International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

From raw scores to cohort standing: a systematic review of large language model performance on Korean health professions licensing examinations

INPLASY202690111
doi: 10.37766/inplasy2026.9.0111
Received: 24 September 2026
Published: 24 September 2026

Lee, MS; Lee, HY.

Corresponding author:
Myeong Soo Lee

drmslee@gmail.com

Author Affiliation:
Korea Institute of Oriental Medicine.

ADMINISTRATIVE INFORMATION

Support - Korea Institute of Oriental Medicine (KSN2123211).

Review Stage at time of this submission - Piloting of the study selection process.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202690111

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 24 September 2026 and was last updated on 24 September 2026.

INTRODUCTION

Review question / Objective Where do the accuracies reported for large language models (LLMs) on Korean national health personnel licensing examinations fall within the score distributions of the human candidates who sat the same examination sessions, and what does an LLM "pass" therefore mean under Korea's fixed cut score (60% of the total, 40% in every subject, applied to unequated raw scores)?
Secondary questions: (a) How often does a reported LLM accuracy exceed the session mean of human candidates, and how does the model's within-cohort percentile vary across examinations of different difficulty? (b) Does the same model receive discordant pass/fail verdicts across sessions that differ only in difficulty? (c) How do subject-level accuracies, particularly in the health-law subject, relate to the 40% subject minimum? (d) What is the design and reporting quality of the included studies?

Rationale Since 2023, studies have reported that generative LLMs "pass" Korean national licensing examinations for physicians, dentists, Korean medicine doctors, pharmacists, nurses, and allied health professions. All of these examinations apply the same fixed cut score to unequated raw scores, so the meaning of a pass depends on the difficulty of the particular session: in an easy session most human candidates pass, in a hard session many fail, and a model's accuracy is interpretable only relative to the human cohort of the same session. No review has relocated the reported model accuracies within the human score distributions of the same sessions, which are available from the Korea Health Personnel Licensing Examination Institute (KHPLEI) open data. This review will do so for every published evaluation of an LLM on a KHPLEI examination.

Condition being studied Health professions education and assessment: national licensing examinations for health personnel in the Republic of Korea administered by KHPLEI, and the

evaluation of generative large language models on those examinations. No clinical condition is studied; the human reference population is the candidate cohort of each examination session.

METHODS

Search strategy PubMed (NLM, via the PubMed web interface); Korea Citation Index (KCI, National Research Foundation of Korea, via kci.go.kr); RISS (Research Information Sharing Service, KERIS, via riss.kr — journal articles and theses/dissertations). All from inception to the date of the final search. No language or date restriction.
Search terms combine (1) generative language models (ChatGPT, GPT-3.5, GPT-4, GPT-4o, o1, Claude, Gemini, Bard, Copilot, DeepSeek, "large language model", "generative artificial intelligence", and the Korean terms 챗GPT, 생성형 인공지능, 대형언어모델) with (2) Korean national licensing examinations (국가시험, 국가고시, 면허시험, "licensing examination", "Korean Medical Licensing Examination", and each KHPLEI profession). Because KCI and RISS support only short queries, profession-specific strings are run separately.
PubMed (5 strings), e.g.: ("large language model*"[tiab] OR ChatGPT[tiab] OR GPT-4[tiab] OR GPT-3.5[tiab] OR Gemini[tiab] OR Claude[tiab]) AND (Korea*[tiab]) AND (licens*[tiab] OR "national exam*"[tiab]); the remaining strings substitute profession terms (dental, pharmacist, nurse, Korean medicine, therapist, hygienist, technologist).
KCI (18 strings): ChatGPT 국가시험; GPT 국가고시; 인공지능 국가시험; 대형언어모델 국가시험; 생성형 인공지능 국가시험; ChatGPT 면허시험; and profession-specific strings (의사·치과의사·한의사·약사·간호사·치과위생사·작업치료사·물리치료사·방사선사·언어재활사 + ChatGPT).
RISS (24 journal strings and 5 thesis strings): the same Korean strings as KCI, restricted to article title/abstract fields, plus the thesis collection.

Participant or population Studies of generative LLMs tested on KHPLEI licensing examination items. The human reference population is the cohort of candidates who sat the same examination session(s); its session mean, score dispersion, and pass rate are taken from KHPLEI open data (item-analysis and applicant-status files, 2018–2025 sessions), not from the included studies.

Intervention Administration of examination items to a generative LLM (any model, version, prompt language, or prompting strategy) and the reported proportion of items answered correctly. Multiple models, sessions, or prompting conditions within one study are recorded as separate model–session records.

Comparator (a) The human candidate cohort of the same session(s): session mean total score, approximate SD, and observed pass rate from KHPLEI open data. (b) The fixed cut score: 60% of the maximum total score and 40% in every subject, applied to unequated raw scores.

Study designs to be included Primary evaluation studies of LLM performance on examination items (cross-sectional benchmarking studies), published as journal articles or full conference papers. No randomized or controlled designs are expected.

Eligibility criteria Inclusion: primary studies in which a generative large language model answered the items of a Korean national health personnel licensing examination (any KHPLEI examination) and reported an overall accuracy for a complete examination session or for all written subjects of that session; journal articles or full conference papers; any language; any year.
Exclusion: examinations of other countries; specialist board, in-training, or school-level examinations; evaluation of a single subject or a subset of sections only; studies in which the model generated or explained items without being scored; reviews, meta-analyses, editorials, letters, and correction notices (corrections are used only to correct the data of an included article); abstracts without a full paper; theses superseded by a journal article of the same study (the journal article is included instead).

Information sources PubMed (NLM, via the PubMed web interface); Korea Citation Index (KCI, National Research Foundation of Korea, via kci.go.kr); RISS (Research Information Sharing Service, KERIS, via riss.kr — journal articles and theses/dissertations). All from inception to the date of the final search. No language or date restriction.
Grey literature: theses and dissertations and full conference papers indexed in RISS; reference lists of included studies and relevant reviews; correction notices linked to included articles; contact with authors when a report or a session-level accuracy is unavailable.

Main outcome(s) The accuracy of each model on each examination session (percentage of items correct) and its position within the human cohort of that session, expressed as the standardized distance from the session mean, z = (accuracy – session mean)/SD, and the corresponding within-

cohort percentile; pass or fail under the fixed 60% cut; agreement between the verdict reported by the study and the verdict implied by the fixed cut; the difference between model accuracy and the human session mean.

Additional outcome(s) Subject-level accuracy where reported, with attention to the health-law subject and the 40% subject minimum; discordance of pass/fail verdicts for the same model across sessions; the correlation between model accuracy and the human session mean across records; study characteristics (model and version, number of items, prompt language, prompting strategy, test date, use of images); design and reporting quality (METRICS).

Data management Records exported from the three databases will be de-duplicated and screened by title and abstract, then by full text, by two reviewers independently, using a piloted screening form with predefined exclusion codes (S1 no LLM tested; S2 LLM study but no licensing-examination performance; S3 examination of another country; S4 unrelated; E1 single subject or section; E2 specialist or in-training examination; E3 review; E4 correction or letter; E5 not a health personnel examination; E6 grey literature not obtainable; E7 thesis superseded by journal article). Disagreements will be resolved by discussion, with a third reviewer if needed. Reasons for exclusion at full text will be reported for every record, and the process summarized in a PRISMA 2020 flow diagram.

Data will be extracted independently by two reviewers into a piloted structured form (a spreadsheet with fixed-choice fields; Microsoft Excel) covering bibliographic data; examination and profession; session(s) and year(s); items tested and total items; scope (whole examination, written subjects, text items only); model and version; prompt language and strategy; test date; accuracy per session and per subject; the authors' pass verdict; and correction notices. Values reported only graphically will be retained for description but excluded from the positional synthesis. Authors will be contacted when a report or session-level accuracy is unavailable.

Quality assessment / Risk of bias analysis Design and reporting quality will be appraised with the METRICS checklist for studies of generative AI-based models in health care education and practice (Sallam M, Barakat M, Sallam M. Interact J Med Res 2024;13:e54704). Each of the nine items (Model, Evaluation, Timing, Transparency, Range, Randomization, Individual factors, Count, Specificity of prompts and language) is scored 1 (suboptimal) to 5 (excellent); the study average (sum of applicable items / number of applicable items) is classified as excellent (4.21–5.00), very good (3.41–4.20), good (2.61–3.40), satisfactory (1.81–2.60), or suboptimal (1.00–1.80).

Two prespecified adaptations: Randomization is scored 5 when the complete item set of the session was tested; Individual factors is scored "not applicable" when scoring was fully objective against the official answer key.

Two reviewers will score independently and blind to each other; percent agreement and Cohen's kappa (unweighted and quadratic-weighted) will be reported; disagreements will be resolved by consensus. Quality will not be used to exclude studies. GRADE will not be applied because the outcome is a descriptive position rather than an intervention effect.

Strategy of data synthesis No pooled estimate of LLM accuracy will be computed, because models, versions, examinations, and item sets differ in ways that make one summary uninformative. Instead, a positional synthesis will relocate each model–session record within the human cohort of the same session.

For each session the mean total score (percent of maximum) and pass rate are taken from KHPLEI open data. The session SD is approximated from the mean item difficulty and a fixed mean inter-item correlation ($\rho = 0.33$, derived from published KR-20 reliabilities) and validated against empirical SDs where individual-level score data exist (Korean medicine doctor examination). For records spanning several sessions, a pooled SD combines the mean within-session variance and the variance of session means.

The percentile of each record is $\Phi(z)$ under a normal approximation, with empirical percentiles reported where individual scores exist. Results will be tabulated per record and displayed as (i) model positions superimposed on the human distributions of each examination and (ii) percentile against accuracy across all records. Pearson's r between model accuracy and the human session mean will be reported with a 95% CI. Subject-level accuracies will be compared with the 40% subject minimum. Analyses will use Python (pandas, SciPy, matplotlib); code and data will be deposited in a public repository.

Subgroup analysis Records will be described by examination (profession); model generation (GPT-3.5-class vs GPT-4-class and later; other developers); prompt language (Korean, English, both); prompting strategy (plain vs engineered); item scope (all items vs text-only); and presence of a health-law subject. Given the small number of

records, subgroups will be described, not tested inferentially.

Sensitivity analysis (1) Alternative values of the mean inter-item correlation (ρ = 0.25 and 0.40) for the SD approximation; (2) empirical rather than normal percentiles where individual-level scores exist; (3) exclusion of records based on text-only item subsets; (4) exclusion of the conference paper (grey literature).

Language restriction None. Studies in Korean and English are expected; other languages will be translated if found.

Country(ies) involved Republic of Korea.

Other relevant information Session-level human data are obtained from KHPLEI open data (aggregate item-analysis and applicant-status files) and involve no human participants; ethical approval is not required. Two related manuscripts by the same group on the fixed cut score of the Korean medicine doctor examination and of all KHPLEI examinations are under submission; they do not review LLM studies. Reporting will follow PRISMA 2020.

Keywords large language model; ChatGPT; licensing examination; cut score; standard setting; Korea; health personnel; METRICS.

Dissemination plans Publication in a peer-reviewed journal; deposit of the search strategies, screening log, extraction table, METRICS scores, session statistics, and analysis code in Harvard Dataverse.

Contributions of each author

Author 1 - Myeong Soo Lee.
Email: drmslee@gmail.com
Author 2 - Hye Won Lee.
Email: hwlee@kiom.re.kr