

TRIPOD-LLM checklist

Manuscript: Referring uncertain answers by decision-model confidence: Jev, OpenAI Decisions and Clef on Korean and US medical licensing examination questions. Sangzin Ahn.

Guideline: Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-69. doi:10.1038/s41591-024-03425-5. Item list taken from the published checklist (PMC12104976) on 30 September 2026.

Research design category. E, LLM evaluation. Public examination questions were used. No healthcare setting, patients or clinical records were involved, so H-only items are marked not applicable.

LLM task category. C, classification (single-best-answer multiple choice).

Locations: **MS** is the main manuscript and **S** the supplementary material. NA means not applicable to this design or task.

Abstract (TRIPOD-LLM for Abstracts)

Item	Topic	Reported	Location and note
2a	Title	Yes	Title names the use (referring uncertain answers by confidence), the three decision models and the data (Korean and US licensing examination questions)
2b	Background	Yes	Abstract, Background (referring uncertain answers to a clinician or a stronger model)
2c	Objectives	Yes	Abstract, Objective (evaluation only, no development or tuning)
2d	Study setting	Yes	Abstract, Methods (in silico study of public examination datasets)
2e	Data and splits	Yes	Abstract, Methods (all test questions). The referral share needs no labels. Development questions served pilots and request checks (S Supplementary methods)
2f	LLM name and version	Yes	Abstract, Methods (Jev, OpenAI Decisions, Clef, GPT-6 Luna, GPT-6.1 Sol). Jev version and all reference configurations in S Table 1
2g	Model-building steps	NA	No development, fine-tuning or alignment
2h	Task, inputs and outputs	Yes	Abstract, Objective (question and options in, option probabilities and selected answer out)
2i	Evaluation data, endpoint, measures	Yes	Abstract, Methods (official-key accuracy, error-detection AUROC, errors caught at a 20% referral share, accuracy after referral relative to random and oracle referral, option-order stability)
2j	Main results	Yes	Abstract, Results
2k	Implications	Yes	Abstract, Conclusions (no clinical safety claim)
2l	Registry	NA	H only. Not registered (MS Declarations)

Main text

Item	Topic	Reported	Location and note
1	Title	Yes	Title
2	Abstract	Yes	See the abstract table above
3a	Background and rationale	Yes	MS Introduction, paragraphs 1 to 3
3b	Target population, intended use and users	Yes	MS Introduction, paragraph 2 (a less expensive model answers first and refers uncertain answers to a clinician or a stronger model). The Conclusions propose cautious use as a first reader in well-bounded clinical decision tasks. No clinical deployment is proposed

Item	Topic	Reported	Location and note
4	Objectives	Yes	MS Introduction, final paragraph. MS Methods, Study design and datasets (evaluation only)
5a	Data sources	Yes	MS Methods, Study design and datasets. S Supplementary methods, Datasets (pinned KorMedMCQA and MedQA revisions). Published MedQA physician ratings [16] in MS Methods, Error analysis. No training or tuning data
5b	Data description, languages, countries	Yes	MS Methods (Korean doctor examination with five options, US MedQA with four options). S Supplementary methods, Datasets
5c	Dates of oldest and newest items	Partly	KorMedMCQA test items come from the 2022 to 2024 examinations (MS Methods). MedQA does not record examination dates. The dataset was published in 2021 [23]
5d	Preprocessing and quality checks	Yes	MS Methods (text, options and keys used as released, no translation). S Supplementary methods, Datasets (every English item cross-checked against a second source)
5e	Missing and imbalanced data	Yes	MS Methods, Outcomes and statistical analysis (invalid outputs and refusals counted as incorrect, probability analyses on valid answers with stated denominators). MS Results (refusals and invalid outputs). S Table 2. Question-level quality concerns are analyzed in MS Results, Where errors remained
6a	LLM name, version, training date	Partly	MS Methods, Decision models and Generative comparisons. S Table 1 (identifiers, available versions and collection dates). Training data cut-off dates are not reported
6b	Development details	NA	M and D only
6c	Prompting and inference settings	Yes	S Table 1 (reasoning settings and output allowances, sampling parameters at provider defaults). S Answering prompts
6d	Initial and post-processed output	Yes	MS Methods, Decision models (option probabilities and selected option, bounded normalization). S Probability validation and calibration
6e	Probability and threshold determination	Yes	MS Methods, Decision models and Referral (selected-option probability rather than the separate confidence field, compared in S Table 3, least confident 20% of each model's answers, curves over every possible share). S Supplementary methods, Referral, oracle and cost
7a	Generative output quality metrics	NA	Generative tasks only. Output is a classification scored against the official key
7b	Relevance of metrics to downstream use	Yes	MS Introduction and Methods, Referral (AUROC for error ranking, errors caught as the yield of a review workload, referral to a stronger model bracketed by random and oracle referral). No evaluation with real clinicians was performed
7c	Outcome definition and inference dates	Yes	MS Methods, Outcomes and statistical analysis (agreement with the official key). MS Methods, Study design and datasets (model requests between 17 September and 7 October 2026). S Table 1 (collection date of each model)
7d	Subjective assessment	Yes	Primary outcomes need no subjective judgement. Published physician ratings of MedQA were used descriptively [16]. Korean content labels were model assigned with blinded author review of a subset, and selected cases are illustrative rather than independently adjudicated outcomes (MS Methods, Error analysis, S Error analysis and S Table 9)
7e	Comparison with other models and benchmarks	Yes	MS Methods (three decision models, a letter-probability baseline from the model offered through one of them, and six reference models from three providers). MS Tables 1 to 3, S Tables 1 to 4. No comparison with human examinees
8a	Annotation process	Yes	Author review of 113 Korean content labels (S Error analysis). Existing physician ratings [16] were matched to every MedQA question with a predefined majority rule (MS Methods, Error analysis, and S Error analysis). Official keys were used unchanged
8b	Number of annotators, agreement	Yes	Three or four physicians per MedQA question in the published ratings (S Error analysis). A strict-majority rule summarized their judgements

Item	Topic	Reported	Location and note
8c	Annotator characteristics	Yes	US physicians in the published ratings [16]. Korean content labels and draft characterizations of selected errors by GPT-6.1 Sol without keys or model answers for the labels, with the labels checked against a blinded review by the author, a physician (MS Methods, S Supplementary methods and Declarations)
9a	Prompt design and selection	Yes	MS Methods (one fixed instruction per language, no prompt optimization on test outcomes). S Answering prompts
9b	Data used to develop prompts	Yes	MS Methods (one fixed instruction, no prompt tuning). S Supplementary methods, Datasets (development questions used only for pilot and request checks)
10	Summarization preprocessing	NA	Summarization tasks only
11	Instruction tuning or alignment	NA	M and D only
12	Compute	Yes	MS Tables 1 to 3 and Figure 3 (estimated cost per 1000 questions). Latency was recorded but not reported. Hardware is not applicable to hosted APIs
13	Ethical approval	Yes	MS Declarations (no human participants or patient data, IRB review not required)
14a	Funding and funder role	Yes	MS Declarations (NRF, MSIT, RS-2025-02214129, no funder role)
14b	Conflicts of interest	Yes	MS Declarations
14c	Protocol	Yes	H only, reported anyway. Protocols and analysis code are in the code repository (MS Declarations)
14d	Registration	Yes	H only, reported anyway. Not registered (MS Declarations)
14e	Data availability	Yes	MS Declarations (outputs public, question text retrieved from upstream releases under their licences, physician ratings from their original release)
14f	Code availability	Yes	MS Declarations (GitHub repository)
15	Patient and public involvement	Yes	H only, reported anyway. None (MS Declarations)
16a	Data flow	Yes	Question flow instead of patient flow. MS Results and S Table 2 (435 and 1273 planned questions with valid answers per model), S Table 6 (complete blocks in the order experiment)
16b	Characteristics by source	Yes	MS Methods and S Supplementary methods, Datasets (cohort sizes, option counts, examination years)
16c	Development vs evaluation distribution	Partly	Development and test sizes reported (S Supplementary methods, Datasets). Clinical variables are not applicable, and subject mix was not compared
16d	Numbers per analysis phase	Yes	MS Results and Tables 1 to 4 (denominators stated for every analysis)
17	Performance	Yes	MS Methods, Figure 1. MS Results, Tables 1 to 4, Figures 2 to 4. S Tables 2 to 10
18	Model updating	NA	No model updating was performed. Identifiers and available version information are reported in S Table 1
19a	Interpretation and fairness	Yes	MS Discussion, including comparison with prior work [6,13,14,15,16,17,25,26,27]. Fairness across patient groups is not applicable because no patient data were used. Differences by content area (Korean law and policy) are reported in MS Results and S Table 8
19b	Limitations	Yes	MS Discussion, limitations paragraph (analyses defined after earlier results, examination questions as a proxy with possible training exposure and historical keys, commercial services with undisclosed implementations and list-price costs, simulated referral)
19c	Known data challenges	Yes	MS Results and Discussion (physician-flagged MedQA questions [16], historical keys, Korean jurisdiction-specific legal content). S Tables 8 to 10
19d	Intended use and oversight	Yes	MS Discussion and Conclusions (examination-question evaluation only, human oversight still needed, no clinical use or autonomy supported)

Item	Topic	Reported	Location and note
19e	Poor-quality or unavailable input	Yes	MS Methods (invalid-output handling). MS Methods (questions used as released, including figure references without figures). MS Results and S Table 10 (sensitivity analysis without physician-flagged questions). S Table 9
19f	User interaction and expertise	NA	No user interaction was evaluated
19g	Future research	Yes	MS Discussion, limitations paragraph (prospective evaluation on new questions with clinicians reviewing a prespecified share of answers)