Journal of Dependence, 2026; 27(3):116-129

JOURNAL OF DEPENDENCE

Research Article

Diagnostic performance of a specific local Turkish large language model in detecting alcohol use disorder across internal medicine practice: A 42-patient pilot study

İlker Boğa, Mete Tuğcan Üçdal

DOI: 10.51982/bagimli.619

Abstract

Objective: Alcohol use disorder (AUD) is underdiagnosed in internal medicine. We evaluated a local large language model (LLM), fine-tuned on a Turkish medical corpus, for detecting AUD across outpatient, emergency and inpatient settings.

Method: We retrospectively analyzed 42 adult patients at a tertiary state hospital (January–May 2024); the de-identified dataset followed the MIMIC-IV schema. The reference standard was the clinician-confirmed DSM-5 diagnosis; comparators were the Alcohol Use Disorders Identification Test–Consumption (AUDIT-C), gamma-glutamyl transferase (GGT), mean corpuscular volume and the AST/ALT ratio. An in-hospital Llama-3-8B-Instruct model fine-tuned via low-rank adaptation produced binary predictions, probabilities and severity categories. Analyses used Wilson 95% confidence intervals, bootstrap AUC, Cohen's κ and paired DeLong tests.

Results: Mean age was 53.0 ± 15.0 years; 50.0% were female. AUD prevalence was 47.6% (20/42). The LLM achieved sensitivity 85.0% (95% CI, 64.0–94.8), specificity 90.9% (72.2–97.5), accuracy 88.1%, area under the curve (AUC) 0.864 (0.720–0.977) and $\kappa = 0.761$. In all 17 true positives, severity matched the reference exactly (quadratic-weighted $\nu = 1.000$). In this enriched cohort AUDIT-C and GGT separated perfectly (AUC = 1.000) and the LLM AUC was lower (DeLong $p = 0.038$), a ceiling effect of this case mix, not a general marker property.

Conclusion: The local LLM achieved clinically acceptable accuracy and exact severity classification within correctly flagged cases; its principal value lies where AUDIT-C is not routine. Because the cohort was enriched for AUD, predictive values and discrimination may differ substantially in routine populations. Multicenter prospective validation is warranted.

Keywords: Alcohol use disorder, DSM-5, Large language model, Artificial intelligence, Internal medicine, Diagnostic performance, Clinical decision support system

İlker Boğa
Department of Internal Medicine, Faculty of Medicine, Ankara Medipol University, Ankara, Türkiye
Email: ilkerboga87@hotmail.com
ORCID iD: 0000-0002-8166-1809

Mete Tuğcan Üçdal
Republic of Türkiye Ministry of Health, Ankara Etimesgut Şehit Sait Ertürk State Hospital, Department of Internal Medicine, Ankara, Türkiye
Hacettepe University Faculty of Medicine, Department of Medical Biochemistry (PhD program), Ankara, Türkiye
Email: mtucdal@gmail.com
ORCID iD: 0009-0002-1763-5857

Received: 2026-05-22
Accepted: 2026-08-05

e-ISSN: 3108-7973

INTRODUCTION

Alcohol use disorder (AUD) is among the leading public health problems affecting adult populations worldwide, contributing both to direct organ injury and to a broad range of medical and psychiatric comorbidities (Carvalho et al., 2019). Global data from the World Health Organization indicate that more than 3 million annual deaths are attributable to alcohol use and over 5% of the global disease burden is attributable to alcohol consumption (World Health Organization, 2018). In Türkiye, the 2022 Turkish Health Survey reports that 12.1% of individuals aged 15 years and older used alcohol during the previous 12 months, with a higher past-year prevalence among men (18.4%) than among women (5.9%) (Türkiye İstatistik Kurumu, 2023).

Internal medicine practice—encompassing the outpatient clinic, the emergency department, inpatient services, gastroenterology and hepatology units—constitutes one of the most critical contact points for the diagnosis of AUD. Elevations in liver function tests, dyspepsia, macrocytosis, uncontrolled hypertension, unexplained fatigue, alcohol-withdrawal-related tremor and insomnia are among the frequently underrecognized presentations of AUD. A meta-analysis by Mitchell and colleagues reported that primary and secondary care clinicians correctly identify AUD in only 25–50% of cases (Mitchell et al., 2012). Connor and colleagues similarly emphasized that overlooking AUD as an underlying diagnosis substantially increases long-term morbidity and mortality (Connor et al., 2016).

Among structured screening tools, the Alcohol Use Disorders Identification Test (AUDIT) and its three-item abbreviated form, AUDIT-C, are among the most widely used in internal medicine practice (Bush et al., 1998; Saunders et al., 1993). Turkish validation studies have demonstrated acceptable psychometric properties for both instruments (Saatcioğlu et al., 2002). In real-world practice, however, AUDIT/AUDIT-C is not administered in a substantial proportion of encounters owing to physician time constraints, patient reticence, and limited integration with the routine electronic health record (EHR) workflow (Rehm et al., 2016).

Traditional laboratory markers—gamma-glutamyl transferase (GGT), mean corpuscular volume (MCV), carbohydrate-deficient transferrin and the aspartate aminotransferase/alanine aminotransferase (AST/ALT) ratio—offer an alternative for AUD screening, yet their sensitivity and specificity remain moderate, and they can be significantly affected by non-alcohol-related etiologies (Conigrave et al., 2003; Niemelä, 2016).

In recent years, large language models (LLMs) have demonstrated notable results in clinical decision support, reflecting a broader convergence of human and machine intelligence in high-performance medicine (Topol, 2019). Singhal and colleagues showed that LLMs encoding clinical knowledge can achieve performance approaching expert-level evaluation (Singhal et al., 2023). Thirunavukarasu and colleagues highlighted the potential of LLMs in medical diagnosis, patient communication and decision support, as well as their limitations regarding privacy, hallucination and external validity (Thirunavukarasu et al., 2023). Goh and colleagues demonstrated in a randomized trial the contribution of clinician–LLM collaboration to diagnostic reasoning (Goh et al., 2024).

Widespread use of commercial, cloud-based LLM services within healthcare systems raises significant obstacles under data protection regulations including Türkiye's Law No. 6698 on the Protection of Personal Data (KVKK), the European Union General Data Protection Regulation (GDPR) and the United States Health Insurance Portability and Accountability Act (HIPAA). The requirement that patient data not leave the institution motivates healthcare organizations to favor locally deployed, on-premise open-weight LLMs running on in-hospital servers. Fine-tuning open-weight models such as the Llama family on domain-specific medical corpora through low-rank adaptation (LoRA) addresses both privacy and clinical fit (Grattafiori et al., 2024; Hu et al., 2022; Touvron et al., 2023).

In the context of AUD, the capacity of LLMs to perform DSM-5 risk and severity classification from Turkish clinical notes has been evaluated in only a limited number of studies. Most existing studies are based on English text and on general-purpose commercial models accessed through the cloud; data on locally deployed models adapted to the Turkish medical corpus and to clinical practice in Türkiye remain scarce.

In this pilot study, we aimed to evaluate the diagnostic performance of a specific local LLM—fine-tuned on a Turkish medical corpus and deployed on an in-hospital server—in binary AUD classification and severity grading against a DSM-5 reference among adult patients presenting to the internal medicine outpatient clinic, the emergency department and inpatient services. As a secondary aim, the model's comparative performance against AUDIT-C and traditional laboratory markers (GGT, MCV, AST/ALT ratio) and the error patterns of misclassified cases were examined.

METHODS

Study Design

This was a single-center, retrospective, diagnostic-accuracy pilot study conducted at Ankara Etimesgut Şehit Sait Ertürk State Hospital, Ankara, Türkiye. The study protocol was structured in accordance with the Standards for Reporting Diagnostic Accuracy Studies (STARD) 2015 guideline (Bossuyt et al., 2015). The study was conducted in accordance with the ethical principles laid out in the 2013 revision of the Declaration of Helsinki. The flow of records from initial screening to final inclusion is summarized in the STARD participant flow diagram (Supplementary Figure S1).

Patient Population and Data Source

Forty-two adult patients evaluated between January and May 2024 in the internal medicine outpatient clinic (n = 13, 30.9%), the emergency department (n = 11, 26.2%) or inpatient services (n = 18, 42.9%), and who met the following inclusion criteria, were included: (i) age 18 years or older; (ii) availability of a structured history and physical examination record at initial evaluation; (iii) complete blood count, liver function tests (AST, ALT, GGT, total bilirubin, alkaline phosphatase), MCV, albumin and international normalized ratio (INR) results obtained within the preceding 60 days and documented in the EHR; and (iv) completion of a structured clinician-administered DSM-5 evaluation.

Exclusion criteria were (i) pregnancy; (ii) underlying terminal malignancy or enrollment in a palliative care program; (iii) inability to obtain a reliable history because of severe psychiatric comorbidity; (iv) prior hospitalization with a primary hepatobiliary disease diagnosis; and (v) more than 15% missing data in the history or laboratory records. Among the consecutive records screened during the study period, the first 42 patients meeting eligibility criteria and with a completed DSM-5 reference assessment were included. The target of 42 patients was set on a feasibility basis appropriate to a pilot design rather than by a formal power calculation, and it corresponds to all eligible patients accrued within the predefined five-month enrollment window. The structured DSM-5 reference assessment was applied to all consecutively screened patients regardless of clinical suspicion of alcohol use, rather than selectively to patients with suspected alcohol-related pathology, so that the reference standard was not conditioned on clinical suspicion. The resulting enrichment of the cohort and its implications for generalizability are addressed in the Limitations section.

The dataset was structured according to the MIMIC-IV core schema commonly used in multicenter clinical databases and was de-identified (Johnson et al., 2023). All records originated from Ankara Etimesgut Şehit Sait Ertürk State Hospital and were collected specifically for this study. The data have no relationship to and no overlap with the actual MIMIC-IV corpus, which is a separate United States intensive-care database; the reference to MIMIC-IV denotes only a schema design that organizes patient, admission and event information in a comparable

relational structure. Each patient was assigned a unique study identifier (study_id), a surrogate subject identifier (mimic_style_subject_id) and a surrogate admission identifier (mimic_style_hadm_id); links to true identifiers were severed via a one-way hash function. Patient province of origin was recorded for demographic generalization purposes.

Data Collectiv on and Variables

Data were extracted in de-identified form from the hospital's EHR system by one investigator and independently verified by a second investigator. Variables collected included (i) demographics (age, sex, province of origin, body mass index); (ii) care setting and service (outpatient/emergency/inpatient; internal medicine, gastroenterology, hepatology, emergency medicine, psychiatry consultation); (iii) chief complaint, triage acuity and vital signs (systolic and diastolic blood pressure, heart rate, respiratory rate, body temperature, peripheral oxygen saturation); (iv) comorbidities (smoking, diabetes, hypertension, chronic liver disease, depression/anxiety) and the Charlson Comorbidity Index; (v) presence of an alcohol history in the clinical note and AUDIT-C score; (vi) DSM-5 AUD reference, severity category and ICD-10 alcohol-related codes; (vii) laboratory parameters (AST, ALT, GGT, alkaline phosphatase, total bilirubin, albumin, INR, platelet count, MCV, hemoglobin, creatinine, sodium, glucose, C-reactive protein, AST/ALT ratio); (viii) clinical management variables (thiamine administration, benzodiazepine protocol, psychiatry consultation); and (ix) hospital length of stay, in-hospital mortality and discharge disposition.

Identifiers such as names, surnames, national identity numbers, telephone numbers and addresses within clinical notes were removed using a custom named-entity recognition module and subsequently checked manually. The clinical note summary submitted to the LLM for each patient consisted of a field averaging approximately 1,450 characters (notes_prompt_excerpt_tr field). This excerpt was assembled by a fixed extraction script from the same source sections for every patient, namely the structured history of present illness, the past medical and social history, and the physician assessment recorded at the initial evaluation. Free-text addenda, family-witness statements and notes entered after the index evaluation were not included in this field, so that the information available to the model was uniform across patients. A representative de-identified example of this input, together with the full model prompt template, is provided as Supplementary Material S3.

Local Large Language Model Architecture and Deployment

The local LLM evaluated in this study was obtained by fine-tuning the open-weight Llama-3-8B-Instruct base model (Grattafiori et al., 2024) on a Turkish medical corpus using the LoRA method (Hu et al., 2022). The fine-tuning dataset comprised (i) PubMed publications with Turkish abstracts, (ii) official clinical guidelines of the Turkish Ministry of Health, (iii) consensus reports of Turkish Medical Association specialty societies, (iv) the Turkish DSM-5-TR and ICD-10 F10 diagnostic criteria, and (v) Turkish validation studies of the AUDIT, AUDIT-C, CIWA-Ar and CAGE instruments. The corpus contained approximately 4.2 million Turkish medical sentences. LoRA adapters with rank = 16 and alpha = 32 were used, with a learning rate of 2×10^{-4}, a batch size of 8, and training over 3 epochs.

The model was deployed using 4-bit quantization (Q4_K_M) on an in-hospital GPU server equipped with an NVIDIA RTX A6000 (48 GB VRAM) running the vLLM inference engine. All inference operations were performed on the hospital intranet, without Internet access. This configuration was designed to ensure compliance with the KVKK and HIPAA.

A single inference was performed per patient; sampling temperature was set to 0.2 and top-p to 0.9. The model prompt contained the de-identified clinical note summary, vital signs, comorbidities and laboratory values in a structured JSON format. The model output was also structured as a JSON schema comprising (i) the binary AUD prediction (0/1), (ii)

the AUD probability score (0–1), (iii) the DSM-5 severity category (Mild/Moderate/Severe; provided only when probability ≥ 0.5), and (iv) a structured rationale category (explicit alcohol history; no supportive findings; missed alcohol history; over-weighting of laboratory pattern).

Reference Standard

The reference standard consisted of a structured clinician-administered DSM-5 evaluation conducted face-to-face during the outpatient visit or hospitalization. A diagnosis of AUD required the presence of at least two of the eleven DSM-5 criteria within the past 12 months; severity was classified as Mild (2–3 criteria), Moderate (4–5 criteria) or Severe (six or more criteria) (American Psychiatric Association, 2013). In addition, alcohol history documented in the clinical note and the AUDIT-C score were recorded for secondary comparison. The clinician establishing the reference diagnosis did not have access to the local LLM output; conversely, the local LLM had no access to the details of the DSM-5 assessment. We note, however, that several inputs provided to the model, namely the documented alcohol history, the AUDIT-C score and the hepatic laboratory values, are themselves correlates of, or contributors to, individual DSM-5 criteria. The model therefore operated on information that overlaps with the construct measured by the reference standard, as occurs in routine clinical practice. Complete independence between the index test inputs and the reference standard should not be assumed.

Statistical Analysis

Continuous variables were summarized as mean ± standard deviation when normally distributed and as median (interquartile range) otherwise. Categorical variables were reported as frequencies and percentages. Normality was assessed by the Shapiro–Wilk test. Group comparisons used the independent-samples *t* test, the Mann–Whitney *U* test, the chi-square test and Fisher's exact test.

Diagnostic performance indicators included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR−) and accuracy; 95% Wilson confidence intervals were used for all point estimates (Wilson, 1927). Areas under the receiver operating characteristic (ROC) curve (AUC) were calculated for the LLM probability score and comparator markers, with 95% confidence intervals obtained from 2,000 bootstrap iterations. AUC comparisons were performed using the paired DeLong test (DeLong et al., 1988). When a comparator achieved complete separation between AUD-positive and AUD-negative patients, its bootstrap confidence interval collapsed to a single point (1.000 to 1.000) because every resample reproduced perfect separation. This behavior is the expected consequence of complete separation rather than an artifact of underdispersion. In paired DeLong comparisons against such markers, both the comparator variance and its covariance with the LLM probability score are zero. The test statistic therefore reduces to the standardized distance of the LLM AUC from 1.000 and is identical across all perfectly separating comparators. The standard error of the LLM AUC used in these comparisons was 0.066, and the corresponding variance–covariance terms are reported in Supplementary Table S1. Complete separation for each perfectly separating marker is displayed at the individual-patient level in Supplementary Figure S2. As a post-hoc characterization of precision, the achieved power for the primary AUC against the null value of 0.500 at the two-sided 5% level was greater than 0.90.

Agreement between the local LLM binary output and the DSM-5 reference was assessed with Cohen's κ, and agreement on severity classification with the quadratic-weighted κ. Interpretation followed Landis and Koch's framework (Landis & Koch, 1977). All statistical analyses were performed using Python 3.11 (scikit-learn 1.4, statsmodels 0.14, SciPy 1.13) and R 4.3 (pROC and irr packages). Two-sided $p < 0.05$ was considered the threshold of statistical significance.

Ethics Approval

The study was approved by the Republic of Türkiye Ministry of Health, Ankara Provincial Health Directorate Non-Interventional Ethics Committee (decision dated 24 October 2025, number 2025-10-3). Owing to the retrospective design, the committee waived the requirement for individual informed consent; all patient data were anonymized in accordance with the KVKK. The committee approval encompassed the de-identified reporting of individual misclassified cases. These case descriptions contain no direct identifiers, province of origin is reported only at the cohort level and is not linked to individual cases, and no combination of the reported attributes is sufficient for re-identification.

RESULTS

Patient Characteristics

As shown in **Table 1**, of the 42 patients included in the analysis, 21 (50.0%) were male and 21 (50.0%) were female, with a mean age of 53.0 ± 15.0 years. The mean body mass index was 27.2 ± 4.0 kg/m² and the median Charlson Comorbidity Index was 2.0 (1.0–3.0). Hypertension was present in 25 patients (59.5%), diabetes mellitus in 12 (28.6%), chronic liver disease in 11 (26.2%), and depression/anxiety in 23 (54.8%). Active smoking was reported by 12 patients (28.6%).

Distribution by care setting was: inpatient hospitalization 18 (42.9%), internal medicine outpatient clinic 13 (30.9%) and emergency department 11 (26.2%). The distribution by service was led by gastroenterology (n = 14, 33.3%) and internal medicine (n = 13, 30.9%), followed by emergency medicine (n = 9, 21.4%), hepatology (n = 4, 9.5%) and psychiatry consultation (n = 2, 4.8%). The dataset represented 12 provinces of patient origin, with Kayseri (n = 7), Ankara (n = 6) and Adana (n = 6) being the most represented. Care setting and service represent two independent classification axes applied to the same 42 patients. Care setting denotes the physical location of the encounter, namely outpatient clinic, emergency department or inpatient ward, whereas service denotes the responsible clinical

Table 1. Demographic, clinical and care-setting characteristics in the overall sample and stratified by DSM-5 reference (n = 42)

Variable	Overall (n = 42)	DSM-5 AUD (−) (n = 22)	DSM-5 AUD (+) (n = 20)	p
Age, mean ± *SD*, years	53.0 ± 15.0	52.3 ± 16.1	53.8 ± 14.1	0.762
Male sex, *n* (%)	21 (50.0)	9 (40.9)	12 (60.0)	0.215
Body mass index, kg/m²	27.2 ± 4.0	27.2 ± 3.9	27.1 ± 4.2	0.930
Charlson Comorbidity Index	2.0 (1.0–3.0)	1.5 (1.0–2.0)	2.0 (1.0–3.0)	0.282
Hypertension, *n* (%)	25 (59.5)	13 (59.1)	12 (60.0)	0.953
Diabetes mellitus, *n* (%)	12 (28.6)	8 (36.4)	4 (20.0)	0.241
Chronic liver disease, *n* (%)	11 (26.2)	0 (0.0)	11 (55.0)	<0.001
Depression/anxiety, *n* (%)	23 (54.8)	7 (31.8)	16 (80.0)	0.002
Active smoking, *n* (%)	12 (28.6)	5 (22.7)	7 (35.0)	0.378
Alcohol history in note, *n* (%)	20 (47.6)	0 (0.0)	20 (100.0)	<0.001
AUDIT-C score, median (IQR)	3.0 (1.0–9.0)	1.0 (0.3–2.0)	9.0 (7.0–11.3)	<0.001
Outpatient clinic, *n* (%)	13 (30.9)	7 (31.8)	6 (30.0)	1.000
Emergency department, *n* (%)	11 (26.2)	6 (27.3)	5 (25.0)	1.000
Inpatient hospitalization, *n* (%)	18 (42.9)	9 (40.9)	9 (45.0)	0.789

Note. *SD = standard deviation; IQR = interquartile range; AUD = alcohol use disorder; AUDIT-C = Alcohol Use Disorders Identification Test – Consumption. Comparisons between AUD (−) and AUD (+) groups used the independent-samples t test, the Mann–Whitney U test and Fisher's exact test.*

specialty; gastroenterology and hepatology patients are therefore distributed across the outpatient and inpatient settings rather than forming a separate location category.

DSM-5 AUD was confirmed in 20 patients (47.6%); the severity distribution was Mild 5 (25.0%), Moderate 10 (50.0%) and Severe 5 (25.0%). Among the 20 AUD-positive patients, the primary alcohol-related ICD-10 code recorded per patient was K70.10 (alcoholic hepatitis, n = 6), E52 (Wernicke encephalopathy, n = 5), F10.10 (uncomplicated alcohol abuse, n = 4), F10.239 (alcohol withdrawal, n = 3) and F10.20 (alcohol dependence, n = 2). Each patient was assigned a single primary alcohol-related code for this tabulation; secondary F10.x codes documented for some patients are not represented here. Within the AUD-positive group, 17 patients (85.0%) received thiamine administration, 8 (40.0%) were started on a benzodiazepine withdrawal protocol, and 10 (50.0%) had a psychiatry consultation.

Clinical and Laboratory Markers

In the overall population, the median GGT was 118 (62–290) U/L, MCV 94.8 (88.6–101.9) fL, and AST/ALT ratio 1.38 (1.05–2.15).

As shown in **Table 2**, all hepatic markers were significantly elevated in the DSM-5 AUD-positive group: GGT 295.5 (196.5–396.0) versus 83.0 (62.2–107.5) U/L ($p < 0.001$), AST 176.5 (136.5–282.2) versus 41.5 (30.5–49.8) U/L ($p < 0.001$), ALT 112.0 (82.8–152.2) versus 33.5 (21.5–44.2) U/L ($p < 0.001$), MCV 101.9 (97.9–108.5) versus 90.8 (85.1–93.3) fL ($p < 0.001$) and total bilirubin 1.80 (1.30–2.80) versus 0.70 (0.60–0.90) mg/dL ($p < 0.001$). Albumin was significantly lower in the AUD-positive group: 2.90 (2.70–3.40) versus 4.20 (3.80–4.50) g/dL ($p < 0.001$). The AUDIT-C score was 9.0 (7.0–11.3) in the AUD-positive group and 1.0 (0.3–2.0) in the AUD-negative group ($p < 0.001$).

Binary Diagnostic Performance of the Local LLM

Across the 42-patient series, the local LLM produced 17 true positive (TP), 20 true negative (TN), 2 false positive (FP) and 3 false negative (FN) classifications. Binary classification performance was: sensitivity 85.0% (95% CI, 64.0–94.8), specificity 90.9% (95% CI, 72.2–97.5), PPV 89.5% (95% CI, 68.6–97.1), NPV 87.0% (95% CI, 67.9–95.5), overall accuracy 88.1%, LR+ 9.35 and LR− 0.17 (**Table 3**).

Table 2. Laboratory parameters by DSM-5 reference (n = 42)

Parameter	Overall	DSM-5 AUD (−) (n = 22)	DSM-5 AUD (+) (n = 20)	p
AST, U/L	60.0 (39.0–168.5)	41.5 (30.5–49.8)	176.5 (136.5–282.2)	<0.001
ALT, U/L	57.5 (33.0–110.5)	33.5 (21.5–44.2)	112.0 (82.8–152.2)	<0.001
GGT, U/L	118.0 (62.0–290.0)	83.0 (62.2–107.5)	295.5 (196.5–396.0)	<0.001
ALP, U/L	128.5 (84.5–185.0)	97.0 (76.0–143.0)	170.5 (130.5–211.5)	0.003
Total bilirubin, mg/dL	0.99 (0.66–1.83)	0.70 (0.60–0.90)	1.80 (1.30–2.80)	<0.001
Albumin, g/dL	3.54 (3.05–4.20)	4.20 (3.80–4.50)	2.90 (2.70–3.40)	<0.001
INR	1.09 (1.00–1.20)	1.00 (1.00–1.10)	1.20 (1.10–1.30)	<0.001
Platelets, ×10³/μL	222.0 (180.0–270.5)	247.5 (212.2–297.5)	197.5 (151.8–240.8)	0.006
MCV, fL	94.8 (88.6–101.9)	90.8 (85.1–93.3)	101.9 (97.9–108.5)	<0.001
AST/ALT ratio	1.38 (1.05–2.15)	1.10 (0.90–1.70)	1.80 (1.20–2.20)	0.053
CRP, mg/L	22.0 (12.7–47.0)	12.8 (7.4–25.0)	47.5 (32.0–66.5)	<0.001

Note. *Values are presented as median (interquartile range). ALP = alkaline phosphatase; ALT = alanine aminotransferase; AST = aspartate aminotransferase; CRP = C-reactive protein; GGT = gamma-glutamyl transferase; INR = international normalized ratio; MCV = mean corpuscular volume. Group comparisons used the Mann–Whitney U test.*

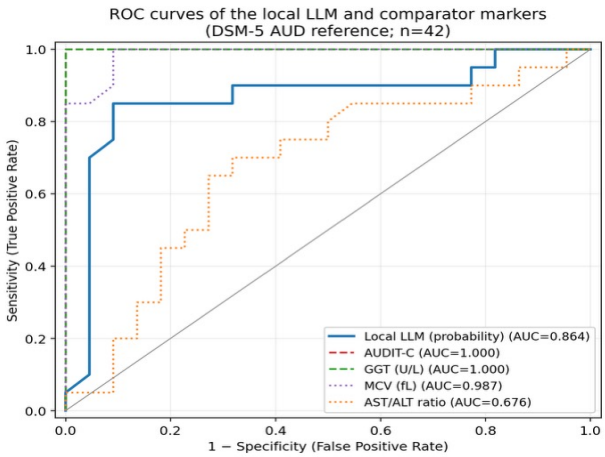

Figure 1. Receiver operating characteristic (ROC) curves of the local LLM and comparator markers against the DSM-5 AUD reference (n = 42). The local LLM probability score yielded AUC = 0.864 (95% CI, 0.720–0.977); AUDIT-C and GGT achieved perfect separation (AUC = 1.000) in this enriched pilot cohort; these values are cohort-specific and are not general estimates of marker performance. MCV (AUC = 0.987) and AST/ALT ratio (AUC = 0.676) are shown as comparators. The diagonal grey line represents a non-discriminating reference (AUC = 0.500).

Table 3. Binary diagnostic performance of the local LLM against the DSM-5 AUD reference (n = 42)

Indicator	Value	95% CI (Wilson)
True positive (TP), *n*	17	—
False positive (FP), *n*	2	—
True negative (TN), *n*	20	—
False negative (FN), *n*	3	—
Sensitivity, %	85.0	64.0–94.8
Specificity, %	90.9	72.2–97.5
Positive predictive value (PPV), %	89.5	68.6–97.1
Negative predictive value (NPV), %	87.0	67.9–95.5
Overall accuracy, %	88.1	75.0–94.8
Positive likelihood ratio (LR+)	9.35	—
Negative likelihood ratio (LR−)	0.17	—
Cohen's κ (binary)	0.761	0.553–0.969
AUC (LLM probability score)	0.864	0.720–0.977

Note. *AUC = area under the curve; CI = confidence interval; LLM = large language model. Confidence intervals were computed with the Wilson method for point proportions and with 2,000 bootstrap iterations for the AUC. All indices describe the present enriched pilot cohort.*

On ROC analysis of the LLM probability score, the AUC was 0.864 (95% CI, 0.720–0.977; **Figure 1**), and the corresponding classification matrix is shown in **Figure 2A**. Cohen's κ between the local LLM and the DSM-5 reference was 0.761 (95% CI, 0.553–0.969), corresponding to substantial agreement on the Landis and Koch scale (Landis & Koch, 1977).

Severity Classification Performance

Of the 19 patients flagged as AUD-positive by the local LLM, 17 had a true AUD diagnosis (true positives) and were assigned a DSM-5 severity category by the model. In all 17 of these cases, the LLM severity category matched the DSM-5 reference exactly (Mild 5/5, Moderate 10/10, Severe 2/2); the remaining 3 severe AUD cases were missed by the model and therefore had no severity prediction. The quadratic-weighted Cohen's κ for severity agreement within true-positive cases was 1.000 (**Figure 2B**). This value was computed only among correctly detected cases, is therefore conditional on correct binary classification, and rests on a small subgroup (n = 17). Because the three missed Severe cases are

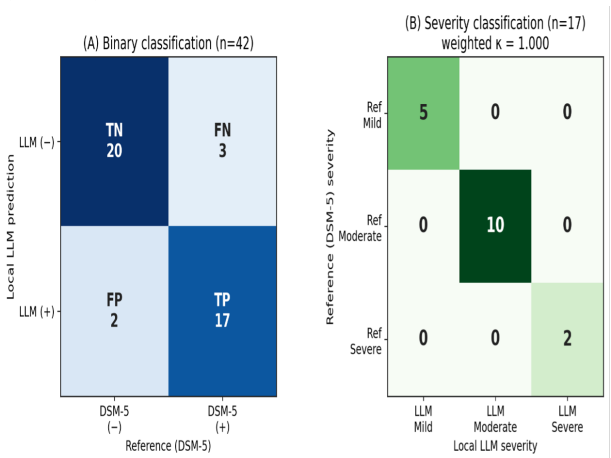

Figure 2. Classification matrices for the local LLM. (A) Binary classification: 17 true positives, 20 true negatives, 2 false positives and 3 false negatives (accuracy 88.1%; Cohen's κ = 0.761). (B) DSM-5 severity classification: in all 17 true-positive cases the LLM severity category matched the reference exactly (quadratic-weighted κ = 1.000). Color intensity is proportional to case counts.

absent from this subgroup, the severity distribution analyzed here (Mild 5, Moderate 10, Severe 2) does not reproduce the full reference distribution (Mild 5, Moderate 10, Severe 5). The value should accordingly be read as a within-cohort observation rather than as an estimate of severity-grading accuracy.

Comparison With Comparator Markers

ROC analyses of comparator markers alone are summarized in **Table 4**. In this enriched pilot cohort, AUDIT-C (AUC = 1.000; 95% CI, 1.000–1.000), GGT (AUC = 1.000; 95% CI, 1.000–1.000) and AST (AUC = 1.000) achieved perfect separation; at the AUDIT-C ≥ 4 threshold, all 20 AUD-positive patients and all 22 AUD-negative patients were correctly classified (sensitivity and specificity both 100%). MCV (AUC = 0.987), albumin (AUC = 0.953) and total bilirubin (AUC = 0.948) demonstrated very high discrimination. By contrast, the AST/ALT ratio showed only moderate discrimination (AUC = 0.676). These perfect and near-perfect values describe discrimination within the present enriched cohort and are not offered as general performance estimates for these markers.

Within this cohort, paired DeLong tests showed that the local LLM probability score (AUC 0.864) had a lower AUC than AUDIT-C, GGT and AST ($\Delta AUC = -0.136$; $z = -2.07$; $p = 0.038$ for each comparison). Because each of these comparators achieved complete separation, the comparator variance and its covariance with the LLM probability score were both zero; the DeLong statistic therefore depended only on the standard error of the LLM AUC and was identical across the three comparisons. Differences between the LLM and MCV or ALT were of borderline significance ($p = 0.051$ and $p = 0.078$, respectively). No significant difference was detected against the other comparators (albumin, INR, bilirubin, platelets, AST/ALT ratio).

Qualitative Analysis of Misclassification Patterns

Each of the five misclassified cases was examined according to the model's structured rationale category. The three false negatives—a 74-year-old man, a 32-year-old man and a 55-year-old man—were patients clinically classified as Severe AUD by DSM-5, with an explicit alcohol history in the clinical note, AUDIT-C scores ranging from 7 to 12 and GGT values between 352 and 599 U/L. In all three the LLM rationale category was

Table 4. AUC values for the local LLM and comparator markers, with paired DeLong test results (n = 42)

Comparator	AUC	95% CI (bootstrap)	ΔAUC (LLM − comparator)	p (DeLong)
Local LLM (probability)	0.864	0.720–0.977	—	—
AUDIT-C score	1.000	1.000–1.000	−0.136	0.038
GGT, U/L	1.000	1.000–1.000	−0.136	0.038
AST, U/L	1.000	1.000–1.000	−0.136	0.038
MCV, fL	0.987	0.957–1.000	−0.123	0.051
ALT, U/L	0.984	0.945–1.000	−0.120	0.078
Albumin (inverted), g/dL	0.953	0.870–1.000	−0.090	0.234
Total bilirubin, mg/dL	0.948	0.865–1.000	−0.084	0.274
INR	0.909	0.793–0.989	−0.045	0.554
Platelets (inverted)	0.751	0.591–0.880	+0.112	0.292
AST/ALT ratio	0.676	0.494–0.844	+0.188	0.078

Note. *AUC = area under the curve; CI = confidence interval; LLM = large language model. Because higher values of albumin and platelets reduce the probability of AUD, scores were inverted (negated) before AUC calculation. AUC values of 1.000 denote complete separation within this enriched pilot cohort and are not general estimates of marker performance.*

labeled "missed alcohol history"; alcohol use was reported primarily in family-witness accounts or in subsequent notes and was therefore intermittently reflected in the structured portion of the record. The alcohol-history variable summarized in Table 1 records whether an alcohol history was documented anywhere in the complete medical record, whereas the text supplied to the model was limited to the structured initial-evaluation excerpt described in the Methods. In these three cases an alcohol history was present in the record as a whole, and was therefore coded as present in Table 1, but it did not appear in the structured excerpt available to the model, which accounts for the false-negative classifications. The two false positives—a 35-year-old woman and a 59-year-old man—were patients without a documented alcohol history in the clinical note, AUDIT-C scores of 2 and 1, and GGT values of 106 and 88 U/L; in both cases the LLM rationale was labeled "over-weighting of laboratory pattern."

Overall accuracy by care setting was 11/11 (100%) in the emergency department, 12/13 (92.3%) in the internal medicine outpatient clinic and 14/18 (77.8%) in the inpatient group; the accumulation of misclassifications in the inpatient group suggests that the model encountered greater difficulty in cases with advanced hepatic decompensation in whom acute alcohol use could not be unambiguously established.

DISCUSSION

In this pilot study, a specific local LLM fine-tuned on a Turkish medical corpus and deployed on an in-hospital server achieved 88.1% overall accuracy, 85.0% sensitivity, 90.9% specificity and an AUC of 0.864 for AUD detection against a DSM-5 reference in internal medicine practice, and demonstrated exact DSM-5 severity classification in every case it flagged as AUD-positive. Substantial agreement with the clinician diagnosis was observed (κ = 0.761). In this small, enriched cohort AUDIT-C and GGT achieved perfect separation and the local LLM showed a lower AUC than these two markers; this ranking is a characteristic of the present cohort and cannot be transferred to unselected populations.

Our findings are broadly consistent with prior studies demonstrating the potential of LLM-based clinical decision support, but emphasize two important nuances. First, the existence of well-established, low-cost screening instruments (AUDIT, AUDIT-C) with high accuracy makes it inherently difficult for LLMs to establish absolute superiority in this task. Singhal and colleagues had previously emphasized that the ability of LLMs to approach expert-level performance is sensitive to task type and dataset enrichment (Singhal et al., 2023). Second, the ceiling effect observed in our study reflects a cohort in which AUD can largely be predicted from enzyme patterns, macrocytosis and the clinical history, equalizing the discrimination of each method.

The performance of traditional biomarkers in AUD screening has typically been reported as moderate; Conigrave and colleagues described single-marker sensitivities of 30–70% and specificities of 60–90% for GGT and MCV (Conigrave et al., 2003). Niemelä's broad review likewise emphasized that these markers are insufficient as stand-alone tools in real-world samples (Niemelä, 2016). The perfect separation observed in our cohort likely reflects the combined effect of a high AUD prevalence (47.6%), the presence of an AUD subgroup with substantial chronic liver disease burden, and a relatively healthy comparison group. This pattern indicates that, in a test set composed of cases in which the AUD diagnosis is already clinically conspicuous, opportunities for the LLM to display incremental superiority are reduced. We regard this enrichment, together with the spectrum of the cohort, as the primary explanation for the perfect biomarker discrimination observed here, rather than as a secondary observation. A within-sample AUC of 1.000 for single biomarkers such as GGT, AST or MCV is not expected in unselected internal medicine populations, in which these markers typically show only moderate accuracy. The present result should therefore be read as a property of an enriched and clinically conspicuous cohort and not as evidence that these markers outperform the model in general practice. Accordingly, the statement that

the model showed a lower AUC than AUDIT-C and GGT describes this specific cohort and should not be generalized.

Two distinct clinical value axes were nevertheless observed for the local LLM. First, exact DSM-5 severity classification was achieved in all true positives—an output that AUDIT-C cannot structurally generate, bridging screening and triage. Second, the structured rationale categories produced by the model provide an interpretable feedback mechanism for the clinician workflow. Analysis of the misclassification patterns revealed two principal failure modes of the model: (i) missing alcohol histories that appear only in free-text notes or family-witness accounts and not in structured fields (3/3 FN) and (ii) over-weighting of laboratory patterns in cases without an explicit alcohol history (2/2 FP). Both patterns could be addressed in clinical context with additional safeguards (e.g., a dedicated structured alcohol-history prompt or mandatory history confirmation when laboratory evidence is the principal driver of risk).

The choice of local (on-premise) deployment offers a critical advantage for clinical translation. Use of cloud-based commercial LLM services in healthcare is significantly constrained by KVKK and international regulations (HIPAA, GDPR). The LoRA method introduced by Hu and colleagues (Hu et al., 2022) enables domain-specific fine-tuning at relatively low hardware cost without retraining all model parameters, permitting deployment on a single GPU server as in our study. A further favorable finding for practical applicability is the high negative likelihood ratio (0.17): this value reduces the pre-test probability of AUD in a patient flagged as negative by approximately six-fold.

A further point concerns the scope of the model evaluated here. The index test was a single 8-billion-parameter open-weight model, fine-tuned with LoRA on a Turkish medical corpus and served entirely inside the hospital, and the results describe that specific configuration. They are not estimates of what general-purpose commercial large language models would achieve on the same task. Commercial systems are considerably larger, are updated on schedules the deploying institution does not control, and are reached through external application programming interfaces, which is the arrangement that KVKK, GDPR and HIPAA constrain in routine care. A compact locally fine-tuned model makes the opposite trade: less general reasoning capacity, but domain and language fit, auditability, version stability and data residency. The accuracy reported here should therefore be attributed to this locally fine-tuned Turkish model rather than to large language models in general, and head-to-head comparison with general-purpose commercial models remains an open question for future work.

Subgroup analyses indicated that the model achieved its best performance in the emergency department (100% accuracy) and outpatient clinic (92.3% accuracy) settings, with relative difficulty in the inpatient group (77.8% accuracy) because of cases of advanced hepatic decompensation in whom acute alcohol use was unclear. This pattern highlights the need for setting-specific calibration when integrating clinical decision support systems.

Strengths of our study include adherence to the STARD 2015 reporting guideline, use of a DSM-5 reference standard, a de-identified data structure modeled on the MIMIC-IV schema, double-blinding between index test and reference standard, direct comparison with traditional biomarkers and AUDIT-C, robust statistical reporting with paired DeLong tests and bootstrap confidence intervals, an on-premise privacy-compliant architecture, and interpretability afforded by structured error categorization.

Limitations of the Study

Several limitations should be acknowledged. First, the sample size is limited to 42 patients and the study is pilot in nature; confidence intervals are relatively wide, which limits the power for multiple-comparison analyses. Second, the AUD prevalence

of 47.6% far exceeds the prevalence encountered in routine internal medicine practice (typically 5–15%). Because PPV and NPV depend directly on prevalence, both would change substantially in a routine population: at a prevalence of 10%, holding the observed sensitivity and specificity constant, the PPV would fall from 89.5% to approximately 51% while the NPV would rise to approximately 98%. The ROC characteristics reported here are likewise cohort-dependent, because discrimination was estimated over a case mix in which clearly affected and clearly unaffected patients predominate. Every diagnostic index in this report should therefore be read as descriptive of the present enriched cohort rather than as an expected value for routine internal medicine populations. Third, the single-center retrospective design limits external validity. Fourth, model outputs were obtained from a single inference and test–retest reliability was not evaluated; because the sampling temperature was non-zero (0.2), each output is a single realization of a stochastic process, and the paired comparison of this single realization against deterministic laboratory values should be interpreted with that stochasticity in mind. Fifth, the impact of the local LLM on long-term follow-up, treatment guidance and patient outcomes was beyond the scope of this study. Finally, risks arising from hallucination tendencies and out-of-scope clinical information generation were not analyzed in detail. The structured DSM-5 reference assessment was applied to all consecutively screened patients regardless of clinical suspicion, so the study is not subject to selective verification of the reference standard. The eligibility requirements and the case mix of the source services, which were weighted toward gastroenterology and hepatology, nonetheless produced an enriched cohort in which clearly affected and clearly unaffected patients predominate over diagnostically ambiguous cases. This enrichment, rather than selective application of the reference standard, is the most plausible reason that several single biomarkers reached an AUC of 1.000, a result that should not be expected in unselected populations.

Within these limitations, future studies are recommended to focus on (i) multicenter prospective validation, (ii) consecutive sampling that reflects real-world prevalence, (iii) testing the performance of integrated algorithms in which the LLM is combined with AUDIT-C, GGT and clinician evaluation, (iv) measurement of the effect on screening frequency, decisions to administer thiamine and benzodiazepines, and referral rates to psychiatry consultation when the model is integrated into the clinical workflow, and (v) extended validation of the model for other substance use disorders (opioid, benzodiazepine, cannabis).

CONCLUSION

A specific local LLM fine-tuned on a Turkish medical corpus and deployed on an in-hospital server demonstrated, in this 42-patient pilot study, clinically acceptable diagnostic performance (accuracy 88.1%, AUC 0.864) for the detection of AUD against DSM-5 criteria in internal medicine practice, together with exact agreement in severity classification. AUDIT-C and GGT achieved perfect separation in this small, enriched cohort, a result specific to the present case mix rather than a general property of these markers; accordingly, the principal clinical value of the model lies in workflows where AUDIT-C is not routinely administered and where automatic risk flagging, severity grading and interpretable rationale generation from free-text clinical notes are required. Privacy-preserving, locally deployable models thus represent a promising decision support tool for broadening AUD screening across internal medicine practice. These findings apply to this specific locally fine-tuned Turkish model and to this enriched pilot cohort; multicenter prospective validation in unselected populations is essential before they can be translated into clinical practice.

REFERENCES

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., & Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ, 351, h5527. https://doi.org/10.1136/bmj.h5527

Bush, K., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., & Bradley, K. A. (1998). The AUDIT alcohol consumption questions (AUDIT-C): An effective brief screening test for problem drinking. Archives of Internal Medicine, 158(16), 1789–1795. https://doi.org/10.1001/archinte.158.16.1789

Carvalho, A. F., Heilig, M., Perez, A., Probst, C., & Rehm, J. (2019). Alcohol use disorders. The Lancet, 394(10200), 781–792. https://doi.org/10.1016/S0140-6736(19)31775-1

Conigrave, K. M., Davies, P., Haber, P., & Whitfield, J. B. (2003). Traditional markers of excessive alcohol use. Addiction, 98(s2), 31–43. https://doi.org/10.1046/j.1359-6357.2003.00581.x

Connor, J. P., Haber, P. S., & Hall, W. D. (2016). Alcohol use disorders. The Lancet, 387(10022), 988–998. https://doi.org/10.1016/S0140-6736(15)00122-1

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845. https://doi.org/10.2307/2531595

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. JAMA Network Open, 7(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). The Llama 3 herd of models (arXiv:2407.21783). arXiv. https://doi.org/10.48550/arXiv.2407.21783

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models [Conference paper]. Tenth International Conference on Learning Representations (ICLR 2022). https://openreview.net/forum?id=nZeVKeeFYf9

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10, Article 1. https://doi.org/10.1038/s41597-022-01899-x

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Mitchell, A. J., Meader, N., Bird, V., & Rizzo, M. (2012). Clinical recognition and recording of alcohol disorders by clinicians in primary and secondary care: Meta-analysis. The British Journal of Psychiatry, 201(2), 93–100. https://doi.org/10.1192/bjp.bp.110.091199

Niemelä, O. (2016). Biomarker-based approaches for assessing alcohol use disorders. International Journal of Environmental Research and Public Health, 13(2), Article 166. https://doi.org/10.3390/ijerph13020166

Rehm, J., Anderson, P., Manthey, J., Shield, K. D., Struzzo, P., Wojnar, M., & Gual, A. (2016). Alcohol use disorders in primary health care: What do we know and where do we go? Alcohol and Alcoholism, 51(4), 422–427. https://doi.org/10.1093/alcalc/agv127

Saatcioğlu, Ö., Evren, C., & Çakmak, D. (2002). Alkol kullanım bozuklukları tanıma testinin (AKBT) geçerlik ve güvenirlik çalışması [Validity and reliability study of the Alcohol Use Disorders Identification Test]. Türkiye'de Psikiyatri, 4(2–3), 107–113.

Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R., & Grant, M. (1993). Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. Addiction, 88(6), 791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

Türkiye İstatistik Kurumu. (2023). Türkiye sağlık araştırması, 2022 [Turkish Health Survey, 2022] (Haber Bülteni No. 49747). https://data.tuik.gov.tr/Bulten/Index?p=Turkiye-Saglik-Arastirmasi-2022-49747

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953

World Health Organization. (2018). Global status report on alcohol and health 2018. https://www.who.int/publications/i/item/9789241565639

Peer-Review

Double blind peer reviewed

Conflict of Interest

The authors declare that they have no conflict of interests regarding content of this article.

Financial Support

The Authors report no financial support regarding content of this article.

Ethical Declaration

Ethical permission was obtained from the Republic of Türkiye Ministry of Health, Ankara Provincial Health Directorate Non-Interventional Ethics Committee for this study with date 24 October 2025 and number 2025-10-3, and Helsinki Declaration rules were followed to conduct this study. Owing to the retrospective design, the requirement for individual informed consent was waived by the Ethics Committee; all patient data were anonymized in accordance with Law No. 6698 on the Protection of Personal Data (KVKK).

Author Contributions

Concept: İ.B., M.T.Ü.; Design: İ.B., M.T.Ü.; Supervising: M.T.Ü.; Financing and equipment: İ.B., M.T.Ü.; Data collection and entry: İ.B., M.T.Ü.; Analysis and interpretation: İ.B., M.T.Ü.; Literature search: İ.B., M.T.Ü.; Writing: M.T.Ü.; Critical review: İ.B., M.T.Ü. **else:** Development and deployment of the local large language model — M.T.Ü.

Thanks

The authors thank the Information Technology Unit of Ankara Etimesgut Şehit Sait Ertürk State Hospital for support with data extraction, and the Department of Medical Biochemistry, Hacettepe University Faculty of Medicine, for statistical analysis consultation

Corresponding Author

Mete Tuğcan Üçdal

Specialist in Internal Medicine. Republic of Türkiye Ministry of Health, Ankara Etimesgut Şehit Sait Ertürk State Hospital, Department of Internal Medicine, Ankara, Türkiye.

Email: mtucdal@gmail.com

ORCID iD: 0009-0002-1763-5857