatment guidelines consistently, and their output depended on the quantity and order of information (6). Across 10,000 emergency department visits, GPT-3.5 and GPT-4 were overly cautious, gaining sensitivity at the cost of specificity (7). Four commercial LLMs reproduced race-based medical misconceptions and answered identical questions inconsistently (8). Two vision–language models were less accurate than radiologists on image-based diagnostic cases, with accuracy varying by subspecialty (9). In each study, the clinically important behaviour became visible only when performance was broken down by model, task, or subgroup.

Pre-deployment silent trials—running a model in the background without acting on its output—have emerged as one defensive layer, catching drift and dataset shift before clinical exposure (10). Yet silent trials do not decompose the failure surface into interpretable clinical regions, and clinicians reviewing LLM outputs at the bedside still need to know: in which decision regions does this model behave as intended, where does it diverge, and how would that divergence look at the point of care? For general-purpose LLMs whose training and prompt-response behaviour cannot be fully audited, upstream boundary-zone probing becomes even more important, because the downstream defensive layers of quality assurance—Root Cause Analysis, Failure Mode and Effects Analysis, alert-fatigue monitoring (11)—have less mechanistic understanding to work from.

Renoprotective medication review in chronic kidney disease (CKD) is a decision domain where LLM-assisted tools are actively being proposed (12, 13). Three drug classes—sodium-glucose cotransporter-2 (SGLT2) inhibitors, finerenone, and renin-angiotensin system inhibitors (RASi)—reduce kidney and cardiovascular events in eligible patients (14–22) but remain substantially underutilised (23, 24). The 2024 KDIGO guideline (25, 26) is explicit in the guideline-rich zone (RASi in albuminuric G3 CKD, SGLT2 inhibitors under defined eGFR/UACR thresholds) but becomes progressively contested in the guideline-silent gray zone at advanced kidney failure (eGFR < 15), where evidence is sparser, contraindication concerns rise, and clinician judgment carries the most weight. Frontier LLMs perform well on nephrology assessments in aggregate (27, 28) yet recent benchmarks show reasoning gaps (29, 30), and safety dimensions beyond accuracy—replicate-query reproducibility (31) and clinically incorrect reasoning rates (32–35)—are rarely measured.

We characterised three axes of behavioural divergence—boundary-zone rule adherence, replicate-query reproducibility, and clinically incorrect reasoning—across four frontier LLMs (GPT-5.4, Gemini 3.1 Pro, Grok 4.1 Fast, DeepSeek R1) using a stratified synthetic benchmark of 100 CKD vignettes spanning G3a to G5D. Our aim was not to rank models. Part of the benchmark deliberately sits where guidelines do not define a correct answer. In that zone we did not ask whether a model chose correctly. We asked whether contemporaneous models adopt different decision policies when guidance is silent, whether they state those policies, and whether aggregate

Abbreviations

ACEi, angiotensin-converting-enzyme inhibitor; ARB, angiotensin receptor blocker; CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; HFrEF, heart failure with reduced ejection fraction; HbA1c, glycated haemoglobin; KDIGO, Kidney Disease: Improving Global Outcomes; LLM, large language model; RASi, renin-angiotensin system inhibitor; SGLT2, sodium-glucose co-transporter-2; UACR, urine albumin-to-creatinine ratio.

accuracy conceals the difference. A decision-support tool whose behaviour in such zones is unknown cannot be governed, whichever policy is ultimately preferred.

2 Materials and methods

2.1 Study design and ethics

We benchmarked four large language models (LLMs) for detection of renoprotective therapy omissions in chronic kidney disease (CKD), following the TRIPOD+AI reporting guideline (36) and its LLM-specific extension, TRIPOD-LLM (37). Completed TRIPOD+AI and TRIPOD-LLM checklists are provided as separate files. The complete study workflow is shown in Figure 1. Only investigator-constructed synthetic vignettes were used; no patient data or human subjects were involved. Institutional review board approval was therefore not required. Eligibility rules and gold-standard reasoning were derived from the 2024 KDIGO Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease (25, 26) and contemporary CKD-related trial evidence [DAPA-CKD (14), EMPA-KIDNEY (15), FIDELIO-DKD (18), FIGARO-DKD (19)].

2.2 Case construction

One hundred synthetic case vignettes were constructed and stratified across CKD stages: 30 cases at G3a–G3b (estimated glomerular filtration rate [eGFR] 30–59 mL/min/1.73 m^2), 33 cases at G4 (eGFR 15–29), and 37 cases at G5/G5D (eGFR < 15 or on dialysis). Each vignette specified demographics, eGFR and CKD stage, UACR, serum potassium, blood pressure, HbA1c, comorbidities, current medications, allergies, and a brief clinical question requesting identification of renoprotective gaps. Three omission criteria were adapted from KDIGO 2024, plus an investigator-defined gray-zone RASi rule at eGFR < 15:

- **SGLT2 inhibitor**—type 2 diabetes or heart failure with reduced ejection fraction, eGFR ≥20 mL/min/1.73 m^2, UACR ≥ 200 mg/g, no contraindication recorded (n = 36).
- **Finerenone**—high-risk type 2 diabetic CKD, eGFR ≥25 mL/min/1.73 m^2, UACR ≥300 mg/g, serum potassium ≤4.8 mEq/L, on concurrent renin-angiotensin system blockade (n = 34). The primary endpoint was class-level omission detection; background RASi dose optimisation was not a prerequisite, and starting dose was not scored.
- **RASi (ACEi or ARB)**—albuminuric CKD (UACR ≥30 mg/g), serum potassium ≤5.5 mEq/L, no contraindication (n = 35). This included a prespecified eGFR < 15 boundary subgroup (n = 20) designed as *de novo* initiation scenarios (no ACEi/ARB in current medications), in which eligibility was operationally preserved when no explicit contraindication was recorded. Initiating RASi at eGFR < 15 is not a settled guideline-endorsed recommendation; these cases were included as a deliberate gray-zone benchmark to test conservative heuristics beyond stated criteria. The boundary subgroup comprised 11 non-dialysis G5 cases and 9 dialysis-dependent G5D cases (5 haemodialysis, 4 peritoneal dialysis); the operational rule did not distinguish dialysis status. The separate question of whether to *continue* existing RASi at eGFR < 15 (STOP-ACEi-relevant) was not tested here.

Thirty cases contained two prespecified gaps each (most commonly SGLT2i + finerenone or SGLT2i + RASi) and 45 contained one gap, yielding 105 prespecified gaps across 75 gap-bearing cases. A further 25 decoy cases without true gaps were constructed using two strategies in combination—current regimens already satisfying all renoprotective indications, and clinical features that operationally precluded one or more candidate classes (e.g., dialysis, documented intolerance)—to assess specificity. The N = 100 sample size was prespecified to populate three CKD severity strata (30/33/37) with adequate per-gap counts (36/34/35) for stratified pairwise model comparison while keeping per-case PI annotation time tractable. All vignettes were generated with complete laboratory and medication data by design; no missing values were permitted in the evaluator input. Anthropic Claude Sonnet 4.6 instantiated vignettes and gold-standard reasoning under investigator-defined eligibility rules specified *before* case generation, and was therefore excluded from the evaluator set under a strict generator/evaluator separation.

2.2.1 Relationship of operational criteria to KDIGO 2024

The operational eligibility rules used to construct gaps and score model outputs are deliberate simplifications of KDIGO 2024, and in two places extensions of it. Supplementary Table S14 sets out, for each drug class, the KDIGO 2024 recommendation, the operational rule used for the gold standard, and the exact wording given to the models.

For SGLT2 inhibitors, we applied a UACR ≥ 200 mg/g floor to both the type 2 diabetes and the non-diabetic CKD indications. KDIGO 2024 recommends SGLT2 inhibitors for any adult with type 2 diabetes and CKD at eGFR ≥ 20 without a UACR floor; it applies the UACR ≥ 200 threshold only to non-diabetic CKD. The heart-failure arm was restricted to heart failure with reduced ejection fraction (HFrEF). Benefit in heart failure with preserved ejection fraction is now supported by EMPEROR-Preserved and DELIVER, but no such cases were included.

For finerenone, the operational rule required UACR ≥ 300 mg/g and serum potassium ≤ 4.8 mEq/L, matching FIDELIO-DKD entry criteria. KDIGO 2024 supports finerenone in the broader UACR ≥ 30 population. The tighter rule was chosen to reflect trial entry criteria rather than the KDIGO recommendation ceiling.

For RASi, KDIGO 2024 recommends starting an ACEi or ARB in albuminuric CKD only for GFR categories G1–G4 (Recommendations 3.6.1–3.6.3) (25). A practice point also suggests considering dose reduction or discontinuation of RASi to reduce uraemic symptoms when eGFR is <15. Below

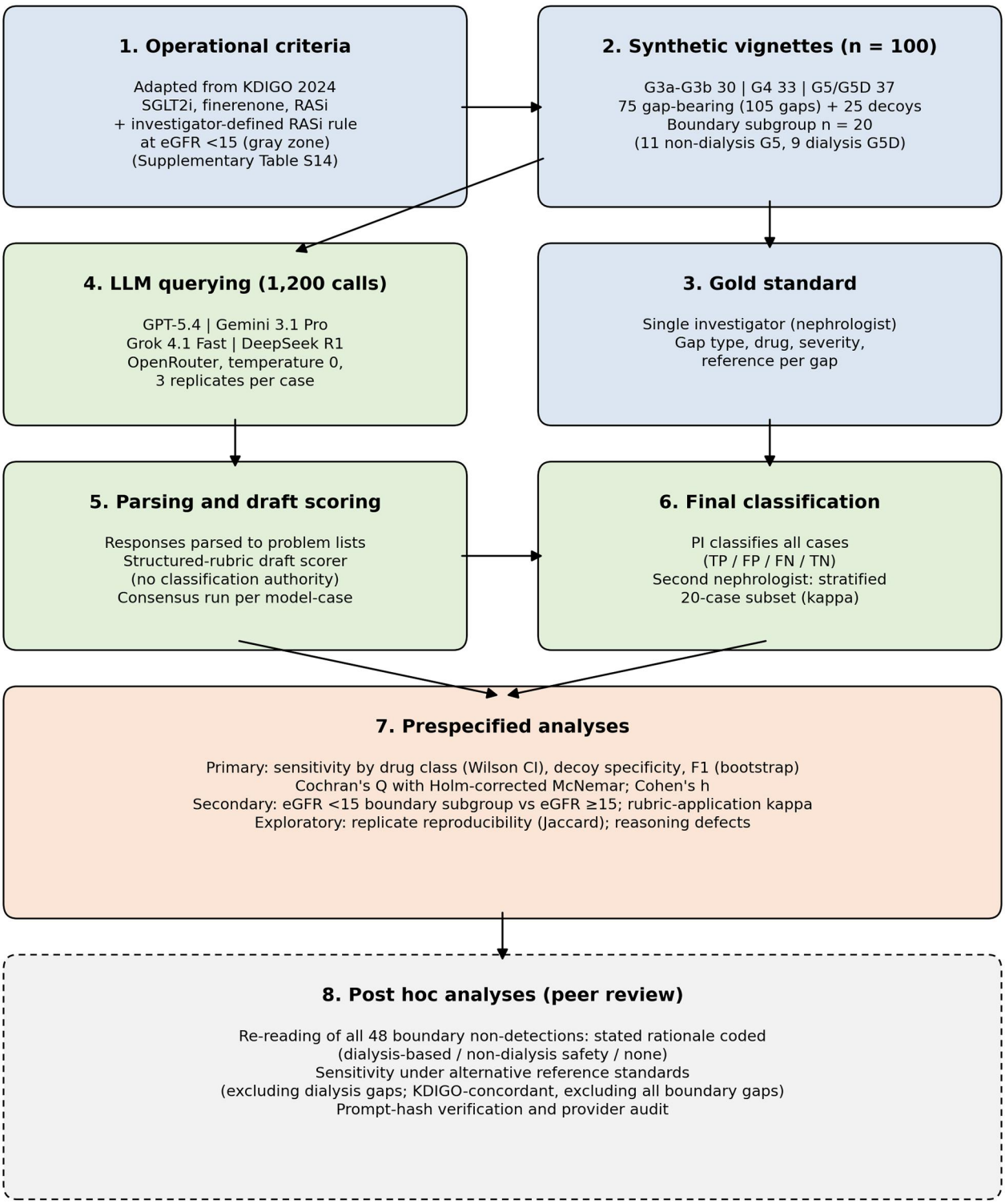


FIGURE 1
Study workflow. One hundred synthetic CKD vignettes were constructed against operational criteria adapted from KDIGO 2024, with a single-investigator gold standard. Each case was submitted three times to each of four LLMs through OpenRouter (temperature 0; 1,200 calls). Responses were parsed, draft-scored by a structured-rubric scorer, and finally classified by the principal investigator, with a second nephrologist applying the rubric to a stratified 20-case subset. Analyses covered drug-class sensitivity and specificity, the eGFR <15 boundary subgroup, replicate reproducibility, and reasoning defects. Post hoc analyses added during peer review are shown in the dashed box.

eGFR 15, initiation is therefore neither recommended nor discouraged by a graded recommendation. We selected this cut-off because it marks the point at which guideline coverage ends. The eGFR < 15 initiation rule is thus a deliberate probe of model behaviour, not a clinical recommendation. Results in Sections 3.4, 4.4 interpret non-adherence at this cut-off as behaviour, not as clinical error. Lower adherence is not evidence of lower clinical safety.

2.3 Models and evaluation protocol

Four LLMs were evaluated: OpenAI GPT-5.4 (openai/gpt-5.4-20260305), Google Gemini 3.1 Pro (google/gemini-3.1-pro-preview-20260219), xAI Grok 4.1 Fast (x-ai/grok-4.1-fast), and DeepSeek R1 (deepseek/deepseek-r1). All models were accessed on 2026-05-03 through the OpenRouter unified API. A shared request interface does not guarantee shared provider infrastructure. We therefore audited the serving provider recorded in every response. All 300 GPT-5.4 calls were served by OpenAI, all 300 Grok 4.1 Fast calls by xAI, and all 300 DeepSeek R1 calls by Novita. Gemini 3.1 Pro calls were served by Google (299/300) and Google AI Studio (1/300).

Each model received the same system prompt, which asked it to act as a nephrology clinical pharmacist. The model was to identify any renoprotective gap (drug class, drug, dose, severity, justification) or state explicitly that none existed. Each case was rendered into a fixed structured user message listing demographics, kidney function, laboratory values, dialysis status, comorbidities, medications, and allergies. The verbatim system prompt and a fully rendered user message are given in Supplementary Section S0. We recomputed the SHA-256 hash of each rendered prompt and compared it with the hash logged at run time; the results are reported in Limitations. The prompt instructed models to follow current KDIGO 2024 guidelines and provided simplified eligibility criteria for each drug class. SGLT2 inhibitor and finerenone thresholds (UACR $\geq$ 200 mg/g; UACR $\geq$ 300 mg/g, eGFR $\geq$ 25, and K $\leq$ 4.8 mEq/L) were stated explicitly. Ceiling-level detection of those two classes therefore mainly reflects prompt adherence rather than independent clinical reasoning. The RASi criterion was stated only as "indicated for CKD with albuminuria (UACR $\geq$30 mg/g)". The eGFR < 15 initiation rule and the potassium thresholds used in scoring were not disclosed. The RASi boundary result is therefore the class-comparative signal of interest (see Limitations).

Generation parameters were fixed at temperature 0 and max tokens 8192. Other sampling parameters (top_p, presence_penalty, frequency_penalty) were left at API defaults (1.0, 0, 0). No provider routing constraints were sent. Temperature 0 reduces but does not eliminate output variability in hosted LLMs, because of batching and floating-point non-determinism on the provider side. Each case was therefore queried three times per model (1,200 calls in total) so that residual variability could be measured. All 1,200 stored responses were non-empty, and all were analysed without exclusion or manual editing. Transient API errors and empty responses triggered automatic retry (up to three attempts); per-attempt logs for the main run were not retained.

2.4 Reproducibility assessment

Free-text responses from the three replicate runs were reduced to the set of issue types identified (SGLT2i, finerenone, RASi, or no-gap), since the clinical unit of interest is the gap type rather than the specific agent. Intra-model consistency was quantified by pairwise Jaccard similarity and full three-way agreement; a consensus run was selected by majority issue-type set (Supplementary Material).

2.5 Outcome classification

All cases were classified against the predefined gold standard by the principal investigator (H-M.L., board-certified nephrologist). A co-investigator nephrologist (H-J.L.) independently classified a stratified 20-case × 4-model sample covering all gap types, the eGFR < 15 boundary subgroup, multi-gap cases, and decoy cases. Both raters were blinded to each other but provided the prespecified gold-standard gap labels. Cohen's κ was computed for three mutually exclusive scoring layers. Gap detection was scored at the issue level (n = 88 ratings), whereas unsupported recommendations and decoy specificity were scored at the model-case level (n = 68 and n = 12, respectively).

An LLM-assisted structured-rubric scorer (GPT-5.4) generated draft TP/FP/FN/TN proposals; both raters made all final classifications, and the PI re-rated 70 ratings under blinded conditions with the draft hidden to confirm no anchoring (Supplementary Material). PI adjudication resolved disagreements.

We additionally recorded *clinically incorrect reasoning*—materially incorrect or unsupported clinical claims in a model's reasoning text, defined as fabricated guideline thresholds, drug-class properties contradicted by the drug label, or contraindication logic unsupported by trial evidence. Statements matching the study's prespecified operational criteria (including the UACR $\geq$300 mg/g finerenone rule) and defensible clinical simplifications were not counted; PI adjudication resolved rater disagreements. Identified instances were stratified post hoc into **factual pharmacology errors** (verifiable against external pharmacology or label text, independent of our operational criteria) and **boundary-discordant conservative reasoning** (a more restrictive potassium or eGFR threshold than our criterion, possibly reflecting defensible conservatism). The former is the headline safety finding; the latter is reported separately.

2.6 Statistical analysis

Per-omission-type sensitivity and overall specificity were estimated with Wilson 95% confidence intervals (38). Overall precision was computed across all FPs in the cohort. Overall F1 was estimated by case-level non-parametric bootstrap (n = 2,000 resamples; random seed = 42 for reproducibility). Inter-model differences in detection of each omission type were tested by Cochran's Q test; if omnibus $p < 0.05$, all six pairwise model comparisons were performed by McNemar's test and adjusted within each gap type by the Holm step-down procedure; each drug-class contrast was treated as an independent primary hypothesis (pairwise tests were conditional on the omnibus test, i.e., a protected procedure). Any within-boundary-subgroup pairwise tests reported below are treated as a separate exploratory family with within-family Holm correction only; no cross-family adjustment was applied and boundary-subgroup

p-values are reported as descriptive supporting evidence rather than as protected inference. Effect sizes were reported as Cohen's h. Because both raters were provided the prespecified operational gap labels, Cohen's κ between the two nephrologist raters quantifies rubric-application agreement—the reproducibility of applying a fixed operational rubric—rather than independent gold-standard construction. The independent construction of the gold standard was performed by a single investigator; no second-independent-rubric IRR was computed. We report κ with $\kappa > 0.80$ interpreted as almost perfect rubric-application agreement following convention.

Two prespecified secondary analyses were conducted: RASi sensitivity against the operational rule (reported as rule adherence) in the eGFR < 15 boundary subgroup vs. eGFR ≥ 15, and intra-rater κ between the PI's anchored and blinded passes (Supplementary Material). Clinically incorrect reasoning rate was analysed as an exploratory secondary safety endpoint with case-level proportions and Wilson 95% CIs; formal inferential testing was not prespecified for this exploratory endpoint and is reported descriptively. This analysis is hypothesis-generating and requires real-world replication.

Three analyses were added post hoc during peer review. First, every consensus-run non-detection in the eGFR < 15 boundary subgroup (48 model–case pairs) was re-read in full. The stated reason for withholding RASi was coded as dialysis-based, dialysis-based with potassium also cited, or a non-dialysis safety concern (hyperkalaemia, acute kidney injury, or haemodynamic risk). We also recorded whether the model proposed RASi for another indication or deferred it conditionally. Coding was performed with AI assistance. Per-case codes and representative quotations are given in Supplementary Table S15, and the full extracted rationale text is included in the reproducibility package. Second, sensitivity was recomputed under two alternative reference standards: excluding the 9 dialysis-dependent boundary gaps, and excluding all 20 boundary gaps (a KDIGO-concordant standard, because KDIGO 2024 does not recommend initiation below eGFR 15). Third, the run-time prompt hashes were recomputed from the current case files.

Case generation enforced a male:female ratio of 67:33 balanced across CKD strata (Table 1). Sex-stratified inferential testing of LLM behaviour was not performed: at the headline eGFR < 15 RASi boundary subgroup (n = 20), the cohort is underpowered to detect sex-based differences.

All analyses were performed in Python 3.13 with NumPy 2.4, SciPy 1.17, and Matplotlib 3.10. Reproducibility materials are described in the Data Availability statement.

TABLE 1 Case characteristics (n = 100 synthetic CKD vignettes).

Characteristic	Overall (n = 100)	G3a/G3b (n = 30)	G4 (n = 33)	G5/G5D (n = 37)
Demographics				
Age, years, median (IQR)	68 (58–74)	69 (59–77)	65 (58–74)	68 (60–74)
Male sex, n (%)	67 (67%)	20 (67%)	22 (67%)	25 (68%)
BMI, kg/m^2, median (IQR)	26 (23–30)	27 (23–30)	26 (22–28)	27 (25–30)
Renal function & labs				
eGFR, mL/min/1.73 m^2, median (IQR)	24 (11–36)	48 (41–53)	25 (21–27)	9 (6–12)
UACR, mg/g, median (IQR)	270 (118–358)	350 (285–420)	350 (280–400)	100 (70–170)
Serum potassium, mEq/L, median (IQR)	5 (4–5)	4 (4–5)	5 (4–5)	5 (5–5)
HbA1c, %, median (IQR)	8 (7–8)	8 (7–8)	8 (7–8)	8 (7–8)
Systolic BP, mmHg, median (IQR)	149 (141–156)	139 (132–146)	150 (145–156)	155 (150–166)
Comorbidities, n (%)				
Type 2 diabetes mellitus	78 (78%)	27 (90%)	28 (85%)	23 (62%)
Hypertension	100 (100%)	30 (100%)	33 (100%)	37 (100%)
Heart failure	5 (5%)	3 (10%)	2 (6%)	0 (0%)
Dyslipidemia	19 (19%)	10 (33%)	6 (18%)	3 (8%)
On dialysis, n (%)	20 (20%)	0 (0%)	0 (0%)	20 (54%)
Medications per case, median (IQR)	10 (8–11)	8 (8–8)	10 (10–10)	11 (11–11)

BMI, body mass index; eGFR, estimated glomerular filtration rate; UACR, urine albumin-to-creatinine ratio; HbA1c, glycated haemoglobin.
In 3 cases (BMI 20.5, 25.5, and 27.1 kg/m^2), the case text sent to the models listed obesity as a comorbidity, discordant with the recorded BMI. This label was removed from the released case files after the model runs; no other field was changed (see Limitations). All cases were generated according to KDIGO 2024-informed eligibility rules with prespecified boundary criteria. Cases were assigned to 36 SGLT2 inhibitor, 34 finerenone, and 35 RASi (ACEi/ARB) gap types, plus 25 decoys with no true gap; a case could contain more than one gap type. The full per-case gap-type breakdown is in Supplementary Table S12.

3 Results

3.1 Cohort

The 100 case vignettes (Table 1) had a median eGFR of 24 mL/min/1.73 m^2 (IQR 11–36), with stratified representation across G3a–G3b (n = 30), G4 (n = 33), and G5/G5D (n = 37); type 2 diabetes was present in 78 cases and hypertension in 100. The 75 gap-bearing cases contained 105 prespecified gaps (36 SGLT2 inhibitor, 34 finerenone, 35 RASi), with 30 multi-gap cases; 25 decoy cases were included to assess specificity. Per-gap recall in multi-gap cases was 100% for three models and 98.3% for DeepSeek R1 (Supplementary Table S1).

3.2 Overall performance differs by drug class and model

Across all 100 cases, the four LLMs differed substantially in overall classification performance (Figure 2; Table 2). Sensitivity

ranged from 97.1% (Grok 4.1 Fast, 95% CI 92%–99%) to 80.0% (DeepSeek R1, 71%–87%), with GPT-5.4 91.4% (85%–95%) and Gemini 3.1 Pro 81.9% (73%–88%); F1 followed the same ranking (Grok 0.981, GPT-5.4 0.932, Gemini 0.887, DeepSeek 0.853). Decoy specificity was 100.0% (Gemini), 96.0% (Grok), 92.0% (GPT-5.4), and 88.0% (DeepSeek) (Supplementary Figure S2).

3.3 Rule-based omissions show near-ceiling detection

Sensitivity for SGLT2 inhibitor and finerenone omissions was near-perfect across all four models: 100% (95% CI 90.4%–100.0%) for GPT-5.4, Gemini 3.1 Pro, and Grok 4.1 Fast on both classes, and 97.2% (95% CI 85.8%–99.5%) for SGLT2 inhibitor and 97.1% (95% CI 85.1%–99.5%) for finerenone for DeepSeek R1 (Supplementary Figure S1). Cochran's Q was uninformative given ceiling detection for these two omission types (SGLT2i Q = 3.00, df = 3, $p = 0.39$; finerenone Q = 3.00, $p = 0.39$).

In contrast, RASi rule adherence diverged markedly between models: sensitivity 91.4% (95% CI 77.6%–97.0%) for Grok 4.1 Fast, 74.3% (95% CI 57.9%–85.8%) for GPT-5.4, 45.7% (95% CI 30.5%–61.8%) for both Gemini 3.1 Pro and DeepSeek R1; Cochran's Q = 36.19, df = 3, $p < 0.001$. Holm-corrected pairwise McNemar tests confirmed significant differences between Grok 4.1 Fast and Gemini 3.1 Pro ($p = 0.0002$), Grok 4.1 Fast and DeepSeek R1 ($p = 0.0002$), GPT-5.4 and DeepSeek R1 ($p = 0.008$), and GPT-5.4 and Gemini 3.1 Pro ($p = 0.019$). The Gemini 3.1 Pro vs. DeepSeek R1 comparison (1 discordant pair in each direction; $p = 1.000$) and the Grok 4.1 Fast vs. GPT-5.4 comparison (Grok detected 6 cases that GPT-5.4 missed and zero in the reverse direction; $p = 0.063$) did not meet significance after correction.

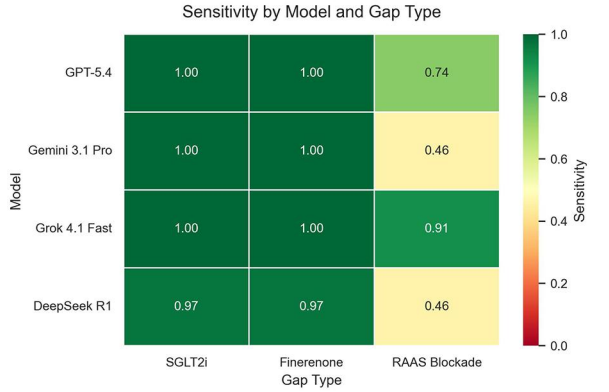


FIGURE 2
Overall performance: sensitivity by model and gap type. Heatmap of per-omission-type sensitivity across the four large language models evaluated. Each cell shows the proportion of prespecified gaps of that omission type correctly identified. SGLT2 inhibitor and finerenone omissions were detected at near-ceiling sensitivity by all four models; RASi detection diverged substantially between models, driven primarily by the eGFR < 15 boundary subgroup (see Figure 3). Colour gradient: red (low), yellow (intermediate), green (high).

3.4 eGFR < 15 boundary subgroup

The inter-model gap in RASi rule adherence was localised to the prespecified eGFR < 15 boundary subgroup (Figure 3). Among 20 boundary cases—operationally defined as RASi-eligible under the study's gray-zone criterion (UACR $\geq$ 30, K $\leq$ 5.5, no contraindication)—adherence to the operational rule was 85% (17/20) for Grok 4.1 Fast, 55% (11/20) for GPT-5.4, 10% (2/20) for Gemini 3.1 Pro, and 10% (2/20) for DeepSeek R1. Among 15 non-boundary RASi cases (eGFR $\geq$ 15), adherence was uniformly high: 100% for GPT-5.4 and Grok 4.1 Fast, 93% for both DeepSeek R1 and Gemini 3.1 Pro. The boundary–non-boundary difference was −15 (Grok), −45 (GPT-5.4), −83 (DeepSeek), and −83 (Gemini) percentage points, identifying eGFR < 15 as the principal locus of RASi non-initiation (Supplementary Table S13).

In a post-hoc secondary analysis conducted in response to statistical peer-review, pairwise Holm-corrected McNemar tests were applied within the eGFR < 15 boundary subgroup as a separate exploratory family (not cross-corrected against the primary drug-class family). Four of the six between-model contrasts were significant at $p_{Holm} < 0.05$: Gemini 3.1 Pro vs. Grok 4.1 Fast (b = 0, c = 15, Holm $p = 0.0004$), Grok 4.1 Fast vs. DeepSeek R1 (b = 15, c = 0, Holm $p = 0.0004$), GPT-5.4 vs. DeepSeek R1 (b = 9, c = 0, Holm $p = 0.0156$), and GPT-5.4 vs.

TABLE 2 Model performance: overall and by drug class.

Model	Overall sensitivity	SGLT2i	Finerenone	RASi	Specificity
GPT-5.4	91% (85–95)	100% (90–100)	100% (90–100)	74% (58–86)	92% (75–98)
Gemini 3.1 Pro	82% (73–88)	100% (90–100)	100% (90–100)	46% (30–62)	100% (87–100)
Grok 4.1 Fast	97% (92–99)	100% (90–100)	100% (90–100)	91% (78–97)	96% (80–99)
DeepSeek R1	80% (71–87)	97% (86–100)	97% (85–99)	46% (30–62)	88% (70–96)

Values are point estimates as percentages with 95% Wilson confidence intervals in parentheses. RASi denotes ACEi or ARB. RASi sensitivity in this table includes the 20 eGFR < 15 boundary gaps (the operational reference standard); adherence under KDIGO-concordant alternative reference standards is reported in Table 4. RASi sensitivity differed significantly across the four models (Cochran's Q $p < 0.001$; pairwise McNemar with Holm correction significant in 4 of 6 contrasts). Detection of SGLT2 inhibitor and finerenone omissions did not differ between models (Cochran's Q $p = 0.39$ for both). Detailed precision and F1 metrics are in Supplementary Table S7; Cochran's Q with pairwise Cohen's h is in S8; full pairwise McNemar contrasts are in S6.
Sensitivity for omission detection across the full case set ($n = 75$ cases with prespecified gaps; 105 total gaps) and stratified by drug class. Specificity is computed against the 25 decoy cases (no true gap).

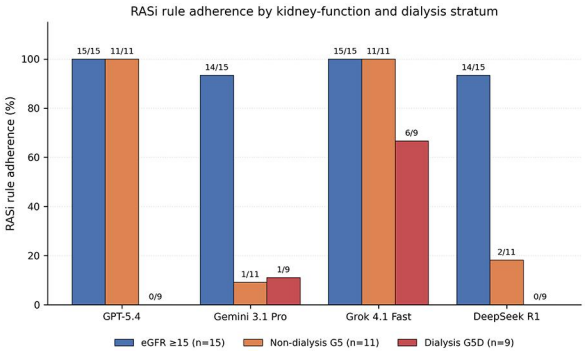


FIGURE 3
RASi rule adherence by kidney-function and dialysis stratum. Adherence to the operational RASi initiation rule in non-boundary cases (eGFR ≥ 15, n = 15) and in the prespecified eGFR < 15 boundary subgroup, split into non-dialysis G5 (n = 11) and dialysis-dependent G5D (n = 9) cases. Labels give detected/eligible gaps. Combined boundary adherence was 85% (Grok 4.1 Fast), 55% (GPT-5.4), 10% (Gemini 3.1 Pro), and 10% (DeepSeek R1). GPT-5.4 detected every non-dialysis G5 gap and no dialysis gap. KDIGO 2024 does not recommend RASi initiation below eGFR 15, so lower adherence is presented as model behaviour at a guideline-silent boundary, not as a safety failure. Stated rationales for non-detection are summarised in Table 3.

Gemini 3.1 Pro (b = 10, c = 1, Holm $p = 0.0352$). The GPT-5.4 vs Grok 4.1 Fast contrast fell short of the Holm threshold (b = 0, c = 6, Holm $p = 0.0625$), consistent with the overlapping Wilson CIs at n = 20. The Gemini 3.1 Pro vs. DeepSeek R1 contrast was not significant (b = 1, c = 1, Holm $p = 1.000$). Full pairwise counts are in Supplementary Table S16.

3.5 Stated rationale for boundary non-detection (post hoc)

All 48 boundary non-detections gave an explicit reason for not starting RASi; none omitted the drug class silently (Table 3; Supplementary Table S15). Detection split sharply by dialysis status. GPT-5.4 detected all 11 non-dialysis G5 gaps and none of the 9 dialysis gaps. Each of its 9 non-detections stated that, once the patient is dialysis-dependent, starting RASi to slow CKD progression is no longer a clear guideline-supported omission; in GC-03, GC-04, and GC-18 the response added a general comment that RASi could still be considered for blood pressure or cardiovascular indications. These were not accompanied by a named agent or dose, and are therefore not counted in the "RASi proposed for another indication" row of Table 3 (see Table 3 footnote). Grok 4.1 Fast detected 11/11 non-dialysis and 6/9 dialysis gaps; its 3 non-detections were also dialysis-based. Gemini 3.1 Pro (1/11 non-dialysis, 1/9 dialysis) and DeepSeek R1 (2/11, 0/9) withheld RASi in most non-dialysis G5 cases as well. In those cases they cited hyperkalaemia, acute kidney injury, or haemodynamic risk (Gemini 10, DeepSeek 9). Gemini proposed RASi for an alternative indication in 3 dialysis cases and DeepSeek in 1. DeepSeek deferred RASi pending potassium control in 3 cases.

Sensitivity under alternative reference standards is shown in Table 4. Excluding the 9 dialysis gaps raised GPT-5.4 and Grok 4.1 Fast to 100%. Under the KDIGO-concordant standard, which excludes all 20 boundary gaps, sensitivity was 96.5%–100% for every model.

TABLE 3 Stated rationale for RASi non-detection in the eGFR < 15 boundary subgroup.

Outcome / stated rationale	GPT-5.4	Gemini 3.1 Pro	Grok 4.1 Fast	DeepSeek R1
Detected, non-dialysis G5	11/11	1/11	11/11	2/11
Detected, dialysis G5D	0/9	1/9	6/9	0/9
Non-detections	9	18	3	18
Dialysis-based rationale	9	8	3	9
of which potassium also cited	2	2	2	3
Non-dialysis safety rationale	0	10	0	9
No stated rationale	0	0	0	0
RASi proposed for another indication	0	3	0	1
RASi deferred until potassium controlled	0	0	0	3

Dialysis-based rationale: the model stated that, once the patient is dialysis-dependent, starting RASi for kidney protection is not indicated or is no longer a clear guideline-supported omission. Non-dialysis safety rationale: the model cited hyperkalaemia, acute kidney injury, or haemodynamic risk. The two rationale categories are mutually exclusive; the last two rows are not. "RASi proposed for another indication" counts responses that recommended RASi with a named agent, dose, or specific alternative therapeutic indication (e.g., blood pressure target); general comments that RASi "could be considered" for blood pressure or cardiovascular risk without a named agent or dose (as in GPT-5.4 GC-03, GC-04, and GC-18) were not counted here. Post hoc analysis; per-case codes and representative quotations are in Supplementary Table S15.
Consensus run, n = 20 boundary cases per model. Detection is shown separately for non-dialysis G5 (n = 11) and dialysis-dependent G5D (n = 9; 5 haemodialysis, 4 peritoneal dialysis) cases.

3.6 Robustness and extension of the boundary finding

Two prespecified supplementary analyses tested robustness of the eGFR < 15 finding and a second clinical boundary (Supplementary Tables S3, S4). RASi non-detection was graded rather than threshold-defined: as the cutoff widened from eGFR < 10 to <25, RASi sensitivity rose monotonically for the three lower-recall models (e.g., GPT-5.4 0/9 → 18/27), while Grok remained ≥67%. Finerenone detection was unaffected by potassium stratum (100% in K 4.6–4.8). RASi detection in the K 5.1–5.5 yellow zone (n = 7) dropped further for Gemini and DeepSeek (14% each); stratum size precludes formal inference. RASi non-detection is therefore graded with respect to eGFR but selective with respect to drug class and potassium.

TABLE 4 Overall sensitivity under three reference standards (post hoc).

Model	Operational	Excluding dialysis	KDIGO-concordant
GPT-5.4	96/105 (91.4%)	96/96 (100%)	85/85 (100%)
Gemini 3.1 Pro	86/105 (81.9%)	85/96 (88.5%)	84/85 (98.8%)
Grok 4.1 Fast	102/105 (97.1%)	96/96 (100%)	85/85 (100%)
DeepSeek R1	84/105 (80.0%)	84/96 (87.5%)	82/85 (96.5%)

Operational: all 105 prespecified gaps. Excluding dialysis: the 9 dialysis-dependent eGFR < 15 RASi gaps removed. KDIGO-concordant: all 20 eGFR < 15 RASi gaps removed, because KDIGO 2024 recommends RASi initiation only for G1–G4.

3.7 Repeated queries reveal model-specific reproducibility

Across the three replicate queries per case, intra-model agreement on the identified issue-type set differed substantially between models (Supplementary Material). Full agreement across all three runs occurred in 89% of cases for Grok 4.1 Fast, 82% for GPT-5.4, 81% for Gemini 3.1 Pro, and 60% for DeepSeek R1. Mean pairwise Jaccard similarity was 0.93 (Grok), 0.88 (GPT-5.4), 0.87 (Gemini), and 0.74 (DeepSeek). Five DeepSeek R1 cases (5%) showed three-way disagreement; no other model produced any three-way discordant case. All primary analyses used a single consensus decision per model–case.

3.8 Rubric-application agreement (stratified 20-case subset)

On the 20-case inter-rater subset, gap detection agreement was perfect (κ = 1.000, 88/88). For unsupported recommendations, observed agreement was 97.1% (66/68); the near-absence of positive events precluded meaningful κ (prevalence-dependent paradox). All 12 decoy-model pairs were unanimously true negatives. The PI's blinded re-rating yielded intra-rater κ = 0.971 vs. the draft-anchored pass, confirming no draft anchoring (Supplementary Material).

3.9 Clinically incorrect reasoning (exploratory secondary endpoint)

Across 100 cases, 52 instances of reasoning defects were identified. We separate these into two distinct categories with different clinical interpretations. First, 33/52 (63%) were **clinically incorrect reasoning**—external factual pharmacology errors independent of the study's operational criteria (e.g., fabricated sulfonamide cross-reactivity, unsupported ARB-over-ACEi hyperkalemia claims, inverted sodium bicarbonate pharmacology). Second, 19/52 (37%) were **operational-criterion discordance**—the model applied stricter potassium or eGFR thresholds than the study's operational criterion, which can represent defensible clinical caution rather than factual error.

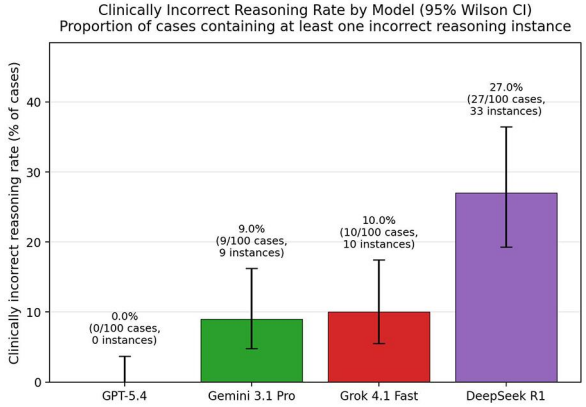


FIGURE 4
Reasoning-defect rate by model: exploratory secondary endpoint. Case-level rate of reasoning defects, defined as the proportion of cases containing at least one clinically incorrect claim OR at least one instance of operational-criterion discordance in the model reasoning text, with Wilson 95% confidence intervals. GPT-5.4 produced no such defects across 100 cases; DeepSeek R1 produced them in 27% of cases. Two distinct categories are combined here: clinically incorrect reasoning (factual pharmacology errors independent of the study rubric—33 instances across 29 unique cases) and operational-criterion discordance (application of stricter thresholds than the study rubric, representing defensible clinical caution—19 instances across 18 unique cases). Per-category rates are reported separately in Supplementary Table S2. This analysis is positioned as an exploratory secondary endpoint; pairwise rate differences are descriptive, and formal inferential testing was not prespecified.

A stated rationale, however, is not evidence that the rationale is correct: on internal audit of this category, two DeepSeek R1 responses in dialysis-dependent cases (GC-04 and GC-18) that initially appeared to express boundary caution were reclassified as clinically incorrect reasoning because they attributed to KDIGO 2024 a recommendation to discontinue or avoid ACEi/ARB on initiation of dialysis that is not actually present in the guideline (final per-case classification in Supplementary Table S2). The case-level rate for the combined pool (either category) differed markedly across models (Figure 4): 0/100 (0.0%, 95% CI 0.0%–3.7%) for GPT-5.4, 10/100 (10.0%) for Grok 4.1 Fast, 9/100 (9.0%) for Gemini 3.1 Pro, and 27/100 (27.0%, 19.3%–36.4%) for DeepSeek R1. The clinically-incorrect-reasoning-only case rate (unique cases containing at least one factual pharmacology error, excluding operational-criterion discordance) is 0/100 (0.0%) for GPT-5.4, 9/100 (9.0%) for Grok 4.1 Fast, 2/100 (2.0%) for Gemini 3.1 Pro, and 18/100 (18.0%) for DeepSeek R1; corresponding instance counts (which exceed case counts when a case contains more than one factual error) are 0, 9, 2, and 22, giving the 33-instance total above (Supplementary Table S2). These rates are specific to this benchmark and prompt and should not be read as general properties of the models. Non-overlapping Wilson CIs between GPT-5.4 and other models on the combined pool suggest meaningful between-model differences, but formal pairwise inferential testing was not prespecified for this exploratory endpoint and is reported descriptively; findings require real-world replication.

4 Discussion

All four LLMs performed near-ceiling on rule-based SGLT2 inhibitor and finerenone omissions. Three diverged substantially from an investigator-defined eGFR < 15 RASi initiation rule, in a zone not covered by KDIGO 2024 recommendations. Every boundary non-detection carried an explicit rationale, and the rationales revealed distinct decision policies rather than oversight. Under a KDIGO-concordant reference standard, all models exceeded 96% sensitivity. Reproducibility (60%–89% full agreement across identical replicate queries) and reasoning-defect rates (combined-pool case rate 0%–27%; clinically-incorrect-reasoning-only case rate 0%–18%; see Section 3.9 for the pharmacology-error vs conservative-boundary split) varied markedly between models. Each of these three axes reflects a distinct behavioural divergence that would be invisible to an evaluation reporting aggregate sensitivity alone.

4.1 Rule-based vs. boundary detection

For SGLT2 inhibitor and finerenone omissions, all models achieved 97%–100% sensitivity with no inter-model difference (Cochran's Q $p = 0.39$), reflecting unambiguous eligibility logic. For RASi, overall sensitivity ranged from 46% to 91% (Q < 0.001), localised to the eGFR < 15 boundary subgroup: rule adherence was 85% for Grok 4.1 Fast, 55% for GPT-5.4, and 10% for both Gemini and DeepSeek. The supplementary threshold analysis confirmed that this is graded, not threshold-defined: GPT-5.4 and DeepSeek reached 0% at eGFR < 10.

The STOP-ACEi trial (39) demonstrated safety of *continuing* existing RASi at eGFR < 30, but all 20 boundary cases tested *de novo initiation* (no ACEi/ARB prescribed). Models did not violate trial evidence; they diverged from a prespecified operational gray-zone criterion extending initiation into a zone without dedicated randomised evidence. Characterising this selective operational-rule discordance is important because decision-support systems must behave predictably when local rules extend beyond explicit trial evidence.

The stated rationales clarify what the divergence represents. GPT-5.4 applied a single, consistent policy: it recommended RASi in every non-dialysis G5 case and withheld it in every dialysis case, on the grounds that slowing CKD progression is no longer the therapeutic target once dialysis has begun. This position is clinically defensible for the progression endpoint used in our gold standard. RENAAL and IDNT, which that gold standard cites, excluded dialysis patients (20, 21). Separate small trials in peritoneal dialysis have reported slower loss of residual renal function with an ACEi or an ARB (40, 41). Those trials do not create a KDIGO initiation recommendation, but they mean one dialysis rule is coarser than the case record. Supplementary Table S15 marks haemodialysis and peritoneal dialysis separately; it does not summarise detection by modality. The dialysis-dependent gaps therefore reflect a limitation of our reference standard as much as of the models. Gemini 3.1 Pro and DeepSeek R1 applied a broader policy that also withheld RASi from non-dialysis G5 patients because of hyperkalaemia or acute kidney injury risk. KDIGO 2024 practice points on hyperkalaemia management argue against potassium alone as a reason to withhold therapy when RASi is otherwise indicated, although the direct guidance addresses continuation rather than *de novo* initiation (25). Neither policy is unambiguously wrong at eGFR < 15. What matters for deployment is that the policies differ, are stated, and are invisible in aggregate sensitivity.

Potassium-stratum analysis showed no conservative behaviour for finerenone at borderline K (4.6–4.8: 100% detection) yet further RASi non-detection at K 5.1–5.5—a real-world contraindication zone in which many nephrologists would defer RASi initiation and treat the potassium first, so this may reflect appropriate clinical caution rather than model deficit—consistent with drug-eligibility-specific rather than universal boundary conservatism (30).

4.2 Reproducibility and clinically incorrect reasoning

Full agreement across three replicate queries ranged from 89% (Grok) to 60% (DeepSeek), with five DeepSeek cases showing three-way disagreement. Temperature 0 is often assumed to produce deterministic output. In hosted LLMs it does not, because provider-side batching and floating-point arithmetic introduce residual variation. Each model was served by a single provider throughout (apart from one Gemini call), so the variation we observed cannot be explained by switching between providers. Reasoning models such as DeepSeek R1 generate long intermediate reasoning, which may give small early differences more opportunity to change the final answer; we did not test this mechanism. Because each model was served by a single provider throughout, model and provider effects on the observed Jaccard spread are confounded in our data. This differential has direct deployment implications: a model whose classification varies across identical queries presents a barrier to quality-improvement workflows requiring reproducible recommendations.

GPT-5.4 produced no reasoning defects (0.0%, CI 0.0%–3.7%), whereas DeepSeek R1 produced defects in 27% of cases. These rates describe behaviour on this benchmark and prompt; they are not general error rates for the models. Such instances were more common in gap-bearing than decoy cases, suggesting that clinically incorrect reasoning is most likely at the moment a recommendation is being constructed rather than during null-case review.

4.3 No single model dominates

The four models occupied distinct positions: Grok 4.1 Fast led on sensitivity (97%) and reproducibility (89%) but produced reasoning defects in 10% of cases, while GPT-5.4 produced none (0% vs 9%–27% in the other three models) and adhered less often to the boundary rule (55%), withholding RASi only in dialysis-dependent cases. Relative performance in real clinical documentation, and relative to nephrologist reasoning, was not measured here and remains to be established. Within this

benchmark, models disagreed on eGFR < 15 RASi decisions, so any of them would require clinician oversight in this zone.

4.4 Boundary-zone behaviour of clinical LLMs

Our findings sit alongside established literature on hidden model behaviour in clinical AI, but with an important scoping caveat. The pulse-oximetry racial bias described by Sjoding and colleagues (1), the Epic sepsis-prediction shortfall reported by Wong and colleagues (2), the chest-radiograph shortcut learning shown by Zech and colleagues (3), and the care-management algorithm bias documented by Obermeyer and colleagues (4) each represented a device or model failure against ground-truth patient outcomes, invisible in aggregate performance yet concentrated in a demographic or geographic subgroup. Our benchmark documents a related but distinct phenomenon: model-specific divergence in adherence to an investigator-defined operational rule at a clinical decision boundary—the eGFR < 15 RASi initiation cut-off—where KDIGO 2024 does not endorse initiation and where clinician judgement therefore drives real-world practice. The 10%–85% spread across models is a genuinely subgroup-concentrated behavioural divergence, but its safety direction cannot be resolved from a synthetic benchmark alone: models with lower boundary adherence may be practising defensible clinical caution against hyperkalaemia, AKI-transition risk, and futile initiation late in the disease course. The transferable methodological point is that boundary-zone probing surfaces subgroup-specific model divergence that aggregate accuracy metrics cannot, whether or not each specific divergence maps directly onto a bedside-safety hazard. Even under a safety-agnostic reading, the fact that four contemporaneous frontier models spread across a 75-percentage-point range at the same clinical cut-off is itself the evidence that aggregate accuracy hides model-specific decision policies, which is what deployment-side governance most needs to know.

4.5 Implications for pre-deployment LLM safety assurance

Two implications follow. First, aggregate accuracy is not a sufficient pre-deployment behaviour criterion for LLM-assisted medication review. Silent trials before clinical exposure (10) catch some failure modes—drift, dataset shift, integration friction—but do not decompose the behaviour surface into interpretable clinical decision regions. Deliberate boundary-zone probing, which queries the model in guideline-uncertain regions, surfaces the most model-specific divergences at low cost and complements silent-trial designs at both pre-deployment and post-deployment monitoring stages. Second, deployment-side quality-assurance workflows [Root Cause Analysis triggered by adverse events, Failure Mode and Effects Analysis for new tools, alert-fatigue monitoring (11)] are downstream defensive layers in Reason's Swiss-Cheese model (5) that presume the tool's upstream behaviour profile has been characterised. When the tool is a general-purpose LLM whose training and prompt-response behaviour cannot be fully audited, upstream boundary-zone probing becomes even more critical because downstream defensive layers have less mechanistic understanding to draw from. Guideline-development bodies, medical-device regulators, and hospital AI-governance committees may therefore need reporting standards that require boundary-stratified performance, replicate reproducibility, and reasoning-error rates alongside aggregate metrics for any LLM tool proposed for medication-review use (42). Such reporting would also need to address explainability and data privacy; recent clinical prediction frameworks have built both into model design, combining explainable-AI outputs with privacy-preserving federated training (43, 44). Stated rationales, such as those analysed here, are one form of explanation that LLM tools provide natively. Prospective validation of these three behaviour axes against real-world clinical documentation, and characterisation of the direction of safety implication at each boundary, are required and are planned as separate studies.

4.6 Limitations

First, the vignettes are synthetic, structured, and in English. Every case was complete and internally consistent. Real records contain missing or conflicting laboratory values, free-text notes, non-English documentation, and patient-reported barriers. The vignettes also gave a single time point, without laboratory trends, prior drug intolerance, or treatment history, all of which shape real prescribing decisions. Synthetic cases allow controlled placement of boundary conditions, but they cannot show how models behave with real documentation. Our findings therefore support the evaluation method, not the readiness of any model for deployment. A real-world replication is planned pending institutional review board approval.

Second, rubric-application agreement was assessed on a stratified 20-case subset; cases outside the subset were classified by the principal investigator alone. Because both raters applied the same prespecified labels, the perfect kappa shows that the rubric can be applied reproducibly; it does not validate the rubric itself.

Third, four LLMs were tested at a single time point with one prompt at temperature 0. Hosted models change over time, and a different prompt, prompting strategy, or temperature could change both accuracy and reproducibility. Replicate querying measured residual variability at temperature 0 but not sensitivity to sampling temperature.

Fourth, the eGFR < 15 RASi criterion was investigator-defined and tested a gray-zone de novo initiation scenario; boundary non-detections reflect operational-rule discordance rather than proven clinical-reasoning failure. The boundary subgroup also combined non-dialysis G5 and dialysis-dependent G5D cases under one rule. Our gold standard cited RENAAL and IDNT for the dialysis cases, although those trials excluded dialysis patients. The post hoc analyses in Tables 3, 4 show how the conclusions change when these cases are removed.

Fifth, three case files (GA-16, GB-16, and GC-30) were edited after the model runs to remove an obesity label that was

discordant with the recorded BMI. No gap-relevant field was changed. Recomputed prompt hashes matched the run-time hashes for all other 97 cases; for these 3 cases, the pre-edit versions in the version-control history reproduce the run-time hashes exactly. The pre-edit versions are available on request.

4.6.1 Cohort composition

The vignette cohort is deliberately CKD-severity-stratified rather than epidemiologically representative. Hypertension prevalence is 100%, heart failure is 5%, and dyslipidaemia is 19%; a real Taiwan CKD registry cohort would have lower hypertension prevalence (∼85%–90%), higher heart-failure prevalence (∼25%–35% in stages G3–G5), and higher dyslipidaemia prevalence (∼60%–70%). Coronary artery disease, prior stroke, and atrial fibrillation are not represented. The 67:33 male:female ratio (67% male) is higher than the ∼55:45 male-female distribution in the Taiwan CKD registry. Non-diabetic CKD accounts for only 22 of the 100 vignettes; because both RASi and SGLT2 inhibitor indications extend to non-diabetic albuminuric CKD, external generalisability to that population is not supported by the current cohort. External generalisability claims should therefore be tempered by the acknowledgement that the vignette cohort under-represents heart-failure, cardiovascular comorbidity, and non-diabetic CKD relative to a real CKD clinic.

4.6.2 Model version pinning and API-layer audit

GPT-5.4 (`openai/gpt-5.4-20260305`) and Gemini 3.1 Pro (`google/gemini-3.1-pro-preview-20260219`) were queried at pinned date-stamped snapshot IDs. Grok 4.1 Fast (`x-ai/grok-4.1-fast`) and DeepSeek R1 (`deepseek/deepseek-r1`) are floating model tags whose underlying weights may change without notice, and were queried via OpenRouter. A post-hoc audit of the `provider_response` field across all 1,200 runs confirmed that provider routing was consistent within each model over the request window: GPT-5.4 resolved to OpenAI for 300/300 runs; Gemini 3.1 Pro resolved to Google for 299/300 runs and to Google AI Studio for 1 run; Grok 4.1 Fast resolved to xAI directly for 300/300 runs; DeepSeek R1 resolved to the Novita provider (a specific OpenRouter serving stack) for 300/300 runs. Reproducibility of the specific numeric findings therefore depends on continued availability of Grok on xAI's own inference and of DeepSeek on Novita's serving stack. The observed 3-replicate reproducibility ranges (Jaccard 0.74–0.93) capture the small residual API-layer non-determinism within these routes. A second post-hoc audit confirmed that no run reached the `max_tokens = 8192` output-token budget (DeepSeek maximum 7,498 tokens; Grok maximum 7,814 tokens; all other models well under) and no scored response reached the scoring pipeline's 10,000-character response cap (maximum observed response length 5,485 characters, GPT-5.4), so neither budget was a source of silent truncation in this benchmark.

4.6.3 Single-author rubric and construction dependency

The principal investigator designed the operational eligibility criteria (including the gray-zone eGFR < 15 RASi rule), generated the vignettes and their prescribing gaps against those criteria, wrote the gold standard, and served as the primary rater. The co-rater applied the prespecified gold-standard labels rather than constructing an independent rubric. Reported inter-rater agreement therefore represents rubric-application reproducibility, not rubric-validity agreement. A third-nephrologist audit of criteria validity on a small sample would be a natural next-step sensitivity check but is beyond the scope of this benchmark.

4.6.4 Automated scoring by an evaluator model

OpenAI GPT-5.4 served both as one of the four benchmarked evaluators and as the automated draft-scorer that produced first-pass true-positive/false-positive/false-negative/true-negative and reasoning-defect classifications against the gold standard. Although the principal investigator adjudicated every case and a blinded intra-rater re-rating with the LLM draft hidden yielded $\kappa = 0.971$ against the anchored pass, an epistemic asymmetry remains: reviewers who see "no issues flagged" in GPT-5.4's draft vs. multiple flagged issues in another model's draft are not fully symmetric even under blinded conditions. GPT-5.4's headline 0/100 clinically-incorrect-reasoning rate should therefore be interpreted with caution pending an independent-scorer sensitivity analysis (planned as a separate work using Claude Sonnet or an unaided rater re-scoring of the 100 GPT-5.4 responses).

This manuscript reports a benchmark in which large language models (LLMs) were the experimental object; the four evaluated LLMs (GPT-5.4, Gemini 3.1 Pro, Grok 4.1 Fast, DeepSeek R1) did not contribute to writing or analysis. Anthropic Claude Sonnet 4.6 was used to instantiate the synthetic vignettes and gold-standard reasoning under investigator-fixed eligibility rules specified before case generation, and was therefore excluded from the evaluated set under a strict generator/evaluator separation. Anthropic Claude Haiku parsed model responses into structured problem lists without classification authority, and OpenAI GPT-5.4 produced draft TP/FP/FN/TN proposals against the gold standard without final classification authority; all final classifications were made by the principal investigator (board-certified nephrologist). A second nephrologist independently applied the operational criteria to a stratified 20-case rubric-application-agreement subset; rubric-application agreement was Cohen's kappa = 1.000.

During peer review, an AI coding assistant (Anthropic Claude) was used to write the scripts for the post hoc analyses and to draft the rationale codes for the 48 boundary non-detections; the codes and analyses were reviewed by the authors, and the scripts are included in the reproducibility package. In addition to the research-workflow uses above, the authors used AI-assisted tools during manuscript drafting and revision for translation and proofreading on iterative human-directed prompts; the authors originated all scientific content, methods,

analyses, and interpretation, and reviewed and approved every version of the text. The authors take full responsibility for the manuscript.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This study used only investigator constructed synthetic case vignettes; no patient-level data, no chart review, and no human subjects were involved. Institutional review board approval was therefore not required.

Author contributions

S-EY: Writing – review & editing, Writing – original draft, Validation, Visualization. H-JL: Investigation, Writing – review & editing, Validation. W-WL: Writing – review & editing, Writing – original draft. H-ML: Visualization, Writing – review & editing, Supervision, Data curation, Investigation, Software, Conceptualization, Methodology, Resources, Project administration, Formal analysis, Validation.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

We thank two anonymous nephrology colleagues for informal feedback during methodology development.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. Generative AI was used to assist manuscript preparation as follows. First, in the research workflow itself, generative AI played roles that are the subject of study rather than authoring aids: Anthropic Claude Sonnet 4.6 instantiated the synthetic clinical vignettes and gold-standard reasoning under investigator-fixed eligibility rules specified before case generation and was excluded from the evaluated panel under a strict generator/evaluator separation; Anthropic Claude Haiku parsed model responses into structured problem lists without classification authority; and OpenAI GPT-5.4 produced draft true-positive/false-positive/false-negative/true-negative classifications against the gold standard without final classification authority. All final case classifications were performed by the principal investigator (a board-certified nephrologist), and a second nephrologist independently applied the operational criteria to a stratified 20-case subset (Cohen's κ = 1.000). The four evaluated large language models (GPT-5.4, Gemini 3.1 Pro, Grok 4.1 Fast, DeepSeek R1) were the objects of the benchmark and did not contribute to manuscript writing or analysis. Second, during manuscript drafting and revision, AI-assisted tools (Anthropic Claude and OpenAI Codex) were used for translation and proofreading on iterative human-directed prompts. Further details are given at the end of Section 4.6.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2026.2003897/full#supplementary-material

References

1. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. *N Engl J Med.* (2020) 383:2477–8. doi: 10.1056/NEJMc2029240

2. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* (2021) 181:1065–70. doi: 10.1001/jamainternmed.2021.2626

3. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* (2018) 15:e1002683. doi: 10.1371/journal.pmed.1002683

4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* (2019) 366:447–53. doi: 10.1126/science.aax2342

5. Reason J. Human error: models and management. *BMJ*. (2000) 320:768–70. doi: 10.1136/bmj.320.7237.768

6. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. (2024) 30(9):2613–22. doi: 10.1038/s41591-024-03097-1

7. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large languagemodels to provide clinical recommendations in the emergency department. *Nat Commun*. (2024) 15:8236. doi: 10.1038/s41467-024-52415-1

8. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *npj Digit Med*. (2023) 6:195. doi: 10.1038/s41746-023-00939-z

9. Suh PS, Shim WH, Suh CH, Heo H, Park CR, Eom HJ, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini pro vision using image inputs from diagnosis please cases. *Radiology*. (2024) 312(1):e240273. doi: 10.1148/radiol.240273

10. Kwong JCC, Erdman L, Khondker A, Skreta M, Goldenberg A, McCradden MD, et al. The silent trial—the bridge between bench-to-bedside clinical AI applications. *Front Digit Health*. (2022) 4:929508. doi: 10.3389/fdgth.2022.929508

11. Murad DA, Tsugawa Y, Elashoff DA, Baldwin KM, Bell DS. Distinct components of alert fatigue in physicians' responses to a noninterruptive clinical decision support alert. *J Am Med Inform Assoc*. (2022) 30:64–72. doi: 10.1093/jamia/ocac191

12. Ong JCL, Jin L, Elangovan K, Lim GYS, Lim DYZ, Sng GGR, et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. *Cell Rep Med*. (2025) 6:102323. doi: 10.1016/j.xcrm.2025.102323

13. Hu Y, Liu J, Jiang W. Large language models in nephrology: applications and challenges in chronic kidney disease management. *Ren Fail*. (2025) 47:2555686. doi: 10.1080/0886022x.2025.2555686

14. Heerspink HJL, Stefánsson BV, Correa-Rotter R, Chertow GM, Greene T, Hou F-F, et al. Dapagliflozin in patients with chronic kidney disease. *N Engl J Med*. (2020) 383:1436–46. doi: 10.1056/nejmoa2024816

15. The EMPA-KIDNEY Collaborative Group. Long-term effects of empagliflozin in patients with chronic kidney disease. *N Engl J Med*. (2023) 388:117–27. doi: 10.1056/nejmoa2204233

16. The EMPA-KIDNEY Collaborative GroupLong-term effects of empagliflozin in patients with chronic kidney disease. *N Engl J Med*. (2025) 392:777–87. doi: 10.1056/nejmoa2409183

17. Perkovic V, Jardine MJ, Neal B, Bompoint S, Heerspink HJL, Charytan DM, et al. Canagliflozin and renal outcomes in type 2 diabetes and nephropathy. *N Engl J Med*. (2019) 380:2295–306. doi: 10.1056/nejmoa1811744

18. Bakris GL, Agarwal R, Anker SD, Pitt B, Ruilope LM, Rossing P, et al. Effect of finerenone on chronic kidney disease outcomes in type 2 diabetes. *N Engl J Med*. (2020) 383:2219–29. doi: 10.1056/nejmoa2025845

19. Pitt B, Filippatos G, Agarwal R, Anker SD, Bakris GL, Rossing P, et al. Cardiovascular events with finerenone in kidney disease and type 2 diabetes. *N Engl J Med*. (2021) 385:2252–63. doi: 10.1056/nejmoa2110956

20. Brenner BM, Cooper ME, de Zeeuw D, Keane WF, Mitch WE, Parving H-H, et al. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *N Engl J Med*. (2001) 345:861–9. doi: 10.1056/nejmoa011161

21. Lewis EJ, Hunsicker LG, Clarke WR, Berl T, Pohl MA, Lewis JB, et al. Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *N Engl J Med*. (2001) 345:851–60. doi: 10.1056/nejmoa011303

22. Agarwal R, Green JB, Heerspink HJL, Mann JFE, McGill JB, Mottl AK, et al. Finerenone with empagliflozin in chronic kidney disease and type 2 diabetes. *N Engl J Med*. (2025) 393:533–43. doi: 10.1056/nejmoa2410659

23. Wanner C, Zhao M-H, Amin AN, De Nicola L, Sauer AJ, Allum AM, et al. Guideline-recommended disease-modifying therapies for patients with cardiorenal disease: a call-to-action narrative review. *Adv Ther*. (2025) 42:3141–57. doi: 10.1007/s12325-025-03228-1

24. Yau K, Dharia A, Alrowiyti I, Cherney DZI. Prescribing SGLT2 inhibitors in patients with CKD: expanding indications and practical considerations. *Kidney Int Rep*. (2022) 7:1463–76. doi: 10.1016/j.ekir.2022.04.094

25. Stevens PE, Ahmed SB, Carrero JJ, Foster B, Francis A, Hall RK, et al. KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int*. (2024) 105:S117–314. doi: 10.1016/j.kint.2023.10.018

26. Levin A, Ahmed SB, Carrero JJ, Foster B, Francis A, Hall RK, et al. Executive summary of the KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease: known knowns and known unknowns. *Kidney Int*. (2024) 105:684–701. doi: 10.1016/j.kint.2023.10.016

27. Wu S, Koo M, Blum L, Black A, Kao L, Fei Z, et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. *NEJM AI*. (2024) 1:AIdbp2300092. doi: 10.1056/aidbp2300092

28. McCoy LG, Swamy R, Sagar N, Wang M, Bacchi S, Fong JMN, et al. Assessment of large language models in clinical reasoning: a novel benchmarking study. *NEJM AI*. (2025) 2:AIdbp2500120. doi: 10.1056/aidbp2500120

29. Unger Z, Soffer S, Efros O, Chan L, Klang E, Nadkarni GN. Clinical applications and limitations of large language models in nephrology: a systematic review. *Clin Kidney J*. (2025) 18:sfaf243. doi: 10.1093/ckj/sfaf243

30. Rao AS, Esmail KP, Lee RS, Jiang S, Arraiza Carlo B, Gill J, et al. Large language model performance and clinical reasoning tasks. *JAMA Netw Open*. (2026) 9:e264003. doi: 10.1001/jamanetworkopen.2026.4003

31. Windisch P, Koechli C, Dennstädt F, Aebersold DM, Zwahlen DR, Förster R, et al. Is one run enough? Reproducibility of flagship large language models across temperature and reasoning settings in biomedical text processing. *J Am Med Inform Assoc*. (2026) 33:1179–84. doi: 10.1093/jamia/ocag039

32. Kim Y, Jeong H, Chen S, Li SS, Park C, Lu M, et al. Medical hallucinations in foundation models and their impact on healthcare. *arXiv* [Preprint]. *arXiv:2503.05777[cs.CL]* (2025). Available online at: https://arxiv.org/abs/2503.05777

33. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med*. (2025) 8:274. doi: 10.1038/s41746-025-01670-7

34. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun Med*. (2025) 5:330. doi: 10.1038/s43856-025-01021-3

35. Chen S, Gao M, Sasse K, Hartvigsen T, Anthony B, Fan L, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digit Med*. (2025) 8:605. doi: 10.1038/s41746-025-02008-z

36. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. (2024) 385:e078378. doi: 10.1136/bmj-2023-078378

37. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. (2025) 31:60–9. doi: 10.1038/s41591-024-03425-5

38. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc*. (1927) 22:209–12. doi: 10.1080/01621459.1927.10502953

39. Bhandari S, Mehta S, Khwaja A, Cleland JGF, Ives N, Brettell E, et al. Renin–Angiotensin system inhibition in advanced chronic kidney disease. *N Engl J Med*. (2022) 387:2021–32. doi: 10.1056/nejmoa2210639

40. Li PKT, Chow KM, Wong TYH, Leung CB, Szeto CC. Effects of an angiotensin-converting enzyme inhibitor on residual renal function in patients receiving peritoneal dialysis: a randomized, controlled study. *Ann Intern Med*. (2003) 139(2):105–12. doi: 10.7326/0003-4819-139-2-200307150-00010

41. Suzuki H, Kanno Y, Sugahara S, Okada H, Nakamoto H. Effects of an angiotensin II receptor blocker, valsartan, on residual renal function in patients on CAPD. *Am J Kidney Dis*. (2004) 43(6):1056–64. doi: 10.1053/j.ajkd.2004.01.019

42. Tao X, Zhou S, Ding K, Li S, Li Y, Wu B, et al. An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. *Nat Med*. (2026) 32:934–42. doi: 10.1038/s41591-025-04176-7

43. Kumar D, Bakariya B, Verma C, Illés Z. LivXAI-Net: an explainable AI framework for liver disease diagnosis with IoT-based real-time monitoring support. *Comput Methods Programs Biomed*. (2025) 270:108950. doi: 10.1016/j.cmpb.2025.108950

44. Kumar D, Verma C, Illés Z. Federated learning with explainable AI for liver disease prediction: a privacy-preserving approach. *Intell-Based Med*. (2025) 12:100285. doi: 10.1016/j.ibmed.2025.100285